J Mol Model (2012) 18:1255–1262
DOI 10.1007/s00894-011-1155-8

ORIGINAL PAPER

# Density functional theory analysis of carbonyl sulfide hydrolysis: effect of solvation and nucleophile variation

**Ri-Guang Zhang · Li-Xia Ling · Bao-Jun Wang**

Received: 8 November 2010 / Accepted: 15 June 2011 / Published online: 7 July 2011


**Abstract** The detailed mechanisms of the hydrolysis of carbonyl sulfide (OCS) by nucleophilic water and hydroxide ion in both the gas phase and bulk water solvent have been investigated using density functional theory. Various reaction channels on the potential surface have been identified. The thermodynamic results demonstrate that the hydrolysis of OCS by nucleophilic water and hydroxide ion should proceed more favorably at low temperature. The hydrolysis of OCS by the hydroxide ion is the main reaction channel from thermodynamic and kinetic perspectives, and the bulk solvent can influence the rate-determining step in this channel. However, the solvent barely modifies the activation energy of the rate-determining step. For the hydrolysis of OCS by nucleophilic water, the solvent does not modify the rate-determining step, and the corresponding activation energy of the rate-determining step barely changes. This bulk solvent effect suggests that most of the contribution of the solvent is accounted for by considering one water molecule and a hydroxide ion.

**Keywords** Carbonyl sulfide · Hydrolysis · $OH^-$ · $H_2O$ · Thermodynamics · Kinetics · Density functional theory

## Introduction

Over the last two decades, it has become increasingly apparent that emissions of sulfur compounds into the atmosphere have been unacceptably high. Recently, increasingly stringent emission standards have been introduced around the world in order to reduce emissions of sulfur-containing compounds into the atmosphere [1]. Carbonyl sulfide (OCS), which is the by-product of many chemical processes involved in the conversion of fossil fuels, is thought to be one of the major components of organic sulfur compounds [2]. The hydrolysis of OCS is considered a promising process due to its mild reaction conditions and high conversion efficiency [3]. There have been numerous experimental studies of the hydrolysis mechanism of OCS [4–12]. The distinct channels for the hydrolysis of OCS by nucleophilic water and hydroxide ion can be represented stoichiometrically using the reactions below:

$$OCS + H_2O = H_2S + CO_2 \quad (1)$$

$$OCS + OH^- = SH^- + CO_2. \quad (2)$$

However, to gain a detailed understanding of the OCS hydrolysis mechanism, experimental data is not always sufficient; theoretical calculations can be helpful in order to clarify some essential questions. Quantum chemical methods have become useful tools for determining reaction mechanisms. With recent developments in theoretical methodology, density functional theory (DFT) is now capable of providing qualitative and quantitative insights into reaction mechanisms [13, 14]. While there have already been some theoretical studies of the hydrolysis mechanism of OCS, they have all modeled the reaction at metal oxide surfaces [15, 16]. To the best of our knowledge, few theoretical studies of the hydrolysis of OCS in pure water have been reported; such studies would be helpful for

R.-G. Zhang · L.-X. Ling · B.-J. Wang (✉)
Key Laboratory of Coal Science and Technology of Ministry of Education and Shanxi Province,
Taiyuan University of Technology,
Taiyuan 030024, China
e-mail: wangbaojun@tyut.edu.cn

deeply probing the mechanism of hydrolysis from the viewpoint of quantum chemistry. Deng et al. [3] have investigated the mechanism for hydrolysis reaction (1) in the presence of up to five water molecules, but, as we shall see, this reaction is somewhat different from those presented in this study. Meanwhile, reaction (2) was not mentioned in the studies of Deng et al. In fact, reaction (2) is central to OCS hydrolysis, and our present studies show that reaction (2) is the main reaction path in the OCS hydrolysis. Further, several theoretical studies of the elementary processes and reaction mechanisms associated with reaction (2) are often contradictory. For example, studies of the gas-phase reaction of OCS with OH by high-level ab initio characterization of the potential surface have shown that the production of $CO_2$ and SH is the main pathway to the destruction of OCS by OH radicals [17, 18]. However, studies by Hu et al. [19] into the gas-phase reaction of OCS with OH radicals using DFT and the G3 method indicated that the production of CO and SOH was the main pathway. Furthermore, there is no mention in the literature [17–19] of the effect of bulk water solvent on the hydrolysis of OCS.

In the present work, to complete a comprehensive theoretical study of the hydrolysis mechanism of OCS and to gain a good understanding of the hydrolysis of OCS, we carried out a detailed theoretical investigation of the hydrolysis of OCS by reactions (1) and (2) in the presence of explicitly described $OH^-$ and $H_2O$. In addition, the effect of the bulk water solvent on the reaction was also investigated using the conductor-like screening model (COSMO).

## Computational methods

The structures of all stationary points (reactants, intermediates, transition states and products) involved in the reaction pathways of OCS hydrolysis were fully optimized at the level of the general gradient approximation (GAA) using the Becke–Lee–Yang–Parr correlation functional (BLYP) [20, 21]. The double-numeric quality basis set with polarization functions (DNP) was used; the size of the DNP basis set is comparable to Gaussian 6-31 G**, but the DNP basis set is more accurate than the Gaussian basis set of the same size. Moreover, this basis set is known to yield only a small basis set superposition error (BSSE) [22]. All stationary points were identified as minima (number of imaginary frequencies NIMAG=0) or transition states (NIMAG=1). The zero-point energy (ZPE) of each stationary point was also determined. The reaction pathways were examined by performing TS confirmation on most of the transition state structures to confirm that they lead to the desired reactants and products [23]. All calculations are performed with Dmol$^3$ using the Materials Studio 4.4 software package on an HP Proliant DL 380 G5 server system [24, 25].

The effect of the bulk solvent was investigated using the conductor-like screening model (COSMO) as implemented in Dmol$^3$ [26, 27]. This is a dielectric continuum solvation model in which the mutual polarization of the solute and solvent is represented by screening charges on the surface of the solute cavity. These charges are derived under the simplified boundary condition that the electrostatic potential vanishes for a conductor ($\varepsilon = 0$), and the charges are scaled to account for the finite dielectric permittivity of a real solvent. In this case, bulk water solvent is represented by a dielectric permittivity $\varepsilon = 78.54$.

## Results and discussion

### Evaluation of the computational accuracy

To evaluate the reliability of the chosen level of theory, we calculated the bond lengths and bond dissociation energies (BDE) for several species involved in reactions (1) and (2) using the GGA-BLYP functional. The relevant results are listed in Table 1, where the available experimental and calculated values are also presented for comparison. It is clear that the GGA-BLYP functional can provide satisfactory results, considering that our main goals in the present work are to examine the detailed hydrolysis mechanism and to calculate the relative energies of the species involved, not to calculate accurate bond energies.

**Table 1** Theoretical and experimental bond lengths (Å) and bond dissociation energies (kJ mol$^{-1}$)

| Species | Calculated | | Experimental |
|---|---|---|---|
| | GGA-BLYP | GGA-BLYP [a] | |
| OC–S | 318.6 | 316.1 | 308.4 [b] |
| $R_{C–S}$ | 1.579 | 1.580 | 1.561 [c] |
| O–CS | 689.6 | 687.8 | 664.9±0.85 [d] |
| $R_{C–O}$ | 1.176 | 1.172 | 1.156 [c] |
| C–S | 701.0 | 699.6 | 712.2±0.85 [e] |
| O–C | 1079.3 | 1071.1 | 1076.4±0.67 [b] |
| HS–H | 384.1 | – | 381.4±0.50 [b] |
| HO–H | 484.1 | – | 497.1±0.29 [b] |

[a] From [28]
[b] From [29]
[c] From [30]
[d] From [31]
[e] From [32]

## Reaction mechanism

There are two possible reaction pathways for the hydrolysis of OCS (O=C=S) in reaction (1). In the first, $H_2O$ attacks across the C=O bond; in the other, $H_2O$ attacks across the C=S bond. Research by Deng et al. [3] indicated that the addition of $H_2O$ across the C=S bond is more favorable than its addition across the C=O bond. Thus, we will discuss the addition of an $H_2O$ molecule across the C=S bond below. The optimized geometries for the reactants, intermediates, transition states, and products in reactions (1) and (2) are shown in Fig. 1, and the imaginary frequency of each transition state in the gas phase and bulk water solvent are summarized in Table 2. All bond distances are in Å and bond angles are in degrees, respectively.

### *OCS hydrolysis in the presence of $H_2O$*

The reaction and activation energies for reaction (1) are listed in Table 3. The overall energetic profile is shown in Fig. 2. The nucleophilic attack of $H_2O$ on the C atom of OCS leads to the formation of the precoordination complex IM1. IM1 is the reactant-like intermediate in which the geometries of two reactant molecules remain almost the same as for the corresponding free molecules. IM1 is 8.8 kJ $mol^{-1}$ more stable than the original reactants OCS and $H_2O$. In the first step of the hydrolysis, the addition of $H_2O$ to OCS is completed upon full hydrogen (H6) transfer from $H_2O$ to S on OCS to form the thiocarbonic acid IM2 via a concerted transition state TS1 involving the cleavage of the O4–H6 bond and 1,3-hydrogen migration from O4 to S3. In TS1, the length of the O4–H6 bond that is broken is

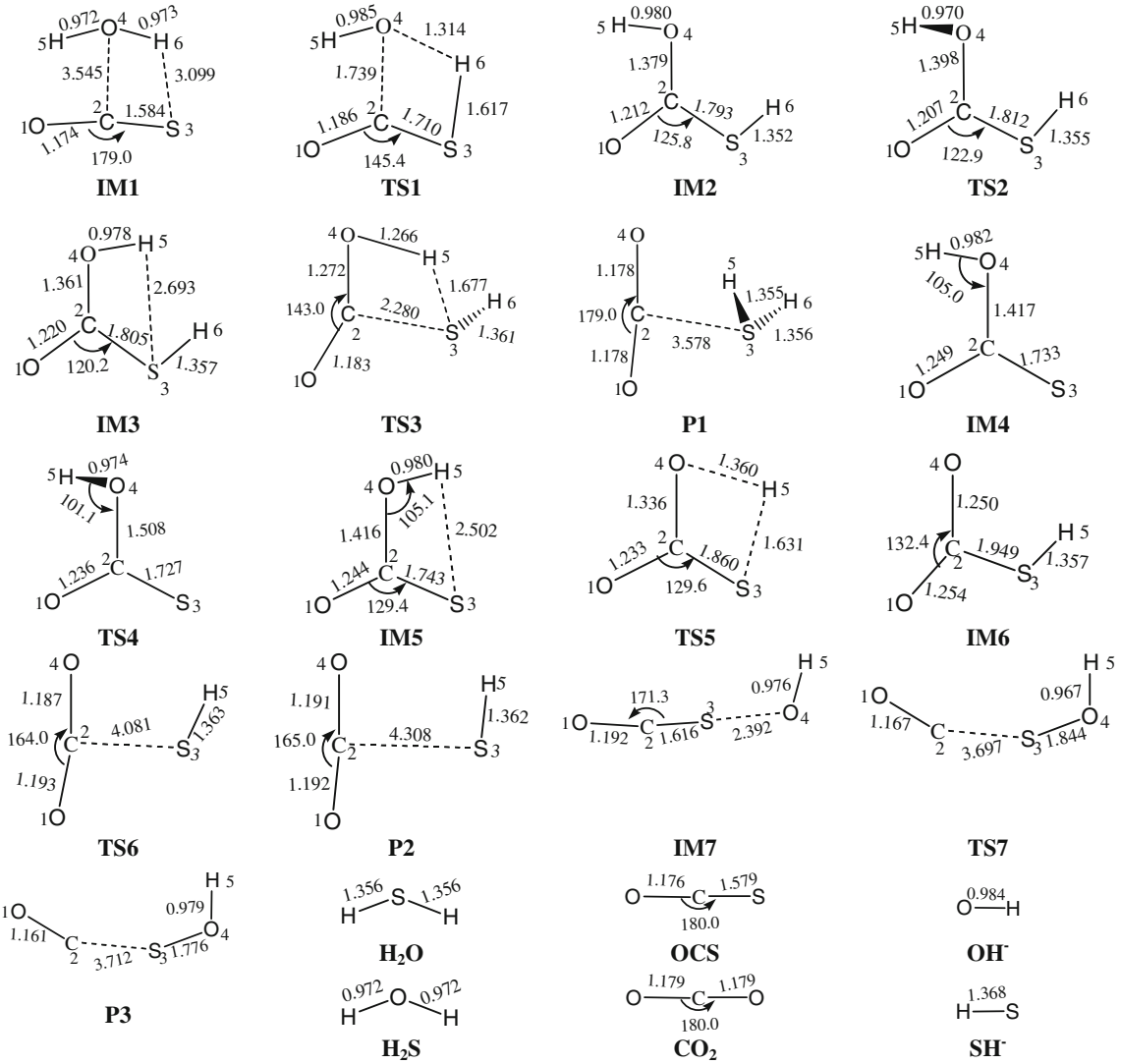



**Fig. 1** Optimized geometries for the reactants, intermediates, transition states, and products involved in reactions (1) and (2) (distances in Å and angles in degrees)

**Table 2** Imaginary frequency of each transition state in the gas phase and bulk water solvent

| Transition states | Imaginary frequency ($cm^{-1}$) | |
|---|---|---|
| | Gas phase | Bulk water solvent |
| TS1 | −1383.9i | −1499.45i |
| TS2 | −612.9i | −383.2i |
| TS3 | −1521.0i | −1563.5i |
| TS4 | −514.2i | −556.0i |
| TS5 | −1409.1i | −1477.7i |
| TS6 | −135.1i | −154.9i |
| TS7 | −211.1i | −307.6i |

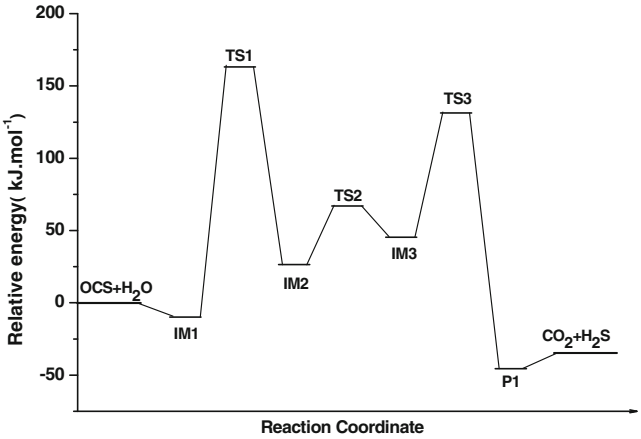



**Fig. 2** Reaction energy profiles for the attack of $H_2O$ on OCS

1.314 Å, and the angle O1–C2–S3 changes from 180.0° in OCS to 145.4°. The original linear OCS molecule is distorted. The H6–S3 and C2–O4 bonds that are formed have lengths of 1.617 and 1.739 Å, respectively. The C2–O4 bond shortens from 3.545 to 1.739 Å. The barrier height of TS1 relative to IM1 is 173.2 kJ $mol^{-1}$. Then, through a rotation along the O–H bond, IM2 is further converted to isomer IM3 via the transition state TS2. TS2 is 66.5 kJ $mol^{-1}$ higher in energy than the reactants and 40.2 kJ $mol^{-1}$ higher than the intermediate IM2. Finally, IM3 converts to P1, leading to the final products $CO_2$ and $H_2S$ via a concerted transition state TS3 involving C2–S3 bond cleavage and 1,3-hydrogen (H5) migration from O4 to S3. TS3 is predicated to be 86.5 kJ $mol^{-1}$ higher in energy than IM3.

According to the relative energies shown in Fig. 2, the activation barrier for TS1 is 173.2 kJ $mol^{-1}$, which is higher than those for TS2 and TS3 by 133.0 and 86.7 kJ $mol^{-1}$, respectively, indicating that the first step IM1→IM2 via TS1 is the rate-determining step (RDS) of reaction (1). Furthermore, Williams et al. [1] reported that at lower temperatures (303–333 K), the formation of thiocarbonic acid in the reaction of OCS with $H_2O$ was experimentally found to be the RDS, suggesting that our calculated results are consistent with experimental results. The reverse reaction—the attack of $H_2S$ on $CO_2$—appears to be similar to the attack of $H_2O$ on OCS, and the corresponding activation barrier for the RDS is 177.0 kJ $mol^{-1}$ (Fig. 2). Moreover, Fig. 2 also shows that the energy of (OCS + $H_2O$) is 26.7 kJ $mol^{-1}$ higher than that of ($CO_2$ + $H_2S$), indicating that the reaction of OCS with $H_2O$ is thermodynamically more favorable than the reverse reaction of $H_2S$ with $CO_2$.

The mechanism for the hydrolysis of OCS with $H_2O$ was found to be somewhat similar to that published earlier by Deng et al. [3]: the nucleophile $H_2O$ approaches OCS, leading to the adduct IM1, and IM1 can then form the intermediate IM2 by hydrogen transfer from $H_2O$ to S on OCS via a concerted transition state TS1, which is still the RDS. Then, IM2 evolves into another intermediate (IM3) and two transition states (TS2 and TS3), leading to the intermediate IM4, in which hydrogen transfer from O4 to S3 leads to the final product P1 via the transition state TS4. However, in our present studies, the calculated results show that the intermediate IM3 can directly form the final product P1 via a concerted transition state, and TS confirmation verifies that every transition state encountered in our studies is connected directly to the corresponding product and reactant. The activation barrier is 86.5 kJ $mol^{-1}$, which is much lower than the 126.1 kJ $mol^{-1}$ obtained by Deng et al. Therefore, we can confirm that the mechanism of reaction (1) in our studies has some improvements, and differs to some degree from that obtained in the studies of Deng et al.

### *OCS hydrolysis in the presence of $OH^-$*

The reaction mechanism of OCS with $OH^-$ is different from that with $H_2O$; it is somewhat analogous to that of OCS with

**Table 3** Activation and reaction energies for reaction (1) in the gas phase and bulk water solvent

| Elementary step | Gas phase | | Bulk water solvent | |
|---|---|---|---|---|
| | $E_a$ (kJ $mol^{-1}$) | $\Delta E$ (kJ $mol^{-1}$) | $E_a$ (kJ $mol^{-1}$) | $\Delta E$ (kJ $mol^{-1}$) |
| IM1→IM2 | 173.2 | 17.6 | 178.8 | 38.7 |
| IM2→IM3 | 40.2 | 19.7 | 29.3 | 4.6 |
| IM3→P1 | 86.5 | −90.5 | 110.7 | −63.5 |

the OH radical proposed by Danielache et al. [18], involving common isomerization and a hydrogen transfer step. However, the reaction process and intermediates are mostly different; in particular, the hydrogen thiocarbonate ($HSCO_2^-$) intermediate is not obtained, which is a key intermediate in experimental studies of OCS adsorbed on an alumina surface at basic hydroxyl groups [9, 10] and magnesium oxide with surface hydroxyls [33]. The reaction and activation energies related to reaction (2) are listed in Table 4.

Figure 3 shows an energy profile for the carbon-bonded adducts, as OCS is a polar molecule with high positive charge on the C atom, while there is a strong negative charge on the O atom in $OH^-$. Thus, the addition of $OH^-$ to the carbon gives the carbon-bond adduct IM4. The formation of IM4 is highly exothermic (by 268.5 kJ $mol^{-1}$) relative to the reactants OCS and $OH^-$. A transition state for the formation of IM4 from the reactants could not be located; all attempts to locate this species invariably converged to the structure of IM4. Our calculations therefore confirm that the attack of $OH^-$ on the C atom of OCS can directly produce the carbon-bonded adduct IM4.

Starting from IM4, an internal rotation of the OH group in IM4 results in the rotational isomer IM5 via the transition state TS4 with an activation energy of 32.3 kJ $mol^{-1}$. IM5 is planar with the OH oriented toward sulfur, which is 13.5 kJ $mol^{-1}$ lower in energy than IM4. TS4 is the only nonplanar structure; in this, the OH projects above the plane with a dihedral angle (H5–O4–C2–S3) of −97.9°, and the angle between OH and the C atom (H5–O4–C2) is 101.0°. Internal transfer of the hydrogen atom from the OH in IM5 to sulfur via a planar four-membered ring transition state TS5 where the hydrogen atom is shared between the oxygen and sulfur atoms leads to hydrogen thiocarbonate IM6 ($HSCO_2^-$), which is in agreement with previous experimental studies [9, 10]. In TS5, the O4–H5 and C2–S3 bonds elongate from 0.980 to 1.360 Å and from 1.743 to 1.860 Å, respectively. The C2–O4 and S3–H5 distances shorten from 1.416 to 1.336 Å and from 2.502 to 1.631 Å, respectively. This step has an activation barrier of 56.6 kJ $mol^{-1}$, and IM6 is 35.1 kJ $mol^{-1}$ more stable than IM5. Finally, a transition state, TS6, with C–S bond cleavage for the decomposition of IM6 to P2 was located. This step was found to be endothermic by 77.9 kJ $mol^{-1}$, with an activation energy of 73.1 kJ $mol^{-1}$. As shown in Fig. 3,

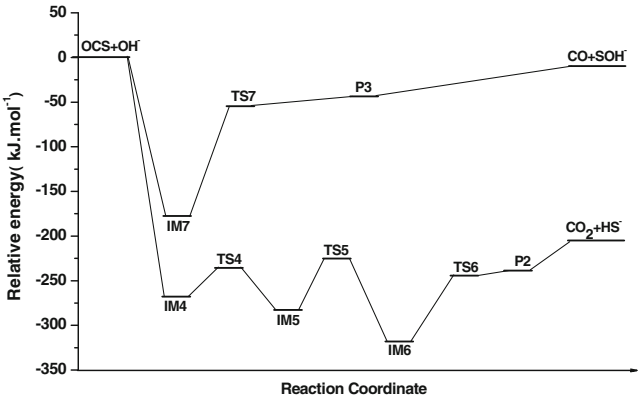



**Fig. 3** Reaction energy profiles for the attack of $OH^-$ on OCS

the activation energy for TS6 is 73.1 kJ $mol^{-1}$, which is higher than TS4 and TS5 by 40.8 and 16.5 kJ $mol^{-1}$, respectively, indicating that the IM6→P2 step via the transition state TS6 with C–S bond cleavage is the RDS for reaction (2). This is in agreement with the reported experimental results. For example, at a higher temperature (523 K), the reaction follows Langmuir–Hinshelwood kinetics, where the surface hydrolysis of hydrogen thiocarbonate with C–S bond cleavage is the RDS [1, 16].

In addition, Fig. 3 also shows an energy profile for the sulfur-bonded adduct. The addition of $OH^-$ to the sulfur of OCS directly gives sulfur-bond adduct IM7, and P3 is formed from IM7 via the transition TS7. IM7 is highly exothermic (by 176.9 kJ $mol^{-1}$) relative to the reactants OCS and $OH^-$. This reaction has an activation energy of 122.4 kJ $mol^{-1}$. Given the wide-scale rearrangement of atoms in this reaction, the formation of the sulfur-bonded adducts CO and $SOH^-$ is endothermic and not stable with respect to IM7. Thus, any collision sulfur-bond complex of OCS and $OH^-$ would be likely to decompose spontaneously back to IM7. Experiments by Leu and Smith [34] detected that SOH was only a minor product of the reaction of OCS and OH. The results calculated above show that $CO_2$ and $SH^-$ are the main products for reaction (2). Moreover, the experiment by Leu and Smith indicated that SH was formed as a major product of the reaction of OCS with OH at 517 K.

Based on the above analyses of the hydrolysis mechanisms of OCS, we can see that the RDSs of

**Table 4** Activation and reaction energies of reaction (2) in the gas phase and bulk water solvent

| Elementary step | Gas phase | | Bulk water solvent | |
|---|---|---|---|---|
| | $E_a$ (kJ $mol^{-1}$) | $\Delta E$ (kJ $mol^{-1}$) | $E_a$ (kJ $mol^{-1}$) | $\Delta E$ (kJ $mol^{-1}$) |
| IM4→IM5 | 32.3 | −13.5 | 32.8 | −6.3 |
| IM5→IM6 | 56.6 | −35.1 | 73.6 | −29.6 |
| IM6→P2 | 73.1 | 77.9 | 37.8 | 64.8 |

**Table 5** Effect of temperature on the Gibbs free energy changes $\Delta G$ and equilibrium constants $K$ for reactions (1) and (2)

| Reaction | | Temperature (K) | | | | |
|---|---|---|---|---|---|---|
| | | 298.15 | 300 | 400 | 500 | 600 |
| 1 | $\Delta G$ (kJ mol$^{-1}$) | −41.0 | −41.0 | −38.2 | −35.1 | −31.9 |
| | $K$ | $1.53\times10^{7}$ | $1.35\times10^{7}$ | $9.68\times10^{4}$ | $4.68\times10^{3}$ | $5.98\times10^{2}$ |
| 2 | $\Delta G$ (kJ mol$^{-1}$) | −198.0 | −198.0 | −198.0 | −198.4 | −199.3 |
| | $K$ | $4.91\times10^{34}$ | $3.00\times10^{34}$ | $7.12\times10^{25}$ | $5.39\times10^{20}$ | $2.26\times10^{17}$ |

reactions (1) and (2) are the elementary steps via TS1 and TS6, respectively. The corresponding activation energies for these steps are 173.2 and 73.1 kJ mol$^{-1}$, respectively. The activation energy of TS1 is much higher than that of TS6, suggesting that reaction (2) is more favorable in terms of dynamics than reaction (1). Therefore, it can be concluded that reaction (2) is the main reaction path for the hydrolysis of OCS.

## Effect of temperature on the thermodynamic quantities of the reactions

The total Gibbs free energy changes and equilibrium constants for reactions (1) and (2) calculated at different temperatures are listed in Table 5. Our study covers the temperature range 298.15–600 K.

From Table 5, it can be seen that reaction (2) is always the easiest at any temperature as it has the smallest $\Delta G$ at different temperatures. It always makes the greatest contribution to the hydrolysis of OCS, and reaction (1) always makes the next largest contribution; this sequence does not change with increasing temperature. In addition, for reactions (1) and (2), the calculated equilibrium constants decrease as the temperature increases, suggesting that OCS hydrolysis should proceed more favorably at low temperatures. Taking the reported experimental results [1, 16] into account, we can conclude that at lower temperatures, reactions (1) and (2) are responsible for the hydrolysis of OCS (as supported by thermodynamic quantities), and that reaction (2) is the main reaction. At higher temperatures, reaction (2) mainly contributes to the hydrolysis of OCS.

Using the calculated results for the above thermodynamic quantities, the specific mechanism of hydrolysis for OCS was elucidated in the temperature range 298.15–600 K, and this hydrolysis mechanism was found to be very close to that found experimentally. However, we cannot only consider the effect of temperature on the thermodynamic quantities associated with the hydrolysis of OCS; we must also study the effect of temperature on the kinetic quantities associated with this hydrolysis reaction.

## The rate constants for reactions (1) and (2)

As mentioned above, reaction (2) may be the main pathway for the hydrolysis of OCS. To understand the hydrolysis of OCS at different temperatures from a kinetics perspective, the rate constants for the RDSs of reactions (1) and (2) were calculated in the temperature range 298.15–600 K. The rate constants were obtained using Eyring's transition state theory (TST) [35, 36], and the relevant data are listed in Table 6.

According to the data in Table 6, the rate constant $k$ increases rapidly with the temperature, but the rate of increase drops at higher temperatures. The rate constant of reaction (2) is much larger than that of reaction (1) at the same temperature, and this does not change if the temperature is modified, which implies that reaction (2) is faster than reaction (1).

The results calculated above show that the hydrolysis of OCS in the presence of $OH^-$ is the main reaction pathway from thermodynamic and kinetic standpoints, and is responsible for the hydrolysis of OCS. $OH^-$ plays a critical role in the hydrolysis of OCS, as confirmed by the experimental observation that $OH^-$ triggers and encourages the hydrolysis of OCS [16, 37]. These calculated results provide us with a microscopic illustration of and theoretical guidance for the hydrolysis of OCS.

**Table 6** The calculated rate constants $k$ (cm$^3$ mol$^{-1}$ K$^{-1}$) for the reactions (1) and (2)

| Reaction | Temperature (K) | | | | |
|---|---|---|---|---|---|
| | 298.15 | 300 | 400 | 500 | 600 |
| 1 | $3.66\times10^{-19}$ | $5.50\times10^{-19}$ | $7.06\times10^{-12}$ | $1.32\times10^{-7}$ | $9.49\times10^{-5}$ |
| 2 | $3.72\times10^{3}$ | $4.40\times10^{3}$ | $3.78\times10^{6}$ | $2.20\times10^{8}$ | $3.28\times10^{9}$ |

### Analysis of the effect of temperature on the thermodynamic and kinetic results

Based on the above thermodynamic and kinetic results, we can conclude that the percentage conversion of the reactants is the major consideration for the thermodynamics of the hydrolysis reaction; a lower temperature favors the hydrolysis of OCS. The reaction velocity of the reactants is the major consideration for the kinetics of the reaction; a higher temperature can accelerate the hydrolysis of OCS. Generally, there is not necessarily a relationship between the reaction velocity and the percentage conversion of the reactant. However, we can be sure that the hydrolysis of OCS is influenced by the thermodynamics and kinetics of the reaction. Therefore, for the hydrolysis of OCS, the reaction temperature should lie in a reasonable range that will lead to not only a high conversion rate but also a fast reaction.

### Effect of the solvent

The literature [38–41] has shown that chemical characteristics and reactions can be influenced by the nature of the solvent used. In order to study the effect of the solvent, OCS hydrolysis via reactions (1) and (2) was examined in bulk water solvent using the COSMO model (dielectric constant $\varepsilon$=78.54), employing the geometries optimized in the gas phase. The reaction and activation energies after correcting for bulk solvent effects are listed in Tables 3 and 4.

As listed in the tables, for reaction (1), when the activation energy values of the RDS in the gas phase are compared with those obtained after COSMO treatment, it becomes apparent that the bulk solvent does not modify the RDS, and it barely modifies its activation energy. For example, the activation energy is 173.2 kJ $mol^{-1}$ in the gas phase but 178.8 kJ $mol^{-1}$ in bulk water solvent. This similarity in values probably arises because most of the solvent effect is already taken into account when just one water molecule is considered. Our calculated results are in agreement with the results of Deng et al. [3], who found that the bulk water solvent does not change the calculated energy barriers of the OCS hydrolysis reaction with five waters significantly when using the PCM solvation model. However, for reaction (2), the effect of the bulk solvent can change the RDS from the elementary step IM6→P2 in the gas phase to the elementary step IM5→IM6 in bulk water solvent. Again, however, the corresponding activation energies barely change: 73.1 kJ $mol^{-1}$ for IM6→P2 in the gas phase to 73.6 kJ $mol^{-1}$ for IM5→IM6 in bulk water solvent. The calculated results for reaction (2) suggest that the solvent can influence the RDS, but it still barely modifies the activation energy of the RDS. The above results on the effect of the solvent on the hydrolysis of OCS show that the effect of the bulk water solvent on the hydrolysis of OCS can be taken into account by simply making OCS react with a single water molecule or a hydroxyl ion.

## Conclusions

In this study, a comprehensive mechanistic investigation of the hydrolysis of OCS, including the reactions of OCS with $H_2O$ and $OH^-$, was undertaken by performing DFT calculations in both the gas phase and bulk water solvent. The effect of the bulk water solvent was taken into account using the COSMO model. The calculated results show that the reaction of OCS with $H_2O$ proceeds via an addition–elimination mechanism controlled by the formation of hydrogen thiocarbonic acid, as observed in the reported experiment. For the reaction of OCS with $OH^-$, the calculated results suggest that carbon-bonded adduct reaction is the main pathway for the destruction of OCS by $OH^-$. The reaction of OCS with $OH^-$ is mainly responsible for the hydrolysis of OCS. When the activation energies for the RDSs of reactions (1) and (2) in the gas phase were compared with those obtained after COSMO treatment, we found that the bulk solvent has a relatively small effect, probably because most of the solvent effect is already taken into account when we consider a water molecule or a hydroxyl ion, thus indicating that the hydrolysis of OCS can be considered to be the reaction of OCS with just a single water molecule or a hydroxyl ion.

Based on our results, which indicate that the reaction of OCS with $OH^-$ is the main contributor to the hydrolysis of OCS, we believe that in the hydrolysis of OCS under realistic reaction conditions on metal oxide surfaces such as $Al_2O_3$, the intrinsic properties of the $Al_2O_3$ and the pretreatment conditions will affect the hydroxyl coverage. Furthermore, varying the reaction conditions will change the partial pressure of water, thus affecting the hydroxyl coverage of the $Al_2O_3$ surface. As a result, higher coverage of hydroxyls on the $Al_2O_3$ surface will promote the hydrolysis of OCS, leading to the elimination and removal of environmental sulfur-containing pollutants like carbonyl sulfide as $CO_2$.

**Acknowledgments** This work was supported financially by the National Basic Research Program of China (no. 2005CB221203), the National Natural Science Foundation of China (no. 20976115 and 20776093) and the Younger Foundation of Shanxi Province (no. 2009021015).

## References

1. Williams BP, Young NC, West J, Rhodes C, Hutchings GJ (1999) Catal Today 49:99–104
2. Deng C, Wu XP, Sun XM, Ren Y, Sheng YH (2009) J Comput Chem 30:285–294

3. Deng C, Li QG, Ren Y, Wong NB, Chu SY, Zhu HJ (2008) J Comput Chem 29:466–480
4. Fiedorow R, Leaute R, Dalla LIG (1984) J Catal 85:339–348
5. Liang MS, Li CH, Guo HX, Xie KC (2002) Chin J Catal 23:357–362
6. Nnmba S, Shiba T (1968) Kogyo Kagaku Zasshi 71:93–96
7. Zhang YQ, Xiao ZB, Ma JX (2004) Appl Catal B Environ 48:57–63
8. Zhang QL, Guo HX (1988) Chin J Catal 9:14–24
9. He H, Liu JF, Mu YJ, Yu YB, Chen MX (2005) Environ Sci Technol 39:9637–9642
10. Liu JF, Yu YB, Mu YJ, He H (2006) J Phys Chem B 110:3225–3230
11. Wu HB, Wang X, Cheng JM (2004) Sci China Ser B 32:127–132
12. Wu HB, Wang X, Cheng JM, Yu HK, Xue HX, Pan XX, Hou HQ (2004) Chin Sci Bull 49:739–743
13. Dykstra C, Frenking G, Kim K, Scuseria G (2005) Theory and applications of computational chemistry. Elsevier, Amsterdam
14. Yang T, Wen XD, Huo CF, Li YW, Wang JG, Jiao HJ (2009) J Mol Catal A Chem 30:129–136
15. Aboulayt A, Maugé F, Hoggan PE, Lavalley JC (1996) Catal Lett 39:213–218
16. Hoggan PE, Aboulayt A, Pieplu A, Nortier P, Lavalley JC (1994) J Catal 149:300–306
17. Wilson C, Hirst DM (1995) J Chem Soc Faraday Trans 91:793–798
18. Danielache SO, Johnson MS, Nanbu S, Grage MML, McLinden C, Yoshida N (2008) Chem Phys Lett 450:214–220
19. Hu WH, Shen W, Xu JH (2006) J Sichuan Normal Univ (China) 29:336–339
20. Becke AD (1988) J Chem Phys 88:2547–2553
21. Lee C, Yang W, Parr RG (1988) Phys Rev B 37:785–789
22. Sauer J (1992) In: Catlow CRA (ed) Modeling of structure and reactivity in zeolites. Academic, London, pp 206–207
23. Zhang RG, Huang W, Wang BJ (2007) Chin J Catal 28:641–645
24. Delley B (1990) J Chem Phys 92:508–517
25. Delley B (2000) J Chem Phys 113:7756–7764
26. Klamt A, Schrmann G (1993) J Chem Soc Perkin Trans 2:799–805
27. Klamt A, Jonas V, Bürger T, Lohrenz JCW (1998) J Phys Chem A 102:5074–5085
28. Gao LG, Song XL (2007) J Mol Struc (THEOCHEM) 820:12–17
29. Luo YR (2005) Handbook of chemical bond dissociation energies. Science, Beijing
30. Lahaye JG, Vandenhaute R, Fayt A (1987) J Mol Spectrosc 123:48–83
31. Pedley JB, Naylor RD, Kirby SP (1986) Thermochemical data of organic compounds. Chapman and Hall, London
32. Prinslow DA, Armentrout PB (1991) J Chem Phys 94:3563–3567
33. Liu YC, He H, Xu WQ, Yu YB (2007) J Phys Chem A 111:4333–4339
34. Leu MT, Smith RH (1981) J Phys Chem 85:2570–2575
35. Fu XC, Shen WX, Yao TY (1990) Physical chemistry, 4th edn. Higher Education, Beijing
36. Wang BJ, Wei XY, Xie KC (2004) Chin J Chem Eng 55:569–574
37. Liu JF, Liu YC, Xue L, Yu YB, He H (2007) Acta Phys Chim Sin 23:997–1002
38. de Lima EF, de Carneiro MJW, Fenollar-Ferrer C, Miertus S, Zinoviev S, Om Tapanes NC, Aranda DAG (2010) Fuel 89:685–690
39. Tomasi J, Mennucci B, Cammi R (2005) Chem Rev 105:2999–3094
40. Sanna N, Chillemi G, Grandi A, Castelli S, Desideri A, Barone V (2005) J Am Chem Soc 127:15429–15436
41. Sebek J, Kejik Z, Bour P (2006) J Phys Chem A 110:4702–4711

J Mol Model (2012) 18:1263–1271
DOI 10.1007/s00894-011-1098-0

ORIGINAL PAPER

# A combined experimental and computational study on the material properties of shape memory polyurethane

**Cuili Zhang · Jinlian Hu · Fenglong Ji · Ying Fan · Yan Liu**

Received: 8 February 2011 / Accepted: 15 April 2011 / Published online: 7 July 2011


**Abstract** A type of shape memory polyurethane with 60 wt% hard segments (SMPU60) was prepared. Its material properties were tested by dynamic mechanical analysis (DMA) and Instron, and simulated using fully atomistic molecular dynamics (MD). The glass transition temperature ($T_g$) of SMPU60 determined by DMA is 316 K, which is slightly lower than that estimated through MD simulations ($T_g$=328 K) , showing the calculated $T_g$ is in good agreement with experimental data. A complex hydrogen bonding network was revealed with the calculation of radial distribution functions (RDFs). The C═O···H bond is the predominant hydrogen-bonding interaction. With increasing temperature, both the hydrogen bonding and the moduli decreased, and the dissociation of intermolecular hydrogen bonding induced the decrease of the moduli.

**Keywords** Polyurethane · Dynamic mechanical analysis · Fully atomistic molecular dynamics · Material property

## Introduction

Shape memory polymers (SMPs) can recover their original shape by being heated above their transition temperature, which is defined by different phases in the material. This property has attracted the attention of researchers from various fields over the past decade [1–3]. In particular, shape memory polyurethanes (SMPUs) have attracted a great deal of attention recently due to their unique properties, such as a wide range of shape recovery temperatures (from −30 to 70°C), high shape recoverability, good processing ability and excellent biocompatibility [3–5]. Polyurethanes are usually composed of polyether, or a polyester soft segment and diisocyanate-based hard segment. Because of the immiscibility between the hard segment and the soft segment, polyurethanes undergo microphase separation resulting in a hard-segment domain, soft-segment matrix, and a poorly characterized interphase. The hard-segment domains act as physical cross-links as well as filler particles for the soft-segment matrix. The primary driving force for domain formation is the strong intermolecular interaction of hydrogen bonding between the hard–hard segments of urethane and/or urea linkages [6].

Hydrogen bonding in polyurethanes has been studied extensively using infrared spectroscopy [7–11], which has proved a useful tool in characterizing hydrogen-bonding characteristics. Our group used density functional theory (DFT) to study the hydrogen bonding interactions on 4,4′-diphenylmethane diisocyanate (MDI)-based polyurethane in a previous study [12], and gained a deeper understanding of the hydrogen bonding of polyurethanes. Hydrogen bonding is evidenced by a frequency shift to values lower than that observed when these groups are free (i.e., not hydrogen bonded). The magnitude of the frequency shift is a measure of hydrogen–bonding strength. Various studies on the effect of temperature on hydrogen–bonding indicate that the average strength of the hydrogen bonds decreases with increasing temperature, which is observed as an infrared absorbance peak shifted

C. Zhang · J. Hu (✉) · F. Ji
Institute of Textiles and Clothing,
The Hong Kong Polytechnic University,
Hung Hom, Hong Kong, China
e-mail: tchujl@inet.polyu.edu.hk

Y. Fan · Y. Liu
State Key Laboratory of Polymer Materials Engineering,
College of Polymer Science and Engineering, Sichuan University,
Chengdu 610065, China

to a higher frequency. Meanwhile, the absorption coefficient decreases, leading to a reduction in the intensity of the band [7, 13]. Working on a series of MDI–based polyurethanes, Cooper et al. [14] concluded that the thermal behavior of hydrogen bonding is independent of morphology, and depends primarily on the glass transition temperature ($T_g$) of the hard segments. The $T_g$ is significant in determining the mechanical properties of amorphous polymers. Many physical properties change sharply during glass transition.

To obtain good performance of materials, and in order to guide experimental research, a few more experiments combined with theory are necessary to understand some important experimental processes and phenomena. In striving to attain the ultimate object, namely designing molecular structures according to property requirements, the goal is to reduce the number of experimental trials and to develop more accurate theories. To this end, computer simulation provides an attractive complementary tool or substitute for experimental studies. Towards developing better products, computer simulation can decrease the effort involved in synthesis and characterization, thus reducing energy requirements dramatically and largely avoiding environmental and economic impacts. Two popular arithmetics are the Monte Carlo simulation based bond–fluctuation model [15, 16] and the molecular dynamics (MD) simulation based bead–spring model [17, 18]. Both these are coarse-grained models that ignore detailed chemical structures, and therefore cannot be used to study the effect of the structure of a specific molecule on the properties of a given material. On the other hand, an atomistic level simulation shows great promise because it allows several practical problems to be explored in detail. Atomistic molecular simulations have proved very successful in the pharmaceutical industry [19, 20], and have also been applied successfully to predicting material properties throughout the last decade and become widely accepted as a powerful technique for understanding some of the properties exhibited by materials [21, 22]. Many such studies have been devoted to synthetic polymers of interest for industrial applications [23–26].

In the present work, a type of SMPU with 60% hard segment was prepared, which is composed of polycaprolactone (Mw=400, PCL400) as soft segments, MDI and 1,4–butanediol (BDO) as hard segments, and its thermo and mechanical properties were investigated by dynamic mechanical analysis (DMA) and Instron. At the same time, a simple corresponding fully atomistic model with the mole ratio of 1:1.78:0.78 for PCL400, MDI and BDO was built, and the physical and mechanical properties calculated from MD simulations were compared with experimental data to explore the relationship between the different properties of this material.

## Methodology

### Experimental

#### *Materials*

SMPU60 was synthesized with PCL diol (Mw=400, PCL400, Solvay; http://www.solvaychemicals.com) as soft segments, MDI (Acros; http://www.acros.com) and BDO (International Laboratory; http://online1.ispcorp.com) as hard segments. The PCL400 was dried and dehydrated at 80°C under vacuum for 6 h prior to use. The solvent N, N′–dimethylformamide (DMF) from Aldrich was dried with 4 Å molecular sieves.

#### *Synthesis and film preparation*

The isocynate–terminated prepolymers were first prepared by reacting the dried PCL400 with the mole excessive MDI to ensure complete reaction in the DMF solution at 65°C for 3 h. The prepolymers were then chain extended with BDO at 80°C for another 4 h. The feed ratio (mole) for the three kinds of monomers was 1:1.78:0.78. SMPU60 was then prepared.

#### *Characterization*

The $T_g$ and storage modulus of SMPU60 were obtained by DMA. Tests were carried out on a Perkin-Elmer Diamond dynamic mechanical analyzer in the tensile mode, at a heating rate of 3°C $min^{-1}$ from −56 to +140°C, an oscillation amplitude of 5.0 μm and a fixed frequency of 1 Hz. The dimensions of the SMPU60 DMA sample were 10 mm in length, 5 mm in width and 0.163 mm in thickness.

Tensile tests were performed according to ASTM-D 882–09 specification with a loading rate of 2 mm $min^{-1}$, using a tensile tester Instron 5566 accompanied by a temperature constant chamber at 293 K, 313 K and 333 K, respectively. A clip type extensometer was used to measure strain within the specimen gauge length. Properties of tensile strength and tensile modulus were calculated after testing. The samples were 20 mm in length, 5 mm in width and 0.163 mm in thickness.

### Simulation

The chemical structures of the monomers and the corresponding repeat units of SMPU60 are shown in Figs. 1 and 2, respectively. In this simulation, the full atomic model was adopted without the approximation of united atoms. It has been found by Fan [27] that cell size has no evident effect on the calculated properties.

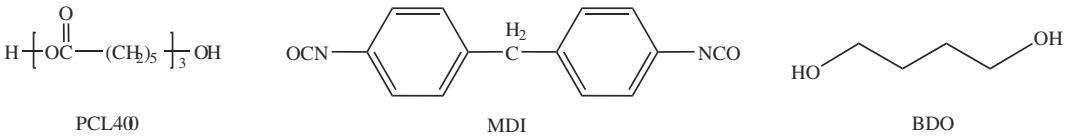



**Fig. 1** Chemical structure of monomers

Therefore, to reduce the computational requirement, a moderate size of cell was used in the following simulation. Based on a mole ratio of 1:1.78:0.78 for PCL400, MDI and BDO of SMPU60 (60.0 wt % hard segments) in the present design, three random SMPU60 chains each with 8 PCL400, 14 MDI and 6 BDO repeating units were generated according to a procedure implemented in Materials Studio's [28] random copolymer builder program by the stochastic choice of torsion angles. For molecular simulation of the polymer system, the periodic boundary condition with an amorphous cell was applied. A total of 2,856 atoms was assembled in a cubic cell, with a density equal to the experimental value of 1.10 g $cm^{-3}$ at 298 K (see Fig. 3). The system size of 28 repeat units is not enough to represent conformations of a real polymer chain. However, previous works [23, 29] using 10–15 repeat units for simulation have reported reasonable results.

The potential energy of the system was first minimized by the use of the Smart Minimer method, and then equilibrated using NVT ensemble (constant number of particles, volume and temperature) at 1,000 K for 1,000 ps. A constant–temperature, constant–pressure MD simulation (NPT) MD relaxation at 500 K lasting 1,000 ps was then performed. The pressure (0.0001 GPa) was controlled according to the Parrinello-Rahman [30] algorithm, while the temperature was imposed by Nosé [31] method. This high-temperature relaxation procedure was suggested by Lee and Mattice [32]. The following NPT MD simulations were performed from 400 K to 250 K for 1,000 ps at each temperature with an interval of 10 K. The final structure from each simulation was used as the starting structure for the next simulation. Cooling technique is preferred to heating technique for polymer systems comprising large molecules as relaxation phenomena are faster at higher temperatures [33]. If a heating approach is used to determine $T_g$ instead, much more time is needed to equilibrate the system before heating, and similar values are obtained as in the cooling approach with sufficient relaxations, or with some hysteresis due to insufficient relaxation as shown in previous work by Ding et al. [34]. The temperature was decreased rather than increased in an attempt to smooth out the effects of high energy transitions. Each trajectory was sampled every 100 fs, and the last 5,000 structures were subjected to analysis.

## Forcefield

To construct the initial molecular structures and implement all ensemble simulations and post-processes, the commercial molecular simulation program Materials Studio 4.0 [28] was used; all calculations were performed using ab initio condensed-phase optimized molecular potentials for atomistic simulation studies (COMPASS) forcefield [35], which have been employed to study compounds containing elements C, N, O, H [36–38]. COMPASS forcefield was used to calculate the properties of the polymer material. This forcefield describes intramolecular and intermolecular interactions in chemical systems (see Eq. 1).

Here, the first four terms are bond stretching, bending, torsion, and out-of-plane potentials, known as the valence term. The next six terms are the cross terms describing the interactions between each valence term, and the last two terms are the non-bonded coulomb and van der Waals potentials. In calculating the non-bonded potentials, the atom-based summation with a cutoff radius of 9.5 Å was used in the van der Waals summation with long range correction. Electrostatic interaction by Coulomb potential

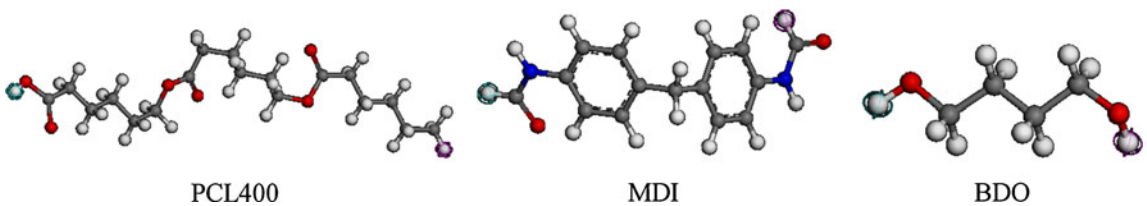



**Fig. 2** Repeat units for shape memory polyurethane with 60 wt% hard segments (SMPU60). *PCL400* polycaprolactone (Mw=400), *MDI* 4, 4′-diphenylmethane diisocyanate, *BDO* 1,4–butanediol

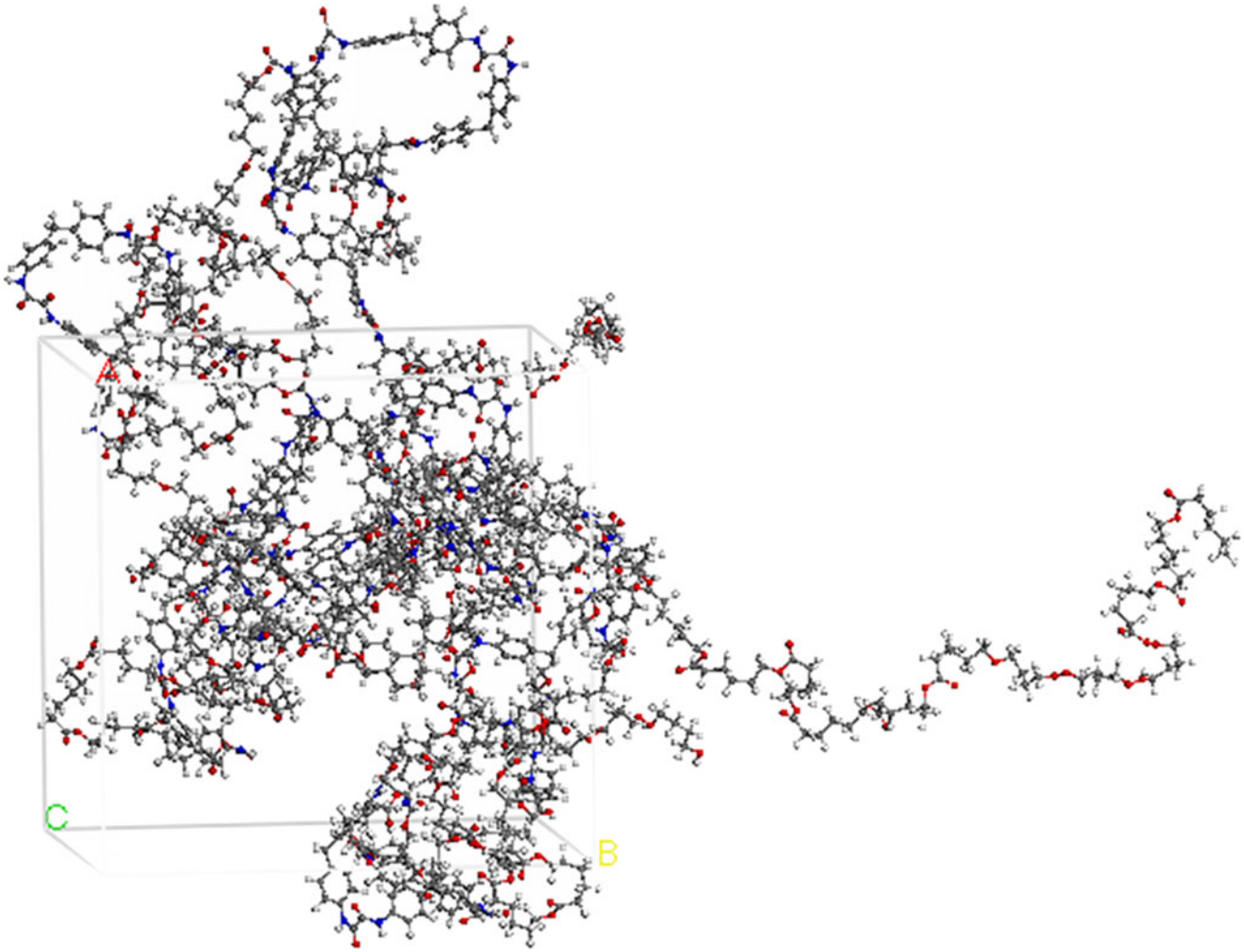



**Fig. 3** Amorphous cell for the SMPU60 with three polymer chains (*red* O, *blue* N, *gray* C, *white* H)

was calculated using the Ewald summation method, despite requiring excessive computing cost.

$$
\begin{aligned}
E_{total} = & \sum_{b}\left[k_2(b-b_0)^2 + k_3(b-b_0)^3 + k_4(b-b_0)^4\right] \\
& + \sum_{\theta}\left[k_2(\theta-\theta_0)^2 + k_3(\theta-\theta_0)^3 + k_4(\theta-\theta_0)^4\right] \\
& + \sum_{\phi}\left[k_1(1-\cos\phi) + k_2(1-\cos 2\phi) + k_3(1-\cos 3\phi)\right] \\
& + \sum_{\chi} k_2\chi^2 + \sum_{b,b'} k(b-b_0)\left(b'-b'_0\right) + \sum_{b,\theta} k(b-b_0)\left(\theta-\theta'_0\right) \\
& + \sum_{b,\phi} k(b-b_0)[k_1\cos\phi + k_2\cos 2\phi + k_3\cos 3\phi] \\
& + \sum_{\theta,\phi} (\theta-\theta_0)[k_1\cos\phi + k_2\cos 2\phi + k_3\cos 3\phi] \\
& + \sum_{b,\theta} k\left(\theta'-\theta'_0\right)(\theta-\theta_0) + \sum_{\theta,\theta',\phi} k\left(\theta'-\theta'_0\right)(\theta-\theta_0)\cos\phi \\
& + \sum_{i,j}\frac{q_i q_j}{r_{ij}} + \sum_{i,j}\varepsilon_{ij}\left[2\left(\frac{r^0_{ij}}{r_{ij}}\right)^9 - 3\left(\frac{r^0_{ij}}{r_{ij}}\right)^6\right]
\end{aligned}
\tag{1}
$$

## Method for calculating elastic constants

In the MD simulations, the mechanical properties of all atomistic unit cells were calculated using the Parrinello-Rahman fluctuation method [30, 39], where uniform external stress is applied to the unit cell and the deviations of the resultant strain values at each simulation step are monitored. All the components of the stiffness matrix can be obtained from the fluctuation formulation given below. During the NσT (constant stress) ensemble, the ensemble average of the strain perturbations is determined from the deformation tensor and the metric tensor, and the stiffness matrix of the unit cell can be calculated as shown below

$$\sigma_{ij} = C_{ijkl}\varepsilon_{kl} \tag{2}$$

$$C_{ijkl} = \frac{kT}{\langle V\rangle}\left\langle \delta\varepsilon_{ij}\delta\varepsilon_{kl}\right\rangle^{-1} \tag{3}$$

where $k$ is Boltzmann's constant, $T$ is the temperature during the ensemble and<V> is the ensemble average of the cell volume.

Based on the stress–strain relationship, the elastic stiffness coefficient can be obtained. For isotropic material, stress–strain behavior can be described by only two independent coefficients. Lame coefficients λ and $\mu$, which are the same as that described at a macro-scale level, can be calculated from the stiffness matrix related to

$$\lambda = \frac{1}{3}(C_{11} + C_{22} + C_{33}) - \frac{2}{3}(C_{44} + C_{55} + C_{66})$$
$$\mu = \frac{1}{3}(C_{44} + C_{55} + C_{66}) \quad (4)$$

Once the two lame constants have been obtained, several related mechanical properties of material can be derived:

$$E = \mu\left(\frac{3\lambda + 2\mu}{\lambda + \mu}\right) \quad G = \mu \quad (5)$$

where $E$ and $G$ represent Young's modulus and shear modulus, respectively.

## Results and discussion

The glass transition temperature ($T_g$) and moduli of SMPU60 were obtained from both experiments and MD simulations. The hydrogen bonding interactions were examined by radical distribution functions (RDFs). It is necessary to ensure that the conformation of the system is stable before calculating any static properties. The equilibration procedure described above was applied to the SMPU60 model.

### Glass transition temperature

A DMA thermogram of the SMPU60 is shown in Fig. 4. The loss tangent (tan δ) peak that emerges in the vicinity of 316 K is ascribed to the $T_g$ of SMPU60, where the polyurethane passes from a glassy to a rubbery state.

The specific volume and density data from MD simulations in the trajectory files were collected when the density and energy of the system reached a balanced value. Specific volume is mostly used to determine the value of $T_g$ [40–42]. The curves of the specific volume as a function of temperature are shown in Fig. 5. A steady increase in the specific volume with increasing temperature and the simulated $T_g$ is found at the point where the intersection of slopes occurs, which is about 328 K.

Density of SMPU60 at each temperature was calculated from the average specific volume of the system, as shown in Fig. 6. A steady decrease in density with decreasing temperature, and a clear change in the slope of the density curve is observed. The change in the slope of the density curve defines the value of $T_g$ at approximately 328 K.

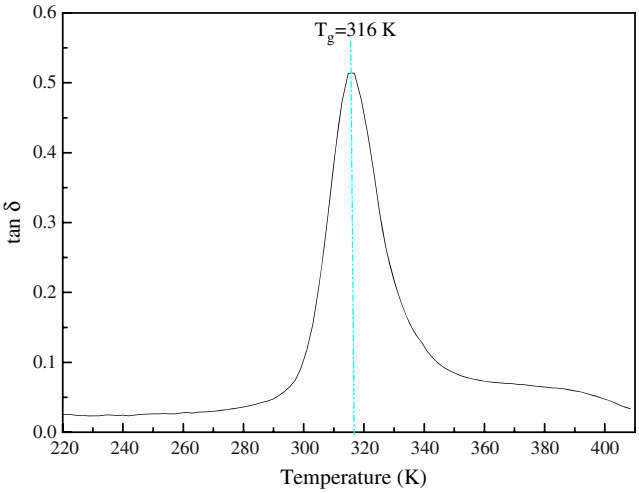



**Fig. 4** Dynamic mechanical analysis (DMA) thermogram of SMPU60

The assigned cooling rate was observed to affect the predicted value from MD simulations [43, 44]. The mobility of the polymer chains decreases rapidly when the polymer melt is cooled below $T_g$, and a sufficiently long time is needed to reach the equilibrium structure. A higher cooling rate in the simulation can increase departure from the equilibrium state, which results in a higher value of $T_g$ in MD simulations. The cooling rate chosen in present study was 10 K/1,000 ps, a relatively slow cooling rate attainable based on the scale of experimental $T_g$ and the computational time for the system, so that the predicted value of $T_g$ from MD simulation is expected to be very close to the experimental value within 15 K of the actual experimental value, providing confidence in the prediction of $T_g$ of other polymers. It should be noted

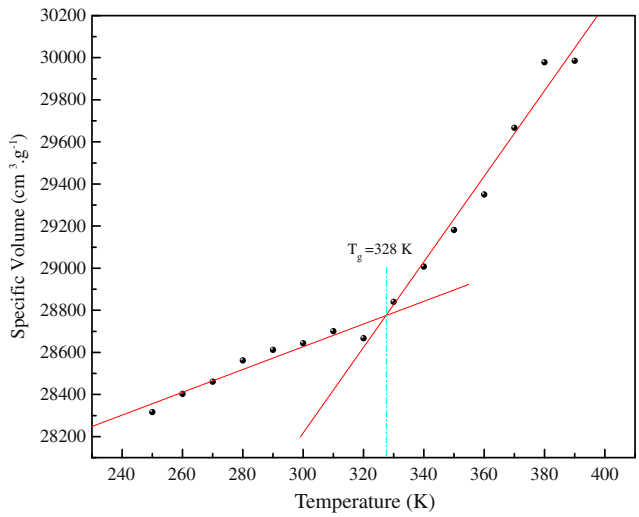



**Fig. 5** Molecular dynamics (MD) curves of the specific volume vs temperature for SMPU60

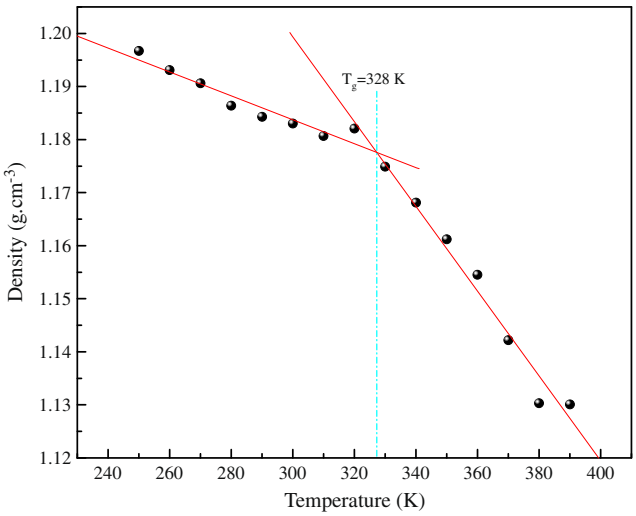



**Fig. 6** MD curves of density vs temperature for SMPU60

that the simulated values represent $T_g$ of a 'perfect' polymer and the model does not take into account any defects or areas of voids that might yield a $T_g$ at variance with the simulation. Nevertheless, accurate results can be achieved with relatively simple models such as this present work.

## Hydrogen bonding

For the polyurethane SMPU60 containing one hydrogen–bonding donor amidogen hydrogen (N–H) and three possible acceptors (carbonyl oxygen (C═O), ester oxygen (C–O–), and amidogen nitrogen (N–H), it is widely speculated that there are three types of hydrogen–bonding interactions in the SMPU60. Since the forcefield COMPASS was used, where no hydrogen bonding energy terms are explicitly included and hydrogen bonding interactions were implicitly considered, this method is not applicable to our system. On the contrary, we examined these interactions by comparing the radical distribution functions (RDFs) [45] of polar atom pairs at four different temperatures (250 K, 300 K, 350 K and 400 K). The intermolecular and intramolecular RDFs between carbonyl O atoms or ester O atoms or N atoms and H atoms on the N–H were calculated by averaging the last 500 ps of the MD trajectory, respectively. The corresponding RDFs are depicted in Figs. 7, 8 and 9.

The intermolecular and intramolecular RDFs of carbonyl O atoms and H atoms on the N–H at four different temperatures are shown in Figs. 7a, 7b, respectively. It can be seen that there is one main peak within 3.0 Å (generally, hydrogen bonds are generally less than 3.0 Å in length [46]). For the intermolecular RDFs, the peak at around 2.0 Å comes from hydrogen bonds between C═O and N–H, while, for the intramolecular RDFs, a much

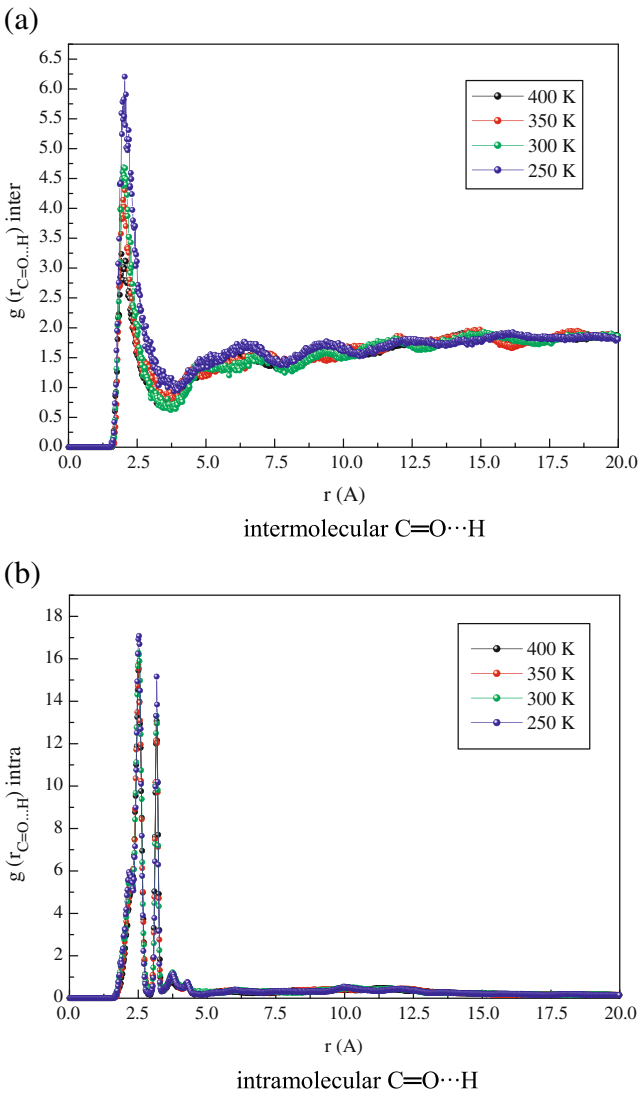



**Fig. 7** Radial distribution functions (RDFs) for the hydrogen bonding between carbonyl O atoms and H atoms at four temperatures

higher peak for hydrogen bonding is located at 2.5 Å, which may be due to the proximity of H atoms on N–H and carbonyl O atoms in one same urethane unit. Both the intermolecular and intramolecular hydrogen bonding RDFs decrease with increasing temperature, and the variation for the intermolecular hydrogen bonding is more obvious, which indicates that the intermolecular hydrogen bonding is more sensitive to temperature.

The RDFs for ester O atoms and H atoms on N–H at four different temperatures are given in Fig. 8. At a distance close to 2.1 Å, we can find a small peak with an intensity lower than 0.2 arising from intermolecular hydrogen bonding (see Fig. 8a). For the intramolecular RDFs, a peak around 2.3 Å with the intensity about 2.0 is assigned to the intramolecular hydrogen bonding (see Fig. 8b). Similar trends can be found as in the carbonyl hydrogen bonding, with the intramolecular hydrogen bonds being much more

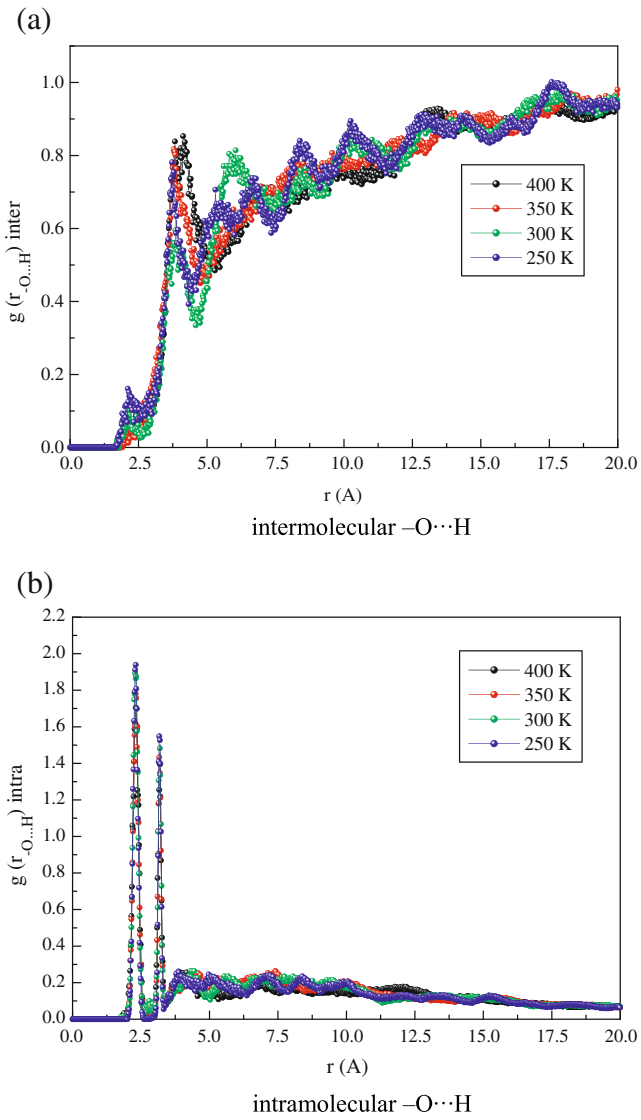



**Fig. 8** RDFs for the hydrogen bonding between ester O atoms and H atoms at four temperatures

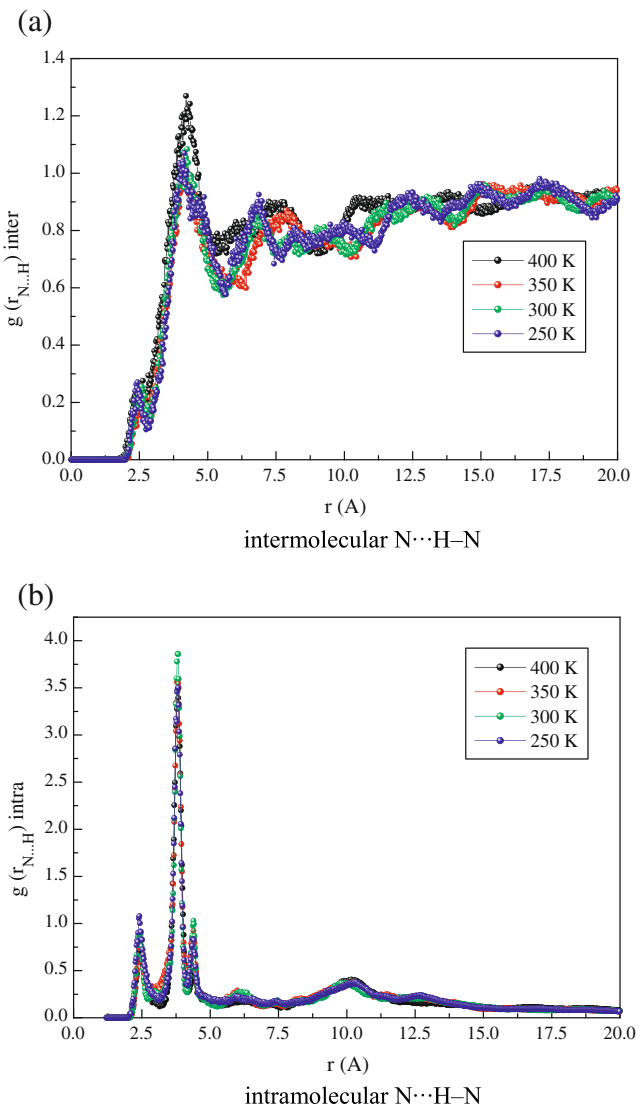



**Fig. 9** RDFs for the hydrogen bonding between N atoms and H atoms at four temperatures

than the intermolecular ones, and hydrogen bonding decreasing with increasing temperature.

Compared with the carbonyl hydrogen bonding, hydrogen bonding between ester O and H is negligible and can be ignored. According to [12], the binding energy for the carbonyl hydrogen bond (C═O···H) is twice as big as that of the ester hydrogen bond (−O···H), which may be the main reason for the large difference in the amounts of the two kinds of hydrogen bonds.

The RDFs for N atoms and H atoms on the N–H at four temperatures are shown in Fig. 9. For the intermolecular RDFs, the peak around 2.4 Å with an intensity of ~0.3 corresponds to N···H–N hydrogen bonds. The distance is longer than that of the hydrogen bonds between O and H atoms, which indicates that the bond distance for N···H is longer than that of O···H. For the intramolecular RDFs, the hydrogen bond peak can also be found at near 2.4 Å, and the intensity is about 1.0, clearly lower than that of the intramolecular ester hydrogen bonds (−O···H) (the intensity is nearly 2.0), indicating that intramolecular hydrogen bonds for the ester O form much more easily. Due to the chain connectivity and the steric hindrance of the bulky phenyl groups in the amorphous phase, the intermolecular approach for a hydrogen atom to a nitrogen atom experiences more difficulty than its approach to an oxygen atom. The intensities of the hydrogen bonds (N···H–N) at higher temperatures are also generally lower than those at lower temperatures.

The peak for the hydrogen bond N···H–N, just as that of the –O···H, is also much lower than that of C═O···H, indicating that the hydrogen bond of C═O···H is the predominant hydrogen–bonding interaction, and the contri-

bution of the hydrogen bonds of –O···H and N···H–N is probably small. From the RDF analysis we can investigate the hydrogen bonding of polymer materials in microcosmic scale easily, which can help us understand hydrogen bonding distribution clearly.

## Mechanical properties of the models

This section presents the results for the modulus of SMPU60 as a function of temperature. The moduli is expected to decrease with increasing temperature ($T$), because, as we have shown above, a lot of hydrogen bonds, especially intermolecular C═O···H hydrogen bonds, dissociate when the temperature increases, while the high values of polyurethane modulus is reinforced through hydrogen bonding.

Figure 10 shows the tensile behavior of SMPU60 at three different temperatures. We found that the stress decreases dramatically with higher temperature at the same strain. From the stress–strain curves, we can obtain one part of Fig. 11—the Young's modulus of SMPU60 at three temperatures. The other two parts of Fig. 11 are the simulation Young's modulus and DMA storage modulus as a function of temperature.

All the moduli including the Young's modulus and the storage modulus decline with increasing temperature, and the trends of the simulation results are consistent with the experiment. For Young's modulus by simulations and storage modulus by DMA test (see Fig. 11), the decrease is slight below 290 K, while it becomes more obvious between 290 K and 340 K, and after becoming slight again, the modulus finally reaches a balanced value above 360 K. Comparing the Young's modulus obtained by tensile test and MD simulations, the difference in value is relatively large. The calculated values are much higher than those

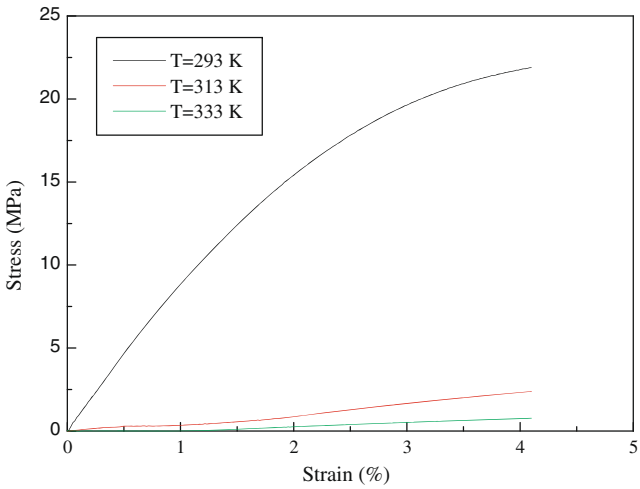



**Fig. 10** Tensile behavior of SMPU60 at different temperatures

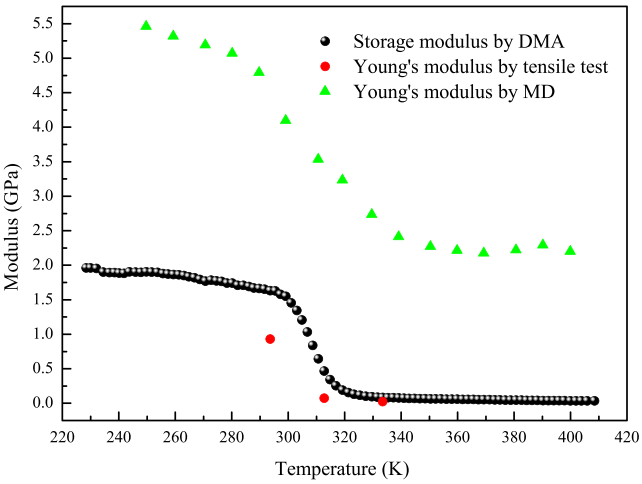



**Fig. 11** Mechanical properties as a function of temperature $T$, obtained from simulations and experiment

obtained by experiments, with more than 2 GPa at the same temperature. As discussed before, one reason for this discrepancy is that the simulated system represents ideal structures, whereas the experimental sample may have voids or defects in the structures. A rigorous analysis of the structure of SMPU60 within the matrix might yield more insight, but might also be limited by the statistical sampling inherent in atomistic simulation. This will require further analysis, but is beyond the scope of this paper. From these results, COMPASS reproduced well experimental measurements. Although more accurate values can be obtained by averaging several different structures [47], no repeats were done in this work for the purposes of validation. Good agreement between the calculated elastic constants and the experimental values confirmed that the models are representative.

## Conclusions

This paper presents a study of a type of shape memory polyurethane with 60 wt% hard segments (SMPU60) prepared from PCL400, MDI and BDO, and its material properties were studied both experimentally and by fully atomistic MD simulations.

The $T_g$ from the loss tangent (tan δ) peak tested by DMA is a little lower (by only 12 K) than that of the characteristic change in slope from plots of specific volume vs temperature and density vs temperature. The hydrogen bonding interactions were examined by comparing RDFs between carbonyl O atoms or ester O atoms or N atoms and H atoms on the N–H. Due to the correlation hole effect, the intramolecular contributions overwhelm intermolecular contributions for hydrogen bonding. The hydrogen bonding is dissociated with increasing temperature, with intermo-

lecular hydrogen bonding decreasing especially sharply. There are three kinds of hydrogen bonds (C═O···H, –O···H and N···H–N) in SMPU60, while the hydrogen bond C═O···H is the predominant hydrogen–bonding interaction, while the probability of hydrogen bonds –O···H and N···H–N is very low, which results in the high modulus of SMPU60. The moduli were found to decrease consistently with increasing temperature arising from the sharp decrease in hydrogen bonds. The calculated Young's moduli were generally higher than the experimental values due to differences between the simulated system representing ideal structures and the real sample that may contain defects. In spite of the limited size of the models, these results prove reasonable insight into understanding these experimental processes and phenomena. In summary, this work has demonstrated the usefulness of MD simulations as a tool to understand the structure-mechanical property relationships of polyurethane and give us the confidence to study other polymer systems using this computational method.

**Acknowledgments** The authors wish to express their gratitude for the support of a Ph.D. studentship from the Hong Kong Polytechnic University, and to the State Key Laboratory of Polymer Materials Engineering at Sichuan University, which provided the Materials studio software.

## References

1. Yang ZH, Hu JL, Liu YQ, Yeung LY (2006) Mater Chem Phys 98:368–372
2. Lendlein A, Kelch S (2002) Angew Chem Int Edn 41:2034–2057
3. Lee BS, Chun BC, Chung YC, Sul KI, Cho JW (2001) Macromolecules 34:6431–6437
4. Lin JR, Chen LW (1998) J Appl Polym Sci 69:1563–1574
5. Takahashi T, Hayashi N, Hayashi S (1996) J Appl Polym Sci 60:1061–1069
6. Teo LS, Chen CY, Kuo JF (1997) Macromolecules 30:1793–1799
7. Coleman MM, Skrovanek DJ, Hu J, Painter PC (1988) Macromolecules 21:59–65
8. Sung CSP, Schneider NS (1975) Macromolecules 8:68–73
9. Brunette CM, Hsu SL, MacKnight WJ (1982) Macromolecules 15:71–77
10. Lee HS, Wang YK, Hsu SL (1987) Macromolecules 20:2089–2095
11. Zharkov VV, Strikovsky AG, Verteletskaya TE (1993) Polymer 34:938–941
12. Zhang CL, Hu JL, Chen SJ, Ji FL (2010) J Mol Model 16:1391–1399
13. Skrovanek DJ, Howe SE, Painter PC, Coleman MM (1985) Macromolecules 18:1676–1683
14. Seymour RW, Estes GM, Cooper SL (1970) Macromolecules 3:579–583
15. Carmesin I, Kremer K (1988) Macromolecules 21:2819–2823
16. Jo WH, Ko MB (1994) Macromolecules 27:7815–7824
17. Stevens MJ (2001) Macromolecules 34:1411–1415
18. Tsige M, Stevens MJ (2004) Macromolecules 37:630–637
19. Perun TJ, Propst CL (1989) In: Perun TJ, Propst CL (eds) Computer–aided drug design: methods and applications. Dekker, New York
20. Charifson PS (1997) Practical application of computer–aided drug design. Dekker, New York
21. Madkour TM, Barakat AM (1997) Comput Theor Polym Sci 7:35–46
22. Kremer K (2003) Macromol Chem Phys 204:257–264
23. Kang JW, Choi K, Jo WH, Hsu SL (1998) Polymer 39:7079–7087
24. Pavel D, Ball J, Bhattacharya S (1999) J Polym Sci Part B Polym Phys 37:2334–2352
25. Liang T, Yang X, Zhang X (2001) J Polym Sci Part B Polym Phys 39:2243–2251
26. Eichinger BE, Rigby D, Stein J (2002) Polymer 43:599–607
27. Fan HB, Yuen MMF (2007) Polymer 48:2174–2178
28. Materials studio 4.0 (2006) visualizer, amorphous cell and discover modules Accelrys Inc, San Diego, CA
29. Vasudevan VJ, McGrath JE (1996) Macromolecules 29:637–645
30. Parrinello M, Rahman A (1982) J Chem Phys 76:2662–2666
31. Nosé S (1984) J Chem Phys 81:511–519
32. Lee KJ, Mattice WL (1992) Comput Polym Sci 2:55–63
33. Wagner KG, Maus M, Kornherr A, Zifferer G (2005) Chem Phys Lett 406:90–94
34. Ding F, Bolton K, Rosén A (2004) J Vac Sci Technol A 22:1471–1476
35. Sun H (1998) J Phys Chem B 102:7338–7364
36. Barton JM, Buist GJ, Deazle AS, Hamerton I, Howlin BJ, Jones JR (1994) Polymer 35:4326–4333
37. Mijovic J, Zhang H (2003) J Phys Chem B 107:2557–2563
38. Yu S, Yang S, Cho M (2009) Polymer 50:945–952
39. Parrinello M, Rahman A (1980) Phys Rev Lett 45:1196–1199
40. Abu-Sharkh BF (2001) Comput Theor Polym Sci 11:29–34
41. Karl GW, Martin M, Andreas K, Gerhard Z (2005) Chem Phys Lett 406:90–94
42. Kurt B, Jorg B, Paul W (2003) Prog Polym Sci 28:115–172
43. Yoshioka S, Aso Y, Kojima S (2003) Pharm Res 20:873–878
44. Yang L, Srolovitz DJ, Yee AF (1999) J Chem Phys 110:7058–7069
45. Leach AR (1996) Molecular modeling principles and applications. Addison Wesley Longman, London
46. Mayo SL, Olafson BD, Goddard WA III (1990) J Phys Chem 94:8897–8909
47. Suter UW, Eichinger BE (2002) Polymer 43:575–582

J Mol Model (2012) 18:1273–1284
DOI 10.1007/s00894-011-1152-y

ORIGINAL PAPER

# Evaluation of natural and nitramine binding energies to 3-D models of the S1S2 domains in the *N*-methyl-D-aspartate receptor

**Jason Ford-Green · Olexandr Isayev · Leonid Gorb · Edward J. Perkins · Jerzy Leszczynski**

Received: 23 March 2011 / Accepted: 10 June 2011 / Published online: 7 July 2011


**Abstract** Overactivation of the *N*-methyl-D-aspartate receptor (NMDAR) in postsynaptic neurons leads to glutamate-related excitotoxicity in the central nervous system of mammals. We have built 3-D models of each domain for the universal screening of potential toxicants and their binding mechanisms. Our docking results show that the calculated p$K_i$ values of glycine and L-glutamate significantly increase (>1) when the NR1 and NR2A S1S2 domains are closing, respectively. Inversely, D-cycloserine (DCS) and 5,7-dichlorokynurenic acid (5,7-DCKA) do not show such a dependence on domain closure. Replica exchange molecular dynamics (REMD) confirmed 5 different conformational states of the S1S2 domain along the 308.2 K temperature trajectory. Analysis of residue fluctuations during this temperature trajectory showed that residues in loop 1, loop 2, the amino terminal domain (ATD), and the area linked to ion channel α-helices are involved in this movement. This further implicates the notion that efficacious ligands act through S1S2 lobe movement which can culminate in the opening or closing of the ion channel. We further tested this by docking hexahydro-1,3,5-trinitro-1,3,5-triazine (RDX) and octahydro-1,3,5,7-tetranitro-1,3,5,7-tetrazocane (HMX) to the S1S2 domain. Our results predict that these nitramines are not efficacious and thus do not produce excitoxicity when they bind to the S1S2 domain of the NMDAR.

**Keywords** Homology modeling · Molecular dynamics · Molecular docking · NMDAR · RDX · HMX

This work was performed at the Environmental Laboratory of the US Army Engineer Research and Development Center (ERDC) in Vicksburg, MS 39180, USA

**Electronic supplementary material** The online version of this article (doi:10.1007/s00894-011-1152-y) contains supplementary material, which is available to authorized users.

J. Ford-Green (✉) · J. Leszczynski
Interdisciplinary Center for Nanotoxicity,
Jackson State University,
Jackson, MS 39217, USA
e-mail: jgreen@icnanotox.org

J. Ford-Green · E. J. Perkins · J. Leszczynski
Environmental Laboratory, US Army ERDC,
Vicksburg, MS 39180, USA

E. J. Perkins
e-mail: Edward.j.Perkins@usace.army.mil

O. Isayev
Department of Chemistry, Case Western Reserve University,
Cleveland, OH 44106, USA

L. Gorb
Badger Technical Services,
San Antonio, TX 78216, USA

## Introduction

Excitotoxic (glutamate-related) neuronal-cell injury and cell death has been cited as one of the causative factors in virtually every major neurodegenerative disorder in humans [1–5]. In such events there is an overactivation of the *N*-methyl-D-aspartate receptor (NMDAR) in postsynaptic neuronal membranes. NMDA-ergic overactivation results in an influx of $Ca^{2+}$ into the cytoplasm via the NMDAR's coupled ion channel. As $Ca^{2+}$ and some $K^+$ translocate into the cytoplasm from the synaptic cleft, there is a buildup of positive charge inside the postsynaptic neuron. This triggers the voltage-dependent moiety of the NMDAR by expelling the divalent $Mg^{2+}$ from its position in the ion channel via positive-charge repulsion [6–8]. Glutamate-related excitotoxicity plays a significant role in various neurodegenera-

tive diseases, such as Alzheimer's, Huntington's, and Parkinson's disease, which disrupt the synaptic plasticity underlying Hebbian learning [9, 10].

In mammals, the apparent form of the obligatory NMDAR involves a heterotetrameric stoichiometry [11, 12]. The main binding mechanism of the receptor is to cooperatively bind the neurotransmitters glycine (NR1) and L-glutamate (NR2A) (Fig. 1a), within the clamshell-like binding pocket S1S2 domain (Fig. 1b). This domain may also bind the partial agonist D-cycloserine (DCS) or the full antagonist 5,7-dichlorokynurenic acid (5,7-DCKA) and relay an event that opens or closes its ion channel. It has been established that S1S2 binding is largely dependent upon the ligand type and availability of the binding pocket moieties.

Many sequenced-based site-directed mutagenesis and X-ray crystallography studies have been conducted on the S1S2 domain [13–16]. Previous molecular modeling studies have focused on the opening and closing dynamics of the S1S2 domains in glutamate receptors [17, 18]. Others have utilized homology modeling approaches to produce the obligatory structure of the NMDAR to detect polarization changes in the ion channel before and after ligand docking [19, 20]. However, there has been little study of the effect of S1S2 lobe movement on ligand docking.

Environmental pollutants such as hexanitrohexaazaisowurtzitane (CL-20) can disrupt neural transmission systems [21]. However, some of these toxicants are specific binders to the molecular target NMDAR. The heavy-metal cation $Pb^{2+}$ is a pollutant that is found to directly inhibit this receptor, causing the reduction of NR2A-composed receptors [22]. Contamination by 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (dioxin) in rats also significantly downregulated NR1 mRNA expression, which is directly correlated with NMDAR targeting [23].

RDX (hexahydro-1,3,5-trinitro-1,3,5-triazine) and HMX (octahydro-1,3,5,7-tetranitro-1,3,5,7-tetrazocane) are the premier chemical agents in a class of high density per molecule explosives classified as nitramines. The widespread use of RDX and HMX after World War II has resulted in their presence in at least 16 of the 1397 hazardous waste sites on the National Priorities List [24], and they are found as pollutants in soil [25, 26] and the air [27] around or near munitions manufacturing sites. In mammals, the clinical outcome of oral administration of RDX is the rapid onset of seizures (skeletal muscle convulsions) and increased salivation in humans [28], rats [29–31], mice [32], pigs [33], northern bobwhite quail [34], and western fence lizards [35]. Recently, evidence that RDX binds to the γ-aminobutyric acid type A receptor ($GABA_AR$) instead of the NMDAR was presented [36]. Until now, studies conducted at the genetic level have contributed the most toward the elucidation of the mammalian mechanism of nitramine-induced excitotoxicity.

To this end, we have built viable molecular models of the NR1 and NR2A S1S2 domains for the screening of potentially excitotoxic ligands. This approach is quantitative, and involves comparing the −log of the computational inhibition constants ($pK_i^{cal}$) of DCS and 5,7-DCKA to

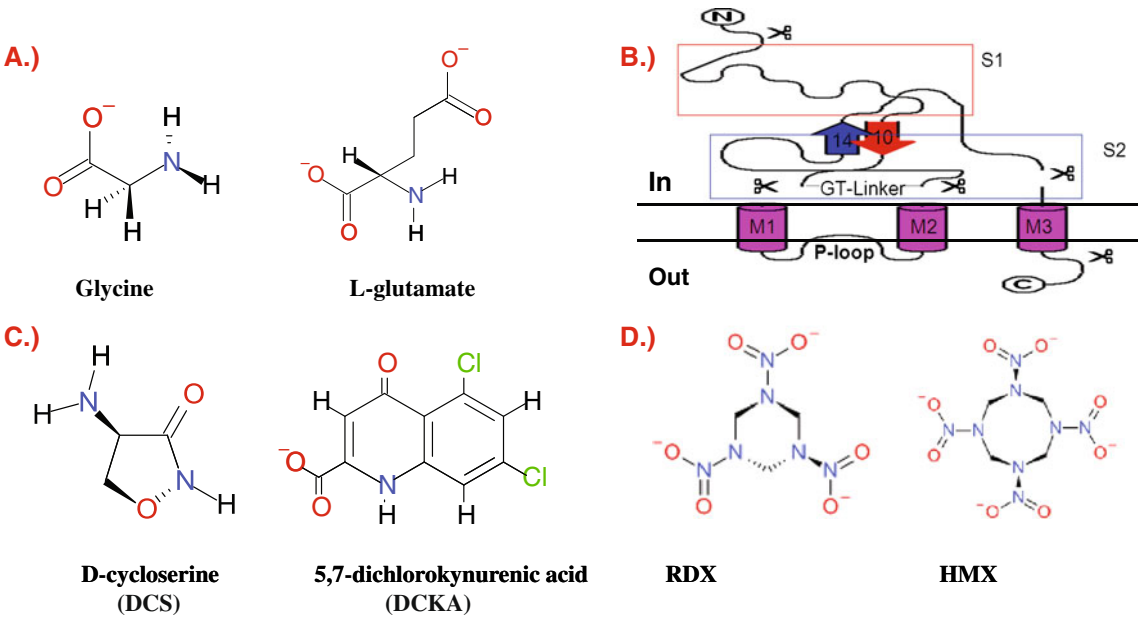



**Fig. 1** **a** Two-dimensional structures of the natural ligands glycine and L-glutamate. **b** Illustration of a generic NMDAR monomer with an amino terminus (*N*), carboxyl terminus (*C*), S1 lobe (in the *red box*), S2 lobe (in the *blue box*), sites where the sequence has been cleaved (indicated by *scissors*), GT-linker region, and transmembrane portion formed by M1/P loop/M2/M3. **c** Two-dimensional structures of the partial agonist DCS and antagonist DCKA. **d** Stereostructures of cyclic nitramine explosives: RDX in the axial-axial-equatorial (AAE) conformation and HMX in the α (boat) conformation

those of the munitions RDX and HMX (Fig. 1d). Recently, our lab has reported transcriptional events implicating the ionotropic glutamate receptor signaling pathway machinery at 3.5 h post-RDX exposure in rats (18 mg $kg^{-1}$ gel capsule) [37], and downregulation of the genes coding for the obligatory NMDAR at 48 h postexposure (24 mg $kg^{-1}$ gel capsule). This prompted us to screen RDX and the structurally similar explosive HMX. The known binding mechanisms of these different ligands can be used to characterize previously unstudied molecules as agonists, antagonists, or nonbinders. Thus, the S1S2 binding mechanism can aid in determining if foreign compounds might be involved in glutamate-related excitotoxicity.

## Methods

### Construction of 3-D models via homology modeling

NR1 [38] and NR2A [39] S1S2 sequences were extracted from the Protein Knowledgebase (UniProtKB). Our target sequences were extracted from two separate genes, GRIN1 (database ID: Q05586-1 isoform-3) and GRIN2A (database ID: Q12879). The portion of the sequence specifying the NR1 and NR2A S1S2 domains from *Homo sapiens* were extracted, respectively. The NR1 sequence was inserted into the program SwissModel using the First Approach Mode with a BLAST E-value limit of $1.0 \times 10^{-5}$ for homologous detection [40] and aligned to the crystal structure of an NR1 S1S2 domain (pdbcode: 1pbq chain A), at 1.90 Å resolution, to provide a quality initial state [13]. The NR2A sequence was also inserted into SwissModel. This time we used Alignment Mode to achieve an equally viable initial state for the NR2A structure and aligned to Furukawa and Gouaux's NR1 S1S2 crystal structure. In this utility we chose to perform our own sequence alignment using the MUSCLE [41] algorithm for optimal alignment of two different yet homologous structures. The quality of the model structure was verified using ANOLEA software [42], Verify3D software [43], and Ramachandran plots [44].

### Energy minimizations and molecular dynamics (MD) simulations

The initial and final states of the models were used to detect any change in ligand binding upon domain movement. The models of NR1 and NR2A were protonated with the Xleap program in the Amber10 package [44]. We utilized the ff03 force field with the Generalized Born (GB) OBC implicit solvent model with an external dielectric constant equal to 78.5 to recover the bulk behavior of water. All simulations contained a 99.0 Å cutoff for optimal evaluation of nonbonded interactions along with the SHAKE algorithm for decoupling forces arising from the inherently high hydrogen bond vibrational modes. We relaxed the 3-D models in the initial state with 10,000 cycles of conjugate gradient minimization. For the final state, a rigid minimization followed by a short MD run protocol was followed and achieved by using heavy restraints on heavy atoms (100 kcal $mol^{-1}$ $Å^{-2}$) to find local minima for hydrogens with 10,000 cycles of conjugate gradient minimization.

We then conducted additional MD sampling of hydrogen positions using the previous coordinates from the last minimization cycle as starting velocities (25 ps with a 1 fs time step) and by heating up to 300 K using Langevin dynamics, where the collision frequency $\gamma$ was set to 1.0 $ps^{-1}$. Subsequently we minimized the sampled structure with 10,000 cycles of conjugate gradient minimization, this time holding only the backbone ($C_\alpha$, C, N, O) atoms rigid by coupling their motion to a 100 kcal $mol^{-1}$ $Å^{-2}$ restraint weight. Next we started allowing the side chains to relax via another MD simulation using the coordinates from the last minimization cycle as the reference structure and by heating up to 300 K using Langevin dynamics and keeping the previous restraint weight over 25 ps. Lastly, we slowly decreased the previous backbone restraints to zero over 25 ps of MD simulation time, i.e., 10 kcal $mol^{-1}$ $Å^{-2}$→ 1.0 kcal $mol^{-1}$ $Å^{-2}$→free, with a 1 fs time step (a 2 fs time step was used for the NR1 model) along with the above thermostat, cutoff, and SHAKE parameters. The final states of the NR1 and NR2A models were chosen during the production phase of the classical simulation based upon minimum total energy, domain S1S2 cleft opening, and binding affinity values ("Electronic supplementary material," ESM, Fig. S3 and Table S1).

Replica exchange molecular dynamics (REMD) simulations were applied to systematically perturb the total energy of the NR1 S1S2 domain by varying the overall temperature of the system. This metropolis-type calculation was simulated in the GBOBC solvent model with the dielectric constant equal to 78.5. A simulation with 64 replicas was evenly distributed between the temperatures 270 and 570 K. Here we also used the ff03 force field along with a 12.0 Å cutoff for nonbonded interactions with the SHAKE algorithm. We used Langevin dynamics with $\gamma$ set to 1.0 $ps^{-1}$, due to the fact that there could possibly be exchanges of replicas within two temperature extremes. Lastly, we ran this simulation with a 2 fs time step for a total of 5 ns. This enhanced-sampling method was employed to identify a more time-efficient computational approach to determine the conformational flexibility of the NR1 S1S2 model. To ensure the binding affinity accuracy of the open and closed models, we used all prevalent conformers from this REMD simulation and compared them to classical MD conformers (see ESM Fig. S3 and Table S1).

### Molecular docking

Blind docking was done on each initial and final state of the S1S2 using the Lamarckian genetic algorithm (LGA). For this initial run, we used 0.375 Å of grid spacing. The lowest-energy docked conformation of each ligand was then used in a subsequent docking calculation using a grid spacing of 0.18 Å. The maximum number of evaluations was set to 5 million, the number of generations was 27,000, the rate of gene mutation was 0.02, and the crossover rate was 0.8. Clustering analysis was used to determine specific docked low-energy conformations. We also used a rigorous method for ligand preparation where atomic charges fitted to reproduce electrostatic potentials were fitted to all ligands using the Merz–Kollman–Singh method [45], and were prepared at the B3LYP/cc-pvtz level of theory in the Gaussian 09 package [46]. All geometries, binding constants, and intermolecular forces were calculated using the Autodock 4.2 program with the AD4 scoring function [47].

## Results and discussion

### Sequence alignment and sequence-based homology modeling

We aligned and modeled structural homology in the amino acid sequences of the NR1 (Q05586-1 isoform-3) and NR2A (Q12879) from *Homo sapiens* to construct the 3-D models of the NR1 and NR2A S1S2 domains (Fig. 3a and b). This provided us with a biologically relevant starting structure to allow for a lower amount of steric hindrance when docking ligands. Table 1 shows that the NR1 and NR2A S1S2 models contained 293 and 225 amino acids using 1pbq chain A as the 3-D template. The NR1 S1S2 sequence had high (99.20%) sequence identity with its template, the 3-D crystal structure from the rat NR1 S1S2 complexed with the partial agonist 5,7-dichlorokynerenic acid (DCKA) [13]. Previous molecular modeling efforts for the NR1 S1S2 chose to directly use the metabotropic glutamate receptor 2 (GluR2) [18, 19] and bacterial histidine-binding protein (HBP) [20] instead of creating homology models. The alpha-carbon (Cα) alignment of the co-crystal complex glycine-NR1 S1S2 aligned to the complex DCKA-NR1 S1S2 had a domain closure of ~21°.

As a result, the authors designated the structure 1pbq the open form of the domain. However, considerable molecular dynamics effort is required if we are to designate a particular form the open form. A study utilizing this approach was done on the NR1 S1S2, showing fast and slow closure of the S1S2 lobes over a period of 20 ns [17]. This behavior was attributed to the lack of opposing force from the transmembrane domain (Fig. 2c). We cannot dismiss the notion that this force, in addition to the effect of the domain leading from the amino terminus (AT), may work to further open the S1S2 lobes in the absence of a ligand in addition to its closure. The NMDAR extracellular environment is known to be a site for redox modulation from endogenous nitric oxide molecules [48, 49]. Therefore, the cysteine residues on the NR1 (Cys744 and Cys798) and NR2A (Cys87 and Cys320, both in the AT domain) should not be relied upon to restrain the degree of opening of the S1S2 lobes.

The NR2A S1S2 sequence had a marked difference (43.02%) in sequence identity with its template after initial alignment using the MUSCLE algorithm. Four genes code for the NR2 subunit (NR2A-D) compared to 1 gene for the NR1. This diversity adds to the difficulty in producing a homology model with large spanning regions of conserved amino acids. Additional difficulty, unlike for the NR1, arises from the lack of an NR2A S1S2 X-ray crystal structure with varying degrees of cleft opening. Indeed, it is a harder task to produce a feasible open form of the NR2A S1S2. The most viable NR2A S1S2 crystal is complexed with its agonist L-glutamate (pdb code: 2a5s) [15]. A previous homology model for this domain was built using the glycine-bound NR1 S1S2 complex (pdb code: 1pb7) as a template [50]. However, using 1pb7 as a structural

**Table 1** Sequenced-based homology modeling of 3-D models

| Model | | Template | | | Homology |
|---|---|---|---|---|---|
| Domain[a] | Total number of residues | Structure[c] | Chain | Resolution (Å) | % Sequence identity |
| NR1 S1S2[§b] | 293 | 1pbq | A | 1.9 | 99.20 |
| NR2A S1S2[§b] | 225 | 1pbq | A | 1.9 | 43.02 |

[§] Extracellular (main) binding domain of each NMDAR monomer

[a] Sequence taken from the Protein Knowledgebase (UniProtKB)

[b] Ultimate alignment done using the ClustalW algorithm in SwissModel [40]

[c] Protein databank ID (pdb ID)

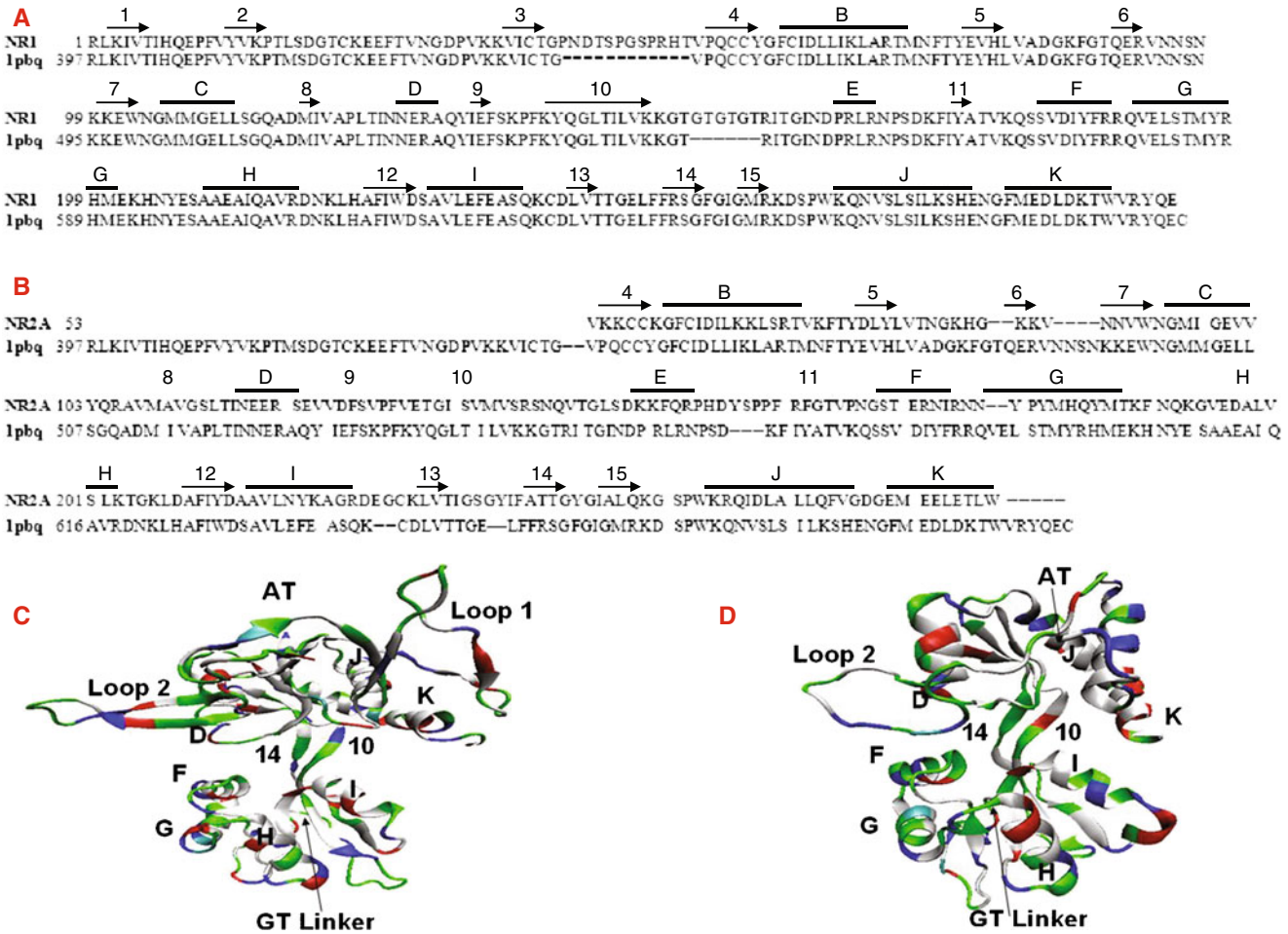



**Fig. 2** Sequences and illustrations of the 3-D models produced by SwissModel, where the *arrows above the sequence* indicate β-sheets and the *rectangles* indicate α-helices. **a** NR1 S1S2 main (clamshell-like) binding domain sequence. **b** NR2A S1S2 main (clamshell-like) binding domain sequence, **c** 3-D model of the NR1 S1S2 main (clamshell-like) binding domain with the amino terminus (AT), the glycine-threonine (GT) linker normally coupled to its monomer forming the TM domain, and the major loops 1 and 2, which were previously found to be coupled to the NR2A S1S2 domain [16], **d** 3-D model of the NR2A S1S2 main (clamshell-like) binding domain, like the NR1 S1S2 model, showing β-sheets 14 and 10 forming the interdomain cleft and α-helices D, F, G, H, I, J, and K

template limits the ability to review the efficacy of a ligand. One study aligned the sequences of the NR2A and NR2B S1S2 domains to the atomic coordinates of 1pbq in order to structurally screen similar drugs [51].

Figure 2 shows the SwissModel sequence alignment used to create the S1S2 models. The NR1 sequence builds a 3-D model starting at Arg397 of 1pbq chain A and includes the first β-sheet formed by the sequence $^{399}$KIVT$^{402}$. In total there are 15 β-strands and 10 α-helices numbered from B-K on the NR1 model, whereas the NR2A model contains 12 β-strands and also 10 α-helices numbered from B-K. This model starts at Val451 of 1pbq chain A located on β-strand 4, and as a result the first 54 amino acids are not modeled. Such residues correspond to β-strands 1–3 and are structurally responsible for correlating AT domain movement to NR2A S1S2–ligand binding through the loop 1 motif [15]. Since our binding calculations are static in nature, the lack of loop 1 and β-sheets 1–3 could not introduce their artifacts into our study. One residue, Glu413, an electrostatic partner to Asp731 is not modeled, and will not affect the direct binding of any ligand to the pocket of the NR2A S1S2 model. The amine nitrogen of L-glutamate participates in a water-mediated interaction with Glu413, but this interaction is also stabilized by direct bonding with the carbonyl oxygen of Ser511 which is present in the model. Both S1S2 models contain the essential secondary structure elements for binding, namely helices D, F, G, H, I, J, and K along with interdomain antiparallel β-strands 10 and 14. Essential glycine-threonine (GT) linkers were also modeled in both structures which connect the S1 and S2 domains. Ramachandran plots of the models in the initial state reveal that these 3-D models have amino acids in structurally ordered regions and are therefore viable (see ESM Fig. S1).

## Molecular dynamics of the S1S2 models

We have energetically minimized (initial state) and subsequently simulated (final state) each model in preparation for docking studies. The models were effectively equilibrated using a rigorous protocol for simulation providing a valid geometry for the extended-cleft form of the NR1 S1S2 (ESM Fig. S2). In a separate simulation, an RMSD analysis of the backbone atoms of the NR1 S1S2 model was taken from a 64-replica REMD simulation (Fig. 3). Trajectory 283.8 K was used as the reference in the temperature range of 283.8–308.2 K. This range was chosen because previous simulations (not shown here) revealed it to have marked RMSDs. After analyzing Fig. 3, we manually set the conformations into specific RMSD ranges (i.e., 1.0–2.0 Å, 5.0–6.0 Å, 6.0–7.0 Å, 9.0–10.0 Å, and 10.0–11.0 Å) to categorize distinct conformers. The temperature trajectory 308.2 K is the only one that visits all five ranges, and there is a <0.3 Å RMSD difference (in the above ranges) between its snapshots and snapshots from other temperature trajectories. Thus, all five snapshots were taken from the 308.2 K temperature trajectory, as it gave the best representation of the varying degrees of domain flexibility along a single trajectory. Figure 3 clearly shows regions where certain conformations are visited by the model in this temperature range illustrated by A, B, C, D, and E. This trajectory also encompasses conformations A–E.

Characterization of these conformational states shows a combination of a breathing (no S1S2 lobe separation→ vertical S1S2 lobe separation) and hinge-bending (degree of tilt of S1 lobe is>1° with respect to S2 lobe) motion to accommodate partial and full agonist binding or release. These findings are in correspondence with a previous study done on the glutamate receptor (GluR2) S1S2 domain [18].

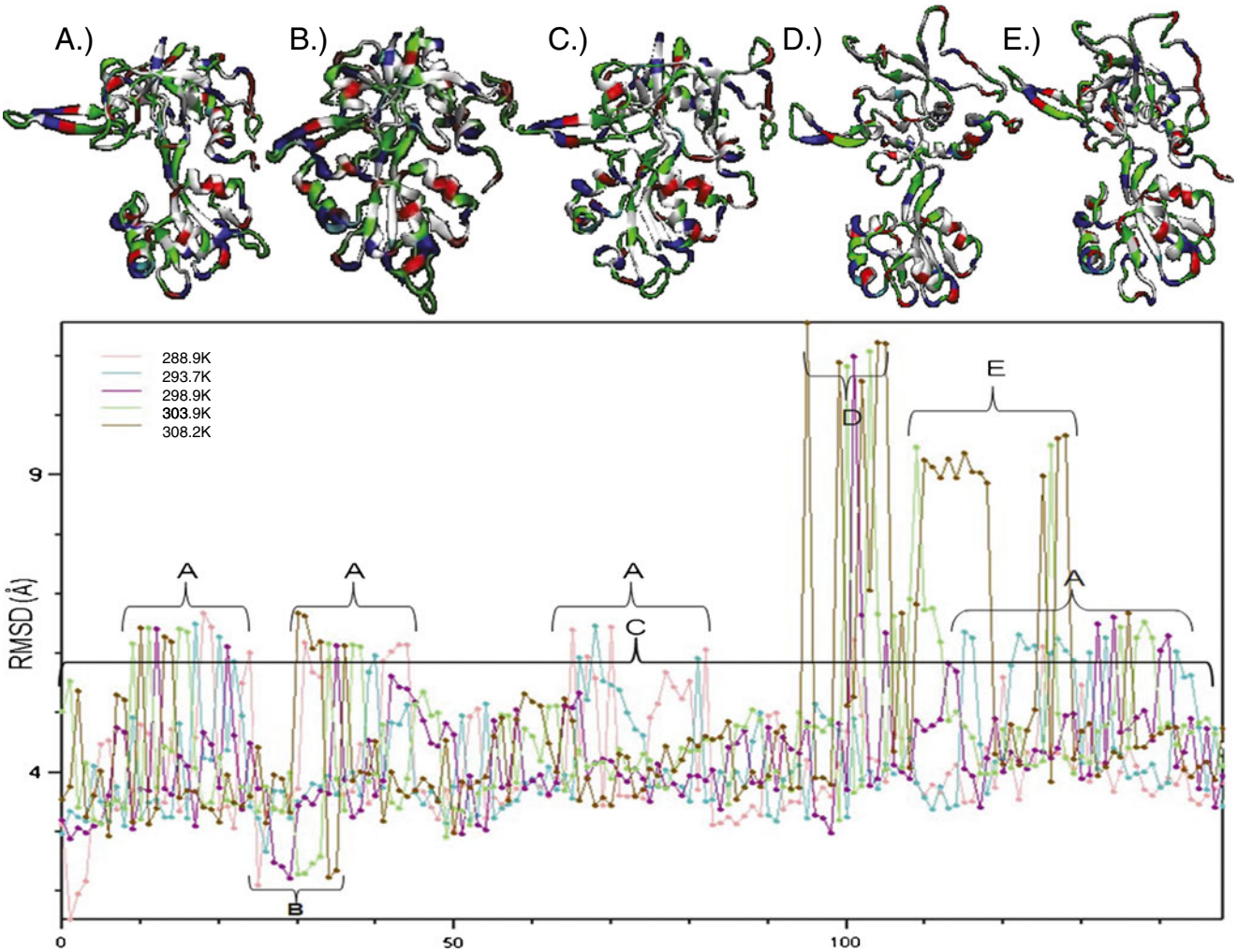



**Fig. 3** Illustration of all temperature trajectories within the 288.9–380.2 K temperature range from a replica exchange molecular dynamics (REMD) simulation using 64 replicas over 5 ns. The RMSD is plotted with respect to the temperature trajectory for 283.8 K. Overall, five different conformational states were visited during the simulation, showing that the NR1 S1S2 3-D model is flexible along the 308.2 K temperature trajectory

However, the presence of a hinge-bending motion is also supported by earlier dynamics done on the NR1 S1S2 without the presence of a lobe-twisting motion [17]. A lobe-twisting motion is governed by the slight movement of the S1 lobe in and out of the plane formed by the overlapping S1 and S2 lobes (the plane of the page). The back-to-back orientation of the NR1 and NR2A S1S2 domains may offer a coupled response for this type of motion. If the NR1 S1 lobe were to move in the plane of the page then the NR2A S1 domain might be perturbed enough to affect its ligand binding. We cannot rule out the existence of this lobe-twisting motion with our current approach. A further study including a broader temperature range, more replicas, and a longer simulation time would further enhance the sampling of conformations on the potential energy landscape to provide a more definitive answer.

We investigated which amino acid segments are responsible for these hinge-bending and breathing motions that in turn allow the docking of glycine and excitotoxicants such as RDX. The GT-linker region is known to be coupled to the α-helices forming the ion channel (Fig. 1b). The results of an atomic fluctuation analysis along the trajectory that visits the most conformational states in the above temperature range during the simulation are shown in Fig. 4. Analysis of the fluctuations of amino acids during the 308.2 K temperature trajectory indicated high fluctuations in the regions near Asp416 (our model Asp21) in loop 1, which is < 30 amino acids away from the amino terminus domain (ATD, Fig. 2a). High-fluctuation regions were also near Gly445 (our model Gly49) and Asn494 (our model Asn96) in loop 2. These findings are in agreement with a study reporting high fluctuations in residues 416–417 (loop 1), 426–427 (loop 1), 443–449 (loop 1), and 492–493 (loop 2) [17]. Finally, there is a high-fluctuation region in the GT-linker segment (our model Thr152), where movement is coupled to NR1 α-helices for channel gating (Fig. 2c) [11, 12]. Ultimately, this supports the notion that S1S2 domain movement is dynamically coupled to the ATD and the TMD [12–16].

It is important to note that, in addition to these calculations, we calculated the binding affinities of two trial ligands (structurally unrelated) to each of the five conformations from REMD (Fig. 3). This was done to verify that our NR1 initial and final state structures are indeed are the best candidates for molecular docking. ESM Table S1 clearly shows that, for both ligands, conformations A–E have binding affinity values that lie between NR1 open and NR1 closed values. This apparently demonstrates that we have properly chosen the most closed and the most open forms of the NR1 using a short classical MD simulation. We have also calculated the RMSD values of each REMD structure with respect to the NR1 open conformation and found that the E conformer is most closely related (ESM Fig. S3). Consequently, the NR1 E conformer was used as a starting point for a classical MD

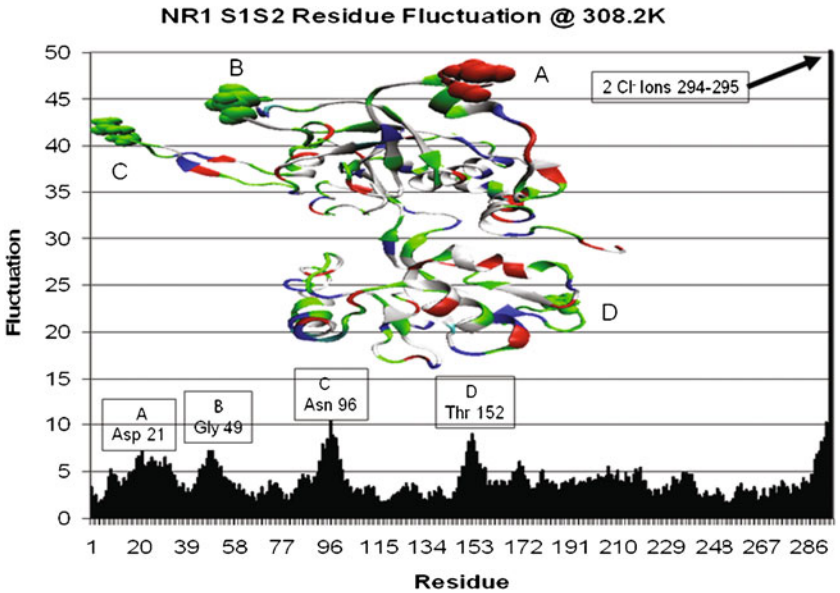



**Fig. 4** Atomic fluctuations based on residue analysis using the ptraj program at 308.2 K show that there are four major areas that collaborate in the overall motion in the NR1 S1S2 model. These areas are (*A*) aspartic acid 21 (Asp416) in loop 1, which is only a few residues (< 30 amino acids) away from the AT end where the amino terminus domain (ATD) sits, (*B*) glycine 49 (Gly445), which is also very close to the ATD [13], (*C*) asparagine 96 (Asp494) is at the tip of loop 2 and its movement has been shown to be highly coupled with the NR2A S1S2 domain [14–16], and (*D*) threonine 152 is located near the GT-linker segment of the 3-D model and appears to verify earlier proposals of NR1 S1S2 movement coupled with NR1 α-helical movement for channel gating [12, 13]

simulation. The structure was prepared under the same conditions that the NR1 open conformer was subjected to. After a 1 ns simulation, there was a less than 1 Å RMSD difference between the two structures, further indicating that the NR1 open model is related to the five REMD conformers and that this is the most computationally open form that can be achieved under such conditions.

## Binding to the S1S2 domains

The −log of the binding constants, p$K_i$, of each ligand to its respective three-dimensional model is given in Table 2. A PDB file was created for each of the lowest-energy conformations of both initial and final state structures. These conformations are designated NR1 initial, NR1 final, NR2A initial, and NR2A final. Each ligand's experimental inhibition constant was compared to the lowest calculated inhibition constant [13, 48]. This experimental value reflects the binding affinity of the ligand once the S1S2 domain cleft has closed up (blocking ligand accommodation), and is termed its efficacy. We have labeled these values $pK_i^{exp}$, while the calculated values reflect the affinity of the ligand for either an initial or final state. It should be noted that since we were investigating the effect of domain closure on the binding affinity of each ligand, we compared the $pK_i^{exp}$ to the initial state in each model. We did not expect to recover an exact agreement between the $pK_i^{cal}$ and $pK_i^{exp}$ values, but we did expect to find the same trends in binding.

There was an order of magnitude binding difference between the final and initial state of the S1S2 domains, 1.54 to 2.85 and 1.24 to 2.91, when binding the small peptide ligands glycine and L-glutamate, respectively (Table 2). This effect is explained by their differing molecular mechanisms of binding in Fig. 5a–d. The final state model of the NR1 was purposefully modeled with reduced disulfide bridges on Cys744 and Cys798, so the S1 lobe separated from the S2 lobe during the production phase of the MD simulation (resembling Fig. 3d). This extended orientation was also formed in the final state of the NR2A model during its production phase. However, this trend did not persist with the partial agonist D-cycloserine (DCS) and the antagonist 5,7-dichlorokynurenic acid (5,7-DCKA), where the largest differences in p$K_i$ values were 2.07 to 1.98 and 3.59 to 3.72, respectively. The dependence (glycine, L-glutamate) and independence (DCS, 5,7-DCKA) of S1S2 cleft closure on the binding affinity supports earlier findings that the natural ligands are more efficacious than the other two ligands [12–16]. This can be explained by the fact that there are significantly fewer electrostatic interaction partners for the small peptides to reach viable energetic minima when the S1S2 cleft is open. The cyclic structures of DCS and 5,7-DCKA do not

**Table 2** Comparison of inhibition constants for the three-dimensional S1S2 models

| Model | Ligand | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Glycine | | | L-glu | | | DCS | | | 5,7-DCKA | | | RDX | | | HMX | | |
| | $\Delta G^c$ | $pK_i^{cal}$ | $pK_i^{exp}$ | $\Delta G$ | $pK_i^{cal}$ | $pK_i^{exp}$ | $\Delta G$ | $pK_i^{cal}$ | $pK_i^{exp}$ | $\Delta G$ | $pK_i^{cal}$ | $pK_i^{exp}$ | $\Delta G$ | $pK_i^{cal}$ | $pK_i^{exp}$ | $\Delta G$ | $pK_i^{calc}$ | $pK_i^{exp}$ |
| NR1 final | −2.10 | 1.54 | – | −3.92 | 2.88 | – | −2.82 | 2.07 | – | −4.87 | 3.57 | – | −2.08 | 1.52 | – | −0.94 | 0.69 | – |
| NR1 initial | −3.89 | 2.85 | 4.58 [a] | −3.76 | 2.75 | >0.52 [b] | −2.7 | 1.98 | 3.62 [a] | −4.74 | 3.47 | 6.27 [a] | −1.24 | 0.91 | – | −0.95 | 0.70 | – |
| NR2A final | −2.24 | 1.64 | – | −1.70 | 1.24 | – | −2.7 | 1.98 | – | −4.9 | 3.59 | – | −3.19 | 2.34 | – | −1.66 | 1.21 | – |
| NR2A initial | −3.10 | 2.27 | 7.39 [b] | −3.97 | 2.91 | 2.02 [b] | −2.58 | 1.89 | 3.56 [b] | −5.07 | 3.72 | 3.19 [b] | −1.9 | 1.39 | – | −1.5 | 1.10 | – |

[cal] Calculated inhibition constants using Autodock 4.2

[exp] Experimental inhibition constants

[a] Values determined from [13]

[b] Values determined from [52]

[c] The change in free energy is calculated once the ligand binds or docks in its lowest-energy conformation

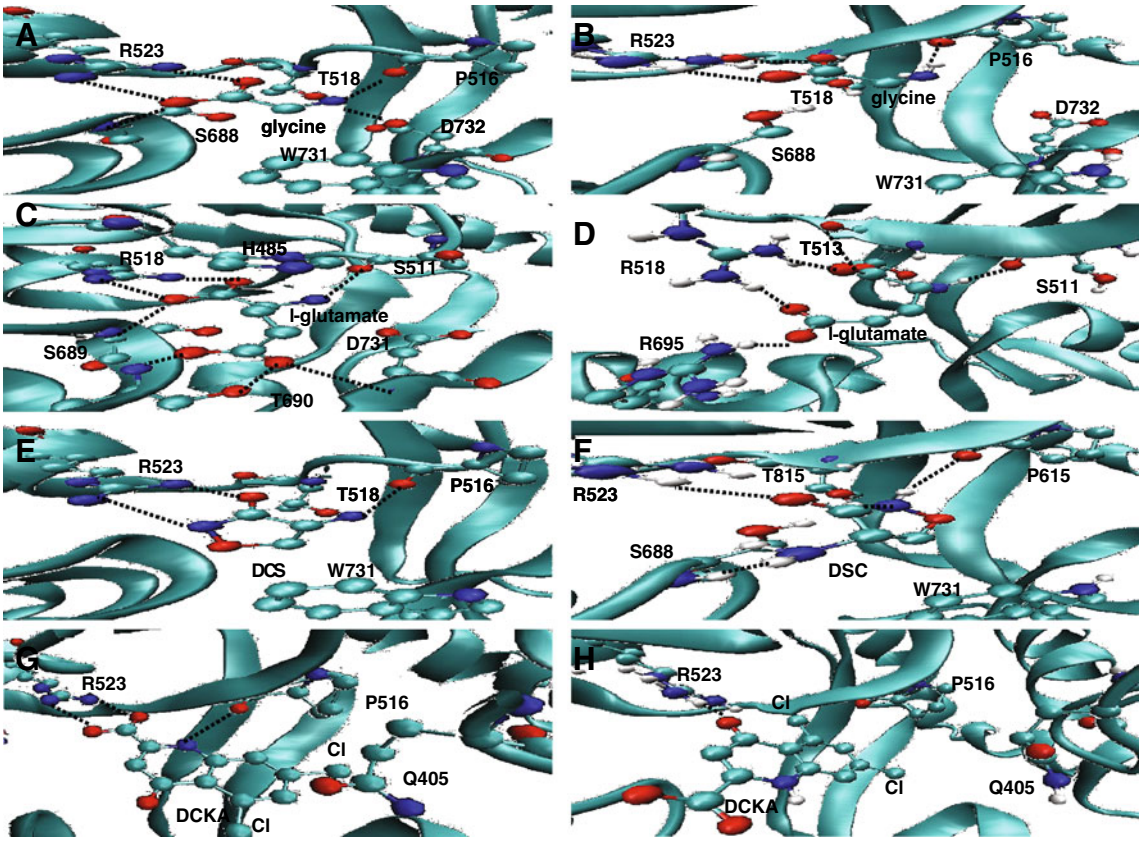



**Fig. 5** Comparisons of the binding mechanisms of certain ligands in the NR1 and NR2A S1S2 domains. **a** Glycine crystallized inside of the binding pocket of the NR1 subunit from 1pb7 [13]. **b** Docked glycine inside the NR1 model. **c** Crystallized L-glutamate in the NR2A subunit from 2a5s [15]. **d** Docked L-glutamate inside the NR2A model. **e** Partial agonist DCS crystallized inside the NR1 subunit from 1pb9 [13]. **f** Docked structure of DCS inside the NR1 model. **g** DCKA crystallized inside the NR1 binding pocket from 1pbq [13]. **h** Docked antagonist DCKA inside the pocket of the NR1 model

experience this limitation, as all of the binding occurs within the S1 domain.

A direct comparison is made in Fig. 5 of the crystal structures 1pb7 [13], 2a5s [15], 1pb9 [13], and 1pbq [13] to the initial states of the lowest-energy docked structures of glycine, L-glutamate, DCS, and DCKA, respectively. The docked structure of glycine in Fig. 5b agrees with the crystal 1pb7 in Fig. 5a, forming electrostatic interactions with Arg523, Pro516, and Thr518. The floor of the active site is formed by Tyr731 and is also conserved in our model. However, Asp732 is turned away from the model's active site, driving glycine to form a tighter hydrogen bond with the carbonyl oxygen of Pro516. Consequently, glycine's carboxyl oxygen cannot hydrogen bond with the backbone amine nitrogen on Ser688. The close correlation between calculation and experiment highlight the accuracy of this model in screening peptides. The −log of the experimental binding affinity of L-glutamate to the NR2A (7.39) is almost twofold higher than that observed when it binds to the NR1 (4.58). However, this trend is not observed in the docking results, where these values increase from 2.27 to 2.85 when L-glutamate is docked to the NR2A and NR1 models, respectively. This can be explained by the residues His485 (S1 domain), Asp731 (S2 domain), and Thr690 (S2 domain) of the NR2A model not being in close enough proximity to the active site (Fig. 5d). Absence of the lid His485 allowed for L-glutamate to bind to an experimentally undetermined residue, Thr513 (S1 domain), via its carbonyl oxygen. The δ carboxyl oxygen of L-glutamate also forms an electrostatic interaction with a nonexperimental Arg695. Such a difference in active site side chains in the NR2A model is attributed to low sequence homology in helix F and gaps inserted by SwissModel between residues $^{486}$GKKVNNV$^{492}$ to compensate for a shorter loop 2 region. The insertion of gaps into loop 2 is essential in order to align the NR2A sequence to any NR1 crystal [51], and a more sophisticated loop refinement technique should be employed if the exact identification of the active site residues is to be achieved. Although these residues are not present, Arg518 (S1 domain) and Ser511 (S1 domain) provide electrostatic interactions that lock peptides into the active site and allow the use of this model as a qualitative screening tool.

The cyclic compounds DCS and DCKA bind in similar fashion to the NR1 active site (Fig. 5e–h). The partial agonist

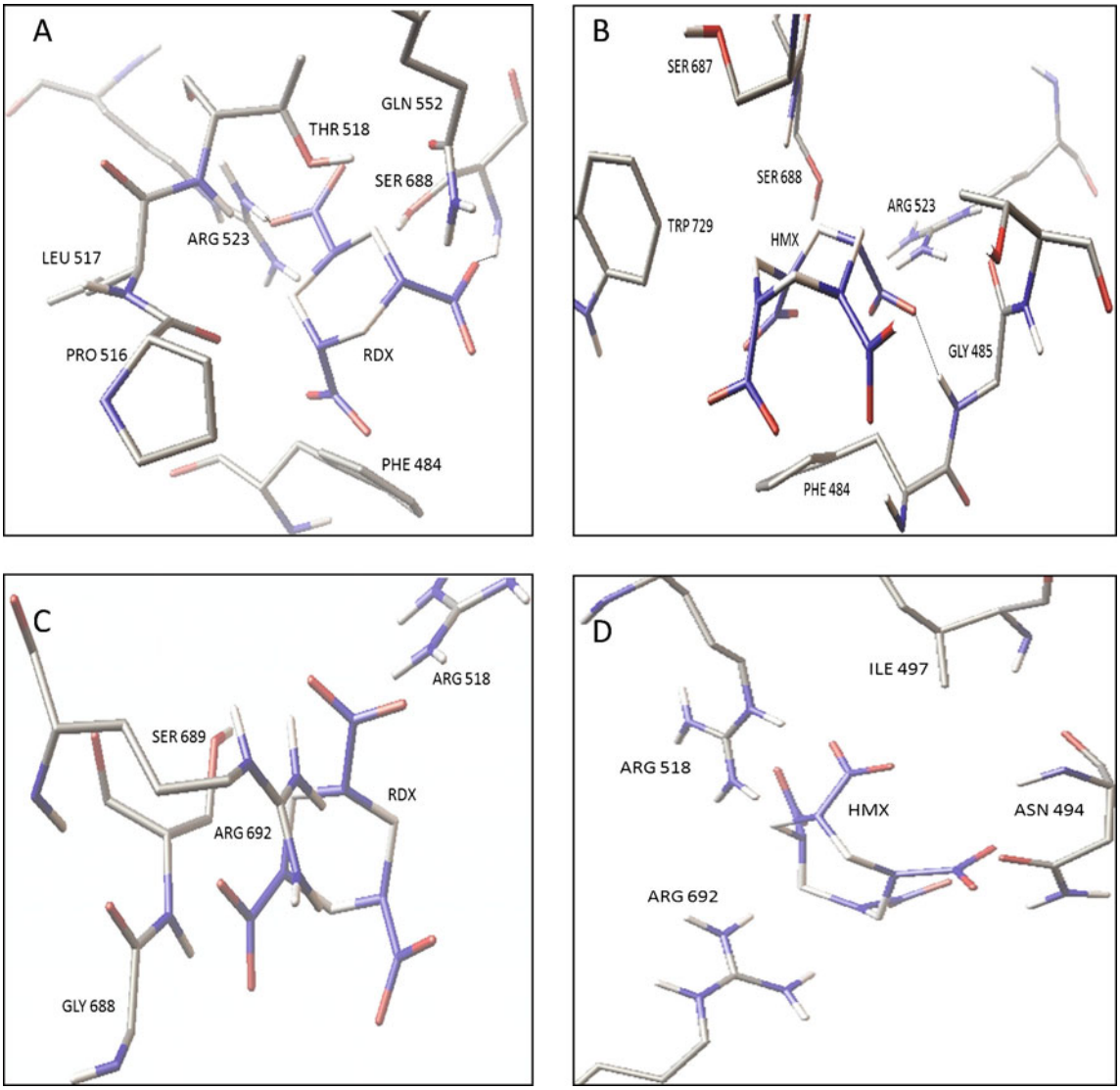



**Fig. 6** **a** Binding mechanism of RDX in the binding pocket of NR1 initial state. **b** HMX binding in the NR1 initial state. **c** RDX in the binding pocket of NR2A initial state. **d** HMX binding in different pockets of the NR2A initial state

DCS requires Arg523 on helix D, the carbonyl oxygen of Pro516 between β-strand 8 and helix D, and the carbonyl oxygen of Thr518 between β-strand 8 and helix D. There are no major binding differences experimentally between the NR1 (3.62) and NR2A (3.56) subunit. Calculation gives the same trend, where the NR1 model has a binding affinity of 1.98 and the NR2A model of 1.89. However, according to calculations, the isoxazolidine ring is inverted and points the amine nitrogen away from the carbonyl oxygen of Pro516. This hydrogen bonding partner is substituted for the guanidinium nitrogen of Ser688 to accommodate for the formation of four electrostatic interactions. The four electrostatic partners at the active site of the NR2A model (i.e., Arg518, Ser511, Thr519, and Arg695) correspond to the functionally similar residues Arg518, Ser511, Asp731, Thr690, and Ser689 in the 1pb9 crystal. We can conclude that a small nitrogen-containing cyclic compound with a nitrogenous substituent must form four electrostatic interactions within the NMDAR S1S2 domain to become efficacious. The antagonist DCKA forms electrostatic interactions between its carboxyl oxygens and the guanidinium nitrogens of Arg523, and the nitrogen member of its bicyclic ring hydrogen bonds to the carbonyl oxygen of Pro516. The docked structure only shows electrostatic interactions between the ligand's carbonyl oxygen and the guanidinium nitrogens of Arg523. This may explain the similar calculated binding affinities between the NR1 and NR2A subunits: 3.47 and 3.22, respectively. The experimental difference between NR1 and NR2A is markedly larger (6.27 and 3.19, respectively). This could be due to steric hindrance at the NR2A active site arising from the less flexible nature of the S1S2 domain [15].

Compared to these previously docked ligands, RDX and HMX both showed less binding affinity to the NR1 initial state of the S1S2 domain, with p$K_i$ values of 0.91 and 0.70, respectively. Slightly higher binding to the NR2A initial state was seen for RDX and HMX at p$K_i$ values of 1.39 and 1.10, but these are still less than the intermediate-strength binding chemical DCS (1.89), and far less than the natural ligand L-glutamate (2.91). These results indicate that RDX and HMX do not exhibit any significant binding to the S12 domains of the NMDAR. What little binding that does occur in the NR1 appears to be mediated by similar side chains to those in peptides, cyclic nitrogen-containing partial agonists, and chlorinated bicyclic antagonists (Fig. 6a–b). RDX binds to the backbone-amide hydrogen of Ser688, the guanidinium nitrogen of Arg523, and is held in the active site by the Phe484 of the NR1 model. HMX participates in a backbone-amide hydrogen bond to Gly485, and also is held in the active site by Phe484. These nitramines found a slightly different active site in the NR2A model, lined by the essential residue Arg518, but they visit a different site on helix F near Arg692. However RDX does not form any hydrogen bonds with these residues or Ser689. HMX also lacks the formation of electrostatic interactions with these essential residues. These may be the reasons why RDX and HMX interact more strongly with the NR1 subunit rather than the NR2A. Low homology in helix F could have offered a poorly constructed active site compared to the NR1 model, but the lack of electrostatic partners pointing towards the directional nitro groups is a more feasible explanation. Although DCS is the most structurally similar ligand to RDX, the directionality of the nitro groups reduces the probability that the nitramine will form the 4–5 electrostatic interactions at the S1S2 active sites.

Figure 7 shows the conservation of binding residues in the S1S2 domain over several species used in experiments [13–15]. This multiple sequence alignment determines residues that are conserved in the ionotropic glutamate receptor (iGluR) family across several species that were found near the S1S2 docking sites. The same helix D residues Arg523 (NR1) and Arg518 (NR2A) are found in or near the active site during both RDX and HMX binding, which is reminiscent of glycine, L-glutamate, DCS, and DCKA binding. A local site sitting at the edge of the main binding pocket between β-sheet 11 and helix F conserves Ser687 (NR1) and Gly688 (NR2A), although each preferentially binds HMX and RDX, respectively. Since nitro-group directionality increases the molecular volumes of RDX and HMX, these residues have small enough side chains to allow effective recruitment of these nitramines within a distance close to the active sites. Once docked, the 4–5 residues that provide electrostatic interactions for efficacy do not fully align with nitramine nitrogens or nitro groups, making their interactions with the NR1 and NR2A S1S2 domains weak.

## Summary

We have performed computational analysis to determine whether or not the nitramine compounds RDX and HMX are significant contributors to NMDAR-ergic excitotoxicy.

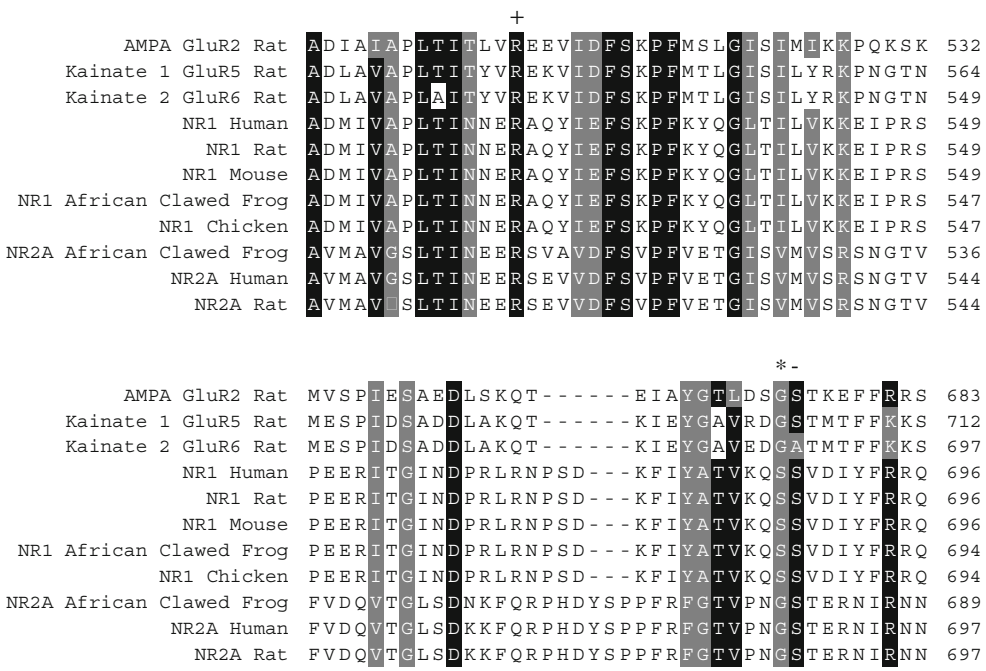



**Fig. 7** Multiple sequence alignment of S1S2 domains from species used in experiments [13–15]. A *plus sign* indicates side chains that were found to aid the docking of both RDX and HMX to the NR1 and NR2A. We placed a *minus sign* over the amino acids unique to either HMX docking to the NR1 or RDX docking to the NR2A. The *asterisk* indicates residues that dock both RDX and HMX to the NR1 and RDX to the NR2A. Identical amino acids are colored with a *black background*, while similar amino acids are colored with a *gray background*. Similar amino acids are categorized into the groups ILV, FWY, KRH, DE, GAS, P, C, and TNQM

The flexibility of the NR1 S1S2 domain, in a hinge-bending and breathing motion, could be the precursor for the opening or closing of the ion channel. We have also verified earlier findings that the natural ligands are more efficacious than the partial agonist and antagonist [13–15]. This further validates the use of our model as a reliable in silico screening tool for the screening of NMDA-ergic excitotoxicity. RDX and HMX binding is very weak, and there seems to be no efficacy contribution to aid in their binding. Therefore, there is no apparent computational evidence of their preferential binding to the S1S2 domains.

**Acknowledgments** The authors would like to thank Jacques Reifmann and the rest of his group at The Biotechnology High Performance Computing Software Application Institute (BHSAI) for their generous allotment of time, and the Army Research Laboratory (ARL) for the use of their computational resources. This work has been done under grant number W912HZ-09-C-0026. The use of trade, product, or firm names in this report is for descriptive purposes only and does not imply endorsement by the U.S. Government. Results in this study were funded and obtained from research conducted under the Environmental Quality Technology Program of the United States Army Corps of Engineers by the US Army ERDC. Permission was granted by the Chief of Engineers to publish this information. The findings of this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

## References

1. Lipton SA, Nicotera P (1998) Cell Calcium 23:165–171
2. Choi YB, Tenneti L, Le D, Ortiz J, Bai G, Chen HSV, Lipton S (2000) Nat Neurosci 3:15–21
3. Meldrum B, Garthwaite J (1990) Trends Pharmacol Sci 11:379–387
4. Rothman SM, Olney JW (1987) Trends Neurosci 10:299–302
5. Dingledine R, Borges K, Bowie D, Traynelis SF (1999) Pharmacol Rev 51(1):7–61
6. Liu Y, Zhang J (2000) Chin Med J 113:948–956
7. Cull-Candy S, Brickley S, Farrant M (2001) Curr Opin Neurobiol 11:327–335
8. Paoletti P, Neyton J (2007) Curr Opin Pharmacol 7:39–47
9. Hebb DO (1949) The organization of behavior. Wiley, New York
10. Hebb DO (1961) Brain mechanisms and learning. Oxford University Press, London
11. Schorge S, Colquhoun D (2003) J Neurosci 23:1151–1158
12. Chatterton E, Awobuluyi M, Premkumar LS (2002) Nature 415:793–798
13. Furukawa H, Gouaux E (2003) EMBO J 22:2873–2885
14. Inanobe A, Furukawa H, Gouaux E (2005) Neuron 47:71–84
15. Furukawa H, Singh SK, Mancusso R, Gouaux E (2005) Nature 438:185–192
16. Chen HSV, Lipton SA (2005) JPET 314:961–971
17. Kaye SL, Sansom MSP, Biggin PC (2006) J Biol Chem 281:12736–12742
18. Mamonova T, Speranskiy K, Kurnikova M (2008) Proteins 73:656–671
19. Chohan KK, Wo ZG, Oswald RE, Sutcliffe MJ (2000) J Mol Model 6:16–25
20. Sutcliffe MJ, Smeeton AH, Wo ZG, Oswald RE (1998) Faraday Discuss 111:259–272
21. Gong P, Basu N, Scheuhammer AM, Perkins EJ (2010) Environ Sci Pollut Res Int 17:181–186
22. Neal AP, Worley PF, Guilarte TR (2011) Neurotoxicology 32:281–289
23. Nayyar T, Wu J, Hood DB (2003) Cell Mol Biol (Noisy-le-grand) 49:1357–1362
24. Agency for Toxic Substances and Disease Registry (1995) Toxicological profile for RDX. Agency for Toxic Substances and Disease Registry, Atlanta
25. Bartha R, Hsu T (1974) Soil Science 115:444–453
26. Jenkins TF, Walsh ME (1992) Talanta 39:419–428
27. Carpenter BH, Liepins R, Sickles J II, Hamilton HL, van Osdell DW (1978) SmartMedia 284:A221060
28. Barsotti M, Crofti G (1949) Med Lav 40:107–112
29. Burdette LJ, Cook LL, Dyer RS (1988) Toxicol Appl Pharmacol 92:436–444
30. Crouse LC, Michie MW, Major MA et al (2006) Subchronic oral toxicity of RDX in rats. Elsevier, Amsterdam, pp 619–638
31. Schneider NR, Bradley SR, Andersen ME (1978) Toxicol Appl Pharmacol 46:1163–1171
32. McCain WC, Ferguson JW (1998) Toxicol study 2340-38-95-6-1. US Army Center for Health Promotion and Preventative Medicine, Aberdeen Proving Ground
33. Major MA, Reddy G, Berge MA, Patzer S, Li AC, Gohdes M (2007) J Toxocol Environ Health 70:1191–1202
34. Quinn MJ Jr, McFarland CA, Johnson MS (2009) Toxicol Mech Methods 19:1537–6524
35. McFarland CA, Quinn MJ Jr, Bazar MA, Remick AK, Talent LG, Johnson MS (2008) Environ Toxicol Chem 27:1102–1111
36. Williams LR, Aroniadou-Anderjaska V, Qashu F, Finne H, Pidoplichko V et al (2011) Environ Health Perspect 119:357–363
37. Bannon DI, Dillman JF, Hable MA, Phillips CS, Perkins E (2009) J Chem Res Toxicol 22:620–625
38. Karp SJ, Masu M, Eki T, Ozawa K, Nakanishi SJ (1993) Biol Chem 268:3728–3733
39. Molina H, Horn DM, Tang N, Mathivanan S, Pandey A (2007) Proc Natl Acad Sci USA 104:2199–2204
40. Arnold K, Bordoli L, Kopp J, Schwede T (2006) Bioinformatics 22:195–201
41. Edgar RC (2004) Nucleic Acids Res 32:1792–1797
42. Melo F, Feytmans E (1997) J Mol Biol 267:207–222
43. Lüthy R, Bowie JU, Eisenberg D (1992) Nature 356:83–85
44. Case DA, Darden TA, Cheatham TE III et al (2008) AMBER10. University of California, San Francisco
45. Besler BH, Merz KM Jr, Kollman PA (1990) J Comput Chem 11:431–439
46. Frisch MJ, Trucks GW, Schlegel HB et al (2009) Gaussian 09, revision A.1. Gaussian Inc., Wallingford
47. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ (1998) J Comput Chem 19:1639–1662
48. Lipton SA, Choi YB, Takahashi H, Zhang D, Li W, Godzik A (2002) Bankston LA 25:474–480
49. Takahashi H, Shin Y, Cho SJ, Zago WM, Nakamura T, Gu Z, Ma Y, Furukawa H, Liddington R, Zhang D, Tong G, Chen HSV, Lipton SA (2007) Neuron 53:53–64
50. Chen PE, Geballe MT, Stansfeld PJ, Johnston AR, Yuan H, Jacob AL, Snyder JP, Traynelis SF, Wyllie DJA (2005) Mol Pharmacol 67:1470–1484
51. Blaise MC, Sowdhamini R, Pradhan N (2005) J Mol Model 11:489–502
52. Yao Y, Mayer ML (2006) J Neurosci 26:4559–4566

J Mol Model (2012) 18:1285–1297
DOI 10.1007/s00894-011-1157-6

ORIGINAL PAPER

# Application of information theory to feature selection in protein docking

**Olaf G. Othersen · Arno G. Stefani · Johannes B. Huber · Heinrich Sticht**

Received: 20 May 2011 / Accepted: 21 June 2011 / Published online: 12 July 2011


**Abstract** In the era of structural genomics, the prediction of protein interactions using docking algorithms is an important goal. The success of this method critically relies on the identification of good docking solutions among a vast excess of false solutions. We have adapted the concept of mutual information (MI) from information theory to achieve a fast and quantitative screening of different structural features with respect to their ability to discriminate between physiological and nonphysiological protein interfaces. The strategy includes the discretization of each structural feature into distinct value ranges to optimize its mutual information. We have selected 11 structural features and two datasets to demonstrate that the MI is dimensionless and can be directly compared for diverse structural features and between datasets of different sizes. Conversion of the MI values into a simple scoring function revealed that those features with a higher MI are actually more powerful for the identification of good docking solutions. Thus, an MI-based approach allows the rapid screening of structural features with respect to their information content and should therefore be helpful for the design of improved scoring functions in future. In addition, the concept presented here may also be adapted to related areas that require feature selection for biomolecules or organic ligands.

O. G. Othersen · H. Sticht (✉)
Bioinformatik, Institut für Biochemie,
Friedrich-Alexander-Universität Erlangen-Nürnberg,
Fahrstr. 17,
91054 Erlangen, Germany
e-mail: h.Sticht@biochem.uni-erlangen.de

A. G. Stefani · J. B. Huber
Lehrstuhl für Informationsübertragung,
Friedrich-Alexander-Universität Erlangen-Nürnberg,
Cauerstr. 7,
91058 Erlangen, Germany

**Keywords** Protein interaction · Structure analysis · Feature selection · Protein interface · Mutual information

## Introduction

The specific recognition between proteins plays a crucial role in many biological processes. Examples include signal transduction processes, the regulation of gene expression by interactions between different transcription factors, the repression of enzymes by inhibitor proteins, and the recognition of pathogenic proteins by antibodies. An experimental determination of the respective complex structures is a time- and labor-intensive task, which is frequently hampered by limited availability of the biological material, by poor solubility, or the transient nature of the interaction. Therefore, it is important to develop computational methods that can predict the structures of complexes with proper accuracy. In the absence of a homologous complex of known three-dimensional structure, *docking* provides the only approach that can be used to predict the geometry of protein–protein complexes. In recent years, numerous docking procedures have been developed, and their performance is continuously evaluated in the CAPRI (critical assessment of predicted interactions) challenge [1–4].

Protein docking procedures generally consist of a sampling step and a subsequent scoring step. In the *sampling* step, a large collection of putative structural complexes are calculated, usually by exploiting the geometric fit between both proteins. The aim of the sampling step is to generate at least one docking solution that closely resembles the correct complex geometry. The generation of good docking solutions in the sampling step is necessary but insufficient for the successful completion of a docking

study, because the good docking solution still needs to be identified among the vast excess of bad docking solutions.

Most strategies developed for *scoring* evaluate the shape complementarity and the chemical character of the interacting surface patches [5–7]. Some algorithms also consider hydrogen bonds [8, 9], hydrophobic contacts [10], or electrostatic interactions [9, 11]. In particular, the calculation of the electrostatic interaction energy is sensitive to the positions of the atoms. Therefore, these approaches tend to fail in cases where the docked conformation has structural inaccuracies. Some of these strategies are also quite time-consuming and therefore less appropriate for screening hundreds of complexes.

Knowledge-based scoring functions, which are derived from a statistical analysis of experimentally determined protein complex structures, represent an alternative approach for evaluating docked complexes [12–15]. Most commonly, such functions are based on residue-type properties, such as frequencies of residue–residue pairs in the interface, but more global features such as the size of the interacting surface can also be considered. Advantageously, coarse-grained energy functions result in a smoothing of the energy landscape, which makes the scoring less prone to local structural inaccuracies.

The development of more powerful knowledge-based scoring schemes, however, is hampered by the identification of useful structural features. There are a wide variety of structural features (residue–residue contacts, buried surface area, type of secondary structure, evolutionary conservation of residues, …) which all need to be systematically assessed. In addition, once predictive features have been identified, their combination into a scoring function may be hampered by differences in dimension. As an example, pairwise amino-acid contacts will be counted as a dimensionless number, while accessible surface area buried upon complex formation has the dimension $Å^2$. Therefore, a widely applicable method that allows fast screening and the quantitative comparison of different structural features with respect to their predictive powers would be highly desirable.

In the present study, we adapted the concept of mutual information (MI) from information theory to identify structural features that are powerful in discriminating between good and bad docking solutions. An advantage of such an approach is that mutual information is dimensionless and can directly be compared for highly diverse structural features and also between datasets of different size. The utility of this approach will be demonstrated by calculating the mutual information in two model datasets of different sizes and for different structural features. Finally, the derived MI values will be converted into a simple scoring function, and we will demonstrate that features with a higher MI are actually more powerful for the identification of good docking solutions.

## Methods

### The concept of mutual information

The basic aim of this work is the identification of structural features that are informative for discriminating between good and bad docking solutions. To address this problem, the formalism of mutual information was adopted from information theory (Fig. 1a, Eq. 1).

Mutual information (MI) in information theory is a measure of the coupling between source $X$ and receiver $Y$ (Fig. 1a), and it tells us what we can learn about $X$ when we observe $Y$ (and vice versa) [16]. The mutual information $I(X; Y)$ depends only on the probability distributions $\Pr(X)$ and $\Pr(Y)$ of source $X$ and receiver signal $Y$, as well as their joint probability function $\Pr(XY)$:

$$I(X;Y) = \sum_{i=1}^{M_x} \sum_{j=1}^{M_y} \Pr\left(X = x_i, Y = y_j\right) \log_2 \left( \frac{\Pr\left(X = x_i, Y = y_j\right)}{\Pr(X = x_i)\Pr\left(Y = y_j\right)} \right). \quad (1)$$

In our scenario, the binary random variable $X$ expresses whether a protein docking solution is correct ("true situation") or wrong ("false situation"). Consequently, $X$ can adopt two values $X \in \{t,f\}$, where t designates the true and f the false docking solutions.

A second random variable $Y_c$ (where c stands for continuous) specifies the structural properties exhibited by these true and false docking solutions. There are multiple types of structural features (e.g., the number of salt bridges or the buried surface area upon complex formation) that were investigated separately to obtain their MI values (Fig. 1b). To avoid statistical problems resulting from the limited number of docked complexes available, the absolute values observed for a structural feature were grouped into discrete intervals (termed "$Y_c$ bins") according to their absolute values (Fig. 1c). This introduces the random variable $Y$, with realizations $y_j$ corresponding to the $Y_c$ bins.

The number of $Y_c$ bins required was qualitatively deduced from the following considerations. A minimum of two $Y_c$ bins is required to separate descriptor values corresponding to true and false docking solutions (Fig. 2a). However, two bins only work well if the structural feature investigated is clearly distinct in true and false docking solutions (e.g., all false solutions exhibit four or less salt bridges and all true solutions exhibit five or more salt bridges). In such a situation, the observation of a value within one of the two bins would be sufficient to conclude unambiguously whether the docking solution is true or false, respectively.

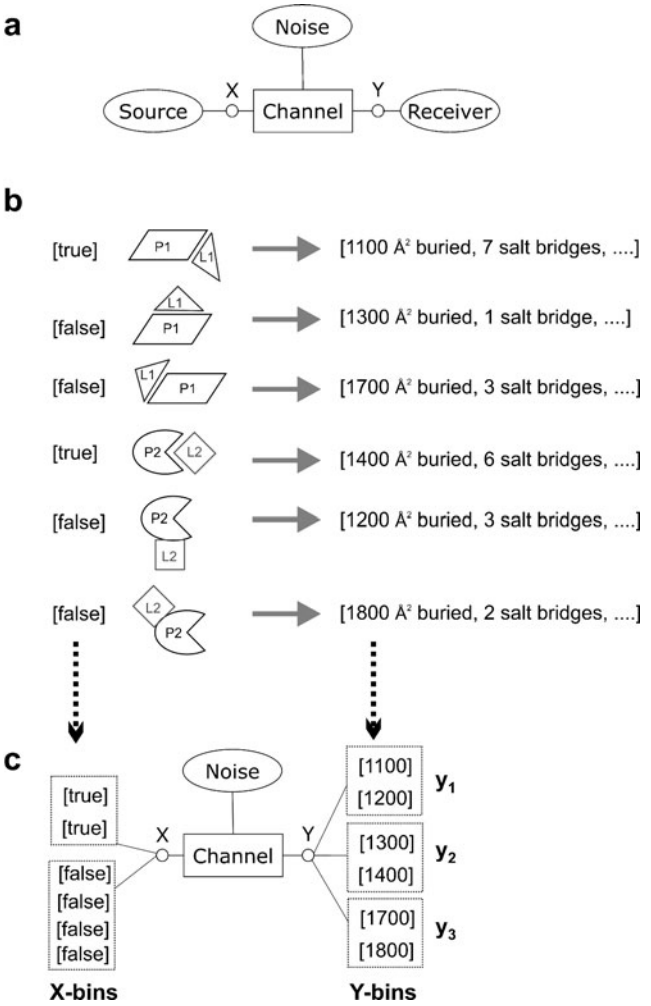



**Fig. 1** **a** Principle of digital communication over a noisy channel. By adapting this concept from information theory to protein docking, $X$ and $Y$ are defined as follows. The binary variable $X$ expresses whether a docking solution for the protein pair is correct or incorrect. $Y$ specifies a structural feature of the protein–protein interface that is observed in a docking solution. The concept of mutual information is used to deduce whether this feature results from a true or false docking solution in source variable $X$. Structural inaccuracies, the algorithms of docking and sampling, and the natural heterogeneity of protein–protein interfaces represent possible reasons for noise. **b, c** Implementation of the concept of MI in the present algorithm. In the first step, structural features are extracted from a set of docked complexes for both the correct ("true") and incorrect ("false") docking geometries. Panel **b** exemplarily shows one true and two false docking solutions for two docked complexes (P1-L1 and P2-L2). From these solutions, different structural features (e.g., buried surface area, number of salt bridges) are extracted (right side of panel **b**). **c** The true/false information (random variable $X$) serves as input for the channel, while the structural features are grouped into distinct $Y_c$ bins according to the absolute value. For this setup, the mutual information can be calculated over all combinations of $X$ and the $Y_c$ bins according to Eq. 1. The bins are shown as dotted boxes in panel **c**

In a more realistic scenario (Fig. 2b), the descriptor values of the true docking solutions are distributed over a particular range, and the features of false docking solutions exhibit higher and lower values but also overlap to a certain extent with the true solutions. To treat such a situation with the MI formalism, five bins ($y_1$, …,$y_5$) separated by four borders ($b_1$, …,$b_4$) appear to be appropriate (Fig. 2b). In this scenario, $y_1$ and $y_5$ predominantly contain false docking solutions, while $y_3$ predominantly contains true docking solutions. $y_2$ and $y_4$ correspond to $Y_c$ values at which both true and false docking solutions are observed.

For every structural feature, the four boundaries ($b_1$, …, $b_4$) were numerically optimized as described in the legend of Fig. 2. As a result of our binary source variable $X$ and the fivefold binning of the $Y_c$ values, there are five times two summands in the equation for the mutual information:

$$I(X;Y) = \sum_{j=1}^{5} I(\mathrm{t};y_j) + I(\mathrm{f};y_j) = \sum_{j=1}^{5} I(X;y_j). \quad (2)$$

An obvious advantage of the MI is that it is dimensionless, allowing a direct comparison of the information contents of different structural features *within* one dataset. MI values, however, depend on the ratio between the true and false docking solutions present in the dataset. To allow a comparison between different datasets, we have calculated a normalized MI ($MI_{norm}$) that takes into account the maximal MI that can be obtained in the respective dataset:

$$\mathrm{MI}_{\mathrm{norm}} = I(X;Y)/I_{\max}(X;Y), \quad (3)$$

and $I_{max}(X;Y)$ can be calculated according to

$$I_{\max}(X;Y) = \Pr(\mathrm{t})\log_2\frac{1}{\Pr(\mathrm{t})} + \Pr(\mathrm{f})\log_2\frac{1}{\Pr(\mathrm{f})}. \quad (4)$$

For example, for a t:f ratio of 1:100, Pr(t) is 1/101 and Pr(f) is 100/101. Therefore, $I_{max}$ is 0.0801 for the respective dataset. This maximal MI is achieved in a situation in which the descriptor values of true and false docking solutions are perfectly separated (Fig. 2a).

## Conversion of the MI into a scoring function

The $I(X;y_j)$ calculated according to Eq. 2 are always positive (as shown in the "Appendix"), and the higher $I(X;y_j)$, the higher the discriminative power of the $Y_c$ bin $y_j$. Therefore, high $I(X;y_j)$ values indicate a strong coupling between $X$ and $Y$ but do not state the origin of the coupling (e.g., a particular $Y_c$ bin might either be typical of true or for false docking solutions). Consequently, a high value of $I(X;y_j)$ does not necessarily indicate a true docking solution.

Given a certain feature value falling into $Y_c$ bin $y_j$, the idea here is to use $I(X;y_j)$ in an appropriate way for scoring. $I(X;y_j)$ itself cannot be directly used for scoring because it does not state if the actual $Y_c$ bin represents a true or a false solution. We therefore introduced a negative leading sign for $I(X;y_j)$ if false solutions are overrepresented in the $Y_c$ bin $y_j$ ($\Pr(y_j|\mathrm{f}) > \Pr(y_j)$), and left the sign positive if true

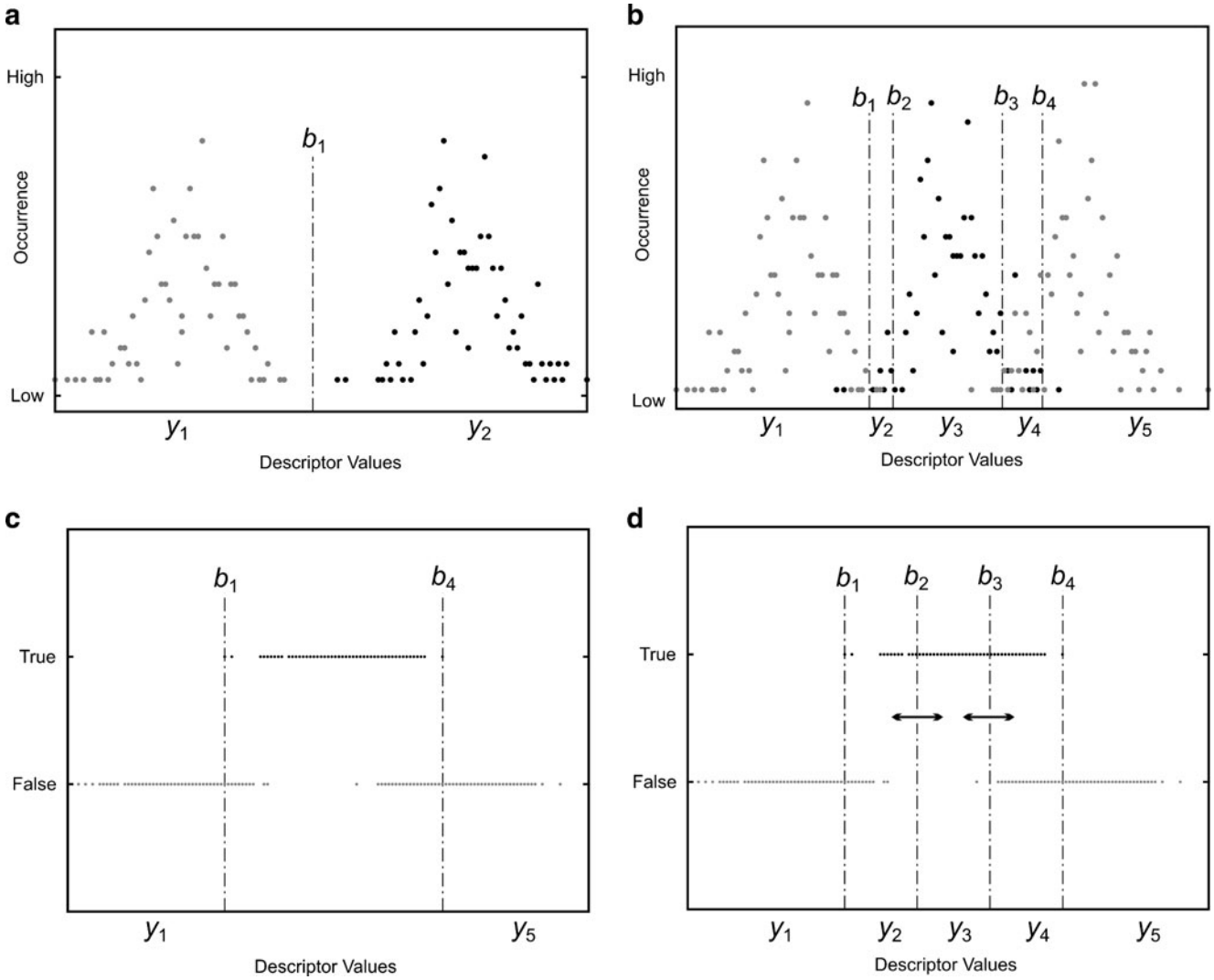



**Fig. 2** Different scenarios for the distribution of good and bad docking solutions, and strategy for determining suitable boundaries to dissect the dataset. **a** Ideal scenario, in which the true (*black*) and false (*gray*) docking solutions are perfectly separated. In such a situation, a single boundary ($b_1$) which creates two bins ($y_1$,$y_2$) is sufficient and results in the maximal possible MI. **b** Realistic scenario, in which the descriptor values of the true docking solutions are distributed over a particular range, and the false docking solutions exhibit higher and lower descriptor values but also overlap to a certain extent with the true solutions. Such a situation can be treated by assigning five $Y_c$ bins ($y_1$, …,$y_5$) separated by four borders ($b_1$, …,$b_4$). This setup was chosen in the present study. **c, d** Strategy for determining the boundaries of the $Y_c$ bins. The first borders are assigned to separate high and low $Y_c$ bins if these are populated exclusively by true or false docking solutions (see $b_1$ and $b_4$ in panel **c**, which separate bins $y_1$ and $y_5$). The unassigned remaining borders are then distributed equidistantly over the remaining descriptor range and subsequently systematically varied by one step towards higher and lower descriptor values (see *black arrows* in panel **d**). Within each optimization passage, the border variation leading to the largest increase in MI is kept while the other changes are discarded. This process is repeated until varying the borders yields no further increase in the mutual information

solutions were overrepresented. The resulting value is named $MI_{score}$. Consequently, a positive value of $MI_{score}$ is indicative of true solutions, whereas a negative value indicates false docking solutions, while the absolute value states in a way how sure we can be about the decision.

Using this formalism, a simple scoring function was created by assigning an $MI_{score}$ to every $Y_c$ bin. The subsequent scoring of a structural feature then only requires the identification of the $Y_c$ bin corresponding to the feature value and the assignment of the $MI_{score}$ of this $Y_c$ bin.

## Dataset used for method development

The MI-based approach was developed and tested using the Dockground dataset [17–19]. This dataset was designed as an unbound benchmark set for the development of novel scoring procedures and contains non-native (“false”) and near-native (“close”) docking solutions for 61 different protein–protein complexes. The near-native solutions are characterized by an RMSD of less than 5 Å with respect to the correct complex geometry. The MI values of different

structural features were investigated for two different scenarios:

1) Comparison of physiological (true) and nonphysiological (false) complex geometries. For this purpose, our dataset included 100 false docking solutions as well as the correct complex crystal structure for each of the 61 complexes of the Dockground dataset. The respective set is denoted "t/f" throughout the manuscript.
2) Additional consideration of near-physiological (close) complex geometries. This corresponds to a more realistic scenario in protein–protein docking, because docking algorithms may fail to generate the physiological complex geometry in the sampling step, but generally produce a number of near-physiological solutions.

For this second dataset, the 61 correct complex geometries were supplemented by the 505 close docking solutions provided by Dockground. The set of 6100 false docking solutions (100 for each of the 61 complexes investigated) remained unchanged. This dataset, which compares true and close versus false solutions, is henceforth termed the "tc/f" dataset.

For the final investigation of the performance of the MI approach, a fivefold cross-validation was performed. For this purpose, the original set was divided into five sets each containing twelve docking cases with all the respective true, close and false docking solutions. Four of these five sets were combined to form the training set, and the performance of the approach was measured using the fifth set as a test set. This procedure was repeated five times for different combinations of training and test sets.

### Types of structural features investigated

Within this paper, we basically investigated two types of structural features: (i) pairwise interface contacts between different classes of amino acids, and (ii) the size of the solvent-accessible surface area buried upon complex formation. This setup was chosen to assess the ability of an MI-based approach to allow a direct comparison between diverse features differing in their dimensions.

To reduce the types of pairwise interface contacts and to improve statistics, the 20 different amino acids were clustered into four groups according to their physicochemical properties. These groups were $AA^+$ (Arg, His, Lys), $AA^-$ (Asp, Glu), $AA_{hb}$ (Asn, Gln, Ser, Thr, Trp, Tyr), and $AA_{vdw}$ (Ala, Cys, Gly, Ile, Leu, Met, Phe, Pro, Val). The grouping of amino acids into four classes resulted in a total of ten different types of contacts, which were investigated separately with respect to their information content. According to the concept of a residue-based potential, only one contact was counted per pair of interface residues. In this context, the distance of such a residue–residue contact is defined as the closest atomic distance between the amino acids involved. Interface residues are defined as having at least one atom that is <7 Å from the docking partner. The buried surface area was calculated as the difference between the solvent accessible surface areas (SASA) of the free binding partners and their complex using DSSP [20].

## Results and discussion

### Properties of protein interfaces in docking solutions: accessible surface

The properties of physiological (true), near-physiological (close), and false protein interfaces were analyzed for 61 different complex types, as described in "Methods." Figure 3a shows a comparison of the surface area buried upon complex formation in the true and false docking solutions. It can be seen that the true interfaces tend to have smaller buried surfaces than the false solutions, indicating that molecular recognition does not involve the largest complementary patches in shape. This trend is also preserved but less pronounced when close docking solutions are included in the dataset (Fig. 3b). The vertical lines in Fig. 3 indicate the borders of the $Y_c$ bins, which were numerically optimized to maximize the MI.

The MI, which represents the sum over all bins (see Eq. 2), reaches values of 0.0163 and 0.0471 for the situation in shown Fig. 3a and b, respectively (Table 1, second column). The values correspond to 20.3% and 11.2% of the MI that is maximally attainable in the respective datasets (see the $MI_{norm}$ values in Table 1). When comparing the $MI_{norm}$ values, it becomes evident that the buried surface is more appropriate for discrimination in the t/f compared to the tc/f dataset. This is also consistent with a qualitative interpretation of Fig. 3, which shows that the close solutions overlap more with the false solutions than the true solutions.

### Properties of protein interfaces in docking solutions: pairwise residue contacts

Contacts between different types of amino acids represent a key feature guiding complex formation. In the present study, intermolecular contacts were analyzed between four different classes of amino acids: positively charged ($AA^+$), negatively charged ($AA^-$), polar ($AA_{hb}$), and nonpolar ($AA_{vdw}$).

The number of interface contacts detected in true and false docking solutions is exemplarily shown in Fig. 3c for the interaction of negatively charged and nonpolar residues. Generally, fewer contacts of this type are found in true

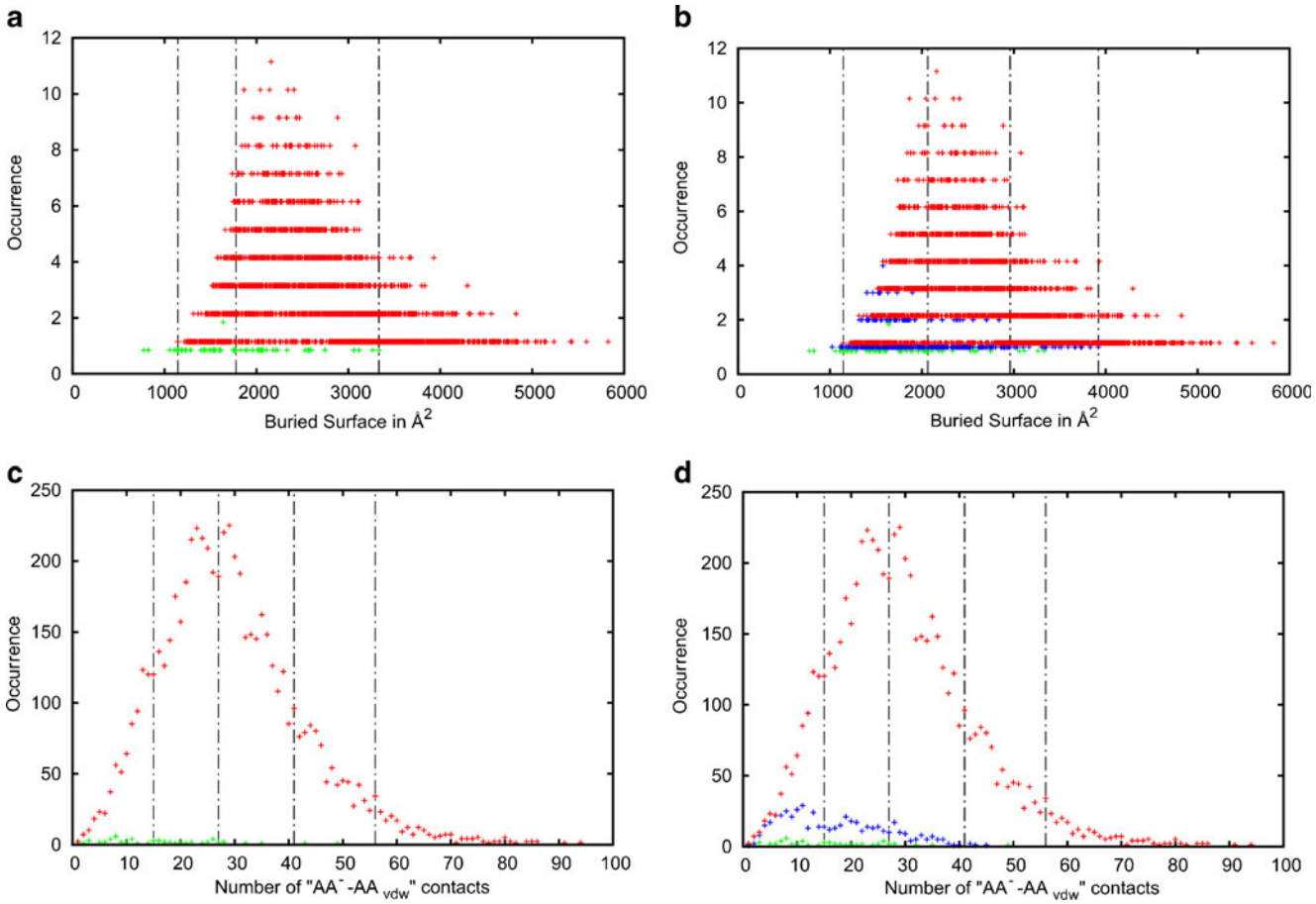



**Fig. 3** Distribution of different structural features in the good and bad docking solutions. **a** Size of the surface area that is buried upon complex formation in the true (*green*) and false (*red*) docking solutions. Occurrence values larger than 1 indicate that the respective numerical value for the buried surface area was observed in more than one docking solution. *Dotted vertical lines* indicate the $Y_c$ boundaries that were obtained by the strategy described in Fig. 2. **b** Same presentation as in **a**, but for the tc/f dataset, which also includes near-physiological docking solutions (*blue*). Four $Y_c$ boundaries were calculated in both **a** and **b**, but two of the boundaries are very close together in the t/f dataset shown in **a**. **c** Number of $AA^-$–$AA_{vdw}$ contacts that are formed upon complex formation in the true (*green*) and false (*red*) docking solutions. Occurrence values larger than 1 indicate that the respective number of contacts was observed in more than one docking solution. *Dotted vertical lines* indicate the $Y_c$ boundaries that were obtained by the strategy described in Fig. 2. **d** Same presentation as in **c**, but for the tc/f dataset, which also includes near-physiological docking solutions (*blue*)

compared to false docking solutions. This is consistent with biophysical considerations that the burial of a negative charge within a hydrophobic surface is highly unfavorable because it requires a high desolvation energy [21]. A very similar distribution is obtained when the close docking solutions are considered in addition to the true complex geometry (Fig. 3d). Calculation of the MI over the five bins shown in Fig. 3c and d results in $MI_{norm}$ values of 12.7%

**Table 1** Mutual information $I(X;Y)$ of the 11 structural features investigated in the present study. MI values were calculated separately for the t/f and tc/f datasets. Since these two datasets exhibit different maximal MI values, the normalized MI values ($MI_{norm}$) are also listed to allow a comparison between both datasets

| Structural feature | | Buried surface | $AA^+$-$AA^+$ | $AA^+$–$AA^-$ | $AA^+$-$AA_{hb}$ | $AA^+$-$AA_{vdw}$ | $AA^-$-$AA^-$ | $AA^-$-$AA_{hb}$ | $AA^-$-$AA_{vdw}$ | $AA_{hb}$-$AA_{hb}$ | $AA_{hb}$-$AA_{vdw}$ | $AA_{vdw}$-$AA_{vdw}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| t/f | $I(X;Y)$ | 0.0163 | 0.0053 | 0.0035 | 0.0072 | 0.0058 | 0.0075 | 0.0080 | 0.0102 | 0.0033 | 0.0036 | 0.0016 |
| | $MI_{norm}$ (%) | 20.3% | 6.6% | 4.4% | 9.0% | 7.2% | 9.4% | 10.0% | 12.7% | 4.1% | 4.5% | 2.0% |
| tc/f | $I(X;Y)$ | 0.0471 | 0.0385 | 0.0197 | 0.0291 | 0.0531 | 0.0429 | 0.0450 | 0.0636 | 0.0073 | 0.0114 | 0.0019 |
| | $MI_{norm}$ (%) | 11.2% | 9.2% | 4.7% | 6.9% | 12.7% | 10.2% | 10.7% | 15.2% | 1.7% | 2.7% | 0.5% |

and 15.2% for the t/f and tc/f datasets, respectively (Table 1).

All other structural features were analyzed accordingly and the results are summarized in Table 1. When comparing the structural features, it becomes immediately apparent that they differ significantly in their information contents for dissecting physiological and nonphysiological protein interfaces. A high information content is generally observed for interactions involving charged amino acids, while interactions involving only uncharged amino acids in both binding partners have minor information content (see the last three columns in Table 1).

Interestingly, for most of the contact types investigated, the $MI_{norm}$ obtained is quite similar for the t/f and tc/f datasets (Table 1). This observation suggests that the MI values are quite robust with respect to small structural inaccuracies present in the near-physiological (close) solutions. This property is considered useful for a realistic docking scenario, in which the true complex geometry might not be present among the docking solutions.

All of the MI values above were calculated using a cutoff distance of 7 Å , which was chosen based on initial test calculations showing that this threshold generally results in high MI values. This can, for example, be seen by plotting the $MI_{norm}$ values of the $AA^{-}$-$AA_{vdw}$ feature as a function of the cutoff distance (Fig. 4). For both datasets investigated, the maximal $MI_{norm}$ is observed at cutoff distances very close to 7 Å. The plot is smoother for the tc/f compared to the t/f dataset, which can be attributed to the larger number of good solutions in the tc/f dataset, resulting in better statistics. The 7 Å cutoff distance determined in the present study is also in good agreement with a previous statistical analysis of protein interfaces, which used cutoff distances in the range 5–10 Å [22–25].

Furthermore, the high MI values obtained in the present analysis demonstrate that the clustering of the 20 amino acids into four classes still allows meaningful analysis of intermolecular contact preferences. This is in line with a previous study showing that a reduced alphabet of six amino acid classes is sufficient to encode the specificity of protein–protein interactions [26].

## MI-based scoring scheme

The investigations above demonstrate that even simple structural features exhibit a significant amount of mutual information in discriminating between good and bad docking solutions. An advantage of the MI values in Table 1 is that they are dimensionless and can be directly compared and combined for entirely different structural features, like buried surface and pairwise residue contacts. The conversion of mutual information into a scoring function, however, is hampered by the fact that MI values are always positive. Therefore, it is not immediately apparent whether an observed structural feature is indicative of a good or a bad docking solution. This information, however, can be obtained by inspecting the individual summands $I(x_i;y_j)$ contributing to the MI. The use of two $X$ symbols representing the true and false docking solutions and five $Y_c$ bins, which cluster docking solutions with similar structural properties, results in a total of ten individual MI summands (two for each $Y_c$ bin; see Eq. 2).

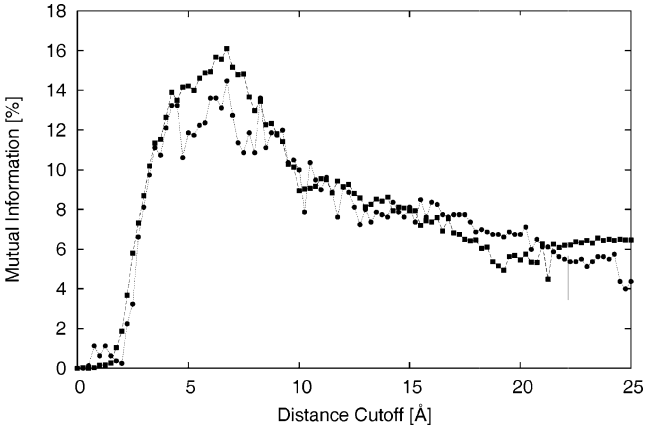



**Fig. 4** Dependence of the mutual information on the cutoff distance applied. The graph shows changes in the normalized MI values for the $AA^{-}$-$AA_{vdw}$ feature when the cutoff for counting intermolecular contacts is systematically varied from 0 Å to 25 Å. MI values for the t/f and tc/f datasets are presented as *circles* and *squares*, respectively

The values for these individual summands are listed explicitly in Table 2 for the $AA^{-}$-$AA_{vdw}$ contacts in the t/f dataset. The first row in Table 2 shows the range of the five $Y_c$-bins with respect to the number of $AA^{-}$-$AA_{vdw}$ contacts clustered in these bins. These bins and their population with true and false docking solutions are graphically represented in Fig. 3c.

For example, $y_1$ clusters all docking solutions that exhibit fewer than twelve $AA^{-}$-$AA_{vdw}$ contacts in their interface. This bin contains 28 (45.9%) of the 61 true and 375 (6.15%) of the 6100 false docking solutions (Table 2; Fig. 3c). These values indicate that true docking solutions are enriched in the respective bin, resulting in a higher information content $I(t;y_1) = +0.0128$ for the true docking solutions compared to the false docking solutions $I(f;y_1) = -0.0054$ (Table 2).

To allow scoring of the docked solutions, we defined a variable $MI_{score}$. In the current implementation, $MI_{score}$ has the same absolute value as $I(X;y_j)$ but additionally contains a leading sign that indicates whether true or false docking solutions are overrepresented in the respective bin (see "Methods" for details). The $MI_{score}$ of $y_1$, which contains many of the true docking solutions, is +0.0073, while $y_5$, which only contains false docking solutions, has a negative $MI_{score}$ of −0.0010.

**Table 2** Conversion of mutual information $I(X;Y)$ into a scoring function using the $AA^{-}$-$AA_{vdw}$ feature in the t/f dataset as an example. The table lists the distribution of the true ($X$ = t) and false ($X$ = f) docking solutions in the five bins $y_1$ to $y_5$. $N$ denotes the number of $AA^{-}$-$AA_{vdw}$ interface contacts present in the docking solutions clustered in the respective bin. The overrepresented $X$ value is colored in *gray* and its contribution $I(x_i;y_j)$ to the MI is generally positive. For each bin, the magnitude of the $MI_{score}$ is defined by the value of $I(X;y_j) \geq 0$, but a negative leading sign is added if false solutions are overrepresented in the respective $Y_c$ bin (see the "Methods" section for details and Fig. 3c for a graphical presentation of the docking raw data)

| | $y_1$: $N<12$ | | $y_2$: $12\leq N<26$ | | $y_3$: $26\leq N<38$ | | $y_4$: $38\leq N<50$ | | $y_5$: $50\leq N$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $X$=t | $X$=f | $X$=t | $X$=f | $X$=t | $X$=f | $X$=t | $X$=f | $X$=t | $X$=f |
| No. cases | 28 | 375 | 21 | 2243 | 10 | 2095 | 2 | 940 | 0 | 447 |
| % Cases | 45.9% | 6.1% | 34.4% | 36.8% | 16.4% | 34.3% | 3.3% | 15.4% | 0.0% | 7.3% |
| $I(x_i;y_j)$ | 0.0128 | −0.0054 | −0.0003 | 0.0003 | −0.0017 | 0.0025 | −0.0007 | 0.0017 | 0.0000 | 0.0010 |
| $I(X;y_j)$ | 0.0073 | | 0.0000 | | 0.0008 | | 0.0010 | | 0.0010 | |
| $MI_{score}$ | +0.0073 | | −0.0000 | | −0.0008 | | −0.0010 | | −0.0010 | |
| $I(X;Y)$ | 0.0073+0.0000+0.0008+0.0010+0.0010=0.0102 (12.7% of max. MI) | | | | | | | | | |

Comparing the magnitude of the $MI_{score}$ for $y_1$ and $y_5$ reveals an additional interesting feature of the MI that differs from most conventional scoring functions: $y_5$ exhibits a smaller absolute $MI_{score}$ value than $y_1$, although it contains exclusively false docking solutions. Intuitively, one might expect that this situation leads to the maximal MI for $y_5$, therefore unambiguously indicating a false docking solution.

However, the absolute value of the individual summand $I(x_i;y_j)$ also depends on the number of observations made in the respective bin. Since $y_5$ contains only 7.3% of the false docking solutions, $I(X;y_j)$ only reaches a value of 0.0010 (Table 2). Thus, the $MI_{score}$ reflects not only the ratio of true and false docking solutions within a bin but also the population of the bin itself. The latter feature stops high scores from being assigned to lowly populated bins, thus stopping true/false decisions from being made on the basis of structural properties that are only very rarely observed.

## $MI_{score}$ values of different structural features

The $MI_{score}$ has been calculated individually for all structural features investigated, and is summarized in graphical form in Fig. 5. The properties observed for the t/f (Fig. 5a) and the tc/f (Fig. 5b) datasets are qualitatively similar. In both datasets, the most informative features for the identification of a good docking solution are the surface area buried upon complex formation as well as the contacts between charged and nonpolar amino acids ($AA^{-}$-$AA_{vdw}$, $AA^{+}$-$AA_{vdw}$).

The high MI scores observed for small buried surface areas reflect the observation that a physiological complex is not necessarily formed by the largest interface that would be feasible based on the shape complementarity of both binding partners. Therefore, docking algorithms, which mainly select complex structures based on the size of the interacting surface, tend to miss good docking solutions. The $MI_{score}$ of the buried surface area may be particularly large in the present study, because we have deliberately chosen the DOCKGROUND dataset, which was generated without extensive post-processing of the docked solutions, thus providing a raw dataset for own method development.

Our analysis also revealed that a small number of contacts between charged and nonpolar amino acids is indicative of a good docking solution (Fig. 5). This observation reflects the fact that burying a charged amino acid in a nonpolar interface is energetically very unfavorable because it requires desolvation of the charged amino acid prior to complex formation, while no compensating favorable interactions can be formed by the charged residue with its nonpolar counterpart.

The particular role of charged residues in protein interfaces is also supported by independent computational studies, which indicate that binding sites used in nature are characterized by a low desolvation penalty [21], and this property has already been exploited for scoring docked protein complexes [27]. Our finding is also in line with recent experimental studies which demonstrated that replacing nonpolar residues with charged ones represents a highly efficient way to disrupt protein interfaces [28, 29].

The remaining types of contacts investigated in the present study exhibit a lower MI than the charged–nonpolar contacts discussed above. Nevertheless, the magnitudes of the $MI_{score}$ values observed for the different intervals (Fig. 5) give additional insight into the structural properties that are favorable for complex formation. For all interactions involving at least one polar interaction partner, a relatively low number of contacts (generally less than 20) is indicative of good docking solutions. At first sight this might be surprising, because one may expect that, at least for the attractive interactions (e.g., $AA^{+}$-$AA^{-}$), maximizing the number of interactions would be desirable in a physiological protein interface. The observation that only a limited number of these types of interactions are formed

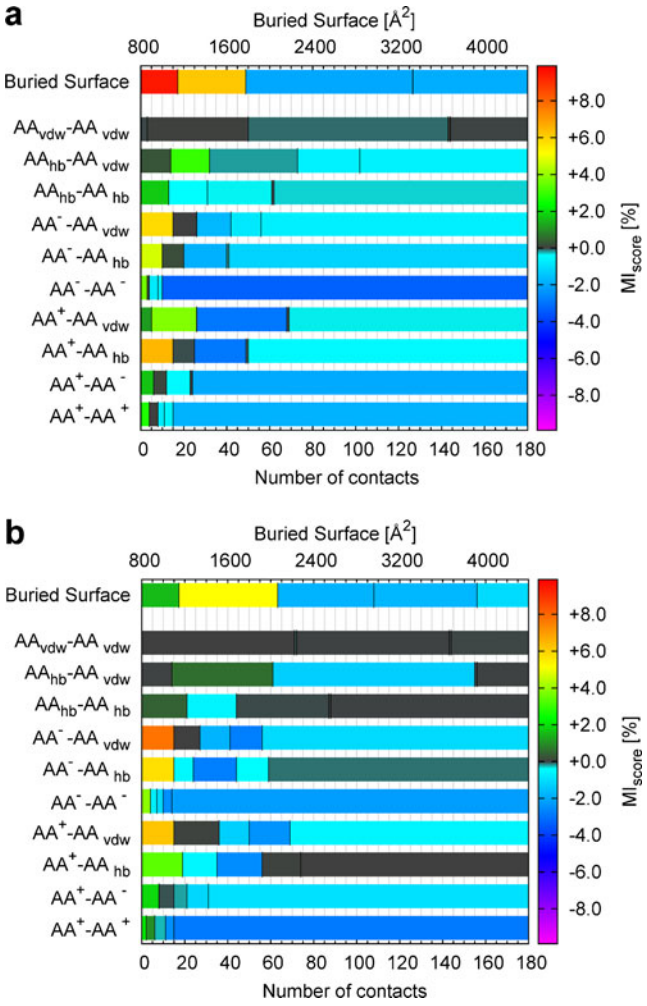



**Fig. 5** $MI_{score}$ values of different structural features in the t/f (**a**) and tc/f (**b**) datasets. The diagrams show the range of the $Y_c$ bins for each of the 11 structural features investigated. All bins are color coded according to their $MI_{score}$. $MI_{score}$ values have been normalized to allow direct comparison of the datasets shown in **a** and **b**

in physiological interfaces might be explained by the fact that most interfaces are not designed to achieve the maximal stability, but rather to allow for transient protein–protein interactions [30, 31]. Our approach takes this consideration into account by defining five $Y_c$ bins, which group the number of contacts observed in a protein interface into five intervals. The optimization of the ranges of these bins in order to maximize the MI allows the identification of intervals that are characteristic of good docking solutions.

This concept is different from conventional types of analysis, in which a type of interaction is considered either favorable or unfavorable. Our investigation, however, shows that none of the contact types investigated tend to be maximized in good docking solutions, and that the number of contacts found in good solutions is actually distributed over a distinct range (Fig. 5). In summary, the results above suggest that an MI-based score should be able to distinguish between good and bad docking solutions. To address this point in more detail, we applied the $MI_{score}$ to identify good docking solutions in a cross-validated dataset.

## Application of the scoring scheme to a cross-validated dataset

For cross-validation, the original dataset of 61 complexes was divided into five sets each containing twelve docking cases with all the respective true, close and false docking solutions (one of the 61 original complexes was randomly discarded to obtain five datasets of equal size). Four of these five sets were combined to form the training set, and the performance of the approach was measured using the fifth set as test set. MI scores for the training set of 48 complexes were calculated in an identical fashion to that described in the section above and subsequently used to score the remaining 12 complexes. This procedure was repeated five times for different combinations of training and test sets, finally resulting in unbiased predictions for 60 types of complexes.

**Table 3** Performance of the $MI_{score}$ in the identification of good docking solutions in fivefold cross-validated datasets. The $MI_{score}$ was derived from either single structural features ($AA^{-}$-$AA_{vdw}$, $AA_{hb}$-$AA_{vdw}$) or as an additive score that takes into account all of the structural features investigated ($\Sigma MI_{score}$; see text for details). The c/f dataset was generated from the tc/f dataset by removing the true docking solutions from the test set prior to prediction. A total of 60 different complexes were investigated in the cross-validated approach, and the docking solutions were ranked according to their $MI_{score}$. The numbers listed for each rank state the number of complexes for which a good docking solution was found on the respective rank. For example, the $AA^{-}$-$AA_{vdw}$ based $MI_{score}$ identified the correct docking solution for 11 of the 60 complexes investigated, while the correct solution was ranked worse than rank 9 for 37 of the docking cases

| Feature | Dataset | Rank | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | >9 |
| $AA^{-}$-$AA_{vdw}$ | t/f | 11 | 3 | 1 | 1 | 2 | 1 | 1 | 1 | 2 | 37 |
| $AA_{hb}$-$AA_{vdw}$ | t/f | 3 | 4 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 51 |
| All 11 features | t/f | 28 | 3 | 4 | 2 | 1 | 0 | 1 | 2 | 0 | 19 |
| All 11 features | tc/f | 41 | 4 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 11 |
| All 11 features | c/f | 37 | 5 | 1 | 1 | 0 | 1 | 0 | 2 | 1 | 12 |

Initially, we tested the performance of *single* structural features for the identification of correct docking solutions in the cross-validated t/f dataset. The results of using MI scores derived from single features are shown in Table 3 exemplarily for $AA^{-}$-$AA_{vdw}$ and $AA_{hb}$-$AA_{vdw}$. Using $AA^{-}$-$AA_{vdw}$ as the sole criterion allows the identification of the correct complex geometry as the best docking solution (first rank) among 100 false solutions in 11 of the 60 docking cases tested. $AA_{hb}$-$AA_{vdw}$ performs worse and the correct solution is ranked best for only 3 of the 60 docking cases studied (Table 3). These findings are interesting for two reasons:

1) They reveal a correlation between the MI (Table 1) of a particular feature and the performance of this feature as a scoring function. Features with a higher MI also prove to be more successful in the identification of good docking solutions. Therefore, the calculation of the MI represents a suitable formalism to identify the most predictive features to be incorporated in a scoring function.
2) The findings above also demonstrate that single criteria only perform well for a rather small subset of the complexes investigated. For example, $AA^{-}$-$AA_{vdw}$ works perfectly for 11 of the 60 docking cases, but fails completely for 37 of the complexes.

Consequently, we defined a novel scoring function (termed $\Sigma MI_{score}$), which simply takes into account the $MI_{score}$ of all 11 structural features in an additive fashion, weighting all 11 features uniformly. As an example, the individual MI scores and the $\Sigma MI_{score}$ are shown for close and false docking solutions of the 1akj docking case in Table 4.

Figure 6 shows the cross-validated $\Sigma MI_{score}$ for all docking solutions in the t/f and the tc/f datasets. From the distribution of the good and bad docking solutions, it becomes evident that the $\Sigma MI_{score}$ does indeed have the predictive power to discriminate between good and bad docking solutions. This property becomes even more evident from an individual inspection of the 60 docking cases (Table 3). The third line shows the result for the t/f dataset when the $\Sigma MI_{score}$ derived from 11 features is used. The improved performance compared to the scoring based on individual features becomes immediately apparent. While the $MI_{score}$ values derived from single features rank the true solution first in 3 or 11 of the 60 complexes, the $\Sigma MI$ score succeeds in 28 of 60 cases.

The $\Sigma MI_{score}$ also performs even better for the tc/f dataset (Table 3). In this dataset, the correct or a near-physiological (close) docking solution is ranked first for 41 of the 60 complexes. This number, however, must be approached with caution, since realistic docking experiments frequently do not reproduce the correct geometry in the sampling step. Therefore, a good scoring function should also be able to perform well when only near-

**Table 4** Example showing the $\Sigma MI_{score}$ values for four representative solutions of the 1akj docking case from the Dockground dataset. For each docking solution, the 11 structural features that contribute to the $\Sigma MI_{score}$ are listed; the first line in the respective cell states the descriptor value (e.g., 7 $AA^{+}$-$AA^{-}$ contacts) and the second line gives the corresponding $MI_{score}$. Ranking was performed based on the $\Sigma MI_{score}$.The four solutions shown correspond to the best close (rank 1), best false (rank 4), worst close (rank 16), and worst false (rank 110) docking solution

| Rank | Close/false | $AA^{+}$ - $AA^{+}$ | $AA^{+}$ - $AA^{-}$ | $AA^{+}$ - $AA_{hb}$ | $AA^{+}$ - $AA_{vdw}$ | $AA^{-}$ - $AA^{-}$ | $AA^{-}$ - $AA_{hb}$ | $AA^{-}$-$AA_{vdw}$ | $AA_{hb}$-$AA_{hb}$ | $AA_{hb}$-$AA_{vdw}$ | $AA_{vdw}$-$AA_{vdw}$ | Buried surface | $\Sigma MI_{score}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | c | 2 | 7 | 14 | 16 | 1 | 31 | 13 | 39 | 24 | 2 | 2057.0 | |
| | | 0.0039 | 0.0003 | 0.0169 | 0.0055 | 0.0147 | −0.0184 | 0.0003 | −0.0028 | 0.0044 | 0.0000 | 0.0216 | 0.0464 |
| 4 | f | 9 | 13 | 22 | 39 | 2 | 18 | 22 | 18 | 65 | 38 | 2831.0 | |
| | | −0.0009 | −0.0012 | −0.0019 | −0.0047 | 0.0147 | 0.0049 | 0.0003 | −0.0028 | −0.0061 | 0.0003 | −0.0075 | −0.0048 |
| 16 | c | 6 | 18 | 27 | 25 | 3 | 38 | 14 | 43 | 45 | 7 | 2838.0 | |
| | | 0.0039 | −0.0012 | −0.0019 | −0.0047 | 0.0001 | −0.0184 | 0.0003 | −0.0001 | 0.0044 | 0.0000 | −0.0075 | −0.0252 |
| 110 | f | 15 | 27 | 36 | 51 | 15 | 35 | 56 | 20 | 70 | 41 | 3951.0 | |
| | | −0.0156 | −0.0044 | −0.0102 | −0.0098 | −0.0140 | −0.0184 | −0.0086 | −0.0028 | −0.0061 | 0.0003 | −0.0044 | −0.0940 |

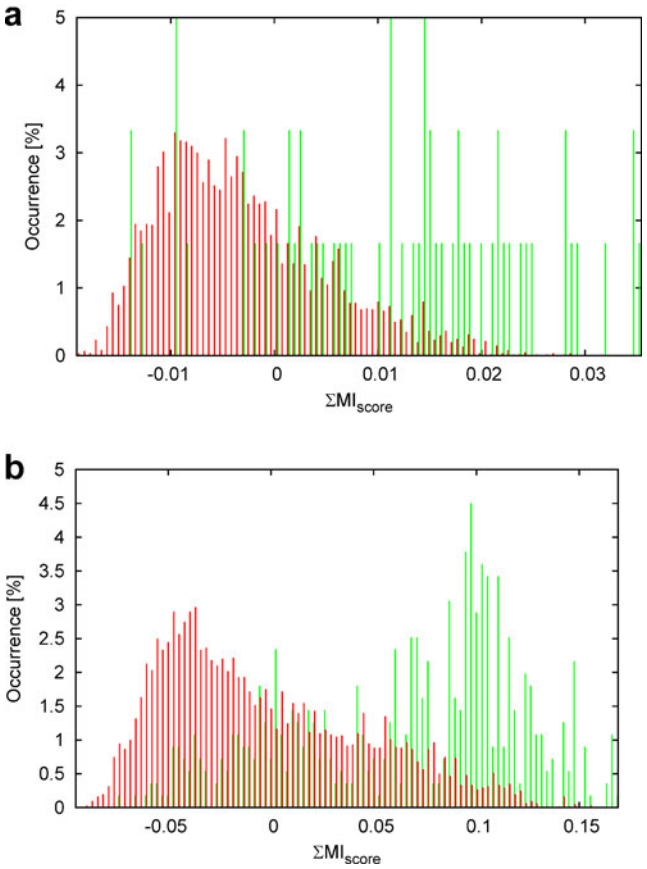



**Fig. 6** $\Sigma MI_{score}$ values for good (*green*) and bad (*red*) docking solutions in the t/f (**a**) and tc/f (**b**) datasets. Scoring was performed on the basis of all 11 structural features by a fivefold cross-validated approach, as described in "Methods." To enhance readability, the heights of the bars have been scaled with respect to the overall number of complexes; e.g., a true docking solution represents 1/60 (1.66%) and a false docking solution represents 1/6000 (0.0166%) of the respective dataset. The diagram summarizes the results for all 60 complexes investigated. A dissection of the performance with respect to the 60 individual docking cases is given in Table 3

physiological docking solutions are produced in the sampling step. We have assessed this scenario via the following setup. A fivefold cross-validation was performed as described above, but the true docking solutions were removed from the test set prior to prediction. The last line in Table 3 shows that the results are only marginally affected by this step, and that a near-physiological complex geometry is identified as the best docking solution for 37 of the 60 complexes studied. This result demonstrates that the MI-based scoring function is capable of coping with structural inaccuracies present in the near-physiological solutions, and allows easy combination of different structural features.

However, we want to emphasize that the implementation above does not represent a final scoring function; it is only intended to demonstrate the utility of MI for the characterization of protein interfaces. Future steps towards a final scoring function will include additional larger docking datasets, the consideration of additional structural features, a different weighting of individual structural features, or the development of a more rigid formalism for determining the size and number of the $Y_c$ bins.

## Conclusions

Initial biological applications of information theory used Shannon's methods to measure the information content in strings of monomers such as genes, RNA, and proteins (see [32] for a review), but more recent studies have also demonstrated the utility of information theory in structural biology. Mutual information has been used to identify amino acids relevant to DNA binding [33], and for the prediction of active sites [34]. Information theory has also been applied to calculate the binding sites [35] and the binding affinities [36] of small organic ligands. In addition, mutual information proved to be successful for the identification of class-specific molecular descriptors [37]. Information theory has been applied for fold recognition and for the identification of correct protein models [38–41].

In the present work, the concept of mutual information was adapted from information theory to address another biological problem: the identification of a good docking solution among a vast excess of false complex geometries. Key features of this method, which differ from most traditional approaches, include the explicit consideration of false docking solutions, as well as the clustering of structural properties into discrete ranges (termed "$Y_c$ bins").

As an advantage, mutual information is dimensionless and can be directly compared for datasets of different sizes and for highly diverse structural features (e.g., number of salt bridges and the buried accessible surface area). In addition, the MI reflects not only the ratio of true and false docking solutions but also the number of observations itself. This property means that true/false decisions are not made solely on the basis of structural properties that are generally only very rarely observed.

Conversion of the mutual information into a simple scoring function reveals that features with a higher MI also prove to be more successful at identifying good docking solutions. Therefore, the calculation of the MI represents a suitable formalism to identify the most predictive features to be incorporated in a scoring function. The present scoring scheme performs almost equally well for the identification of physiological and near-physiological complex geometries, indicating that the MI values are quite robust with respect to small

structural inaccuracies. This property should be helpful when treating realistic docking scenarios, for which the true complex geometry might not be present among the docking solutions. Thus, the present approach offers an alternative avenue to the generation of improved scoring functions in future, which should be helpful for addressing the docking problem.

In addition, the concept presented here is not limited to structural features; it can also be applied to a variety of molecular properties like mass, solubility, toxicity, or other physical quantities. Therefore, an MI-based approach should also prove useful in related areas that require feature selection for biomolecules or organic ligands.

**Acknowledgments** The authors thank Kristin Kassler and Dr. Christophe Jardin for critically reading the manuscript. The project was funded within the DFG (Deutsche Forschungsgemeinschaft) priority program (SPP 1395) by grants to JH and HS.

## Appendix

Proof that $\underline{I(X;y_j)}$ is always positive

We will show that

$$I(X;y_j) = \sum_{i=1}^{M_x} \Pr(x_i,y_j) \log_2 \frac{\Pr(x_i,y_j)}{\Pr(x_i)\Pr(y_j)}$$

is always greater than or equal to 0. For this purpose, the log sum inequality is quoted first [16]:

$$\sum_{i=1}^{n} a_i \log_2 \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i\right) \log_2 \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \text{ for any } a_i, b_i \geq 0, i = 1 \ldots n.$$

With $a_i = \Pr(x_i,y_j)$, $b_i = \Pr(x_i)\Pr(y_j)$ and $n = M_x$, it follows that

$$I(X;y_j) = \sum_{i=1}^{M_x} \Pr(x_i,y_j) \log_2 \frac{\Pr(x_i,y_j)}{\Pr(x_i)\Pr(y_j)} \geq \left(\sum_{i=1}^{M_x} \Pr(x_i,y_j)\right) \underbrace{\log_2 \frac{\overbrace{\sum_{i=1}^{M_x} \Pr(x_i,y_j)}^{\Pr(y_j)}}{\underbrace{\sum_{i=1}^{M_x} \Pr(x_i)\Pr(y_j)}_{\Pr(y_j)}}}_{\log_2 1 = 0} = 0$$

$$\Rightarrow I(X;y_j) \geq 0$$

## References

1. Lensink MF, Mendez R, Wodak SJ (2007) Docking and scoring protein complexes: CAPRI 3rd edition. Proteins 69:704–718
2. Lensink MF, Wodak SJ (2010) Docking and scoring protein interactions: CAPRI 2009. Proteins 78:3073–3084
3. Janin J (2010) Protein–protein docking tested in blind predictions: the CAPRI experiment. Mol Biosyst 6:2351–2362
4. Janin J (2010) The targets of CAPRI rounds 13–19. Proteins 78:3067–3072
5. Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA (1992) Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci USA 89:2195–2199
6. Walls PH, Sternberg MJ (1992) New algorithm to model protein–protein recognition based on surface complementarity. Applications to antibody–antigen docking. J Mol Biol 228:277–297
7. Jones S, Thornton JM (1996) Principles of protein–protein interactions. Proc Natl Acad Sci USA 93:13–20
8. Meyer M, Wilson P, Schomburg D (1996) Hydrogen bonding and molecular surface shape complementarity as a basis for protein docking. J Mol Biol 264:199–210
9. Ausiello G, Cesareni G, Helmer-Citterich M (1997) Escher: a new docking procedure applied to the reconstruction of protein tertiary structure. Proteins 28:556–567
10. Vakser IA, Aflalo C (1994) Hydrophobic docking: a proposed enhancement to molecular recognition techniques. Proteins 20:320–329
11. Gabb HA, Jackson RM, Sternberg MJ (1997) Modelling protein docking using shape complementarity, electrostatics and biochemical information. J Mol Biol 272:106–120
12. Robert CH, Janin J (1998) A soft, mean-field potential derived from crystal contacts for predicting protein–protein interactions. J Mol Biol 283:1037–1047
13. Moont G, Gabb HA, Sternberg MJ (1999) Use of pair potentials across protein interfaces in screening predicted docked complexes. Proteins 35:364–373
14. Zhang C, Liu S, Zhou H, Zhou Y (2004) An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. Protein Sci 13:400–411
15. Pons C, Talavera D, de la Cruz X, Orozco M, Fernandez-Recio J (2011) Scoring by intermolecular pairwise propensities of exposed residues (sipper): a new efficient potential for protein–protein docking. J Chem Inf Model 51:370–377

16. Cover TM, Thomas JA (2006) Elements of information theory. Wiley-Interscience, Hoboken
17. Douguet D, Chen HC, Tovchigrechko A, Vakser IA (2006) Dockground resource for studying protein–protein interfaces. Bioinformatics 22:2612–2618
18. Gao Y, Douguet D, Tovchigrechko A, Vakser IA (2007) Dockground system of databases for protein recognition studies: unbound structures for docking. Proteins 69:845–851
19. Liu S, Gao Y, Vakser IA (2008) Dockground protein–protein docking decoy set. Bioinformatics 24:2634–2635
20. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637
21. Fiorucci S, Zacharias M (2010) Prediction of protein–protein interaction sites using electrostatic desolvation profiles. Biophys J 98:1921–1930
22. Aloy P, Russell RB (2002) Interrogating protein interaction networks through structural biology. Proc Natl Acad Sci USA 99:5896–5901
23. Ansari S, Helms V (2005) Statistical analysis of predominantly transient protein–protein interfaces. Proteins 61:344–355
24. Melo F, Feytmans E (1997) Novel knowledge-based mean force potential at atomic level. J Mol Biol 267:207–222
25. Melo F, Sanchez R, Sali A (2002) Statistical potentials for fold assessment. Protein Sci 11:430–448
26. Launay G, Mendez R, Wodak S, Simonson T (2007) Recognizing protein–protein interfaces with empirical potentials and reduced amino acid alphabets. BMC Bioinforma 8:270
27. Fiorucci S, Zacharias M (2010) Binding site prediction and improved scoring during flexible protein–protein docking with attract. Proteins 78:3131–3139
28. Ohlson MB, Huang Z, Alto NM, Blanc MP, Dixon JE, Chai J, Miller SI (2008) Structure and function of salmonella sifa indicate that its interactions with skip, ssej, and rhoa family gtpases induce endosomal tubulation. Cell Host Microbe 4:434–446
29. Diacovich L, Dumont A, Lafitte D, Soprano E, Guilhon AA, Bignon C, Gorvel JP, Bourne Y, Meresse S (2009) Interaction between the sifa virulence factor and its host target skip is essential for salmonella pathogenesis. J Biol Chem 284:33151–33160
30. Perkins JR, Diboun I, Dessailly BH, Lees JG, Orengo C (2010) Transient protein–protein interactions: structural, functional, and network properties. Structure 18:1233–1243
31. Dey S, Pal A, Chakrabarti P, Janin J (2010) The subunit interfaces of weakly associated homodimeric proteins. J Mol Biol 398:146–160
32. Gatenby RA, Frieden BR (2007) Information theory in living systems, methods, applications, and challenges. Bull Math Biol 69:635–657
33. Kauffman C, Karypis G (2008) An analysis of information content present in protein–DNA interactions. Pac Symp Biocomput:477–488
34. Sterner B, Singh R, Berger B (2007) Predicting and annotating catalytic residues: an information theoretic approach. J Comput Biol 14:1058–1073
35. Magliery TJ, Regan L (2005) Sequence variation in ligand binding sites in proteins. BMC Bioinforma 6:240
36. Kulharia M, Goody RS, Jackson RM (2008) Information theory-based scoring function for the structure-based prediction of protein–ligand binding affinity. J Chem Inf Model 48:1990–1998
37. Wassermann AM, Nisius B, Vogt M, Bajorath J (2010) Identification of descriptors capturing compound class-specific features by mutual information analysis. J Chem Inf Model 50:1935–1940
38. Cline MS, Karplus K, Lathrop RH, Smith TF, Rogers RG Jr, Haussler D (2002) Information-theoretic dissection of pairwise contact potentials. Proteins 49:7–14
39. Shackelford G, Karplus K (2007) Contact prediction using mutual information and neural nets. Proteins 69(Suppl 8):159–164
40. Miller CS, Eisenberg D (2008) Using inferred residue contacts to distinguish between correct and incorrect protein models. Bioinformatics 24:1575–1582
41. Solis AD, Rackovsky S (2008) Information and discrimination in pairwise contact potentials. Proteins 71:1071–1087

J Mol Model (2012) 18:1299–1311
DOI 10.1007/s00894-011-1153-x

ORIGINAL PAPER

# Insights into $ET_A$ subtype selectivity of benzodiazepine endothelin receptor antagonists by 3D-QSAR approaches

**Jun Xia · Jiabin Li · Hongbin Sun**

Received: 29 March 2011 / Accepted: 10 June 2011 / Published online: 12 July 2011


**Abstract** $ET_A$ subtype selective antagonists constitute a novel and potentially important class of agents for the treatment of pulmonary hypertension, heart failure, and other pathological conditions. In this paper, 60 benzodiazepine derivatives displaying potent activities against $ET_A$ and $ET_B$ subtypes of endothelin receptor were selected to establish the 3D-QSAR models using CoMFA and CoMSIA approaches. These models show excellent internal predictability and consistency, external validation using test-set 19 compounds yields a good predictive power for antagonistic potency. Statistical parameters of models were obtained with CoMFA-$ET_A$ ($q^2$=0.787, $r^2$=0.935, $r^2_{pred}$=0.901), CoMFA-$ET_B$ ($q^2$=0.842, $r^2$=0.984, $r^2_{pred}$=0.941), CoMSIA-$ET_A$ ($q^2$=0.762, $r^2$=0.971, $r^2_{pred}$=0.958) and CoMSIA-$ET_B$ ($q^2$=0.771, $r^2$=0.974, $r^2_{pred}$=0.953) respectively. Field contour maps (CoMFA and CoMSIA) corresponding to the $ET_A$ and $ET_B$ subtypes reflects the characteristic similarities and differences between these types. The results of this paper provide valuable information to facilitate structural modifications of the title compounds to increase the inhibitory potency and subtype selectivity of endothelin receptor.

J. Xia · H. Sun (✉)
Center for Drug Discovery, College of Pharmacy,
China Pharmaceutical University,
24 Tongjia Xiang,
Nanjing 210009, China
e-mail: hbsun2000@yahoo.com

J. Li (✉)
Department of Inorganic Chemistry, College of Science,
China Pharmaceutical University,
24 Tongjia Xiang,
Nanjing 210009, China
e-mail: jbli@cpu.edu.cn

**Keywords** Benzodiazepine derivatives · CoMFA · CoMSIA · Endothelin receptor antagonists · Subtype selectivity · 3D-QSAR

## Introduction

Endothelin-1(ET-1) originally isolated from cultured porcine endothelial cells by Yanagisawa and co-workers in 1988, is the most potent vasoconstrictor peptide and has a long duration of action [1]. ET-1 exerts its physiological effects via two specific G-protein coupled receptors termed endothelin type-A ($ET_A$) and type-B ($ET_B$) [2, 3]. The $ET_A$ receptors are primarily expressed on vascular smooth muscle cells and cardiac myocytes; in contrast, $ET_B$ receptors are predominantly localized on endothelial cells, and to a lesser extent, on smooth muscle cells and on fibroblasts. Activation of the $ET_A$ receptor (and the $ET_B$ receptor on smooth muscle cells) induces vasoconstriction and proliferation of vascular smooth muscle cells; $ET_B$ receptors may induce vasodilatation via release of NO and prostacyclin from the endothelial cells, and are involved in the clearance of ET-1 from plasma additionally [4].

Recently, a number of selective as well as nonselective endothelin receptor antagonists are being actively developed as new therapeutic agents [5–8]. Extensive preclinical and clinical studies, mainly with $ET_A$ receptor antagonists, have shown excellent therapeutic benefits in disease such as pulmonary hypertension [9], heart failure [10], atherosclerosis [11], systemic hypertension [12], chronic renal failure [13] and solid tumors [14]. These results indicate endothelin receptor antagonists, especially those selective for the $ET_A$ receptor subtype, may constitute a novel and potentially important class of agents for the treatment of the aforementioned pathological conditions. Selective inhibition of $ET_A$

receptors may be preferential to non-selective receptor antagonism by permitting maintenance of vasodilator and clearance functions specific to $ET_B$ receptors on the endothelial cells, while preventing the vasoconstriction and cellular proliferation mediated by $ET_A$ [7].

Due to the lack of any exact 3D structure of endothelin recptor, it is very difficult to clarify the action model of drug-receptor, which restrains the application of receptor-guided drug design to find promising lead compounds. In this case, ligand-based 3D-QSAR analysis may be one of the most effective tools for optimizing a lead compound and designing new potential $ET_A$ antagonists. A careful literature survey revealed that there are only a few reports on 3D-QSAR studies of endothelin receptor antagonists. Krystek and colleagues first studied the effects of simple charge calculations on the CoMFA models for 36 aryl sulfonamide antagonists of $ET_A$ receptor [15]; Xu et al. carried out CoMFA models on a series of gamma- hydroxy butenolide $ET_A$ receptor antagonists in the presence of additional H-bond fields [16]; Recently, Zhu and co-workers analyzed 3D-QSAR models of a series of 63 butenolide $ET_A$ receptor antagonists [17]. Up to now, there is a lack of comparative studies between $ET_A$ and $ET_B$ 3D-QSAR models and a lack of 3D-QSAR studies of benzodiazepine derivatives as endothelin receptor antagonists.

In this paper, a profound QSAR analysis toward both $ET_A$ and $ET_B$ antagonistic activities was performed with CoMFA and CoMSIA on title compounds. It is expected that our efforts would afford some useful information for further development of new effective and $ET_A$-selective endothelin antagonists.

## Methods

### Data set

The data used in this QSAR study consist of endothelin receptor affinities ($IC_{50}$ values), determined in radioligand competition binding studies on membranes of Chinese hamster ovary (CHO) cells expressing the human recombinant $ET_A$ or $ET_B$ receptor using $^{125}$I-labeled ET-1. Sixty benzo[1, 4]diazepin- 2-one derivatives reported by Bolli et al. [18] were selected as training set and another 19 molecules were employed as external test set in the current study. The affinity activity data shown in Table 1 were converted into the corresponding $pIC_{50}$ values ($pIC_{50}=-log\ IC_{50}$) and then used for subsequent 3D-QSAR (CoMFA and CoMSIA) analysis. The training and the test sets represented respectively by these 60 and 19 compounds (Table 1) were selected randomly, and the distribution of $ET_A$ and $ET_B$ $pIC_{50}$ values for the training and test sets is shown in Fig. 1. The selected compounds showed good structural variation and range of $pIC_{50}$ affinities from 5.000 to 8.959 and 5.000 to 8.149 (Fig. 1) for the $ET_A$ and $ET_B$ receptors, respectively. This makes them suitable for 3D-QSAR studies.

### Molecular modeling

All the molecular modeling studies, CoMFA and CoMSIA reported herein were performed using SYBYL 6.9 molecular modeling software from Tripos, Inc., St. Louis, MO. All the compounds were built from fragments in the SYBYL database. Each structure was fully geometry-optimized using the standard Tripos force field with a distance-dependent dielectric function until a Root Mean Square (RMS) deviation of 0.001 kcal $mol^{-1}$ $A^{-1}$ was achieved. The least energy conformations were selected for each compound and further energy minimized using Powell (100 iterations) and conjugation gradient (10,000 iterations) methods Gasteiger–Huckel charges were computed for all molecules including the template after energy minimization.

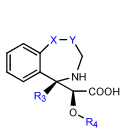
### Alignment procedure

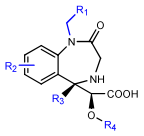
A proper alignment of the structures is critical for obtaining valid 3D-QSAR models. Furthermore, it is vital that all compounds are aligned in a bioactive orientation since the 3D-QSAR model assumes that each structure exhibits activity at the same binding site of the receptor. In this study, the most active molecule (rac-39ao) was used as the template for superimposition, assuming that its bioactive conformation represents the most probable conformation of the antagonists at the putative receptor. The atoms used for alignment were the common substructure of each molecule. The atoms used for alignment were the common substructure of each molecule (atoms 2, 8, 9, 10, 11 and 12, shown in Fig. 2). An automatic alignment method was carried out by using Database alignment and Field-fit alignment. The molecules were aligned and put into a new database.

### CoMFA analysis

Following alignment, the molecules are placed one by one into a 3D cubic lattice with 2Å grid. The steric (van der Waals) and electrostatic (Coulombic with 1/r dielectric) fields are calculated at each grid point using a $sp^3$-hybridized carbon probe with a +1.0 charge. The steric and electrostatic energy values were truncated at 30.0 kcal $mol^{-1}$. CoMFA $q^2$ value were obtained using the partial least-squares (PLS) methodology [19–21], which is an extension of multiple regression analysis. The cross-validation with Leave-One-Out (LOO) [22] option was carried out to obtain the optimum number of components

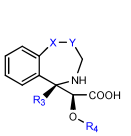
**Table 1** Molecular structures used in the construction of CoMFA and CoMSIA models. The molecules used as test set are highlighted with an asterisk (*)

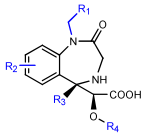



| Compd. No. | -X-Y- | $R_1$ | $R_2$ | $R_3$ | $R_4$ |
|---|---|---|---|---|---|
| 1) rac-30 | $-CH_2-CH_2-$ | - | - | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 2) rac-31 | $-O-CH_2-$ | - | - | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 3) rac-33* |  | - | - | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 4) rac-34 |  | - | - | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 5) rac-39a | - | $-C_5H_{11}$ | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 6) rac-39b* | - |  | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 7) rac-39c | - | cyclopropyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 8) rac-39e | - |  | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 9) rac-39f | - |  | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 10) rac-39h | - | methylol | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 11) rac-39i | - | carboxyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 12) rac-39j | - |  | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 13) rac-39k | - |  | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 14) rac-39n | - |  | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 15) rac-39l | - | phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 16) rac-39o | - | 2-F-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 17) rac-39p | - | 2-Cl-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 18) rac-39q | - | 2-$CH_3$-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 19) rac-39s | - | 2-$OCF_3$-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 20) rac-39t | - | 3-F-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 21) rac-39u | - | 3-Cl-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 22) rac-39v | - | 3-Br-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 23) rac-39w | - | 3-$CH_3$-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 24) rac-39x | - | 3-$CF_3$-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 25) rac-39y | - | 3-$OCF_3$-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 26) rac-39z | - | 4-F-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 27) rac-39aa* | - | 4-Cl-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 28) rac-39ab | - | 4-Br-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 29) rac-39ac | - | 4-$CH_3$-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 30) rac-39ad | - | 4-$CF_3$-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 31) rac-39af | - | 3-$OCH_3$-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 32) rac-39ag | - | 4-$COOCH_3$-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 33) rac-39ai | - | 4-*n*-Bu-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 34) rac-39aj* | - | 4-*tert*-Bu-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 35) rac-39al* | - | 2,3-$F_2$-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 36) rac-39am | - | 2,4-$F_2$-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 37) rac-39an* | - | 2,5-$F_2$-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 38) rac-39ao | - | 2,6-$F_2$-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 39) rac-39ap* | - | 3,5-$F_2$-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 40) rac-39aq* | - | 3,4-$F_2$-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |

**Table 1** (continued)




| Compd. No. | -X-Y- | $R_1$ | $R_2$ | $R_3$ | $R_4$ |
|---|---|---|---|---|---|
| 41) rac-39ar* | - | 2,3,4-$F_3$-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 42) rac-39as* | - | 2,3,6-$F_3$-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 43) rac-39at* | - | 2,4,5-$F_3$-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 44) rac-39au* | - | 2,4,6-$F_3$-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 45) rac-39av | - | 3,4,5-$F_3$-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 46) rac-39ay* | - | 2,3,4,5,6-$F_5$-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 47) rac-39az | - | 2,4,6-$(CH_3)_3$-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 48) rac-5a* | - | H | H | phenyl | 3,5-$(H_3CO)_2$-Phenyl |
| 49) rac-5b | - | H | 7-Cl | phenyl | 3,5-$(H_3CO)_2$-Phenyl |
| 50) rac-5c | - | H | 8-Cl | phenyl | 3,5-$(H_3CO)_2$-Phenyl |
| 51) rac-5d | - | H | 8-$OCH_3$ | phenyl | 3,5-$(H_3CO)_2$-Phenyl |
| 52) rac-5e | - | H | 7,8-$(OCH_3)_2$ | phenyl | 3,5-$(H_3CO)_2$-Phenyl |
| 53) rac-68 | - | 4-$H_3CO$-phenyl | H | *n*-pentyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 54) rac-69 | - | 4-$H_3CO$-phenyl | H | cyclohexyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 55) rac-70* | - | 4-$H_3CO$-phenyl | H | 2-F-phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 56) rac-71 | - | 4-$H_3CO$-phenyl | H | 3-F-phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 57) rac-72 | - | 4-$H_3CO$-phenyl | H | 3-Cl-phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 58) rac-73* | - | 4-$H_3CO$-phenyl | H | 3-Me-phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 59) rac-75 | - | 4-$H_3CO$-phenyl | H | 3-Me-4-F-phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 60) rac-76 | - | 4-$H_3CO$-phenyl | H | 3-$F_3C$-Phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 61) rac-77 | - | 4-$H_3CO$-phenyl | H | 3-$H_3CO$-phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 62) rac-78* | - | 4-$H_3CO$-phenyl | H | 3-Et-phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 63) rac-79 | - | 4-$H_3CO$-phenyl | H | 3-*n*-Bu-phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 64) rac-80 | - | 4-$H_3CO$-phenyl | H | 3-biphenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 65) rac-84b | - | 2,6-$Cl_2$-phenyl | H | phenyl | 2-Cl-phenyl |
| 66) rac-84c* | - | 2,6-$Cl_2$-phenyl | H | phenyl | 2-$CH_3$-phenyl |
| 67) rac-84d | - | 2,6-$Cl_2$-phenyl | H | phenyl | 2-$OCH_3$-phenyl |
| 68) rac-84i | - | 2,6-$Cl_2$-phenyl | H | phenyl | 3,5-$(OCH_3)_2$-phenyl |
| 69) rac-84k | - | 2,6-$Cl_2$-phenyl | H | phenyl | 4-*i*-pr-phenyl |
| 70) rac-84l | - | 2,6-$Cl_2$-phenyl | H | phenyl | 2,3-( $CH_3)_2$-phenyl |
| 71) rac-85 | - | 2,4,6-$(CH_3)_3$-phenyl | H | 3-*n*-Bu-phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 72) rac-87 | - | 2-Cl-6-F-phenyl | H | 3-*n*-Bu-phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 73) rac-88 | - | 2,6-$Cl_2$-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 74) rac-89 | - | 2,4,6-$F_3$-phenyl | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 75) rac-9a | - | H | H | phenyl | -2-pyrimidine |
| 76) rac-9b | - | H | H | phenyl | 4,6-$(H_3C)_2$-2-pyrimidine |
| 77) rac-9c | - | H | H | phenyl | 4,6-$(H_3CO)_2$-2-pyrimidine |
| 78) rac-9f | - | H | H | phenyl | 5-$H_3CS$-2-pyrimidine |
| 79) rac-9g* | - | H | H | phenyl | 5-Br-2-pyrimidine |

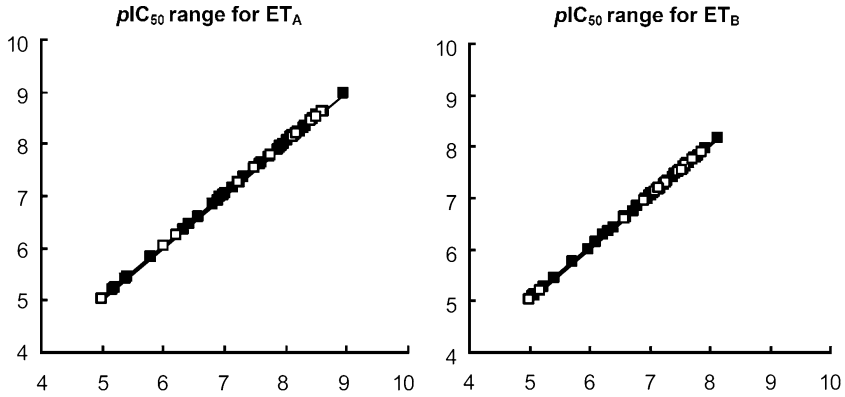



**Fig. 1** Distribution of binding affinity $pIC_{50}$ values of the molecules used to compose the training and the test sets for $ET_A$ and $ET_B$ receptors. The 60 compounds selected as training set are highlighted in black and the 19 compounds used as test set are highlighted in white

(ONC), which was determined on the basis of the first minimum cross-validated standard error of prediction. Using the optimum number of components, the final non-cross-validated model was developed. The cross-validated coefficient $q^2$ was evaluated as follows:

$$q^2 = 1 - \frac{\sum (Y - Y_{pred})^2}{\sum (Y - Y_{mean})^2} \quad (1)$$

where $Y$, $Y_{pred}$, $Y_{mean}$ are the observed, predicted and mean values of the $pIC_{50}$, respectively.

To speed up the analysis and reduce noise, column filtering was set at 2.0 kcal $mol^{-1}$ so that only those steric and electrostatic energies with values greater than 2.0 kcal $mol^{-1}$ are considered in the PLS analysis. In the above stage, the predictive quality of the "best" correlation model is determined. Then, the optimal number of components (*ONC*) is employed to do nonvalidation PLS analysis to get the final model parameters such as correlation coefficient ($r^2$), standard error of estimation (*SEE*) and $F$ value. At the same time, the CoMFA color contour maps are derived for the steric and electrostatic fields. The quality of the final CoMFA model is measured by two statistical parameters: $r^2$ and $q^2$. The value of $q^2$, which indicates the predictive capacity of the model, should be greater than 0.40; and the value of $r^2$, which shows the self-consistency of the model, should be greater than 0.90.

## CoMSIA analysis

CoMSIA descriptors were derived according to Klebe et al. [23]. Being anextension to the CoMFA approach, CoMSIA not only compute steric and electrostatic fields, but also calculates additional hydrophobic, hydrogen-bond donor, and hydrogen-bond acceptor fields. In the present study, CoMSIA descriptors were calculated with the same lattice box and probe atom as that used for the CoMFA calculations. The similarity indices between the compounds and the probe atom were calculated according to:

$$A^q_{F,k}(j) = \sum_i w_{probe,k} w_{ik} e^{-\alpha r^2_{iq}}, \quad (2)$$

where $A$ is the similarity index at grid point $q$, summed over all atoms $i$ of the molecule $j$ under investigation; $w_{probe,k}$ is the probe atom with radius 1 Å, charge +1, hydrophobicity +1, hydrogen bond donating +1, hydrogen bond accepting +1; $w_{ik}$ is the actual value of the physicochemical property $k$ of atom $i$; $r_{iq}$ is the mutual distance between the probe atom at grid point $q$ and atom $i$ of the test molecule; $\alpha$ is the attenuation factor, and the default value of $\alpha$ is 0.3 [24].

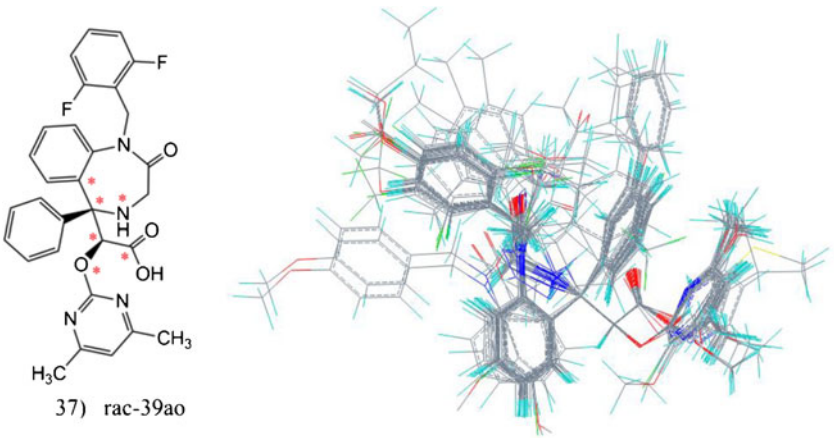



**Fig. 2** All molecules of dataset were aligned on the compound rac-39ao

Larger values of α will result in a steeper Gaussian function and an increasing attenuation of the distance-dependent effects of molecular similarity. On the other hand, reducing α to smaller values means that a probe placed at a particular lattice point detects molecular similarity in its neighborhood more globally.

In this study, five physicochemical properties were evaluated: steric, electrostatic, hydrophobic, and hydrogen-bond donor or acceptor properties. These fields were selected to cover the major contributions to ligand binding. Using all the five CoMSIA descriptors for the explanatory variables, a LOO run and a no validation PLS analysis were performed.

## Results of CoMFA and CoMSIA models

### $ET_A$ receptor antagonists

The statistical parameters of CoMFA model of $ET_A$ receptor antagonists are given in Table 2. LOO analysis gives a result that the cross-validated $q^2$ is 0.787 with six components. The non-cross-validated PLS analysis results in a conventional $r^2$ of 0.935, $F$=126.436, and a standard error of estimation (*SEE*) of 0.275. The steric field descriptors explain 56.3% of the variance, while the electrostatic descriptors explain 43.7%. The predicted activities, experimental activities and their residues of the 79 antagonists are listed in Table 3, and the correlation between the predicted activities and the experimental activities is depicted in Fig. 3.

**Table 2** Statistical results for the CoMFA and CoMSIA models of affinity toward $ET_A$ and $ET_B$ receptor obtained for the training set and test set

| | $ET_A$ | | $ET_B$ | |
|---|---|---|---|---|
| | CoMFA | CoMSIA | CoMFA | CoMSIA |
| Parameters | | | | |
| $q^2$ | 0.787 | 0.762 | 0.842 | 0.771 |
| *ONC* | 6 | 10 | 10 | 10 |
| $r^2$ | 0.935 | 0.971 | 0.984 | 0.974 |
| $F$ | 126.44 | 164.08 | 307.51 | 185.77 |
| *SEE* | 0.275 | 0.190 | 0.126 | 0.161 |
| $r^2_{pred}$ | 0.901 | 0.958 | 0.941 | 0.953 |
| Contribution | | | | |
| Steric | 0.563 | 0.128 | 0.544 | 0.136 |
| Electrostatic | 0.437 | 0.302 | 0.456 | 0.317 |
| Hydrophobic | | 0.319 | | 0.328 |
| H-bond donor | | 0.102 | | 0.096 |
| H-bond acceptor | | 0.149 | | 0.123 |

For the CoMSIA model of $ET_A$ receptor antagonists, some possible combinations of different fields were performed to determine the CoMSIA model. The highest cross-validated $q^2$ was obtained by using the combination of steric, electrostatic, hydrophobic, hydrogen-bond donor and hydrogen-bond acceptor fields ($q^2$=0.762, $r^2$=0.971, $F$=164.078, *SEE*=0.190) with ten components. The corresponding field contributions are 12.8%, 30.2%, 31.9%, 10.2% and 14.9%, respectively. The result indicates that electrostatic and hydrophobic fields are very important for the present series of molecules. CoMSIA analysis results are also summarized in Table 2. Figure 3 shows the relationship between the predicted and the experimental $pIC_{50}$ values for the CoMSIA model. From the cross-validation results, it can be seen that the CoMSIA model has a better predictive ability than CoMFA model, suggesting that are liable CoMSIA models successfully constructed.

### $ET_B$ receptor antagonists

CoMFA analysis of $ET_B$ receptor antagonists was carried out for the training set, and the results are listed in Table 2. The LOO analysis gives a result that the cross-validated $q^2$ is 0.842 with ten components. The noncross-validated PLS analysis results in a conventional $r^2$ of 0.984, $F$=307.508, and a standard error of estimation (*SEE*) of 0.126. The steric field descriptors explain 54.4% of the variance, while the electrostatic descriptors explain 45.6%. The predicted activities, experimental activities and their residues of the 79 inhibitors are listed in Table 3, and the correlation between the predicted activities and the experimental activities are depicted in Fig. 4.

CoMSIA analysis results of $ET_B$ receptor antagonists were also summarized in Table 2. A CoMSIA model with a cross-validated $q^2$ of 0.771 for ten components and non-cross-validated $r^2$ of 0.974 were obtained, $F$=185.771 and *SEE*=0.161. The corresponding field contributions of steric, electrostatic, hydrophobic, hydrogen-bond donor and hydrogen-bond acceptor are 13.6%, 31.7%, 32.8%, 9.6%, and 12.3%, respectively. The result indicates that electrostatic and hydrophobic fields are very important for the present series of molecules. Figure 4 shows the relationship between the predicted and the experimental $pIC_{50}$ values for the CoMSIA model. From Table 3, it can be concluded that the activities predicted by the constructed CoMSIA model are in good agreement with the experimental data, suggesting that a reliable CoMSIA model is successfully constructed.

## Discussion

The steric contour maps indicate green and yellow contours as steric bulk favored and disfavored areas. Blue and red

**Table 3** Predicted activities (PA) versus experimental activities ($pIC_{50}$) with residues (⊿) by CoMFA and CoMSIA for $ET_A$ and $ET_B$ receptor antagonists of the dataset

| Compd. No. | $ET_A$ | | | | | $ET_B$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $pIC_{50}$ | CoMFA | | CoMSIA | | $pIC_{50}$ | CoMFA | | CoMSIA | |
| | | PA | ⊿ | PA | ⊿ | | PA | ⊿ | PA | ⊿ |
| 1) rac-30 | 7.038 | 6.982 | 0.055 | 6.930 | 0.108 | 5.106 | 4.988 | 0.118 | 5.147 | -0.041 |
| 2) rac-31 | 7.156 | 7.456 | -0.300 | 7.156 | 0.001 | 5.439 | 5.443 | -0.004 | 5.420 | 0.019 |
| 3) rac-33* | 7.231 | 7.687 | -0.456 | 7.192 | 0.038 | 6.583 | 6.956 | -0.373 | 6.816 | -0.232 |
| 4) rac-34 | 7.253 | 7.618 | -0.366 | 7.322 | -0.069 | 6.602 | 6.668 | -0.066 | 6.593 | 0.010 |
| 5) rac-39a | 7.625 | 7.606 | 0.019 | 7.568 | 0.057 | 7.399 | 7.525 | -0.126 | 7.268 | 0.131 |
| 6) rac-39b* | 8.125 | 7.820 | 0.305 | 7.736 | 0.389 | 7.321 | 7.443 | -0.123 | 7.106 | 0.215 |
| 7) rac-39c | 8.137 | 7.842 | 0.295 | 8.257 | -0.120 | 7.057 | 7.065 | -0.008 | 7.182 | -0.126 |
| 8) rac-39e | 6.836 | 6.765 | 0.071 | 6.944 | -0.108 | 6.815 | 6.711 | 0.104 | 6.808 | 0.007 |
| 9) rac-39f | 6.900 | 6.800 | 0.099 | 7.161 | -0.262 | 6.129 | 6.358 | -0.229 | 6.279 | -0.150 |
| 10) rac-39 h | 8.161 | 7.920 | 0.241 | 7.993 | 0.168 | 6.967 | 6.976 | -0.010 | 6.953 | 0.013 |
| 11) rac-39i | 7.886 | 7.836 | 0.050 | 7.800 | 0.086 | 6.253 | 6.302 | -0.048 | 6.312 | -0.059 |
| 12) rac-39j | 8.181 | 7.921 | 0.260 | 8.331 | -0.151 | 7.183 | 7.164 | 0.019 | 7.333 | -0.150 |
| 13) rac-39 k | 7.338 | 7.553 | -0.215 | 7.222 | 0.116 | 6.597 | 6.586 | 0.011 | 6.391 | 0.205 |
| 14) rac-39n | 7.767 | 7.522 | 0.245 | 7.687 | 0.080 | 7.738 | 7.903 | -0.166 | 7.600 | 0.137 |
| 15) rac-39 l | 8.194 | 8.236 | -0.042 | 8.216 | -0.022 | 7.324 | 7.399 | -0.074 | 7.305 | 0.019 |
| 16) rac-39o | 8.456 | 8.249 | 0.207 | 8.275 | 0.181 | 7.480 | 7.420 | 0.060 | 7.388 | 0.092 |
| 17) rac-39p | 7.991 | 8.071 | -0.079 | 8.033 | -0.042 | 7.318 | 7.379 | -0.061 | 7.458 | -0.140 |
| 18) rac-39q | 8.051 | 7.825 | 0.226 | 7.811 | 0.240 | 7.260 | 7.277 | -0.017 | 7.281 | -0.021 |
| 19) rac-39 s | 7.006 | 7.731 | -0.724 | 7.511 | -0.505 | 7.126 | 7.119 | 0.006 | 7.328 | -0.202 |
| 20) rac-39 t | 8.174 | 8.302 | -0.128 | 8.241 | -0.067 | 7.572 | 7.578 | -0.006 | 7.492 | 0.080 |
| 21) rac-39u | 8.237 | 8.307 | -0.071 | 8.202 | 0.035 | 7.670 | 7.589 | 0.081 | 7.654 | 0.015 |
| 22) rac-39v | 8.194 | 8.316 | -0.122 | 8.180 | 0.014 | 7.594 | 7.635 | -0.042 | 7.781 | -0.187 |
| 23) rac-39w | 8.301 | 8.306 | -0.005 | 8.192 | 0.109 | 7.777 | 7.702 | 0.075 | 7.505 | 0.272 |
| 24) rac-39x | 8.041 | 8.329 | -0.288 | 8.146 | -0.105 | 7.807 | 7.844 | -0.037 | 7.862 | -0.055 |
| 25) rac-39y | 7.959 | 8.295 | -0.336 | 8.155 | -0.196 | 7.511 | 7.537 | -0.026 | 7.657 | -0.145 |
| 26) rac-39z | 8.229 | 8.237 | -0.008 | 8.102 | 0.128 | 7.304 | 7.266 | 0.037 | 7.191 | 0.113 |
| 27) rac-39aa* | 7.762 | 8.103 | -0.341 | 7.824 | -0.062 | 6.914 | 7.135 | -0.221 | 7.070 | -0.156 |
| 28) rac-39ab | 7.588 | 8.053 | -0.465 | 7.646 | -0.057 | 6.925 | 7.091 | -0.167 | 6.990 | -0.065 |
| 29) rac-39 ac | 7.516 | 8.000 | -0.484 | 7.979 | -0.463 | 7.162 | 7.192 | -0.030 | 7.283 | -0.121 |
| 30) rac-39ad | 7.770 | 8.015 | -0.246 | 7.739 | 0.031 | 7.327 | 7.073 | 0.254 | 7.194 | 0.133 |
| 31) rac-39af | 8.337 | 8.183 | 0.154 | 8.432 | -0.094 | 7.650 | 7.576 | 0.073 | 7.514 | 0.136 |
| 32) rac-39ag | 8.131 | 7.994 | 0.137 | 8.121 | 0.010 | 7.153 | 7.174 | -0.021 | 7.233 | -0.080 |
| 33) rac-39ai | 8.456 | 8.284 | 0.172 | 8.292 | 0.164 | 7.955 | 7.961 | -0.006 | 7.813 | 0.142 |
| 34) rac-39aj* | 7.767 | 8.227 | -0.460 | 7.894 | -0.127 | 7.618 | 7.798 | -0.180 | 7.580 | 0.038 |
| 35) rac-39al* | 8.432 | 8.272 | 0.160 | 8.319 | 0.113 | 7.733 | 7.627 | 0.105 | 7.488 | 0.245 |
| 36) rac-39 am | 8.222 | 8.210 | 0.012 | 8.175 | 0.047 | 6.963 | 7.321 | -0.359 | 7.177 | -0.215 |
| 37) rac-39an* | 8.174 | 8.208 | -0.034 | 8.260 | -0.087 | 7.483 | 7.449 | 0.034 | 7.375 | 0.108 |
| 38) rac-39ao | 8.959 | 8.209 | 0.749 | 8.348 | 0.610 | 7.752 | 7.474 | 0.278 | 7.383 | 0.370 |
| 39) rac-39ap* | 8.102 | 8.298 | -0.196 | 8.198 | -0.096 | 7.883 | 7.566 | 0.317 | 7.559 | 0.324 |
| 40) rac-39aq* | 8.187 | 8.241 | -0.054 | 8.054 | 0.133 | 7.484 | 7.264 | 0.220 | 7.250 | 0.234 |
| 41) rac-39ar* | 8.187 | 8.250 | -0.063 | 8.108 | 0.079 | 7.304 | 7.291 | 0.013 | 7.331 | -0.027 |
| 42) rac-39as* | 8.620 | 8.216 | 0.404 | 8.308 | 0.312 | 7.879 | 7.471 | 0.408 | 7.449 | 0.431 |
| 43) rac-39at* | 8.143 | 8.217 | -0.074 | 8.134 | 0.009 | 7.139 | 7.320 | -0.182 | 7.244 | -0.105 |
| 44) rac-39au* | 8.602 | 8.218 | 0.384 | 8.222 | 0.380 | 7.078 | 7.345 | -0.267 | 7.251 | -0.173 |
| 45) rac-39av | 7.928 | 8.307 | -0.379 | 8.072 | -0.144 | 7.597 | 7.438 | 0.159 | 7.432 | 0.165 |

**Table 3** (continued)

| Compd. No. | $ET_A$ | | | | | $ET_B$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $pIC_{50}$ | CoMFA | | CoMSIA | | $pIC_{50}$ | CoMFA | | CoMSIA | |
| | | PA | ⊿ | PA | ⊿ | | PA | ⊿ | PA | ⊿ |
| 46) rac-39ay* | 8.168 | 8.285 | -0.117 | 8.195 | -0.028 | 7.541 | 7.519 | 0.022 | 7.493 | 0.048 |
| 47) rac-39az | 8.174 | 8.036 | 0.138 | 8.123 | 0.051 | 7.810 | 7.754 | 0.056 | 7.694 | 0.115 |
| 48) rac-5a* | 6.218 | 6.382 | -0.164 | 5.881 | 0.338 | 5.202 | 5.674 | -0.472 | 5.364 | -0.162 |
| 49) rac-5b | 6.582 | 6.225 | 0.357 | 6.434 | 0.147 | 5.742 | 5.607 | 0.135 | 5.743 | -0.001 |
| 50) rac-5c | 5.416 | 6.012 | -0.596 | 5.679 | -0.264 | 5.252 | 5.284 | -0.033 | 5.434 | -0.182 |
| 51) rac-5 d | 5.403 | 5.329 | 0.075 | 5.485 | -0.082 | 5.109 | 5.074 | 0.035 | 5.101 | 0.008 |
| 52) rac-5e | 6.354 | 6.453 | -0.099 | 6.305 | 0.049 | 5.095 | 5.117 | -0.022 | 5.097 | -0.001 |
| 53) rac-68 | 6.450 | 6.538 | -0.088 | 6.473 | -0.023 | 7.133 | 7.071 | 0.062 | 7.191 | -0.058 |
| 54) rac-69 | 6.951 | 6.994 | -0.043 | 6.992 | -0.041 | 6.979 | 6.938 | 0.040 | 7.117 | -0.139 |
| 55) rac-70* | 7.519 | 7.931 | -0.412 | 7.930 | -0.411 | 7.096 | 7.294 | -0.198 | 7.273 | -0.177 |
| 56) rac-71 | 8.041 | 7.873 | 0.168 | 7.899 | 0.142 | 7.307 | 7.239 | 0.069 | 7.230 | 0.078 |
| 57) rac-72 | 7.602 | 7.787 | -0.185 | 7.740 | -0.138 | 7.107 | 7.170 | -0.063 | 7.125 | -0.018 |
| 58) rac-73* | 8.201 | 7.919 | 0.282 | 8.231 | -0.030 | 7.295 | 7.220 | 0.075 | 7.292 | 0.003 |
| 59) rac-75 | 8.137 | 7.880 | 0.257 | 8.166 | -0.029 | 7.479 | 7.193 | 0.286 | 7.254 | 0.224 |
| 60) rac-76 | 7.604 | 7.665 | -0.061 | 7.862 | -0.258 | 6.618 | 6.711 | -0.093 | 6.856 | -0.238 |
| 61) rac-77 | 7.991 | 7.972 | 0.019 | 8.026 | -0.035 | 7.026 | 6.986 | 0.040 | 7.066 | -0.041 |
| 62) rac-78* | 8.509 | 8.022 | 0.487 | 8.340 | 0.169 | 7.168 | 7.185 | -0.018 | 7.524 | -0.356 |
| 63) rac-79 | 8.102 | 7.838 | 0.265 | 7.886 | 0.216 | 7.438 | 7.411 | 0.027 | 7.505 | -0.068 |
| 64) rac-80 | 8.523 | 8.180 | 0.343 | 8.317 | 0.206 | 7.062 | 7.049 | 0.013 | 7.263 | -0.202 |
| 65) rac-84b | 5.000 | 4.911 | 0.089 | 4.821 | 0.179 | 5.000 | 5.159 | -0.159 | 5.150 | -0.150 |
| 66) rac-84c* | 5.000 | 4.943 | 0.057 | 4.967 | 0.033 | 5.000 | 5.339 | -0.339 | 5.316 | -0.316 |
| 67) rac-84 d | 5.228 | 4.799 | 0.429 | 5.103 | 0.125 | 5.000 | 4.951 | 0.049 | 5.060 | -0.060 |
| 68) rac-84i | 7.335 | 7.337 | -0.001 | 7.330 | 0.006 | 6.407 | 6.399 | 0.008 | 6.400 | 0.007 |
| 69) rac-84 k | 5.000 | 5.028 | -0.028 | 4.926 | 0.074 | 5.065 | 5.156 | -0.091 | 5.057 | 0.007 |
| 70) rac-84 l | 5.000 | 5.220 | -0.220 | 5.259 | -0.260 | 6.000 | 5.833 | 0.167 | 5.830 | 0.170 |
| 71) rac-85 | 7.738 | 7.693 | 0.045 | 7.764 | -0.027 | 7.867 | 7.756 | 0.111 | 8.022 | -0.156 |
| 72) rac-87 | 8.119 | 7.990 | 0.129 | 8.059 | 0.061 | 8.149 | 8.255 | -0.106 | 8.034 | 0.115 |
| 73) rac-88 | 7.772 | 7.781 | -0.008 | 7.824 | -0.052 | 7.602 | 7.641 | -0.039 | 7.380 | 0.222 |
| 74) rac-89 | 8.102 | 8.092 | 0.010 | 8.178 | -0.076 | 7.047 | 7.274 | -0.227 | 7.349 | -0.303 |
| 75) rac-9a | 5.804 | 5.943 | -0.138 | 5.841 | -0.037 | 5.000 | 5.125 | -0.125 | 5.111 | -0.111 |
| 76) rac-9b | 7.910 | 7.676 | 0.234 | 7.903 | 0.007 | 6.735 | 6.701 | 0.035 | 6.740 | -0.005 |
| 77) rac-9c | 7.631 | 7.324 | 0.307 | 7.607 | 0.024 | 6.325 | 6.304 | 0.021 | 6.195 | 0.130 |
| 78) rac-9f | 5.181 | 5.311 | -0.131 | 4.921 | 0.259 | 5.000 | 4.969 | 0.031 | 4.647 | 0.353 |
| 79) rac-9 g* | 6.033 | 5.598 | 0.435 | 6.066 | -0.033 | 5.023 | 4.916 | 0.108 | 5.188 | -0.164 |

contours in the electrostatic maps indicate areas where positive and negative charge substituents favor activity. Hydrophobic, hydrogen-bond donor and acceptor contour maps indicate favorable by yellow, cyan, magenta contours and unfavorable by white, violet/purple and red contours, respectively. Although the contour maps cannot be used as receptor maps, useful interpretations can be derived. To aid in visualization, several compounds are displayed in the respective figures and the contour maps are discussed with the reference compounds.

## Contour maps of CoMFA and CoMSIA

### *Steric fields*

1) $ET_A$ receptor antagonists

CoMFA steric fields for $ET_A$ affinity based on PLS analysis are represented as 3D contour plots in Fig. 5a. Two bulky favored regions (marked in green) are located around the pyrimidine ring on the $R_4$ position. This may be the

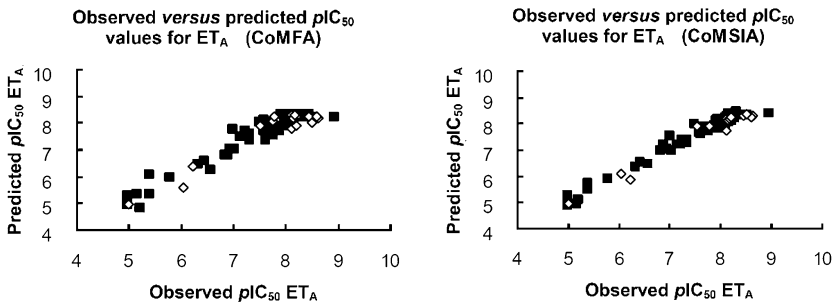



**Fig. 3** Plot of predicted versus observed $pIC_{50}$ values derived from the CoMFA and CoMSIA affinity model of the training set (in black) and test set (in white) for $ET_A$ antagonists

reason that compounds with 4,6-substituents on the pyrimidine ring, e.g., compounds rac-9b ($pIC_{50}$=7.910), rac-9c ($pIC_{50}$=7.631), are more potent than molecules without any substituent, such as rac-9a ($pIC_{50}$=5.804). However, when the substituent was at 3,5-position, where there are two bulky disfavored regions (marked in yellow), such as compounds rac-9f ($pIC_{50}$=5.181), rac-84c ($pIC_{50}$= 5.000), there is an unfavorable contribution to the potency. One bulky favored region (marked in green) located at the $R_1$ position of phenyl ring (e.g., compound rac-39 l, $pIC_{50}$= 8.194) suggests favorable steric interactions of this structural moiety on the endothelin receptor active site. Compounds with 3-substituent on this phenyl ring give better activities (e.g., compound rac-39w, $pIC_{50}$=8.301) than others with 4-substituent (e.g., compound rac-39aj, $pIC_{50}$=7.767). Another bulky disfavored region (marked in yellow) is observed in the cyclohexane position such as compound rac-39e ($pIC_{50}$= 6.836). This indicates that bulky groups herein will reduce the binding affinity for $ET_A$ receptor.

Figure 5b shows the steric field distributions of CoMSIA for $ET_A$ affinity. Compared with Fig. 5a, it can be seen that the CoMSIA steric contour map is very similar to the corresponding CoMFA contour map. There are two green polyhedrons around the pyrimidine ring on the $R_4$ position, a large green polyhedron on the $R_1$ position, a large yellow polyhedron corresponding to cyclohexane position of compound rac-39e, a small yellow polyhedron at the position of 4-substituent of the phenyl ring on the $R_1$ position of compound rac-39aj. However, there is just one yellow polyhedron around the pyrimidine ring on the $R_4$ position, rather than two yellow polyhedrons in the CoMFA steric fields. It can be seen that the CoMFA model gives more detailed steric field information than CoMSIA model in steric field.

2) $ET_B$ receptor antagonists

The steric contour maps for $ET_B$ affinity by the CoMFA and CoMSIA models (Fig. 5c, d) are similar. Four bulky favored regions (marked in green) are observed, of which two are located around the pyrimidine ring on the $R_4$ position, one is located at the phenyl ring on the $R_1$ position, and another one is located at the long chain of compound rac-87 ($ET_A$: $pIC_{50}$=8.119, $ET_B$: $pIC_{50}$=8.149). The difference between CoMFA and CoMSIA models is the distributions of yellow polyhedrons. It should be noted that the trifluoromethoxy substituent at the phenyl ring on the $R_1$ position of compound rac-39 s ($ET_A$: $pIC_{50}$=7.006, $ET_B$: $pIC_{50}$=7.126), appears in the green polyhedron of the CoMFA model (Fig. 5c), but is included in the yellow polyhedron of the CoMSIA model (Fig. 5d). In fact, compound rac-39 s has a higher affinity with $ET_B$ than $ET_A$. It gives further evidence that the CoMFA model may have a better predictive ability than the CoMSIA model in steric field.

*Electrostatic fields*

1) $ET_A$ receptor antagonists

CoMFA electrostatic fields for $ET_A$ affinity are shown in Fig. 6a. The electropositive groups favored region (marked in blue) adjacent to pyrimidine ring, indicates that the

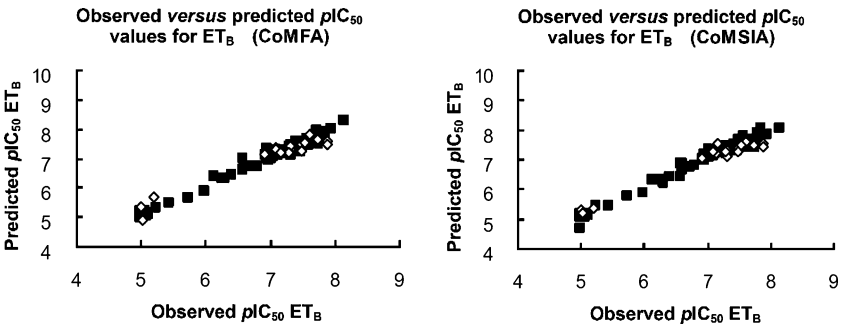



**Fig. 4** Plot of predicted versus observed $pIC_{50}$ values derived from the CoMFA and CoMSIA affinity model of the training set (in black) and test set (in white) for $ET_B$ antagonists

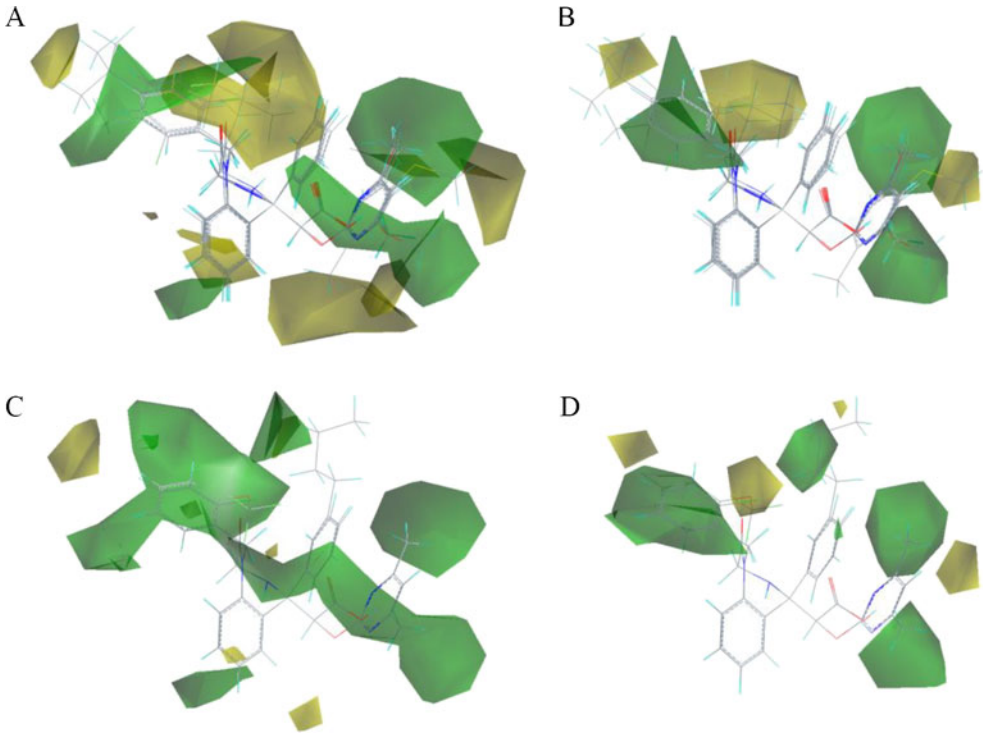



**Fig. 5** Steric Contour maps of the CoMFA and CoMSIA regions are shown. (**a**) Steric contour maps of the CoMFA affinity model for $ET_A$ receptor. (**b**) Steric contour map of the CoMSIA affinity model for $ET_A$ receptor. (**c**) Steric contour map of the CoMFA affinity model for $ET_B$ receptor. (**d**) Steric contour map of the CoMSIA affinity model for $ET_B$ receptor. (Green indicates steric bulk favored, yellow indicates bulk disfavored. Compounds rac-9a, rac-9b, rac-9c, rac-9f, rac-39e, rac-39 l, rac-39w, rac-39aj and rac-84c are shown in **a** and **b**. Compound rac-87 and rac-39 s are shown in **c** and **d**)

substituent should be electron deficient for high binding affinity with protein. The electronegative groups favored region (marked in red) distributed at the 2-substituent of the phenyl ring on the $R_1$ position, e.g., compound rac-39ao ($pIC_{50}$=8.959) shows that compounds with electron-rich substituent at this position are highly active. In addition, two electronegative groups favored regions (marked in red) located on the two sides of the pyrimidine ring without any corresponding compound, indicates that it may be an undiscovered motif which can be modified.

The CoMSIA contour map of electrostatic field for $ET_A$ affinity is shown in Fig. 6b. Compared with Fig. 6a (CoMFA), three major red regions are absent, and one blue region is added, which is at the 4-substituent of the phenyl

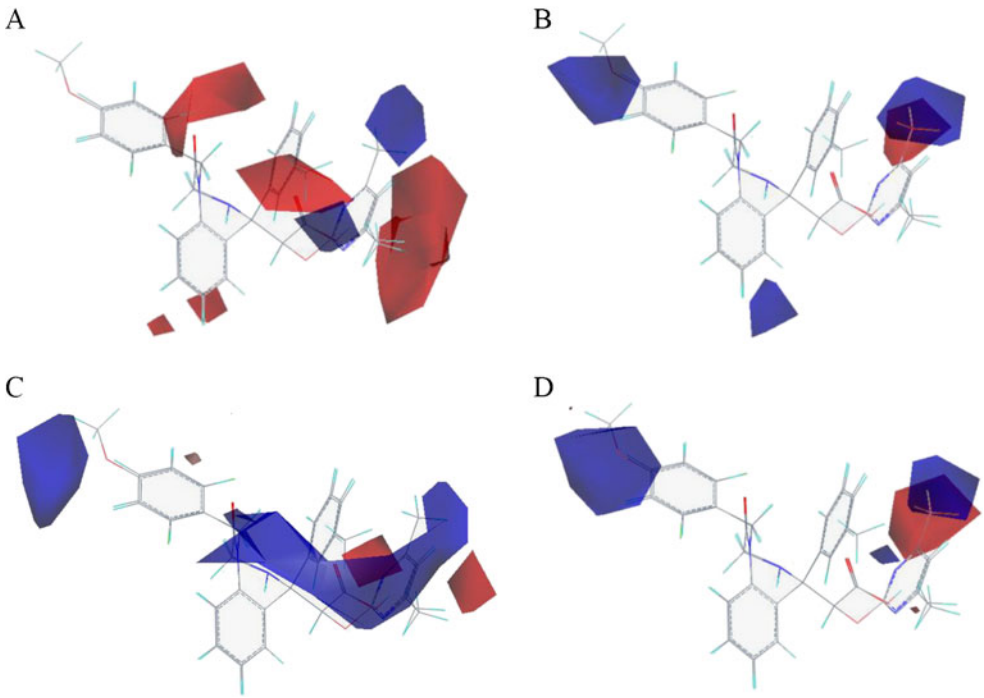



**Fig. 6** Electrostatic contour maps of the CoMFA and CoMSIA regions are shown. (**a**) Electrostatic contour map of the CoMFA affinity model for $ET_A$ receptor. (**b**) Electrostatic contour map of the CoMSIA affinity model for $ET_A$ receptor. (**c**) Electrostatic contour map of the CoMFA affinity model for $ET_B$ receptor. (**d**) Electrostatic contour map of the CoMSIA affinity model for $ET_B$ receptor. (Blue indicates electropositive groups favored, red indicates electronegative groups favored. Compounds rac-73 and rac-39ao are shown inside the fields)

ring on the $R_1$ position of compound rac-73 ($pIC_{50}$=8.201). In this case, we can see that CoMFA and CoMSIA models may be complementary in electrostatic fields.

2) $ET_B$ receptor antagonists

Figure 6c and d show the electrostatic fields for $ET_B$ affinity by the CoMFA and CoMSIA models, respectively. These two contour maps are similar, with two major electropositive groups favored regions (marked in blue) on the corresponding position. One is at 4-substituent of phenyl ring on the $R_1$ position of compound rac-73 ($pIC_{50}$=8.201), and the other is beside 4-substituent of the pyrimidine ring on the $R_4$ position.

*Hydrophobic, hydrogen-bond donor and acceptor fields*

1) $ET_A$ receptor antagonists

CoMSIA calculates both steric and electrostatic fields, as in CoMFA, but additionally uses hydrophobic, hydrogen-bond donor and acceptor fields. The contour map of hydrophobic fields is displayed in Fig. 7a. One hydrophobic favorable region (marked in yellow) located at the phenyl ring on the $R_1$ position of compounds rac-39 l ($pIC_{50}$=8.194) suggests that hydrophobic groups in this region are beneficial to enhance the activity. Three hydrophobic unfavorable regions (marked in white) observed around the phenyl ring on the $R_1$ position indicate the affinity will decrease when the linker between the phenyl ring and the benzodiazepine scaffold is too long, such as compound rac-39n ($pIC_{50}$=7.767). Besides, there are also two hydrophobic favorable regions (marked in yellow) around the pyrimidine ring on the $R_4$ position, corresponding to two bulky favored regions (marked in green) in steric contour maps.

The graphical interpretation of the hydrogen-bond donor and acceptor interaction in the CoMSIA model is represented in Fig. 7b. Compound rac-34 was set as an example to illustrate the contour map. A hydrogen-bond acceptor disfavored contour (marked in red) close to the triazole of compound rac-34 ($pIC_{50}$=7.253) represents the region where hydrogen-bond acceptor group decreases activity.

2) $ET_B$ receptor antagonists

Figure 7c shows the hydrophobic fields for $ET_B$ affinity. Likely to $ET_A$, two hydrophobic regions (marked in yellow) are located around the pyrimidine ring on the $R_4$ position. Besides, there are two other hydrophobic regions (marked in yellow) and two hydrophilic regions (marked in white) inside the fields, which are distributed around the phenyl

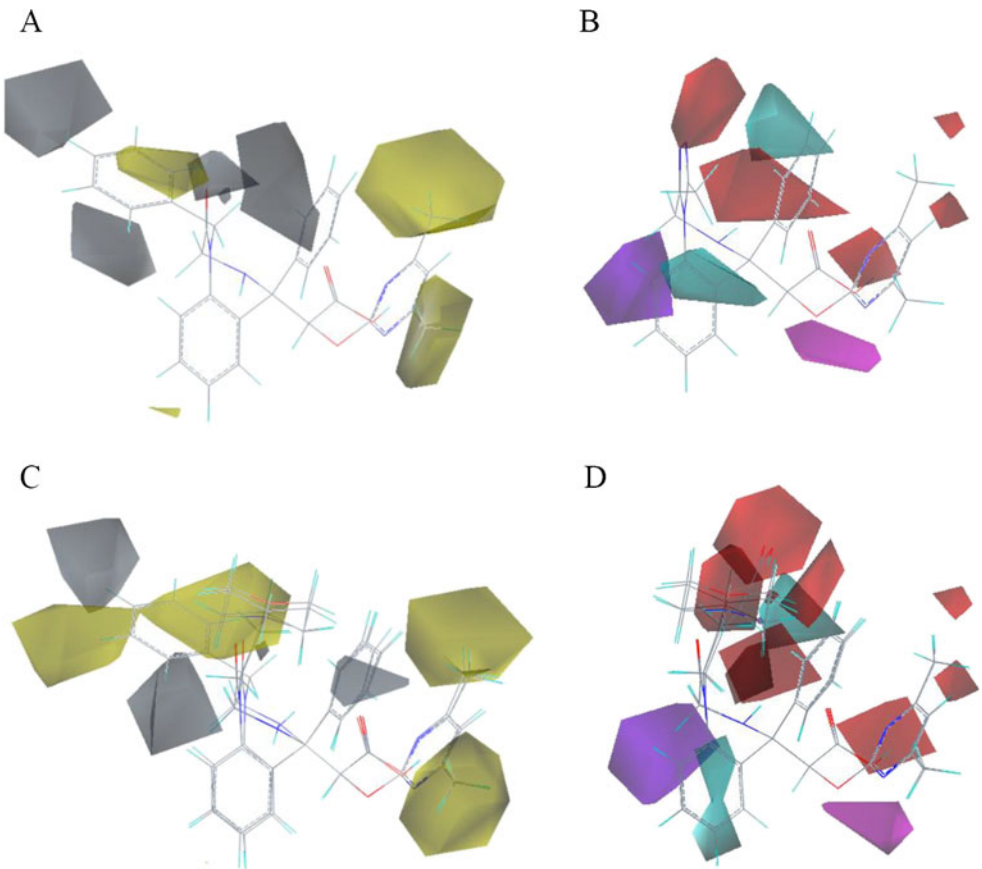



**Fig. 7** Hydrophobic and H-bond donor and acceptor contour maps of the CoMSIA regions are shown. (**a**) Hydrophobic contour map of the CoMSIA affinity model for $ET_A$ receptor. Compound rac-39 l is shown in A. (**b**) H-bond donor and acceptor contour map of the CoMSIA affinity model for $ET_A$ receptor. Compounds rac-39 l, rac-39e and rac-39f are shown in B. (**c**) Hydrophobic contour map of the CoMSIA affinity model for $ET_B$ receptor. Compound rac-34 is shown in C. (**d**) H-bond donor and acceptor contour map of the CoMSIA affinity model for $ET_B$ receptor. Compounds rac-39e, rac-39f and rac-39 k are shown in D. (Hydrophobic, hydrogen-bond donor and acceptor contour maps indicate favorable by yellow, cyan, magenta contours and unfavorable by white, violet/purple and red contours, respectively)

ring on the $R_1$ position. The hydrophobic cyclohexyl of compound rac-39e ($pIC_{50}$=6.815) and the hydrophilic morpholinyl of rac-39f ($pIC_{50}$=6.129) are mainly included in one of the above hydrophobic regions (marked in yellow).

Hydrogen-bond donor and acceptor fields for $ET_B$ affinity are shown in Fig. 7d. Two major hydrogen-bond acceptor disfavored contours (marked in red) are observed. One is at the carboxyl position of compound rac-39 k ($pIC_{50}$=6.597). The other one is occupied by the cyclohexyl position of compound rac-39e ($pIC_{50}$=6.815) and the morpholinyl of compound rac-39f ($pIC_{50}$=6.129). These red contour maps show that the occupancy of these regions would decrease the inhibitory activity. For instance, compound rac-39f ($pIC_{50}$=6.129) with hydrogen-bond acceptor group on the above area results in lower inhibitory activity than compound rac-39e ($pIC_{50}$=6.597).

## A comparison of $ET_A$ and $ET_B$ models

### *Steric fields*

As shown in Fig. 5c, compound rac-39 s was overlaid as reference to distinguish the affinities toward $ET_A$ ($pIC_{50}$= 7.006) and $ET_B$ ($pIC_{50}$=7.126). Its 2-substituent of the phenyl ring at the $R_1$ position is trifluoromethoxy (a sterically bulky group), which locates in the green region of Fig. 5c but in the yellow one of Fig. 5a. This indicates that a bulky group, at 2-substituent of phenyl ring at $R_1$ position, is sterically favored with $ET_B$ but disfavored with $ET_A$. Another example is the compound rac-87 ($ET_A$: $pIC_{50}$= 8.119, $ET_B$: $pIC_{50}$=8.149) with the highest affinity for $ET_B$ among title compounds. In Fig. 5c and d ($ET_B$ model), a green polyhedron (bulky favored) was adjacent to *n*-butyl group at $R_3$ position, whereas, in Fig. 5a, b ($ET_A$ model), a yellow polyhedron (bulky disfavored) located here. It shows that a sterically bulky group at $R_3$ position should be favored by $ET_B$ rather than $ET_A$, which can afford that compound rac-87 has a higher affinity to $ET_B$ than $ET_A$.

### *Electrostatic fields*

The electrostatic contour maps for $ET_A$ and $ET_B$ affinity by the CoMSIA model are almost the same (Fig. 6b, d). The distinction of electrostatic fields between $ET_A$ and $ET_B$ receptors is incarnated in CoMFA model. Compared with Fig. 6c, the electropositive groups favored region (marked in blue) at 4-substituent of phenyl ring on the $R_1$ position is absent, while the electronegative groups favored region (marked in red) appears at 2-substituent of phenyl ring on the $R_1$ position in Fig. 6a. In fact, compounds with 4-substituent of phenyl ring on the $R_1$ position also have high affinity for $ET_A$, such as compound rac-73 ($pIC_{50}$=8.201), compound rac-39af ($pIC_{50}$=8.337), which coincides with the result of CoMSIA model.

### *Hydrophobic, hydrogen-bond donor and acceptor fields*

The difference between hydrophobic contour maps for $ET_A$ (Fig. 7a) and $ET_B$ (Fig. 7c) is the distribution of hydrophobic favorable regions (marked in yellow) and hydrophobic unfavorable regions (marked in white) around the phenyl ring on the $R_1$ position. A small hydrophobic favorable region (marked in yellow) at the phenyl ring is besieged by four hydrophobic unfavorable regions (marked in white) in Fig. 7a, while the hydrophobic favorable region (marked in yellow) is larger than the hydrophobic unfavorable regions (marked in white) and not encircled in Fig. 7c. It is obvious that compound rac-39 l ($ET_A$: $pIC_{50}$=8.194) with the limited length of linker between the phenyl ring and the benzodiazepine scaffold is more potent than compound rac-39n ($ET_A$: $pIC_{50}$=7.767) with a long chain stretching to the white hydrophilic region for $ET_A$ affinity, while the result is opposite for $ET_B$ affinity of compound rac-39 l ($ET_B$: $pIC_{50}$=7.324), rac-39n ($ET_B$: $pIC_{50}$=7.738).

The distinction between the H-bond donor and acceptor fields for $ET_A$ and $ET_B$ affinity (Fig. 7b and d) is incarnated in the hydrogen-bond acceptor disfavored contours (marked in red). The red contour in Fig. 7b ($ET_A$ model) is close to the triazole of compound rac-34. The distribution of two red contours in Fig. 7d ($ET_B$ model) is unlike. One is at the carboxyl of compound rac-39 k, and the other one is occupied by the cyclohexyl of compound rac-39e and the morpholinyl of compound rac-39 f. That may be why the $ET_B$ affinity of compounds rac-39e and rac-39f is distinct, while the $ET_A$ affinity is nondistinctive.

## Conclusions

In summary, 79 benzodiazepine derivatives have been employed to develop CoMFA and CoMSIA models with the biological activities against the two subtypes of endothelin receptor: $ET_A$ and $ET_B$. The developed models have proven to predict the inhibitory potencies of related structures. The 3D-QSAR analysis reveals that the contribution of steric and hydrophobic fields may play a major role in ligand receptor interactions of title compounds. The contour analyses for the CoMFA and CoMSIA models have provided useful insights on designing novel antagonists for the respective endothelin receptors: the phenyl ring on the $R_1$ position is beneficial for both subtypes; moderate steric substituents like methyl at 4,6-substituents of the pyrimidine ring on the $R_4$ position are expected to improve the affinities to both $ET_A$ and $ET_B$; a more electronegative substitution at 2-substituent of phenyl ring on the $R_1$ position is favorable

for $ET_A$, while a more electropositive substitution at 4-substituent of phenyl ring on the $R_1$ position is likely to improve the inhibitory potency of both $ET_A$ and $ET_B$; considering the discrepancy between $ET_A$ and $ET_B$ models in the distribution of hydrophobic polyhedrons around the phenyl ring of $R_1$ position, it will be very helpful to replace with different hydrophobic substituents and adjust its spatial location to search for new antagonists with good subtype selectivity of $ET_A$ or $ET_B$ in the future. In summary, the results from 3D-QSAR models could be employed to bias the selectivity to either subtype, especially to $ET_A$, which should be more significant for the therapy of diseases caused by endothelins.

## References


1. Yanagisawa M, Kurihara H, Kimura H et al. (1988) A novel potent vasoconstrictor peptide produced by vascular endothelial cells. Nature 332:411–415
2. Arai H, Hori S, Aramori I et al. (1990) Cloning and expression of a cDNA encoding an endothelin receptor. Nature 348:730–732
3. Sakurai T, Yanagisawa M, Takuwa Y et al. (1990) Cloning of a cDNA encoding a nonisopeptide-selective subtype of the endothelin receptor. Nature 348:732–735
4. Opitz CF, Ewert R, Kirch W et al. (2008) Inhibition of endothelin receptors in the treatment of pulmonary arterial hypertension:does selectivity matter. Eur Heart J 29:1936–1948
5. Casserly B, Klinger JR (2008) Ambrisentan for the treatment of pulmonary arterial hypertension. Drug Des Devel Ther 2:265–280
6. Murugesan N, Gu Z, Spergel S (2003) Biphenylsulfonamide endothelin receptor antagonists. 4. Discovery of N-[[2'-[[(4,5-Dimethyl-3-isoxazolyl)amino]sulfonyl]-4-(2-oxazolyl)[1,1'-biphenyl]-2-yl]methyl]-N,3,3-trimethylbutanamide (BMS-207940), a highly potent and orally active ETA selective antagonist. J Med Chem 46:125–137
7. Ishizuka N, Matsumura K, Sakai K (2002) Structure-activity relationships of a novel class of endothelin-A receptor antagonists and discovery of potent and selective receptor antagonist, 2-(Benzo[1,3]dioxol-5-yl)-6-isopropyloxy-4-(4- methoxyphenyl)-2H-chromene-3-carboxylic Acid (S-1255). 1. Study on structure-activity relationships and basic structure crucial for ETA antagonism. J Med Chem 45:2041–2055
8. Kim NHS, Rubin LJ (2002) Endothelin in health and disease: Endothelin receptor antagonists in the management of pulmonary artery hypertension. Cardiovasc Pharmacol Ther 7:9–19
9. Girgis RE (2010) Emerging drugs for pulmonary hypertension. Expert Opin Emerging Drugs 15:71–85
10. Greenberg BH (2002) Endothelin and endothelin receptor antagonists in heart failure. Congest Heart Fail 8:257–261
11. Skalska AB, Grodzicki TJ (2010) Carotid atherosclerosis in elderly hypertensive patients:potential role of endothelin and plasma antioxidant capacity. Hum Hypertens 24:538–544
12. Prasad VS, Palaniswamy C, Frishman WH (2009) Endothelin as a clinical target in the treatment of systemic hypertension. Cardiol Rev 17:181–191
13. Moore R, Linas S (2010) Endothelin antagonists and resistant hypertension in chronic kidney disease. Curr Opin Nephrol Hypertens 19:432–436
14. Russo A, Bronte G, Rizzo S (2010) Anti-endothelin drugs in solid Tumors. Expert Opin Emerging Drugs 15:27–40
15. Krystek SR Jr et al. (1995) Three-dimensional quantitative structure-activity relationships of sulfonamide endothelin inhibitors. J Med Chem 38:659–668
16. Gu C, Hou T, Xu X (2001) Comparative molecular field analysis of γ-hydroxy butenolide endothelin antagonists. Chem J Chinese U 22:1864–1868
17. Zhu L, Xu X (2003) Three-dimensional quantitative structure-activity relationship analyses of a series of butenolide ETA antagonists. Chinese J Chem 21:261–269
18. Bolli MH, Marfurt J, Grisostomi C et al. (2004) Novel Benzo[1,4] diazepin-2-one derivatives as endothelin receptor antagonists. J Med Chem 47:2776–2795
19. Wold S, Rhue A, Wold H et al. (1984) The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM J Sci Stat Comput 5:735–743
20. Clark M, Cramer RD (1993) The probability of chance correlation using partial least squares (PLS). Quant Struct-Act Relat 12:137–145
21. Bush BL, Nachbar RB (1993) Sample-distance partial least squares:PLS optimized for many variables, with application to CoMFA. J Comput Aided Mol Des 7:587
22. Wold S (1978) Cross-validatory estimation of the number of components in factor and principal component analysis. Technometric 20:397–405
23. Klebe G, Abraham U, Mietzner T (1994) Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J Med Chem 37:4130–4146
24. Klebe G, Abraham U (1999) Comparative molecular similarity index analysis (CoMSIA) to study hydrogen-bonding properties and to score combinatorial libraries. J Comput Aided Mol Des 13:1–10

J Mol Model (2012) 18:1313–1324
DOI 10.1007/s00894-011-1156-7

ORIGINAL PAPER

# Proton transfer from $H_2O$ to p-substituted anilide anion: can the size of water cluster influence the $N^-\cdots H–OH\rightarrow N–H\cdots OH^-$ switching

**Hossein Roohi · Behnaz Moghadam**

Received: 18 April 2011 / Accepted: 15 June 2011 / Published online: 12 July 2011


**Abstract** The effects of the $H_2O$ cluster size and the substituent (X=H, Me, OMe, CHO, NO and $NO_2$) in the *para* position of anilide ion on the $HN^-\cdots H–OH\rightarrow HN–H\cdots OH^-$ switching were investigated by means of B3LYP and MP2 methods. The changes in the H-bond strength due to variation of the substituent and $H_2O$ cluster size were well monitored by changes in the interaction energy, structural parameter, electron density topology, natural charge, charge transfer and percentage of p-character of N atom in the C–N bond. The $HN^-\cdots H–OH\rightarrow HN–H\cdots OH^-$ switching was enhanced by an increase in the extent of solvation and the insertion of the electron-donating substituents in the *para* position. The results demonstrate that the strength of the $HN^-$ HO H-bond increases and that of NH $OH^-$ H-bond decreases by electron-donating substituents. The reverse situations were found for electron-accepting substituents.

**Keywords** Anilide · $H_2O$ cluster · Proton transfer · Quantum chemical methods

H. Roohi (✉)
Department of Chemistry, Faculty of Science,
University of Guilan,
Namjoh Street, P.O. Box 41335–1914, Rasht, Iran
e-mail: hroohi@guilan.ac.ir

H. Roohi
e-mail: hrouhi@gmail.com

B. Moghadam
Department of Chemistry, Faculty of Science,
University of Sistan & Baluchestan,
Zahedan, Iran

## Introduction

Proton transfer is a common phenomenon in the chemical and biological sciences [1–8]. Ionic hydrogen bond interactions are implicated in ionic crystals and clusters, ions solvation, electrolytes, and acid–base chemistry. The strength of ionic hydrogen bond ranges from 5 to 35 kcal $mol^{-1}$. The importance of this interaction in proton solvation, surface phenomenon, self-assembly process in supramolecular chemistry, and biomolecular structure has also been recognized. The formation of an ionic hydrogen bond in simple H-bonded complexes involves partial proton transfer from the donor to the acceptor [9].

Aniline and its derivatives are convenient systems to study intermolecular H-bonding and its influence on the molecular structure. Aniline can operate as a strong Bronsted base ($pK_b$=4.596 in water) and also as a weak acid ($pK_a$=30.6 in DMSO) [10]. In acidic media, aniline can form the anilinium cation in which N-H may also be involved in H-bonding as a proton donor. In strongly basic media, aniline can form the anilide anion that the nitrogen atom acts as a strongly basic center. Szatyłowicz et al. have investigated H-bonded complexes of aniline with HF/$F^-$ and anilide with HF at B3LYP/6-311+G(d,p) level [11] and the effects of H-bonding on the aromaticity of the ring in variously substituted complexes aniline/anilinium/anilide with bases and acids [10]. In addition, interrelations between the mesomeric and electronegativity effects in *para*-substituted derivatives of phenol/phenolate and aniline/anilide H-bonded complexes have been studied by Szatyłowicz et al. [12].

Because $OH^-$ in the gas phase is a stronger base than $C_6H_5NH^-$, proton transfer takes place from $C_6H_5NH_2$ to $OH^-$. Thus, proton-transfer chemical reaction $C_6H_5NH_2\cdots OH^- \rightarrow C_6H_5NH^-\cdots H_2O$ is thermodynamically favorable. As a result, in reaction between anilide anion and $H_2O$ in the gas phase, proton transfer from isolated $H_2O$ to anilide anion does not occur so that favorable structure in the gas phase is $C_6H_5NH^-\cdots H_2O$. There is an important question about the gas phase acidity of $H_2O$: Can the anilide/$H_2O$ system switches completely from $C_6H_5NH^-\cdots H_2O$ to $C_6H_5NH_2\cdots OH^-$? How does the system this switching? Wolf et al. [13] have studied photoelectron spectra of $C_6H_5NH^-(H_2O)_{0-6}$. They have shown that the spectra of the clusters corresponding to n=1 and 2 is similar to the spectra of the bare anilide anion but they are shifted to higher ionization energy for n>2 which can be attributed to the spectra of $OH^-$ solvated in polar-molecule clusters.

To the best of our knowledge, influences of the size of $H_2O$ cluster and substituent in the *para* position on the structural and electronic properties correspond to the transfer of the proton from the $H_2O$ to p-substituted anilide anion have not been reported. The aim of the present study is to investigate the effect of size of $(H_2O)_n$ network and substituents in the *para* position of anilide anion on the transfer of proton from $H_2O$ to anilide anion.

## Computational details

The geometries were optimized at the B3LYP/6-311++G(d,p) and B3LYP/6-311++G(2 d,2p) levels of theory. The B3LYP calculated vibrational frequencies have been used to characterize the stationary points and calculation of the zero-point vibrational energy (ZPVE). The vibrational frequencies were calculated without scaling. In addition, single-point calculations have been carried out at the MP2/6-311++G(2 d,2p) level of theory. The counterpoise procedure (CP) [14] was used to correct the basis set superposition error (BSSE) in the calculation of interaction energies. All calculations were performed using the GAMESS quantum chemistry package [15]. Topological properties were calculated at B3LYP/6-311++G(2 d,2p) level of theory by the AIM2000 program package [16]. The NBO analysis [17] was carried out on the B3LYP/6-311++G(2 d,2p) wave functions using version 3.1 of NBO package [18].

## Results and discussion

### Geometry and energy

Analysis of variously substituted anilide/aniline in H-bonded complexes **X–An-(W)$_{1,2}$** and **X–Ph-NH$_2$(OH$^-$)(W)$_{2,3}$**

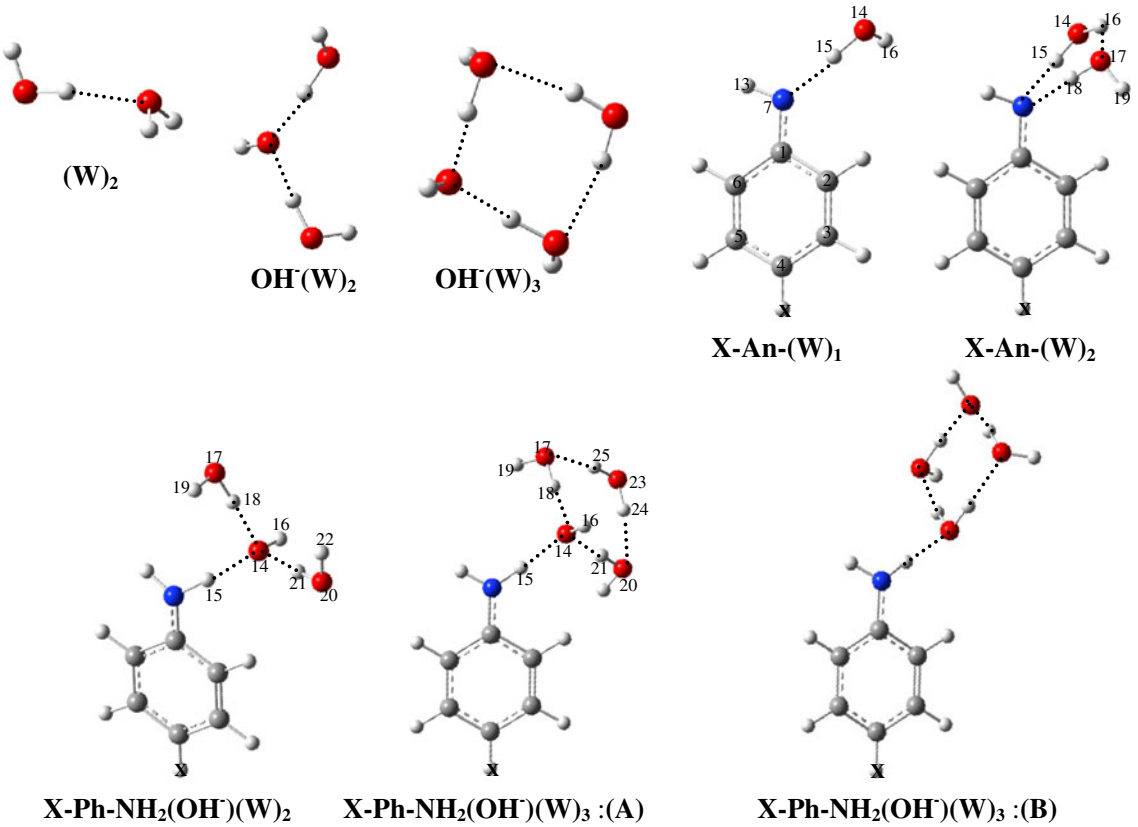



**Fig. 1** The optimized geometries in the gas phase for monomers and complexes

(X=NO, $NO_2$, CHO, H, $CH_3$ and $OCH_3$) revealed that the H-bond distances, *ipso-ortho* C–C and C–N bonds lengths and the ipso angles are strongly affected by change in the cluster size of water and substituents in *para* position of anilide ion. The above-mentioned parameters describe geometrical pattern of the close vicinity of the H-bonding. Structural consequences of H-bonding can also affect on the aromaticity of ring. The optimized structures of the most stable complexes **X–An-(W)$_n$** are depicted in Fig. 1. The selected structural parameters for **X–An-(W)$_n$** complexes are given in Table 1. The **X–An-(W)$_1$** complexes are stabilized by a near linearly N7 H15–O H-bonding (176°). Two N7 H15–O and N7 H18–O H-bonds are formed in cyclic **X–An-(W)$_2$** complexes. As can be seen in Fig. 1, interaction between p-

**Table 1** The optimized geometrical parameters (distances in Å and angles in °) at B3LYP/6-311++G(2 d,2p) level

| | **X-An-(W)$_1$** | | | | | | **X-An-(W)$_2$** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | X=H | Me | OMe | CHO | NO | $NO_2$ | X=H | Me | OMe | CHO | NO | $NO_2$ |
| C1-C6 | 1.437 | 1.434 | 1.436 | 1.444 | 1.453 | 1.447 | 1.430 | 1.427 | 1.429 | 1.442 | 1.448 | 1.441 |
| C1-C2 | 1.436 | 1.435 | 1.435 | 1.447 | 1.445 | 1.445 | 1.430 | 1.429 | 1.429 | 1.437 | 1.440 | 1.440 |
| C1-N7 | 1.338 | 1.340 | 1.340 | 1.322 | 1.317 | 1.319 | 1.350 | 1.352 | 1.352 | 1.332 | 1.326 | 1.328 |
| N7-H13 | 1.016 | 1.016 | 1.016 | 1.017 | 1.017 | 1.017 | 1.016 | 1.016 | 1.016 | 1.016 | 1.017 | 1.017 |
| O14-H15 | 1.018 | 1.018 | 1.017 | 1.003 | 0.999 | 0.999 | 1.012 | 1.012 | 1.012 | 1.000 | 0.997 | 0.997 |
| O14-H16 | 0.960 | 0.960 | 0.960 | 0.960 | 0.960 | 0.960 | 0.960 | 0.960 | 0.960 | 0.960 | 0.960 | 0.960 |
| O17-H18 | | | | | | | 0.989 | 0.990 | 0.989 | 0.980 | 0.976 | 0.977 |
| O17-H19 | | | | | | | 0.964 | 0.964 | 0.964 | 0.964 | 0.965 | 0.965 |
| N7 H15 | 1.690 | 1.687 | 1.690 | 1.757 | 1.778 | 1.775 | 1.729 | 1.727 | 1.729 | 1.781 | 1.800 | 1.800 |
| N7 H18 | | | | | | | 1.892 | 1.885 | 1.889 | 1.995 | 2.044 | 2.025 |
| O17 H16 | | | | | | | | | | | | |
| C1-C2-C6 | 114.4 | 114.0 | 114.3 | 115.0 | 115.5 | 115.1 | 114.9 | 114.5 | 114.8 | 115.4 | 115.8 | 115.4 |
| O14-H15-N7 | 175.8 | 175.9 | 175.9 | 176.1 | 176.1 | 176.2 | 176.5 | 176.5 | 176.4 | 176.4 | 176.2 | 176.2 |
| O17-H18-N7 | | | | | | | 161.1 | 160.2 | 161.8 | 164.1 | 164.2 | 164.1 |
| | **X-Ph-NH$_2$(OH$^-$)(W)$_2$** | | | | | | **X-Ph-NH$_2$(OH$^-$)(W)$_3$** | | | | | |
| C1-C6 | 1.412 | 1.409 | 1.409 | 1.426 | 1.426 | 1.425 | 1.411 | 1.407 | 1.409 | 1.420 | 1.423 | 1.423 |
| C1-C2 | 1.414 | 1.413 | 1.412 | 1.422 | 1.433 | 1.426 | 1.412 | 1.411 | 1.411 | 1.425 | 1.431 | 1.424 |
| C1-N7 | 1.364 | 1.368 | 1.373 | 1.339 | 1.332 | 1.335 | 1.368 | 1.372 | 1.372 | 1.343 | 1.337 | 1.340 |
| N7-H13 | 1.006 | 1.006 | 1.006 | 1.005 | 1.007 | 1.006 | 1.005 | 1.006 | 1.006 | 1.005 | 1.006 | 1.006 |
| N7-H15 | 1.055 | 1.054 | 1.051 | 1.076 | 1.086 | 1.084 | 1.049 | 1.047 | 1.048 | 1.071 | 1.085 | 1.082 |
| O14 H15 | 1.660 | 1.672 | 1.689 | 1.558 | 1.521 | 1.526 | 1.697 | 1.709 | 1.702 | 1.575 | 1.527 | 1.536 |
| O14 H18 | 1.629 (1.533) | 1.628 | 1.632 | 1.651 | 1.654 | 1.655 | 1.590 (1.497) | 1.588 | 1.591 | 1.626 | 1.640 | 1.637 |
| O14 H21 | 1.641 (1.545) | 1.637 | 1.601 | 1.659 | 1.673 | 1.669 | 1.593 (1.500) | 1.589 | 1.591 | 1.622 | 1.636 | 1.634 |
| O14-H16 | 0.961 | 0.961 | 0.960 | 0.961 | 0.961 | 0.961 | 0.961 | 0.961 | 0.961 | 0.961 | 0.961 | 0.961 |
| O17-H18 | 1.013 (1.039) | 1.014 | 1.012 | 1.008 | 1.006 | 1.006 | 1.020 (1.047) | 1.021 | 1.020 | 1.012 | 1.009 | 1.010 |
| O17-H19 | 0.960 | 0.960 | 0.960 | 0.960 | 0.960 | 0.960 | 0.960 | 0.960 | 0.960 | 0.960 | 0.960 | 0.960 |
| O17 H25 | | | | | | | 1.986 (1.962) | 1.984 | 1.984 | 2.013 | 2.018 | 2.008 |
| O20-H21 | 1.010 (1.035) | 1.011 | 1.016 | 1.005 | 1.003 | 1.003 | 1.019 (1.046) | 1.020 | 1.019 | 1.012 | 1.009 | 1.009 |
| O20-H22 | 0.960 | 0.960 | 0.960 | 0.959 | 0.959 | 0.959 | 0.960 | 0.960 | 0.960 | 0.960 | 0.960 | 0.960 |
| O20 H24 | | | | | | | 1.982 (1.965) | 1.983 | 1.987 | 1.977 | 1.982 | 1.990 |
| O23-H24 | | | | | | | 0.973 (0.975) | 0.973 | 0.973 | 0.973 | 0.973 | 0.972 |
| O23-H25 | | | | | | | 0.973 (0.975) | 0.973 | 0.973 | 0.972 | 0.971 | 0.972 |
| C1-C2-C6 | 117.3 | 116.8 | 117.2 | 117.4 | 117.7 | 117.3 | 117.4 | 117.0 | 117.3 | 117.5 | 117.8 | 117.4 |
| O14-H15-N7 | 176.6 | 176.4 | 177.3 | 175.2 | 172.6 | 173.8 | 178.5 | 179.0 | 179.4 | 176.4 | 176.9 | 178.2 |
| O17-H18-O14 | 175.5 | 175.4 | 175.4 | 175.3 | 174.1 | 174.6 | 176.3 | 176.3 | 176.4 | 176.4 | 176.6 | 176.9 |
| O20-H21-O14 | 175.2 | 175.3 | 176.7 | 175.9 | 175.5 | 176.0 | 176.7 | 176.7 | 177.0 | 176.6 | 176.6 | 176.9 |
| O17-H25-O23 | | | | | | | 166.1 | 166.2 | 166.2 | 164.3 | 164.0 | 165.0 |
| O20-H24-O23 | | | | | | | 166.2 | 166.1 | 165.9 | 166.2 | 166.1 | 165.9 |

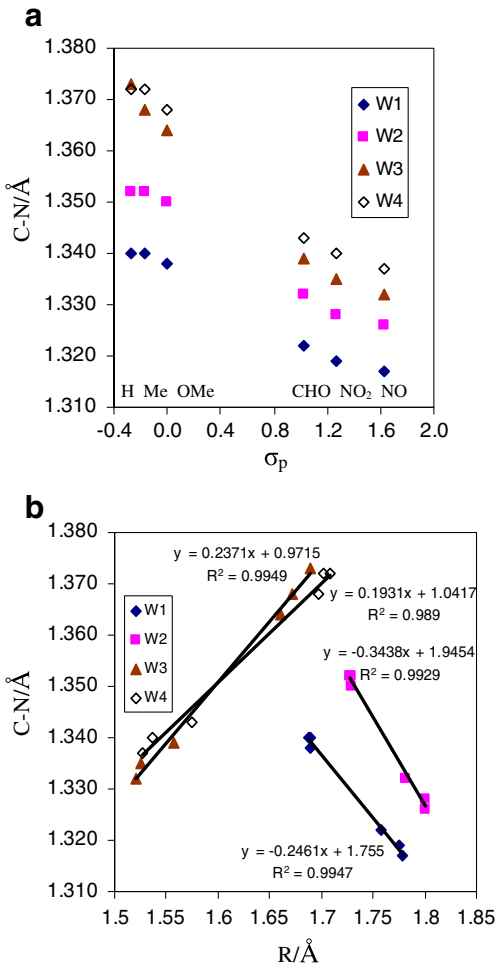



**Fig. 2** (**a**) Relation between C–N bond distances and substituent constants, $\sigma_p$ in **X-An-(W)$_{1,2}$** (symbolized as W1 and W2) and **X-Ph-NH$_2$(OH$^-$)(W)$_{2,3}$** (symbolized as W3 and W4) complexes. (**b**) Correlation between C–N and H-bond (R) distances

substituted anilide anions and one and two $H_2O$ molecules forms X-$C_6H_4$–N $(H_2O)_{1,2}$ H-bonded complexes without proton transfer. In **X–An-(W)$_{3,4}$** complexes, proton is transferred from $H_2O$ to anilide ion through hydrogen bond and consequently N7 H15–O H-bond is exchanged by N7–H15 O one. Thus, N7 H15 H-bond is converted to N7–H15 covalent one and negative charge is transferred from anilide ion to $OH^-$. In these complexes, $OH^-$ ions have been surrounded by aniline and $H_2O$ molecules. Thus, it is reliable better that the **X–An-(W)$_{3,4}$** complexes are written as **X–Ph-NH$_2$(OH$^-$)(W)$_{2,3}$**. For **X–Ph-NH$_2$(OH$^-$)(W)$_{2,3}$** complex, two configurations were predicted (Fig. 1) in which structure **A** is

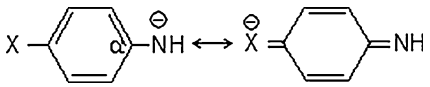



**Chart 1** Resonance delocalization of the negative charge

the most stable structure. The $OH^-$ in structure **A** is surrounded by aniline and water molecules whereas in structure **B** by water molecules only.

The most varying structural parameters in **X–An-(W)$_{1–4}$** complexes are H-bond distances and C–N bond lengths. Correlation between C–N bond lengths and substituent constants, $\sigma_p$, as well as H-bond distances are shown in Fig. 2. The C–N bond length is 1.331 Å in anilide ion, 1.338 Å in **H–An-(W)$_1$**, 1.350 Å in **H–An-(W)$_2$**, 1.364 Å in **H–Ph-NH$_2$(OH$^-$)(W)$_2$** and 1.368 Å (1.373 Å in complex **B**) in **X–Ph-NH$_2$(OH$^-$)(W)$_3$** (complex **A**). As a result, an increase in size of $H_2O$ cluster is associated with the increase in C–N bond length. The changes in the C–N bond length may be used as a measure of the H-bond strength for **X–An-(W)$_{1–4}$** complexes. From Fig. 2b, it is found that the decrease in H-bond distance in **X–An-(W)$_{1,2}$** complexes for which the nitrogen atom is a proton acceptor is accompanied by an increase in C–N bond length. A reverse trend is observed for **X–Ph-NH$_2$(OH$^-$)(W)$_{2,3}$** complexes which the nitrogen atom is a proton donor. In free anilide anion, negative charge on the nitrogen atom is mesomerically delocalized over the ring, leading to a decrease in C–N bond length. In **X–An-(W)$_{1,2}$** complexes, shortening of N7 H15 H-bond distance is associated with a decrease in the negative charge of N atom and consequently with an increase in C–N bond length and decrease in weight of the quinoid-like structure (Chart 1). The presence of electron-accepting substituents leads to the transfer of charge from the N atom to the ring and then decrease in C–N bond length. In contrast, the presence of electron-donating groups in the *para* position increases the C–N bond lengths and in turn decreases weight of the quinoid-like structure.

The H-bond distances in monomers **OH$^-$(W)$_{2,3}$** changes upon complex formation. Obviously, corresponding HB strength changes upon complexation, depending on power of electron withdrawing or donating substituents. As can be seen in Table 2, all H-bonding distances in the **OH$^-$(W)$_{2,3}$** increase, consequently strengths of H-bonding interactions decrease upon complexation.

In anilide/aniline systems, the *ipso* angle, α, changes upon complex formation. It has been proposed [19] that α angle can be used as a sufficient geometrical parameter describing the electronegativity of groups involved in the interaction. As can be seen in Table 2, value of α angle in all series of X–anilide (W=1–4) substituted by electron accepting groups (X=CHO, NO and $NO_2$) is greater than those of electron donating groups (X=Me, OMe) and in NO-anilide is greatest. In addition, increase in size of water cluster is accompanied by an increase in value of α angle.

The interaction energies (IEs) computed according to reactions X-anilide+$(H_2O)_{1,2}$→X-anilide-$(H_2O)_{1,2}$ and X-aniline+$(OH^-)(H_2O)_{2,3}$→X-aniline-$(OH^-)(H_2O)_{2,3}$ are given

in Table 2. Figure 3 shows the correlation between BEs and substituent constants at the B3LYP/6-311++G(2 d,2p) level of theory. It can be seen that the absolute values of BEs for **X–An-(W)$_{1,2}$** complexes are greater for electron donating substituents than those of electron-accepting ones. In other words, the strong electron-donating/accepting groups of the substituents increase/decrease the BEs. An increase in the BE (as the absolute value) is associated with an immobilization of the lone pair of nitrogen toward the ring because it is involved in H-bond formation. For **X–Ph-NH$_2$(OH$^-$)(W)$_{2,3}$**) complexes situation is reverse, *i.e.*, hydrogen bonding strength decreases as the substituent changes from electron-acceptor to electron-donor. For **X–Ph-NH$_2$(OH$^-$)(W)$_{2,3}$**) complexes, an increase of BE (as the absolute value) is associated with a mobilization of the lone pair of nitrogen toward the ring. Therefore, it is expected electron-accepting substituents stimulate performance of this process.

Dependence on the size of water cluster, correlation between BE and size of the cluster is different. For the NH$^-$ HO interactions observed in **X–An-(W)$_{1,2}$** com-

**Table 2** Interaction energies (kJ mol$^{-1}$) calculated at various levels of theory

| | **X-An-(W)$_1$** | | | | **X-An-(W)$_2$** | | |
|---|---|---|---|---|---|---|---|
| X– | Basis set | $\Delta E_{elec}$ | $\Delta E_{ele}^{BSSE}$ | $\Delta E_{ele,ZPE}^{BSSE}$ | $\Delta E_{elec}$ | $\Delta E_{ele}^{BSSE}$ | $\Delta E_{ele,ZPE}^{BSSE}$ |
| H | L1[a] | −78.4 | −75.7 | −66.3 | −118.7 | −113.0 | −101.7 |
| | L2[a] | −74.7 | −72.8 | −64.0 | −114.0 | −110.9 | −100.6 |
| | L3[a] | −78.2 | | | −125.3 | | |
| Me | L1 | −78.8 | −76.1 | −68.1 | −119.6 | −113.6 | −103.6 |
| | L2 | −75.2 | −73.1 | −65.1 | −115.0 | −111.7 | −102.3 |
| | L3 | −78.6 | | | −126.5 | | |
| OMe | L1 | −77.7 | −75.0 | −66.9 | −117.8 | −111.8 | −102.0 |
| | L2 | −74.1 | −72.2 | −64.4 | −113.3 | −110.1 | −100.9 |
| | L3 | −77.9 | | | −122.5 | | |
| CHO | L1 | −64.9 | −62.4 | −54.2 | −94.3 | −89.4 | −79.8 |
| | L2 | −61.6 | −59.8 | −52.0 | −90.3 | −87.5 | −78.6 |
| | L3 | −65.6 | | | −100.2 | | |
| NO | L1 | −60.5 | −58.1 | −49.8 | −85.4 | −80.5 | −71.3 |
| | L2 | −57.5 | −55.7 | −47.8 | −81.7 | −79.0 | −70.4 |
| | L3 | −58.9 | | | −89.6 | | |
| $NO_2$ | L1 | −60.9 | −58.3 | −50.0 | −86.1 | −81.1 | −71.7 |
| | L2 | −57.8 | −56.1 | −47.9 | −82.4 | −79.7 | −70.8 |
| | L3 | −60.5 | | | −92.7 | | |
| | **X-Ph-NH2-(OH)(W)$_2$** | | | | **X-Ph-NH2-(OH)(W)$_3$:A(B)** | | |
| H | L1 | −83.2 | −78.9 | −74.6 | −72.8 | −69.8 | −62.9 |
| | L2 | −77.1 | −73.8 | −69.6 | −67.6(−53.2) | −65.9 | −61.6 |
| | L3 | −90.8 | | | −76.6 | | |
| Me | L1 | −77.4 | −72.7 | −68 | −66.9 | −63.1 | −58.6 |
| | L2 | −74.1 | −71.5 | −67.1 | −65.0(−50.7) | −63.2 | −58.6 |
| | L3 | −88.4 | | | −74.5 | | |
| OMe | L1 | −82.7 | −77.8 | −72.9 | −70.8 | −66.9 | −62.2 |
| | L2 | −79.8 | −76.4 | −71.3 | −68.7(−54.3) | −67.9 | −63.2 |
| | L3 | −90.1 | | | −77.6 | | |
| CHO | L1 | −119 | −13.8 | −108.7 | −101.3 | −97.7 | −91.9 |
| | L2 | −114.8 | −111.9 | −105.2 | −98.6(−82.50 | −96.7 | −90.6 |
| | L3 | −121.4 | | | −105.7 | | |
| NO | L1 | −130.9 | −127.0 | −123 | −115 | −111.3 | −104.9 |
| | L2 | −129.1 | −125.9 | −119.0 | −112.8(−93.7) | −111.0 | −104.7 |
| | L3 | −132.3 | | | −116.3 | | |
| $NO_2$ | L1 | −131.4 | −127.5 | −123.6 | −115.6 | −111.9 | −105.8 |
| | L2 | −130.3 | −127.2 | −120.8 | −114.1(−95.2) | −112.2 | −106.2 |
| | L3 | −131.5 | | | −116.4 | | |

[a] L1=B3LYP/6-311++G(d,p). L2=B3LYP/6-311++G(2 d,2p), L3=MP2/6-311++G(2 d,2p)

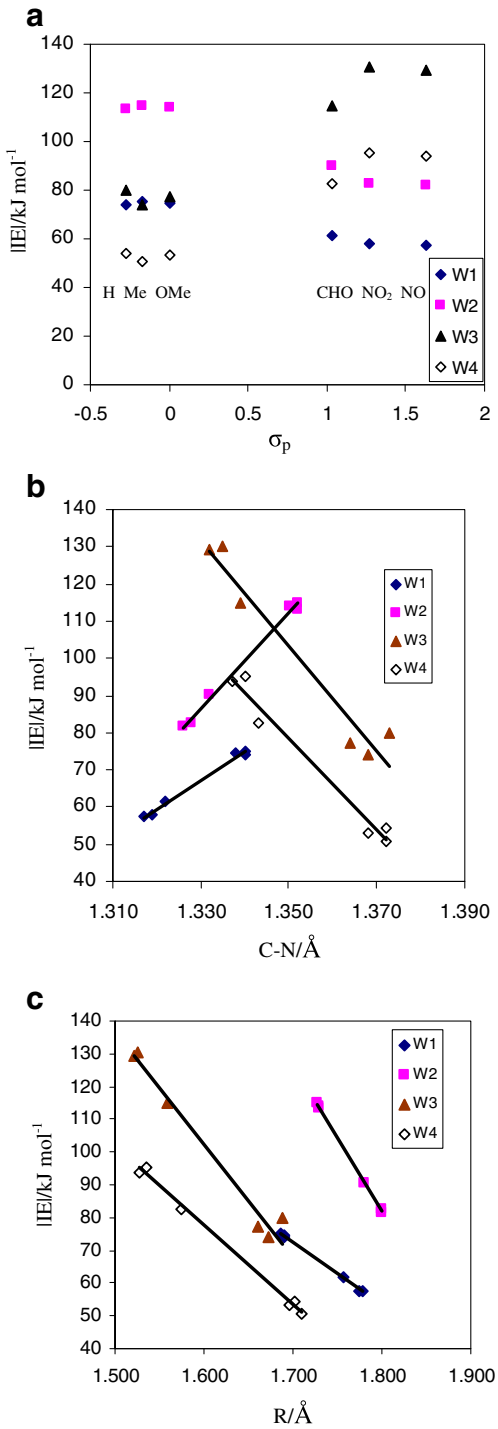



**Fig. 3** Correlation between the absolute value of interaction energy | IE| and (**a**) substituent constants $\sigma_p$, (**b**) C–N bond distances and (**c**) H-bond distances (R)

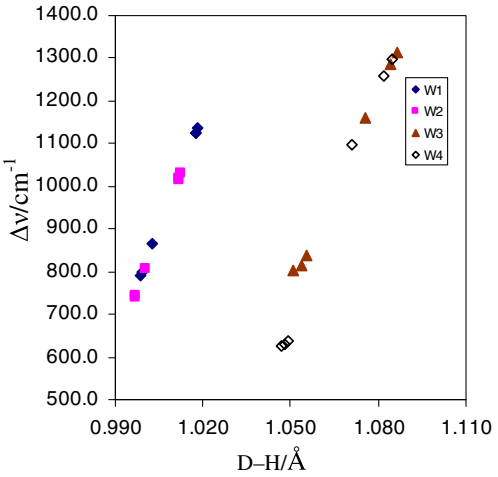



**Fig. 4** Correlation between red-shift of D–H stretching frequencies ($\Delta\nu$) and D–H bond distances

plexes, BE of complexes increases as the number of $H_2O$ in complex increases. A reverse trend is observed for **X–Ph-NH$_2$(OH$^-$)(W)$_{2,3}$** complexes. Increase in size of water cluster is accompanied by the decrease in BE, i.e., $OH^-$ is better solvated by water molecules as the size of cluster increases, in turn interaction $OH^-$ with aniline decreases. Thus, base strength depends not only on the availability of the unshared pairs for sharing with a hydrogen ion but also the environment in which it is surrounded.

From Table 2, BEs predicted by B3LYP method are lower than MP2 one. Besides, although zero-point vibrational energy (ZPVE) correction has a significant effect on the BE of complexes, inclusion of ZPVE does not change the stability order of complexes.

Correlation between BE and C–N as well as H-bond distances are shown in Figs. 3b, c. As can be seen, increase in BE of **X–An-(W)$_{1,2}$** complexes is associated with the lengthening of C–N bond and reduction of H-bond distance. Electron-accepting substituents stimulate electron delocalization toward the ring, consequently, C–N bond decreases as the BE decreases. The reverse situation is observed for **X–Ph-NH$_2$(OH$^-$)(W)$_{2,3}$** complexes. Electron-accepting substituents strengthen H-bond interaction, consequently, decrease in C–N bond is associated with the increase in BE.

Inspection of the vibrational frequencies shows that the complex formation results in a marked change of vibrational frequency of D–H (D denotes donor) bonds involved in H-bonding. Red-shift of O–H and N–H stretching frequencies ($\Delta\nu$) as a function of D–H bond distance in **X–An-(W)$_{1,2}$** and **X–Ph-NH$_2$(OH$^-$)(W)$_{2,3}$** complexes are depicted in Fig. 4. As expected in conventional H-bonding, interaction leads to increase in D–H bond distance and consequently, a substantial red-shift in the fundamental D–H stretching vibrational frequency. Analysis of frequencies shows that the amount of red-shifts of O–H stretching frequencies in **X–An-(W)$_{1,2}$** complexes are greater for electron-donating than electron-accepting substituents. In **X–An-(W)$_{1,2}$** complexes that nitrogen atom has a negative

charge, electron-attracting substituents lead to the decrease in H-bond strength and hence decrease in red shift of frequency. For **X–Ph-NH$_2$(OH$^-$)(W)$_{2,3}$** complexes, red-shift of N–H stretching frequency is reverse.

## AIM analysis

The quantum theory of atoms in molecules (QTAIM) is a useful tool to characterize the hydrogen bonding. The presence of bond critical point is used as one of the criteria for identifying the presence of H-bonding in any intermolecular complexes [20, 21]. The topological parameters derived from the Bader theory are often used as descriptors of H-bond strength. The calculated values of electron density, ρ(r), Laplacian of electron density, $\nabla^2\rho(r)$, and electronic energy density, H(r), at the bond critical points (BCPs) at the B3LYP/6-311++G (2 d,2p) level of theory are listed in Table 3. The molecular graphs (including the critical points and bond paths) of the complexes are shown in Fig. 5. Besides all of the expected BCPs, the topological features of electron density reveal the additional BCPs in H-bond distances.

Since two H-bonds exist between anilide and **(W)$_2$** in the case of **X–An-(W)$_2$**, hereafter, sum of two ρ(r) at HBCPs (N H15 and N H18) are considered. From Table 3 and Fig. 6a, a correlation can be found between the ρ(r) at the C–N BCPs and hydrogen bond critical points (HBCPs). For NH$^-$ HO interactions found in **X–An-(W)$_{1,2}$**, regressions

**Table 3** Topological properties of the electron density (au) at B3LYP/6-311++G(2 d,2p) level

| | **X–An-(W)$_1$** | | | **X–An-(W)$_2$** | | | **X–Ph-NH$_2$(OH$^-$)(W)$_2$** | | | **X–Ph-NH$_2$(OH$^-$)(W)$_3$** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| X- | ρ(r) | $\nabla^2\rho(r)$ | H(r) | ρ(r) | $\nabla^2\rho(r)$ | H(r) | ρ(r) | $\nabla^2\rho(r)$ | H(r) | ρ(r) | $\nabla^2\rho(r)$ | H(r) |
| | | | | | | C1-N7 | | | | | | |
| H | 0.3453 | −1.0693 | −0.4387 | 0.3369 | −1.0162 | −0.4122 | 0.3229 | −1.0076 | −0.4054 | 0.3202 | −0.9949 | −0.3999 |
| Me | 0.3442 | −1.0604 | −0.4346 | 0.3356 | −1.0063 | −0.4079 | 0.3210 | −0.9937 | −0.3983 | 0.3177 | −0.9775 | −0.3912 |
| OMe | 0.3444 | −1.0574 | −0.4330 | 0.3356 | −1.0048 | −0.4072 | 0.3173 | −0.9711 | −0.3879 | 0.3176 | −0.9739 | −0.3891 |
| CHO | 0.3573 | −1.1547 | −0.4848 | 0.3504 | −1.1147 | −0.4600 | 0.3389 | −1.1029 | −0.4640 | 0.3361 | −1.0899 | −0.4562 |
| NO | 0.3607 | −1.1766 | −0.4994 | 0.3547 | −1.1449 | −0.4774 | 0.3440 | −1.1262 | −0.4825 | 0.3405 | −1.1109 | −0.4718 |
| NO$_2$ | 0.3596 | −1.1688 | −0.4936 | 0.3533 | −1.1335 | −0.4703 | 0.3424 | −1.1197 | −0.4738 | 0.3388 | −1.1021 | −0.4626 |
| | | | | | | C1-C2 | | | | | | |
| H | 0.2898 | −0.7638 | −0.2706 | 0.2936 | −0.7839 | −0.2783 | 0.3021 | −0.8333 | −0.2974 | 0.3029 | −0.8373 | −0.2994 |
| Me | 0.2905 | −0.7681 | −0.2720 | 0.2942 | −0.7878 | −0.2795 | 0.3027 | −0.8368 | −0.2985 | 0.3035 | −0.8410 | −0.3004 |
| OMe | 0.2901 | −0.7643 | −0.2712 | 0.2938 | −0.7844 | −0.2787 | 0.3027 | −0.8357 | −0.2989 | 0.3033 | −0.8385 | −0.3000 |
| CHO | 0.2850 | −0.7424 | −0.2604 | 0.2898 | −0.7668 | −0.2704 | 0.2973 | −0.8102 | −0.2878 | 0.2959 | −0.8030 | −0.2847 |
| NO | 0.2857 | −0.7457 | −0.2620 | 0.2884 | −0.7599 | −0.2675 | 0.2921 | −0.7831 | −0.2762 | 0.2930 | −0.7880 | −0.2784 |
| NO$_2$ | 0.2859 | −0.7473 | −0.2620 | 0.2887 | −0.7616 | −0.2677 | 0.2955 | −0.8007 | −0.2835 | 0.2966 | −0.8063 | −0.2860 |
| | | | | | | N7-H15 (N7-H18) | | | | | | |
| H | 0.0548 | 0.0853 | −0.0119 | 0.0507 (0.0347) | 0.0823 (0.0801) | −0.0098 (−0.0022) | 0.3010 | −1.6123 | −0.4552 | 0.3070 | −1.6608 | −0.4672 |
| Me | 0.0552 | 0.0852 | −0.0121 | 0.0512 (0.0354 | 0.0822 (0.0807) | −0.0101 (−0.0024) | 0.3028 | −1.6214 | −0.4576 | 0.3087 | −1.6649 | −0.4684 |
| OMe | 0.0549 | 0.0854 | −0.0119 | 0.0508 (0.0350) | 0.0823 (0.0805) | −0.0099 (−0.022) | 0.3055 | −1.6406 | −0.4623 | 0.3080 | −1.6598 | −0.4672 |
| CHO | 0.0465 | 0.0877 | −0.0072 | 0.0442 (0.0274) | 0.0845 (0.0714) | −0.0063 (0.0000) | 0.2822 | −1.4760 | −0.4223 | 0.2860 | −1.5112 | −0.4308 |
| NO | 0.0442 | 0.0877 | −0.0060 | 0.0421 (0.0244) | 0.0846 (0.0665) | −0.0053 (0.0006) | 0.2736 | −1.3922 | −0.4024 | 0.2748 | −1.4032 | −0.4050 |
| NO$_2$ | 0.0445 | 0.0877 | −0.0062 | 0.0421 (0.0255) | 0.0844 (0.0685) | −0.0053 (0.0004) | 0.2755 | −1.4013 | −0.4044 | 0.2773 | −1.4199 | −0.4089 |
| | | | | | | O14-H15 | | | | | | |
| H | 0.3076 | −2.1608 | −0.6141 | 0.3392 | −2.4705 | −0.6912 | 0.0522 | 0.1229 | −0.0071 | 0.0478 | 0.1183 | −0.0053 |
| Me | 0.3071 | −2.1536 | −0.6124 | 0.3384 | −2.4618 | −0.6891 | 0.0508 | 0.1216 | −0.0065 | 0.0465 | 0.1169 | −0.0048 |
| OMe | 0.3078 | −2.1622 | −0.6145 | 0.3392 | −2.4689 | −0.6909 | 0.0486 | 0.1194 | −0.0056 | 0.0473 | 0.1178 | −0.0050 |
| CHO | 0.3228 | −2.3302 | −0.6553 | 0.3505 | −2.5669 | −0.7160 | 0.0671 | 0.1312 | −0.0150 | 0.0642 | 0.1306 | −0.0133 |
| NO | 0.3271 | −2.3757 | −0.6666 | 0.3545 | −2.5968 | −0.7239 | 0.0739 | 0.1306 | −0.0194 | 0.0726 | 0.1302 | −0.0185 |
| NO$_2$ | 0.3267 | −2.3714 | −0.6655 | 0.3533 | −2.5890 | −0.7219 | 0.0727 | 0.1327 | −0.0185 | 0.0709 | 0.1320 | −0.0173 |

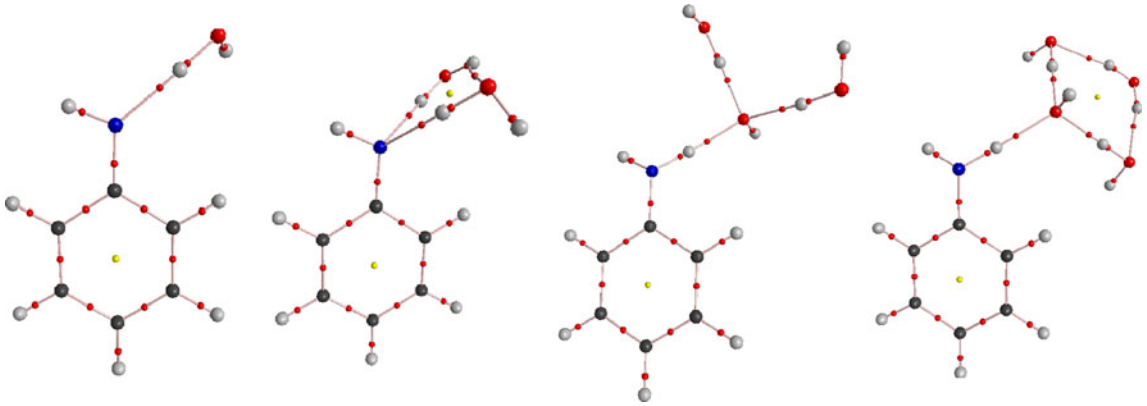
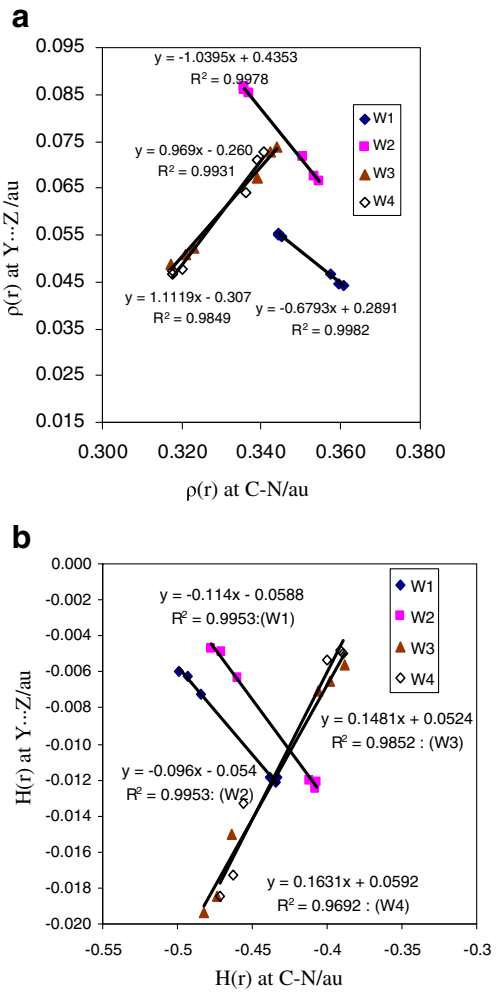


**Fig. 5** Molecular graphs of the of **H-An(W)$_{1,2}$** and **H-Ph-NH$_2$OH(W)$_{2,3}$** complexes. Nuclei and critical points (bond and ring) are represented by big and small circles, respectively




**Fi\g. 6** (**a**) Correlation between electron densities at C-N BCPs and Y Z HBCPs (Y=NH$^-$ or NH$_2$ and Z=H$_2$O or OH$^-$). (**b**) Relationship between H(r) at C-N BCPs and Y Z HBCPs

have negative slopes. There is a simple explanation: the formation of NH$^-$ HO bonding is associated with a decrease of mobility of lone-pair toward the ring because it becomes a partner for H-bond formation. The more electron density at HBCP, the less mobile the lone pair; hence, the decrease in ρ(r) of the C–N bond may be observed. If in the *para* position, an electron-donating substituent is attached then due to its donating power, decrease in ρ(r) of the C–N bond may be enhanced. Increase in the H-bond strength is accompanied by the lower mobility of the electron pair. This trend is opposite for electron-accepting substituents. However, for interactions in which the NH$_2$ group is proton-donating (i.e., NH$_2$ OH$^-$ ones in **X–Ph-NH$_2$(OH$^-$)(W)$_{2,3}$**), the correlation between ρ(r) at the C–N and NH$_2$ OH$^-$ BCPs has a positive slope. In this case, increase in NH$_2$ OH$^-$ H-bond strength (i.e., increase in ρ(r) of NH$_2$ OH$^-$ HBCP) is associated with the decrease in the C–N bond and consequently increase in ρ(r) of its BCP. The strong NH$_2$ OH$^-$ H-bonding interaction augments mobility of lone-pair toward the ring leading to increase in ρ(r) at the C–N BCP. An electron-accepting substituent in the *para* position enhances lone-pair mobility, which causes the H-bonding strength increases.

The analysis of the electron density property shows that the H(r) and $\nabla^2\rho(r)$ at both NH$^-$ HOH (in **X–An-(W)$_{1,2}$**) and NH$_2$ OH$^-$ (in **X–Ph-NH$_2$(OH$^-$)(W)$_{2,3}$**) HBCPs have negative and positive values, respectively, indicating partially covalent nature of H-bonds in both cases. The presence of an electron-accepting substituent in the *para* position increases the covalent nature of H-bonds in **X–Ph-NH$_2$(OH$^-$)(W)$_{2,3}$** complexes and decreases **in X–An-(W)$_{1,2}$** ones. This situation for electron-donating substituents is opposite. The relationship between H(r) at C–N BCP and their values at NH$^-$ HOH and NH$_2$ OH$^-$ HBCPs are depicted in Fig. 6. As can be seen, this dependence has a negative slope for **X–An-(W)$_{1,2}$** complexes and a positive slope for **X–Ph-NH$_2$(OH$^-$)(W)$_{2,3}$** ones.

The ρ(r) values (ρ(r)>0.04 au) at main HBCPs are greater than those determined for typical HBs (0.002 to 0.04 au) [20]. Thus, they could be considered as the strong hydrogen bonds. From H(r) and $\nabla^2\rho(r)$ values given in Table 3, all H-bonds are partially covalent in nature.

There are correlations between BE and ρ(r) at $NH^-$ HOH and $NH_2$ $OH^-$ HBCPs as well as the sum of ρ(r) at all HBCPs. These relationships are shown in Fig. 7. As can be seen, in all cases (regardless of the nature of X) BE increases as the ρ(r) increases. However, BEs for electron-donating substituents are greater than those of electron-accepting ones. An increase of BE (as its absolute value) for electron-donating substituents is associated with an immobilization of the lone pair of nitrogen toward the ring because it becomes a partner for H-bond formation.

For the $NH^-$ HO interactions, ρ(r) at the HBCPs and BE of corresponding complexes increases as the number of $H_2O$ molecule changes from one to two. In the case of $NH^-$ HOH interactions, increasing of ρ(r) at HBCP is associated with an increase of O–H distance (see Table 1) and consequently with an increase of the acidity of $H_2O$. Therefore, acidity of $H_2O$ in **X–An-(W)$_2$** is greater than **X–An-(W)$_1$**.

As mentioned before, proton transfer occurs from water to anilide when the size of water cluster increases from two to three and four. Therefore, $OH^-$ is better solvated than anilide ion as the size of water cluster increases. Relationship between ρ(r) at $NH_2$ $OH^-$ HBCPs and size of water cluster is opposite of $NH^-$ HOH; that is, ρ(r) at $NH_2$ $OH^-$ HBCPs and BE of corresponding complexes decreases on going from **X–Ph-NH$_2$(OH$^-$)(W)$_2$** to **X–Ph-NH$_2$(OH$^-$)(W)$_3$** complexes. As a result, acidity of aniline decreases on going from former to later, in agreement with the greater value of N–H bond distance in former with respect to later (see Table 1). In other words, $OH^-$ is better solvated in **X–Ph-NH$_2$(OH$^-$)(W)$_3$** complexes than the **X–Ph-NH$_2$(OH$^-$)(W)$_2$** ones. As can be seen, BEs for electron-accepting substituents are greater than electron-donating ones because the increase in mobility of lone pair electrons when X is an electron-accepting group.

Correlation between BE and ρ(r) at C–N BCP is shown in Fig. 7c. For $NH^-$ HO interactions in **X–An-(W)$_{1,2}$** com-

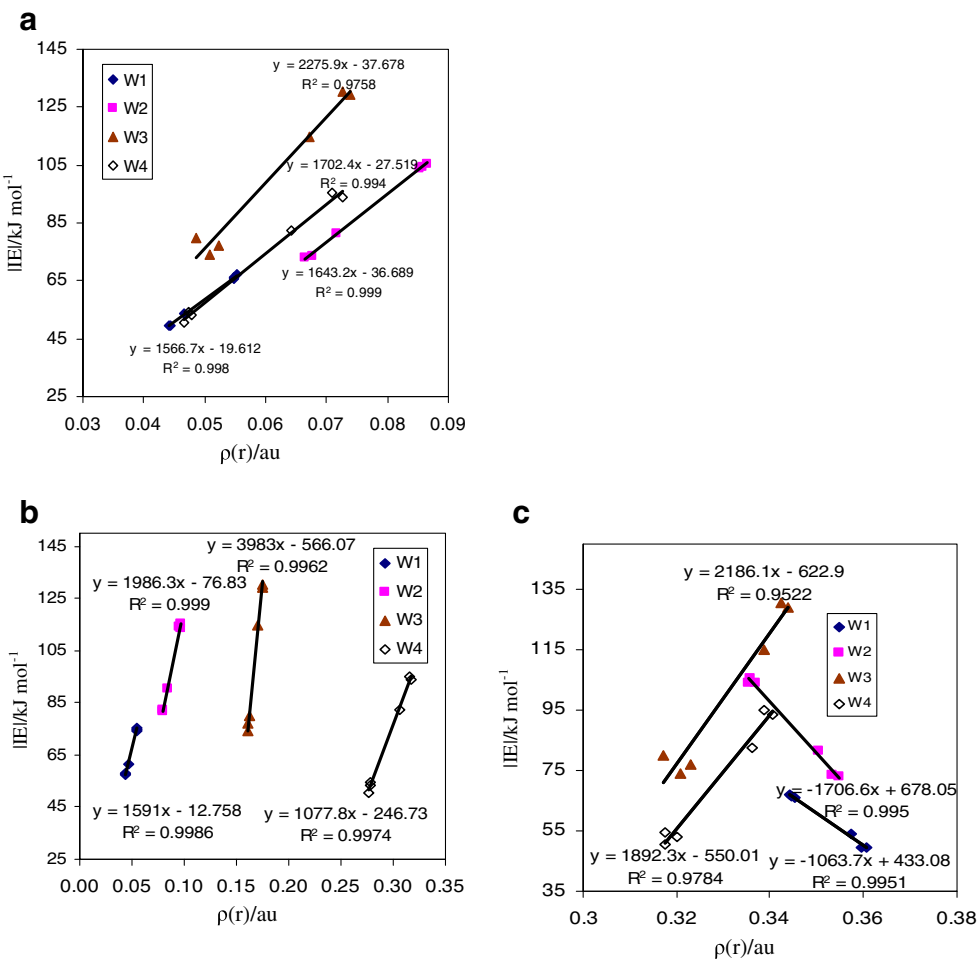



**Fig. 7** Correlation between the absolute value of interaction energy |IE| and (**a**) ρ(r) at Y Z HBCPs (Y=$NH^-$ or $NH_2$ and Z=$H_2O$ or $OH^-$), (**b**) sum of ρ(r) at all HBCPs and (**c**) ρ(r) at C-N BCPs

plexes, which $H_2O$ is a proton-donor, dependence has a negative slope and for $NH_2$ $OH^-$ interactions in **X–Ph-$NH_2(OH^-)(W)_{2,3}$** complexes, which $NH_2$ is a proton-donor, dependence has a positive slope. Formation of the NH $OH^-$ interaction is accompanied by an increase in ρ(r) at the C–N BCP. The electron-accepting and electron-donating substituents increase and decrease the H-bond strength, respectively. The situation for $NH^-$ HO interactions is reverse. As can be seen in Table 3, in contrast with the C–N bond, ρ(r) of ipso C1–C2 bond decreases when the *para* position is substituted by electron-accepting groups.

## NBO analysis

The results of NBO analysis including charge transfer (CT), natural charge and occupancy of NBOs at B3LYP/6-311++G (2 d,2p) level of theory are given in Table 4.

Charge transfer, CT, term corresponds to the shift of electron charge between interacting subunits. Here, the charge transfer has been defined as the sum of atomic charges on the proton donor subunit. Population analysis shows that the charge transfer taking place from acceptor proton subunit to donor proton one. The values of charge transfer from anilide ion to water cluster in **X–An-$(W)_2$** complexes are greater than **X–An-$(W)_1$** ones. In addition, their values increase as the substituents change from electron-accepting to electron-donating. The electron-accepting substituents cause the electron delocalization of lone pairs of N atom in the ring and then decrease in CT from anilide ion to water cluster. A reverse trend is observed for the **X–Ph-$NH_2(OH^-)(W)_{2,3}$** complexes. As can be seen, charge transfer values for **X–Ph-$NH_2(OH^-)(W)_3$** complexes are smaller than **X–Ph-$NH_2(OH^-)(W)_2$** ones, indicating that the intermolecular $NH_2$ $OH^-$ H-bonding interactions in **X–Ph-$NH_2(OH^-)(W)_2$** complexes are stronger than **X–Ph-$NH_2(OH^-)(W)_3$** ones, in agreement with the lower H-bond distances found in **X–Ph-$NH_2(OH^-)(W)_2$** complexes. Besides, CT values for complexes including electron-accepting substituents are greater than electron-donating ones. There is a correlation between BE and CT values. As can be seen in Fig. 8, the greater the BE values complexes are, the greater the CT ones.

Increase in the size of water cluster and change in substituents in the *para* position can influence the electron density of N atom. Correlation between natural charge of N atom and BE is depicted in Fig. 8b. For **X–An-$(W)_{1,2}$** complexes, increase in BE is accompanied with a decrease of electronic charge on the nitrogen atom. The stronger the electron-attracting power of the substituent is, the lesser the electron density at the N atom, and thus the weaker interaction with a water molecule. This situation is reverse for **X–Ph-$NH_2(OH^-)(W)_{2,3}$** complexes. The stronger electron-attracting

**Table 4** NBO data calculated at B3LYP/6-311++G(2 d,2p) level of theory

| | X-An-$(W)_1$ | | | | | | X-An-$(W)_2$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | X=H | Me | OMe | CHO | NO | $NO_2$ | X=H | Me | OMe | CHO | NO | $NO_2$ |
| charge | | | | | | | | | | | | |
| N7 | −0.9205 | −0.9266 | −0.9246 | −0.8456 | −0.8115 | −0.8240 | −0.9668 | −0.9715 | −0.9724 | −0.8981 | −0.8658 | −0.8792 |
| H15 | 0.4992 | 0.4990 | 0.4991 | 0.4998 | 0.4994 | 0.4994 | 0.4950 | 0.4950 | 0.4949 | 0.4964 | 0.4965 | 0.4963 |
| O14 | −1.0167 | −1.0169 | −1.0164 | −1.0017 | −0.9968 | −0.9974 | −1.0118 | −1.0122 | −1.0113 | −0.9998 | −0.9961 | −0.9963 |
| CT (A→D)[a] | 0.0869 | 0.0876 | 0.0867 | 0.0603 | 0.0603 | 0.0610 | 0.1201 | 0.1226 | 0.1207 | 0.0881 | 0.0773 | 0.0792 |
| %p-character of N | 60.01 | 60.16 | 60.27 | 59.02 | 58.80 | 58.95 | 61.07 | 61.24 | 61.23 | 59.84 | 59.47 | 59.69 |
| **X-Ph-$NH_2(OH^-)(W)_2$** | | | | | | | **X-Ph-$NH_2(OH^-)(W)_3$** | | | | | |
| charge | | | | | | | | | | | | |
| N7 | −0.8104 | −0.8154 | −0.8225 | −0.7595 | −0.7372 | −0.7451 | −0.7595 | −0.7504 | −0.7689 | −0.8213 | −0.8197 | −0.8153 |
| H15 | 0.4502 | 0.4493 | 0.4486 | 0.4545 | 0.4541 | 0.4527 | 0.4530 | 0.4542 | 0.4540 | 0.4468 | 0.4466 | 0.4482 |
| O14 | −1.2064 | −1.2072 | −1.2075 | −1.1957 | −1.1900 | −1.1944 | −1.1883 | −1.1848 | −1.1910 | −1.1989 | −1.1989 | −1.1980 |
| CT (A→D)[a] | 0.0795 | 0.0771 | 0.0773 | 0.1074 | 0.1199 | 0.1161 | 0.0719 | 0.0696 | 0.0714 | 0.1015 | 0.1173 | 0.1133 |
| %s-character of N | 62.09 | 62.39 | 62.65 | 60.45 | 60.10 | 60.28 | 62.15 | 62.50 | 62.63 | 60.55 | 60.38 | 60.58 |

[a] D=proton donor and A=proton acceptor

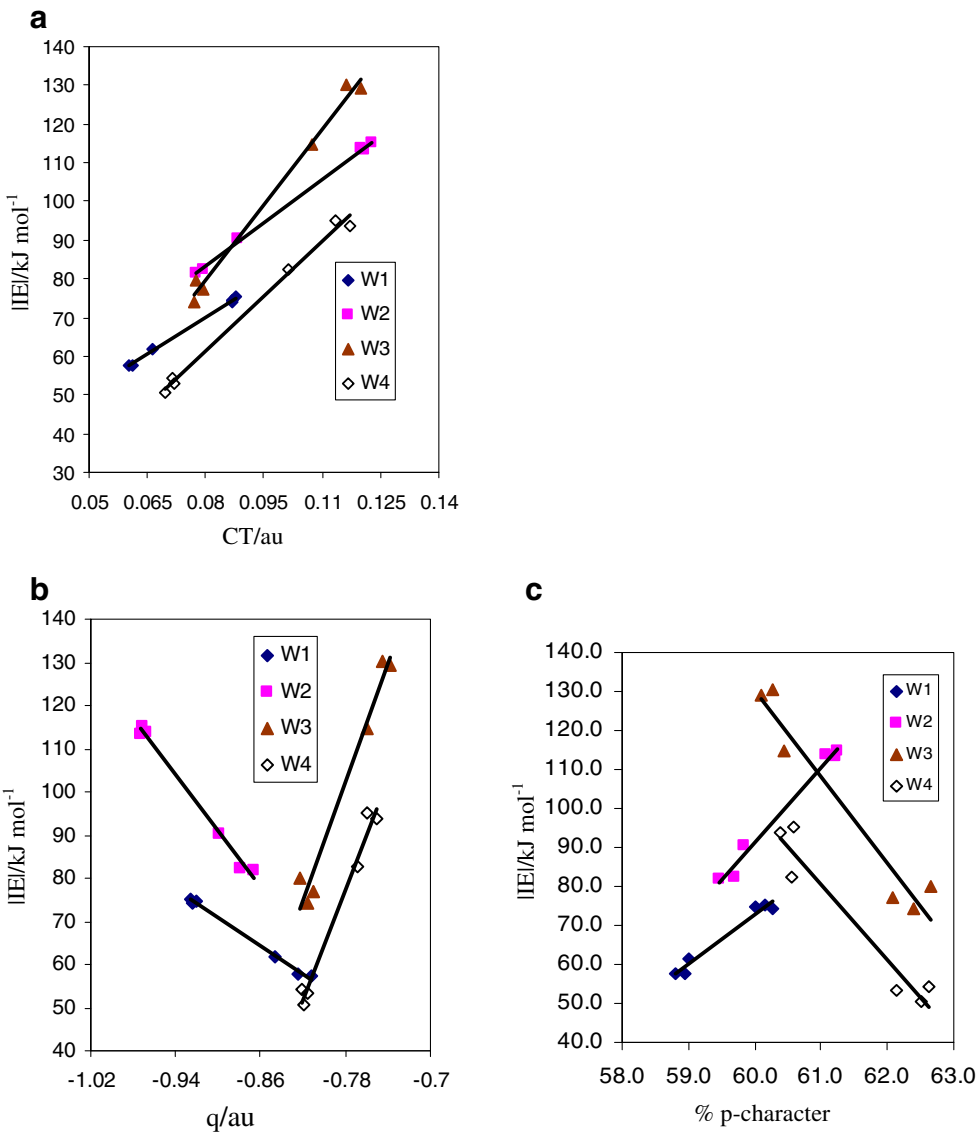



**Fig. 8** Correlation between the absolute value of interaction energy |IE| and (**a**) CT values, (**b**) electronic charge (q) of N atom and (**c**) % p-character of N atom in C–N bond calculated at B3LYP/6-311++G(2 d,2p) level

power of the substituent is, the stronger the H-bond interaction, and thus the lower the electron density at the N atom. There is a relationship between the electronic charge of N atom and the C–N bond distance. This correlation is displayed in Fig. 9a. The strength of the C–N bond increases with a decrease of electron density on N atom.

For the examined complexes, a good correlation was found between the percentage of the p-character of N atom in C–N bond and the corresponding bond distance (Fig. 9b) as well as BE (Fig. 8c). Increase in C–N bond distance is accompanied by an increase in percentage of the p-character of N atom. For electron-accepting substituents, percentage of the p-character is smaller than electron-donating ones. In addition, percentage of the p-character increases on going from **X–An-(W)$_1$** to **X–An-(W)$_2$**. The strength of the CN bond decreases (its shortening is observed) with an increase of the H-bonding strength (i.e., BE), which may be expressed by an increase of p-character. This situation is reverse for **X–Ph-NH$_2$(OH$^-$)(W)$_{2,3}$** complexes (Fig. 8c).

## Conclusions

The effects of the $H_2O$ cluster size and the substituent (X=H, Me, OMe, CHO, NO and $NO_2$) in the *para* position of anilide ion on the proton transfer between the anilide anion and water clusters via intermolecular H-bonding

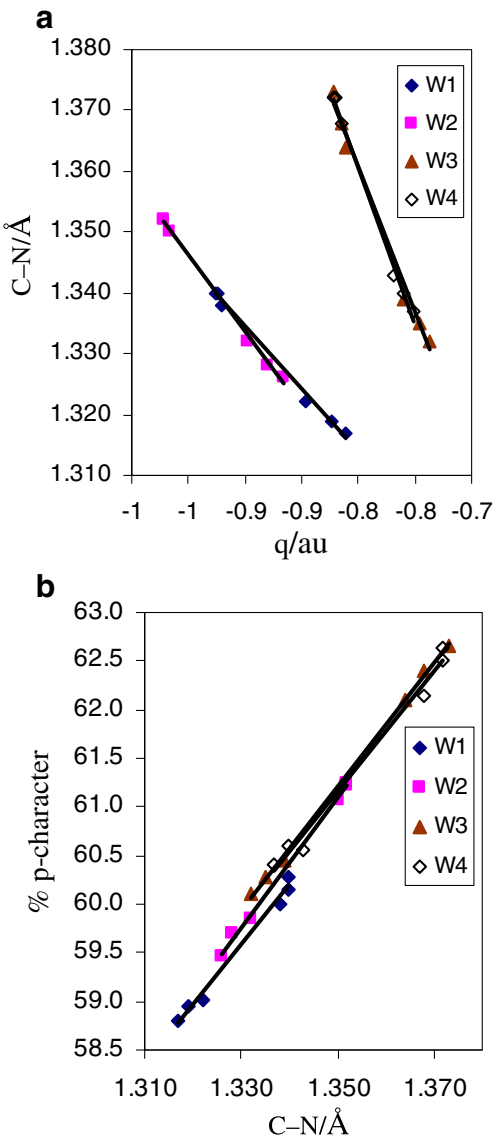



**Fig. 9** Correlation between the C–N bond distance and (**a**) electronic charge (q) of N atom, (**b**) % p-character of N atom in C–N bond calculated at B3LYP/6-311++G(2 d,2p) level

interaction were investigated in the gas phase. The change in H-bond strength due to variation of the substituents and $H_2O$ cluster size were well monitored by change in interaction energy, structural parameter, electron density topology, natural charge, charge transfer, and percentage of p-character of N atom in C–N bond. The strength of $HN^-$ HO H-bonds increases by insertion of electro-donating substituents in the *para* position of anilide ion and using increase in the size of water cluster. Therefore, $HN^-\cdots H–OH \rightarrow HN–H\cdots OH^-$ switching enhances by an increase in the extent of solvation and by the electron-donating substituents. The critical size for proton transfer from water cluster $\mathbf{(W)_n}$ to anilide anion is when n>2. The topological properties of electron density at H-bond critical points demonstrate that all H-bonds involved in complexes are partially covalent in nature.

## References


1. Cha Y, Murray CJ, Klinman JP (1989) Science 243:1325–1330
2. Paddock ML, Rongey SH, Feher G, Okamura MY (1989) Proc Natl Acad Sci USA 86:6602–6606
3. Bell RL, Truong TN (1994) J Chem Phys 101:10442–10451
4. Kim TG, Lee SI, Jiang DJ, Kim YI (1995) J Phys Chem 99:12698–12700
5. Podolyan Y, Gorb L, Leszczynski J (2002) J Phys Chem A 106:12103–12109
6. Hu X, Li H, Liang W, Han S (2004) J Phys Chem B 108:12999–13007
7. Kim Y (1998) J Phys Chem A 102:3025–3036
8. Ahn DS, Lee S, Kim B (2004) Chem Phys Lett 390:384–388
9. Meot-Ner (Mautner) M (2005) Chem Rev 105:213–284, and references cited therein
10. Szatyłowicz H, Krygowski TM, Zachara-Horeglad JE (2007) J Chem Inf Model 47:875–886
11. Szatyłowicz H, Krygowski TM, Panek JJ, Jezierska A (2008) J Phys Chem A 112:9895–9905
12. Szatyłowicz H, Krygowski TM, Jezierska A, Panek JJ (2009) J Phys Chem A 113:5800–5805
13. Wolf I, Shapira A, Giniger R, Miller Y, Gerber RB, Cheshnovsky O (2008) Angew Chem Int Ed 47:6272–6274
14. Boys SF, Bernardi F (1970) Mol Phys 19:553–566
15. Granovsky AA, PC GAMESS Version 7.1. <http://classic.chem.msu.su/gran/gamess/index.html>
16. Biegler-König FF, Schönbohm J, Bayles D (2001) AIM2000-A program to analyze and visualize atoms in molecules. J Comp Chem 22:545–559
17. Reed AE, Curtiss LA, Weinhold F (1988) Chem Rev 88:899–926
18. Glendening DE, Reed AE, Carpenter JE, Weinhold F, NBO, Version 3.1
19. Krygowski TM, Szatyłowicz H (2006) J Phys Chem A 110:7232–7236
20. Koch U, Popelier PLA (1995) J Phys Chem 99:9747–9754
21. Popelier PLA (1998) J Phys Chem A 102:1873–1878

J Mol Model (2012) 18:1325–1331
DOI 10.1007/s00894-011-1164-7

ORIGINAL PAPER

# Molecular design of new nitramine explosives: 1,3,5,7-tetranitro-8-(nitromethyl)-4-imidazolino[4,5-b] 4-imidazolino-[4,5-e] pyridine and its N-oxide

**Hui Liu · Hongchen Du · Guixiang Wang · Yan Liu · Xuedong Gong**

Received: 3 May 2011 / Accepted: 24 June 2011 / Published online: 12 July 2011


**Abstract** Two new nitramine compounds containing pyridine, 1,3,5,7-tetranitro-8-(nitromethyl) -4-imidazolino[4,5-b]4-imidazolino-[4,5-e]pyridine and its N-oxide 1,3,5,7-tetranitro-8- (nitromethyl)-4-imidazolino[4,5-b]4-imidazolino-[4,5-e]pyridine-4-ol were proposed. Density functional theory (DFT) has been employed to study the molecular geometries, electronic structures, infrared spectra, and thermodynamic properties at the B3LYP/6-31G* level. Their detonation performances evaluated using the Kamlet-Jacobs equations with the calculated densities and heats of formation are superior to those of HMX. The predicted densities of them were ca. 2 g*cm$^{-3}$, detonation velocities were over 9 km*s$^{-1}$, and detonation pressures were about 40 GPa, showing that they may be potential candidates of high energy density materials (HEDMs). The natural bond orbital analysis indicated that N-$NO_2$ bond is the trigger bond during thermolysis process. The stability of the title compounds is slightly lower than that of 2,4,6,8,10,12-hexanitro-2,4,6,8,10,12-hexaazaisowurtzitane (CL-20). The results of this study may provide basic information for the molecular design of new HEDMs.

**Keywords** Density functional theory · Detonation performance · Nitramine compound · Pyridine

H. Liu (✉) · H. Du · G. Wang · Y. Liu · X. Gong
Department of Chemistry,
Nanjing University of Science and Technology,
Nanjing 210094, China
e-mail: lhgbx@yahoo.com.cn

X. Gong
e-mail: googxd325@mail.njust.edu.cn

## Introduction

There is an ongoing need for powerful but insensitive high energetic materials. A specific goal [1] is to synthesize the materials that have performance better than that of 1,3,5,7-tetranitro-1,3,5,7-tetraazacyclooctane (HMX) and stability close to or higher than that of 1,3,5-triamino-2,4,6-trinitrobenzene (TATB). Heterocycles generally have higher heat of formation and density and better oxygen balance than their carbocyclic analogues [2]. So the synthesis of energetic heterocyclic compounds has received a great amount of interest in recent years [3–9].

Pyridine is a six-membered heterocyclic compound with one nitrogen atom. Compared to carbocyclic analogues, the derivatives of pyridine have higher nitrogen contents, enabling these compounds to release more nitrogen gases in explosion. On the other hand, since the presence of N atom in the aromatic ring can increase the density effectively, e.g., the density of benzene is only 0.897 g·cm$^{-3}$, but that of pyridine is 0.982 g·cm$^{-3}$, the presence of pyridine in molecule is supposed to improve its density and correspondingly increases the detonation velocity and pressure [10]. Therefore, the search for new potential high-energy density materials containing pyridine has attracted many attentions [11–21]. In addition, studies show that introducing nitramine group (N-$NO_2$) into a molecule can improve its detonation performance, for example, there are 3, 4, and 6 N-$NO_2$ groups in hexahydro-1,3,5-Trinitro-1,3,5-triazine (RDX), HMX, and 2,4,6,8,10,12-hexanitro-2,4,6,8,10,12-hexaazaisowurtzitane (CL-20) respectively, and their deto- nation performance increase in the same order. In this study, two new nitramine molecules containing pyridine ring, i.e., 1,3,5,7-tetrani-

tro-8-(nitromethyl)-4-imidazolino[4,5-b]4-imidazolino-[4,5-e]pyridine(I) and its N-oxide 1,3,5,7-tetranitro-8-(nitromethyl)-4-imidazolino[4,5-b]4-imidazolino-[4,5-e] pyridine-4-ol(II) were proposed to benefit from the high density characteristics of pyridine and high energy characteristics of nitramine. The molecular structures are shown in Fig. 1. To the best of our knowledge, these kinds of compounds have not been synthesized and reported to date, though their possible synthetic routines have been suggested [22]. Theoretical studies have not been performed yet. We believe it is of great meaning to carry out a systematic theoretical investigation before synthesis on their geometrical structures and various properties, such as heat of formation, infrared spectrum, density, detonation properties and sensitivity, which will be helpful for understanding the candidate to be synthesized.

## Computational methods

The density functional theory (DFT)-B3LYP method [23, 24] in combination with the 6-31G* [25] basis set which has been proved [26–32] to be able to give quite reliable energies, molecular structures, and other properties was used to fully optimize the molecular geometries of I and II with the Gaussian03 program package [33]. To characterize the nature of the stationary points and to determine the zero-point vibrational energies, harmonic vibrational analyses were performed subsequently on each optimized structure at the same level. According to the previous studies, the computed harmonic vibrational frequencies were scaled uniformly by 0.96 to take into account the systematic overestimation of B3LYP/6-31G* calculations [34]. Based on the principle of statistical thermodynamics [35], heat capacity ($C_{p,m}^{\theta}$), entropy ($S_m^{\theta}$), and thermal enthalpy ($H_m^{\theta}$) ranging from 200 to 800 K were derived.

Detonation velocity and pressure, the most important parameters for evaluating the detonation characteristics of energetic materials, were calculated using the Kamlet-Jacbos(K-J) equations [36, 37], which have been verified by many studies [26–32] to be suitable for organic explosives. The K-J equations are shown as follows:

$$D = 1.01\left(N\overline{M}^{1/2}Q^{1/2}\right)^{1/2}(1 + 1.30\rho) \quad (1)$$

$$P = 1.558\rho^2 N\overline{M}^{1/2}Q^{1/2} \quad (2)$$

where P is the detonation pressure (GPa), D is the detonation velocity (km/s), $\rho$ is the packed density (g/cm$^3$), N is the moles of gas products per gram of explosives, M is the average molar weight of detonation products, and Q is the chemical energy of detonation (kJ/g). N, M, and Q are determined according to the largest exothermic principle [32].

Since Q and $\rho$ are unknown for the unsynthesized hypothetical compounds, to predict their detonation properties, the modified K-J equations based on the calculation results of quantum chemistry were recommended [23, 24, 26, 38–43], in which the density of the explosive $\rho$ was replaced by the crystal theoretical density ($\rho_c$) computed by the following equation:

$$\rho_c = \frac{M}{V(0.001)}. \quad (3)$$

In which M is the molecular mass (g/molecule), and V (0.001) is the volume (cm$^3$/molecule) defined as the space inside a contour of electron density of 0.001e · Bohr$^{-3}$. However, the results obtained using this equation may have quite big errors for some systems, for example, molecules that can form strong hydrogen bonds. Politzer et al. [44] suggested that Eq. 3 should be corrected to better reflect the effects of intermolecular interactions in crystals. They have proposed two modified equations and the better one is shown as follows [44]:

$$\rho_p = \alpha\left[\frac{M}{V(0.001)}\right] + \beta\left(v\sigma_{tot}^2\right) + \gamma \quad (4)$$

where $\alpha$, $\beta$ and $\gamma$ are regression coefficients and their values are taken from ref. [44].

In this paper, both (3) and (4) have been used to predict the density of the title compounds for comparison.

The chemical energy of the detonation reaction Q was calculated as the difference between the heats of formation (HOFs) of products and reactants. In this paper, the HOFs of the title compounds were calculated with the help of the following reactions at the B3LYP/6-31G* level:

$$7c(s) + 2H_2(g) + 5O_2(g) + 5N_2(g) \rightarrow C_7H_4O_{10}N_{10} \text{ (I)}$$
$$7C(s) + 2H_2(g) + 5.5O_2(g) + 5N_2(g) \rightarrow C_7H_4O_{11}N_{10} \text{ (II)}$$

With the calculated enthalpies of all species and the experimental sublimation enthalpy of graphite, it is easy to obtain the HOFs of the title compounds.

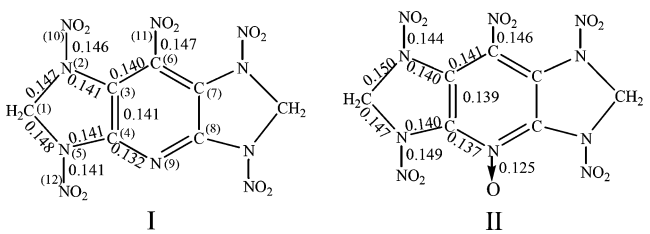



**Fig. 1** The molecular structures of the designed nitramines containing pyridine ring (bond lengths are in nm)

It should be noted, however, Q obtained in this way is the detonation enthalpy of the compound in gas phase, which is somewhat higher than that in solid phase used in K-J equations. The detonation velocity and pressure for the title compounds in solid state should be a little lower.

## Results and discussion

### Molecular geometry

With regard to the new designed compound, it is especially necessary to examine the geometric structures before discussing the other properties. Some bond lengths obtained from the B3LYP/6-31G*calculations are shown in Fig. 1.

According to the calculation results, the pyridine ring and the imidazole rings are essentially but not completely planar, e.g., C(4)-C(3)-C(6)-C(7) and C(1)-N(2)-C(3)-C(4) in I are −5.3° and −12.3°, respectively, and corresponding values are −11.1° and 5.8° for II. Dihedral angles about C (6)-N(11) in I are about 5° and in II are about 10°, which indicates the nitro groups linked with pyridyl slightly depart the plane of pyridine because of the steric hindrance effect.

The corresponding bond lengths of the two compounds are quite close to each other. Some discrepancies are obviously raised by the formation of N→O coordination bond in II. Because of the conjugation effect, the bond lengths of C-C (0.139~0.141 nm) in the pyridine ring are between the usual lengths of C=C double bond (0.134 nm) and C-C single bond (0.154 nm). The same is true for the C-N bonds. In addition, the C-N bond in the pyridyl unit of II (0.137 nm) is a little longer than that of I (0.132 nm) because of the electron withdrawing inductive effect of the coordination O atom. The length of N→O bond is 0.125 nm. The lengths of these bonds are not very different from those in pyridine and N-oxide of pyridine. In pyridine, the experimental lengths of C-C and C-N bonds are 0.139~0.140 nm and 0.134 nm [45], respectively. The corresponding results at the B3LYP/6-31G* level are completely the same. For the N-oxide of pyridine, the experimental determined bond lengths are 0.138~0.139 nm for C-C, 0.138 nm for C-N, and 0.129 nm for N→O bond [46]. The corresponding calculated results are 0.138~0.140 nm, 0.137 nm, and 0.127 nm. These results demonstrate once again that the method used in this study is reliable. The lengths of all N-$NO_2$ bonds in the title compounds are longer than the usual N-N bond lengths (0.135-0.140 nm) in nitramines [47]. The N-$NO_2$ bonds on the side of the N atom of pyridine in I (N(5)-N(12)) are shorter than those on the other side (N(2)-N(10)). However, this is not the case for II, in which the N-$NO_2$ on the side of pyridyl N (0.149 nm) is longer than others (0.144 and 0.146 nm). Obviously, this change is caused by the introduction of the coordination O atom.

### IR spectrum and thermodynamic property

It is well known that infrared spectrum (IR) is the basic property of a compound and an effective measure to analyze or identify the compound. It is also directly related with the thermodynamic properties. The simulated IR spectra of I and II based on the scaled harmonic vibrational frequencies are shown in Fig. 2. As there are no corresponding experimental results, no comparison and evaluation of the calculated results can be made. However, many previous studies [31, 38, 48] have shown that the calculated IR are usually in reasonable agreement with the experimental data. Our results may be used as reference data for future investigations of the title compounds.

For the complexity of the vibrational modes, it is difficult to assign all bands, so only some typical vibrational modes are analyzed. In the range of 1200–1300 $cm^{-1}$, the characteristic symmetrical stretching vibrations of N=O in the nitro groups and out of plane wagging of C-H have the highest intensities. They appear at 1240 $cm^{-1}$ and 1270 $cm^{-1}$ in I and II, respectively, not much different from that in CL-20 and HMX (1260 $cm^{-1}$ and 1250 $cm^{-1}$, respectively). The next strongest absorption (1650 $cm^{-1}$ in I and 1660 $cm^{-1}$ in II) is raised by the asymmetrical stretching vibrations of nitro N=O and is also close to that in CL-20 and HMX (1650 $cm^{-1}$ and 1630 $cm^{-1}$ respectively). These two bands are both blue-shifted from I to II. The absorption band at 1580 $cm^{-1}$ in II is resulted from the stretch of N→O coordination bond. It is blue-shifted compared with the N-oxide of pyridine for which the calculated value is 1320 $cm^{-1}$ and the experimental value is 1265 $cm^{-1}$ [49], which is consistent with the fact that the N→O length in II is shorter and stronger. The absorptions around 3000 $cm^{-1}$ raised by the C-H symmetric and asymmetric stretching are very weak. The N-N stretching vibrations locate at the fingerprint region (800–900 $cm^{-1}$). It is red-shifted from 850 $cm^{-1}$ in I to 820 $cm^{-1}$ in II due to the coordination of O to N. The corresponding values are 920 $cm^{-1}$ and 870 $cm^{-1}$ for CL-20 and HMX. Once again, this agrees with the fact that N-N in the title compounds are longer and weaker than those in CL-20 and HMX. In the range of 1000–1200 $cm^{-1}$, there are absorptions raised by the symmetrical stretching vibrations of C-N in the imidazole ring and scissoring vibrations of C-H.

Thermodynamic functions, such as heat capacity, entropy, and enthalpy, which are important parameters and are necessary in predicting reactive properties, have been

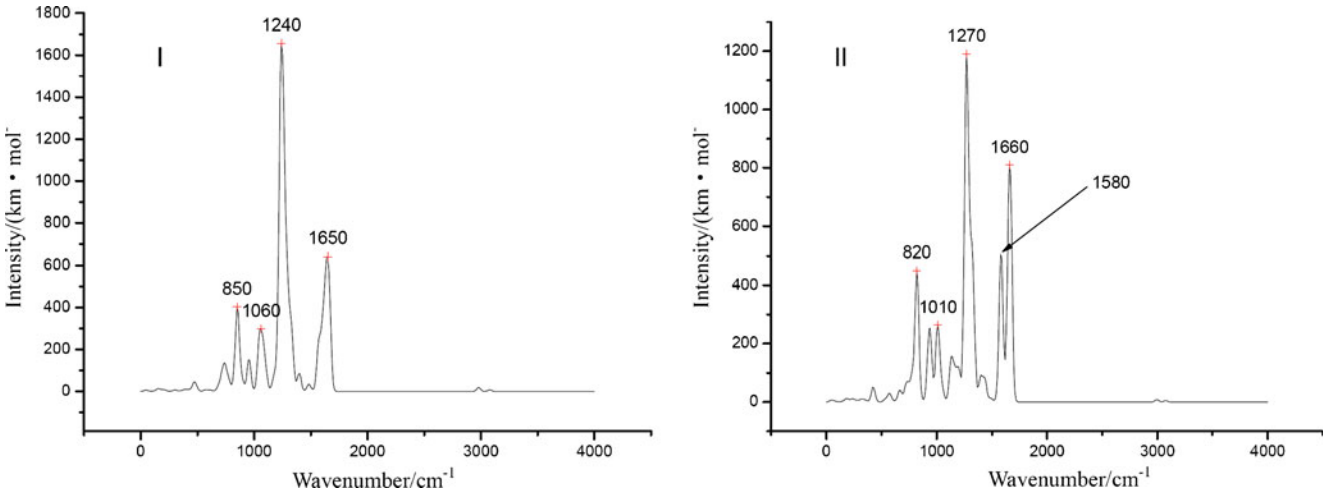



**Fig. 2** The calculated infrared spectra of I and II at the B3LYP/6-31G* level

evaluated based on the scaled vibrational frequencies and the principle of statistic thermodynamics for I and II. The results are listed in Table 1.

At room temperature, the predicted $C_{p,m}^{\theta}$, $S_m^{\theta}$, and $H_m^{\theta}$ of I are 345.95 J·mol$^{-1}$·K$^{-1}$, 644.43 J·mol$^{-1}$·K$^{-1}$, and 516.51 kJ·mol$^{-1}$ respectively and corresponding values of II are 361.35 J·mol$^{-1}$·K$^{-1}$, 640.06 J·mol$^{-1}$·K$^{-1}$, and 528.86 kJ·mol$^{-1}$. All these thermodynamic functions increase with temperature evidently. In addition, the values of $C_{p,m}^{\theta}$ and $H_m^{\theta}$ of II are all larger than that of I in the temperature range of 200~800 K. However, the values of $S_m^{\theta}$ are larger for II than for I only when the temperature is higher than 400 K. Analyzing the contribution of various movements, we see that this is mainly attributed to the contribution of vibration to entropy ($S_v$). The contribution of rotation ($S_r$) and translation ($S_t$) to entropy are always larger for II than for I, but the $S_v$ of II is smaller than that of I when the temperature is lower than 400 K, so the sum of $S_v$, $S_r$, and $S_t$ is smaller for II than for I at lower temperature. When the temperature increases, the $S_v$ of II becomes bigger than that of I, and therefore the total entropy of II is larger. The temperature-dependent relationships for $C_{p,m}^{\theta}$, $S_m^{\theta}$, and $H_m^{\theta}$ in the range of 200~800 K can be expressed as shown in Table 2 and Fig. 3 (where SD is the standard deviation, the correlation coefficients ($R^2$) are 1.000 for all equations). We think these will be helpful for the further studies on the other physical, chemical, and explosive properties of these two compounds.

The changes in energy (ΔH) and Gibbs free energy (ΔG) have been calculated for the oxidation reactions of I and pyridine, i.e., I + O → II, py + O → pyo, where py and pyo stand for pyridine and the N-oxide of pyridine, respectively. Results show that both are exothermic and spontaneous process with the ΔH being −131.86 and −259.75 kJ·mol$^{-1}$, and ΔG being −141.83 and −219.75 kJ·mol$^{-1}$, respectively. Since the oxidation from py to pyo has been achieved [50], the oxidation of I to II should also be feasible though it may be more difficult.

## Detonation performance

To find a new HEDM, it is important to evaluate its detonation performance. Table 3 collects the predicted ρ, D, and P of the title compounds. For comparison, the calculated and experimental results of the famous explosives HMX [51] and CL-20 [52] are also listed. From

**Table 1** Thermodynamic functions of the title compounds at different temperatures

| | T/K | 200 | 298.15 | 300 | 400 | 500 | 600 | 700 | 800 |
|---|---|---|---|---|---|---|---|---|---|
| I | $C_{p,m}^{\theta}$/ J·mol$^{-1}$·K$^{-1}$ | 258.87 | 345.95 | 347.51 | 425.01 | 487.83 | 536.82 | 574.72 | 604.23 |
| | $S_m^{\theta}$/ J·mol$^{-1}$·K$^{-1}$ | 524.59 | 644.43 | 646.57 | 757.49 | 859.35 | 952.81 | 1038.53 | 1117.28 |
| | $H_m^{\theta}$/ kJ·mol$^{-1}$ | 486.77 | 516.51 | 517.15 | 555.90 | 601.66 | 653.00 | 708.66 | 767.66 |
| II | $C_{p,m}^{\theta}$/ J·mol$^{-1}$·K$^{-1}$ | 269.68 | 361.35 | 362.97 | 443.20 | 507.72 | 557.87 | 596.57 | 626.66 |
| | $S_m^{\theta}$/ J·mol$^{-1}$·K$^{-1}$ | 514.97 | 640.06 | 642.30 | 758.07 | 864.18 | 961.38 | 1050.41 | 1132.12 |
| | $H_m^{\theta}$/ kJ·mol$^{-1}$ | 497.81 | 528.86 | 529.53 | 569.96 | 617.64 | 671.02 | 728.83 | 790.05 |

**Table 2** The correlation equations between the thermodynamic functions and temperature for the title compounds (the units of $C_{p,m}^{\theta}$, $S_m^{\theta}$, and $H_m^{\theta}$ are J·mol$^{-1}$·K$^{-1}$, J·mol$^{-1}$·K$^{-1}$, and kJ·mol$^{-1}$, respectively)

| I | II |
|---|---|
| $C_{p,m}^{\theta} = 47.43 + 1.25T - 6.58 \times 10^{-4}T^2$ SD = 1.959 | $C_{p,m}^{\theta} = 45.69 + 1.19T - 6.21 \times 10^{-4}T^2$ SD = 1.598 |
| $S_m^{\theta} = 260.75 + 1.41T - 4.27 \times 10^{-4}T^2$ SD = 1.106 | $S_m^{\theta} = 239.12 + 1.48T - 4.52 \times 10^{-4}T^2$ SD = 1.159 |
| $H_m^{\theta} = 436.47 + 0.19T - 2.84 \times 10^{-4}T^2$ SD = 1.575 | $H_m^{\theta} = 444.93 + 0.20T - 2.93 \times 10^{-4}T^2$ SD = 1.659 |

Table 3 it can be seen that the densities obtained with the Eq. 4 is usually a little smaller than that obtained with the Eq. 3 and correspondingly, D and P predicted using the former are usually lower. For HMX, the densities obtained with Eqs. 3 and 4 are very close and both are smaller than the experimental value. The predicted D are a little bigger and P somewhat smaller than the corresponding experimental results. For CL-20, the density obtained from Eq. 3 is slightly larger and that from Eq. 4 is smaller than the experimental value with the former being somewhat better. The predicted D and P of CL-20 are all larger than the corresponding experiment values and those obtained using the density from Eq. 4 are better. As for the title compounds, II has better performances than I due to the improvement of oxygen balance. The detonation performances of II (D=9.26 km•s$^{-1}$, P=39.70 GPa) are superior to HMX, though slightly inferior to CL-20. According to the criteria as an HEDM, i.e., ρ≈1.9 g•cm$^{-3}$, D≈9 km•s$^{-1}$, and P≈40 GPa, compounds I and II are potential HEDMs and worth further investigations.

## Stability

Besides high energy, a good candidate of HEDM should also have a high stability. Bond overlap populations reflecting the electron accumulations in the bonding region can provide us detailed information about the strength of the bond and the thermal stability of the molecule. As a whole, the smaller the populations a bond has, the easier the bond breaks, and the lower stability the molecule has. In this paper, both Mulliken population analysis (MPA) [53], which is known to suffer from some shortcomings, especially the basis set dependence, and natural bond orbital analysis (NBO) [54] have been performed. The results are listed in Table 4.

According to the MPA results, C(6)-N(11) in I and C(3)-N(2) in II have the smallest populations, which is obviously not reasonable and not in concordance with the conclusion of theoretical and experimental studies [38, 55] that N-$NO_2$ bond is commonly weaker than C-N and C-C bonds and is prone to cleave. From the NBO analysis, N(2)-N(10) of I and N(5)-N(12) of II have the smallest populations, which is consistent with their longer lengths and is believed to be more reliable. This indicates that N(2)-N(10) of I and N(5)-N(12) of II are prone to cleave upon impact or other stimuli, and their breaking may be the initial steps in the decomposition or detonation. This is in line with the thermal decomposition mechanism of nitramine explosives such as HMX and CL-20 and seems to show that NBO analysis is more reliable than MPA, at least for these kinds

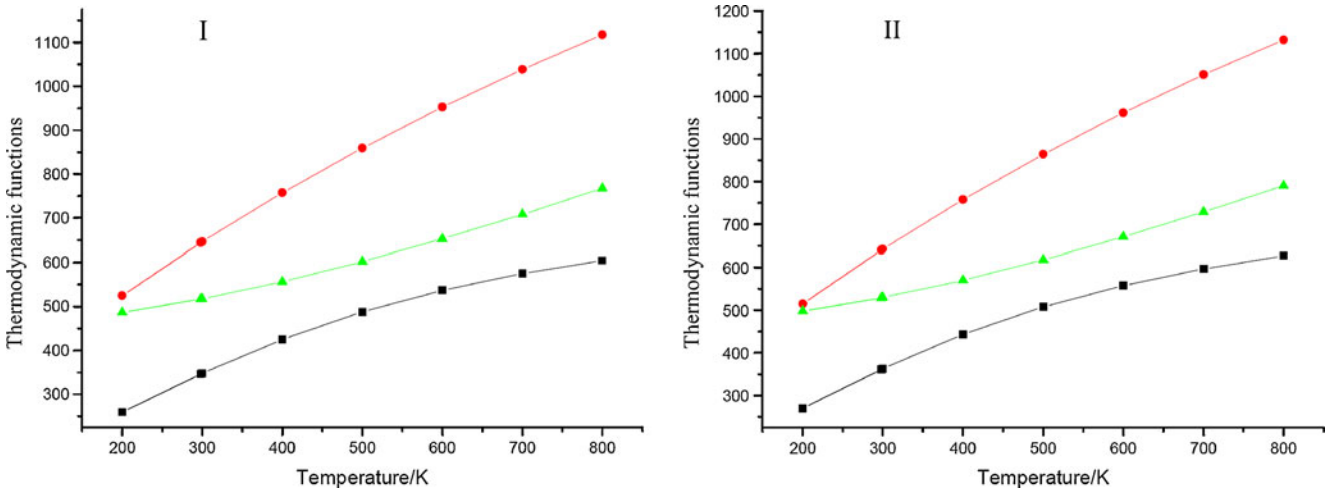



**Fig. 3** Relationships between the thermodynamic functions and temperature for I and I (■,●, and▲ stand for $C_{p,m}^{\theta}$, $S_m^{\theta}$, and $H_m^{\theta}$ and their units are J·mol$^{-1}$·K$^{-1}$, J·mol$^{-1}$·K$^{-1}$,and kJ·mol$^{-1}$, respectively)

**Table 3** Predicted detonation properties of the title compounds and two famous nitramine explosives[a]

| Compound | $\Delta H_f$/(kJ•mol$^{-1}$) | Q(kJ•g$^{-1}$) | ρ/(g•cm$^{-3}$) | D/(km•s$^{-1}$) | P/GPa |
|---|---|---|---|---|---|
| I | 620.6 | 1649.81 | 1.97(1.92) | 9.17(9.01) | 39.26(37.43) |
| II | 693 | 1743.69 | 1.98(1.94) | 9.40(9.26) | 41.42(39.70) |
| HMX | 309 | 1666.03 | 1.82(1.82)[1.90] | 9.12(9.13)[9.10] | 37.11(37.25)[39.00] |
| CL-20 | 641.9 | 1712.44 | 2.06(1.98)[2.04] | 9.90(9.62)[9.40] | 46.98(43.38)[42.00] |

[a] Data in the parentheses are the densities ($\rho_p$) predicted using Eq. 4 and the detonation parameters evaluated using $\rho_p$. Those in the brackets are the experimental values taken from Ref. [51] and [52]

of compounds. In addition, it can be seen that because of the electron-withdrawing effect of the coordination oxygen atom and the steric hindrance between O and the neighboring $NO_2$ groups, the weakest bond changes from N(2)-N(10) in I to N(5)-N(12) in II. Compared with CL-20, the strength of the N-$NO_2$ bonds and the stability of I and II are a little lower.

## Conclusions

From the DFT studies on the title compounds, the following conclusions can be drawn:

(1) In the stable structures of the title compounds obtained at the B3LYP/6-31G* level, the pyridine and imidazole ring are essentially but not completely planar. The nitro groups departure from the plane slightly. The N-$NO_2$ bonds on the side of pyridine nitrogen in II have the longest bond lengths due to the steric hindrance effect between the N→O coordination bond and the $NNO_2$ groups.

(2) The IR spectra were computed and assigned. The strongest absorption (1240 cm$^{-1}$ in I and 1270 cm$^{-1}$ in II) corresponds to the symmetrical stretching vibrations of N=O in the nitro groups and out of plane wagging of C-H. The asymmetrical stretching vibration of N=O in the nitro groups is the next strongest and is located at 1650 cm$^{-1}$ for I and 1660 cm$^{-1}$ for II. Both absorption bands are blue-shifted by the oxidation of N. The N-N stretching vibration locates at 850 cm$^{-1}$ for I and 820 cm$^{-1}$ for II and is red-shifted when a coordination O atom is introduced. The peak at 1580 cm$^{-1}$ for II corresponds to the symmetrical stretching vibration of N→O coordination bond.

(3) Heat capacity, entropy, and enthalpy have been predicted and all increase quantitatively with the increasing temperature. The correlation equations have been obtained by fitting the calculated results.

(4) Natural bond orbital analysis is more reliable than Mulliken population analysis for these kinds of compounds. The N-$NO_2$ bonds are weaker and more fragile than the other bonds. The N-$NO_2$ bonds on the imidazole rings on the side of pyridyl aza nitrogen become weaker than the other N-$NO_2$ bonds after oxidation and they will rupture initially in the thermal decomposition processes. II is less stable than I and both are less stable than CL-20.

(5) The title compounds may be novel potential candidates of HEDM according to their predicted densities (ca.2 g•cm$^{-3}$), detonation velocities (over 9 km•s$^{-1}$), and detonation pressures (about 40 GPa), which are all superior to HMX and only slightly inferior to CL-20. They are worth further investigations.

**Table 4** MPA and NBO results of populations obtained at the B3LYP/6-31G* level

| Bond | MPA | | NBO | |
|---|---|---|---|---|
| | I | II | I | II |
| C(3)-C(4) | 0.370 | 0.343 | 1.278 | 1.353 |
| C(3)-C(6) | 0.337 | 0.264 | 1.342 | 1.289 |
| C(4)-N(9) | 0.398 | 0.224 | 1.391 | 1.131 |
| C(3)-N(2) | 0.157 | 0.158 | 1.033 | 1.068 |
| C(4)-N(5) | 0.151 | 0.191 | 1.052 | 1.040 |
| C(1)-N(2) | 0.227 | 0.199 | 0.956 | 0.914 |
| C(1)-N(5) | 0.217 | 0.242 | 0.922 | 0.962 |
| C(6)-N(11) | 0.125 | 0.163 | 0.927 | 0.956 |
| N(2)-N(10) | 0.172 | 0.180 | 0.884 | 0.937 |
| N(5)-N(12) | 0.181 | 0.173 | 0.977 | 0.859 |
| N→O | | 0.245 | | 1.364 |

**Acknowledgments** Authors thank the NSAF Foundation of National Natural Science Foundation of China (Grant No. 11076017) for support of this project.

## References


1. Stine JR (1993) Mater Research Society Symposium Proc 296:3
2. Hiskey M, Chavez D (2001) Insensitive high-nitrogen compound. NTIS, No:DE220012776133
3. Hiskey M, Chavez D, Darren lN (2000) Proceedings of 27th International Pyrotechnics Seminar USA. Colorado p3-14
4. Huynh MHV, Hiskey M, Ernest L (2004) Angew Chem Int Edn 43:4924–4928
5. All AN, Son SF, Hiskey M (2004) J Propuls Power 20:120–126
6. Huang M, Li HZ (2006) Chin J Energetic Mater 14:457–462
7. Yang SHQ, Xu SL, Lei YP (2006) Chin J Energetic Mater 14:475–484
8. Zhou Y, Long XP, Wang X (2006) Chin J Energetic Mater 14:315–320
9. Li XT, Pang SP, Luo YJ (2007) Chin J Org Chem 27:1050–1059
10. Wang GX, Xiao HM (2007) Acta Chim Sinica 65:517–524
11. Ritter H, Licht HH (1995) J Heterocycl Chem 32:585–590
12. Hollins RA, Merwin LM, Nissan RA (1996) J Heterocycl Chem 33:895–904
13. Ritter H, Licht HH (1988) Propellants Explosives Pyrotechnics 13:25–29
14. Cheng J, Yao QZ (2008) Chin J Org Chem 28:1943–1947
15. Chavez DE, Hiskey MA, Gilardi RD (2000) Angew Chemie 39:1971–1973
16. Turker L, Atalar T (2006) J Hazardous Mater 137:1333–1344
17. Johnson MA, Truong TN (1999) J Phys Chem B 103:9392–9393
18. Zeman S, Trcinski WA, Matyas R (2008) J Hazardous Mater 154:192–198
19. Turker L (2009) J Energetic Mater 27:94–109
20. Turker L, Atalar T, Gümüs S, Amur YC (2009) J Hazardous Mater 167:440–448
21. Fried LE, Manaa MR, Pagoria PF, Simpson RL (2001) Annu Rev Mater Res 31:291–321
22. Lu M, Nie FD (2010) Chin J Energetic Mater 18:618–622
23. Lee C, Yang W, Parr RG (1988) Phys Rev B 37:785–789
24. Becke AD (1993) J Chem Phys 98:5648–5652
25. Hariharan PC, Pople JA (1973) Theor Chim Acta 28:213–222
26. Xiao HM, Xu XJ, Qiu L (2008) Theoretical design of high energy density materials. Science Press, Beijing
27. Xiao HM (2004) Structures and Properties of Energetic Compounds. National Defence Industry Press, Beijing
28. Chen ZX, Xiao JM, Xiao HM, Chiu YN (1999) J Phys Chem A 103:8062–8066
29. Zhang J, Xiao HM (2002) J Chem Phys 116:10674–11683
30. Xu XJ, Xiao HM, Ju XH, Gong XD, Zhu WH (2006) J Phys Chem A 110:5929–5933
31. Wang GX, Gong XD, Xiao HM (2008) Chin J Chem 26:1357–1362
32. Wang GX, Gong XD, Liu Y, Xiao HM (2010) Int J Quantum Chem 110:1691–1701
33. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Zakrzewski VG, Montgomery JA, Stratmann RE, Burant JC, Dapprich S, Millam JM, Daniels AD, Kudin KN, Strain MC, Farkas O, Tomasi J, Barone V, Cossi M, Cammi R, Mennucci B, Pomelli C, Adamo C, Clifford S, Ochterski J, Petersson GA, Ayala PY, Cui Q, Morokuma K, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Cioslowski J, Ortiz JV, Baboul AG, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Gomperts R, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Gonzalez C, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Andres JL, Head-Gordon M, Replogle ES, Pople JA (2003) Gaussian03. Gaussian Inc, Pittsburgh, PA
34. Scott AP, Radom L (1996) J Phys Chem 100:16502–16513
35. Hill TL (1960) Addision-Wesley Publishing Company New York
36. Kamlet MJ, Jacobs SJ (1968) J Chem Phys 48:23–35
37. Zhang XH, Yun ZH (1989) Explosive chemistry. National Defence Industry Press, Beijing
38. Xu XJ, Xiao HM, Gong XD, Ju XH, Chen ZX (2005) J Phys Chem A 109:11268–11274
39. Qiu L, Xiao HM, Gong XD, Ju XH, Zhu WH (2006) J Phys Chem A 110:3797–3807
40. Qiu L, Xiao HM, Ju XH, Gong XD (2005) J Quantum Chem 105:48–56
41. Xu XJ, Xiao HM, Ma XF, Ju XH (2006) J Quantum Chem 106:1561–1568
42. Xu XJ, Xiao HM, Wang GX, Ju XH (2006) Chin J Chem Phys 19:395–396
43. Xiao HM, Zhang J (2002) Sci China Ser B 45:21–29
44. Politzer P, Martinez J, Murray JS, Concha MC, Toro-Labbé A (2009) Mol Phys 107:2095–2101
45. Lide DR (ed) (2005) CRC Handbook of Chemistry and Physics, Internet Version. CRC Press, Boca Raton, FL
46. Qi Y, Chen X, Wang Y, Ming M (2005) Tianjin Chemical Industry 19:1–4
47. Archibald TG, Gilardi R, Baum K, George C (1990) J Org Chem 55:2920–2924
48. Wang GX, Gong XD, Xiao HM (2009) Spectrochim Acta Part A 74:569–574
49. Ochiai E (1967) Aromatic Amine Oxides. Elsevier, Amsterdam, p 119
50. Chiang JF (1974) J Chem Phys 61:1280–1283
51. Qiu L, Xiao HM, Zhu WH, Ju XH, Gong XD (2006) Chin J Chem 24:1538–1546
52. Ghule VD, Jadhav PM, Patil RS, Radhakrishnan S, Soman T (2010) J Phys Chem A 114:498–503
53. Mulliken RS (1955) J Chem Phys 23:1833–1840
54. Glendening ED, Reed AE, Carpenter JE, Weinhold F (1998) NBO, version 3.1. Gaussian Inc, Pittsburgh, PA
55. Qiu L, Gong XD, Xiao HM (2008) Chin J Chem 26:2165–2172

J Mol Model (2012) 18:1333–1343
DOI 10.1007/s00894-011-1154-9

ORIGINAL PAPER

# A systematic search for the structures, stabilities, and electronic properties of bimetallic $Ca_2$-doped gold clusters: comparison with pure gold clusters

**Ya-Ru Zhao · Xiao-Yu Kuang · Peng Shao · Cheng-Gang Li · Su-Juan Wang · Yan-Fang Li**

Received: 14 March 2011 / Accepted: 14 June 2011 / Published online: 7 July 2011


**Abstract** The local meta-GGA exchange correlation density functional (TPSS) with a relativistic effective core potential was employed to systematically investigate the geometric structures, stabilities, and electronic properties of bimetallic $Ca_2Au_n$ ($n$=1–9) and pure gold $Au_n$ ($n$≤11) clusters. The optimized geometries show that the most stable isomers for $Ca_2Au_n$ clusters have 3D structure when $n$>2, and that one Au atom capping the $Ca_2Au_{n-1}$ structure for different-sized $Ca_2Au_n$ ($n$=1–9) clusters is the dominant growth pattern. The average atomic binding energies and second-order difference in energies show that the $Ca_2Au_4$ isomer is the most stable among the $Ca_2Au_n$ clusters. The same pronounced even–odd alternations are found in the HOMO–LUMO gaps, VIPs, and hardnesses. The polarizabilities of the $Ca_2Au_n$ clusters show an obvious local minimum at $n$=4. Moreover, the inverse corrections to the polarizabilities versus the ionization potential and hardness were found for the gold clusters.

**Keywords** Ca-Au cluster · Geometric configuration · Density functional theory

## Introduction

Clusters containing between a few and tens of thousands of atoms show particular physical and chemical properties that are different from their bulk counterparts [1–3]. Among simple small-sized clusters, pure and doped gold clusters have aroused considerable interest due to their potential technological applications in the fields of solid-state chemistry, materials science, optics, nanotechnology, catalysis, biology, and medicine [4–20].

A great number of investigations on single impurity-doped gold clusters have been performed by many groups. For instance, Heinebrodt et al. [21] investigated bimetallic $Au_nX$ ($X$=Cu, Al, Y, In, Cs) clusters. Yuan et al. [22] found that the atoms of the impurity can markedly influence the geometric and electronic properties of the gold cluster when they investigated $Au_nM$ ($n$=1–7, $M$=Ni, Pd, Pt) clusters. Zhang et al. [23] reported $M$@$Au_6$ clusters ($M$=Sc–Ni), where the transition metal atom is located at the center of the $Au_6$ ring. The structures and electronic properties of $Au_{19}X$ clusters (X=Li, Na, K, Ru, Cs, Cu, and Ag) have been studied by Ghanty et al. [24]. However, little systemic research has been done on gold clusters doped with two types of impurity. In previous works, Li et al. [25] reported the structure and bonding of $Si_2Au_n^-$ and $Si_2Au_n$ ($n$=2 and 4) clusters; they found that the dibridged and monobridged structure are the most stable structures of $Si_2Au_2$ and $Si_2Au_4$ clusters, respectively. Doubly bridged square planar ($D_{4h}$) coinage metal clusters ($M_4Li_2$, $M$=Cu, Ag, Au) were investigated by Wannere et al. [26]. Recently, Guo et al. studied the structures and stabilities of $Au_nPd_2$ and $Au_nPt_2$ ($n$=1–4) clusters [27, 28], and their results indicated that the gold–doped atoms interaction is strong enough to enhance the cluster stability, and the larger the $Au_n$ cluster, smaller the distortions caused by the two Pd or Pt atoms. We have systematically reported investigations on small $M_2Au_n$ ($M$=Ag, Cu; $n$=1–10) clusters, comparing them with their equivalent (in terms of size) pure gold clusters [29]; we found that the ground-state

Y.-R. Zhao · X.-Y. Kuang (✉) · P. Shao · C.-G. Li · S.-J. Wang · Y.-F. Li
Institute of Atomic and Molecular Physics, Sichuan University, Chengdu 610065, China
e-mail: scu_kuang@163.com

X.-Y. Kuang
International Centre for Materials Physics, Academia Sinica, Shenyang 110016, China

structures show 3D structure at $n$=6, 9, 10 for $Ag_2Au_n$ and $n$=6, 7, 9, 10 for $Cu_2Au_n$ clusters. This reveals that the dopant atoms dramatically affect the geometries of the ground states of the equivalent $Au_n$ clusters. Also, differences in the properties of pure and double-doped gold clusters were found.

For calcium-gold clusters, Balducci et al. [30] studied the dissociation energy and bond energy of bimetallic AuCa using mass spectrometry and density functional theory. However, to the best of our knowledge, few systematic studies on the structure and stability of double calcium doped gold clusters has been carried out so far. It would be interesting to establish how the structures and properties of the pure gold clusters differ from their equivalent double-doped gold clusters? Hence, we decided that it would be useful to carry out a density functional theory investigation of the geometries, stabilities, and properties of small bimetallic $Ca_2Au_n$ ($n$=1–9) clusters and compare them with those of pure gold clusters. More importantly, to obtain more accurate results, the Tao–Perdew–Staroverov–Scuseria (TPSS) metal-GGAs functional [31] was used instead of the traditional GGA functional. Because the metal-GGA functional includes the kinetic energy density in the functional expression, it yields more accurate results for the atomization energy and the relative stabilities of competing isomers. This has been proven in lots of previous works [32–35], especially Johansson et al. [36], who reported the first metal-GGAs calculation on a gold cluster. In the current work, our optimized geometries reveal that the most stable isomers have 3D structure when $n$>2. The averaged atomic bonding energies and the second-order difference inenergies indicate that the 3D $Ca_2Au_4$ isomer is the most stable structure for $Ca_2Au_n$ clusters. The same pronounced even–odd alternations are found in the HOMO–LUMO gaps, VIPs, and hardnesses. The polarizabilities of $Ca_2Au_n$ clusters depend sensitively on the cluster size, and an obvious minimum occurs at $n$=4. Moreover, the inverse corrections to the polarizabilities versus the ionization potential and hardness were found for the gold clusters.

## Computational details

All optimizations of the $Au_{n+2}$ and $Ca_2Au_n$ ($n$=1–9) clusters were performed using the GAUSSIAN 03 program package [37] with the TPSS functional. For the Au atom, the full electron calculation is rather time-consuming, so it is better to introduce an effective core potential including relativistic effects (RECP) to describe the inner-core electrons. The basis sets (termed GENECP) were then combinations of the LANL2TZ (f) [38, 39] and 6-311 + G (d) [40] basis sets employed for the Au and Ca atoms, respectively. When searching for the lowest-energy structures, lots of possible initial structures—including one-, two- and three-dimensional configurations—were considered, starting from the previously optimized $Au_n$ and $Au_nX_m$ geometries [10–12, 36, 41–48], and all clusters were relaxed fully without any symmetry constraints. For the small clusters, the default convergence thresholds were used. When the clusters were large, the convergence thresholds of the maximum force, the root-mean-square (RMS) force, the maximum atomic displacement, and the RMS displacement were set to 0.00045, 0.0003, 0.0018, and 0.0012 a.u., respectively. In this way, a large number of optimized isomers for $Ca_2Au_n$ ($n$=1–9) clusters were obtained. Here, we only report on a few energetically low-lying isomers, but many higher energy structures are shown in the "Electronic supplementary material" (ESM).

To test the reliability of our calculations, the bond lengths, vibration frequencies, and dissociation energies of the $Au_2$ and CaAu dimers were calculated. The bond length (2.66 Å) and vibration frequency (221.2 $cm^{-1}$) for CaAu dimers and the dissociation energy (2.307 eV) for the $Au_2$ cluster are in good agreement with the experimental results (2.67 Å, 220.9 $cm^{-1}$ and 2.30 eV, respectively) [30, 49–51]. Furthermore, the bond length of $Au_2$ (2.516 Å) is close to that obtained in a previous coupled-cluster calculation (2.512 Å) by Wesendrup et al. [52], and the value observed experimentally (2.47 Å) [50]. In the following calculations, the vertical ionization potential $\left(\mathrm{VIP} = E_{\text{cation at optimized neutral geometry}} - E_{\text{optimized neutral}}\right)$, the highest occupied–lowest unoccupied molecular orbital (HOMO–LUMO) energy gap, and natural population analysis (NPA) of the most stable configurations are also evaluated based on the TPSS level.

## Results and discussion

### Bare gold clusters: $Au_n$ ($n$=3–11)

In order to compare the effects of the dopant atoms on the gold clusters, we first performed some optimization of $Au_n$ ($n$=3–11) clusters. The ground-state gold clusters for each cluster size are shown in Fig. 1. From the figure, it is clear that the structures and electronic states are in good agreement with the previous results [10, 11, 47, 48]. In addition, the average atomic binding energies, the second-order differences in energy, the HOMO–LUMO gaps, the VIPs, the EAs and the polarizabilities of the gold clusters were studied, and the results were compared with the available experimental values.

### Bimetallic calcium-gold clusters: $Ca_2Au_n$ ($n$=1–9)

For the $Ca_2Au_n$ ($n$=1–9) clusters, the ground-state isomer and a few low-lying structures for each cluster

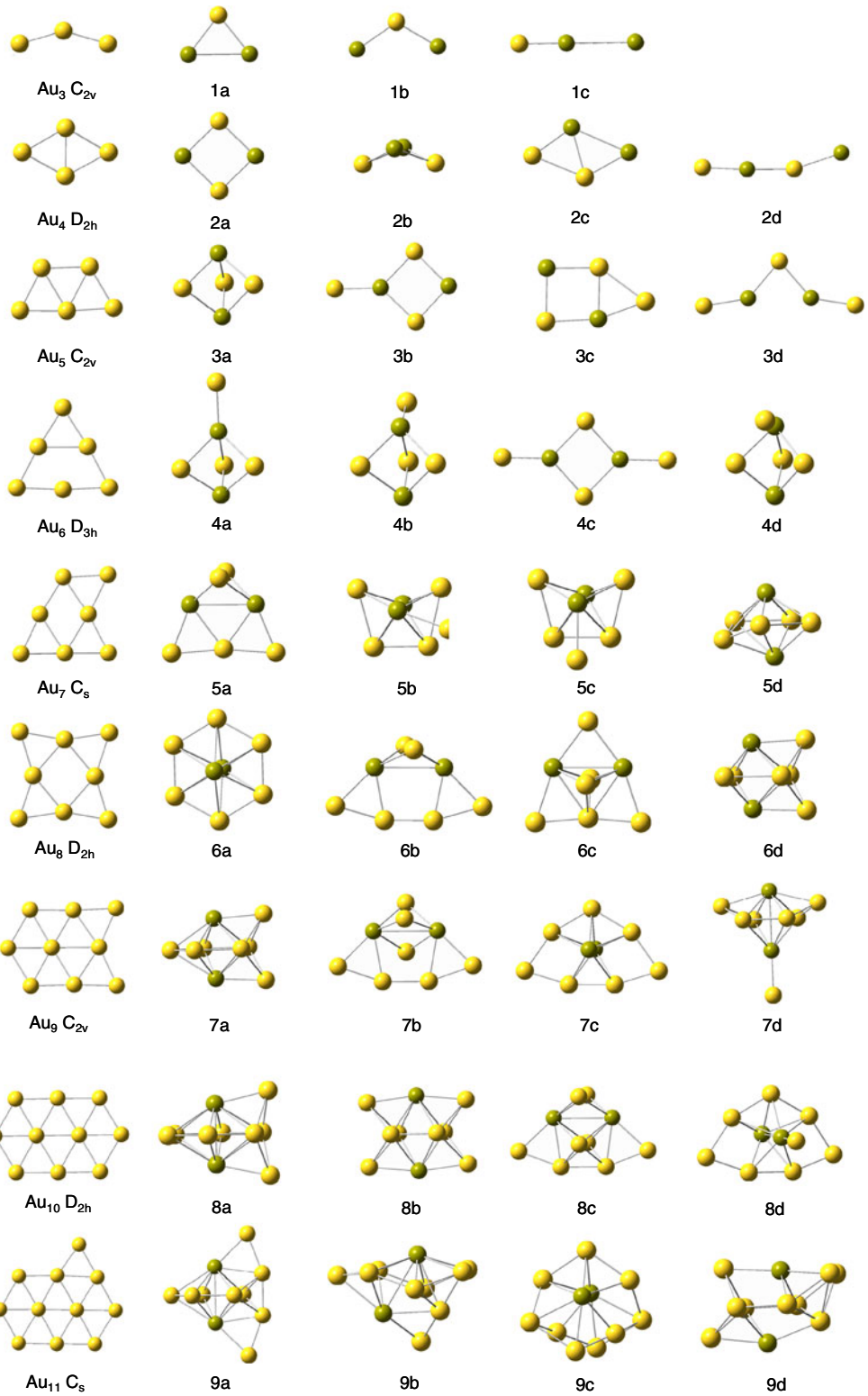



**Fig. 1** The lowest-energy structures of $Au_{n+2}$ and $Ca_2Au_n$ ($n$=1–9) clusters, and a few low-lying isomers of doped clusters. The *yellow and green balls* represent Au and Ca atoms, respectively

size are displayed in Fig. 1. These isomers were designated ***n*a**, ***n*b**, ***n*c**, and ***n*d** (in order over increasing total energy; $n$ is the number of Au atoms in the $Ca_2Au_n$ cluster). Meanwhile, Table 1 lists the electronic states, symmetries, relative energies, HOMO and LUMO energies, as well as the vibration frequencies of the $Ca_2Au_n$ ($n$=1–9) clusters.

All possible initial structures of $Ca_2Au$ clusters—linear structures ($D_{\infty h}$, $C_{\infty v}$), and triangular structures (acute angle or obtuse angle)—were optimized as stable structures with

**Table 1** Electronic states, symmetries, relative energies ($\Delta E$), HOMO energies, LUMO energies, and vibration frequencies of $Ca_2Au_n$ ($n$ = 1–9) clusters

| Isomer | State | Symmetry | $\Delta E$ (eV) | HOMO (hartree) | LUMO (hartree) | Frequency ($cm^{-1}$) |
|---|---|---|---|---|---|---|
| **1a** | $^2B_2$ | $C_{2v}$ | 0.00 | −0.12456 | −0.10260 | 120.4, 127.6, 190.5 |
| **1b** | $^2A_1$ | $C_{2v}$ | 0.29 | −0.12676 | −0.10509 | 23.8, 113.3, 177.1 |
| **1c** | $^2\sum$ | $C_{\infty v}$ | 0.47 | −0.11707 | −0.09811 | 20.7, 86, 231.7 |
| **2a** | $^1A_g$ | $D_{2h}$ | 0.00 | −0.11911 | −0.08503 | 49.5, 56.6, 135.6, 147, 174.9 |
| **2b** | $^3B_1$ | $C_{2v}$ | 0.34 | −0.11970 | −0.11151 | 29.4, 79.9, 134.3, 158.7, 167.2 |
| **2c** | $^1A'$ | $C_s$ | 0.68 | −0.13121 | −0.09129 | 30.9, 54.2, 87.5, 143.6, 202.4 |
| **2d** | $^2\sum_g$ | $D_{\infty h}$ | 0.78 | −0.15224 | −0.07785 | 17.7, 37.7, 51.3, 219.4, 258.1 |
| **3a** | $^2A_2''$ | $D_{3h}$ | 0.00 | −0.12071 | −0.10703 | 26.8, 27, 64.6, 130.3, 131 |
| **3b** | $^2A_1$ | $C_{2v}$ | 0.45 | −0.13363 | −0.12045 | 15.9, 19.2, 45.3, 55.1, 77.4 |
| **3c** | $^2A'$ | $C_s$ | 0.59 | −0.12016 | −0.10674 | 21.3, 21.4, 51.1, 54.1, 76.5 |
| **3d** | $^2A_1$ | $C_{2v}$ | 1.06 | −0.14779 | −0.12461 | 13.3, 23.3, 25.7, 43.8, 53.2 |
| **4a** | $^1A_1$ | $C_{3v}$ | 0.00 | −0.18657 | −0.10912 | 16.6, 17.4, 28.1, 28.2, 58.3 |
| **4b** | $^1A'$ | $C_s$ | 0.02 | −0.16734 | −0.09949 | 19.5, 23.1, 29.2, 56, 65.8 |
| **4c** | $^1A_g$ | $D_{2h}$ | 0.16 | −0.20480 | −0.10441 | 7.2, 13.2, 20.7, 40.5, 41.7 |
| **4d** | $^1A'$ | $C_s$ | 0.16 | −0.15922 | −0.09673 | 26.5, 27.3, 29.3, 44.9, 64.7 |
| **5a** | $^2A_1$ | $C_{2v}$ | 0.00 | −0.14460 | −0.12736 | 13, 18.6, 21.7, 30.5, 53.6 |
| **5b** | $^2A$ | $C_s$ | 0.16 | −0.14143 | −0.12971 | 25, 29.1, 34.8, 42.6, 58.9 |
| **5c** | $^2A''$ | $C_s$ | 0.22 | −0.14350 | −0.13159 | 15.9, 26.3, 33.6, 52.1, 53.8 |
| **5d** | $^2A'$ | $C_s$ | 0.32 | −0.14017 | −0.12872 | 5.9, 24.5, 30.9, 36.2, 43.3 |
| **6a** | $^1A'$ | $D_{6h}$ | 0.00 | −0.16645 | −0.12185 | 7.7, 8, 39.1, 41.2, 41.6 |
| **6b** | $^1A'$ | $C_s$ | 0.06 | −0.19926 | −0.11191 | 13.9, 16.8, 19.5, 20.6, 49.6 |
| **6c** | $^1A_1$ | $C_{2v}$ | 0.26 | −0.17139 | −0.12503 | 12.7, 13.4, 24, 34.8, 38 |
| **6d** | $^1A_1$ | $C_{2v}$ | 0.48 | −0.15389 | −0.12808 | 24.2, 28.6, 38.1, 43, 44.5 |
| **7a** | $^2B_2$ | $C_{2v}$ | 0.00 | −0.16237 | −0.15379 | 24, 32.2, 32.5, 39.1, 40.9 |
| **7b** | $^2A'$ | $C_s$ | 0.36 | −0.15379 | −0.14092 | 12.6, 15.1, 18.1, 35.7, 40.3 |
| **7c** | $^2A'$ | $C_{2v}$ | 0.51 | −0.13490 | −0.12447 | 9.1, 11.5, 30.9, 34.4, 43.1 |
| **7d** | $^2A_2$ | $C_{2v}$ | 0.61 | −0.17954 | −0.17032 | 7.7, 14.5, 23, 24.3, 39.6 |
| **8a** | $^1A_1$ | $C_{2v}$ | 0.00 | −0.17768 | −0.13169 | 9.1, 22.3, 26.4, 27, 29.1 |
| **8b** | $^1A_g$ | $D_{2h}$ | 0.18 | −0.17576 | −0.14020 | 18.1, 27, 37, 42.8, 43.3 |
| **8c** | $^1A_1$ | $C_{2v}$ | 0.43 | −0.16688 | −0.13796 | 15.8, 18.4, 33.2, 35, 38.4 |
| **8d** | $^1A'$ | $C_s$ | 0.57 | −0.18688 | −0.13005 | 9.2, 12, 15.3, 16, 23.2 |
| **9a** | $^2B_1$ | $C_{2v}$ | 0.00 | −0.17598 | −0.16782 | 14.4, 15.9, 19.4, 23.2, 40.2 |
| **9b** | $^2A$ | $C_1$ | 0.05 | −0.15643 | −0.14351 | 18.5, 21.5, 29.4, 31, 35.8 |
| **9c** | $^2A$ | $C_2$ | 0.10 | −0.16550 | −0.15682 | 17.2, 30.7, 38, 43.8, 46.5 |
| **9d** | $^2A'$ | $C_s$ | 0.17 | −0.17664 | −0.16828 | 19.2, 24.5, 31, 33.9, 37.8 |

different spin multiplicities. The calculated results illustrate that the lowest-energy isomer is an acute-angle triangular structure (**1a**) with an angle of 75.9° and a Ca–Au bond length of 2.85 Å. The next isomer, **1b**, with an apex angle of 111.4° and $C_{2v}$ symmetry, can be viewed as a substituted version of the structure of the ground-state $Au_3$ cluster. The linear structure **1c** with an Au atom at one end is 0.47 eV higher in energy than the **1a** isomer. The calculated VIP values are 4.94 eV (**1a**), 5.06 eV (**1b**), and 4.80 eV (**1c**), respectively.

For $Ca_2Au_2$ clusters, the lowest-energy isomer, **2a**, is a square structure with $D_{2h}$ symmetry, like the ground-state $Ag_2Au_2$ and $Cu_2Au_2$ clusters. This isomer, with equal Ca–Au bonds (2.81 Å), can be viewed as a substituted version of the structure of the $Au_4$ cluster. Li et al. [25] reported that the dibridged structure is the global minimum for $Si_2Au_2$ clusters. However, this structure (**2b**) for $Ca_2Au_2$ clusters is 0.34 eV higher in energy than the planar isomer **2a**. This indicates that the different dopant atoms have significant effects on the gold clusters. The distorted rhombus **2c** is

another planar configuration with $C_s$ symmetry. The linear isomer, labeled **2d**, has an electronic state of $^1\Sigma_g$, and exhibits a Au–Ca bond length of 2.66 Å. The VIPs (4.88 eV for **2a**, 4.99 eV for **2b**, 5.28 eV for **2c**, and 5.93 eV for **2d**) are clearly less that that for the $Au_4$ cluster (7.91 eV).

At $n$=3, the lowest-energy structures of the $Ca_2Au_n$ clusters start to include 3D geometries. The trigonal bipyramid (**3a**) was found to be the most stable isomer for $Ca_2Au_3$ clusters; this has $D_{3h}$ symmetry and a VIP of 4.87 eV. When the **2a** isomer is capped with another Au atom, the planar **3b** and **3c** isomers are obtained. However, the isomer **3b** is a scoop-shaped structure with a $C_{2v}$ point group, while **3c** is a house-like structure. Another planar isomer (**3d**) can be viewed as two Au atoms capping the bottom of the triangular structure $Ca_2Au$ (**1a**). The VIP values for **3b**, **3c**, and **3d** are 5.32 eV, 4.89 eV, and 5.91 eV, respectively.

When the **3a** isomer is capped with another Au atom, the three isomers (**4a**, **4b** and **4d**) are generated, as the fourth gold atom can cap the **3a** structure in various different directions. Among these, the isomer **4a** is the lowest-energy structure for $Ca_2Au_4$, and has a higher symmetry $C_{3v}$ point group; **4b** and **4d** are higher in total energy than **4a** by 0.02 and 0.16 eV, respectively. The stable planar structure (**4c**) with $D_{2h}$ symmetry is obtained when the **3b** isomer is capped symmetrically with one Au atom. If we consider the ground states of $M_4Li_2$ ($M$=Cu, Ag, Au) clusters [26], we obtain the doubly bridged square structure ($D_{4h}$) (as shown in the ESM), but this is actually 0.28 eV higher than that of isomer **4a**. It is interesting to note that the Au–Au distance (2.872 Å) in the $Ca_2Au_4$ is very close to that (2.874 Å) in $Li_2Au_4$, and the HOMO–LUMO gap (1.26 eV) for $Ca_2Au_4$ is obviously lower than that (1.77 eV) for $Li_2Au_4$. We also calculated the VIPs for the $Ca_2Au_4$ isomers: 7.93 eV (**4a**), 6.67eV (**4b**), 7.46 eV (**4c**), and 6.37 eV (**4d**).

For $Ca_2Au_5$ clusters, the most stable isomer, **5a**, with $C_{2v}$ symmetry is generated by capping the **4b** structure with one Au atom; its VIP value is 5.87 eV. Similar to the $Ca_2Au_4$ isomers, **5b** and **5c** are two structures derived from the **4f** isomer, where the fifth Au atom can cap the boat-like structure in different directions. Therefore, these two isomers have different relative energies (0.16 eV and 0.22 eV), electronic states ($^2$A and $^2$A″), and VIP values (5.72 eV and 5.80 eV). The pentagonal bipyramid structure **5d** is obtained when two Au atoms in the $Au_7$ cluster with $D_{5h}$ symmetry are replaced with two Ca atoms. Unfortunately, its $D_{5h}$ symmetry drops to $C_s$ symmetry due to Jahn–Teller effects. For the isomer **5d**, the VIP value is 5.65 eV. From $n$=5 onwards, the three-dimensional structures show higher stabilities than the planar clusters. For example, the lowest-energy planar structure (**5g**), which can be viewed as a substituted version of the structure of the $Au_7$ cluster, is higher in energy than **5a** by 0.43 eV.

Numerous possible initial geometries were optimized for $Ca_2Au_6$ clusters. As shown in Fig. 1, the most stable isomer **6a**, with $D_{6h}$ symmetry, can be viewed as the $Au_6$ ring cluster capped by two Ca atoms. When one Au atom is added to the bottom of isomer **5a**, **6b** is obtained. Another two 3D isomers, **6c** and **6d**, which have $C_{2v}$ symmetry, can be viewed as being derived from the doubly bridged square structure ($D_{4h}$). It was found that the substituted version (**6e**) of the $Au_8$ cluster is 0.49 eV higher in energy than the **6a** cluster (see the ESM). The VIP values of these isomers were 6.41 eV (**6a**), 7.31 eV (**6b**), 6.53 eV (**6c**), and 6.05 eV (**6d**).

When we considered stable structures of $Ca_2Au_6$ clusters, two optimized isomers (**7a** and **7b**) were obtained by capping the **6d** and **6b** structures of $Ca_2Au_6$ with a Ca atom. The calculated VIP values for **7a** and **7b** were 6.22 and 5.95 eV, respectively. Isomer **7c**, with $C_{2v}$ symmetry and a VIP of 5.26 eV, can be viewed as the $D_{5h}$ structure of $Ca_2Au_6$ capped by two Au atoms. The VIP for the umbrella-like structure **7d** was 5.47 eV. Noting the sequence of relative energies, it is clear that the stabilities of the three isomers **7i**, **7l**, and **7o**, which can be considered substituted versions of the ground-state $Au_9$ cluster, are obviously weaker than that of the lowest-energy structure **7a** by 1.10, 1.32, and 3.38 eV, respectively.

Similarly, the most stable isomer, **8a**, with $C_{2v}$ symmetry is generated by capping the ground-state $Ca_2Au_6$ cluster with two Au atoms. When four Au atoms are employed to cap the doubly bridged square structure ($D_{4h}$), the stable structure **8b** ($D_{2h}$ symmetry) is generated. This has two different Au–Ca bond lengths: 2.93 Å and 2.97 Å. The optimized isomers **8c** and **8d** were obtained after capping the **7b** and **7c** structures of $Ca_2Au_7$ with a Ca atom. The VIP values for these isomes, 6.58 eV (**8a**), 6.50 eV (**8b**), 6.29 eV (**8c**), and 6.12 eV (**8d**), were also obtained. For $Ca_2Au_8$, only one substituted version (**8l**) of the ground-state $Au_{10}$ cluster was obtained, which was 1.79 eV higher in energy than **8a**.

When the number of gold atoms was increased to 9, the 3D isomer **9a** was found to be the lowest energy structure; this has $C_{2v}$ symmetry and a VIP of 6.51 eV, and can be viewed as the ground-state structure of $Ca_2Au_7$ capped with two Au atoms. Two other configurations (**9b** and **9c**) with $C_1$ and $C_2$ symmetries are 0.05 and 1.10 eV higher than **9a**, respectively. Their calculated VIP values were 5.95 and 6.19 eV. When the front Au atom of **7a** is capped with two extra Au atoms, the 3D structure **9d**, with a distorted $C_s$ symmetry, is obtained, which has a VIP value of 6.50 eV. One planar isomer, **9m**, is obtained based on the ground-state $Au_{11}$ cluster, which can be considered to be the structure of ground-state $Au_{11}$ but with two Au atoms replaced with Ca atoms.

From the above discussion, it is noteworthy that the lowest-energy configurations of $Ca_2Au_n$ clusters tend to be

three-dimensional (3D) structures for $n>2$. This indicates that the two dopant Ca atoms dramatically affect the geometries of the ground-state $Au_n$ clusters. In addition, the $Ca_2Au_{n-1}$ structure capped with one Au atom is the dominant growth pattern for $Ca_2Au_n$ clusters of various sizes ($n$=1–9).

## Relative stabilities

In order to predict and compare the relative stabilities of the most stable $Au_{n+2}$ and $Ca_2Au_n$ ($n$=1–9) clusters, it is worth investigating the average atomic binding energy $E_b(n)$ and second-order difference in energy $\Delta_2E(n)$.

For $Au_n$ clusters, $E_b(n)$ and $\Delta_2E(n)$ are defined using the following formulae:

$$E_b(n) = \frac{nE(\mathrm{Au}) - E(\mathrm{Au}_n)}{n}, \quad (1)$$

$$\Delta_2E(n) = E(\mathrm{Au}_{n-1}) + E(\mathrm{Au}_{n+1}) - 2E(\mathrm{Au}_n), \quad (2)$$

where $E(Au_{n-1})$, $E(Au)$, $E(Au_n)$, and $E(Au_{n+1})$ denote the total energies of the $Au_{n-1}$, Au, $Au_n$, and $Au_{n+1}$ clusters, respectively.

For $Ca_2Au_n$ clusters, $E_b(n)$, $\Delta_2E(n)$ are defined using the following formulae:

$$E_b(n) = \frac{2E(\mathrm{Ca}) + nE(\mathrm{Au}) - E(\mathrm{Ca_2Au}_n)}{n+2}, \quad (3)$$

$$\Delta_2E(n) = E(\mathrm{Ca_2Au}_{n-1}) + E(\mathrm{Ca_2Au}_{n+1}) - 2E(\mathrm{Ca_2Au}_n), \quad (4)$$

where $E(Ca_2Au_{n-1})$, $E(Au)$, $E(Ca)$, $E(Ca_2Au_n)$, and $E(Ca_2Au_{n+1})$ denote the total energies of the $Ca_2Au_{n-1}$, Au, Ca, $Ca_2Au_n$, and $Ca_2Au_{n+1}$ clusters, respectively. The calculated $E_b(n)$ and $\Delta_2E(n)$ values of the $Au_{n+2}$ and $Ca_2Au_n$ clusters are plotted against the cluster size in Fig. 2. The figure shows some interesting results. First, the average atomic binding energy and the second-order difference in energy of $Au_{n+2}$ clusters show even–odd alternation with cluster size. This indicates that the clusters with even numbers of atoms are more stable that their neighboring clusters with odd numbers of atoms, which agrees with the results of previous works [48, 53]. Second, the $E_b(n)$ values of the $Ca_2Au_n$ clusters are significantly higher than those of the $Au_{n+2}$ clusters, which suggests that the presence of impurity (Ca) atoms can strongly enhance the stabilities of small gold clusters. Third, the average atomic binding energy of the $Ca_2Au_n$ clusters increases dramatically as the size of $Ca_2Au_n$ cluster increases from $n$=1 to $n$=4; after that, the average atomic binding energy oscillates slightly with $n$. Therefore, a peak occurs at $n$=4, indicating that $Ca_2Au_4$ more stable than its neighboring clusters. Lastly, the second-order differences in energy for $Ca_2Au_n$ clusters

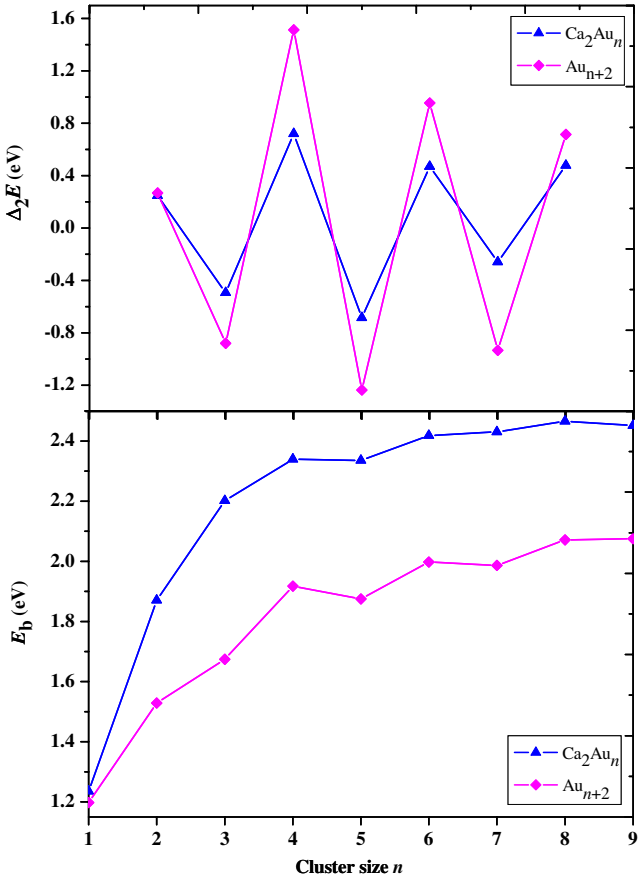



**Fig. 2** Size dependence of the average atomic binding energies and the second-order differences in energies for the lowest-energy structures of $Au_{n+2}$ and $Ca_2Au_n$ ($n$=1–9) clusters

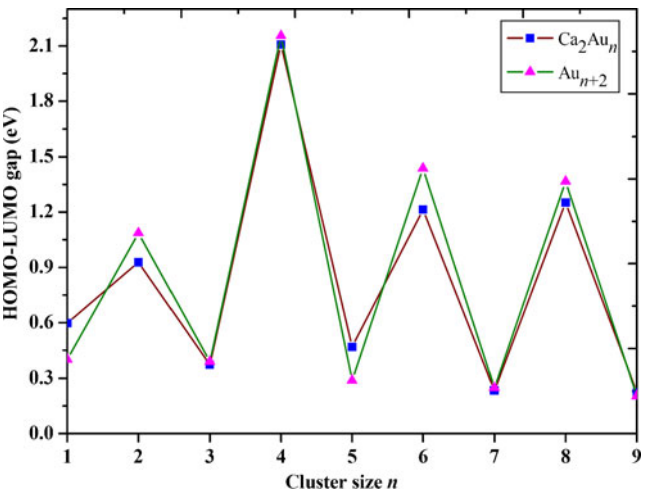



**Fig. 3** Size dependence of the HOMO–LUMO gap for the lowest-energy structures of $Au_{n+2}$ and $Ca_2Au_n$ ($n$=1–9) clusters

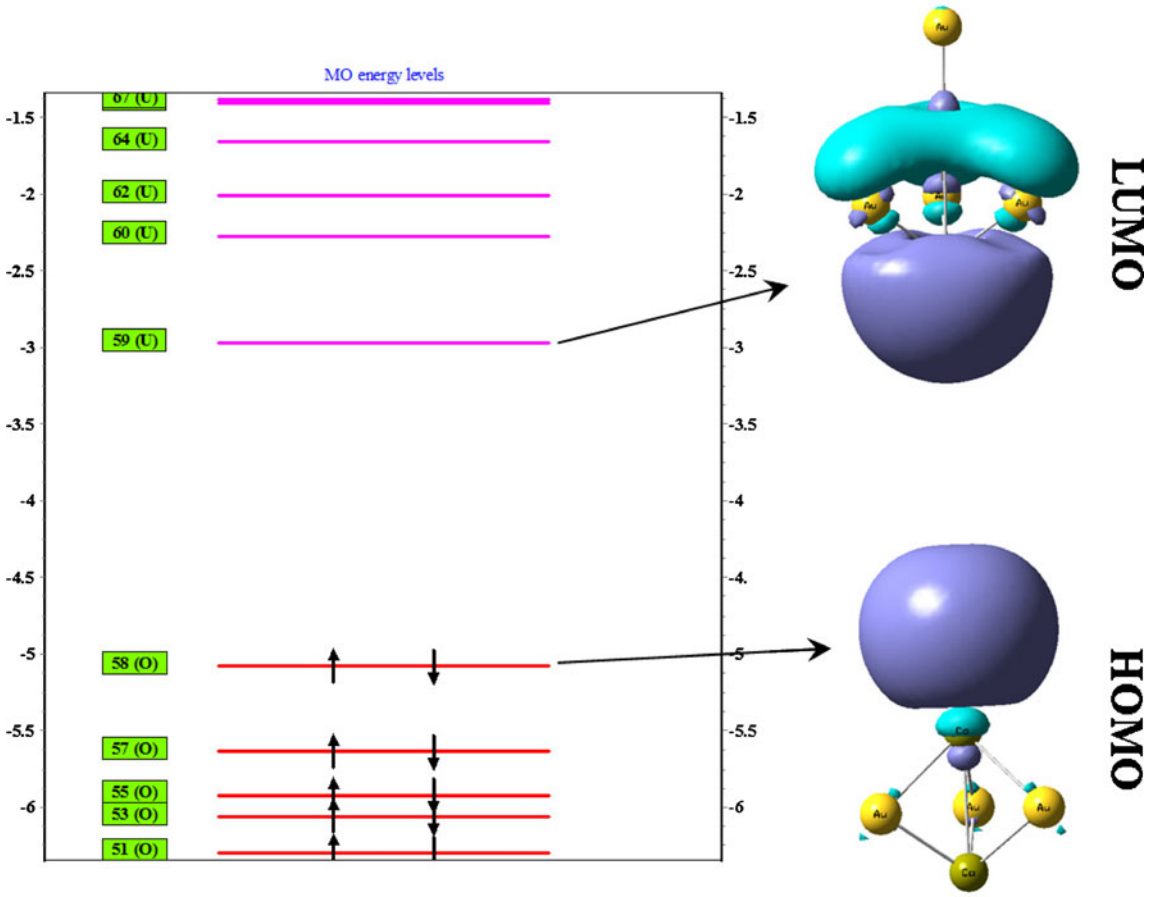



**Fig. 4** Calculated molecular orbital energy levels of $Ca_2Au_4$ together with a molecular orbital map for the HOMO and LUMO; the isovalue is 0.02 a.u.

shows odd–even alternation, as shown in Fig. 2, and a conspicuous maximum appears at $n$=4.

## HOMO-LUMO gaps and charge transfer

The electronic properties of a cluster are reflected in the energy gap between the highest occupied and lowest unoccupied molecular orbitals (HOMO–LUMO), the VIP, the VEA, its chemical hardness, and its polarizability. Among these, the HOMO–LUMO gap is considered to be an important criterion in terms of the electronic stability of a cluster [54]. It represents the ability of molecule to participate in chemical reactions to some degree. A large HOMO–LUMO energy gap is associated with enhanced

**Table 2** Natural charge on each atom in the lowest-energy $Ca_2Au_n$ ($n$=1–9) clusters

| Isomer | Ca-1 | Ca-2 | Au-1 | Au-2 | Au-3 | Au-4 | Au-5 | Au-6 | Au-7 | Au-8 | Au-9 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $Ca_2Au$ | 0.423 | 0.423 | −0.846 | | | | | | | | |
| $Ca_2Au_2$ | 0.752 | 0.752 | −0.752 | −0.752 | | | | | | | |
| $Ca_2Au_3$ | 0.832 | 0.832 | −0.561 | −0.552 | −0.552 | | | | | | |
| $Ca_2Au_4$ | 0.831 | 1.052 | −0.466 | −0.468 | −0.468 | −0.481 | | | | | |
| $Ca_2Au_5$ | 0.984 | 0.984 | −0.284 | −0.284 | −0.450 | −0.450 | −0.500 | | | | |
| $Ca_2Au_6$ | 1.244 | 1.244 | −0.415 | −0.415 | −0.415 | −0.415 | −0.415 | −0.415 | | | |
| $Ca_2Au_7$ | 1.142 | 1.142 | −0.253 | −0.253 | −0.528 | −0.528 | −0.172 | −0.172 | −0.378 | | |
| $Ca_2Au_8$ | 1.136 | 1.136 | −0.028 | −0.202 | −0.202 | −0.028 | −0.341 | −0.341 | −0.564 | −0.564 | |
| $Ca_2Au_9$ | 0.997 | 0.997 | −0.278 | −0.138 | −0.138 | −0.233 | −0.273 | −0.273 | −0.233 | −0.215 | −0.215 |

**Table 3** Natural electronic configurations of the Ca atoms in the lowest-energy $Ca_2Au_n$ systems [Ca (1) and Ca (2) correspond to the top (or left) Ca and the bottom (or right) Ca atoms in Fig. 1, respectively]

| Isomer | Ca (1) | | | | Ca (2) | | | |
|---|---|---|---|---|---|---|---|---|
| | 4*s* | 3*d* | 4*p* | 5*p* | 4*s* | 3*d* | 4*p* | 5*p* |
| $Ca_2Au$ | 1.25 | 0.13 | 0.20 | 0 | 1.25 | 0.13 | 0.20 | 0 |
| $Ca_2Au_2$ | 0.95 | 0.17 | 0.14 | 0 | 0.95 | 0.17 | 0.14 | 0 |
| $Ca_2Au_3$ | 0.61 | 0.24 | 0.33 | 0 | 0.61 | 0.24 | 0.33 | 0 |
| $Ca_2Au_4$ | 0.38 | 0.26 | 0.23 | 0.31 | 0.32 | 0.31 | 0.24 | 0.09 |
| $Ca_2Au_5$ | 0.31 | 0.29 | 0.43 | 0 | 0.31 | 0.29 | 0.43 | 0 |
| $Ca_2Au_6$ | 0 | 0.28 | 0.26 | 0.10 | 0 | 0.28 | 0.26 | 0.10 |
| $Ca_2Au_7$ | 0.24 | 0.26 | 0.39 | 0 | 0.24 | 0.26 | 0.39 | 0 |
| $Ca_2Au_8$ | 0 | 0.24 | 0 | 0.46 | 0 | 0.24 | 0 | 0.46 |
| $Ca_2Au_9$ | 0.23 | 0.25 | 0.55 | 0 | 0.23 | 0.25 | 0.55 | 0 |

chemical stability. HOMO and LUMO energies for the low-lying configurations of the $Ca_2Au_n$ ($n$=1–9) clusters are listed in Table 1. In addition, the HOMO–LUMO gaps for the most stable $Au_{n+2}$ and $Ca_2Au_n$ clusters are plotted against the cluster size in Fig. 3. As seen from the figure, the HOMO–LUMO gaps for $Ca_2Au_n$ and $Au_{n+2}$ clusters exhibit obvious odd-even oscillations: the even-numbered clusters show larger HOMO–LUMO gaps than the odd-numbered ones, as expected for closed/open shell systems. Namely, clusters with even numbers of electrons have enhanced chemical stability as they have larger gaps than their neighbors. In particular $Ca_2Au_4$ has the largest HOMO–LUMO gap of 2.11 eV. This means that $Ca_2Au_4$ possesses dramatically enhanced chemical stability, which agrees with the above analysis based on $E_b$ ($n$) and $\Delta_2E$ ($n$). In order to investigate the $Ca_2Au_4$ cluster, the molecular orbital energy levels and contour maps for the HOMO and LUMO were obtained, and these are displayed in Fig. 4. For $Ca_2Au_4$, the HOMO involves the 4*s* and $5p_z$ orbitals from the Ca atoms and the 6*s* and $5d_z^2$ orbitals from the Au atoms; the LUMO mostly involves the 4*s* and $4p_z$ orbitals of the Ca atoms as well as the 6*s* and $6p_z$ orbitals of the Au atoms. These molecular orbitals indicate that *spd* hybridization occurs between the gold and calcium atoms.

Natural population analysis (NPA) of the $Ca_2Au_n$ ($n$= 1–9) clusters can provide reliable charge-transfer (CT) information [55], and the results from natural population analysis for the lowest energy $Ca_2Au_n$ species are summarized in Table 2. NPA clearly shows the ionic character of the Ca–Au bond in these clusters. The calcium atoms possess positive charges of 0.846–2.488 electrons, suggesting that the charge in the corresponding cluster transfers from the Ca atoms to the $Au_n$ framework due to the larger electronegativity of Au than Ca. In order to understand this internal charge transfer, the natural electron configurations for the Ca atoms in the most stable $Ca_2Au_n$ systems were investigated, and these are tabulated in Table 3. It is shown that the 4*s* states lose electrons 0.75–2.00, while the 3*d* and 4*p* states get 0.24–0.80 electrons. The results are in agreement with the well-known fact in chemistry that internal charges transfer from 4*s* orbital to 3*d* and 4*p* states in Ca atom, which should be related to orbital overlap and Pauli repulsion. For the

**Table 4** Chemical hardnesses, vertical electron affinities and vertical ionization potentials of the lowest-energy $Ca_2Au_n$ and $Au_{n+2}$ ($n$=1–9) clusters

| Cluster size ($n$) | $Ca_2Au_n$ | | | $Au_{n+2}$ | | | |
|---|---|---|---|---|---|---|---|
| | $\eta$ | VEA | VIP | $\eta$ | VEA | VIP | VIP [29] |
| 1 | 3.27 | 1.22 | 4.49 | 5.14 | 3.23 | 8.37 | 7.50 |
| 2 | 3.91 | 0.97 | 4.88 | 5.51 | 2.40 | 7.91 | 8.60 |
| 3 | 3.56 | 1.31 | 4.87 | 4.51 | 2.97 | 7.48 | 8.00 |
| 4 | 6.42 | 1.51 | 7.93 | 6.44 | 2.01 | 8.45 | 8.80 |
| 5 | 4.26 | 1.61 | 5.87 | 4.01 | 3.12 | 7.13 | 7.80 |
| 6 | 4.65 | 1.76 | 6.41 | 5.03 | 2.69 | 7.72 | 8.65 |
| 7 | 3.79 | 2.43 | 6.22 | 3.70 | 3.30 | 7.00 | 7.15 |
| 8 | 4.61 | 1.97 | 6.58 | 4.70 | 2.81 | 7.51 | 8.20 |
| 9 | 3.63 | 2.88 | 6.51 | 3.45 | 3.56 | 7.01 | 7.28 |

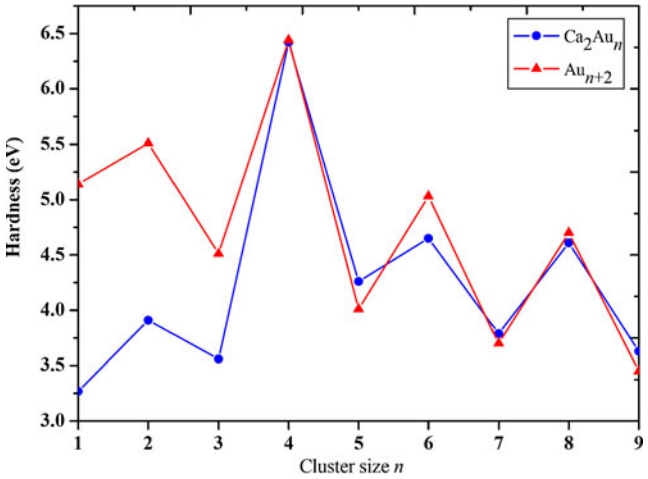



**Fig. 5** Size dependence of the hardnesses of the lowest-energy structures of $Au_{n+2}$ and $Ca_2Au_n$ ($n$=1–9) clusters

contribution from the 5*p* states, it is zero except for $Ca_2Au_4$ $Ca_2Au_6$ and $Ca_2Au_8$ clusters.

## Vertical ionization potential, electron affinity and chemical hardness

Chemical hardness is an electronic quantity that may be applied to characterize the relative stability of molecules and aggregates through the principle of maximum hardness (PMH) proposed by Pearson [56]. In density functional theory (DFT), the second partial derivatives of the ground-state energy $E[\rho]$ of a molecule with respect to the number of electrons $N$ at constant external potential define the chemical hardness $\eta$ of the system:

$$\eta = \left(\frac{\partial^2 E}{\partial N^2}\right)_{v(r)}. \quad (5)$$

Based on the finite-difference approximation and Koopmans' theorem [57], the chemical hardness $\eta$ can be expressed as

$$\eta = \mathrm{VIP} - \mathrm{EA}, \quad (6)$$

where VIP is the vertical ionization potential and EA is the vertical electron affinity [58, 59]. The vertical ionization potential, electron affinity, and chemical hardness were calculated for $Ca_2Au_n$ and $Au_{n+2}$ ($n$=1–9) systems, and they are listed in Table 4. For gold clusters, our calculated values of VIP and VEA show obvious oscillatory behavior with increasing cluster size; the calculated VIP values agree well with experimental data, as expected [60]. For $Ca_2Au_n$ clusters, the VIPs exhibit slight odd-even alternation, in agreement with those for $Au_{n+2}$ clusters, but the same tendency is not found for the VEA values. $Ca_2Au_4$ has the largest VIP of 7.93 eV. This trend, and the maximum VIPs, are also in accord with the above analysis based on the second-order difference in energy and the HOMO–LUMO energy gap. The chemical hardnesses of $Ca_2Au_n$ and $Au_{n+2}$ ($n$=1–9) systems are plotted against the cluster size in Fig. 5. As seen in the figure, the hardnesses for these two types of system show large even-odd cluster size oscillations. Based on the MHP of chemical hardness, these behaviors indicate that the even-numbered isomers

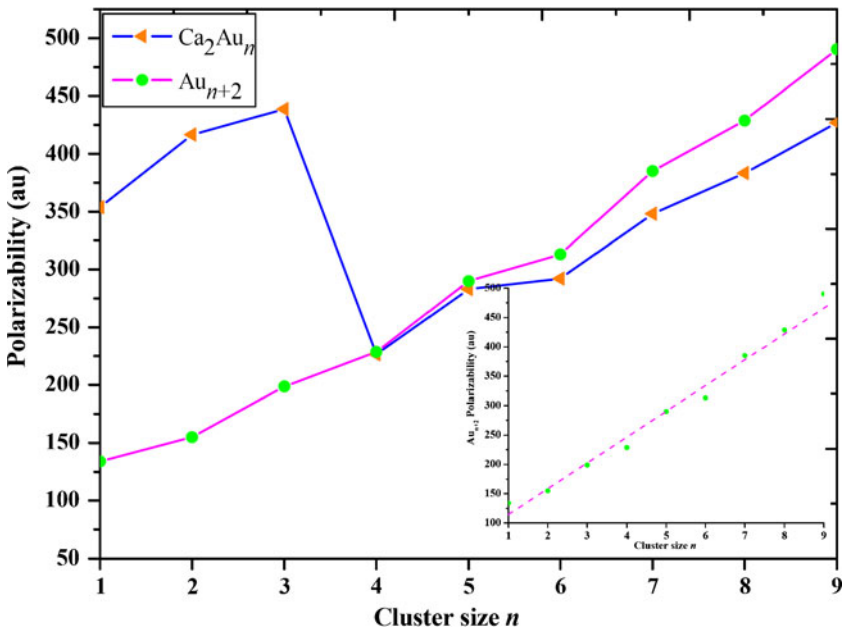



**Fig. 6** Size dependence of the polarizabilities of the lowest-energy structures of $Au_{n+2}$ and $Ca_2Au_n$ ($n$=1–9) clusters. The linear fit to the polarizability for $Au_n$ clusters is shown in the *inset*

(which are harder) are more stable than their neighboring odd-numbered isomers. Moreover, the curves of $\eta$ show an obvious minimum at $n$=4.

Polarizability

It is known that the static polarizability is a measure of the distortion of the electronic density, and provides information about the response of the system under the effect of an external static electric field [61]. It is therefore very sensitive to the delocalization of the valence electrons and structure of the cluster. Here, the average static polarizability is defined as [62]

$$\langle a \rangle = \left(a_{xx} + a_{yy} + a_{zz}\right)/3. \quad (7)$$

In light of previous research [63–66], Maroulis et al. determined the level of calculation needed to compute accurate cluster polarizability values; their results showed that the DFT method can provide reasonable values. The effects of the basis set on cluster polarizability were also discussed by the group of Maroulis [66–69]. We obtained the polarizabilities of our systems with the same basis set as used for the optimizations, so that the calculations required were not excessively time-consuming. For the Au and Ca atoms, our calculated values were 34.8 and 173.0 a.u., respectively. Our results are in reasonable agreement with the reported experimental estimate of 39.1±9.8 a.u. [70] for Au atoms and the theoretical value of 152.0 a.u. [71] for Ca atoms. The value for the Au atom was smaller than that for the Ca atom, which was due to the enhanced screening of the *s* electrons by the *d* electrons in Au atoms. The polarizabilities of the small clusters $Au_{n+2}$ and $Ca_2Au_n$ are shown in Fig. 6. It is clear that the polarizabilities of the gold clusters increase as a function of size $n$, although a slight oscillatory behavior can be seen up to $n$=11. These characteristics agree with the results obtained by Li et al. [48]. Moreover, the inverse corrections to the polarizabilities versus the ionization potential and hardness were found, and these agree with the results for lithium and sodium clusters obtained by the Ghanty group [61]. For $Ca_2Au_n$ clusters, the curve for $\langle a \rangle$ displays the same trend as it does for gold clusters when $n$>3. The differences in shape at $n$=1, 2, and 3 correspond to significant increases compared with the equivalent gold clusters. The curve for $\langle a \rangle$ displays an obvious minimum at $n$=4, which indicates that $Ca_2Au_4$ is much more stable electronically than its neighboring clusters.

## Conclusions

The geometric structures, stabilities, growth behaviors, HOMO–LUMO gaps, charge transfer, chemical hardnesses, and polarizabilities of $Ca_2Au_n$ and $Au_{n+2}$ ($n$=1–9) clusters were investigated systematically by the metal-GGA functional at the TPSS level. All of the results obtained can be summarized as follows:l

(i) The optimized geometries show that the $Ca_2Au_{n-1}$ structure capped with an Au atom is the dominant growth pattern for different sized $Ca_2Au_n$ ($n$=1–9) clusters. In addition, the most stable isomers have 3D structures for $n$=3–9, which indicates that doping with two Ca atoms dramatically affects the geometries of the ground-state $Au_n$ clusters.

(ii) The average atomic binding energies and second-order differences in energy show that the 3D $Ca_2Au_4$ isomer is the most stable structure for $Ca_2Au_n$ ($n$=1–9) clusters. The same pronounced even-odd alternations are seen for the HOMO-LUMO gaps, VIPs, and hardnesses as $n$ increases, which indicates that the clusters with even numbers of atoms, especially $Ca_2Au_4$, have enhanced chemical stabilities compared with their neighbors.

(iii) Our natural population analysis (NPA) results show that the charge in $Ca_2Au_n$ ($n$=1–9) clusters transfers from the Ca atoms to the Au framework. For each Ca atom, the internal charge is transferred from the 4*s* orbital to the 3*d* and 4*p* states. Moreover, the inverse corrections to the polarizabilities versus the ionization potential and hardness were obtained for the gold clusters. The polarizabilities of the $Ca_2Au_n$ clusters were found to depend sensitively on the cluster size, and an obvious minimum occurred at $n$=4.

**Acknowledgments** The authors are grateful to the National Natural Science Foundation of China (nos. 10774103 and 10974138).

## References

1. Alosno JA (2000) Chem Rev 100:637–678
2. Baletto F, Ferrando R (2005) Rev Mod Phys 77:371–423
3. Kittel C (1996) Introduction to solid state physics, 7th edn. Wiley, Now York
4. Kim SH, Medeiros-Ribeiro G, Ohlberg DAA, Williams RS, Heath JR (1999) J Phys Chem B 103:10341–10347
5. Pal R, Wang LM, Huang W, Zeng XC (2009) J Am Chem Soc 131:3396–3404
6. Eachus RS, Marchetti AP, Muenter AA (1999) Annu Rev Phys Chem 50:117–123
7. Zhao Y, Li ZY, Yang JL (2009) Phys Chem Chem Phys 11:2329–2334
8. Hou SM, Zhang JX, Li R, Ning J, Han RS, Shen ZY, Zhao XY, Xue ZQ, Wu QD (2005) Nanotechnology 16:239–245
9. Fournier R (2001) J Chem Phys 115:2165–2177
10. Nijamudheen A, Datta A (2010) J Mol Struct Theochem 945:93–96
11. Assadollahzadeh B, Schwerdtfeger P (2009) J Chem Phys 131:064306

12. Wang F, Liu P, Zhang DJ (2010) J Mol Model. doi:10.1007/s00894-010-0815-4
13. Torres MB, Fernández EM, Balbás LC (2008) J Phys Chem A 112:6678–6689
14. Hashmi ASK, Loos A, Littmann A, Braun I, Knight J, Doherty S, Rominger F (2009) Angew Chem 351:576–582
15. Neumaier M, Weigend F, Hamper O, Kappes MM (2006) J Chem Phys 125:104308
16. Autschbach J, Hess BA, Johansson MP, Neugebauer J, Patzschke M, Pyykkö P, Reiher M, Sundholm D (2004) Phys Chem Chem Phys 6:11–22
17. Ackerson CJ, Jadzinsky PD, Jensen GJ, Kornberg RD (2006) J Am Chem Soc 128:2635–2640
18. Shaw CF III (1999) Chem Rev 99:2589–2600
19. Valden M, Lai X, Goodman DW (1998) Science 281:1647–1650
20. Félix C, Sieber C, Harbich W, Buttet J, Rabin I, Schulze W, Ertl G (2001) Phys Rev Lett 86:2992–2995
21. Heinebrodt M, Malinowski N, Tast F, Branz W, Billas IML, Martin TP (1999) J Chem Phys 110:9915–9921
22. Yuan DW, Wang Y, Zeng Z (2005) J Chem Phys 122:114310–114321
23. Zhang M, He LM, Zhao LX, Feng XJ, Luo YH (2009) J Phys Chem C 113:6491–6496
24. Ghanty TK, Banerjee A, Chakrabarti A (2010) J Phys Chem C 114:20–27
25. Li X, Kiran B, Wang LS (2005) J Phys Chem A 109:4366–4374
26. Wannere CS, Corminboeuf C, Wnag ZX, Wodrich MD, King RB, Schleyer PR (2005) J Am Chem Soc 127:5701–5705
27. Guo JJ, Yang JX, Die D (2005) Phys B 367:158–164
28. Guo JJ, Yang JX, Die D (2006) J Mol Struct Theochem 764:117–121
29. Zhao YR, Kuang XY, Zheng BB, Li YF, Wang SJ (2011) J Phys Chem A 115:569–576
30. Balducci G, Ciccioli A, Gigli G (2004) J Chem Phys 121:7748–7755
31. Tao J, Perdew JP, Staroverov VN, Scueria GE (2003) Phys Rev Lett 91:146401
32. Staroverov VN, Scueria GE, Tao J, Perdew JP (2004) Phys Rev B 69:075102
33. Kanai Y, Wang X, Selloni A, Car R (2006) J Chem Phys 125:234104
34. Tao J, Perdew JP (2005) J Chem Phys 122:114102
35. Goll E, Ernst M, Moegle-Hofacker F, Stoll H (2009) J Chem Phys 130:234112
36. Johansson MP, Lechtken A, Schooss D, Kappes MM, Furche F (2008) Phys Rev A 77:053202
37. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JA Jr, Vreven T, Kudin KN, Burant JC, Millan JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C, Pople JA (2004) Gaussian 03, revision D.01. Gaussian Inc., Wallingford
38. Hay PJ, Wadt WR (1985) J Chem Phys 82:299–310
39. Ehlers AW, Bohme M, Dapprich S, Gobbi A, Hollwarth A, Jonas V, Kohler KF, Stegmann R, Veldkamp A, Frenking G (1993) Chem Phys Lett 208:111–114
40. Krishnan R, Binkley JS, Seeger R, Pople JA (1980) J Chem Phys 72:650–654
41. Häkkinen H, Yoon B, Landman U, Li X, Zhai HJ, Wang LS (2003) J Phys Chem B 107:6168–6175
42. Häkkinen H, Landman U (2000) Phys Rev B 62:R2287–R2290
43. Xiao L, Wang L (2004) Chem Phys Lett 392:452–455
44. Pyykkö P (2004) Angew Chem Int Edn 43:4412–4456
45. Bonačić-Koutecky V, Burda J, Mitrić R, Ge M, Zampella G, Fantucci P (2002) J Chem Phys 117:3120–3131
46. Ferrighi L, Hammer B, Madsen GKH (2009) J Am Chem Soc 131:10605–10609
47. Lee HM, Ge M, Sahu BR, Tarakeshwar P, Kim KS (2003) J Phys Chem B 107:9994–10005
48. Li XB, Wang HY, Yang XD, Zhu ZH (2007) J Chem Phys 126:084505
49. Schiltz J (1963) Ann Phys 8:67–106
50. Huber KP, Herzberg G (1979) Constants of diatomic molecules. Van Nostrand Reinhold, New York
51. Morse MD (1986) Chem Rev 86:1049–1109
52. Wesendrup R, Hunt T, Schwerdtfeger P (2000) J Chem Phys 112:9356–9362
53. Die D, Kuang XY, Gou JJ, Zheng BX (2010) J Phys Chem Solids 71:770–775
54. Rajesh C, Majumder C (2007) J Chem Phys 126:244704
55. Sannigrahi AB, Nandi PK, Schleyer PvR (1994) J Am Chem Soc 116:7225–7232
56. Pearson RG (1997) Chemical hardness: applications from molecules to solids. Wiley-VCH, Weinheim
57. Parr RG, Yang W (1989) Density functional theory of atoms and molecules. Oxford University Press, New York
58. Chermette H (1999) J Comput Chem 20:129–154
59. Senet P (1997) Chem Phys Lett 275:527–532
60. Jackslath C, Rabin I, Schulze W (1992) Ber Bunsenges Phys Chem 96:1200–1204
61. Chandrakumar KRS, Ghanty TK, Ghosh SK (2004) J Phys Chem A 108:6661–6666
62. Gupta K, Ghanty TK, Ghosh SK (2010) Phys Chem Chem Phys 12:2929–2934
63. Maroulis G, Begue D, Pouchan C (2003) J Chem Phys 119:794–797
64. Karamanis P, Maroulis G, Pouchan C (2006) J Chem Phys 124:071101
65. Maroulis G (2008) J Chem Phys 129:044314
66. Maroulis G, Karamanis P, Pouchan C (2007) J Chem Phys 126:154316
67. Maroulis G, Pouchan C (2003) Phys Chem Chem Phys 5:1992–1995
68. Karamanis P, Pouchan C, Maroulis G (2008) Phys Rev A 77:013201
69. Maroulis G (2007) Chem Phys Lett 444:44–47
70. Roos BO, Liudh R, Malmqvist PA, Veryazov V, Widmark PO (2005) J Phys Chem A 109:6575–6579
71. Sadlej AJ, Urban M, Gropen O (1991) Phys Rev A 44:5547–5557

J Mol Model (2012) 18:1345–1354
DOI 10.1007/s00894-011-1142-0

ORIGINAL PAPER

# Deciphering the binding mode of Zolpidem to $GABA_A$ $\alpha_1$ receptor – insights from molecular dynamics simulation

**R. S. K. Vijayan · Dhananjay Bhattacharyya · Nanda Ghoshal**

Received: 31 December 2010 / Accepted: 30 May 2011 / Published online: 7 July 2011


**Abstract** To investigate the binding mode of Zolpidem to $GABA_A$ and to delineate the conformational changes induced upon agonist binding, we carried out atomistic molecular dynamics simulation using the ligand binding domain of $GABA_A$ $\alpha_1$ receptor. Comparative molecular dynamics simulation of the apo and the holo form of $GABA_A$ receptor revealed that $\gamma_2/\alpha_1$ interface housing the benzodiazepine binding site undergoes distinct conformational changes upon Zolpidem binding. We notice that C loop of the $\alpha_1$ subunit experiences an inward motion toward the vestibule and the F loop of $\gamma_2$ sways away from the vestibule, an observation that rationalizes Zolpidem as an alpha1 selective agonist. Energy decomposition analysis carried out was able to highlight the important residues implicated in Zolpidem binding, which were largely in congruence with the experimental data. The simulation study disclosed herein provides a meaningful insight into Zolpidem-$GABA_A$R interactions and helps to arrive at a binding mode hypothesis with implications for drug design.

**Keywords** Molecular dynamics · $GABA_A$ α1 receptor · Docking

**Electronic supplementary material** The online version of this article (doi:10.1007/s00894-011-1142-0) contains supplementary material, which is available to authorized users.

R. S. K. Vijayan · N. Ghoshal (✉)
Structural Biology and Bioinformatics Division,
Indian Institute of Chemical Biology, (A unit of CSIR),
4 Raja S.C. Mullick Road, Jadavpur,
Kolkata 700032, India
e-mail: nghoshal@iicb.res.in

D. Bhattacharyya
Biophysics Division, Saha Institute of Nuclear Physics,
Sector-1, Block-AF, Bidhannagar,
Kolkata 700 064, India

## Introduction

γ-Amino butyric acid (GABA), the principal inhibitory neurotransmitter in the mammalian central nervous system [1] exerts its physiological effects by binding to the ionotropic ($GABA_A$, $GABA_c$) and the metabotropic ($GABA_B$) receptors. Ionotropic $GABA_A$ receptor is a pore forming pentameric membrane spanning protein that belongs to the Cys loop super family of ligand-gated ion channels (LGIC) [2].

Structurally, the $GABA_A$ receptor is a hetero pentameric complex, arranged pseudo symmetrically around a central $Cl^-$ selective ion channel [3]. This remarkable macromolecule contains a binding site, a channel, a gate, and most intriguingly a mechanism to couple the binding event to the gating of the channel.

A prototypical $GABA_A$ receptor is predicted to have a highly β sheeted extracellular hydrophilic N terminal domain (extra cellular domain-ECD), which consists of about 220 amino acids. Each of the five subunits that make up the receptor is about 50,000 Daltons in size. The large N terminal extracellular ligand binding domain also contains the signature disulfide linkage, constituting the Cys loop [4]. The extra cellular domain then transverses to trans membrane region to form a 4 α-helical transmembrane domain (TMD), which are designated as TM1-TM4, followed by a partially resolved intracellular domain. The transmembrane region spans approximately 10Å. The loop connecting the TM1 with TM2 helix and the TM2 helix with the TM3 helix are shorter and are included in the TM region. In contrast, the loop connecting the TM3 helix with TM4 helix is very long and it spans in to the cytostolic region. Each subunit has a "plus" and a "minus" side as per convention [5]. Consequently, the subunit interfaces consist of the plus and minus sides of neighboring subunits.

It is well documented that attenuation of GABAergic neurotransmission is involved in the pathophysiology of several CNS disorders in humans, namely anxiety, pain, and epilepsy [6]. Enhancement of GABA mediated synaptic transmission underlies the pharmacotherapy of various neurological and psychiatric disorders.

Zolpidem, an imidazopyridine, is a popularly prescribed sedative-hypnotic used in the treatment of insomnia. Zolpidem, being classified as a Z-class drug, differs from classical BZDs in both chemical structure (Fig. 1) and neuropharmacological profile [7]. Unlike classical benzodazepines (BZDs), Zolpidem binds with high affinity to the BZ site, present at the $\alpha1(+)/\gamma2\ (-)$ subunit interface of the $GABA_A$ receptor ($GABA_AR$) and exhibits relatively low affinity to $GABA_ARs$ containing the $\alpha_2/\alpha_3$ subunits, and no significant affinity for the $\alpha_5$ subtype [8].

Recent investigations have shown that the sedative actions of BZDs are mediated by $\alpha_1$ containing $GABA_AR$, whereas the other BZD effects such as anxiolysis are mediated by the other $\alpha$ subunit isoforms [9]. This rationalizes why ZPM exerts a sedative/hypnotic pharmacological profile. Very little is known about the structural requirements for Zolpidem binding as structural investigations on $GABA_A$ receptor ionophore complex is precluded due to a host of issues like large size (~50 kD), heterogeneity, low abundance (pmol $mg^{-1}$ of protein), together with other inherent difficulties associated in isolation and purification of integral membrane proteins [10].

Residues believed to be involved in Zolpidem binding site are located in at least six non-contiguous regions in the $\alpha_1$and $\gamma_2$ subunits, which are arbitrarily designated as 'loops' A,B,C of $\alpha_1$subunit and D,E,F belonging to the $\gamma_2$ subunit [11]. To date, only six residues Met57$\gamma_2$, Phe77$\gamma_2$, Met130$\gamma_2$and $\alpha_1$Thr162, $\alpha_1$Gly200, $\alpha_1$Ser204 have been implicated as crucial residues necessary for conferring high-affinity ZPM binding [12–14]. The availability of the crystal structure of AChBP [15] has facilitated modeling the structure of $GABA_AR$ and many groups have come out with comparative/homology models of $GABA_A$ receptor and its variant isoforms. Techniques like radioligand binding, site-directed mutagenesis, substituted cysteine accessibility method (SCAM) and photoaffinity labeling studies have indeed facilitated in broadening our knowledge and understanding of the residues, involved in ZPM binding. Some groups have also supplemented such experimental techniques in tandem with *insilico* docking studies to rationalize and consolidate these experimental findings [11].

To date, to the best of our knowledge, there exists no literature evidence ascribing the dynamic phenomenon of ZPM binding characterizing the subsequent sequel of events, induced upon agonist binding. It is against this backdrop this study aims to delineate the binding mechanism of ZPM and to unravel the conformational perturbations induced upon ZPM binding, particularly confined to the ligand binding domain (LBD) of the $GABA_A$ $\alpha_1$ receptor.

Molecular docking of ZPM to the BZ site was carried out on a published homology model of $GABA_A$ $\alpha_1/\gamma^2R$, which is consistent with most of the experimental data and displays a topology expected within this super family [16].

Comparative molecular dynamics (MD) simulations were carried out for the docked complex (holo form) and the unbound (apo form) using a NPT ensemble. The 20 ns MD trajectories, obtained post equilibration, were subjected to extensive analysis. Essential dynamics using PCA was carried out to reduce the dimensions of the trajectory data and to capture the essential movements of a macromolecular system [17].

Comparative analysis of the bound and unbound trajectories revealed subtle difference indicative of agonist mediated conformational changes in the ligand binding domain. Of these, notable difference was particularly witnessed in the C loop of the $\alpha$ + subunit which displays an inward motion upon agonist binding that leads to an occlusion at the entrance of the BZ site.

On the whole the simulation results were in broad agreement with experimental data and in consonance with previous computational results evinced for other related LGIC proteins. The structural plasticity of the LDB along with agonist induced perturbations has been captured from the simulations and perhaps provides the first glimpse toward a detailed understanding of the molecular events.

$E_{Interaction}$ analysis reveals Vdw interaction to be a driving factor for binding, which leads to conclude that binding is largely dictated by shape recognition. Residue interaction decomposition analysis results were consistent with prior findings established by experimental techniques like SCAM, Alanine scanning and mutagenesis. These interaction studies helped in identifying the key residues involved in inter molecular interactions with implications for structure based design.

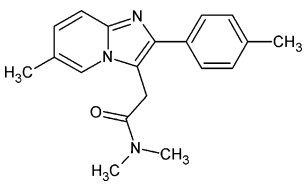



**Fig. 1** 2D Structure of Zolpidem

## Materials and methods

### Molecular docking

To simulate the binding mode of ZPM to the rigid homology modeled receptor, molecular docking was carried

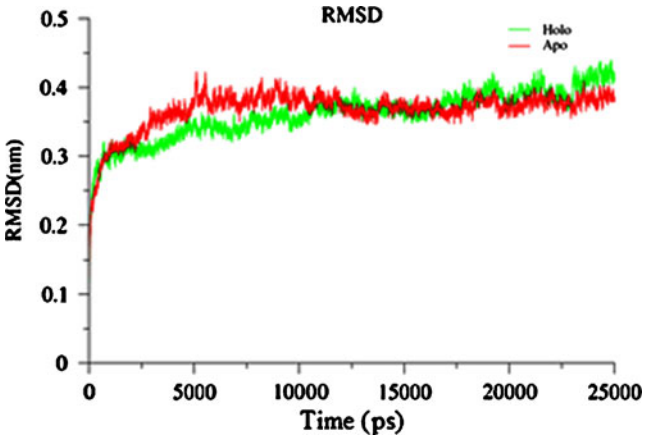



**Fig. 2** Plot of RMSD with respect to time for the apo and the holo form of $GABA_A$ with reference to initial structure

out using GOLD docking program [18]. GOLD uses an evolutionary genetic algorithm (GA) for generating ligand conformations and also imparts partial receptor flexibility in an implicit manner by allowing movement around the side chain dihedrals of protein OH and $NH_3$+. The binding site was defined using the coordinates of HEPES (residues within 10 Å from the ligand). The 'standard default mode' which comprises of 100,000 genetic operations on an initial population of 100 members divided into five subpopulations (number of islands =5) was used. The annealing parameters for the fitness function were set at 4.0 for van der Waals and 2.5 for hydrogen bonding. A niche size of 2, and a selection pressure of 1.1 were used and the early termination option was turned off. The top ranking poses were retained based on Gold fitness function which served as an input for molecular dynamics.

## Molecular dynamics system setup and computational analysis

MD simulation was performed in GROMACS (version 4.0.5) using GROMOS96 43a2 force field [19]. The topologies for Zolpidem were obtained using PRODRG beta server [20]. ZPM was energy optimized in VAMP using the AM1 Hamilitonian type and the RHF formalism [21]. Partial charge calculations were determined using HF/6-31 G* basis set employing the AM1-BCC methods implemented in QUACPAC [22]. The system was solvated in a dodecahedral cell by adding SPC waters keeping a minimum distance of 0.7 nm between the solute and each face of the box. The system was electro neutralized by the addition of counter ions. Leap-frog algorithm was employed for integrating Newton equations of motion. The systems were then energy minimized, initially by 2000 steps of steepest descent followed by conjugate gradient. In order to allow solvent molecules to adjust to the solute, a 50 ps equilibration simulation was carried out by restraining the protein atoms. In the next stage all restraints were removed and a 25 ns MD simulation was initiated at 300 K under NPT ensemble condition using PBC. A Nose-Hoover thermostat [23] and a Parrinello-Rahman barostat [24] were employed for temperature and pressure coupling, respectively. Long-range electrostatic interactions were calculated using particle mesh Ewald (PME) [25] method, the non-bonded interactions were truncated at 12 Å cut-off and the neighbor list updated every 10 steps. The time step for integration was 2 fs, and the coordinates of all atoms were saved every 1 ps. Bonds were constrained using the LINCS algorithm [26]. All MD simulations were run on an Apple X serve cluster. Analyses of the trajectories were performed

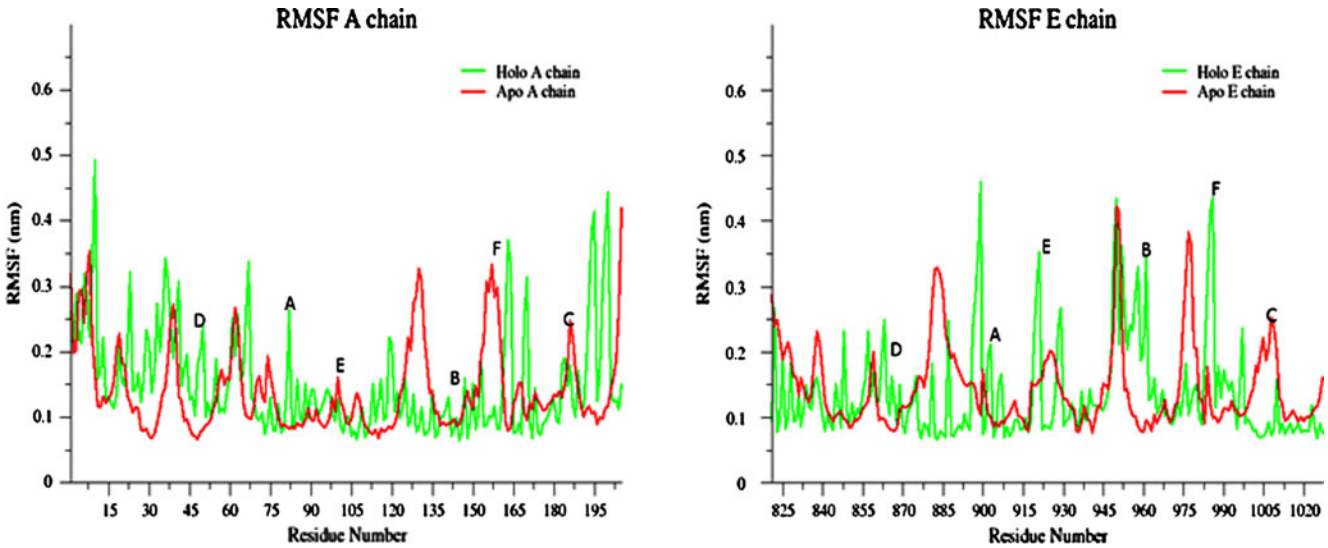



**Fig. 3** RMSF values of the $C_\alpha$ atoms for the chains **A** ($\gamma_2$) and **E** ($\alpha_1$) of the apo form (red) and holo form (green). The loop regions are highlighted using their standard notation

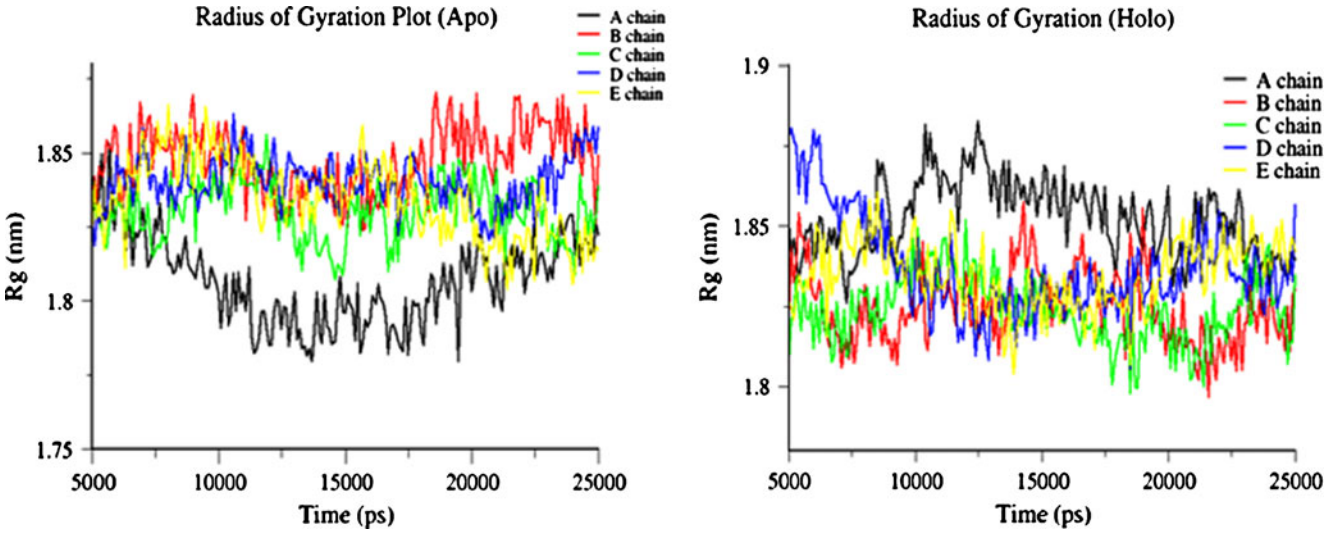



**Fig. 4** Radius of gyration (Rg) for the apo and holo forms, plotted as a function of time for each individual subunit

using analysis tools available in GROMACS suite and the trajectories were visualized using VMD [27]. Hydrogen bonds were determined geometrically with a donor–acceptor distance cutoff of 3.5 Å and an angle cutoff of 30°. The interaction energies were calculated based on the non-bonded interactions, electrostatic ($E_{Coul}$) and van der Waals ($E_{LJ}$) modeled as Coulombic and Lennard-Jones parameters, respectively, between the ligand and the protein. Essential dynamics (ED) was carried out using principal component analysis (PCA) available within the analysis tools of GROMACS package, considering only the backbone atoms for generating the covariance matrix. Prior to performing PCA, all translational and rotational motions were eliminated by fitting the trajectory to the reference average structure. Pore diameter calculations were performed using the HOLE [28] program. In order to visualize the motions, porcupine plots were generated using Tcl scripts in VMD.

## Results and discussion

### Stability analysis of trajectories and mobility of residues

The stability of trajectories was assessed by plotting total energy, pressure and temperature (data provided in supplementary information ((SI), Fig. SI1). Analyses of these graphs signify that the structures were stable during the course of MD simulation.

The root-mean-square displacement (RMSD) of the protein backbone (calculated with respect to the initial structure) is shown in Fig. 2.

The $C_\alpha$ RMSD values plateaued at ~4 Å for both the holo and the apo forms of the receptor. The RMSD plot shows initial drift in the first few hundred ps of the MD simulation. This is attributed to the fact that the production phase was performed without restraining the proteins, whereas the equilibration phase was performed with restraints on the proteins.

The first 5 ns of initial run were considered to be needed for equilibration and the next 5-25 ns of the simulation period was considered as production run. In order to evaluate the dynamic behavior of protein residues, the isotropic temperature (*B*) factor was calculated from the mean square fluctuation. The time-averaged RMSF value for the C-alpha atoms with reference to the average structure was calculated and plotted. The difference in fluctuation for the apo and holo form for the chain A ($\gamma_2$) and E ($\alpha_1$), housing the BZ site, is shown in Fig. 3. The RMSF plots for the other chains (B,C and D) are provided in the supplementary information section (SI2). Regions that report significant fluctuations correspond to the loop

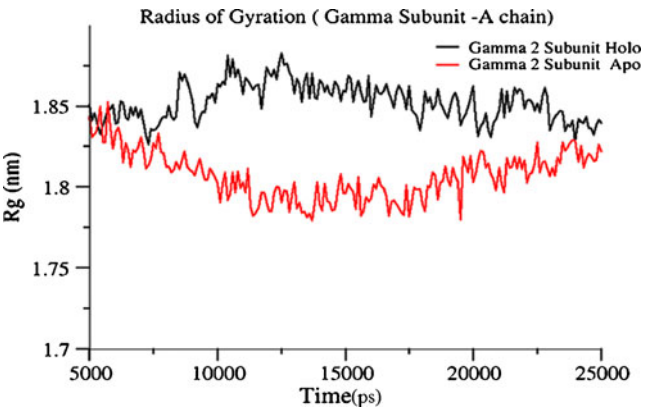



**Fig. 5** Radius of gyration (Rg) for the apo and holo forms plotted as a function of time for the A chain ($\gamma_2$)

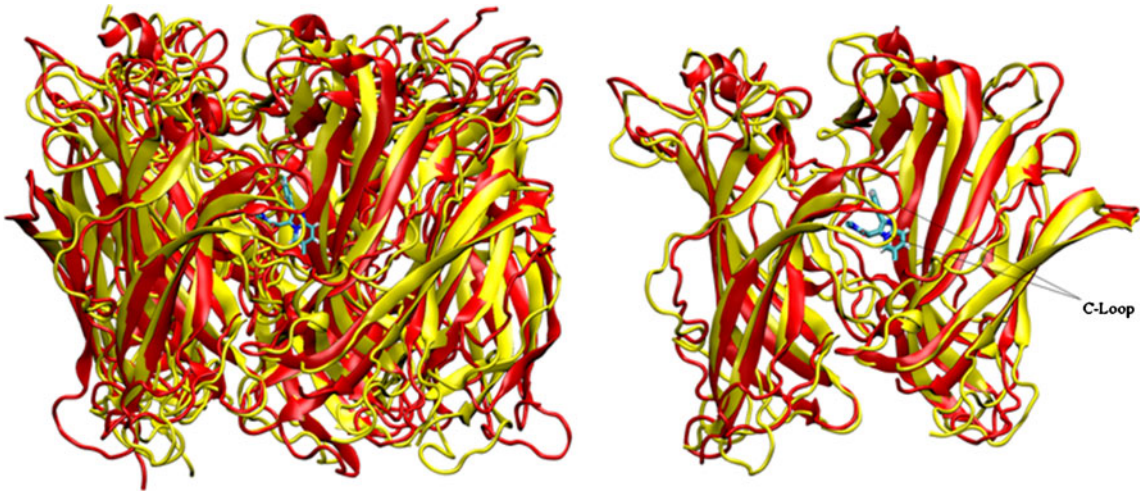



**Fig. 6** Averaged structure apo form (red) and the holo form (yellow) superimposed to signify the degree of overlap. Right panel displays the subunits ($\gamma_2/\alpha_1$), housing the BZ site, highlighting the difference in the orientation of C loop

regions of the $GABA_A$ and the C- terminal region of the individual chains.

Loops exhibiting significant divergence in amplitude for the apo form and holo form are D, A and F of $\gamma_2$ (Chain A) and D, E, B and F loops of $\alpha_1$ (Chain E).

## Zolpidem induced intrinsic conformational changes in the LBD

Agonist induced conformational changes were studied by identifying the breathing motions exerted by the LBD of the pentameric complex. Radius of gyration (Rg) was calculated across the XY plane of the channel, perpendicular to the pore axis (Z axis), to quantify how much each subunit moves inward or outward with reference to the central pore (Fig. 4).

In particular, we noticed that the A chain ($\gamma_2$) subunit displays significant dissimilarity in the nature of their breathing motions. In the holo form, the A chain ($\gamma_2$) subunit undergoes gradual expansion by expanding outward slowly and contracting toward the end. On the contrary, the A chain ($\gamma_2$) subunit of apo form displays an inward breathing motion initially and expands toward the end (Fig. 5).

In general all the subunits in both the forms display a breathing type of motion in small scales, a phenomenon already reported for other related LGIC systems [29]. The resulting pentameric complex of both the apo and the holo forms are pseudo-symmetrical as shown in Fig. 6.

Essential dynamics, an approach used to identify concerted motions by separating irrelevant fluctuations along an MD trajectory, was carried out using a multivariate statistical method termed PCA [17]. This approach consists of diagonalizing the covariance matrix of atomic fluctuations with respect to the average structure, thus obtaining a set of eigenvectors and corresponding eigenvalues, which represent the axes of maximal variance in the protein motion.

On the whole, the first ten eigenvectors extracted are able to explain about 80% of the total relevant motions covered during the MD run for the apo and the holo forms (shown in Fig. 7). Since the first eigenvector corresponds to the most dominant motion exhibited by the system, it is evident from the variance contribution plot that the holo form (34%) displays limited sampling of phase space in comparison to the apo form (38.43%).

The principal modes corresponding to first eigenvector were visualized and represented using porcupine plots (Fig. 8).

The apical part of the C loop of $\alpha 1$ subunit experiences an inward motion upon ligand binding toward the vestibule in the holo form and the F loop experiences an outward

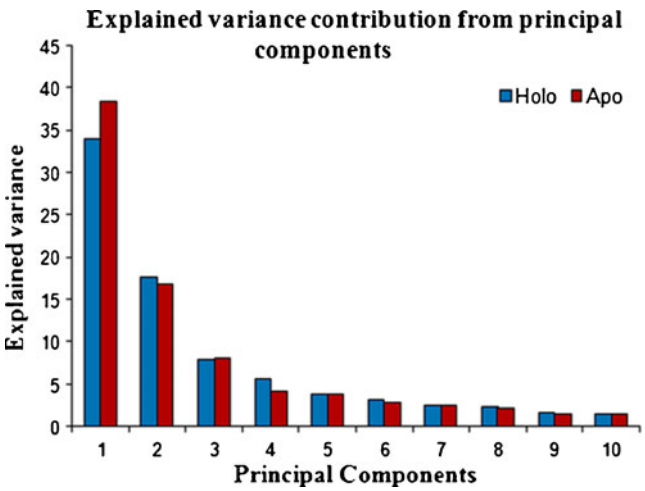



**Fig. 7** Histogram depicting the explained variance contribution of each principal component

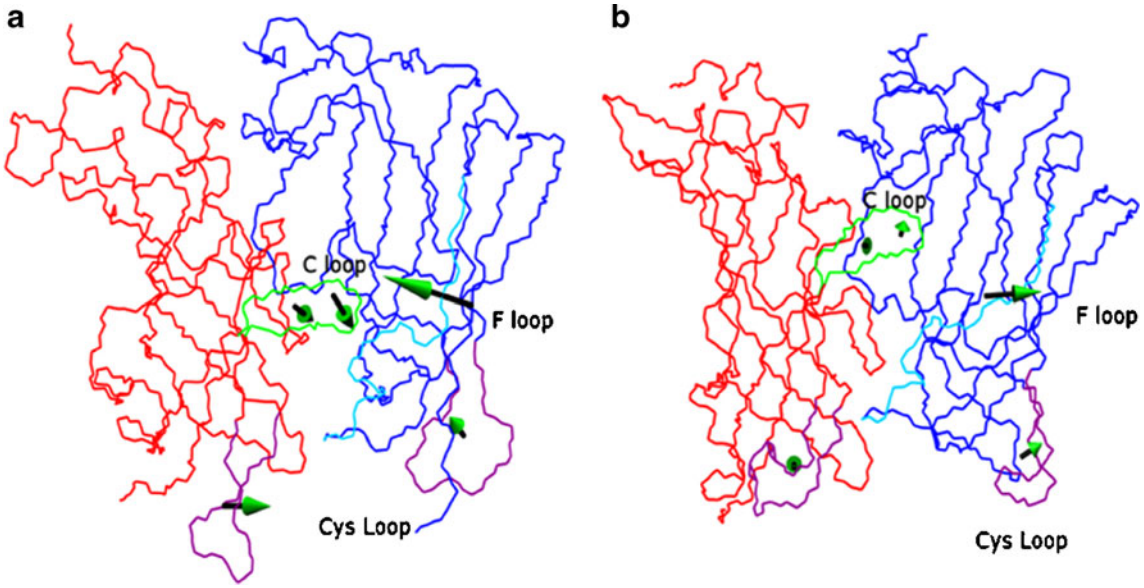



**Fig. 8** Porcupine plot displaying the divergent swaying patterns of the loops evident in the (**a**) holo and apo (**b**) form

motion away from the vestibule in the apo form. This divergent swaying pattern of loop F corroborates with an earlier finding implicating loop F as a structural component that helps to lock the ligand into the agonist binding site [30].

In the holo form the C loop occludes the entrance to the binding site and entraps Zolpidem in to the binding cavity. Further, the C loop of the $\alpha_1$ subunit attains more stability upon complexation with Zolpidem. We also notice that upon binding, the Cys loop of $\alpha$ + subunit and the corresponding Cys loop of the $\gamma$ subunit move concertedly toward each other. As these Cys-loop sway closer, a tightening of the BZ pocket is evident.

From PCA it was found that all the monomeric subunits experience a rotational motion and a global twisting motion envisaged for LGIC, which could be attributed to an amalgamation of all the five individual monomer motions presented at the quaternary level (Fig. 9).

Even in the absence of the TM regions, the motions observed in our study are reminiscence of motions that has been consistently observed for other proteins of the LGIC super family, carried out in the presence of the transmembrane

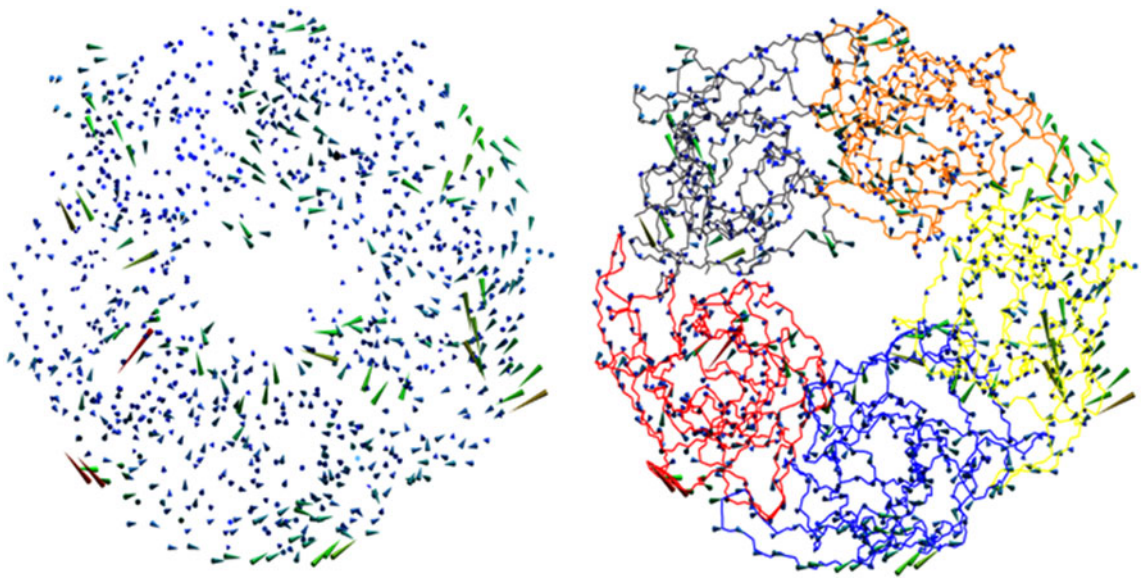


**Fig. 9** Porcupine plot of the principal motions of the holo form calculated using PCA. The right panel shows the directional motion mapped to C$\alpha$ trace of the pentamer

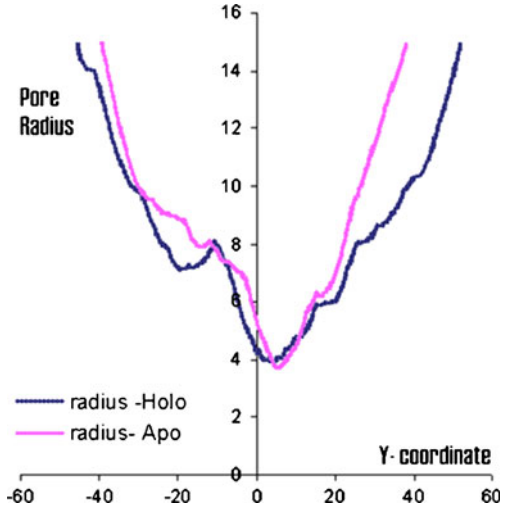



**Fig. 10** Pore diameter calculations for the vestibular region on the averaged apo and holo form

domain (TMD). These observations confide the legitimacy of the evinced motions. In spite of such plausible evidence, we were cautious not to over interpret these motions in light of its simplifications. Studies are currently underway using the entire $GABA_A$ ionophore complex model to provide atomistic details pertaining to channel gating, and will be a part of a future communication.

The vestibular pore diameter was analyzed using the HOLE program. The overall structure of the conductive pore in the vestibular region is well maintained over the course of the MD simulation for both the systems. Comparison of pore radius profiles for the two models (Fig. 10) depicts minimal changes near the pre-transmembrane (pre TM-1) region of the holo form, especially post 15 ns of simulation (data provided vide supplementary information, Fig. SI3). The change should be regarded more as a qualitative tendency. This finding concurs well with experimental results which illustrates that major changes occur only in the transmembrane region that plays a role in gating. The pore diameter calculation for the averaged structure of apo and holo form is shown in Fig. 10.

## Binding mode of zolpidem

Zolpidem remained bound to the BZ site for the whole MD simulation. Aesthetic 2D representation of the non-bonded interactions between ZPM and $GABA_A$ was rendered using Ligplot [31] and is shown in Fig. 11. The imidazopyridine group of ZPM was positioned between the C loop of α1 and the D loop of $\gamma_2$. The aromatic face of the imidazopyridine was found to add more non polar packing by constantly being engaged in a π-π stacking interaction with the aromatic face of Phe77 (D loop of $\gamma_2$ subunit). Interestingly, Phe77 has already been implicated as a major residue for binding [14]. Energy decomposition analysis carried out also revealed that Phe77 has an $E_{int}$ value of −30.091 kcal $mol^{-1}$. The aromatic ring of Tyr159 present in loop B stacks with the pendant phenyl group of ZPM. Removing the aromatic ring of Tyr159 at α1 has been earlier reported to completely abolish Diazepam-mediated potentiation of GABA current in $\alpha_1\beta_2\gamma_2$ receptors [32]. We presumed that the ethyl N, N-dimethyl amide part of ZPM would be highly flexible, on the contrary, it was rather stable due to directional forces exerted by hydrogen bonding between the carbonyl amide and the side chain hydroxyl of Thr206 with an occupancy level of

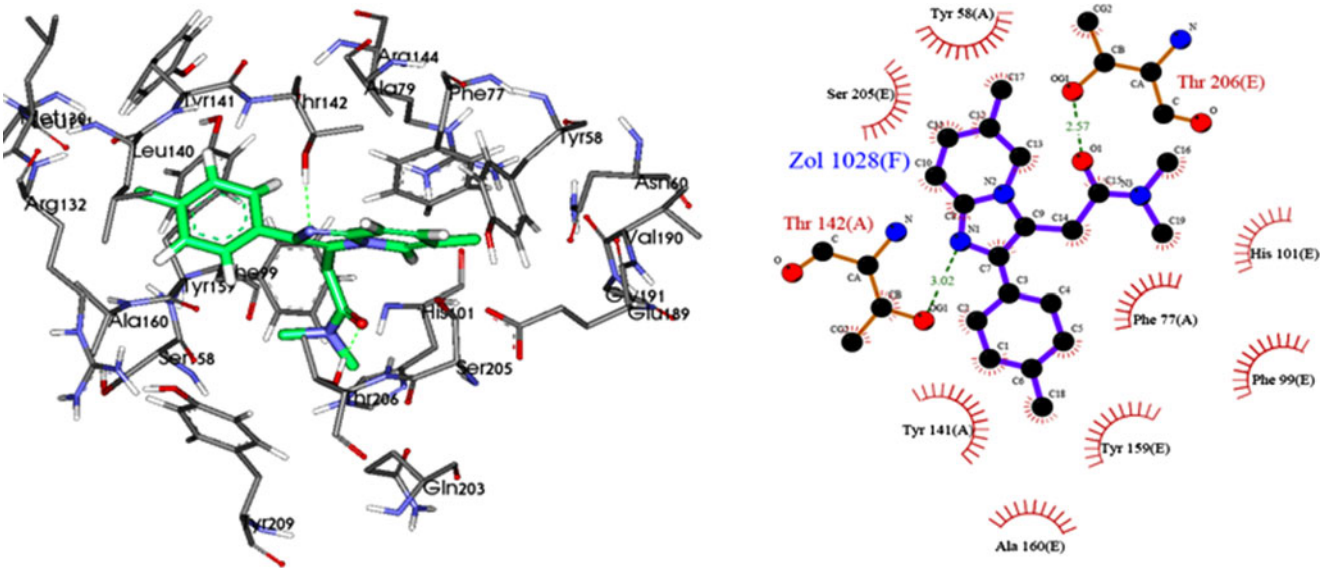



**Fig. 11** The binding mode of Zolpidem, showing the H bonding interactions. For clarity residues within 5Å only are shown. The right panel shows a 2D plot of the interactions, obtained using Ligplot

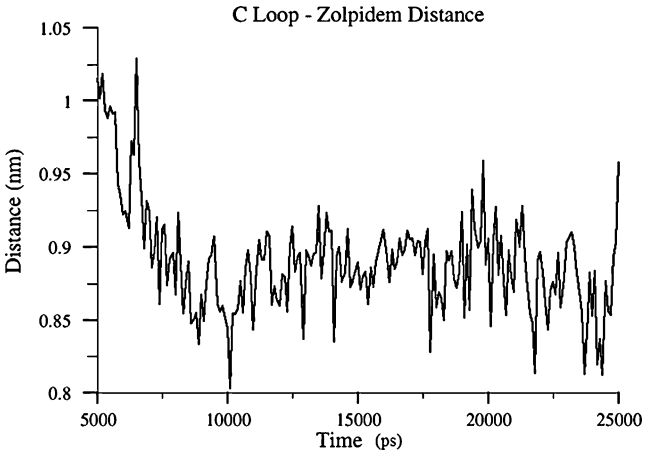



**Fig. 12** Distance between the center of mass of C loop of $\alpha_1$ and Zolpidem plotted as a function of time. This highlights the capping phenomena exerted by C loop

39.26%. This H-bonding opposes a conformational tweak motion which is evident in frames lacking this bonding. The other key BZ site residue His101 of $\alpha_1$ subunit [14], which confers Diazepam sensitivity, was found to be involved in a hydrophobic interaction with one of the N-methyl group of dimethyl amide moiety.

A closed C loop evident from the simulation (Figs. 6, 8) is reminiscent of an earlier observation by Karlin [33], stating an extended loop C being associated with antagonist and a closed C loop associated with agonist. Requirements for this capping motion might well arise from the need to lock the ligand in to the binding site.

Since no major localized reorganizational changes were evident upon complexation in the BZ binding site excluding the C Loop region (Fig. 12), we carried out torsional vibrational motion analysis for those aromatic residues present in the BZ site. The discrepant Chi2 ($\chi^2$) dihedral torsion for aromatic residues Tyr58A, Phe77A and Phe159E implies that that residues Phe77A and Tyr58A had a high *syn* conformation in the apo form, whereas in the holo form, they also had a preference toward high *anti* conformation (vide supplementary information Fig. SI4). We also measured the percentage occupancy of all possible hydrogen bonding interaction atom pairs from the trajectory, which is quantitatively presented in Table 1.

To probe the role of the loops in binding, we carried out an $E_{Interaction}$ analysis between ZPM and the loops. Loops A,B,C of the $\alpha$1 subunit and D,E, F of the $\gamma_2$ subunit anchoring the BZ site were considered for analysis and the results are shown in Table 2.

From the $E_{Interaction}$ values it is obvious that the C loop of $\alpha$1 and the E Loop of $\gamma$2 play critical role in binding. It is equally important to note that a loop can influence ligand binding in multiple ways. Some loops may be directly

**Table 1** Protein–ligand hydrogen bonding interaction pairs extracted from the trajectory

| Donor | Acceptor | Occupancy* |
|---|---|---|
| THR206E-Side | ZPM | 39.26% |
| THR142A-Side | ZPM | 15.28% |
| GLN203E-Side | ZPM | 8.44% |
| SER205E-Side | ZPM | 5.09% |
| THR142A-Main | ZPM | 1.95% |
| ZPM | VAL190A-Main | 1.25% |
| ZPM | TYR58A-Side | 0.55% |
| ZPM | TYR141A-Side | 0.20% |
| ZPM | ARG132A-Side | 0.20% |
| SER04E | ZPM | 0.10% |
| ZPM | THR142A-Main | 0.05% |
| ARG131A -Side | ZPM | 0.05% |
| ZPM | THR206E-Side | 0.05% |
| ZPM | TYR159E-Main | 0.05% |

* Occupancy of H-bonds as % of the simulation time

**Table 2** Energetic contribution of the loops toward binding

| Loop | Residues | $E_{Interaction}$ (kcal/mol) |
|---|---|---|
| A Loop $\alpha_1$ | 99-103 | 0.769084 |
| B Loop $\alpha_1$ | 156-163 | -8.95672 |
| C Loop $\alpha_1$ | 202-208 | -21.8401 |
| D Loop $\gamma_2$ | 74-80 | -10.809 |
| E Loop $\gamma_2$ | 127-138 | -19.0193 |
| F Loop $\gamma_2$ | 186-192 | -8.96245 |

**Table 3** Residue-wise interaction energy contribution analysis

| Residue | $E_{Ele}$ | $E_{Vdw}$ | $E_{Int} = E_{Ele} + E_{Vdw}$ (kcal/mol) |
|---|---|---|---|
| TYR58A | -1.860 | -12.51 | -14.37 |
| ASN60A | -1.523 | -5.00 | -6.523 |
| PHE77A | -0.528 | -30.0961 | -30.6241 |
| ARG 132A | -11.36 | -6.61 | -17.97 |
| THR142A | -20.183 | -20.09 | -40.273 |
| VAL190A | -4.289 | -3.78 | -8.069 |
| ASP192A | -3.89 | -0.77 | -4.66 |
| PHE99E | -1.482 | -13.34 | -14.822 |
| HIS101E | 0.609 | -6.385 | -5.776 |
| TYR159E | -0.979 | -21.923 | -22.902 |
| ALA160E | 0.50 | -4.67 | -4.17 |
| GLN203E | -7.296 | -6.44 | -13.736 |
| SER204E | -1.252 | -3.05 | -4.302 |
| SER205E | -8.59 | -7.79 | -16.38 |
| TYR209E | -11.51 | 0.02 | -11.49 |
| THR206E | -30.05 | -10.79 | -40.84 |

involved in binding with ligand, some may be important for maintaining the structural integrity of the binding site, and others may mediate the necessary allosteric effects. Hence one may conclude that the loop C of $\alpha 1$ and loop E of $\gamma 2$ would participate in direct non-bonded interaction with ZPM. Considering the divergent swaying pattern of loop F and its relevance for benzodiazepine modulation in an earlier experimental study [34], we presume that loop F could be a key loop, crucial for modulation.

### Interaction energy decomposition analysis

Energy contribution of each binding site residue was studied using interaction energy decomposition analysis. The interaction energies were calculated as a summation of electrostatic ($E_{Coul}$) and van der Waals ($E_{LJ}$) modeled as Coulombic and Lennard-Jones parameters, respectively, between the ligand and the protein as shown in Table 3.

Among the binding site residues $\gamma_2$Phe77, $\gamma_2$Thr142, $\alpha_1$Tyr159 and $\alpha_1$Thr206 display strong interactions with Zolpidem. Mutagenesis, radioligand binding and studies using transgenic knockout mice have already implicated $\gamma_2$ Phe77 and $\gamma_2$Thr142 as critical residues that influence the efficacy of classical BZDs as well as Zolpidem [13, 14]. However, the importance of the other ascribed residues is yet to be ascertained experimentally. Analysis of the electrostatic and non-polar contribution reveals that non polar/ van der Waals contribution plays a major role and the affinity of Zolpidem is largely dominated by shape complementarity and hydrophobic interactions.

## Conclusions

In summary, a detailed perspective of the binding mode of Zolpidem to $GABA_A\alpha_1/\gamma_2$ together with the conformational changes, induced upon agonist binding, has been disclosed. The ability of the simulation in reproducing a realistic binding mode, consistent with experimental data, paves the way for rationalizing a binding mode for Zolpidem. Energy decomposition analysis unraveled the role of many critical residues contributing toward binding. Analysis of the energy terms indicates that association is largely driven by favorable van der Waals interactions. The global motions evinced from the simulation further consolidate the validity of the simulation as they are common to many other ligand-gated ion channels. The collective finding obtained from MD simulations provides insight into the binding mode of Zolpidem, and the subsequent perturbations that occur upon agonist binding.

**Acknowledgments** R.S.K.V thanks Council of Scientific and Industrial Research (CSIR), New Delhi for a Research Associateship. N.G. is thankful to CSIR Mission Mode Program (CMM-0017) for providing financial assistance. Software support granted by OpenEye is greatly acknowledged. We acknowledge Mr. Robert Schulz for providing scripts to generate Porcupine Plots.

## References

1. Kuffler SW, Edwards C (1958) Mechanism of gamma amino-butyric acid (GABA) action and its relation to synaptic inhibition. J Neurophysiol 21:589–610
2. Schofield PR, Darlison MG, Fujita N, Burt DR, Stephenson FA, Rodriguez H, Rhee LM, Ramachandran J, Reale V, Glencorse TA et al (1987) Sequence and functional expression of the GABA A receptor shows a ligand-gated receptor super-family. Nature 328:221–227
3. Smith GB, Olsen RW (1995) Functional domains of GABAA receptors. Trends Pharmacol Sci 16:162–168
4. Clayton T, Chen JL, Ernst M, Richter L, Cromer BA, Morton CJ, Ng H, Kaczorowski CC, Helmstetter FJ, Furtmuller R, Ecker G, Parker MW, Sieghart W, Cook JM (2007) An updated unified pharmacophore model of the benzodiazepine binding site on gamma-aminobutyric acid(a) receptors: correlation with comparative models. Curr Med Chem 14:2755–2775
5. Ernst M, Bruckner S, Boresch S, Sieghart W (2005) Comparative models of GABAA receptor extracellular and transmembrane domains: important insights in pharmacology and function. Mol Pharmacol 68:1291–1300
6. Curtis DR (1978) In: Krogsgaard-Larsen P, Scheel-Kruger J, Kofoed H (eds) GABA Neurotransmitters: Pharmacochemical Biochemical and Pharmacological Aspects, Munksgaard, Kopenhagen
7. Darcourt G, Pringuey D, Salliere D, Lavoisy J (1999) The safety and tolerability of zolpidems an update. J Psychopharmacol 13:81–93
8. Sanna E, Busonero F, Talani G, Carta M, Massa F, Peis M, Maciocco E, Biggio G (2002) Comparison of the effects of zaleplon, zolpidem, and triazolam at various GABA(A) receptor subtypes. Eur J Pharmacol 451:103–110
9. Low K, Crestani F, Keist R, Benke D, Brunig I, Benson JA, Fritschy JM, Rulicke T, Bluethmann H, Mohler H, Rudolph U (2000) Molecular and neuronal substrate for the selective attenuation of anxiety. Science 290:131–134
10. Ghoshal N, Vijayan RSK (2010) Pharmacophore models for GABAA modulators: implications in CNS drug discovery. Expert Opin Drug Discov 5:441–460
11. Hanson SM, Morlock EV, Satyshur KA, Czajkowski C (2008) Structural requirement for eszopiclone and zolpidem binding to the gamma-aminobutyric acid type-A (GABAA) receptor are different. J Med Chem 51:7243–7252
12. Renard S, Olivier A, Granger P, Avenet P, Graham D, Sevrin M, George P, Besnard F (1999) Structural elements of the gamma-aminobutyric acid type A receptor conferring subtype selectivity for benzodiazepine site ligands. J Biol Chem 274:13370–13374
13. Buhr A, Baur R, Sigel E (1997) Subtle changes in residue 77 of the gamma subunit of alpha1beta2gamma2 GABAA receptors drastically alter the affinity for ligands of the benzodiazepine binding site. J Biol Chem 272:11799–11804
14. Mihic SJ, Whiting PJ, Klein RL, Wafford KA, Harris RA (1994) A single amino acid of the human gamma-aminobutyric acid type A receptor gamma 2 subunit determines benzodiazepine efficacy. J Biol Chem 269:32768–32773
15. Brejc K, van Dijk WJ, Klaassen RV, Schuurmans M, van Der Oost J, Smit AB, Sixma TK (2001) Crystal structure of an ACh-binding protein reveals the ligand-binding domain of nicotinic receptors. Nature 411:269–276

16. Ernst M, Brauchart D, Boresch S, Sieghart W (2003) Comparative modeling of GABA(A) receptors: limits, insights, future developments. Neuroscience 119:933–943
17. Amadei A, Linssen AB, Berendsen HJ (1993) Essential dynamics of proteins. Proteins 17:412–425
18. Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. J Mol Biol 267:727–748
19. Van der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ (2005) GROMACS: fast, flexible, and free. J Comput Chem 26:1701–1718
20. Schuttelkopf AW, van Aalten DM (2004) PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. Acta Crystallogr D Biol Crystallogr 60:1355–1363
21. TSAR, Version 3.0 (2007) Accelrys Inc, San Diego, CA
22. Jakalian A, Jack DB, Bayly CI (2002) Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. J Comput Chem 23:1623–1641
23. Berendsen HJC, Postma JPM, van Gunsteren WF, Dinola A, Haak JR (1984) Molecular- dynamics with coupling to an external bath. J Chem Phys 81:3684–3690
24. Parrinello M, Rahman A (1981) Polymorphic transitions in single crystals: A new molecular dynamics method. J Appl Phys 52:7182–7190
25. Darden T, York D, Pedersen L (1993) Particle mesh Ewald: An N · log(N) method for Ewald Sums in large systems. J Chem Phys 98:10089–10092
26. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM (1997) LINCS: A linear constraint solver for molecular simulations. J Comput Chem 18:1463–1472
27. Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. J Mol Graph 14:33–38
28. Smart OS, Neduvelil JG, Wang X, Wallace BA, Sansom MS (1996) HOLE: a program for the analysis of the pore dimensions of ion channel structural models. J Mol Graph 14:354–360
29. Henchman RH, Wang HL, Sine SM, Taylor P, McCammon JA (2005) Ligand-Induced conformational change in the alpha7 nicotinic receptor ligand binding domain. Biophys J 88:2564–2576
30. Khatri A, Sedelnikova A, Weiss DS (2009) Structural rearrangements in loop F of the GABA receptor signal ligand binding, not channel activation. Biophys J 96:45–55
31. Wallace AC, Laskowski RA, Thornton JM (1995) LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. Protein Eng 8:127–134
32. Amin J, Brooks-Kayal A, Weiss DS (1997) Two tyrosine residues on the alpha subunit are crucial for benzodiazepine binding and allosteric modulation of gamma-aminobutyric acid A receptors. Mol Pharmacol 51:833–841
33. Karlin A, Akabas MH (1995) Toward a structural basis for the function of nicotinic acetylcholine receptors and their cousins. Neuron 15:1231–1244
34. Padgett CL, Lummis SC (2008) The F-loop of the GABA A receptor gamma2 subunit contributes to benzodiazepine modulation. J Biol Chem 283:2702–2708

J Mol Model (2012) 18:1355–1364
DOI 10.1007/s00894-011-1167-4

ORIGINAL PAPER

# The impact of *Trichoderma reesei* Cel7A carbohydrate binding domain mutations on its binding to a cellulose surface: a molecular dynamics free energy study

**Tong Li · Shihai Yan · Lishan Yao**

Received: 2 March 2011 / Accepted: 27 June 2011 / Published online: 15 July 2011


**Abstract** A critical role of the Family 7 cellobiohydrolase (Cel7A) carbohydrate binding domain (CBD) is to bind to a cellulose surface and increase the enzyme concentration on the surface. Several residues of *Trichoderma reesei* Cel7A CBD, including Y5, N29, Y31, Y32 and Q34, contribute to cellulose binding, as revealed by early experimental studies. To investigate the interactions between these important residues and cellulose, we applied a thermodynamic integration method to calculate the cellulose–Cel7A CBD binding free energy changes caused by Y5A, N29A, Y31A, Y32A and Q34A mutations. The experimental binding trend was successfully predicted, proving the effectiveness of the complex model. For the two polar residue mutants N29A and Q34A, the changes in the electrostatics are comparable to those of van der Waals, while for three Y to A mutants, the free energy differences mainly come from van der Waals interactions. However, in both cases, the electrostatics dominates the interactions between individual residues and cellulose. The side chains of these residues are rigidified after the complex is formed. The binding free energy changes for the two mutants Y5W and Y31W were also determined, and for these the van der Waals interaction was strengthened but the electrostatics was weakened.

**Keywords** Cellobiohydrolase · Cellulose · Thermodynamic integration · Molecular dynamics · Free energy

**Electronic supplementary material** The online version of this article (doi:10.1007/s00894-011-1167-4) contains supplementary material, which is available to authorized users.

T. Li · S. Yan · L. Yao (✉)
Lab of Biofuels, Qingdao Institute of Bioenergy and Bioprocess Technology, Chinese Academy of Sciences,
Qingdao 266101, China
e-mail: yaols@qibebt.ac.cn

## Introduction

Biofuels produced from biomass are a promising and environmentally friendly alternative to fossil fuels, but the process of producing biofuels faces substantial challenges, mainly due to its high cost compared to traditional energy production. One approach is to hydrolyze cellulose—the main polymer in biomass—to sugars by cellulases, which are then fermented to ethanol. Cellulose is the most abundant bioresource on Earth and the major structural component of plant cell walls. In nature, cellulose molecular chains are linear polymers composed of β-1,4 linked D-glucose units.

The enzymes involved in the hydrolysis of cellulose consist of endoglucanases (EG), exoglucanases or cellobiohydrolases (CBH), and β-glucosidases. Among them, endoglucanases are responsible for breaking down the intramolecular β-1,4 glucosidic bonds of cellulose chains to produce new chain ends; exoglucanases hydrolyze cellulose chains from their ends and release soluble cellobiose; and β-glucosidases catalyze cellobiose to glucose [1, 2]. The enzymes secreted by the filamentous fungus *Trichoderma reesei* (*T. reesei*), which are highly effective against native and crystalline cellulose [3], are well studied. *T. reesei* produces a mixture of cellulases consisting of two CBHs (CBH I or Cel7A and CBH II or Cel6A) and several EGs [4, 5]. The enzyme Cel7A is composed of a large catalytic domain (CD) and a small cellulose-binding domain (CBD) connected via a glycosylated peptide linker [6]. The CD of Cel7A, containing a 50 Å tunnel-shaped active site, binds a cellulose chain and hydrolyzes the glycosidic bond to release cellobiose [3, 7–9].

The three-dimensional solution structure of the Cel7A CBD from *T. reesei* has been determined by nuclear magnetic resonance (NMR) spectroscopy [10], which

revealed that several aromatic residues, Y5, Y31 and Y32, form a hydrophobic surface. However, there is no experimental structure of the Cel7A– or CBD–cellulose complex due to the insoluble nature of cellulose. The experimental study by Linder and coworkers [4] has shown that these three tyrosine residues are critical for binding to crystalline cellulose, while N29 and Q34 also contribute but are less important. Recently, the CBD cellulose complex has been modeled in order to investigate the binding mechanism by molecular dynamics (MD) simulations. It has been suggested that CBD may assist with the recognition of the cellulose chain end and facilitate the processivity of Cel7A [11–13]. Beckham et al. predicted that electrostatic interactions dominate the energy changes in the Cel7A CBD processivity [12]. However, it is very important to validate the CBD–cellulose complex model against experimental data.

Free energy methods provide a rigorous way to obtain important and readily measurable thermodynamic observables, such as binding free energy, from MD simulations [14, 15]. Chemical alchemy is an important and accurate class of free energy calculations. This approach involves simulating a transformation of one chemical system to a different one along a nonphysical path to compute the free energy change associated with the transformation [14]. Among the alchemical methods, thermodynamic integration (TI) [15–19] has grown in popularity over the last few years. A parameter $\lambda$, ranging from zero to one, is introduced to define the intermediate states, with $\lambda=0$ (1) corresponding to the initial (final) state. A series of simulations are performed for all the states, and then the partial derivative of the Hamiltonian over $\lambda$ is estimated for each state and integrated over $\lambda$ to yield the free energy change of the transformation. Although the computational demand is still large, TI has come of age with increased computer power and the wide use of parallel computers. One important application of the TI method is the study of the consequences of protein side-chain mutations, which has shown promising results [16, 20–22].

In this work, we calculated the relative binding free energy differences between the mutants (N29A, Q34A, Y5A, Y31A and Y32A) and wild-type CBD that bind to cellulose. Qualitative agreement between the experimental and computed free energy differences was observed, validating the effectiveness of the CBD cellulose-binding model. The absolute binding free energy calculation of CBD involves the annihilation of CBD on the cellulose surface and in water, which is a formidable task, so it was not pursued in this work. On the other hand, the absolute binding free energy was also difficult to determine (if not impossible) experimentally because of the insolubility of cellulose.

## Methods

### Preparation of the system

The starting coordinates for the CBD of Cel7A were obtained from the NMR structure [10] (PDB code 1CBH). The cellulose slab was modeled based on the cellulose Iβ crystal structure presented by Nishiyama et al. [23]. As shown in Fig. 1, the cellulose is three layers deep with five chains in each layer, and each chain includes 12 glucose units. Consistent with early studies [12, 13], the CBD is aligned on a cellulose surface with the aromatic residues Y5, Y31 and Y32 within 3 Å of the cellulose surface simultaneously, and the Y5, Y31 and Y32 rings are stacked over the glucose rings of cellulose. The complex model is stable during the MD equilibration (see the section "Results and discussion").

### Molecular dynamics simulations

The simulations were performed using the SANDER module in the AMBER 10 program [24], with the AMBER99sb [25] force field used for the protein and the GLYCAM_06 force field [26] [27] for the cellulose. The complex was soaked in a rectangular box with 12 Å TIP3P water molecules on each side using the LEAP module in AMBER. $Na^+$ ions were added to neutralize the system. These yielded about 35,000 atoms for the simulation system. The 1–4 electrostatic (SCEE) and nonbonded (SCNB) scaling factors were set to their default values, with SCEE=1.2 and SCNB=2.0. It was realized later that the SCEE and SCNB scaling factors of cellulose should be set to one to yield the correct rotamer population, which is not done properly in AMBER 10. However, in our work, cellulose is restrained by using a harmonic potential with a force constant of 2.0 kcal $mol^{-1}$ $A^{-2}$ on the C1 and C4 atoms, essentially removing the dihedral fluctuations along the C1–O4 and C4–O4 bonds.

Energy minimization was performed to remove bad contacts, and then the system was heated from 0 K to

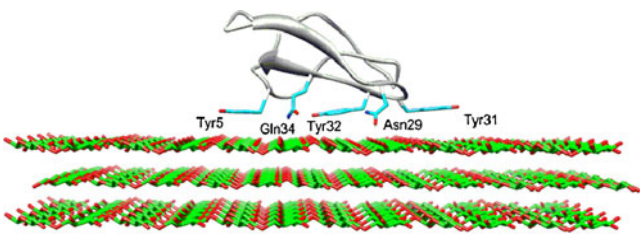



**Fig. 1** Structural model of the CBD and cellulose complex. The side chains of residues Y5, N29, Y31, Y32 and Q34 are shown in stick representation, with the carbon atoms colored in *cyan* and the oxygen atoms in *red*. The cellulose was modeled by three layers of cellulose chains

300 K in a 50 ps constant volume MD simulation, after which a 1 ns NPT simulation was run to equilibrate the water so that the water density reached 1.0 kg/L. The particle mesh Ewald (PME) method [28] was applied to calculate long-range electrostatics interactions. The SHAKE method [29] was applied to constrain all the covalent bonds involving hydrogen atoms. Periodic boundary conditions were applied. A nonbonded pair list cutoff of 12.0 Å was used. The time step used to integrate the equations of motion was 0.002 ps. Coordinates were saved every 0.4 ps along the 20 ns production run, and the trajectories were analyzed with the PTRAJ module of AMBER.

## Free energy calculation by thermodynamics integration

According to the thermodynamic cycle, four states should be considered during the calculation of the relative binding free energy. The relative binding free energy difference $\Delta\Delta G_{\text{bind}}$ between the wild-type CBD cellulose complex (WT-CBD-Cel) and a mutant CBD cellulose complex (MU-CBD-Cel) could be obtained from $\Delta G_{\text{B}} - \Delta G_{\text{A}}$ (Fig. 2). However, as discussed above, directly calculating the binding free energies $\Delta G_{\text{A}}$ and $\Delta G_{\text{B}}$ is a formidable task, which can be avoided by designing a thermodynamic cycle as shown in Fig. 2. We can get $\Delta\Delta G_{\text{bind}}$ by calculating the difference in the free energy changes for the free ($\Delta G_1$) and bound ($\Delta G_2$) states.

$$\Delta\Delta G_{\text{bind}} = \Delta G_{\text{B}} - \Delta G_{\text{A}} = \Delta G_2 - \Delta G_1 \tag{1}$$

$\Delta G_1$ ($\Delta G_2$) corresponds to the free energy change in the alchemical process of converting the wild type to the mutant residue in the absence (presence) of the cellulose. It can be computed using the following equation:

$$\Delta G_{\text{W}\to\text{M}} = \int_0^1 \langle \partial V/\partial\lambda \rangle_\lambda d\lambda, \tag{2}$$

where the subscripts W and M denote the wild type and mutant state, $V$ is the potential energy, and $\lambda$ is a parameter that ranges from 0 (wild type) to 1 (mutant). The binding affinities of the five mutants, including Y5A, N29A, Y31A, Y32A and Q34A, to cellulose were determined experimentally. The free energy calculations for the mutation of each of these five residues to an alanine were performed by dividing the process into three steps, as commonly employed in TI simulations. The first step is to remove the partial charges of the side chain atoms of the residue subject to mutation, except for the $C_\beta$ atom and the two $H_\beta$ hydrogens that exist in the alanine residue. In step two, the van der Waals (VDW) interactions with the atoms whose charges are removed in step one are eliminated (this is described in more detail below). In step three, the third $H_\beta$ hydrogen atom attached to the $C_\beta$ of the alanine residue gets the partial charge back. During the VDW transformation in step two, a linear mixing soft core potential is applied in which nonbonded VDW interactions are represented by a $\lambda$-dependent modified LJ equation [14],

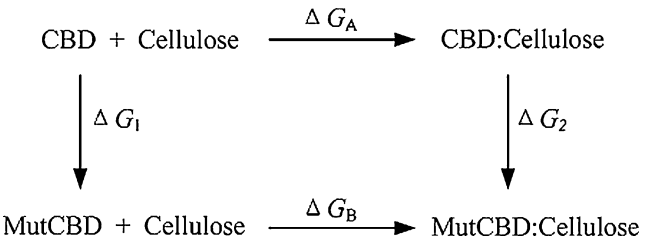



**Fig. 2** Thermodynamic cycle for calculating the binding free energy difference between the wild-type and mutant CBD binding to cellulose

$$V_{V0,disappearing} = 4\varepsilon(1-\lambda)\left[\frac{1}{\left[\alpha\lambda + \left(\frac{r_{ij}}{\sigma}\right)^6\right]^2} - \frac{1}{\alpha\lambda + \left(\frac{r_{ij}}{\sigma}\right)^6}\right], \tag{3}$$

where $\varepsilon$ is the well depth and $\sigma$ is the collision diameter, $\alpha$ is an adjustable constant which is set to 0.5 in this study, and $r_{ij}$ is the distance between atoms $i$ and $j$. Soft core scaling provides smooth $\lambda$ curves that are best suited for commonly used numerical integration schemes [14]. Equation 3 is the potential function used for atoms that disappear during a VDW transformation. When $\lambda = 0$, it has the same form as the normal LJ equation; for $\lambda$ values close to 1, it provides a smooth interaction function that allows other atoms to be situated on top of a vanishing atom, with a finite energy penalty that goes to zero at $\lambda = 1$ [14].

The last configuration of the CBD–cellulose complex from the 20 ns simulation was selected as the starting structure for the alchemical simulations to calculate $\Delta G_2$. A CBD system without cellulose was prepared with a size of ~11000 atoms for the $\Delta G_1$ calculations. In general, for each $\lambda$ in the three-step transformation, we carried out a 100 ps constant-volume equilibration run followed by a 400 ps production run where the energy derivative $\partial V/\partial\lambda$ for each $\lambda$ was collected. $\lambda$ values were set to $0.025n$ ($n$=1, 2, …, 39) in each of the three steps; therefore, there were a total of 117 windows for each mutation. The free energy change associated with each of the three steps was calculated by numerically integrating the derivative data. The integration was done using an in-house perl script. The errors were estimated by splitting the trajectories into four equal blocks and calculating the corresponding free energy differences.

Besides the five single mutations (Y5A, N29A, Y31A, Y32A and Q34A), we also performed two more alchemy simulations (Y5W and Y31W) to determine the binding affinity change after tryptophan substitution. Presumably,

the tryptophan ring prefers to be parallel to the cellulose surface to maximize the VDW interaction, similar to the tyrosine phenol ring conformations on a cellulose surface. From this reasoning, we designed two conformations with ($\chi_1$, $\chi_2$) dihedrals equal to (−36°, 113°) and (−36°, -66°) for W5, and (−69°, 83°) and (−69°, -96°) for W31, respectively (see Fig. 3), as the starting structures, but they were not restrained during the TI simulations. In total, there were nine TI simulations with a total simulation time of ~1 μs.

### Side-chain conformational entropy

The side-chain conformational entropy, $S_{conf}$, was calculated using T-Analyst [30]. $S_{conf}$ is defined as

$$S_{conf} = -R\sum p_i \ln(p_i), \quad (4)$$

where $R$ is the gas constant, $p_i$ is the probability distribution of angle $i$, which refers to the side chain ($\chi_1$, $\chi_2$) of the mutated residues in this work. The distributions were built from 20 ns molecular dynamics trajectories of free and cellulose-bound CBD. For each residue mutation, the entropy change was estimated using the T-Analyst program, and the entropic contribution to the binding free energy was calculated with the temperature set to 300 K. The bin size for each $p_i$ was 5° for side-chain dihedrals [30].

## Results and discussion

### 20 ns MD

A putative binding position for CBD was proposed where the aromatic rings of the three tyrosine residues are nearly collinearly aligned along the cellulose fiber axis [31]. The close contact of the tyrosine and sugar rings is in concert with the fact that three tyrosine residues are critical in the CBD cellulose binding based on the experimental study. The Cel7A processivity action against insoluble highly crystalline cellulose is unambiguously toward glucose nonreducing ends from reducing ends [9], which effectively restrains the alignment of the CBD relative to cellulose. Figure 1 shows the modeled initial structure of the CBD–cellulose complex for the 20 ns simulation. During the simulation, the RMSD values of all atoms compared to the starting structure reach ~0.8 Å after several ps and are maintained at around 1 Å without large deviations (Fig. 4a). In addition, the distance between the center of mass (COM) of the CBD and the COM of cellulose remains stable (Fig. 4b). Figure 4c shows the projection of the CBD COM on the cellulose surface during the 20 ns MD simulation, which is narrowly distributed. The results indicate that the system is equilibrated after the 20 ns simulation and can serve as the input geometry for the TI simulations.

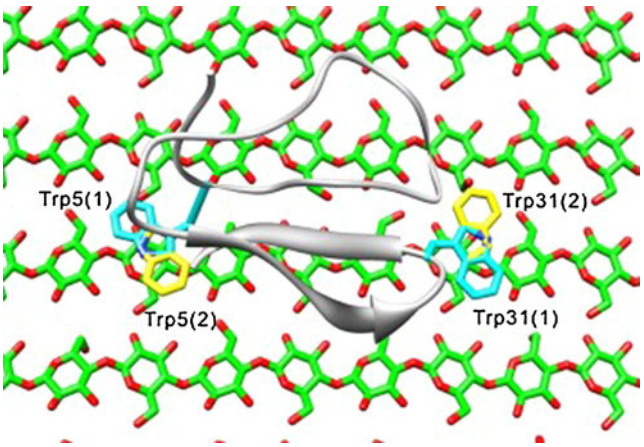



**Fig. 3** Two possible configurations of Y5W and Y31W mutants of the CBD. The *numbers in parentheses* designate different configurations, with ($\chi_1$, $\chi_2$) equal to (−36°, 113°) and (−36°, -66°) for the Y5 configuration 1 (in *cyan*) and 2 (in *yellow*), respectively. For Y31, the corresponding two dihedrals are (−69°, 83°) and (−69°, -96°)

### Free energy calculations

The mutations of functionally important amino acids such as Y5, N29, Q34, Y31 and Y32 in the CBD domain to alanine have been shown experimentally to reduce the binding affinity and activity of Cel7A to crystalline cellulose [4, 32, 33]. To elucidate the fundamental principles governing the interaction between the CBD and cellulose, a three-step TI method was used to calculate the binding free energy differences between the WT-CBD-Cel and MU-CBD-Cel (see "Methods"). The calculations from the first and third steps are related to electrostatic effects, and those from the second step are the VDW interactions. In the second step, soft core Lennard–Jones (LJ) potentials were used to provide smooth $\lambda$ curves that are best suited to commonly used numerical integration schemes [14].

Plots of $\partial V/\partial\lambda$ versus $\lambda$ for steps one and two are depicted in Fig. 5. Switching on the charge of the third hydrogen attached to $C_\beta$, as done in step three, provides a significantly smaller contribution to the free energy changes for all five mutations, so the corresponding $\partial V/\partial\lambda$ plot is not shown here. The $\partial V/\partial\lambda$ values versus time for step two at $\lambda$=0.125 are shown in Fig. S1 as a typical energy derivative plot (see the "Electronic supplementary material"), and this indicates that the systems are equilibrated during the production period. It is worth noting that in the VDW soft core alchemical process, a Y to A mutation generally displays larger $\partial V/\partial\lambda$ fluctuations than a Q or N to A mutation does, especially for the CBD–cellulose complex (Fig. 5). To further investigate the cause of the energy derivative fluctuations, we selected two neighboring windows with the largest $\partial V/\partial\lambda$ difference (13.8 kcal

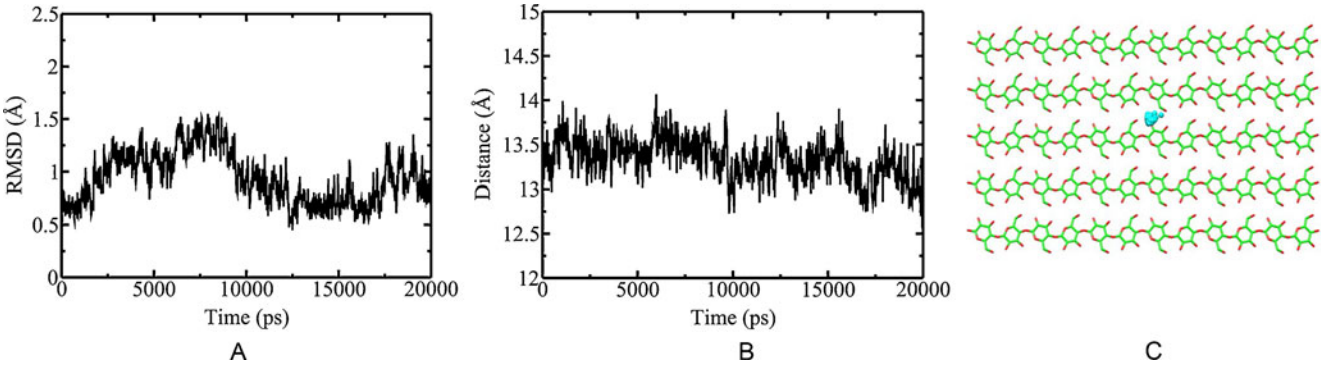



**Fig. 4** **a** All atom RMSDs from the initial configuration, **b** distance between the center of mass of the CBD and cellulose, and **c** projection of the center of mass of CBD onto the cellulose surface during the 20 ns simulation of the WT-CBD-Cel

mol$^{-1}$), corresponding to $\lambda = 0.875$ and 0.900 of the CBD–cellulose complex Y32A VDW alchemy step. It can be seen (Fig. S2) that the side-chain $\chi_2$ distribution is quite broad (ranging from −180° to 180) for both windows, suggesting that tyrosine ring rotation occurs in the simulations, which can cause a dramatic fluctuation in the VDW interaction between the tyrosine and a glucose ring of the cellulose chain. Furthermore, a shift of the $\chi_1$ distribution was observed for the $\lambda = 0.875$ window compared to that for $\lambda = 0.900$ (Fig. S2). Thus, from a structural point of view, the bulky size of tyrosine and a slightly different local environment around the mutated site are the main cause of the large $\partial V/\partial\lambda$ fluctuation in the Y to A VDW alchemical process.

Total free energies were obtained by summing the contributions of all three steps. Table 1 shows the calculated and experimental relative free energy differences between the WT-CBD-Cel and MU-CBD-Cel. According to the TI simulations, the calculated $\Delta\Delta G_{cal}$ values for Q34A, N29A, Y5A, Y31A and Y32A are 3.42, 3.68, 5.40, 6.35 and 10.76 kcal mol$^{-1}$, respectively, suggesting that all five of these residues play roles in the binding to the cellulose surface. The corresponding standard deviations were 0.57, 0.74, 0.82, 0.59 and 0.87 kcal mol$^{-1}$, respectively.

The three aromatic residues Y5, Y31 and Y32 show larger binding free energy contributions than the residues N29 and Q34 do, which is in good agreement with the experimental results [4]. Among the five mutations, Y32A

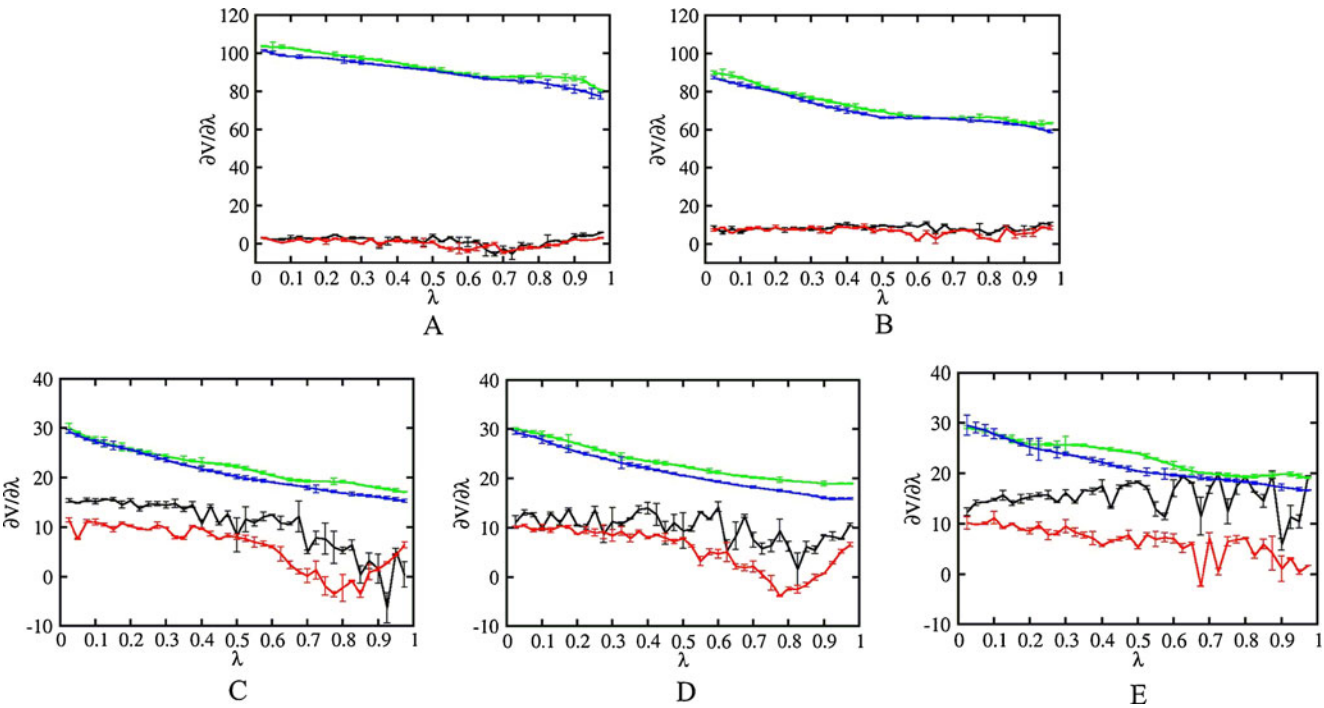



**Fig. 5** $\partial V/\partial\lambda$ for the five single mutations as a function of $\lambda$, for **a** N29A, **b** Q34A, **c** Y5A, **d** Y31A, and **e** Y32A, from the electrostatic (VDW) transformations of the free and bound states of the CBD colored in *blue and green* (*red and black*), respectively. The *error bars* are the standard deviations of the corresponding energy derivatives

**Table 1** Comparison of the predicted and experimental CBD–cellulose binding free energy differences of the single mutations. All values are in kcal mol$^{-1}$

| Mutation | $\Delta\Delta G_{vdw}$ | $\Delta\Delta G_{ele}$ | $\Delta\Delta G_{cal}$ | $\Delta\Delta G_{exp}$[a] |
|---|---|---|---|---|
| Q34A | 1.65±0.43 | 1.77±0.58 | 3.42±0.57 | 1.17 |
| N29A | 1.35±0.41 | 2.32±0.67 | 3.68±0.74 | 0.57 |
| Y5A | 4.13±0.73 | 1.27±0.65 | 5.40±0.82 | >1.74[b] |
| Y31A | 4.39±0.64 | 1.96±0.75 | 6.35±0.59 | 1.74 |
| Y32A | 8.97±0.24 | 1.78±0.77 | 10.76±0.87 | >1.74[b] |

[a] The free energy of binding was deduced from the equation: $\Delta\Delta G = -RT\ln(K_{mut}/K_{wt})$ where $K_{mut}$ and $K_{wt}$ are the partition coefficients of the mutant and wild type

[b] The values cannot be extrapolated from the adsorption isotherms, but they should be higher than 1.74 kcal mol$^{-1}$, according to the adsorption isotherms [4]

shows the largest binding free energy loss. The NMR structure of the CBD suggests that the substitution at position 32 could cause significant disruption of the flat CBD cellulose contact face due to the close proximity of the neighboring side chains, which might account for the binding importance of this residue. In general, the absolute values of the free energy changes calculated from TI simulations are higher than the experimental values extrapolated from the adsorption isotherms. This may be due to a number of factors. (1) The binding of CBD to cellulose is very complex. Lehtio and coworkers [34] showed that the primary binding site is the (110) face of cellulose, which is modeled here. However, as it is stated by Lehtio that "it cannot be ruled out that the initial, perhaps lower affinity, binding occurs in the (100) and (010) faces followed by tighter binding at the (110) plane as the hydrolysis proceeds," the binding on other faces is most likely less sensitive to the mutations. (2) The cellulose model built here has an ideal flat surface composed of polyglucan chains; this is restrained during the TI simulations to simplify the system and yield better convergence; however, in reality, the surface may twist due to molecular motions. The restraints on the C1 and C4 atoms also limit the conformational sampling of glucose units in contact with the CBD side chains. This effect was studied by removing the restraints, except for those on the terminal glucose units. We repeated the TI calculations for Q34A and Y5A. The calculated $\Delta\Delta G$ values decrease slightly, by 0.48 kcal mol$^{-1}$ for the former and 1.08 kcal mol$^{-1}$ for the latter—comparable to the uncertainties of the calculations (Table 1). This effect alone is unlikely to alter the predicted binding trend or account for the differences between experimental and computational $\Delta\Delta G$ values. (3) The calculated binding free energies may also depend on the cellulose binding surface size. Presumably, the smaller surface should yield looser binding. This effect is not considered in our work. (4) The orientation of CBD relative to cellulose sampled in the MD simulations is limited. The TI calculations adopted here are very time-consuming and costly, impeding us from exploring this factor. (5) Force field errors for the cellulose may also contribute to the deviations. The GLYCAM06 force field has not been widely tested for binding free energy calculations of carbohydrates binding to protein receptors. Further investigation is needed to clarify these factors and yield quantitative agreement with the experimental data.

Although free energy decomposition is only an approximate process [35–37], it offers useful information about the molecular mechanism of the CBD–cellulose binding. The three-step TI simulation conveniently separates the electrostatic and VDW contributions to the free energy. We can see that $\Delta\Delta G_{vdw}$ for the three aromatic residue mutations Y5A, Y31A and Y32A are 4.13, 4.39 and 8.97 kcal mol$^{-1}$, respectively, corresponding to 76%, 69% and 83% of the total free energy changes, which indicates that the VDW interactions of the three residues play a more important role in the tight binding to cellulose than the electrostatic interactions. For the N29A and Q34A mutations, $\Delta\Delta G_{vdw}$ accounts for 37% and 48% of the corresponding total free energy differences. Considering that tyrosine residues contribute more to the overall binding, we can conclude that VDW interactions are more important than electrostatics in the binding of CBD to cellulose.

However, if we integrate $\partial V/\partial\lambda$ for the electrostatics and VDW separately (Fig. 5), it becomes apparent that the electrostatic change is much larger than the VDW change in the mutation to alanine for all five residues (Fig. 5) of the CBD–cellulose complex. This result underlines the importance of electrostatics in the CBD–cellulose binding, in concert with an early theoretical study of Cel7A CBD processivity [12] as well as a more recent quantum mechanics (QM) study of a tyrosine methyl β-D-glucopyranoside (OMG) complex, which suggests that dispersive interactions are less important than the hydrogen bond between the tyrosine OH and sugar O4 [38]. This observation seems to suggest that electrostatics is more important than VDW in the binding. However, according to the thermodynamic cycle (Fig. 2), the reference state (CBD alone in solution) must also be considered. In fact, the electrostatics dominates the overall free energy change ($\Delta G_2$) for the free CBD forms (Fig. 5). This electrostatics presumably arises from the change in solvation energy when mutating a polar residue (e.g., N or Y) to a nonpolar alanine. The comparable electrostatic change for a mutation in CBD and the CBD–cellulose complex largely cancels out. In another words, the solvation energy loss due to the binding of CBD to cellulose is compensated by the electrostatic interactions of CBD and cellulose. As a consequence, for $\Delta G_2 - \Delta G_1$ (the binding free energy

**Table 2** Free energy differences of the single mutations Y5W or Y31W. All values are in kcal mol$^{-1}$. Two configurations are considered for each tryptophan side chain (Fig. 3)

| Mutation | $\Delta\Delta G_{vdw}$ | $\Delta\Delta G_{ele}$ | $\Delta\Delta G_{cal}$ |
|---|---|---|---|
| Y5W1 | −1.83±0.41 | 1.59±0.34 | −0.24±0.55 |
| Y5W2 | −0.75±0.54 | 1.63±0.33 | 0.88±0.52 |
| Y31W1 | −0.59±0.30 | 0.93±0.10 | 0.34±0.35 |
| Y31W2 | −0.79±0.51 | 1.42±0.17 | 0.62±0.40 |

difference), the VDW contribution is larger and thus more important for binding. An early experimental isothermal titration microcalorimetry study of the binding of CBD from the β-1,4 exoglucanase Cex of *Cellulomonas fimi* to bacterial microcrystalline cellulose (BMCC) suggests that electrostatics does not drive the adsorption of CBDcex onto cellulose. CBDcex is larger than CBD Cel7A, but both have a flat surface composed of three aromatic residues (tyrosine or tryptophan). Our result is consistent with this experimental finding.

Our work confirms that the hydrophobic effects of the three tyrosine residues Y5, Y31 and Y32 that form the flat surface of the CBD are required for tight binding to cellulose. Experimental data show that the CBD binding free energy difference $\Delta G_{W5} - \Delta G_{Y5}$ is −0.26 kcal mol$^{-1}$, suggesting that the Y5W mutant has a higher binding affinity to cellulose [32]. Our computational values are −0.24±0.55 kcal mol$^{-1}$ for side-chain configuration 1 and 0.88±0.52 kcal mol$^{-1}$ for configuration 2. While both configurations form a flat surface on contact with cellulose, configuration 1 is favored since it is more stable (Table 2). The free energy decomposition suggests that while the VDW interaction becomes stronger for the Y5W mutant, presumably due to its larger hydrophobic surface, the electrostatic interaction is weaker, probably because of the loss of the polar tyrosine hydroxyl group. Compared to the wild type, the Y31W mutant exhibits decreases in the binding free energy by 0.34±0.35 and 0.62±0.40 kcal mol$^{-1}$ for the two configurations. Similar to Y5W, the electrostatic and VDW changes offset each other. The Y32W mutation was not investigated, because substituting the Y32 phenol side chain by the larger indole group of tryptophan in the confined space could disturb the structure and make the surface less flat [39].

The experimental study of the impact of *Humicola grisea* CBHI CBD cellulose binding affinity on CBHI hydrolysis activity against cellulose has shown that the

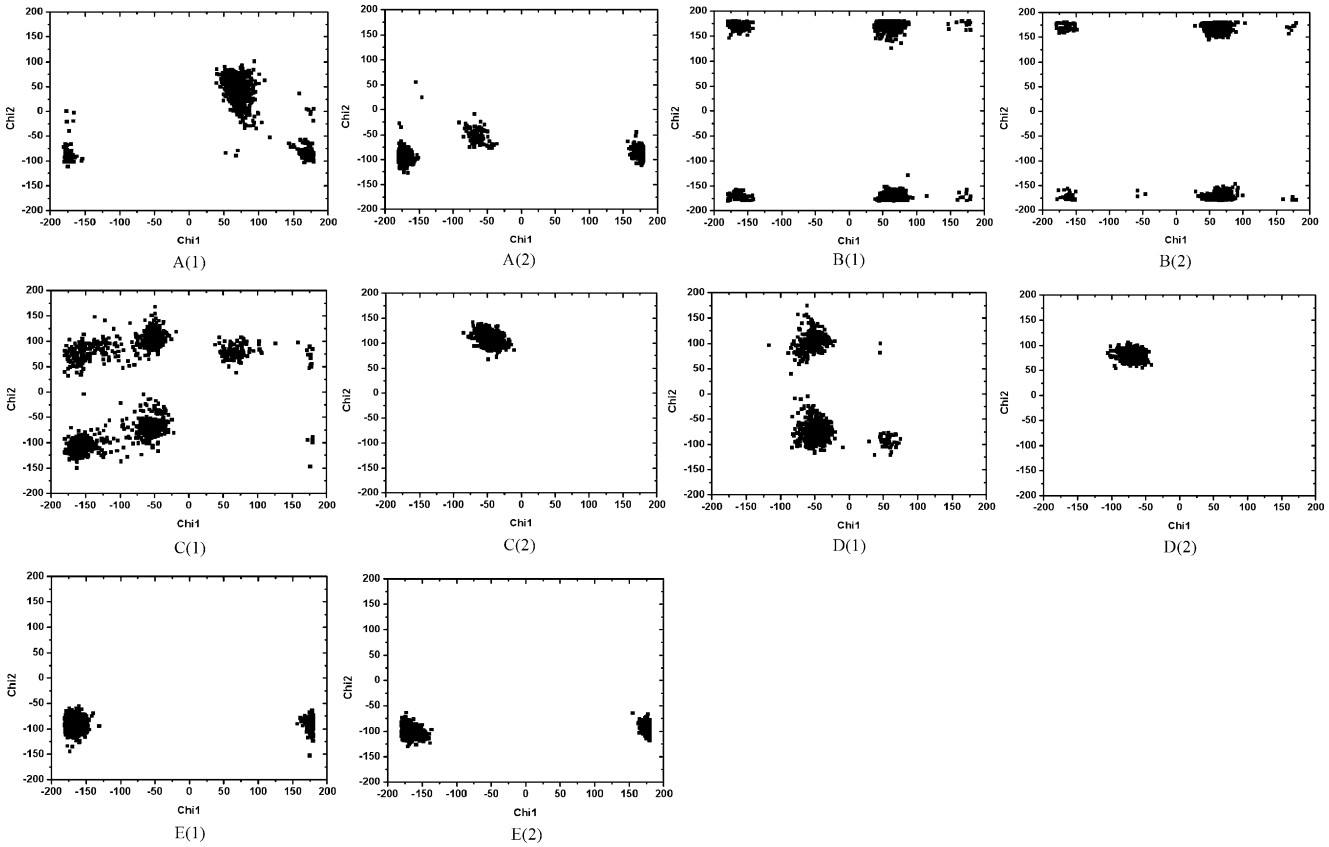



**Fig. 6** Scatter plots of $\chi_1$ versus $\chi_2$ for the side chains of the five residues: **a** N29, **b** Q34, **c** Y5, **d** Y31, **e** Y32. The labels *(1)* and *(2)* stand for free and cellulose-bound CBD states, respectively

tighter the binding, the higher the catalytic activity [39]. Our work suggests that one way to increase the binding affinity is to strengthen either the VDW or the electrostatic interactions, but not to sacrifice the other.

## Changes in the dynamics of side chains due to binding

The protein side-chain dynamics represents a major component of protein conformational entropy [40]. Loss of side-chain conformational entropy could be an important influence that reduces a substrate's binding affinity to an enzyme. The torsional entropy of side chains is thought to provide a good approximation for the side-chain entropy. To investigate the side-chain dynamics, we monitored the changes in side-chain dihedral angles and entropy of the five CBD residues before and after binding to cellulose. Figure 6 shows scatter plots of Chi1 ($\chi_1$) and Chi2 ($\chi_2$) for the five residues Y5, N29, Y31, Y32 and Q34 from the 20 ns free CBD and cellulose-bound CBD simulations. Table 3 lists the averages of and fluctuations in the $\chi_1$ and $\chi_2$ angles of the five residues in free and bound states. In general, as expected, the $\chi_1$ and $\chi_2$ values of the side-chain residues in the free CBD spread over a larger area than those of the cellulose-bound CBD, which indicates that the conformations of the side chains are restrained due to the binding to cellulose.

As shown in Fig. 6, after binding to cellulose, the number of the major side-chain states reduces from 5 and 3 to 1 for Tyr5 and Tyr31, respectively. The fluctuations in $\chi_1$ and $\chi_2$ for both Y5 and Y31 decease to ≤10° (Table 3). The considerably restricted side-chain rotations due to binding are further reflected in the notable drop in the entropic contribution of the side chain to the free energy ($T\Delta S$), with differences between the ligand-free and bound states for Y5 and Y31 of −1.71 and −1.02 kcal mol$^{-1}$, respectively (Fig. 7). Unlike Y5 and Y31, the distributions the of $\chi_1$ and $\chi_2$ values for Y32 are similar in the free and bound states, with only a small fluctuation of ~10° (Table 3), suggesting that this side chain is confined in both states. The $T\Delta S$ value between the free and bound states is only −0.11 kcal mol$^{-1}$ for this residue. In comparison, the side-chain

**Table 3** Averages of and fluctuations in the angles $\chi_1$ and $\chi_2$ for the five residues in the free and bound states

| Residue | Free state | | Bound state | |
|---|---|---|---|---|
| | $\chi_1$(°) | $\chi_2$(°) | $\chi_1$(°) | $\chi_2$(°) |
| N29 | 86±40 | 25±51 | −167±31 | −89±16 |
| Q34 | 93±56 | 178±11 | 76±41 | 177±11 |
| Y5 | −124±77 | −1±91 | −45±10 | 107±10 |
| Y31 | −45±26 | −22±83 | −68±9 | 80±8 |
| Y32 | −169±9 | −90±12 | −173±10 | −98±10 |

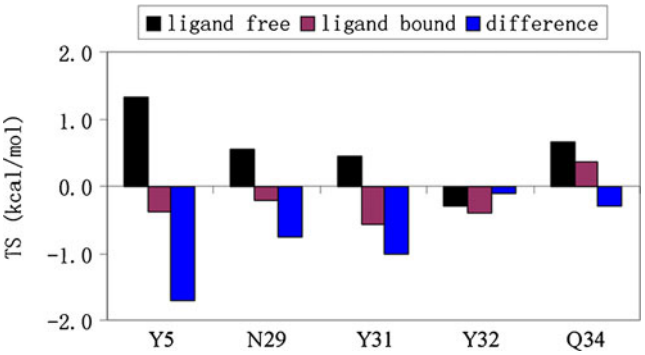



**Fig. 7** Side-chain dihedral entropic contributions to the free energies *TS* (*T*=300 K) of the five residues in the free and cellulose-bound states. The difference between the two states for each residue is shown as a *blue bar*

torsional entropy losses for N29 and Q34 are slightly larger: −0.75 and −0.30 kcal mol$^{-1}$, respectively (Fig. 7). It is worth noting that the entropy change considered here is very rough, and the sampling of the dihedral space with a 20 ns simulation is rather limited. However, qualitatively, it is clear that for some residues the entropy changes are relatively large and partially offset the gain in interaction energy after CBD binding.

## Conclusions

In this paper, we studied the effect of a single mutation of Cel7A CBD on binding to the cellulose surface using a TI simulation method. Analyses of the computed free energy differences provided useful information for understanding the binding mechanism between CBD and cellulose. Our results suggest that residues Y5, Y31 and Y32 are more important than N29 and Q34 for the binding of CBD to cellulose. The VDW interactions between the amino acid Y5, Y31 and Y32 and the cellulose surface account for a large proportion of the binding affinity. Our calculations yielded relative binding affinities with a similar trend, but overestimated the experimental $\Delta\Delta G$ data. The current model yields only qualitative agreement with experiments. A more realistic cellulose model with properly weighted CBD binding modes will likely yield a more quantitative agreement, which will be explored in a further study. The changes in the binding affinities of the mutants Y5W and Y31W to cellulose were calculated. The mutant Y5W shows slightly improved affinity, which is in agreement with the experimental results. Although the VDW interaction between CBD and cellulose is enhanced for the Y to W mutations, weakening of the electrostatics largely offsets this effect. The side-chain torsional entropy decreases for Y5, N29, Y31, Y32 and Q34 due to rotamer rigidification after binding to cellulose.

**Acknowledgments** We are thankful to the Supercomputing Center of Chinese Academy of Sciences (CAS) and the Supercomputing Center of USTC (University of Science and Technology of China) for providing the computer resources and time. This work was supported by the 100 Talent Project, the Knowledge Innovation Program of the CAS (grant no. KSCX2EWJ10), the Director Innovation Foundation of Qingdao Institute of Biomass Energy and Bioprocess Technology, and the Foundation for Outstanding Young Scientists in Shandong Province (no. BS2010NJ020).

## References

1. Henrissat B (1994) Cellulases and their interaction with cellulose. Cellulose 1:169–196
2. Teeri TT (1997) Crystalline cellulose degradation: new insight into the function of cellobiohydrolases. Trends Biotechnol 15:160–167
3. Divne C, Stahlberg J, Teeri TT, Jones TA (1998) High-resolution crystal structures reveal how a cellulose chain is bound in the 50 A long tunnel of cellobiohydrolase I from *Trichoderma reesei*. J Mol Biol 275:309–325
4. Linder M, Mattinen ML, Kontteli M, Lindeberg G, Stahlberg J, Drakenberg T, Reinikainen T, Pettersson G, Annila A (1995) Identification of functionally important amino acids in the cellulose-binding domain of *Trichoderma reesei* cellobiohydrolase I. Protein Sci 4:1056–1064
5. Teeri TT, Koivula A, Linder M, Wohlfahrt G, Divne C, Jones TA (1998) *Trichoderma reesei* cellobiohydrolases: why so efficient on crystalline cellulose? Biochem Soc Trans 26:173–178
6. Tomme P, Van Tilbeurgh H, Pettersson G, van Damme J, Vandekerckhove J, Knowles J, Teeri T, Claeyssens M (1988) Studies of the cellulolytic system of *Trichoderma reesei* QM 9414. Analysis of domain function in two cellobiohydrolases by limited proteolysis. Eur J Biochem 170:575–581
7. Divne C, Stahlberg J, Reinikainen T, Ruohonen L, Pettersson G, Knowles JK, Teeri TT, Jones TA (1994) The three-dimensional crystal structure of the catalytic core of cellobiohydrolase I from *Trichoderma reesei*. Science 265:524–528
8. Barr BK, Hsieh YL, Ganem B, Wilson DB (1996) Identification of two functionally different classes of exocellulases. Biochemistry 35:586–592
9. Imai T, Boisset C, Samejima M, Igarashi K, Sugiyama J (1998) Unidirectional processive action of cellobiohydrolase Cel7A on *Valonia* cellulose microcrystals. FEBS Lett 432:113–116
10. Kraulis J, Clore GM, Nilges M, Jones TA, Pettersson G, Knowles J, Gronenborn AM (1989) Determination of the three-dimensional solution structure of the C-terminal domain of cellobiohydrolase I from *Trichoderma reesei*. A study using nuclear magnetic resonance and hybrid distance geometry-dynamical simulated annealing. Biochemistry 28:7241–7257
11. Bu L, Beckham GT, Crowley MF, Chang CH, Matthews JF, Bomble YJ, Adney WS, Himmel ME, Nimlos MR (2009) The energy landscape for the interaction of the Family 1 carbohydrate-binding module and the cellulose surface is altered by hydrolyzed glycosidic bonds. J Phys Chem B 113:10994–11002
12. Beckham GT, Matthews JF, Bomble YJ, Bu L, Adney WS, Himmel ME, Nimlos MR, Crowley MF (2010) Identification of amino acids responsible for processivity in a Family 1 carbohydrate-binding module from a fungal cellulase. J Phys Chem B 114:1447–1453
13. Nimlos MR, Matthews JF, Crowley MF, Walker RC, Chukkapalli G, Brady JW, Adney WS, Cleary JM, Zhong L, Himmel ME (2007) Molecular modeling suggests induced fit of Family I carbohydrate-binding modules with a broken-chain cellulose surface. Protein Eng Des Sel 20:179–187
14. Steinbrecher T, Mobley DL, Case DA (2007) Nonlinear scaling schemes for Lennard–Jones interactions in free energy calculations. J Chem Phys 127:214108
15. Kirkwood JG (1935) Statistical mechanics of fluid mixtures. J Chem Phys 3:300–313
16. Sattelle BM, Sutcliffe MJ (2008) Calculating chemically accurate redox potentials for engineered flavoproteins from classical molecular dynamics free energy simulations. J Phys Chem A 112:13053–13057
17. Perez MA, Fernandes PA, Ramos MJ (2010) Substrate recognition in HIV-1 protease: a computational study. J Phys Chem B 114:2525–2532
18. Shirts MR, Pitera JW, Swope WC, Pande VS (2003) Extremely precise free energy calculations of amino acid side chain analogs: comparison of common molecular mechanics force fields for proteins. J Chem Phys 119:5740–5761
19. Straatsma TP, Mccammon JA (1992) Computational alchemy. Annu Rev Phys Chem 43:407–435
20. Cheng Y, Cheng X, Radic Z, McCammon JA (2007) Acetylcholinesterase: mechanisms of covalent inhibition of wild-type and H447I mutant determined by computational analyses. J Am Chem Soc 129:6562–6570
21. Nam K, Verdine GL, Karplus M (2009) Analysis of an anomalous mutant of MutM DNA glycosylase leads to new insights into the catalytic mechanism. J Am Chem Soc 131:18208–18209
22. Seeliger D, de Groot BL (2010) Protein thermostability calculations using alchemical free energy simulations. Biophys J 98:2309–2316
23. Nishiyama Y, Langan P, Chanzy H (2002) Crystal structure and hydrogen-bonding system in cellulose 1 beta from synchrotron X-ray and neutron fiber diffraction. J Am Chem Soc 124:9074–9082
24. Case DA, Darden TA, Cheatham TE III, Simmerling CL, Wang J, Duke RE, Luo R, Crowley MF, Walker RC, Zhang W, Merz KM, Wang BC, Hayik S, Roitberg A, Seabra G, Kolossváry I, Wong KF, Paesani F, Vanicek J, Wu X, Brozell SR, Steinbrecher T, Gohlke H, Yang L, Tan C, Mongan J, Hornak V, Cui G, Mathews DH, Seetin MG, Sagui C, Babin V, Kollman PA (2008) AMBER 10. University of California, San Francisco
25. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C (2006) Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins 65:712–725
26. Kirschner KN, Yongye AB, Tschampel SM, Gonzalez-Outeirino J, Daniels CR, Foley BL, Woods RJ (2008) GLYCAM06: a generalizable biomolecular force field. Carbohydrates. J Comput Chem 29:622–655
27. Kirschner KN, Woods RJ (2001) Solvent interactions determine carbohydrate conformation. Proc Natl Acad Sci USA 98:10541–10545
28. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG (1995) A smooth particle mesh Ewald method. J Chem Phys 103:8577–8593
29. Ryckaert JP, Ciccotti G, Berendsen HJC (1977) Numerical-integration of cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Comput Phys 23:327–341
30. Ai R, Qaiser Fatmi M, Chang CE (2010) T-Analyst: a program for efficient analysis of protein conformational changes by torsion angles. J Comput Aided Mol Des 24:819–827
31. Yui T, Shiiba H, Tsutsumi Y, Hayashi S, Miyata T, Hirata F (2010) Systematic docking study of the carbohydrate binding module protein of Cel7A with the cellulose Ialpha crystal model. J Phys Chem B 114:49–58
32. Linder M, Lindeberg G, Reinikainen T, Teeri TT, Pettersson G (1995) The difference in affinity between two fungal cellulose-

binding domains is dominated by a single amino acid substitution. FEBS Lett 372:96–98
33. Srisodsuk M, Lehtio J, Linder M, Margolles-Clark E, Reinikainen T, Teeri TT (1997) *Trichoderma reesei* cellobiohydrolase I with an endoglucanase cellulose-binding domain: action on bacterial microcrystalline cellulose. J Biotechnol 57:49–57
34. Lehtio J, Sugiyama J, Gustavsson M, Fransson L, Linder M, Teeri TT (2003) The binding specificity and affinity determinants of Family 1 and Family 3 cellulose binding modules. Proc Natl Acad Sci USA 100:484–489
35. Smith PE, van Gunsteren WF (1994) When are free energy components meaningful? J Phys Chem 98:13735–13740
36. Mark AE, van Gunsteren WF (1994) Decomposition of the free energy of a system in terms of specific interactions. Implications for theoretical and experimental studies. J Mol Biol 240:167–176
37. Boresch S, Karplus M (1995) The meaning of component analysis: decomposition of the free energy in terms of specific interactions. J Mol Biol 254:801–807
38. Mohamed MN, Watts HD, Guo J, Catchmark JM, Kubicki JD (2010) MP2, density functional theory, and molecular mechanical calculations of C-H…pi and hydrogen bond interactions in a cellulose-binding module-cellulose model system. Carbohydr Res 345:1741–1751
39. Takashima S, Ohno M, Hidaka M, Nakamura A, Masaki H (2007) Correlation between cellulose binding and activity of cellulose-binding domain mutants of *Humicola grisea* cellobiohydrolase 1. FEBS Lett 581:5891–5896
40. Pickett SD, Sternberg MJ (1993) Empirical scale of side-chain conformational entropy in protein folding. J Mol Biol 231:825–839

J Mol Model (2012) 18:1365–1374
DOI 10.1007/s00894-011-1162-9

ORIGINAL PAPER

# A quantum-chemical study of the binding ability of βXaaHisGlyHis towards copper(II) ion

**Żaneta Czyżnikowska · Justyna Brasuń**

Received: 16 February 2011 / Accepted: 24 June 2011 / Published online: 15 July 2011


**Abstract** The present study analyzed binding of $Cu^{2+}$ to tetrapeptides in water solution at several levels of theoretical approximation. The methods used to study the energetic and structural properties of the complexes in question include semiempirical hamiltonians, density functional theory as well as ab initio approaches including electron correlation effects. In order to shed light on the character of interactions between $Cu^{2+}$ and peptides, which are expected to be mainly electrostatic in nature, decomposition of interaction energy into physically meaningful components was applied.

**Keywords** Density functional theory · Intermolecular interaction · Tetrapeptide · Copper binding

## Introduction

Copper$^{2+}$-histidine species have been investigated extensively over the past four decades with an eye towards understanding of the role of $Cu^{2+}$ in cell metabolism (see [1] and references therein). Copper ions have also been proved to be involved in salt-induced peptide formation [2]. Despite the plethora of experimental studies concerning $Cu^{2+}$-polypeptide complexes, theoretical studies are significantly more rare, although molecular modeling and quantum chemistry techniques may provide complementary information, such as data on the structural and exegetic properties of metal-polypeptide complexes [3–25]. Studies have shown that the insertion of β-amino acid residues into peptide chains leads to an increase in their proteolytic stability [26, 27].

Recently, one of us has investigated experimentally the binding of $Cu^{2+}$ to βXaaHisGlyHis, where X=Asp,Ala [28]. Several important observations regarding the formation of complexes at various pH values were presented. However, some issues remained unresolved, e.g., the conformation of deprotonated forms or missing spectroscopic data for some species. Although theoretical modeling of bioinorganic complexes with transition metals might be useful in completing the picture, it is by no means a trivial task. The reasons for this are numerous. Due to the large size of the molecules involved, accurate ab initio methods like coupled cluster formalism [CCSD,CCSD(T)] or CASPT2 cannot be applied. And even if they could, the area of probed conformational space would be very limited. Performing computations of energetic properties with the aid of density functional theory (DFT) is usually feasible for bioinorganic systems involving copper [29, 30] but only a limited number of exchange-correlation functionals have been proved to provide satisfactory results. Turning toward UV-Vis spectra, one finds that most functionals suffer from what is known as self-interaction error. As a result, excitation energies to charge-transfer and Rydberg states are not predicted correctly. Thus, in the field of transition metal chemistry, any computational protocol must be extensively tested against available experimental data before the results can be considered reliable.

The motivation for this study was threefold. First, we aimed to analyze the structural aspects of the binding of $Cu^{2+}$ to polypeptides. Second, we wanted to discuss binding strength together with the analysis of its nature by performing intermolecular interaction decomposition. And a third aim of this study was to compare predictions of

Ż. Czyżnikowska (✉) · J. Brasuń
Department of Inorganic Chemistry, Faculty of Pharmacy,
Wroclaw Medical University,
Szewska 38,
50-139 Wrocław, Poland
e-mail: zaneta.czyznikowska@gmail.com

various exchange-correlation functionals used within the DFT framework for studying $Cu^{2+}$-containing polypeptides. Although the motivation to undertake this latter subject was purely methodological in nature, such comparisons have become an essential part of contemporary computational chemistry.

## Computational methodology

In order to reliably probe the conformational space of the studied complexes, we analyzed several hundred different conformers for all species. In order to find a reasonable compromise between accuracy and computational cost during geometry optimizations, we decided to employ a recently reparametrized NDDO-type semiempirical method called PM6 [31]. As shown recently by Stewart [31], for many properties (including heats of formation, equilibrium geometries or electric-dipole properties) the PM6 method improves upon its predecessor, i.e., the PM3 method. It is also quite successful in determining structures of bioinorganic complexes with transition metals [31]. It has also been shown more recently that this approach can be applied successfully to the modeling of proteins and metaloproteins [32]. Geometry optimization of all conformers, with solvent effects taken into account (see below), was followed by evaluation of a hessian matrix to confirm that stationary points correspond to minima on the potential energy surface (PES). Among all optimized structures, only the lowest-energy conformers for each species are analyzed in the next section. The relative stability of conformers was determined using structures optimized with the PM6 method. In this event, we used several levels of theoretical approximation, including the Hartree-Fock (HF) method, second-order Møller-Plesset (MP2) perturbation theory and the Kohn-Sham formulation of DFT combined with Gaussian basis functions and Hay-Wadt effective core potentials (ECPs) for copper, which incorporate mass-velocity and Darwin relativistic effects [33, 34]. Unless otherwise stated, all computations were performed with inclusion of solvent effects (water). In that event we used a polarizable continuum model (PCM) [35–39]. All the calculations described above were performed within unrestricted formalism using the Gaussian 09 suite of programs [40].

In the present contribution we also made an attempt to analyze the intermolecular interactions between copper ion and the surrounding ligands. In order to compute the total interaction energy, we used a supermolecular approach corrected for what is known as basis set superposition error (BSSE) using the scheme proposed by Boys and Bernardi [41]. To analyze the importance of interaction energy components other than electrostatics (which, in the case of the studied systems, is expected to be main stabilizing factor), we adopted the hybrid variational-perturbational scheme [42–46]. In this approach, the total interaction energy calculated in a supermolecular approach at the MP2 perturbation theory level is partitioned into HF and Coulomb electron correlation interaction energy components:

$$\Delta E^{\mathrm{MP2}} = \Delta E^{\mathrm{HF}} + \varepsilon_{\mathrm{MP}}^{(2)} \tag{1}$$

The HF interaction energy term can be further divided into the Heitler-London interaction energy, which encompasses the electrostatic interactions of unperturbed monomer charge densities as well as the associated exchange repulsion, and the $\Delta E_{\mathrm{del}}^{\mathrm{HF}}$ component comprises the induction and the associated exchange effects [47, 48].

$$\Delta E^{\mathrm{HF}} = \Delta E_{\mathrm{del}}^{\mathrm{HF}} + \Delta E^{\mathrm{HL}} = \Delta E_{\mathrm{del}}^{\mathrm{HF}} + \varepsilon_{\mathrm{el}}^{(10)} + \varepsilon_{\mathrm{ex}}^{\mathrm{HL}}. \tag{2}$$

The second order electron correlation correction term, $\varepsilon_{\mathrm{MP}}^{(2)}$, can be partitioned into the second order dispersion interaction and the correlation corrections to the HF components [44, 45].

$$\varepsilon_{\mathrm{MP}}^{(2)} = \varepsilon_{\mathrm{disp}}^{(20)} + \varepsilon_{\mathrm{el,r}}^{(12)} + \Delta E_{\mathrm{ex-del}}^{(2)}. \tag{3}$$

The $\varepsilon_{\mathrm{el}}^{(10)}$ and $\varepsilon_{\mathrm{disp}}^{(20)}$ contributions are obtained in a standard polarization perturbation theory,[49] whereas the $\varepsilon_{\mathrm{el,r}}^{(12)}$ term is calculated using the formula proposed by Moszynski et al. [50]. In all necessary calculations, the dimer-centered basis set was used consistently and therefore the results are BSSE-free due to the full counterpoise correction [41]. More details about the interaction energy partitioning scheme and the recent implementation adopted in this work can be found elsewhere [51–53]. In the present contribution we compute only $\Delta E_{\mathrm{del}}^{\mathrm{HF}}$, $\varepsilon_{\mathrm{el}}^{(10)}$ and $\varepsilon_{\mathrm{ex}}^{\mathrm{HL}}$.

## Results and discussion

In the present contribution, we report an analysis of numerous conformers of βAspHisGlyHis and βAlaHisGlyHis and their deprotonated forms. As mentioned in the previous section, only the lowest energy conformers (confirmed to correspond to minima on the PES) were considered from among initial set comprising almost 1,000 structures. The species analyzed (Fig. 1) were grouped into five categories according to their terminal amino acid (Asp: **A**,**B**; Ala: **C**, **D**, **E**) and the number of dissociated protons (CuL: **B**, **D**; $CuH_{-1}L$: **A**, **C**; $CuH_{-2}L$: **E**). It is important to stress that various protonation states of the investigated species can be observed experimentally as the pH of the aqueous solution changes. Thus, in quantum-chemical calculations, we account for this effect by considering the three most common protonation states.

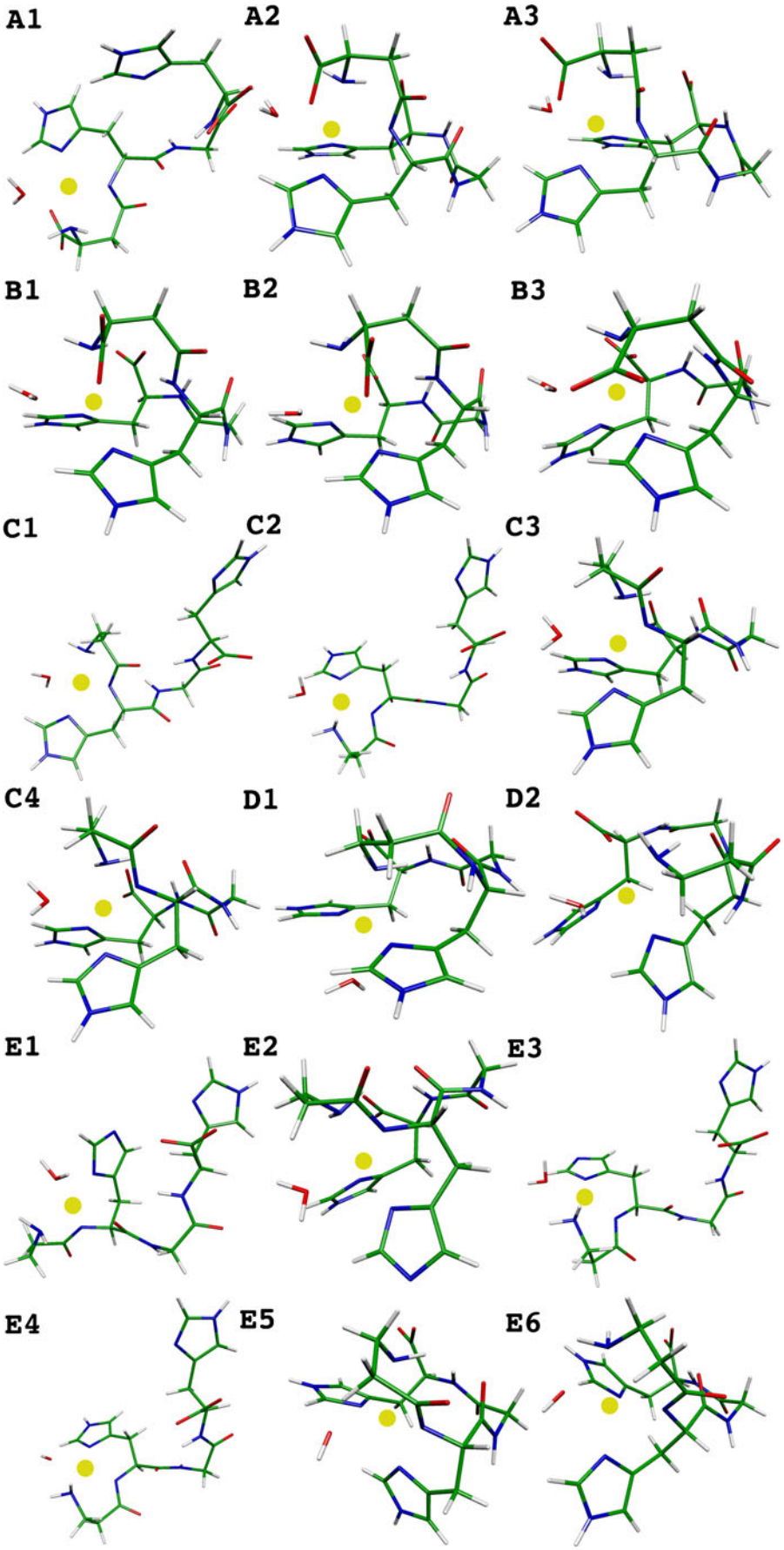



**Fig. 1** Structure of the most stable complexes of the species investigated in the present study

Before proceeding to discuss the relative stabilities of the systems under consideration together with their geometrical aspects, we feel that a few words regarding the reliability of applied computational methods are required. Of the plethora of exchange-correlation functionals, only some are suitable for analysis of thermochemical properties or barrier heights. The other issue is that only a few can be used reliably for studies of bioinorganic complexes with transition metals. One of the functionals that preserves a good balance of the description of bonds between transition metals and metal–ligand bonds is the M05 functional proposed by the Minnesota group [54]. The paper presenting the new M05-2X functional also performed an extensive testing of M05 [55]. The former contains double the amount of HF exchange with respect to the latter, namely 56%.

As is well recognized nowadays, problems in describing systems with significant static (nondynamic) electron correlation have their roots in the amount of HF exchange present in the hybrid functional [56]. As pointed out by Cramer and Truhlar, the multi-reference character seems to be much more important for the failure of HF exchange-rich functionals than the presence of transition metal(s) in the system [56]. In general, the performance of M05-2X in determining the properties of transition metal complexes where multi-reference description is needed is much worse than that of M05. This observation, i.e., the difference between the results computed using M05 and M05-2X, might be sometimes used as a rule of thumb to judge if nondynamic correlation effects are important. In the present study we compute total energies in the presence of a water environment using several exchange-correlation functionals including M05, M06 (a redesigned and reoptimized M05 functional that might be considered more accurate) as well as M06-2X and M06-HF. In particular, the latter contains 100% HF exchange. The results of such computations are presented in Table 1. First, let us note that, for the M05 functional, the inclusion of the Hay-Wadt ECP for copper (labeled as HW) leads to an average difference in relative stabilities of 0.62 kcal/mol. As the picture is quite similar for other employed methods, we present only data determined with the use of ECP. Comparing M05 and M06 functionals reveals that it is the charge of system rather than the amino acid (Asp vs Ala) that seems to determine the difference in relative stabilities. Indeed, we note that, for neutral complexes (**B1**-**B3** and **D1**-**D2**), the differences are not significant and the ordering of conformers is the same.

In order to judge the multi-reference character of the systems let us compare how the relative stabilities vary upon increasing the HF exchange (28%→56%→100%). Doubling the amount of the HF exchange from 28% to 56% in most cases reduces the differences in stability between various conformers. An exception is the group of complexes denoted as **C***n*, where the situation is more complicated (although in terms of magnitude the changes are not significant); in this case increasing the amount HF exchange does not lead to reordering of conformers. It is

**Table 1** Relative energies including solvent effects for the group of compounds investigated. All values were determined with the 6-31G(d) basis set for molecules and are given in kcal/mol

| | B3LYP | B3LYP,HW | M05 | M05,HW | M06,HW | M06-2X,HW | M06-HF,HW | MP2,HW | HF,HW |
|---|---|---|---|---|---|---|---|---|---|
| A1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| A2 | 20.63 | 19.75 | 19.56 | 18.75 | 9.16 | 4.57 | 1.54 | 2.85 | 24.05 |
| A3 | 19.66 | 18.68 | 18.44 | 17.52 | 8.13 | 3.74 | 0.59 | 1.80 | 22.91 |
| B1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.87 |
| B2 | 24.01 | 24.61 | 24.41 | 24.37 | 24.80 | 19.59 | 17.31 | 20.48 | 23.72 |
| B3 | 1.22 | 3.13 | 1.27 | 2.86 | 4.17 | 2.72 | 1.77 | 6.65 | 0.00 |
| C1 | 0.47 | 0.37 | 1.02 | 0.00 | 9.75 | 10.75 | 12.00 | 11.86 | 0.14 |
| C2 | 0.28 | 0.00 | 1.73 | 0.55 | 10.58 | 11.91 | 100.00 | 13.10 | 0.00 |
| C3 | 0.00 | 1.62 | 0.00 | 0.18 | 0.12 | 0.11 | 0.12 | 0.11 | 14.69 |
| C4 | 0.03 | 1.57 | 0.03 | 0.12 | 0.00 | 0.00 | 0.00 | 0.00 | 14.72 |
| D1 | 30.34 | 28.53 | 28.86 | 27.47 | 32.66 | 35.49 | 36.56 | 35.66 | 33.53 |
| D2 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| E1 | 9.34 | 9.36 | 6.89 | 5.87 | 6.53 | 5.89 | 5.87 | 5.66 | 02.30 |
| E2 | 10.77 | 12.03 | 7.66 | 8.05 | 0.00 | 0.00 | 0.00 | 0.00 | 18.64 |
| E3 | 7.56 | 6.47 | 5.96 | 4.59 | 6.31 | 8.22 | 10.77 | 9.58 | 0.94 |
| E4 | 3.51 | 3.24 | 5.37 | 5.10 | 5.85 | 7.97 | 10.49 | 10.10 | 0.00 |
| E5 | 16.25 | 18.21 | 16.48 | 17.45 | 8.98 | 6.99 | 3.32 | 8.91 | 21.58 |
| E6 | 0.00 | 0.00 | 0.00 | 0.00 | 0.22 | 2.28 | 3.28 | 2.77 | 0.24 |

also worth noting that both M06 and M06-2X predict the same most stable conformers in all groups. Likewise, in the large majority of cases, any further increase in the amount HF exchange diminishes the differences in relative stability. The energy difference obtained for the structure denoted as **C2** seems to be an artifact. The spin contamination for this conformer is similar to the values obtained using other functionals. Despite much effort, we did not succeed in locating a lower-energy state. In summary, based on the data presented in Table 1, we may conclude that the results are not sensitive to the amount of exact exchange in the exchange-correlation functional, which might suggest that a single-reference approach is sufficient to describe the studied systems [56]. Hereafter, we shall consider the M06 functional as a reference point for further analysis of the stability of these complexes.

In the following, we discuss the structural aspects of copper coordination by βXaaHisGlyHis peptide in a water environment. The discussion is based on a set of selected geometries of lowest-energy conformers determined using the PM6 method. Considering the size of the molecules studied, it would be a formidable task to probe reliably the potential energy hypersurface using a more accurate method including electron correlation. Thus, to validate the selected set of structures we performed an additional re-optimization of two structures belonging to its smallest subset (denoted as **D**). The differences in geometries optimized using the PM6 and the B3LYP methods [using the 6-31G(d,p) basis set] are presented in Fig. 2 (hydrogens are not shown for the sake of clarity). It should be mentioned that the least-square fit analysis was performed for the three-atom pairs involved in coordination. The relative stabilities obtained (using the M06 method and the 6-31G(d) basis set with the HW pseudopotential for copper) for geometries optimized at the B3LYP/6-31G(d,p) level of theory were 28.58 and 0.0 kcal/mol for **D1** and **D2**, respectively. The energy difference is comparable to that obtained using the PM6 geometry, i.e., 32.66 kcal/mol. In order to unequivocally assess the performance of the PM6 method in geometry prediction, a much more systematic characterization of PES is required.

We included one water molecule in the $Cu^{2+}$ coordination shell in all the investigated complexes. What we observe is that the coordinated water molecule is usually located at longer distances from the metal center in comparison with other ligands ($Cu^{2+}$–O distance lies within the range 2.03–2.16 Å). In general, the bond distances between $Cu^{2+}$ and all ligands lie in the range of 1.9–2.2 Å, which is in good agreement with experimental values [57]. As far as the $CuH_{-1}L$ form of βAspHisGlyHis is concerned, two main ways of coordination are found. The binding of $Cu^{2+}$ to the peptide in the case of **A2** and **A3** involves the $-NH_{2,\ terminal}$ and $-COO^{-}$ groups of aspartic acid, as well as the nitrogen amide and imidazole ring of first and second histidine. Among the structural motifs observed for **A2** and **A3** there are also three chelate rings (one five-, six- and seven-membered ring). The five-membered ring involves the Cu–carboxyl moiety and the amino group while the six-membered ring is formed by the $Cu^{2+}$–nitrogen of the imidazole bond and the nitrogen of first amide donor atom. Both structures are also stabilized by a hydrogen bond involving the C-terminal carboxyl group and the N-terminal amino group at a distance of 2.08 Å. In the **A1** structure, coordination of the metal atom by the amino group, the nitrogen atom of the amide group and the imidazole ring of the first histidine is observed. The second imidazole ring is not involved in copper coordination in this structure. Thus, a pentacoordinated $CuH_{-1}L$ structure is found to be more stable than a hexacoordinated one. It follows from Table 1 that the **A1** conformer is a few

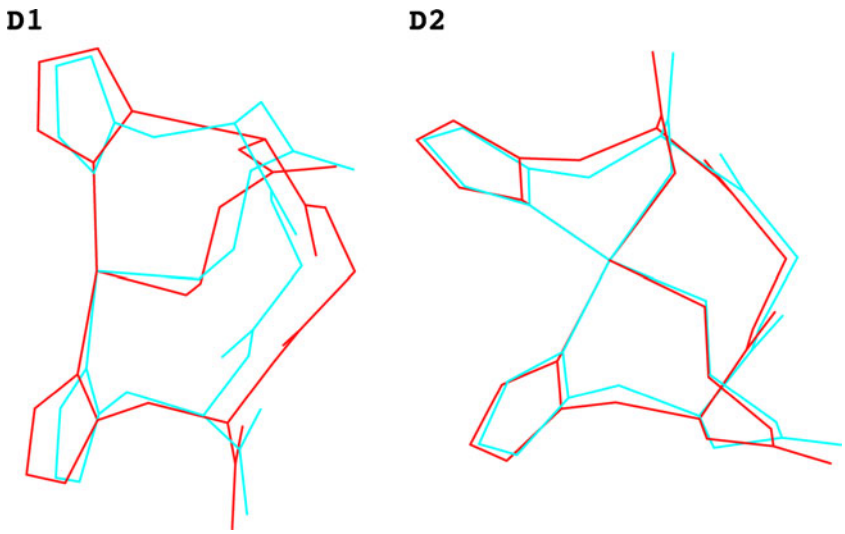
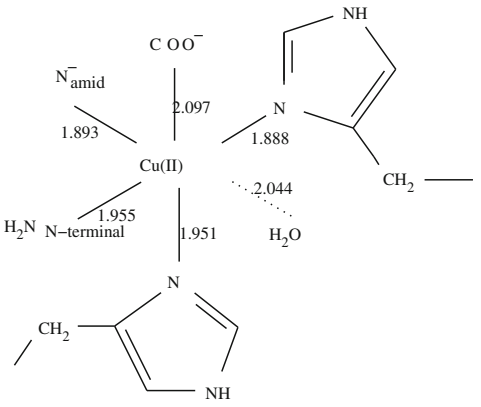
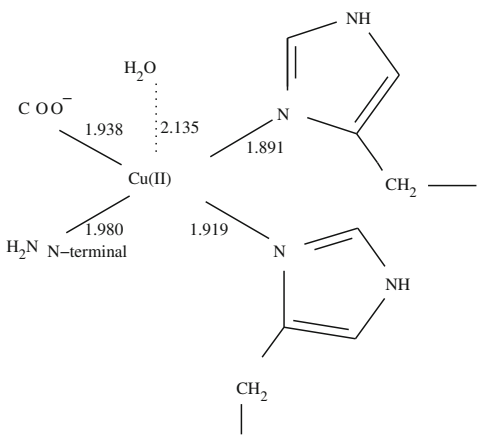


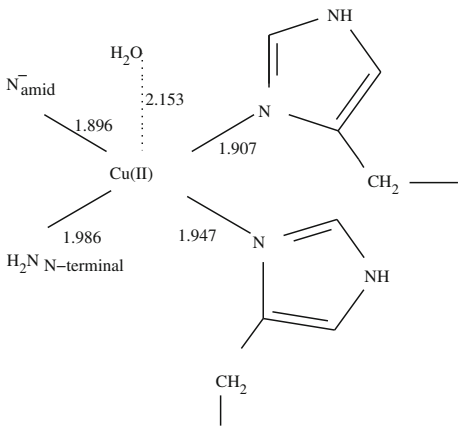

**Fig. 2** Differences in structural parameters determined using PM6 (*cyan*) and B3LYP/HW,6-31G(d,p) (*red*) methods



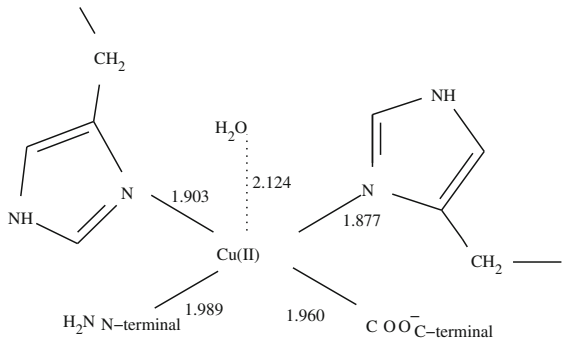

NH2–β Asp–His–Gly–His–C OO⁻ CuH_1L (**A2, A3**)



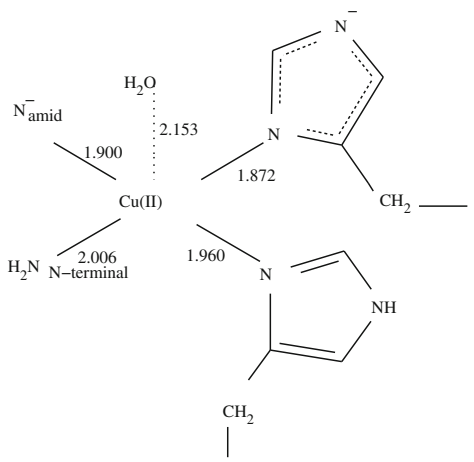

NH2–β Asp–His–Gly–His–C OO⁻ CuL (**B1**)



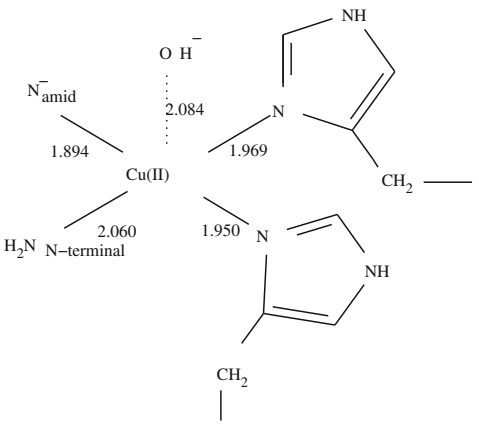

NH2–β Ala–His–Gly–His–C OO⁻ CuH_1L (**C3, C4**)




NH2–β Ala–His–Gly–His–C OO⁻ CuL (**D2**)




NH2–β Ala–His–Gly–His–C OO⁻ CuH_2L (**E2**)




NH2–β Ala–His–Gly–His–C OO⁻ CuH_2L (**E5**)

**Fig. 3** Structural parameters of coordination sites

kcal/mol more stable than **A2** and **A3**. This might be connected with steric effects and the existence of two negatively charged groups (C-terminal carboxyl and $N^-_{amide}$) at a distance of 4 Å, which leads to an unfolded structure. In the peptide with β Ala, the distance is found to be 1 Å longer and the pentacoordinated $CuH_{-1}L$ structure is more stable than the tetracoordinated one. The βAspHis-GlyHis peptide also exists in the CuL form, in which $Cu^{2+}$ is coordinated by the N-terminal amino group, $–COO^-$ and the nitrogens of imidazole ring. Moreover, the formation of the five-membered chelate ring between the $–NH_{2,terminal}$ and the carboxyl group is one of the main structural motifs observed in all complexes with Asp. What we also see is that the **B1** structure is stabilized by three hydrogen bonds formed by the hydrogen of the N-terminal $NH_2$ group and the oxygens of the C-terminal carboxyl group (with distances of 1.93 and 3.12 Å, respectively) and between hydrogen of first amide group and oxygen of the second histidine (1.87 Å). The pentacoordinated complex is the most stable. In the **B2** structure, copper is hexacoordinated by three nitrogen atoms, oxygens from the C-terminal carboxyl group and the side chain carboxyl groups of aspartic acid and water; it is the least stable structure in this set of complexes.

The **B3** complex is about 20 kcal/mol more stable than **B2**, and $Cu^{2+}$ in this structure is bound by the nitrogens of the imidazole rings and amino group and the C-terminal carboxyl group of the peptide. The most stable structures among $CuH_{-1}L$ complexes with β Ala are those denoted as **C3** and **C4** with pentacoordinated copper. **C1** and **C2** are about 10 kcal/mol less stable, and the nitrogen of the second imidazole is not involved in coordination of the copper ion. As far as the βAlaHisGlyHis peptide is concerned, we found only its two stable CuL forms. In the case of the **D1** structure, the 3 N $NH_2$, $2\times N_{Im}$ , O $H_2O$ equatorial coordination is observed. This structure is also stabilized by two hydrogen bonds involving hydrogens of the N-terminal amino group and oxygens belonging to the C-terminal carboxyl group and the carbonyl group of the first histidine. The **D1** complex is about 27 kcal/mol less stable in comparison with **D2**. In the latter structure, we observe that the $Cu^{2+}$ is pentacoordinated, displaying square planar equatorial coordination with three nitrogens and one oxygen belonging to the C-terminal carboxyl group and the axial oxygen from the water molecule. In addition, stabilization also comes from one hydrogen bond (1.933 Å). In the case of all structures belonging to the $CuH_{-2}L$ group, five ligands participate in cooper coordination. The data obtained suggest the existence of two different complexes. The **E1**–**E3** complexes are stabilized mainly by 4N ($NH_2$, $N_{amide}^-$, $N_{Im}$, $N_{Im}^-$), O ($H_2O$) coordination of copper, which results in a square pyramid alignment. There is an experimental evidence for formation of such a structure [28]. The **E4**–**E6** group contains complexes where $Cu^{2+}$ is coordinated by four nitrogens ($NH_2$, $N_{Im}$, $N_{amide}^-$) and oxygen ($OH^-$). Moreover, these

**Table 2** Basis set superposition error (BSSE)-corrected intermolecular interaction energy (given in kcal/mol) for $Cu^{2+}$-βXaaHisGlyHis complexes in the gas phase. Unless otherwise indicated, the results were obtained with 6-31G(d) basis set [the values given in parentheses for structures **A1**–**A3** were computed using the 6-311+G(d,p) basis set]. *HW* stands for the Hay-Wadt pseudopotential for copper [33, 34]

| | HF/HW | HF | B3LYP/HW | B3LYP | M06/HW | M06 | M06-2X/HW | M06 | M06-HF/HW | M06-HF |
|---|---|---|---|---|---|---|---|---|---|---|
| A1 | -808.29 | -810.58 | -868.23 | -870.66 | -863.95 (−836.11) | -866.86 | -861.84 | -856.33 | -854.33 | -861.11 |
| A2 | -861.40 | -862.79 | -913.15 | -916.02 | -908.88 (−879.36) | -912.83 | -912.17 | -906.92 | -910.70 | -916.12 |
| A3 | -863.17 | -864.63 | -914.48 | -917.35 | -909.80 (−881.10) | -914.34 | -913.70 | -908.49 | -912.37 | -917.87 |
| B1 | -699.58 | -701.32 | -756.92 | -759.33 | -752.17 | -756.04 | -754.56 | -747.89 | -749.29 | -754.27 |
| B2 | -719.47 | -721.04 | -768.87 | -771.01 | -763.70 | -768.28 | -769.87 | -764.15 | -767.34 | -772.50 |
| B3 | -698.82 | -701.69 | -751.03 | -755.63 | -745.92 | -752.19 | -749.67 | -745.59 | -745.91 | -753.01 |
| C1 | -607.64 | -609.22 | -708.84 | -698.04 | -705.53 | -691.02 | -688.90 | -652.04 | -653.28 | -658.35 |
| C2 | -621.80 | -580.40 | -707.89 | -699.96 | -704.40 | -693.13 | -679.03 | -669.72 | -667.28 | -672.68 |
| C3 | -712.17 | -713.51 | -773.22 | -775.18 | -769.56 | -772.88 | -767.86 | -760.30 | -762.01 | -766.51 |
| C4 | -712.39 | -713.69 | -773.38 | -775.33 | -769.81 | -773.02 | -767.95 | -760.48 | -762.24 | -766.69 |
| D1 | -504.16 | -438.53 | -575.74 | -572.74 | -571.16 | -567.73 | -562.61 | -553.16 | -552.59 | -556.47 |
| D2 | -565.40 | -566.78 | -621.01 | -622.43 | -616.93 | -619.84 | -617.84 | -610.87 | -613.37 | -618.31 |
| E1 | -832.72 | -834.77 | -890.80 | -891.12 | -887.45 | -887.81 | -883.10 | -876.15 | -874.46 | -880.59 |
| E2 | -865.59 | -867.04 | -925.60 | -928.15 | -922.25 | -926.07 | -920.09 | -912.86 | -913.41 | -918.12 |
| E3 | -788.93 | -790.64 | -858.11 | ... | -854.54 | -850.96 | -844.66 | -835.55 | -831.74 | -837.26 |
| E4 | -859.70 | -862.09 | -912.27 | -919.19 | -909.54 | -915.57 | -909.07 | -903.88 | -900.85 | -908.03 |
| E5 | -906.84 | -908.39 | -956.16 | -961.17 | -951.64 | -958.63 | -955.76 | -950.87 | -953.83 | -959.24 |
| E6 | -858.20 | -860.82 | -913.56 | -919.37 | -911.01 | -916.18 | -909.75 | -904.12 | -900.89 | -908.38 |

**Table 3** Interaction energy components (given in kcal/mol) for selected complexes calculated with the aid of the 6-31G basis set. Values computed using the 3-21G basis set are in parentheses

| | $\varepsilon_{el}^{(10)}$ | $\varepsilon_{HL}^{ex}$ | $\Delta E^{HL}$ | $\Delta E_{del}^{HF}$ | $\Delta E^{HF}$ |
|---|---|---|---|---|---|
| B1 | -686.03 (−723.54) | 209.02 (223.69) | -477.01 (−499.85) | -221.75 (−224.92) | -721.60 (−724.77) |
| C1 | …(−652.68) | …(238.32) | …(−414.36) | …(−210.49) | …(−624.85) |
| E1 | -840.23 (−875.27) | 217.41 (233.75) | -622.83 (−641.52) | -234.36 (−214.67) | -857.19 (−856.20) |
| E5 | -888.81 (−925.22) | 201.04 (215.53) | -687.77 (−709.69) | -244.78 (−227.37) | -932.54 (−937.07) |

latter complexes contain two six-membered chelate rings. Structural data for selected complexes are presented in Fig. 3.

Table 2 lists the results of calculations of intermolecular interaction energies in the gas phase between $Cu^{2+}$ and its surrounding ligands (treated as a whole monomer). The rationale for the neglect of solvent effects (although they are included partially due to the presence of the water molecule in the coordination sphere) is that evaluation of BSSE in the solvent, which is represented as dielectric medium, leads to ambiguity regarding definition of the cavity. One possibility to overcome this difficulty is to evaluate BSSE correction for the complex in the gas phase, and to assume it to be constant in the presence of solvent [58]. However, we have not attempted to use this approach in order to avoid ambiguities in cavity definitions because, in the studied systems, BSSE correction did not exceed 5% of the interaction energy value when using the 6-31G(d) basis set. As seen above, the interaction energies were quite insensitive to the functional employed. On the other hand, the differences are much larger if electron correlation is not taken into account. It is not surprising that the ordering of conformers then differs from that found in Table 1. This is because the interaction energy does not capture the effect of stabilization due to the geometry change in the polypeptide chain. Nevertheless, the values presented in Table 2 might help, although not always, to correlate the binding preferences of $Cu^{2+}$ for complexes of similar conformation with the experimental species distribution. For βAspHisGlyHis and βAlaHisGlyHis, the dominant forms were found to be CuL and $CuH_{-1}L$, respectively. Based on the data in Table 2, one finds that, in the case of βAspHisGlyHis, quantum-chemical calculations predict the CuL form to be less stable than the $CuH_{-1}L$. This disagrees with experimental observations. However, in the case of βAlaHisGlyHis both experiment and theory predict the $CuH_{-1}L$ form to be most stable. The data in Table 2 also confirm another important experimental finding, i.e., species with Asp bind copper more effectively than do species with Ala. The sources of this greater binding ability are due to larger electrostatic interactions [see Table 3 for the results of interaction energy decomposition ($Cu^{2+}$ interacting with ligands in the gas phase)]. In this context, we should mention an attempt to analyze the nature of intermolecular interactions in complexes containing $Cu^{2+}$ recently made by Comba et al. [59]. These latter authors used the KS formulation of DFT to decompose the intermolecular interaction energy into three contributions, namely electrostatic, Pauli repulsive interaction and an electronic stabilization term. As seen in Tables 3 and 4, the electrostatic component is the most important of these, and only 30–50% of its value is cancelled out by associated exchange repulsion. Interestingly, induction effects (present in the delocalization component) are constant and do not depend on the level of deprotonation. The data presented in Table 4 provide us with two important observations. Firstly, the ratio of interaction energy components does not change substantially upon basis set extension. The same is true for the total interaction energy computed using the MP2 method. Secondly, the difference in the HF and the MP2 interaction energies does not exceed 25 kcal/mol. This indicated that, in terms of magnitude, the dispersion interaction is much smaller than the electrostatic term. The reason for presenting the interaction energy components using the 3-21G basis set is that, for some structures, we were not able to converge the dimer wave function to proper solution in larger basis sets. This is because the current implementation of variational-perturbational scheme allows only for the initial guess of orbitals in the form of the Heitler-London wave function. The values of contributions presented in Tables 3 and 4 should be treated with a bit of scepticism as interfragment separations lie in the range 1.9–2.2 Å. For such distances, symmetry adapted

**Table 4** Basis set dependence of the interaction energy components (given in kcal/mol) for the complex denoted as **D1**

| | $\varepsilon_{el}^{(10)}$ | $\varepsilon_{HL}^{ex}$ | $\Delta E^{HL}$ | $\Delta E_{del}^{HF}$ | $\Delta E^{HF}$ | $\Delta E^{MP2}$ |
|---|---|---|---|---|---|---|
| HF/HW,6-31G | -504.26 | 217.89 | -286.37 | -238.04 | -524.40 | -540.33 |
| HF/HW,6-31G(d) | -433.88 | 197.64 | -236.24 | -267.68 | -503.93 | -527.79 |
| HF/HW,6-311G(d) | -428.35 | 195.99 | -232.36 | -270.03 | -502.39 | -523.83 |

perturbation theory is expected to be either not quite applicable or to provide only semiquantitative results.

## Conclusions

In this study we analyzed structural and energetic aspects of the binding of $Cu^{2+}$ to polypeptides. Numerous lowest energy conformers of βAspHisGlyHis and βAlaHisGlyHis and their deprotonated forms were described. It was found that, in most cases, Cu ion is either hexa- or pentacoordinated. Based on stability analysis, several conformers of similar energy were found for each deprotonated form. In most cases, our computations confirm experimental findings regarding the ability to bind $Cu^{2+}$ at various pH. As a part of a model study, we also made an attempt to critically assess the theoretical methods used to describe energetic properties. In particular, it was shown that the results are not sensitive to the amount of exact exchange in an exchange-correlation functional, which might suggest that a single-reference approach is sufficient to describe the studied systems [56]. However, multiconfigurational wave function- based approaches should be applied in order to unequivocally assess the reliability of the applied protocol.

**Acknowledgments** The study was supported by a research fellowship from the "Development program of Wroclaw Medical University" funded by the European Social Fund, Human Capital, national Cohesion Strategy (contract no. UDA-POKL.04.01.01-00-010/08-00). The allotment of CPU time in Wroclaw Center of Networking and Supercomputing (WCSS) is acknowledged. The authors are also grateful to Prof. Wojciech Bartkowiak and anonymous reviewers for useful remarks regarding the manuscript.



## References

1. Deschamps P, Kulkarni PP, Gautam-Basak M, Sarkar B (2005) Coord Chem Rev 249:895
2. Li F, Fitz D, Fraser D, Rode B (2010) Amino Acids 38:287
3. Jaworska M, Lodowski P, Mucha A, Szczepanik W, Valensin G, Cappannelli M, Jeżowska-Bojczuk M (2007) Bioiorg Chem Appl ID 53521:1
4. Wyttenbach T, Liu D, Bowers MT (2008) J Am Chem Soc 130:5993
5. Xu X, Yu W, Huang Z, Lin Z (2010) J Phys Chem B 114:1417
6. Joly L, Antoine R, Albrieux F, Ballivian R, Broyer M, Chirot F, Lemoine J, Dugourd P, Greco C, Mitrić R, Bonačić-Koutecký V (2009) J Phys Chem B 113:11293
7. Rezabal E, Marino T, Mercero JM, Russo N, Ugalde JM (2007) J Chem Theory Comput 3:1830
8. Leopoldini M, Russo N, Toscano M (2007) J Am Chem Soc 129:7776
9. Leopoldini M, Russo N, Toscano M (2009) Chem Eur J 15:8026
10. Marino T, Toscano M, Russo N (2007) J Phys Chem B 111:635
11. Chen SL, Marino T, Fang WH, Russo N, Himo F (2008) J Phys Chem B 112:2494
12. Marino T, Russo N, Toscano M, Pavelka M (2010) Interdiscip Sci Comput Life Sci 2:57
13. Bertrán J, Rodríguez-Santiago L, Sodupe M (1999) J Phys Chem B 103:2310
14. Georgieva I, Trendafilova N, Rodríguez-Santiago L, Sodupe M (2005) J Phys Chem A 109:5668
15. Noguera M, Bertran J, Sodupe M (2004) J Phys Chem A 108:333
16. Noguera M, Bertran J, Sodupe M (2008) J Phys Chem B 112:4817
17. Rimola A, Constantino E, Rodriguez-Santiago L, Sodupe M (2008) J Phys Chem B 112:3444
18. Lamsabhi AM, Alcamí M, Mó O, Yáñez M, Tortajada J, Salpin JY (2007) Chemphyschem 8:181
19. Lamsabhi AM, Alcamí M, Mó O, Yáñez M, Tortajada J (2004) Chemphyschem 5:1871
20. Palacios A, Corral I, Mó O, Martín F, Yáñez M (2005) J Chem Phys 123:014315
21. Trujillo C, Lamsabhi AM, Mó O, Yáñez M (2008) Phys Chem Chem Phys 10:3229
22. Lamsabhi AM, Alcamí M, Mó O, Yáñez M, Tortajada J (2006) J Phys Chem A 110:1943
23. Remko M, Fitz D, Rode B (2010) Amino Acids 39:1309
24. M Remko, D Fitz, R Broer, B Rode (2011) J Mol Model. doi:10.1007/s00894-011-1000-0
25. Poater J, Solà M, Rimola A, Rodríguez-Santiago L, Sodupe M (2004) J Phys Chem A 108:6072
26. Steer DL, Lew RA, Perlmutter P, Smith AI, Aguilar MI (2002) Lett Pept Sci 8:241
27. Steer DL, Lew RA, Perlmutter P, Smith AI, Aguilar MI (2002) Current Med Chem 9:811
28. Brasun J, Czapor H, Matera-Witkiewicz A, Kotynia A, Sochacka A, Cebrat M (2010) Dalton Trans 39:6518
29. Guell M, Luis JM, Sola M, Siegbahn PEM (2009) J Biol Inorg Chem 14:229
30. Garcia-Bosch I, Company A, Frisch JR, Torrent-Sucarrat M, Cardellach M, Gamba I, Guell M, Casella L, Que L, Ribas X, Luis JM, Costas M (2010) Angew Chem Int Ed 49:2406
31. Stewart JJP (2007) J Mol Model 13:1173
32. Stewart JJP (2009) J Mol Model 15:765
33. Hay PJ, Wadt WR (1985) J Chem Phys 82:270
34. Hay PJ, Wadt WR (1985) J Chem Phys 82:299
35. Cancès E, Mennucci B, Tomasi J (1997) J Chem Phys 107:3022
36. Tomasi J, Mennucci B, Cancès E (1999) J Mol Struct (THEOCHEM) 464:211
37. Tomasi J, Mennucci B, Cammi R (2005) Chem Rev 105:2999
38. Mennucci B, Cappelli C, Guido CA, Cammi R, Tomasi J (2009) J Phys Chem A 113:3009
39. Scalmani G, Frisch MJ, Mennucci B, Tomasi J, Cammi R, Barone V (2006) J Chem Phys 124:094107
40. Frisch MJ, et al (2009) Gaussian 09 Revision A.2. Gaussian Inc. Wallingford CT
41. Boys SF, Bernardi F (1970) Mol Phys 19:553
42. Gutowski M, Duijneveldt FB, Chałasiński G, Piela L (1987) Mol Phys 61:223
43. Sokalski WA, Roszak S, Pecul K (1988) Chem Phys Lett 153:153
44. Chałasiński G, Szczęśniak MM (1988) Mol Phys 63:205
45. Cybulski SM, Chałasiński G, Moszynski R (1990) J Chem Phys 92:4357
46. Chałasiński G, Szczęśniak MM (1994) Chem Rev 94:1723
47. Gutowski M, Piela L (1988) Mol Phys 64:337

48. Frey RF, Davidson ER (1989) J Chem Phys 90:5555
49. Hirschfelder J (1967) Chem Phys Lett 1:325
50. Moszynski R, Rybak S, Cybulski S, Chałasiński G (1990) Chem Phys Lett 166:609
51. Gora RW, Bartkowiak W, Roszak S, Leszczynski J (2002) J Chem Phys 117:1031
52. Gora RW, Bartkowiak W, Roszak S, Leszczynski J (2004) J Chem Phys 120:2802
53. Gora RW, Sokalski WA, Leszczynski J, Pett VB (2005) J Phys Chem B 109:2027
54. Zhao Y, Schultz NE, Truhlar DG (2005) J Chem Phys 123:161103
55. Zhao Y, Schultz NE, Truhlar DG (2006) J Chem Theory Comput 2:364
56. Cramer CJ, Truhlar DG (2009) Phys Chem Chem Phys 11: 10757
57. Allen FH, Kennar O, Taylor R (1983) Acc Chem Res 16:146
58. Lipparini F, Scalmani G, Mennucci B (2009) Phys Chem Chem Phys 11:11617
59. Atanasov M, Comba P, Martin B, Müller V, Rajaraman G, Rohwer H, Wunderlich S (2006) J Comp Chem 27:1263

J Mol Model (2012) 18:1375–1388
DOI 10.1007/s00894-011-1165-6

ORIGINAL PAPER

# Signals of LOV1: a computer simulation study on the wildtype LOV1-domain of *Chlamydomonas reinhardtii* and its mutants

**Emanuel Peter · Bernhard Dick · Stephan A. Baeurle**

Received: 9 May 2011 / Accepted: 24 June 2011 / Published online: 15 July 2011


**Abstract** Phototropins are photoreceptors regulating the blue-light response in plants and bacteria. They consist of two LOV (light oxygen voltage sensitive) domains each containing a non-covalently bound flavin-mononucleotide (FMN) chromophore, which are connected to a serine/threonine-kinase. Upon illumination, the LOV-domains undergo conformational changes, triggering a signal cascade in the organism through kinase activation. Here, we present results from molecular dynamics simulations in which we investigate the signal transduction pathway of the wildtype LOV1-domain of *Chlamydomonas reinhardtii* and a methyl-mercaptan (MM) adduct of its Cys57Gly-mutant at the molecular level. In particular, we analyzed the effect of covalent-bond formation between the reactive cysteine Cys57 and the FMN-reaction center, as well as the subsequent charge redistribution, on the spatio-dynamical behavior of the LOV1-domain. We compare the calculation results with experimental data and demonstrate that these adduct state characteristics have an important influence on the response of this photosensor. The light-induced changes implicate primarily an alteration of the surface charge distribution through rearrangement of the highly flexible $C\alpha$-, $D\alpha$- and $E\alpha$-helices including the Glu51-Lys91-salt bridge on the hydrophilic side of the protein domain and a $\beta$-sheet tightening process via coupling of the A$\beta$- and B$\beta$-strands. Our findings confirm the aptitude of the LOV1-domain to function as a dimerization partner, allowing the green alga to adapt its reproduction and growth speed to the environmental conditions.

E. Peter · B. Dick · S. A. Baeurle (✉)
Department of Chemistry and Pharmacy, Institute of Physical and Theoretical Chemistry, University of Regensburg,
93040 Regensburg, Germany
e-mail: stephan.baeurle@chemie.uni-regensburg.de

**Keywords** LOV1-domain of *Chlamydomonas reinhardtii* · Signal transduction mechanism · Blue-light photoreceptor protein · Phototropin · Molecular dynamics simulation

## Introduction

Light oxygen voltage sensitive (LOV)-based fusion proteins possess great potential in photosensor technology for controlling cell behavior and metabolism [1]. This has recently been impressively demonstrated by Wu et al. [2, 3] through connecting the blue light-activable LOV2-J$\alpha$-protein domain from phototropin 1 of *Avena sativa* (oat) with a mutant of the Rac1-GTPase, responsible for regulating the morphology and motility of metazoan cells. This novel protein construct enabled them to modulate Rac1-activity, with which they could bias the direction of motion of HeLa-cells. In biological systems, LOV-domains play an important role as the light-sensitive part of phototropins—the blue light photoreceptors regulating a great diversity of biological processes, e.g. phototropic plant movement, chloroplast relocation, stomatal opening, rapid inhibition of stem growth and gametogenesis, in higher plants as well as in micro-algae [4–7]. Phototropins are composed of two LOV-domains, each containing a non-covalently bound flavin-mononucleotide (FMN)-chromophore attached to a C-terminal kinase. Upon blue-light absorption, a covalent bond between the FMN-chromophore and an adjacent reactive cysteine residue of the apo-protein is formed. This subsequently mediates the activation of the kinase, which induces a signal in the organism through phototropin auto-phosphorylation [8]. A common feature of wildtype phototropins is that the photochemical reactivity of the LOV2-domain is required for the activation of the kinase. In their seminal paper, Harper et

al. [9] first demonstrated that, in the dark, the LOV2-domain interacts with a C-terminal $\alpha$-helix, known as the J$\alpha$-helix. Upon illumination, the J$\alpha$-helix dissociates from the LOV2-domain, which triggers the disruption of the inhibitory interaction between the kinase and the LOV2-domain. This process ultimately causes activation of the kinase. In contrast, the role of the LOV1-domain is still unclear, although several recent experimental investigations suggest that this photosensor may function as a site for photoreceptor dimerization [10–13] and/or dimerization with anchor proteins [14]. Other experimental studies have proposed that the LOV1-domain might be responsible for regulating the lifetime and signal intensity of the LOV2-domain under low light conditions and in this way affect phototropin activity [7, 15, 16]. Through functional analysis of domain swap proteins in a recent study, Kaiserli et al. [17] provided some evidence that the functional role of the LOV1-domain in *Arabidopsis thaliana* phot1 may reside in arresting chloroplast accumulation at high light intensities. Although tremendous progress on the functioning of the LOV2-J$\alpha$-photosensor has lately been made both from the experimental [9, 18–20] as well as theoretical [21] viewpoints, the mechanism of signal transduction at the molecular level and the functional role of the LOV1-domain still remain only poorly understood.

In the past decade, several experimental investigations have attempted to clarify the response mechanism of the LOV1-domain to blue-light photoexcitation. To elucidate the photocycle, Kottke et al. [22] studied the LOV1-domain of phot1 from *Chlamydomonas reinhardtii* with time-resolved absorption spectroscopy. They found that photoexcitation of the dark-state form of the LOV1-domain, designated as LOV1-447, causes transient bleaching and the formation of two spectrally similar red-shifted intermediates that can both be assigned to triplet states of the FMN-chromophore. These triplet states decay with time constants of 800 ns and 4 $\mu$s with an efficiency greater than 90% into a blue-shifted intermediate, designated as LOV1-390, that can be correlated with the formation of a thiol adduct between the reactive cysteine residue Cys57 of the apo-protein and the C4a-atom on FMN (FMN-C4a). Afterwards, the adduct state LOV1-390 reverts to the dark-state form in hundreds of seconds, with the time constant being dependent on pH and salt concentration. By substituting the reactive cysteine Cys57 with glycine, Lanzl et al. [23, 24] demonstrated that adduct formation in the LOV-domain is prevented and that the excited state can be stabilized through generating a photoadduct with methylmercaptan (MM). Fedorov et al. [25] performed accurate measurements of the crystal structure of the LOV1-domain from the same species in the dark and illuminated states (resolution: 1.9 Å) and discovered that blue-light absorption causes the formation of a covalent bond between the FMN-C4a and the thiol group of Cys57. They found only minor conformational changes upon adduct formation, which were restricted to the residues Cys57, Asn56 and Arg58 surrounding the FMN-chromophore. They explained the absence of large conformational changes in their X-ray crystallographic measurements as the result of the confinement of the protein in the crystal lattice. Ataka et al. [26] studied the molecular processes associated with adduct formation on the same species in solution through vibrational spectroscopy.

They concluded from their data that no large rearrangements of the protein backbone take place upon adduct formation. However, they could not exclude displacements of the amino acids near to the FMN-chromophore as well as displacements of the highly flexible peripheral $\alpha$-helices and $\beta$-strands. In a later study, Iwata et al. [27] compared the light-induced structural changes of the LOV1- and LOV2-domains from *Adiantum* phytochrome3 (phy3) by means of UV-visible and Fourier-transform-infrared (FTIR) spectroscopy. They observed the largest differences in secondary structure between phy3-LOV1 and phy3-LOV2 in the amide-I vibration. While in phy3-LOV2 the adduct formation between FMN and the reactive cysteine residue was found to cause a global conformational change in the $\beta$-sheet region, the structural change in phy3-LOV1 was only suspected to take place locally in the loop region, containing the reactive cysteine residue. Moreover, Iwata et al. concluded from their experiments that the lifetime of the adduct state in phy3-LOV1 is the longest among the photocycling LOV-domains, which may be too long for effectively controlling the kinase activity. Finally, they found that the FTIR-spectra of phy3-LOV1 resemble those of the glutamine mutant Gln1029Leu of phy3-LOV2, which lacks temperature-dependent structural changes that are characteristic of phy3-LOV2. Based on these findings, they speculated that the glutamine residue Gln1029 may have different behavior and functional role in phy3-LOV1 compared to phy3-LOV2. Guo et al. [28] performed a comparative study with transient absorption spectroscopy of the photocycles of the isolated LOV1- and LOV2-domains of phot from *C. reinhardtii*, as well as on the tandem LOV1–LOV2 constructs containing the domains in their wildtype form and in the form of Cys57Ser-mutants. They concluded from this investigation that the interaction of the two domains does not manifest itself during the lifetime of the triplet states, but changes the decay behavior of the adduct states, indicating that covalent bond formation within the LOV1-domain affects the interaction between LOV1 and LOV2.

From the theoretical side, there has been a major focus on investigation of the primary steps of photoreaction and the initial structural changes of the photocycle after blue-

light absorption. Fedorov et al. [25] performed first quantum chemical calculations with the crystal structure of the LOV1-domain from *C. reinhardtii*, which provided information about possible pathways of the photoreaction in the binding pocket. Neiss and Saalfrank [29] performed ab initio quantum mechanical calculations to determine the electronic properties of several FMN-related model compounds in different charge and spin states. They deduced from their calculation results that the reaction pathway proceeds via the triplet T1-state of the FMN, and that the process of adduct formation involves intermediate radicals rather than ionic species. Dittrich et al. [30] studied the early steps of the photocycle by performing combined quantum mechanical/molecular mechanical (QM/MM) simulations of a complete LOV1-domain from *C. reinhardtii*. They investigated the electronic properties and initial structural changes that follow blue-light absorption, and concluded from their calculation results that the pathway for cysteinyl-FMN(CFN) adduct formation proceeds via a neutral radical state, generated by H-atom transfer from the reactive Cys57-residue to the FMN-chromophore. In a later work. Freddolino et al. [31] investigated the structural changes taking place at longer times after adduct formation by performing molecular dynamics (MD) simulations with the LOV1-domain of *C. reinhardtii*. To this end, they performed five separate 12 ns-simulations with the light and dark states. Based on their calculation results, they proposed that LOV1 is activated by a change in the H-bond network between the FMN-chromophore and the surrounding apo-protein that destabilizes a highly conserved salt bridge. Although some experimental measurements [27] provide support for rearrangements of the amino acids in the loop region, including the reactive cysteine residue, the pivotal role of the salt bridge in the signal transduction process has still not been confirmed experimentally.

In the present work, we elucidate the early processes of the signal transduction pathway of the wildtype LOV1-domain from *C. reinhardtii* at the molecular level, using MD simulation methods. A particular focus is placed on understanding the effect of covalent-bond formation, as well as the charge redistribution at the FMN-chromophore, on the conformational changes of the CFN-adduct state. To this end, we compare calculation results from the wildtype LOV1-domain in its dark state and CFN-adduct state forms to the MM-adduct state of the Cys57Gly-mutant (LOV1-Cys57Gly). The latter system is characterized by the same charge distribution on the FMN-chromophore as the CFN-adduct state; however, it lacks the covalent bond between the FMN-chromophore and the apo-protein. We compare our calculation results with the available experimental data and demonstrate that covalent-bond formation upon blue-light activation has a major influence on the response of the LOV1-domain. Moreover, our study will provide new insights into the molecular mechanisms and functional role of the LOV1-domain.

## Materials and methods

### Generation of starting structures

As starting structure for our dark state simulation, we used the dark state crystal structure of the wild-type LOV1-domain from *C. reinhardtii*, which was determined by Fedorov et al. [25] through X-ray diffraction measurements (PDB-code: 1N9L) and whose tertiary structure is depicted in Fig. 1a. To create the initial structure for our simulation of the cysteinyl-FMN CFN adduct of the wildtype LOV1-domain, we employed the dark state crystal structure previously mentioned and generated a CFN-adduct by

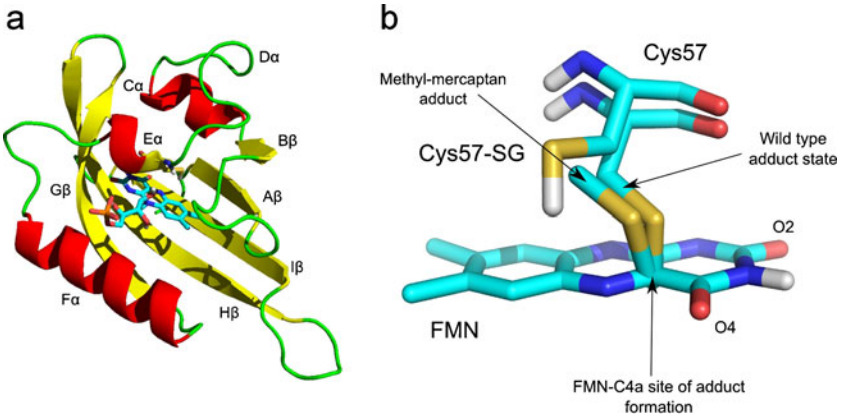



**Fig. 1** **a**,**b** Structural characteristics of the light oxygen voltage sensitive (LOV)1-domain. **a** Protein tertiary structure of the wild-type LOV1-domain, obtained by Fedorov et al. [25] using X-ray diffraction measurements [molecular dynamics (MD)-starting structure]. **b** Flavin-mononucleotide (FMN)-chromophore with surrounding Cys57-residue for the CFN-adduct state as well as methyl-mercaptan (MM)-adduct state. Atoms: *Cyan* C, *yellow* S, *white* H, *red* O, *dark blue* N

forming a covalent bond between the Cys57-S and FMN-C4a. Similarly, we created the MM-adduct of the LOV1-Cys57Gly-mutant by using the same dark state crystal structure and forming a covalent bond between the MM and FMN-C4a. This latter state differs from the CFN-adduct state solely in the lack of the covalent linkage between the reactive cysteine residue Cys57 and the FMN-chromophore (see Fig. 1b). It should be noted that covalent linkage between the reactive cysteine and FMN is required to transmit the stress from the reaction center to the protein and trigger the protein signal, as we will demonstrate through comparison of our simulations with the MM-adduct and CFN-adduct state.

### Simulation details

To resolve the signal transduction pathway from the LOV1-photosensor on a molecular level, we generated MD trajectories making use of the GROMACS MD simulation package version 3.3 in conjunction with the GROMOS43A1-forcefield [32] to describe the interactions, unless explicitly specified otherwise. This widely used forcefield has been tested against NMR-spectroscopic data in case of the hen egg white globular protein lysozyme in water by Soares et al. [33] and has been found to reproduce its solution structure and conformational behavior very well.

In a recent study, Todorova et al. [34] performed extensive MD-simulations on the 51-amino-acid protein insulin and subjected the GROMOS43A1-forcefield to a systematic comparison against other popular biomolecular forcefields, including the CHARMM27-, AMBER03-, OPLS- and GROMOS53A6-forcefields. They analyzed the effect of each forcefield on the conformational evolution and structural properties of the protein in detail and compared the results with the available experimental data. They observed that each forcefield favors different structural trends. Moreover, they found that the united-atom forcefield GROMOS43A1, together with the CHARMM27-forcefield, delivered the best description of the experimentally observed dynamic behavior of the chain B of insulin.

In our simulations we employed full particle-mesh-Ewald electrostatics with a Coulomb cut-off of 1.4 nm and computed the van der Waals (vdW)-interactions using a shift function with a cut-off of 1.4 nm. We placed the protein into a cubic box with a boxlength of 6 nm and filled it with SPC-216 water molecules as well as one supplementary sodium ion, to render the system electrically neutral. For the amino acids as well as the phosphate group at the terminal end of the oxetyl chain of the FMN, we chose the protonation states at the physiological relevant pH value of 8, which means in the latter case a fully deprotonated phosphate group. To generate an isothermal-isobaric ensemble with a temperature of 300 K and a pressure of 1 atm, the system was equilibrated for 1 ns. During this phase, every part of the system was coupled to a Nosé-Hoover-thermostat and a Parrinello-Rahman barostat [35]. We then performed a production run of 20 ns, in which the protein and FMN were decoupled from the thermostat, whereas the solvent and ions remained coupled. This procedure is known as the non-invasive thermostating technique, which allows the protein to sample configurations far from equilibrium and to follow its natural dynamics under solvent-mediated thermostating control [21, 36]. For the numerical integration of the equations of motion, we employed the leapfrog integrator with a timestep of 1 fs. To describe the interactions of CFN as well as MM, we employed the parameters of Neiss and Saalfrank [37], which were determined from B3LYP-6-31 G* -calculation results and through comparison with similar groups in the forcefield to reach consistency. The reliability of the parameterization of the CFN and FMN was tested and confirmed by these authors on the LOV2-domain from *Adiantum capillus-veneris* without the J$\alpha$-helix [37], as well as by us in various studies involving LOV1- and LOV2-domains [21, 36].

## Results and discussion

### Structural fluctuations during the signaling process

We start the analysis of our simulation results by identifying regions of the LOV1-domain of high and low mobility. To this end, we determine the root-mean-square fluctuation (RMSF), which is an indicator of the flexibility or rigidity of the protein structure [38]. The eigenvectors with the largest eigenvalues represent the collective motions that characterize the flexible degrees of freedom, whereas the smallest values represent the most constrained degrees of freedom. To clarify the role of the covalent-bond formation and changes in the electrostatic distribution on the FMN-chromophore upon adduct formation, we compare in Fig. 2a the RMSFs for the dark state and CFN-adduct state of the wildtype LOV1-domain to that of the MM-adduct state of the LOV1-Cys57Gly-mutant. By analyzing the first eigenvector, we observe that, in the range of residue numbers from 20 to 30, significant differences occur between the RMSF of the CFN-adduct state and that of the MM-adduct state. A similar deviation from the RMSF of the MM-adduct state can be observed in case of the wildtype dark state in the second eigenvector. This region corresponds to the linker region between the A$\beta$- and B$\beta$-strands.

A second major deviation occurs in the second eigenvector in the range of residues 36–60, enclosing the reactive cysteine residue Cys57, between the RMSF of the CFN-adduct state and the MM-adduct state of the mutant. The increased mobility in this region in the CFN-adduct state relates to the propagation of the stress to its surrounding

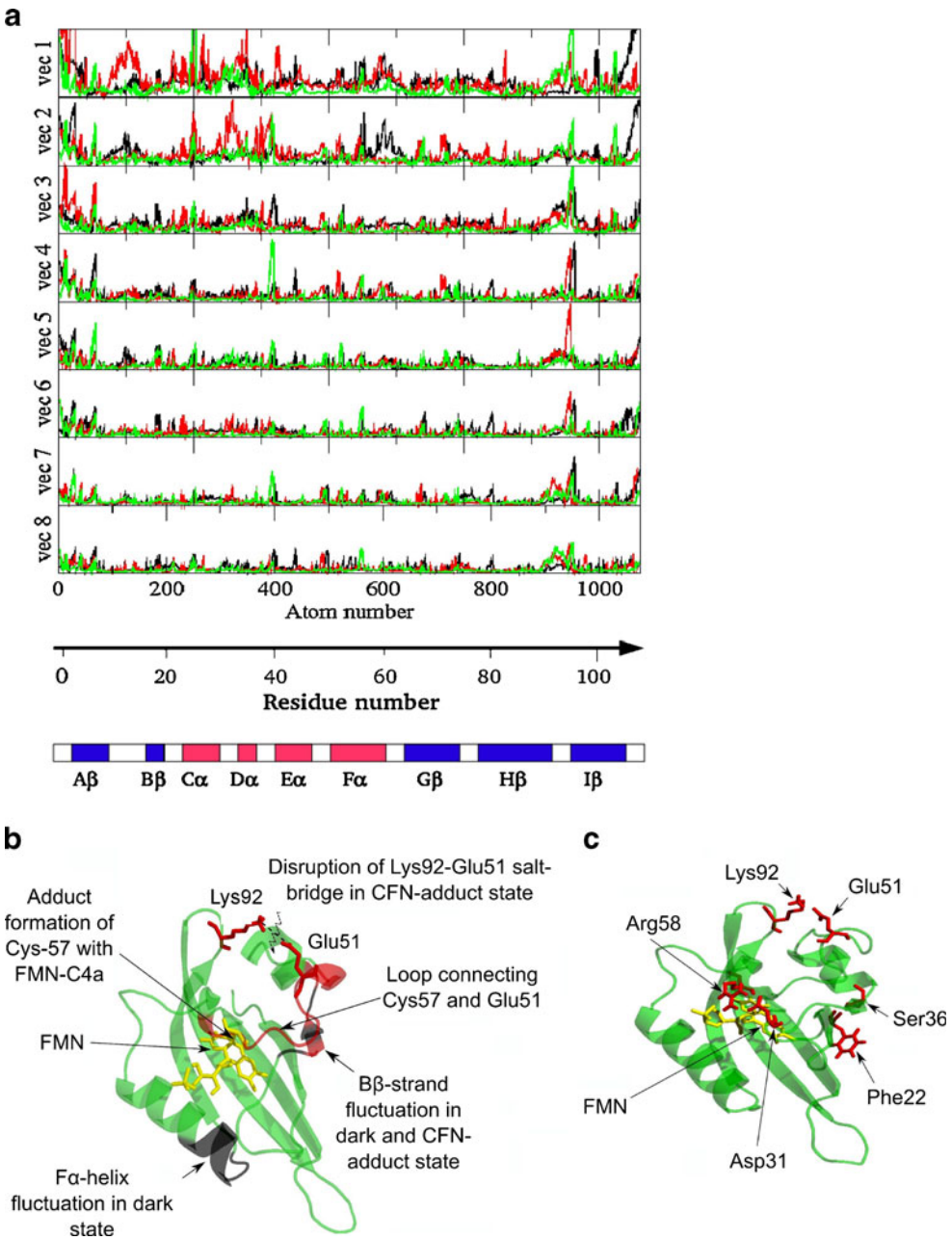



**Fig. 2** **a**–**c** Fluctuations of amino acids in the LOV1-domain. **a** Root-mean square fluctuation (RMSF) of the dark state (*black*) and CFN-adduct (*red*) state of the wildtype LOV1-domain, compared to the MM-adduct state of the LOV1-Cys57Gly-mutant (*green*), as a function of atomic number, residue number as well as secondary-structure elements. **b** Structural elements with largest change in mobility of secondary structure of the dark state (*black*) and CFN-adduct (*red*) state of the wildtype LOV1-domain. **c** Tertiary structure of the LOV1-domain with characteristic residues identified in the RMSF-analysis as being involved in the signal transduction pathway

amino-acid environment that is triggered by covalent-bond formation between Cys57 and FMN-C4a. To provide an overview of the secondary structural elements primarily affected by the adduct formation process, we visualize the regions with the largest RMSF in Fig. 2b. Generation of the CFN-adduct induces increased mobility in the *Cα*-, *Dα*- and *Eα*-helices on the hydrophilic side of the LOV1-domain, extending to a salt bridge between Glu51 and Lys92 at the surface of the protein. This observation is in agreement with the conclusions made by Iwata et al. [27] from FTIR-spectroscopic measurements on the LOV1-domain of *Adiantum* phytochrome3, who found that changes in the secondary structure take place mainly in the region around Cys57, corresponding to the turn structure. An additional difference in the RMSF can be detected in the range of residues 70–82, located on the hydrophobic side of the LOV1-domain, between the dark state and MM-adduct state of the mutant. We conclude from this observation that the change in charge distribution upon adduct formation, but without the covalent-bond constraint, in the MM-case induces a reduction in mobility of the F*α*-helix and the linker region between the F*α*-helix and the G*β*-strand.

## Signal propagation near reactive cysteine Cys57

To study the molecular details of stress propagation following covalent-bond formation and charge redistribution on the FMN-chromophore, we next consider the

distances between characteristic atoms, identified in the previous RMSF-analysis as being involved in the signal transduction pathway, as a function of simulation time (Fig. 3). The graphs compare the results of the dark state (black) and CFN-adduct state (red) to the MM-adduct state of the LOV1-Cys57Gly-mutant (green). The amino acids under consideration are represented in Fig. 2c. Figure 3a shows the distance between Asn56-N and FMN-O4 versus simulation time for all three species. The curve of the CFN-adduct state shows a sudden drop of about 2Å at a simulation time of 5,000 ps, while the curves of the dark state and MM-adduct state remain stable over the whole simulation time. This shift in the distance relates to the stress release originating from covalent-bond formation between Cys57 and FMN-C4a. The stress then continues to evolve along the backbone chain, as can be deduced from jumps in the distance between Glu51-N and FMN-O4 in Fig. 3b as well as Arg58-CZ and Asp31-OD2 in Fig. 3c of the CFN-adduct state curves. Arg58 is found to be located in the center of a H-bond network, and forms H-bonds with the FMN-phosphate as well as with Asp31. It is apparent that both curves start to deviate at around 5,000 ps, but return to their initial values at 15,000 ps after full stress relaxation. In this time range, the temporary increase in the distance between Arg58-CZ and Asp31-OD2 leads to a weakening of the H-bond between both residues. Since, by contrast, these distances in the case of the MM-adduct state do not show a similar response, we conclude that the stress propagation along the backbone chain must be a consequence of covalent-bond formation between Cys57 and FMN-C4a.

We should point out that the implication that residues Arg58 and Asp31 are involved in the signal transduction pathway of the LOV1-domain of *C. reinhardtii* was investigated by Losi et al. [39] through introducing point mutations at both sites. By performing photoacoustic measurements with these mutants, they found that substitution of Arg58 significantly influences H-bonds formed with Asp31 and the FMN-phosphate group, ultimately affecting the light-induced volume changes in the LOV1-domain. Besides Arg58, they recognized that substitution of Asp31 also exhibits well-detectable but smaller effects. Moreover, they concluded from their experiments that covalent-bond formation between Cys57 and FMN-C4a

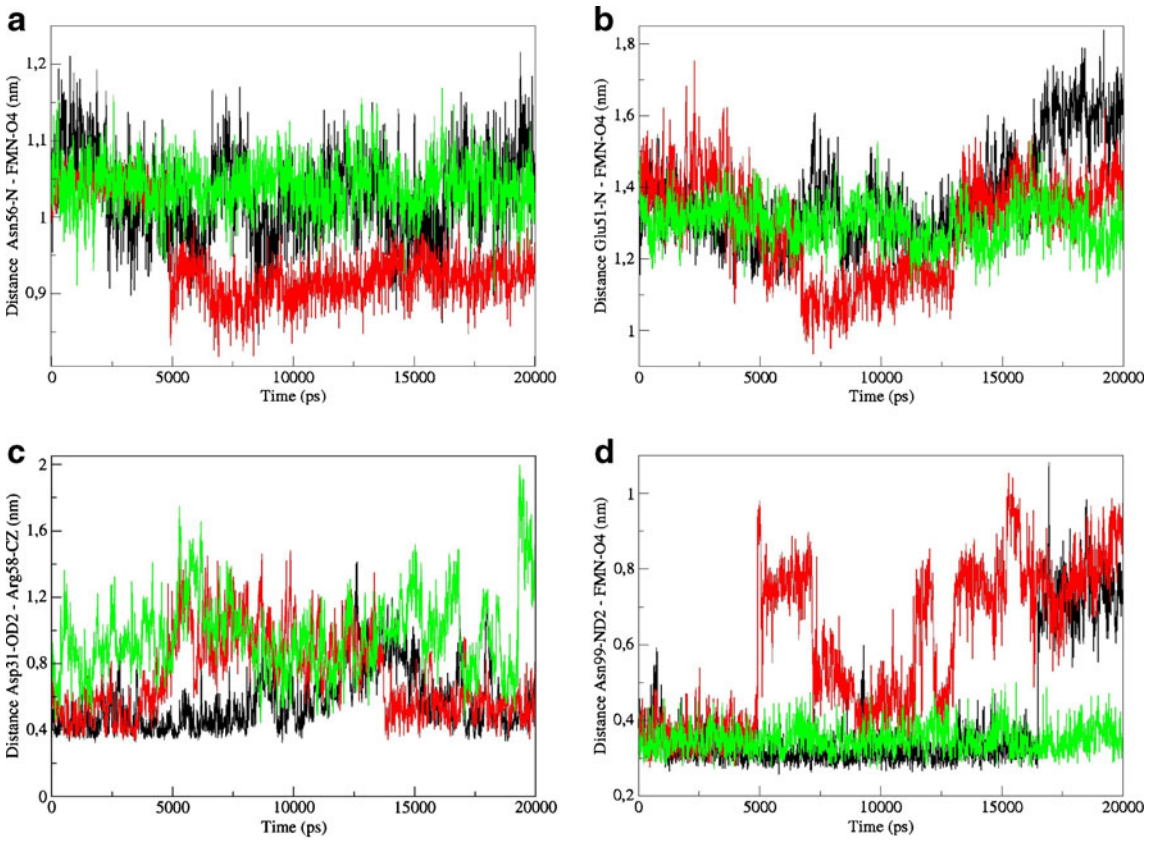



**Fig. 3** **a**–**d** Inter-atomic distances from characteristic residues surrounding the FMN-chromophore and/or located on the hydrophilic side of the LOV1-domain as a function of simulation time. **a** Asn56-N–FMN-O4. **b** Glu51-N–FMN-O4. **c** Asp31-OD2–Arg58-CZ. **d** Asn99-ND2–FMNO4. The graphs compare the curves of the dark state (*black*) and CFN-adduct (*red*) state to the MM-adduct state of the LOV1-Cys57Gly-mutant (*green*). All these residues have been identified in RMSF-analysis as being involved in the signal transduction pathway

causes a larger displacement of Asn56, which follows the Cys57-movement toward the FMN-chromophore, and brings it closer to Arg58. This goes along with the breakage of the H-bond between Asn99-ND2, which is attached to the $H\beta$-strand, and FMN-O4, as can be observed in Fig. 3d. This rearrangement of amino acids helps stabilizing the photoadduct and is reversed during the dark state recovery reaction. This conclusion agrees well with the shifts and the subsequent recovery of the distances between Asn56-N and FMN-O4 as well as between Glu51-N and FMN-O4, observed in the Fig. 3a and b, respectively.

Because the H-bond between Arg58 and Asp31 connects the peripheral $A\beta$- and $B\beta$-strands to the coil region between the $E\alpha$ and $D\alpha$-helices, we conclude that its disruption in the adduct state can lead to a large change in the volume of the protein domain, which results from covalent-bond formation between Cys57 and FMN-C4a. By contrast, in the case of the MM-adduct state of the LOV1-Cys57Gly-mutant, no abrupt changes are observable for the amino acids previously mentioned over the entire simulation run. This indicates that, if the covalent bond between Cys57 and FMN-C4a is lacking, no stress can be built up and propagate along the backbone chain that would permit signal transduction to the peripheral part of the protein. Moreover, it demonstrates that the light-induced spatio-dynamical changes are affected only weakly by a change in the charge distribution on the FMN-chromophore. Similar observations can be made in the case of the dark state. The jump in distance between Glu51-N and FMN-O4 observed around 17,000 ps in Fig. 3b, is caused by displacement of the chromophore in the binding pocket, and therefore cannot be attributed to displacement of the backbone chain, as in case of the adduct state. This is confirmed by a similar jump appearing at the same simulation time in the dark state curve for the distance between Asn99-ND2 and FMN-O4 (Fig. 3d). Therefore, we conclude that the formation of a covalent bond between Cys57 and FMN-C4a induces a characteristic signal, transmitted via stress propagation through the outer periphery of the protein domain, causing a change in its surface properties.

## Salt bridge between Glu51 and Lys92

Previous studies [31, 40] speculated that a salt bridge between Glu51 and Lys92 is involved in the signaling process of the LOV1-domain. To investigate this issue in more detail, we plot in Fig. 4 histograms of the distance between Glu51 and Lys92, performed over different time ranges. Figure 4a compares histograms, taken over the entire simulation run, of the dark state (black) and CFN-adduct state (red) to the MM-adduct state of the LOV1-Cys57Gly-mutant. We see that the dark state curve has one major peak with a maximum centered around a distance of 0.23 nm, as well as a broad tail at larger distances with much smaller magnitude. By contrast, the corresponding histograms of the CFN-adduct and MM-adduct states show two main peaks of almost similar height centered around 0.23 nm and 0.43 nm, which indicates that the salt bridge is partly dissociated in these cases. The agreement of the histograms of both species is rather surprising, since in the case of the mutant no stress is propagated towards the surface that would permit dissociation of the salt bridge. To analyze this issue in more detail, we consider in Fig. 3c the distance characterizing the H-bond between Arg58-CZ and Asp31-OD2 of the CFN-adduct and MM-adduct states. Comparing both curves reveals that, in the case of the mutant, the distance fluctuates with large magnitude, which relates to the fact that Arg58 is rather flexible in the binding pocket and leads to a permanent rupture of the H-bond

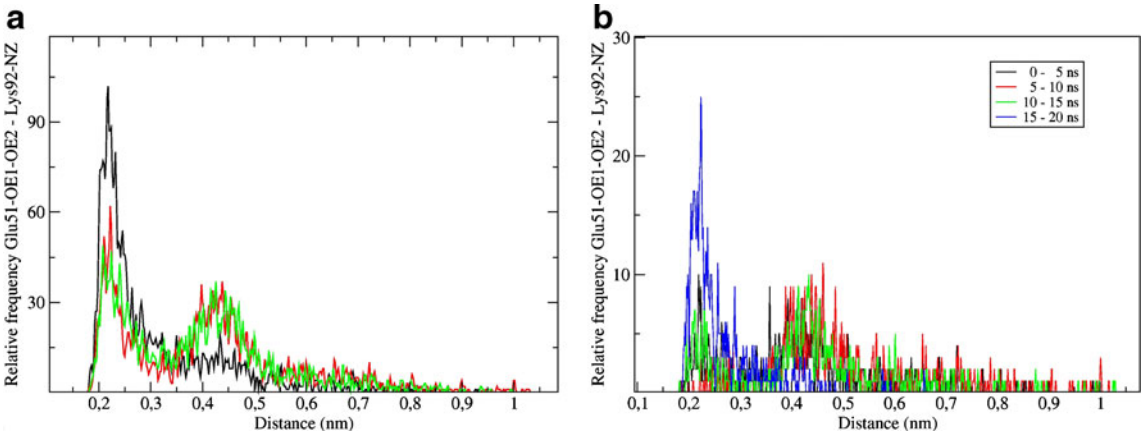



**Fig. 4** **a**,**b** Histogram of the distance between Glu51 and Lys91, characterizing a salt bridge on the hydrophilic side of the LOV1-domain, in different time ranges. **a** 20 ns-trajectory for the dark state (*black*), CFN-adduct state (*red*) and MM-adduct state of the LOV1-Cys57Gly-mutant. **b** CFN-adduct state in the time range between 0–5 ns (*black*), 5–10 ns (*red*), 10–15 ns (*green*) and 15–20 ns (*blue*)

between Arg58-CZ and Asp31-OD2. In the case of the dark state curve, we observe a moderate jump in distance starting at around 8,000 ps, which is much smaller than the jump of the adduct state curve that takes place at 5,000 ps. The increase in the former case correlates with a temporary shortening of the distance between Cys57 and FMN-C4a. This confirms the conclusions of Gardner et al. [41, 42], made from solution NMR-measurements on the LOV2-J$\alpha$-system, that the conformations dominant in the lit state also exist in the dark state, but with lower statistical weight. In our case, the pseudo-lit state conformations can be considered as high-energy dark state conformations, where the Cys57 is temporarily attracted by the FMN-C4a through H-bond formation between the H-atom of the thiol group and FMN-N5.

We next study the time-behavior of the Glu51-Lys92-salt bridge during the signaling process of the LOV1-domain. Figure 4b shows the histogram of the distance between Glu51 and Lys92 of the CFN-adduct state in the time ranges 0–5 ns (black), 5–10 ns (red), 10–15 ns (green) as well as 15–20 ns (blue). Comparing these curves to the histograms of the dark and adduct states in Fig. 4a carried out over the entire simulation run, we conclude that, in the initial phase of the stress propagation, i.e., 0–15 ns, two peaks are generated with an emphasis on the larger distance. This shows that, in this time range, the salt-bridge is essentially broken. Moreover, we observe that in the subsequent time range, i.e., 15–20 ns, the histogram essentially adopts the pattern of the dark state curve in Fig. 4a with a main maximum at around 0.23 nm. This demonstrates that, after 15 ns, the stress has relaxed fully and the salt-bridge readopts its dark state conformation. By further comparing these results to the dark and adduct state curves in Fig. 3, we conclude that the stress originating from the CFN-adduct formation in the binding pocket propagates along the residues on the backbone chain upward and downward in sequence relative to Cys57. This causes dissociation of the Glu51-Lys92-salt bridge at the surface of the LOV1-domain, confirming its implication in the signal transduction pathway, as suggested in various experimental and theoretical studies. Crosson et al. [40] speculated that the conservation of chemically linked residues extending from the FMN to this surface salt bridge is an indication of the involvement of this protein region in the signaling process. Furthermore, they pointed out that, even if the salt bridge persists only 50% of the time, it may still be useful to modulate the structure and dynamics of the B$\alpha$-A$\alpha$'-helices as well as the C$\beta$-D$\beta$-loop in the LOV-domain. Slight shifts in the stability of the bridge upon adduct formation could serve to increase or decrease the structural mobility of the B$\alpha$-A$\alpha$'-helices as well as the C$\beta$-D$\beta$-loop and, thus, affect the interaction with the LOV-partner domains.

To substantiate their hypothesis, they proposed two distinct models for describing light-driven changes in the binding affinity between the LOV-domain and its partner domains, involving the conserved surface salt bridge previously mentioned [40]. In an enthalpically driven interaction model, the process of adduct formation destabilizes the salt bridge and leads to the adoption of a distinct conformation associated with different affinities for different interacting LOV-partners. By contrast, in an entropically driven interaction model, the same process also causes the destabilization of the salt bridge, but does not change the average conformation of the LOV-domain. Instead, it increases its conformational flexibility, resulting in a different affinity for its interacting partners. Although the signaling process of the LOV-domains is likely to combine features of the two models, both favor a dynamic regulation model, where the dynamical state of the LOV-domain determines the interaction with its partner domains. In a later study, Freddolino et al. [31] investigated the structural changes taking place after formation of the CFN-adduct, by performing MD-simulations with the LOV1-domain of *C. reinhardtii*. Based on their simulation results, they proposed that LOV1-activation is induced by a change in the H-bond network between the FMN-chromophore and the surrounding apo-protein, which causes the disruption of the highly conserved salt bridge between Glu51 and Lys91 at the surface of the LOV1-domain. By contrast, our simulation results demonstrate that, although this H-bond network plays an important role, the major influence on the behavior of the Glu51-Lys91-salt bridge results from the covalent-bond formation between Cys57 and FMN-C4a in the CFN-adduct state. We find that the conformational stress generated at the CFN-bond reaches the Glu51-Lys91-salt bridge mainly through the direct route along the backbone chain down to the residue Glu51, as well as to a smaller extent through the route via the H-bond between Arg58 and Asp31.

## Signal propagation on peripheral $\alpha$-helices

To assess how adduct formation affects the secondary structural elements on the hydrophilic side of the LOV1-domain, we next consider in Fig. 5 the center of mass distances between different $\alpha$-helical elements. Figure 5a displays the distance between the E$\alpha$- and C$\alpha$-helices of the CFN-adduct state compared to those of the dark state and MM-adduct state. We observe that the curve of the CFN-adduct state, after an initial slow decrease, stabilizes after 5,000 ps around an average value of 1.35 nm. By contrast, in case of the dark state and MM-adduct state, the quantity remains stable over the whole simulation run and deviates in average from the CFN-adduct state curve by about 0.2 nm. A similar behavior is observed in Fig. 5b, which

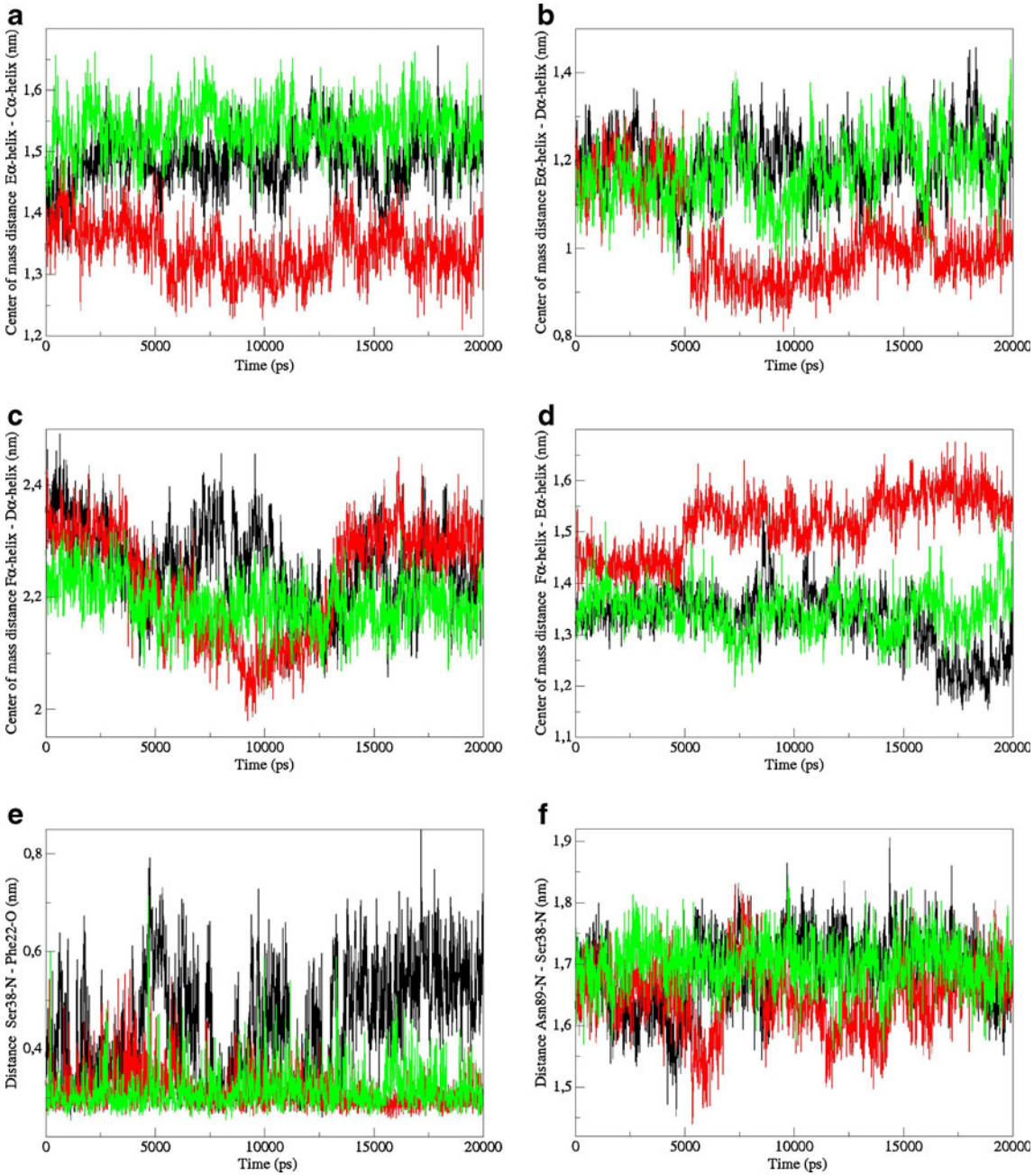



**Fig. 5** **a**–**f** Center of mass distances between $\alpha$-helices on the hydrophilic side and inter-atomic distances from characteristic residues on $\beta$-strands of the dark state (*black*) and CFN-adduct (*red*) state of the wildtype LOV1-domain, compared to the MM-adduct state of the LOV1-Cys57Gly-mutant (*green*). **a** E$\alpha$-helix–C$\alpha$-helix. **b** E$\alpha$-helix–D$\alpha$-helix. **c** F$\alpha$-helix–D$\alpha$-helix. **d** F$\alpha$-helix–E$\alpha$-helix. **e** Ser38-N–Phe22-O. **f** Asn89-N–Ser38-N

visualizes the center of mass distance between the E$\alpha$- and D$\alpha$-helix. We should point out that the D$\alpha$-helix, characterized by residues 49–52, contains Glu51, which forms a salt bridge with Lys92. While the curves of the dark state and MM-adduct state show a stable behavior over the whole simulation run, we see that the curve of the CFN-adduct state undergoes an abrupt drop in distance at around 5,000 ps, which is followed by a slow increase in quantity, indicating a systematic recovery up to the end of the simulation run. Similar perturbations at the same simulation time can also be detected in the distance curves between the F$\alpha$-and D$\alpha$-helix (Fig. 5c), as well as between the E$\alpha$- and F$\alpha$-helix (Fig. 5d).

Comparing these observations with those from the histograms over 5-ns time intervals of the distance between Glu51 and Lys92 in Fig. 4b, we conclude that the

rearrangement at 5,000 ps and subsequent conformational relaxation of the $\alpha$-helical elements correlates with the breakage and subsequent recovery of the Glu51–Lys92 salt bridge, confirming that the latter is indeed involved in the signal transduction process, as suggested in the experimental and theoretical studies discussed previously. We conclude from these observations that covalent bond formation in the CFN-adduct state significantly affects the hydrophilic region of the LOV1-domain, causing the breakage of the Glu51–Lys92 salt bridge and leading to increased mobility of the $\alpha$-helical elements. The larger fluctuations in the hydrophilic region also imply a significant change in the surface charge distribution, causing an attenuation of its aptitude to function as an aggregation partner to other protein domains, such as the LOV2-domain or the kinase. The possibility of the LOV1-domain to act as an attenuator of kinase activity under low light-conditions and reduce its inhibitory effect under high light-conditions has been confirmed by Matsuoka and Tokutomi [43], who performed in vitro phosphorylation experiments with various constructs of bacterially expressed phototropin 2 from *Arabidopsis*. In *C. reinhardtii*, the hydrophilic side of the LOV1-domain might play a similar role in its sexual life cycle [44] by adapting reproduction speed and growth to lighting conditions. Finally, it is also worth pointing out that the importance of the peripheral $\alpha$-helices in light-induced regulation of the aggregation tendency with partner domains has also recently been demonstrated in the case of the BLUF-domain by Nemukhin et al. using MD simulations [45].

## Signal propagation on $\beta$-strands

We should emphasize that the LOV1-domain is involved not only in regulating the signal intensity of phototropin by interacting with the LOV2-domain or the kinase through its hydrophilic side, but has also recently been found to act as a docking partner through face-to-face association with $\beta$-scaffolds of LOV1-domains from other phototropins [46]. The in vitro experiments of Briggs et al. [47] with phototropins from *Arabidopsis* have indicated that cross-phosphorylation can occur between functional phot2 and inactivate phot1 by means of hetero-dimerization through hydrophobic interaction via the $\beta$-sheets. Based on these findings, a response mechanism to account for the phototropic adaptation of the plant to moderate light intensities has been proposed. Recent in vitro experiments by Kutta et al. [13] on the LOV1-domain from *C. reinhardtii* established that adduct formation results in a decrease in the monomer/dimer-ratio; these authors concluded that dimerization of phototropin might play a role in the signal transduction pathway. However, to date it is still unknown whether dimerization might also have a modulating effect on the phototropin function in vivo [47].

To assess how adduct formation affects the dimerization tendency of the LOV1-domain, we next focus on fluctuations of the $\beta$-strands on its hydrophobic side. The RMSF-analysis in Fig. 2 indicates that large deviations are observable in the first eigenvector of the CFN-adduct state and the second eigenvector of the dark state, compared to the MM-adduct state, in the range of residues located between the A$\beta$- and B$\beta$-strands. To analyze this issue in more detail, we next consider the distances of characteristic residues involved within the hydrophobic region of the protein. For this purpose, we first regard the distance between Ser38-N and Phe22-O of the dark state and CFN-adduct state as well as MM-adduct state (Fig. 5e). In the crystal structure, both residues form a H-bond, which connects the A$\beta$-strand and B$\beta$-C$\alpha$-loop. From this figure, we deduce that the distance between these two residues in the dark state is subjected to large fluctuations, whereas the H-bond remains stable in the MM-adduct state. By contrast, in case of the MM-adduct state, curve fluctuations with increased magnitude are observed only within the first 7,500 ps after adduct formation, until the stress has relaxed fully and the H-bond is stabilized. This indicates that the fluctuations of the A$\beta$- and B$\beta$-strands are influenced essentially by the charge redistribution on the FMN-chromophore, but not by the formation of the covalent bond of the CFN. Figure 5f plots the distance between the backbone atoms Asn89-N and Ser38-N, which are part of the G$\beta$- and B$\beta$-strands, respectively. The comparison of the results of all three species demonstrate that neither the covalent-bond formation nor the charge redistribution of the CFN-adduct significantly affect the interaction between the G$\beta$- and B$\beta$-strands. In conclusion, we see that the charge redistribution on the FMN-chromophore is primarily responsible for the damping of the fluctuations of the $\beta$-scaffold through tightening of the A$\beta$- and B$\beta$-strands. Figure 6 shows the arrangement of the $\beta$-strands after 20 ns in the case of the dark state (Fig. 6a,c) as well as CFN-adduct state (Fig. 6b,d). The largest differences between both states are seen in the region around the B$\beta$-strand. In the CFN-adduct state, the B$\beta$-strand is significantly closer to the A$\beta$-strand compared to the dark state, which is consistent with the light-induced $\beta$-sheet tightening process discussed previously. The latter process goes along with an increase of the number of vdW contacts, which can be formed in a dimer between the side chains of residues located on the $\beta$-sheet surfaces of the LOV1-domains, allowing a stronger hydrophobic interaction and increased dimerization tendency. These findings agree well with the experimental results of Kutta et al. [13], who found that adduct formation results in an increase in the monomer interaction within dimers, indicating that dimerization might play a role in the signal transduction pathway. Similar con-

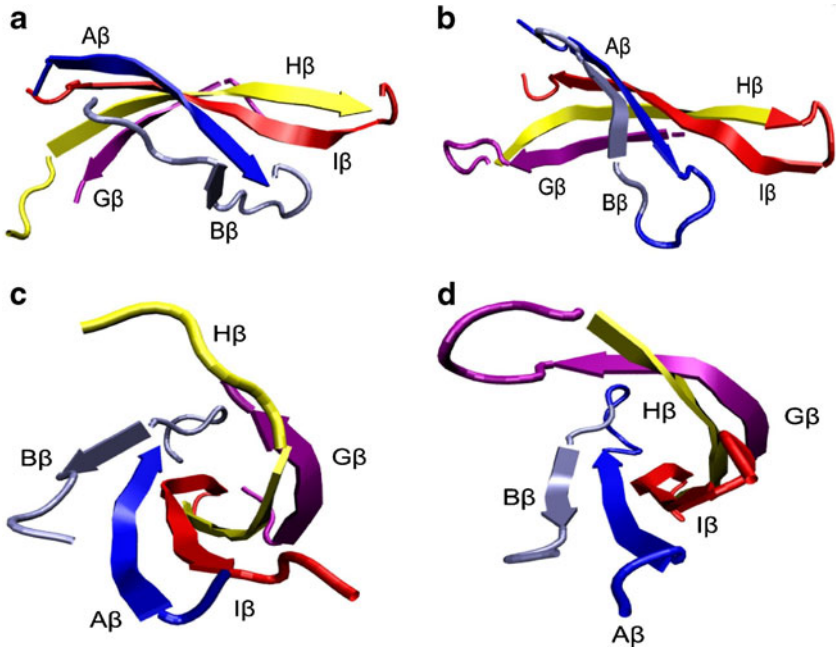



**Fig. 6** Arrangement of $\beta$-strands of the LOV1-domain at the final simulation time of 20 ns. **a** Dark state, side view. **b** CFN-adduct state, side view. **c** Dark state, front view. **d** CFN-adduct state, front view

clusions have also been drawn by Nakasoka et al. [46] through performing small-angle X-ray scattering measurements on the LOV1-domain of phot1 and phot2 from *Arabidopsis*. They established that dimeric association is essentially dominated by vdW-contacts, in addition to some H-bonds, formed between side chains of residues on the $\beta$-scaffolds, stabilizing the dimer structures. Interestingly, similar findings have been made by Iwata et al. [48] on the LOV2-domain of *Adiantum* Phytochrome3 using FTIR- and UV-visible- spectroscopy. They concluded from their experiments that immediately upon formation of the CFN-adduct the system loses the local structure of amino acids surrounding the FMN-chromophore. In subsequent stages, further changes take place, mainly to tighten the $\beta$-sheets as well as to alter the arrangement of $\alpha$-helices on the hydrophilic side of the protein.

## Structural evolution over long times

To assess how covalent bond formation influences the evolution and stability of the secondary structure over time, we next consider secondary structure analysis of all three species in Fig. 7a–c, i.e., dark state, CFN-adduct state, and the MM-adduct state of the LOV1-Cys57Gly-mutant. Whereas the secondary structures of the MM-adduct state and dark state remain almost unaffected by fluctuations, the CFN-adduct state shows large fluctuations in the flexible $\alpha$-helical elements in the initial phase up to 15,000 ps after

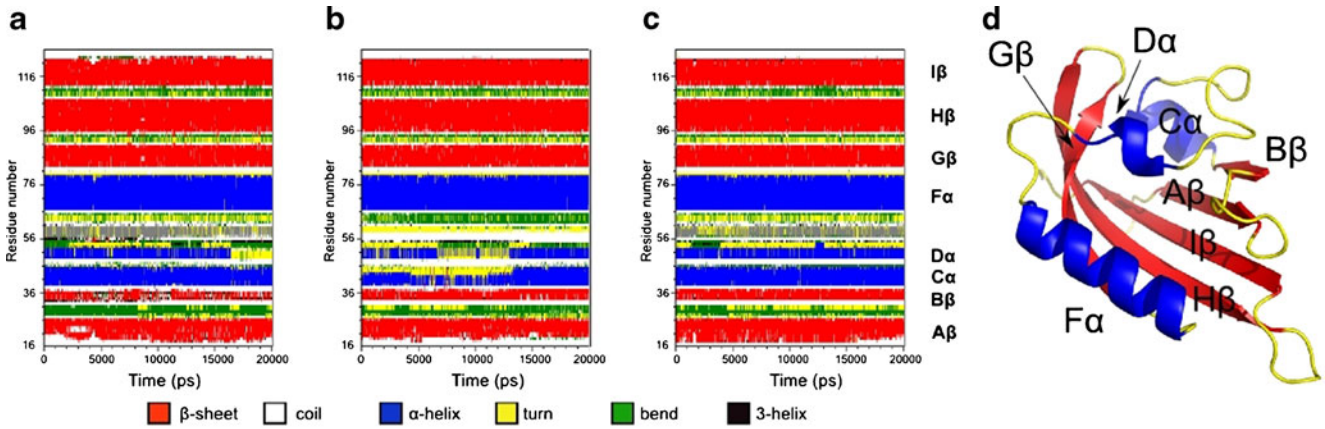



**Fig. 7** Secondary structural analysis from MD simulations. **a** Dark state. **b** CFN-adduct state of the wildtype LOV1-domain. **c** MM-adduct state of the LOV1-Cys57Gly-mutant. **d** Secondary structure elements of the LOV1-domain

adduct formation. In this time range, the CFN-adduct state undergoes significant changes in the hydrophilic region of the protein, extending from the Dα-helix to the Cα-helix. In this context, we should point out that, in the case of the dark state, the Dα-helix also disappears in the time range from 17,000 ps to 20,000 ps, which relates to the formation of a second dark state conformer characterized by a different orientation of the Cys57-residue with respect to the FMN-chromophore. Finally, it is also apparent that the CFN-adduct state shows stronger fluctuations in the Aβ- and Bβ-strands than the dark state as well as the MM-adduct state.

These results demonstrate that it is the formation of the covalent bond upon photoexcitation that mainly induces fluctuations in the residue range, including the Dα-helix and Cα-helix as well as the Aβ-strand. Moreover, we see that the lack of this bond in case of the MM-adduct state leads to a structural stabilization, whereas the charge redistribution on the FMN-chromophore upon adduct formation does not have a significant effect on structural changes. This indicates that the dynamics of signal transduction of the CFN-adduct state is affected primarily by covalent bond formation and to only a minor extent by charge redistribution on the FMN-chromophore.

## Conclusions

This study has elucidated the early stages of the signal transduction pathway of the LOV1-domain of *C. reinhardtii* upon adduct formation at the molecular level using the MD simulation method. To this end, we compared calculation results of the dark state and cysteinyl-FMN state of the wildtype LOV1-domain to a MM-adduct state of its Cys57Gly mutant. This latter system has been generated in recent experiments [23, 24] by replacing the reactive cysteine residue Cys57 with glycine as well as producing an adduct photochemically through addition of MM and irradiation with blue light. Our simulation results demonstrate that covalent-bond formation between Cys57 and FMN-C4a in the adduct state primarily affects the hydrophilic side of the LOV1-domain. More specifically, we find that the conformational stress, generated at Cys57, propagates mainly to the Glu51–Lys91 salt bridge through the direct route along the backbone chain down to the residue Glu51, as well as to a smaller extent through the route via the H-bond between Arg58 and Asp31. This results in increased mobility of the *Cα*-, *Dα*- and *Eα*-helices on the hydrophilic side of the LOV1-domain. The larger fluctuations in the hydrophilic region also imply a significant change in the surface charge distribution, causing a decrease in its aptitude to function as an aggregation partner to other protein domains, such as the LOV2-domain or the kinase. By contrast, the hydrophobic side of the protein is essentially influenced by both the covalent-bond formation between Cys57 and FMN-C4a as well as a change in the charge distribution on the FMN-chromophore. This induces a β-sheet tightening process via coupling of the Aβ-and Bβ-strands. The latter process goes along with an increase of the number of van-der-Waals contacts, which can be formed within a dimer between side chains of residues located on the β-sheet surfaces of the LOV1-domains, allowing a stronger hydrophobic interaction and increase of the dimerization tendency. Therefore, our findings confirm the aptitude of the LOV1-domain to function as a dimerization site, to allow cross-phosphorylation of the phototropins, and thus permit the green alga to adapt its reproduction and growth speed to the lighting conditions. We anticipate that elucidation of the LOV1-signal transduction pathway on a molecular level will provide new perspectives for the development of novel photoactivable fusion proteins with adjustable binding affinity to substrates or membranes [49]. The fine-tuning could be done through the creation of mutants of the LOV1-domain with optimized light-induced structural responses, which might open new avenues for allosteric control of protein activity in living cells [50–52].

**Acknowledgments** The work was supported by the Deutsche Forschungsgemeinschaft (DFG) through the Graduate College GRK 640 “Sensory photoreceptors in natural and artificial systems” as well as the Graduate College GRK 1626 “Chemical photocatalysis”.

## References

1. Scott JD, Pawson T (2009) Signaling through scaffold, anchoring, and adaptor proteins. Science 326:1220–1224
2. Wu YI, Frey D, Lungu OI, Jaehrig A, Schlichting I, Kuhlman B, Hahn KM (2009) A genetically encoded photoactivable Rac controls the motility in living cells. Nature 461:104–108
3. Kaganman I (2009) ’Blue’ lighting cell signaling research. Nat Methods 6:694–695
4. Hegemann P (2008) Algal sensory photoreceptors. Annu Rev Plant Biol 59:167–189
5. Christie JM (2007) Phototropin blue-light receptors. Annu Rev Plant Biol 58:21–45
6. Briggs WR (2007) The LOV domain: a chromophore module servicing multiple pho-toreceptors. J Biomed Sci 14:499–504
7. Kottke T, Hegemann P, Dick B, Heberle J (2006) The photochemistry of the LOV domains in the algal blue light receptor phot. Biopolymers 82:373–378
8. Jones MA, Feeney KA, Kelly SM, Christie JM (2007) Mutational analysis of phototropin 1 provides insights into the mechanism underlying LOV2 signal transmission. J Biol Chem 282:6405–6414
9. Harper SM, Neil LC, Gardner KH (2003) Structural basis of a phototropin light switch. Science 301:1541–1544
10. Salomon M, Lempert U, Rüdiger W (2004) Dimerization of the plant photoreceptor phototropin is probably mediated by the LOV1 domain. FEBS Lett 572:8–10

11. Nakasako M, Iwata T, Matsuoka D, Tokutomi S (2004) Light-induced structural changes of LOV domain-containing polypeptides from *Arabidopsis* phototropin 1 and 2 studied by small-angle X-ray scattering. Biochemistry 43:14881–14890
12. Briggs WR, Tseng T-S, Cho H-Y, Swartz TE, Sullivan S, Bogomolni RA, Kaiserli E, Christie JM (2007) Phototropins and their LOV domains: versatile plant blue-light receptors. J Integr Plant Biol 49:4–10
13. Kutta RJ, Hofinger ESA, Preuss H, Bernhardt G, Dick B (2008) Blue-light induced in-teraction of LOV domains from *Chlamydomonas reinhardtii*. ChemBioChem 9:1931–1938
14. Christie JM, Swartz TE, Bogomolni RA, Briggs WR (2002) Phototropin LOV domains exhibit distinct roles in regulating photoreceptor function. Plant J 32:205–219
15. Kagawa T, Kasahara M, Abe T, Yoshida S, Wada M (2004) Function analysis of phototropin2 using fern mutants deficient in blue light-induced chloroplast avoidance movement. Plant Cell Physiol 45:416–426
16. Sullivan S, Thomson CE, Lamont DJ, Jones MA, Christie JM (2008) In vivo phosphorylation site mapping and functional characterization of *Arabidopsis* phototopin 1. Mol Plant 1:178–194
17. Kaiserli E, Sullivan S, Jones MA, Feeney KA, Christie JM (2009) Domain swapping to assess the mechanistic basis of *Arabidopsis* phototropin 1 receptor kinase activation and endocytosis by blue light. Plant Cell 21:3226–3244
18. Harper SM, Christie JM, Gardner KH (2004) Disruption of the LOV-J$\alpha$ helix interaction activates phototropin kinase activity. Biochemistry 43:16184–16192
19. Halavaty AS, Moffat K (2007) N- and C-terminal flanking regions modulate light-induced signal transduction in the LOV2 domain of the blue light sensor phototropin 1 from *Avena sativa*. Biochemistry 46:14001–14009
20. Nash AI, Ko W-H, Harper SM, Gardner KH (2008) A conserved glutamine plays a central role in LOV domain signal transmission and its duration. Biochemistry 47:13842–13849
21. Peter E, Dick B, Baeurle SA (2010) Mechanism of signal transduction of the LOV2-J$\alpha$ photosensor from *Avena sativa*. Nat Commun 1:122
22. Kottke T, Heberle J, Hehn D, Dick B, Hegemann P (2003) Phot-LOV1: photocycle of a blue-light receptor domain from the green alga *Chlamydomonas reinhardtii*. Biophys J 84:1192–1201
23. Lanzl K, Nöll G, Dick B (2008) LOV1 protein from *Chlamydomonas reinhardtii* is a template for the photoadduct formation of FMN and methylmercaptan. ChemBioChem 9:861–864
24. Lanzl K, von Sanden-Flohe M, Kutta R-J, Dick B (2010) Photoreaction of mutated LOV photoreceptor domains from *Chlamydomonas reinhardtii* with aliphatic mercap-tans: implications for the mechanism of wild type LOV. Phys Chem Chem Phys 12:6594–6604
25. Fedorov R, Schlichting I, Hartmann E, Domratcheva T, Fuhrmann M, Hegemann P (2003) Crystal structures and molecular mechanism of a light-induced signaling switch: the phot-LOV1 domain from *Chlamydomonas reinhardtii*. Biophys J 84:2474–2482
26. Ataka K, Hegemann P, Heberle J (2003) Vibrational spectroscopy of an algal phot-LOV1 domain probes the molecular changes associated with blue-light reception. Biophys J 84:466–474
27. Iwata T, Nozaki D, Tokutomi S, Kandori H (2005) Comparative investigation of the LOV1 and LOV2 domains in *Adiantum* Phytochrome3. Biochemistry 44:7427–7434
28. Guo H, Kottke T, Hegemann P, Dick B (2005) The phot LOV2 domain and its interaction with LOV1. Biophys J 89:402–412
29. Neiss C, Saalfrank P (2003) *Ab initio* quantum chemical investigation of the first steps of the photocycle of phototropin: a model study. Photochem Photobiol 77:101–109
30. Dittrich M, Freddolino PL, Schulten K (2005) When light falls in LOV: a quantum mechanical/molecular mechanical study of photoexcitation in phot-LOV1 of *Chlamydomonas reinhardtii*. J Phys Chem B 109:13006–13013
31. Freddolino PL, Dittrich M, Schulten K (2006) Dynamic switching mechanisms in LOV1 and LOV2 domains of plant phototropins. Biophys J 91:3630–3639
32. Lindahl E, Hess B, van der Spoel D (2001) GROMACS 3.0: a package for molecular simulation and trajectory analysis. J Mol Model 7:306–317
33. Soares T, Daura X, Oostenbrink C, Smith L, Gunsteren W (2004) Validation of the GROMOS force-field parameter set 45A3 against nuclear magnetic resonance data of hen egg lysozyme. J Biomol NMR 30:407–422
34. Todorova N, Legge FS, Treutlein H, Yarovsky I (2008) Systematic comparison of empirical forcefields for molecular dynamic simulation of insulin. J Phys Chem B 112:11137–11146
35. Frenkel D, Smit B (2003) Understanding molecular simulation: from algorithms to applications. Academic, San Diego
36. Peter E, Dick B, Baeurle SA (2011) Effect of the computational methodology on the conformational dynamics of the protein photosensor LOV1 from *Chlamydomonas reinhardtii*. J Chem Biol. doi:10.1007/s12154-011-0060-z
37. Neiss C, Saalfrank P (2004) Molecular dynamics simulation of the LOV2 domain from *Adiantum capillus-veneris*. J Chem Inf Comput Sci 44:1788–1793
38. Tatke SS, Loong CK, D'Souza N, Schoephoerster RT, Prabhakaran M (2008) Large scale motions in a biosensor protein glucose oxidase: a combined approach by QENS, normal mode analysis, and molecular dynamics studies. Biopolymers 89:582–594
39. Losi A, Kottke T, Hegemann P (2004) Recording of blue light-induced energy and volume changes within the wild-type and mutated phot-LOV1 domain from *Chlamydomonas reinhardtii*. Biophys J 86:1051–1060
40. Crosson S, Rajagopal S, Moffat K (2003) The LOV domain family: photoresponsive signaling modules coupled to diverse output domains. Biochemistry 42:2–10
41. Yao X, Rosen MK, Gardner KH (2008) Estimation of the available free energy in a LOV2-J$\alpha$ photoswitch. Nat Chem Biol 4:491–497
42. Vendruscolo M (2008) Protein dynamics under light control. Nat Chem Biol 4:449–450
43. Matsuoka D, Tokutomi S (2005) Blue light-regulated molecular switch of Ser/Thr kinase in phototropin. Proc Natl Acad Sci USA 102:13337–13342
44. Huang K, Beck CF (2003) Phototropin is the blue-light receptor that controls multiple steps in the sexual life cycle of the green alga *Chlamydomonas reinhardtii*. Proc Natl Acad Sci USA 100:6269–6274
45. Khrenova M, Domratcheva T, Grigorenko B, Nemukhin A (2011) Coupling between the BLUF and EAL domains in the blue light-regulated phosphodiesterase BlrP1. J Mol Model 17:1579–1578. doi:10.1007/s00894-010-0842-1
46. Nakasako M, Zikihara K, Matsuoka D, Katsura H, Tokutomi S (2008) Structural basis of the LOV1 dimerization of *Arabidopsis* phototropins 1 and 2. J Mol Biol 381:718–733
47. Cho H-Y, Tseng T-S, Kaiserli E, Sullivan S, Christie JM, Briggs WR (2007) Physiological roles of the light, oxygen, or voltage domains of phototropin 1 and phototropin 2 in *Arabidopsis*. Plant Physiol 143:517–529
48. Iwata T, Nozaki D, Tokutomi S, Kagawa T, Wada M, Kandori H (2003) Light-induced structural changes in the LOV2 domain

of *Adiantum* Phytochrome3 studied by low-temperature FTIR and UVvisible spectroscopy. Biochemistry 42:8183–8191
49. Möglich A, Ayers RA, Moffat K (2009) Structure and signaling mechanism of Per-ARNT-Sim domains. Structure 17:1282–1294
50. Hahn K, Kuhlman B (2010) Hold me tightly LOV. Nat Methods 7:595–597
51. Strickland D, Yao X, Gawlak G, Rosen MK, Gardner KH, Sosnick TR (2010) Rationally improving LOV domain-based photoswitches. Nat Methods 7:623–626
52. Strickland D, Moffat K, Sosnick TR (2008) Light-activated DNA binding in a designed allosteric protein. Proc Natl Acad Sci USA 105:10709–10714

J Mol Model (2012) 18:1389–1399
DOI 10.1007/s00894-011-1166-5

ORIGINAL PAPER

# Effects of the position and manner of hydration on the stability of solvated N-methylacetamides and the strength of binding between N-methylacetamide and water clusters: a computational study

**Xiuchan Xiao · Ying Tan · Lijuan Zhu · Yanzhi Guo · Zhining Wen · Menglong Li · Xuemei Pu · Anmin Tian**

Received: 3 May 2011 / Accepted: 27 June 2011 / Published online: 15 July 2011


**Abstract** This work mainly studies the effects of the position (there are two possible hydrated sites) and the manner (i.e., whether water acts as a proton donor or acceptor) of hydration by various numbers of water molecules on the stability of 14 solvated N-methylacetamide structures, NMA-$(H_2O)_n$ ($n$=1–3), as well as the binding strength between the NMA and the water cluster, using molecular dynamics (MD) and B3LYP methods. Natural bond orbital (NBO) analysis is used to explore the origin of these effects. Some novel observations are obtained from the work. Our results show that monohydration at the carbonyl site favors stability and binding strength compared to monohydration at the amino site. Similarly, the preferred hydration at the carbonyl site is observed for dihydrated NMAs when the second water is added as a proton donor to the C=O group or the first water is H-bonded to the C=O group. However, unfavorable hydration at the C=O site occurs if the second water acts as a proton acceptor. Trihydration by a ring cluster of three water molecules at either the carbonyl site or the amino one yields relatively stable complexes, but significantly disfavors binding strength. The other trihydrated NMAs show similar behavior to dihydrated NMAs. In addition, our results show that the C=O and N–H frequencies can still be utilized to examine the H-bond effects of the water cluster.

**Keywords** N-Methylacetamide · Hydration · Molecular dynamics · B3LYP · NBO

X. Xiao · Y. Tan · L. Zhu · Y. Guo · Z. Wen · M. Li · X. Pu (✉) · A. Tian
Faculty of Chemistry, Sichuan University,
Chengdu 610064, People's Republic of China
e-mail: xmpuscu@scu.edu.cn

## Introduction

N-Methylacetamide (NMA) is one of the simplest models of the main chain of proteins. Its solvated complex has been the subject of theories and experiments [1–15] due to its implication in protein hydration—an important issue in molecular biology and biochemistry. Water molecules can establish hydrogen bonds to the main chain through the C=O and N–H groups of the peptide unit, which plays an important role in the structural properties and functions of proteins. Many theoretical calculations have been performed in order to investigate the hydrogen bonding in solvated NMA [7, 16, 17] and thus to understand protein hydration at the molecular level.

The strength of the hydrogen bond between a water molecule and NMA was studied by Dixon using ab initio molecular orbital theory [18]. Han [19] discussed interaction energies in *trans*- and *cis*-NMA$(H_2O)_n$ ($n$=1–3) complexes upon the formation of hydrogen bonds by means of DFT calculations. Buck used an MD method to study the hydrogen-bond energy between NMA and water [16]. Langley and Allinger [20] performed a computational study of hydrogen-bond parameters for some dimer systems (including monohydrated NMA) at the MP2/6-311+ +G (2d,2p) level. Petr Bouř used MD and DFT methods to analyze amide I vibrations of NMA interacting with a variety of water molecule clusters that were hydrogen bonded to it [21]. Mennucci calculated the geometries and IR and UV spectra of NMA using different solvation models, and discussed the effects of hydrogen bonding on these properties [22]. Asher used a DFT method to study the dependences of the amide I, II and III bands on the H-bonding of NMA with water [17]. Recently, Yang used a newly developed ABEEM/MM potential model to study

NMA-$(H_2O)_n$ ($n$=1–6) gas clusters [13] and to perform a molecular dynamics study of NMA in water [23]; some static properties of clusters (for example, structures, interaction energies, hydrogen-bonding cooperative effects, and so on) and some dynamic properties (radial distribution function, energy distribution, etc.) were discussed.

These studies have already shown that the energy of an H-bond between water and an N–H bond is clearly different from the energy of a C=O bond, and single hydration at the C=O position is more stabilizing than that at the N–H bond. However, as the amount of water increases, bonding occurs to a water cluster instead of a single water molecule. For example, a water dimer ($n$=2) forms a linear structure [24, 25], and we start to see ring structures of water clusters form at $n$=3–5 [26–28]. However, many questions remain unanswered. For example, is the stabilization of a water cluster at the N–H and C=O bond positions similar to or different from that obtained with monohydration? What manner of hydration (i.e., whether water is added as a proton acceptor or proton donor) best favors the stability and strength of binding between NMA and the water cluster? Since the previous studies reported total hydrogen-bond energies or hydration energies calculated for NMA complexes that are hydrated by 1–3 waters at different positions simultaneously, no further information on this subject has been published [20, 29]. Thus, there are still open questions regarding the competition between hydration by the water cluster at the amino site and that at the carbonyl site, and the nature of this competition. Answering these questions would allow us to better understand the hydration of proteins in aqueous solution—especially microsolvation.

Therefore, in this work, we comparatively study the H-bonding of NMA with 1–3 waters that hydrate the N–H and C=O sites. Since the amide links within peptides have mostly *trans* arrangements, with the amino and carbonyl groups opposing each other [30], only *trans*-NMA is considered in this work. Our work focuses on the following two aspects: comparing the stabilities of hydrated NMA complexes and the binding strengths between NMA and the water cluster for the two hydrated sites; gaining insight into how the binding strength varies with the number of water molecules, and the effects of the manner of hydration. In addition, the effects of the water clusters on the C=O and N–H stretching frequencies will be briefly discussed in this work. To explore the physical origin of these effects, we utilized natural bond orbital (NBO) analysis, as it has already been successfully used to investigate the molecular properties of some H-bonded systems [31–34].

## Computational details

One NMA molecule optimized at the level of B3LYP/6-311++G(d,p) was placed into a cubic box consisting of ten water molecules to represent the microhydration of a protein. The all-atom AMBER03 force field [35] was used to model the NMA molecule, while water was represented by the TIP3P model [36]. The nonbonded cutoff distance was defined as 40 Å. After 1000 steps of minimization, an MD simulation was performed in the canonical ensemble at 300 K under nonperiodic boundary conditions using the weak coupling algorithm [37]. A time step of 1 fs was utilized. The length of the MD simulation, performed using the Amber10 software package, was 1 ns [38]. This 1 ns trajectory was applied to analyze the character of the solvent around NMA. Fourteen hydrated NMA conformations were extracted from representative MD snapshots (a detailed description is provided later) for further investigation at the B3LYP/6-311++G(d,p) level with the aid of the Gaussian09 software package [39]. The full counterpoise (CP) method was used to avoid basis set superposition error (BSSE) [40] for the 14 hydrated NMA complexes. Natural bond orbital (NBO) analysis, as introduced by Weinhold and co-workers [41], was carried out in order to analyze orbital interactions in all of the optimized structures at the same level of theory, using the NBO3.1 program [42] incorporated into Gaussian09. The NBO theory considers all possible interactions between occupied donor and empty acceptor NBOs, and estimates their interaction strengths via the second-order perturbation energy $E^{(2)}$. For each donor NBO ($i$) and acceptor NBO ($j$), the $E^{(2)}$ value associated with electron delocalization $i \rightarrow j$ was calculated by

$$E^{(2)} = -q_i \frac{F_{ij}^2}{E_j - E_i} \quad (1)$$

where $q_i$ is the orbital occupancy, $E_i$ and $E_j$ are the electron donor and electron acceptor orbital energies, respectively, and $F_{ij}$ is the off-diagonal NBO Fock matrix element.

## Results and discussion

### Construction of 14 conformations of hydrated NMA complexes

In order to obtain representative configurations of poorly hydrated NMA complexes, we constructed initial structures of NMA-$(H_2O)_n$ ($n$=1–3) complexes by performing an MD simulation of the system of one NMA molecule solvated by ten water molecules. First, we calculated root-mean-square deviations (RMSDs) every 50 fs during the 1 ns MD trajectories (yielding 20,000 RMSD snapshots). The distribution of RMSD values for these 20000 snapshots is shown in Fig. 1a. This clearly shows that the RMSD range with the most snapshots (the highest population or probability)

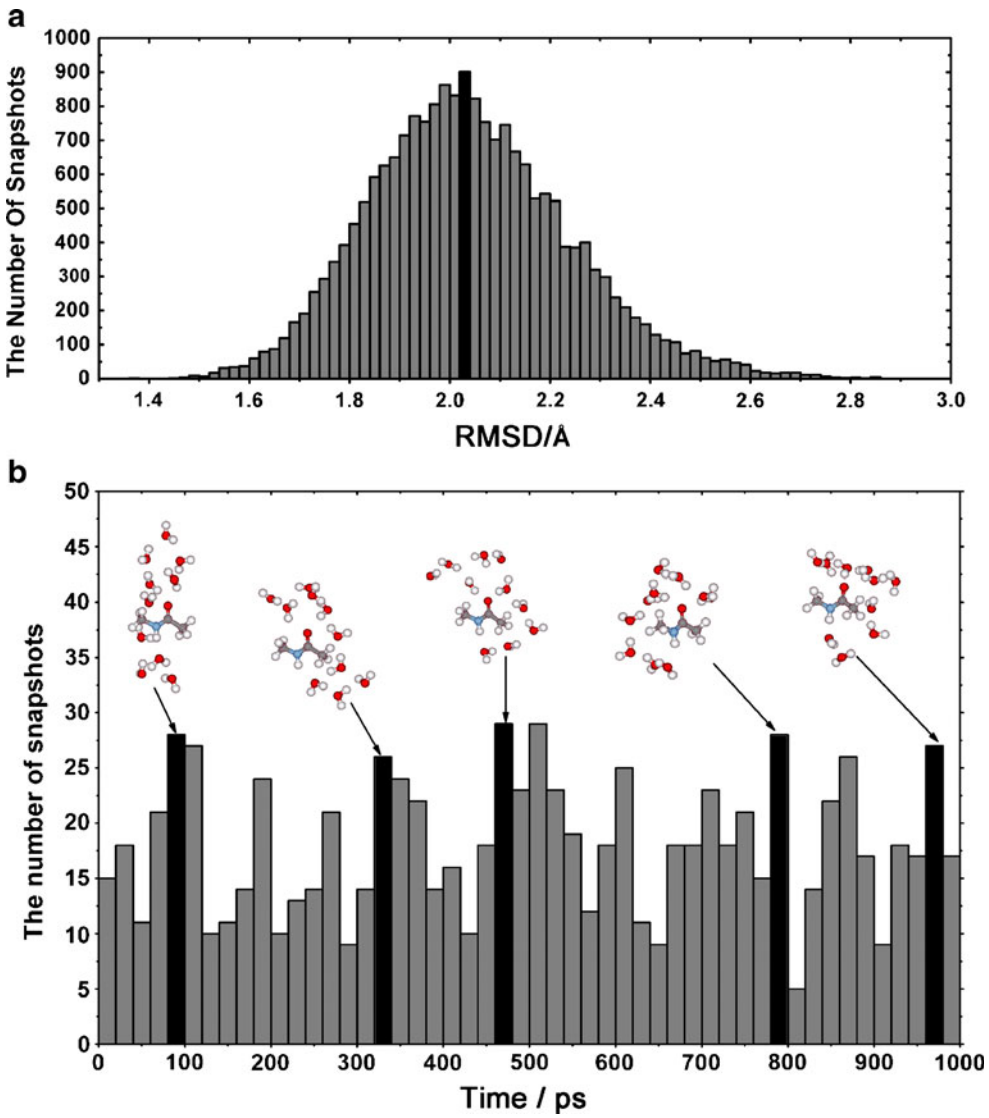



**Fig. 1** **a** Histogram showing the RMSD values of 20,000 snapshots extracted from a 1 ns MD simulation trajectory every 5 psl. The *black bin* denotes the RMSD range with most snapshots. **b** Distribution of the snapshots included in the black bin in **a** with simulation time (ps). The *five black bins* denote the RMSD range with the highest population of snapshots every 200 ps. The *five insets* denote five representative conformations extracted from the snapshots in the five bins, respectively

was from 2.02 to 2.04 Å, as shown by the black bin in Fig. 1a. We then further analyzed the time distributions of these snapshots in the RMSD range of 2.02–2.04 Å (see Fig. 1b). As can be seen from Fig. 1b, these snapshots were distributed across the whole simulation timescale. We randomly selected five snapshots from the five black bins in Fig. 1b, which individually denote the highest population of snapshots every 200 ps intervals, as representative conformations of MD trajectories (see the insets in Fig. 1b). From these five snapshots, we extracted ca. 24 hydrated NMA conformations with one, two, or three waters attached to the carbonyl and amino sites, respectively. Quantum chemistry optimizations were performed at the B3LYP/6-311++G(d,p) level for the 24 hydrated NMAs. Considering the manner of hydration (i.e., whether the waters acted as proton donors or acceptors) of the two functional groups of NMA or the former water H-bonded to NMA, 14 low-energy conformations were ultimately selected from the optimized results to be studied further, as shown in Figs. 2, 3 and 4. We classified them into two sorts

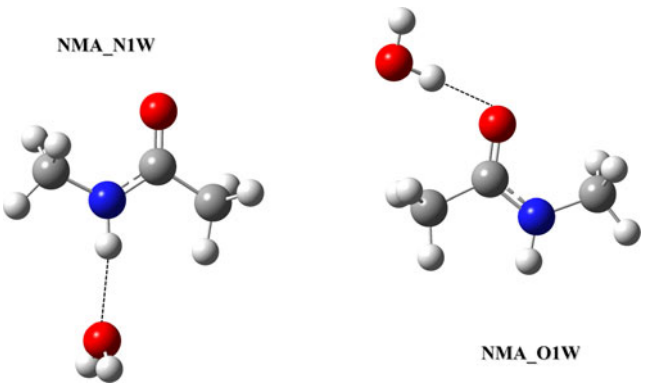



**Fig. 2** Two structures of monohydrated NMAs derived from full CP optimization at the B3LYP/6-311++G(d,p) level. *Black dashed lines* denote H-bonds

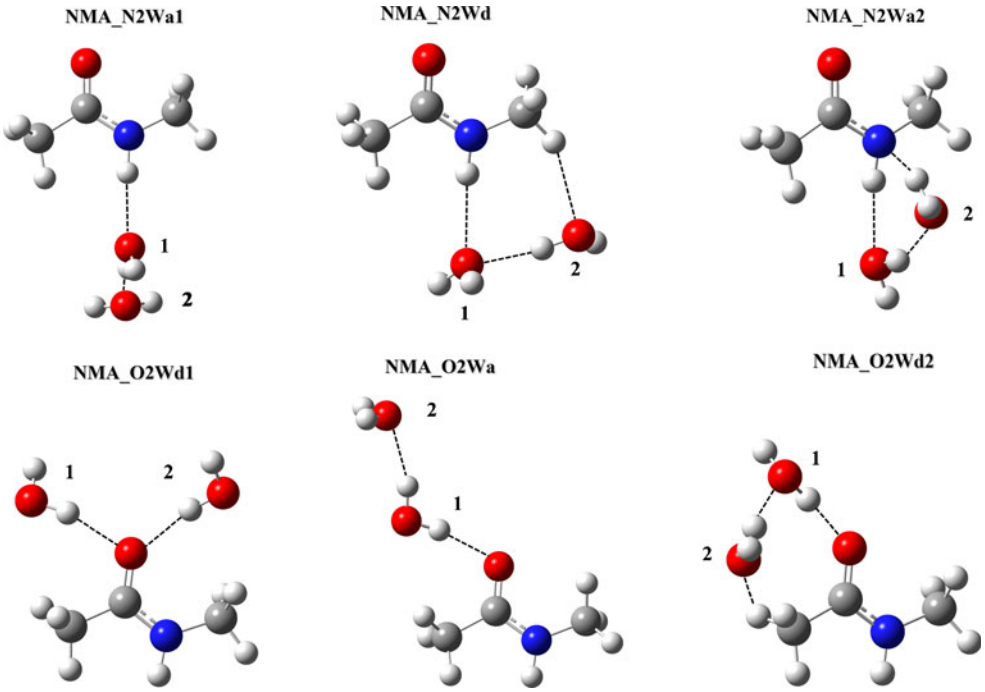



**Fig. 3** Six structures of dihydrated NMAs derived from full CP optimization at the B3LYP/6-311++G(d,p) level. The water molecules are numbered. *Black dashed lines* denote the H-bonds

based on their two hydration positions. One, which mainly has an N–H…O–H hydrogen bond between NMA and the water, is denoted NMA_N*n*W, denoting hydration at the amino group site. The other, which mainly has a C=O…H–O hydrogen bond between NMA and the water, is termed NMA_O*n*W, and represents hydration at the carbonyl group site. The *n* in NMA_N*n*W and NMA_O*n*W designates the number of water molecules in the complex.

Effects of the position and the manner of hydration vary with the number of water molecules

Here, we consider all of the water molecules collectively in order to investigate the binding strength between each water cluster and NMA. The binding energy between NMA and the water cluster, $\Delta E_{\text{cluster}}$, is calculated via

$$\Delta E_{\text{cluster}} = E^{\text{CP}}_{\text{complex}} - E^{\text{CP}}_{\text{cluster}} - E_{\text{NMA}}, \quad (2)$$

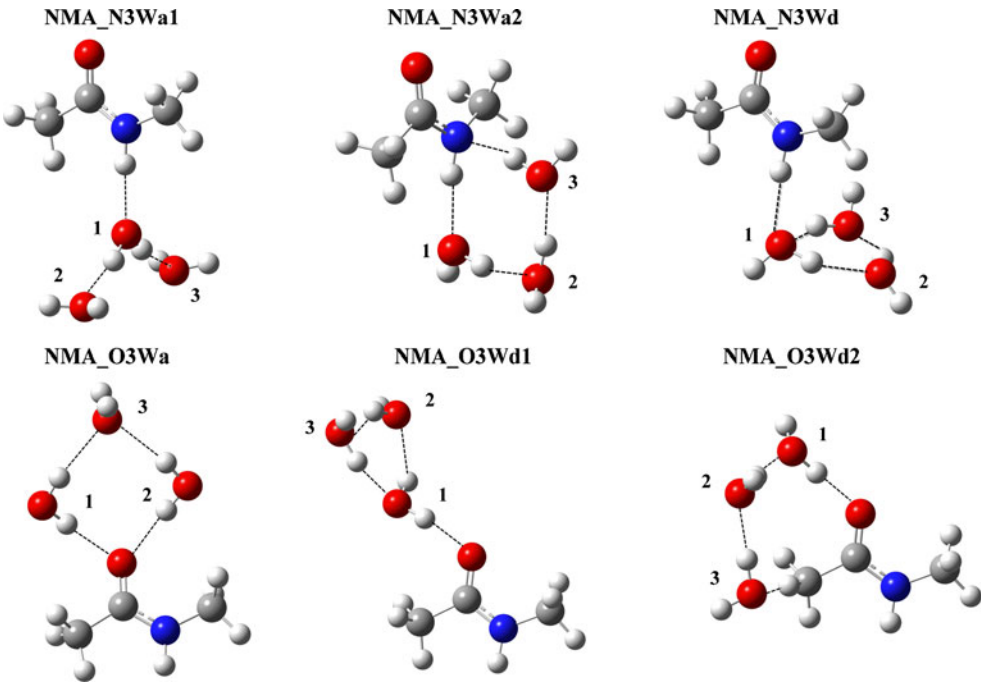



**Fig. 4** Six structures of trihydrated NMAs derived from full CP optimization at the B3LYP/6-311++G(d,p) level. The water molecules are numbered. *Black dashed lines* denote the H-bonds

where $E^{CP}_{complex}$ represents the total energy of the complex, including the BSSE correction, obtained by means of full CP optimization. $E^{CP}_{cluster}$ and $E_{NMA}$ refer to the single-point energies of the water cluster and NMA, respectively, extracted from the CP-optimized NMA-$(H_2O)_n$ structures using frozen geometries. The BSSE in the water cluster ($E^{CP}_{cluster}$) is corrected by single-point counterpoise. The calculated $E^{CP}_{complex}$ and $\Delta E_{cluster}$ values are listed in Table 1.

To gain insight into the nature of the effects of the water cluster on the stabilities of the hydrated NMAs, we now define $E^{(2)}_{total}$ as a descriptor that is used to examine the total orbital interaction strength involved in all intermolecular H-bonding within one complex. The $E^{(2)}_{total}$ value of one complex is obtained by summing the second perturbation energies $E^{(2)}$ of all intermolecular H-bonds. In addition, $E^{(2)}_{W-NMA}$ is defined as a descriptor that is used to examine the total intermolecular orbital interaction strength of the H-bonding between NMA and the water molecules H-bonded to it (i.e., it does not include the $E^{(2)}$ values involved in the H-bonding between water molecules). The $E^{(2)}_{total}$ and $E^{(2)}_{W-NMA}$ values are listed in Table 2; they were calculated using the NBO method at the B3LYP/6-311++ (d,p) level. For monohydrated NMAs, the $E^{(2)}_{W-NMA}$ values are equal to $E^{(2)}_{total}$.

**Table 1** The total energies ($E^{CP}_{complex}$, in hartrees), the binding energies ($\Delta E_{cluster}$, in kcal mol$^{-1}$, as calculated using Eq. 1), the bond lengths (in Å) and the vibrational frequencies (in cm$^{-1}$) of the C=O and N–H bonds in the fourteen hydrated NMA structures

| Structure | $E^{CP}_{complex}$ [b] | $\Delta E_{cluster}$ | Frequency[a] | |
|---|---|---|---|---|
| | | | $\nu_{C=O}$ | $\nu_{N-H}$ |
| NMA_N1W | −325.07160 | −4.52 | 1676.8 | 3467.3 |
| NMA_N2Wa1 | −401.54072 | −6.27 | 1672.7 | 3426.5 |
| NMA_N2Wd | −401.53542 | −3.73 | 1679.2 | 3493.8 |
| NMA_N2Wa2 | −401.54164 | −7.75 | 1693.1 | 3389.8 |
| NMA_N3Wa1 | −478.00728 | −7.86 | 1669.5 | 3380.3 |
| NMA_N3Wa2 | −478.01584 | −10.48 | 1694.7 | 3321.2 |
| NMA_N3Wd | −478.01277 | −4.47 | 1677.4 | 3451.8 |
| NMA_O1W | −325.07577 | −7.28 | 1658.2 | 3537.6 |
| NMA_O2Wd1 | −401.54396 | −14.65 | 1635.4 | 3532.6 |
| NMA_O2Wa | −401.53942 | −5.63 | 1664.8 | 3527.7 |
| NMA_O2Wd2 | −401.54760 | −11.12 | 1647.4 | 3536.6 |
| NMA_O3Wa | −478.01030 | −11.31 | 1648.8 | 3534.6 |
| NMA_O3Wd2 | −478.01940 | −13.03 | 1644.4 | 3532.5 |
| NMA_O3Wd1 | −478.01726 | −7.69 | 1657.5 | 3532.5 |

[a] Frequencies are scaled by 0.9679 [44]

[b] Derived from full counterpoise (CP) optimization at the B3LYP/6-311++(d,p) level

*Monohydrated and dihydrated complexes*

For the two monohydrated complexes (see Fig. 2), the binding energy between NMA and the water in the NMA_O1W structure and the stability of that structure are significantly greater than those of NMA_N1W, as reflected by their $E^{CP}_{complex}$ and $\Delta E_{cluster}$ values (see Table 1). This can be attributed to the fact that the orbital interaction energies associated with the intermolecular H-bonding ($E^{(2)}_{W-MNA}$ and $E^{(2)}_{total}$) between NMA and the water in NMA_O1W are larger than those in NMA_N1W (see Table 2). This observation suggests that hydration at the C=O site leads to enhanced stability and binding strength compared to hydration at the amino position, as noted in a previous investigation [43].

There are six dihydrated NMAs, three of which represent hydration at the amino group site (see NMA_N2Wa1, NMA_N2Wd and NMA_N2Wa2 in Fig. 3). The other three denote hydration at the carbonyl position (see NMA_O2Wd1, NMA_O2Wa and NMA_O2Wd2 in Fig. 3). Their $E^{CP}_{complex}$ and $\Delta E_{cluster}$ values in Table 1 show that the stabilities and the binding strengths of NMA_O2Wd1 and NMA_O2Wd2 are higher than those of the three NMAs that are dihydrated at the N–H bond, again indicating favorable hydration at the carbonyl position. However, the stability and binding strength of NMA_O2Wa are lower than those of NMA_N2Wa1 and NMA_N2Wa2, pointing to a preference for hydration at the amino site.

Comparing with the monohydrated NMA_O1W, the second water molecule in the NMA_O2Wd1 and NMA_O2Wd2 structures is added as a proton donor, in order to interact with the C=O group and with the first water molecule that is already H-bonded to NMA, respectively. For NMA_O2Wd1, the manner of hydration means that the oxygen atom of the carbonyl group acts as a double proton acceptor towards the two water molecules, leading to the formation of two O…H–O hydrogen bonds, as confirmed by the values (8.22 and 6.45 kcal mol$^{-1}$) of the second perturbation energies $E^{(2)}$ associated with the two $n_{O(NMA)} \rightarrow \sigma^*_{O-H(H2O)}$ interactions. The two H-bonds of the two waters with the C=O group are also suggested by the radial distribution function (RDF), since integration of the first well-defined peaks at 1.98Å in the C=O…$H_w$ RDF (see Fig. 5) yields ca. 2 water molecules. Furthermore, a hydrogen-bond analysis based on the 1 ns MD trajectory indicates that the double hydrogen bond in the NMA_O2Wd1 structure is present for over half of the simulation time. For NMA_O2Wd2 structure, the second water acts as a proton donor towards the first water. On the one hand, the manner of hydration significantly enhances the H-bonding between the first water and NMA, as reflected by the value (13.79 kcal mol$^{-1}$) of $E^{(2)} n_{O(NMA)} \rightarrow \sigma^*_{O-H(H2O)}$, which is much larger than its corresponding value in the monohydrated NMA_O1W. On the other hand, it means that

**Table 2** The second-order perturbation energies of the orbital interactions ($E^{(2)}$, in kcal mol$^{-1}$) associated with the H-bonding in the fourteen structures optimized at the B3LYP/6-311++G(d,p) level, obtained via NBO calculation at the same level

| Complex | Interaction[a] | $E^{(2)}$ | $E^{(2)}_{W-NMA}$ [b] | $E^{(2)}_{total}$ [c] |
|---|---|---|---|---|
| NMA_N1W | $n_{O(H2O)} \rightarrow \sigma^*_{N-H(NMA)}$ | 5.17 | 5.17 | 5.17 |
| NMA_N2Wa1 | $n_{O(H2O)} \rightarrow \sigma^*_{N-H(NMA)}$ | 7.89 | 7.89 | 15.61 |
| NMA_N2Wd | $n_{O(H2O)} \rightarrow \sigma^*_{N-H(NMA)}$ | 3.06 | 4.08 | 8.78 |
| NMA_N2Wa2 | $n_{O(H2O)} \rightarrow \sigma^*_{C-H(NMA)}$ | 1.02 | 8.54 | 15.16 |
| | $n_{O(H2O)} \rightarrow \sigma^*_{N-H(NMA)}$ | 5.13 | | |
| NMA_N3Wa1 | $n_{N(NMA)} \rightarrow \sigma^*_{O-H(H2O)}$ | 3.41 | 9.43 | 23.02 |
| | $n_{O(H2O)} \rightarrow \sigma^*_{N-H(NMA)}$ | 9.43 | | |
| NMA_N3Wa2 | $n_{O(H2O)} \rightarrow \sigma^*_{N-H(NMA)}$ | 9.11 | 15.65 | 38.35 |
| NMA_N3Wd | $n_{N(NMA)} \rightarrow \sigma^*_{O-H(H2O)}$ | 6.54 | 6.45 | 28.40 |
| | $n_{O(H2O)} \rightarrow \sigma^*_{N-H(NMA)}$ | 6.45 | | |
| NMA_O1W | $n_{O(NMA)} \rightarrow \sigma^*_{O-H(H2O)}$ | 8.66 | 8.66 | 9.00 |
| NMA_O2Wd1 | $n_{O(NMA)} \rightarrow \sigma^*_{O-H(H2O)\ L}$ | 8.22 | 14.67 | 14.67 |
| NMA_O2Wa | $n_{O(NMA)} \rightarrow \sigma^*_{O-H(H2O)\ R}$ | 6.45 | 7.43 | 12.70 |
| | $n_{O(NMA)} \rightarrow \sigma^*_{O-H(H2O)}$ | 7.43 | | |
| NMA_O2Wd2 | $n_{O(NMA)} \rightarrow \sigma^*_{O-H(H2O)}$ | 13.79 | 15.84 | 26.62 |
| NMA_O3Wa | $n_{O(H2O)} \rightarrow \sigma^*_{C-H(NMA)}$ | 2.05 | 11.15 | 18.40 |
| | $n_{O(NMA)} \rightarrow \sigma^*_{O-H(H2O)\ L}$ | 5.99 | | |
| NMA_O3Wd2 | $n_{O(NMA)} \rightarrow \sigma^*_{O-H(H2O)\ R}$ | 5.16 | 18.88 | 44.35 |
| | $n_{O(NMA)} \rightarrow \sigma^*_{O-H(H2O)}$ | 15.65 | | |
| NMA_O3Wd1 | $n_{O(H2O)} \rightarrow \sigma^*_{C-H(NMA)}$ | 3.23 | 10.00 | 36.01 |
| | $n_{O(NMA)} \rightarrow \sigma^*_{O-H(H2O)}$ | 10.00 | | |

[a] The orbital interaction associated with intermolecular H-bonding between the water molecules and NMA only (i.e., not including the bonding between water molecules)

[b] $E^{(2)}_{W-NMA}$ denotes the sum of the $E^{(2)}$ values associated with H-bonding between NMA and the water molecules only (i.e., not including H-bonding between water molecules) for this particular complex

[c] $E^{(2)}_{total}$ denotes the sum of the $E^{(2)}$ values associated with the intermolecular H-bonding within this particular complex, including the H-bonds between water molecules

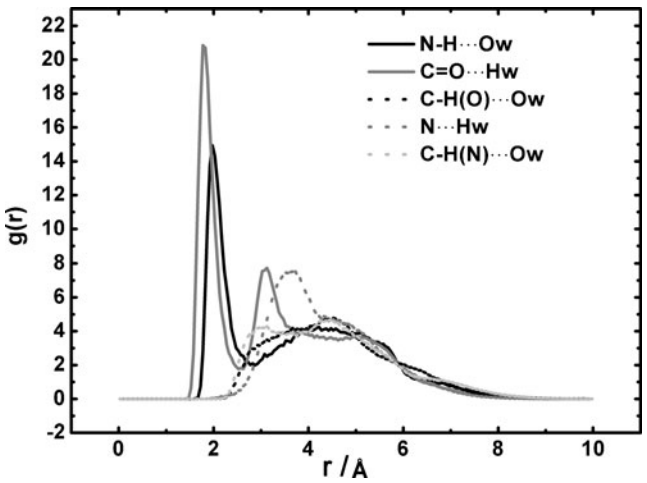



**Fig. 5** RDFs of the hydrogen of water (Hw) around the oxygen of the carbonyl group (C=O…Hw) and the nitrogen of the amino group (N…Hw), the oxygen of water (Ow) around the amino hydrogen (N–H…Ow), the methyl hydrogen at the carbonyl end (C–H(O)…Ow), and the methyl hydrogen at the amino group end (C–H(N)…Ow), as derived from MD simulation

the second water simultaneously acts as a proton acceptor towards the methyl group and induces the formation of another weak C–H…O H-bond between NMA and the water molecule, as demonstrated by the value (2.05 kcal mol$^{-1}$) of $E^{(2)} n_{O(H2O)} \rightarrow \sigma^*_{C-H(NMA)}$. As a result, the $E^{(2)}_{W-MNA}$ value for the NMA_O2Wd2 structure is the largest of those for the six dihydrated complexes, meaning that it also has the largest binding energy $\Delta E_{cluster}$. The weak C–H…O H-bond between NMA and the water in NMA_O2Wd2 is also implied by MD results, since the RDF of C–H(O)…$O_W$ (see Fig. 5) displays less peak structure at 4.53 Å. Furthermore, the H-bond is present for over one-quarter of the MD simulation time. In addition, the weak C–H…O H-bond in the solvated NMA system has already been reported by Zhang [8]. Thus, the NMA_O2Wd2 structure has three intermolecular hydrogen bonds and the largest $E^{(2)}_{total}$, meaning that it has the highest stability among the six dihydrated structures.

In contrast to the manner of hydration for the NMA_O2Wd1 and NMA_O2Wd2 structures, the second water molecule in NMA_O2Wa acts as a proton acceptor towards the first water molecule H-bonded to NMA. The hydration weakens the direct H-bonding between NMA and

the water cluster since its $E^{(2)}n_{O(NMA)}\rightarrow\sigma^*_{O-H(H2O)}$ (i.e., $E^{(2)}_{W-MNA}$) is decreased by 1.23 kcal mol$^{-1}$ with respect to that of the monohydrated NMA_O1W. As a result, its $E^{(2)}_{total}$ and $E^{(2)}_{W-NMA}$ values are the lowest of those of the complexes that are dihydrated at the carbonyl site—even lower than those of the dihydrated NMA_N2Wa1 and NMA_N2Wa2. Accordingly, its stability and binding strength are also lower than those of these dihydrated NMAs. This observation suggests that the second water acting as a proton acceptor disfavors stability and binding strength for NMAs that are dihydrated at the carbonyl site.

Comparing with the monohydrated NMA_N1W, the second water molecules to be added to the NMA_N2Wa1 and NMA_N2Wa2 structures both act as proton acceptors towards the first water molecule to be H-bonded to NMA. For NMA_N2Wa1, the manner of hydration enhances the H-bonding between NMA and the first water, as revealed by its larger $E^{(2)}n_{O(H2O)}\rightarrow\sigma^*_{N-H(NMA)}$ (i.e., $E^{(2)}_{W-MNA}$) than that of NMA_N1W. Thus, its binding strength $\Delta E_{cluster}$ is enhanced compared to that of NMA_N1W. In the NMA_N2Wa2 structure, the second water, which acts as a proton acceptor, induces the formation of another H-bond between it and the amino group nitrogen (O–H…N H-bond), as confirmed by the value (3.41 kcal mol$^{-1}$) of $E^{(2)}n_{N(NMA)}\rightarrow\sigma^*_{O-H(H2O)}$. The O–H…N H-bonding is also implied by the RDF of N…$H_W$, since a less structured peak is observed at 3.87 Å (see Fig. 5). Furthermore, the hydration does not weaken the H-bond between the first water and NMA in the NMA_N2Wa2 structure, since its $E^{(2)}n_{O(H2O)}\rightarrow\sigma^*_{N-H(NMA)}$ value is almost equal to that of its corresponding monohydrated NMA_N1W. Thus, the $E^{(2)}_{W-NMA}$ and $E^{(2)}_{total}$ values of NMA_N2Wa2 are nearly equal to those of NMA_N2Wa1, and both are larger than those of NMA_N1W and NMA_O2Wa. Accordingly, the binding strengths $\Delta E_{cluster}$ and stabilities of NMA_N2Wa1 and NMA_N2Wa2 are higher than those of NMA_N1W and NMA_O2Wa, indicating that enhanced binding strength is obtained with an increasing number of water molecules, and that the preferred hydration site is the amino group. In contrast to the manner of hydration in NMA_N2Wa1 and NMA_N2Wa2, the second water in NMA_N2Wd interacts as a proton donor with the first water to H-bond with NMA. The manner of hydration in the NMA_N2Wd structure induces the formation of a weak H-bond between the oxygen of the second water and the methyl group of NMA, as demonstrated by the $E^{(2)}n_{O(H2O)}\rightarrow\sigma^*_{C-H(NMA)}$ value of 1.02 kcal mol$^{-1}$. However, this significantly weakens the H-bond between the first water and NMA, since the $E^{(2)}n_{O(H2O)}\rightarrow\sigma^*_{N-H(NMA)}$ value decreases from 5.17 kcal mol$^{-1}$ in the monohydrated complex to 3.06 kcal mol$^{-1}$, which is significantly lower than the corresponding values in the NMA_N2Wa1 and NMA_N2Wa2 structures, where the second water acts as a proton acceptor. Accordingly, its $E^{(2)}_{total}$ (8.78 kcal mol$^{-1}$) and $E^{(2)}_{W-NMA}$ (4.08 kcal mol$^{-1}$) values are the lowest among the six dihydrated structures. As a result, its binding strength $\Delta E_{cluster}$ and stability are also the weakest. Furthermore, its $\Delta E_{cluster}$ value is even lower than that of monohydrated NMA_N1W, indicating a decrease in this parameter as the number of water molecules increases. The weak C–H…O H-bond at the amino group end of the NMA_N2Wd structure is implied by the RDF of C–H(N)…$O_W$ (see Fig. 5), which has a less structured peak at 4.48 Å. Furthermore, H-bond analysis based on the MD trajectory shows that the H-bond is present for >17% of the simulation time—less than that of the C–H…O H-bond at the carbonyl site. The observation above suggests that adding the second water as a proton acceptor favors stability and binding strength in NMAs that are dihydrated at the amino group, while adding it as a proton donor has the opposite effect. This is the opposite to the behavior observed for NMAs that are dihydrated at the carbonyl site.

### *Triply hydrated complexes*

Six triply hydrated NMAs are considered in this work, three of which relate to trihydration at the amino group (see NMA_N3Wa1, NMA_N3Wd and NMA_N3Wa2 in Fig. 4). The other three NMAs (see NMA_O3Wd1, NMA_O3Wa and NMA_O3Wd2 in Fig. 4) represent hydration by three water molecules at the carbonyl position. Table 1 shows that the stabilities of the six complexes follow the order NMA_O3Wd2 > NMA_O3Wd1 > NMA_N3Wa2 > NMA_N3Wd > NMA_O3Wa > NMA_N3Wa1, while the binding strength displays a different trend: NMA_O3Wd2 > NMA_O3Wa > NMA_N3Wa2 > NMA_N3Wa1 > NMA_O3Wd1 > NMA_N3Wd. This observation contrasts with that for dihydrated structures, which exhibit stabilities that are consistent with their binding strengths. Similarly, unfavorable hydration at the carbonyl site is observed based on the variations in the stability and binding strength above.

As can be seen from Fig. 4, the three water molecules in the NMA_O3Wd1 and NMA_N3Wd structures form cyclic clusters. This cyclic conformation of the three waters is very stable [26], due to strong H-bond interactions among the three waters. Indeed, the $E^{(2)}_{total}$ values of NMA_O3Wd1 and NMA_N3Wd are found to be larger than those of NMA_O3Wa and NMA_N3Wa1, respectively. Thus, NMA_O3Wd1 and NMA_N3Wd are observed to be more stable than NMA_O3Wa and NMA_N3Wa1, respectively. However, the binding strengths of NMA_O3Wd1 and NMA_N3Wd are the lowest among the NMA structures that are trihydrated at the carbonyl and amino sites, respectively. This can be attributed to the fact that there is only one H-bond between NMA and the cyclic water

cluster, so the strong H-bonding among the three waters in the ring conformation significantly weakens the hydrogen bonding between NMA and the ring water cluster. In fact, the $E^{(2)}_{W-NMA}$ values of NMA_O3Wd1 and NMA_N3Wd are the smallest of the NMAs that are trihydrated at the carbonyl site and at the amino one, respectively, which also leads to the result that their $E^{(2)}_{total}$ values are lower than those of NMA_O3Wd2 and NMA_N3Wa2, respectively. Therefore, the stabilities of NMA_O3Wd1 and NMA_N3Wd are lower than those of NMA_O3Wd2 and NMA_N3Wa2, respectively. In addition, the data in Table 1 show that the stability and binding strength of NMA_O3Wd1 are higher than those of NMA_N3Wd, highlighting the favorable trihydration by the cyclic water cluster at the carbonyl site.

Comparing with NMA_N2Wa1, the third water molecule in the NMA_N3Wa1 structure interacts as a proton acceptor with the first water to H-bond to NMA. The $E^{(2)}n_{O(H2O)}\rightarrow\sigma^*_{N-H(NMA)}$ value (i.e., $E^{(2)}_{W-NMA}$) of NMA_N3Wa1 is larger than that of NMA_N2Wa1, suggesting that the manner of hydration enhances the H-bonding between NMA and the water cluster. As a result, the binding strength between NMA and the cluster of three waters in the NMA_N3Wa1 structure is larger than that of NMA_N2Wa1, thus indicating enhanced binding strength with an increasing number of water molecules. In the NMA_N3Wa2 structure, the third water is introduced as a proton acceptor to form a hydrogen bond with the second water. The $E^{(2)}n_{O(H2O)}\rightarrow\sigma^*_{N-H(NMA)}$ value for NMA_N3Wa2 is higher than that of NMA_N2Wa2, as can be seen from Table 2. Furthermore, the hydration in the NMA_N3Wa2 conformation makes the third water simultaneously act as a proton donor to the amino group nitrogen. The manner of hydration in NMA_N3Wa2 induces the formation of an N…H–O H-bond, with a $E^{(2)}n_{N(NMA)}\rightarrow\sigma^*_{O-H(H2O)}$ value (6.54 kcal mol$^{-1}$) that is also larger than that of NMA_N2Wa2 (3.41 kcal mol$^{-1}$). As a result, the $E^{(2)}_{W-NMA}$(15.65 kcal mol$^{-1}$) and $E^{(2)}_{total}$ (38.35 kcal mol$^{-1}$) values of NMA_N3Wa2 are significantly larger than those of NMA_N2Wa2, and are the largest among the three complexes that are hydrated at the amino group. Thus, its stability and binding strength are highest among the NMAs that are trihydrated at the amino site, and are higher than that of NMA_N2Wa2.

Comparing with NMA_O2Wd1, the third water molecule in NMA_O3Wa, which acts as a double proton acceptor when H-bonding with the first two water molecules, causes the value of $E^{(2)}_{W-NMA}$ to drop by 3.52 kcal mol$^{-1}$, suggesting that the hydration weakens the H-bonding between the water cluster and NMA. As a result, the binding strength in NMA_O3Wa weakens as the number of water molecules increases, as shown by the $\Delta E_{cluster}$ value in Table 1. Furthermore, its $E^{(2)}_{total}$ value is lowest among the three complexes that are trihydrated at the carbonyl end, and is even significantly smaller than those of NMA_N3Wa2 and NMA_N3Wd. Thus, it stability is the lowest among the NMAs that are trihydrated at the carbonyl group, and lower than those of trihydrated NMA_N3Wa2 and NMA_N3Wd, indicating unfavorable hydration at the carbonyl site. Comparing with NMA_O2Wd2, the third water molecule in the NMA_O3Wd2 structure is introduced as a proton donor towards the second water. The hydration also induces C–H…O H-bonding between the third water and the methyl group of NMA, as demonstrated by the value (3.23 kcal mol$^{-1}$) of $E^{(2)}n_{O(H2O)}\rightarrow\sigma^*_{C-H(NMA)}$. Furthermore, the hydration enhances the two H-bonds between NMA and the water cluster, since the $E^{(2)}n_{O(H2O)}\rightarrow\sigma^*_{C-H(NMA)}$ and $E^{(2)}n_{O(NMA)}\rightarrow\sigma^*_{O-H(H2O)}$ values are larger than the corresponding values in NMA_O2Wd2, suggesting an increase in the binding strength $\Delta E_{cluster}$ as the number of water molecules increases. In addition, the hydration gives NMA_O3Wd2 the largest $E^{(2)}_{total}$ and $E^{(2)}_{W-MNA}$ values, meaning that it has the highest stability and strongest binding strength among all of the triply hydrated complexes studied.

In summary, the effects of the manner of hydration (i.e., whether water is added as a proton acceptor or donor) of the trihydrated complexes are only similar to those observed for dihydrated complexes if the three waters do not form a cyclic conformation.

## Effects of the water cluster on the stretching vibrational frequencies of the C=O and N–H bonds

The vibrational properties of the peptide bond can provide useful information on the secondary structures of polypeptides and proteins. On the other hand, it can also reflect the effects of the solvent present—particularly H-bonding with water, which is a common medium for biological samples. Thus, it has attracted much attention [5, 6, 12, 18, 22]. A dependence of the amide I band (i.e., the frequencies of C=O) on hydrogen bonding with water was observed in studies of the model peptide molecule NMA [5, 22]. Here, we also discuss the effect of the water cluster on the stretching frequencies of C=O and N–H bonds in order to check whether these frequency variations still reflect the solvent effect derived from the water cluster. The vibrational frequencies discussed here were corrected by multiplying them by the single scaling factor of 0.9679 [44] (see Table 1). We checked the correlations of the vibrational frequencies of the C=O and N–H bonds with the binding strengths at the two hydration sites. Good linear correlations are observed, as shown in Fig. 6 and Eqs. 3–4.

$$\Delta E_{cluster} = -144.88 - 0.041\nu_{N-H} \quad n = 7, R = 0.9901 \quad (3)$$

$$\Delta E_{cluster} = -554.53 - 0.33\nu_{C=O} \quad n = 7, R = 0.9866. \quad (4)$$

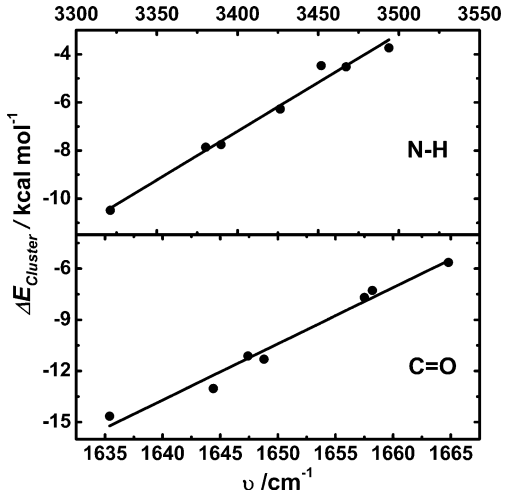



**Fig. 6** Dependences of the vibrational frequencies ($\nu$, in $cm^{-1}$) of C=O and N–H bonds on the binding strength ($\Delta E_{cluster}$, in kcal $mol^{-1}$) between the NMA and the water cluster for seven hydrated NMAs where the water molecules attached to the N–H group and seven hydrated NMAs where the water molecules attached to the C=O group, respectively

The results suggest that the frequency variation of the amide bond can still be utilized as a descriptor to indicate the H-bonding effects of the solvent cluster. To gain insight into the nature of this effect, we correlated the frequencies of the two bonds with the intermolecular orbital interaction involved in the H-bonding of the two groups with the water cluster; i.e., $\nu_{C=O}$ versus $E^{(2)}(n_O(NMA)\rightarrow\sigma^*_{O-H}(H_2O))$, and $\nu_{N-H}$ versus the sum of $E^{(2)}(n_O(H_2O)\rightarrow\sigma^*_{N-H}(NMA))$ and $E^{(2)}(n_N(NMA)\rightarrow\sigma^*_{O-H}(H_2O))$. Good correlation and fair correlation were noted for these frequencies of the two bonds, as shown in Eqs. 5–6:

$$E^{(2)} = 232.48 - 0.066\nu_{N-H} \quad n = 7, R = 0.9859 \tag{5}$$

$$E^{(2)} = 489.68 - 0.29\nu_{C=O} \quad n = 7, R = 0.9549. \tag{6}$$

This reveals that the water cluster influences the frequencies of the C=O and N–H bonds, especially the N–H bond, mainly by changing the intermolecular orbital interaction involved in the H-bonding between the functional group of NMA and the water molecule H-bonded to it, such as the interaction between the lone pair on the oxygen atom of the carbonyl and the $\sigma^*_{O-H}$ orbital of the water H-bonded to it, and the interaction between the amide group and the water H-bonded to it (i.e., $n_{O(H2O)}\rightarrow\sigma^*_{N-H(NMA)}$ and $n_{N(NMA)}\rightarrow\sigma^*_{O-H(H2O)}$).

## Conclusions

We studied 14 NMAs that were hydrated by one, two or three water molecules at the two sites (i.e., the carbonyl and amino sites) using the MD method and the B3LYP/6-311++G(d,p) level of theory. On the one hand, we studied the effects of the hydration site on the stability of the complexes and the binding strength between NMA and the water cluster. On the other hand, we discussed the effects of the manner of hydration by water molecules added as proton acceptors or donors on the binding strength, by comparing the results obtained for NMA-$(H_2O)_n$ and its corresponding NMA-$(H_2O)_{n-1}$. In addition, the effects of the water cluster on the C=O and N–H frequencies were analyzed. NBO analysis was used to explore the origin of these effects in this work.

The result shows that monohydration at the carbonyl site significantly favors complex stability and binding strength with respect to monohydration at the amino site, since the orbital interaction associated with the intermolecular H-bonding is stronger for hydration at the carbonyl site than at the amino group. The results obtained were similar to some previous observations. However, for dihydration, the favorable hydrated site was not always located at the carbonyl end, but displayed a strong correlation with the manner of hydration of the second water (i.e., whether it acts as a proton acceptor or donor). A second water that acts as a proton donor to the carbonyl group, or a first water that is H-bonded to the C=O bond of NMA favors complex stability and binding strength by enhancing the intermolecular H-bonding, while a second water acting as a proton acceptor would weaken the strength of the intermolecular orbital interaction, thus decreasing the stabilities and the binding strengths of NMAs that are dihydrated at the carbonyl site. As a result, unfavorable hydration at the carbonyl site may be observed, in comparison to some dihydrated complexes where the second water is added as a proton acceptor at the amino site. In contrast to the observations made for dihydration at the carbonyl site, a second water acting as a proton acceptor towards the water that is already H-bonded to NMA would favor the stabilities and binding strengths of NMAs dihydrated at the amino group, while the opposite behavior can be seen when the second water acts as a proton donor towards the first water that was H-bonded to NMA. In addition, for the dihydrated complexes, their stabilities are consistent with their binding strengths.

Although hydration by a ring cluster of three waters at either the carbonyl or the amino site enhances the stability of the complex, it significantly disfavors binding strength due to the considerably weakened H-bonding between NMA and the water cluster. This result leads to an inconsistent relationship between the stability and the binding strength for the six trihydrated NMAs. For the other four trihydrated conformations (i.e., excluding the two trihydrated NMA structures with the ring water cluster), the effects of the manner of hydration of the water added as a proton donor or acceptor on the stability and

binding strength are similar to those derived from dihydration. Namely, adding water as a proton donor favors complex stability and binding strength when trihydration occurs at the carbonyl site, while adding the water as a proton acceptor disfavors complex stability and binding strength.

In addition, our results show that the C=O and N–H frequencies can still be utilized as indicators that reflect the H-bond effects of from the water cluster. Furthermore, NBO analysis reveals that the water cluster influences the frequencies mainly through the intermolecular orbital interaction involved in the H-bonding between the two functional groups and the water molecule H-bonded to them.

The observations described above mainly relate to the 14 conformations of NMA hydrated by 1–3 waters, and are limited to the solvated NMA system. However, the fourteen structures should be representative of microsolvated NMA, since they are extracted from representative snapshots of MD trajectories. Though there have been some reports on the solvated NMA system, to our knowledge, this is the first systematic study on the effects of the site and manner of hydration for NMA. This study produced some novel observations that should prove useful for better understanding the protein hydration process.

**Acknowledgments** This project is supported by the National Science Foundation of China (grant no. 20973115).

## References

1. Zielkiewicz J (2000) Preferential solvation of N-methylformamide, N,N-dimethylformamide and N-methylacetamide by water and alcohols in the binary and ternary mixtures. Phys Chem Chem Phys 2:2925–2932
2. Kubelka J, Keiderling TA (2001) Ab initio calculation of amide carbonyl stretch vibrational frequencies in solution with modified basis sets. 1. N-methyl acetamide. J Phys Chem A 105:10922–10928
3. Rocha WR, De Almeida KJ, Coutinho K, Canuto S (2001) The electronic spectrum of N-methylacetamide in aqueous solution: a sequential Monte Carlo/quantum mechanical study. Chem Phys Lett 345:171–178
4. Akiyama M (2002) Study on hydration enthalpy of N-methylacetamide in water. Spectrochim Acta A 58:1943–1950
5. Ham S, Kim JH, Lee H, Cho MH (2003) Correlation between electronic and molecular structure distortions and vibrational properties. II. Amide I modes of NMA-*n*D(2)O complexes. J Chem Phys 118:3491–3498
6. Kijeong K, Minhaeng C (2003) Molecular dynamics simulation study of N-methylacetamide in water. I. Amide I mode frequency fluctuation. J Chem Phys 119:2247–2255
7. Rong Z, Haoran L, Yi L, Shijun H (2004) Structures and interactions in N-methylacetamide–water mixtures studied by IR spectra and density functional theory. J Mol Struct 693:17–25
8. Rong Z, Haoran L, Yi L, Shijun H (2004) Different weak C-H...O contacts in N-methylacetamide–water system: molecular dynamics simulations and NMR experimental study. J Phys Chem B 108:12596–12601
9. Gaigeot MP, Vuilleumier R, Sprik M, Borgis D (2005) Infrared spectroscopy of N-methylacetamide revisited by ab initio molecular dynamics simulations. J Chem Theor Comput 1:772–789
10. Hayashi T, Zhuang W, Mukamel S (2005) Electrostatic DFT map for the complete vibrational amide band of NMA. J Phys Chem A 109:9747–9759
11. Allison SK, Bates SP, Crain J, Martyna GJ (2006) Solution structure of the aqueous model peptide N-methylacetamide. J Phys Chem B 110:21319–21326
12. DeFlores LP, Ganim Z, Ackley SF, Chung HS, Tokmakoff A (2006) The anharmonic vibrational potential and relaxation pathways of the amide I and II modes of N-methylacetamide. J Phys Chem B 110:18973–18980
13. Yang ZZ, Qian P (2006) A study of N-methylacetamide in water clusters: based on atom-bond electronegativity equalization method fused into molecular mechanics. J Chem Phys 125:064311
14. Panuszko A, Gojlo E, Zielkiewicz J, Smiechowski M, Krakowiak J, Stangret J (2008) Hydration of simple amides FTIR spectra of HDO and theoretical studies. J Phys Chem B 112:2483–2493
15. Li ZY, Yu HB, Zhuang W, Mukamel S (2008) Geometry and excitation energy fluctuations of NMA in aqueous solution with CHARMM, AMBER, OPLS, and GROMOS force fields: implications for protein ultraviolet spectra simulation. Chem Phys Lett 452:78–83
16. Buck M, Karplus M (2001) Hydrogen bond energetics: a simulation and statistical analysis of N-methyl acetamide (NMA), water, and human lysozyme. J Phys Chem B 105:11000–11015
17. Myshakina NS, Ahmed Z, Asher SA (2008) Dependence of amide vibrations on hydrogen bonding. J Phys Chem B 112:11873–11877
18. Dixon DA, Matsuzawa N (1994) Density-functional study of the structures and nonlinear-optical properties of urea. J Phys Chem 98:3967–3977
19. Han WG, Suhai S (1996) Density functional studies on N-methylacetamide water complexes. J Phys Chem 100:3942–3949
20. Langley CH, Allinger NL (2003) Molecular mechanics (MM4) and ab initio study of amide–amide and amide–water dimers. J Phys Chem A 107(26):5208–5216
21. Bour P, Keiderling TA (2003) Empirical modeling of the peptide amide I band IR intensity in water solution. J Chem Phys 119:11253–11262
22. Mennucci B, Martinez JM (2005) How to model solvation of peptides? Insights from a quantum-mechanical and molecular dynamics study of N-methylacetamide. 1. Geometries, infrared, and ultraviolet spectra in water. J Phys Chem B 109:9818–9829
23. Oian P, Lu LN, Yang ZZ (2009) Molecular dynamics simulations of N-methylacetamide (NMA) in water by the ABEEM/MM model. Can J Chem 87:1738–1746
24. Buck U, Huisken F (2000) Infrared spectroscopy of size-selected water and methanol clusters. Chem Rev 100:3863–3890
25. Ludwig R (2001) Water: from clusters to the bulk. Angew Chem Int Edn 40:1808–1827
26. Liu K, Cruzan JD, Saykally RJ (1996) Water clusters. Science 271:929–933
27. Pugliano N, Saykally RJ (1992) Measurement of quantum tunneling between chiral isomers of the cyclic water trimer. Science 257:1937–1940
28. Viant MR, Cruzan JD, Lucas DD, Brown MG, Liu K, Saykally RJ (1997) Pseudorotation in water trimer isotopomers using terahertz laser spectroscopy. J Phys Chem A 101:9032–9041
29. Gao JL, Freindorf M (1997) Hybrid ab initio QM/MM simulation of N-methylacetamide in aqueous solution. J Phys Chem A 101:3182–3188
30. Parker LL, Houk AR, Jensen JH (2006) Cooperative hydrogen bonding effects are key determinants of backbone amide proton chemical shifts in proteins. J Am Chem Soc 128:9863–9872

31. Zhang GL, Ma J, Wen J (2007) Interchain impacts on electronic structures of heterocyclic oligomers and polymers containing group 14, 15, and 16 heteroatoms: quantum chemical calculations in combination with molecular dynamics simulations. J Phys Chem B 111:11670–11679
32. Iwaoka M, Komatsu H, Katsuda T, Tomoda S (2004) Nature of nonbonded Se...O interactions characterized by O-17 NMR spectroscopy and NBO and AIM analyses. J Am Chem Soc 126:5309–5317
33. Chen AQ, Pu XM, He SH, Guo YZ, Wen ZN, Li ML, Wong NB, Tian AM (2009) Solvent effects on isolated formamide and its monohydrated complex: observations from PCM study. New J Chem 33:1709–1719
34. Cheng MH, Pu XM, Wong NB, Lia ML, Tian AM (2008) Substituent effects on the hydrogen-bonded complex of aniline–$H_2O$: a computational study. New J Chem 32:1060–1070
35. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA (1996) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules (vol 117, pg 5179, 1995). J Am Chem Soc 118:2309–2309
36. Jorgensen WL, Chandreskhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926–936
37. Berendsen HJC, Postma JPM, Vangunsteren WF, Dinola A, Haak JR (1984) Molecular dynamics with coupling to an external bath. J Chem Phys 81:3684–3690
38. Case DA, Darden TA, Cheatham TE III, Simmerling CL, Wang J, Duke RE, Luo R, Crowley M, Walker RC, Zhang W, Merz KM, Wang B, Hayik S, Roitberg A, Seabra G, Kolossvary I, Wong KF, Paesani F, Vanicek J, Wu X, Brozell SR, Steinbrecher T, Gohlke H, Yang L, Tan C, Mongan J, Hornak V, Cui G, Mathews DH, Seetin MG, Sagui C, Babin V, Kollman PA (2008) Amber 10, 10th edn. University of California San Francisco, San Francisco
39. Frisch MJ et al (2009) Gaussian 09, A.02 edn. Gaussian Inc., Wallingford
40. Boys SF, Bernardi F (1970) Calculation of small molecular interactions by differences of separate total energies—some procedures with reduced errors. Mol Phys 19:553–566
41. Reed AE, Curtiss LA, Weinhold F (1988) Intermolecular interactions from a natural bond orbital, donor–acceptor viewpoint. Chem Rev 88:899–926
42. Glendening ED, Reed AE, Carpenter JE, Weinhold F (1995) NBO version 3.1. University of Wisconsin, Madison
43. Harder E, Anisimov VM, Whitfield TW, MacKerell AD, Roux B (2008) Understanding the dielectric properties of liquid amides from a polarizable force field. J Phys Chem B 112:3509–3521
44. Andersson MP, Uvdal P (2005) New scale factors for harmonic vibrational frequencies using the B3LYP density functional method with the triple-ζ basis set 6-311+ G (d, p). J Phys Chem A 109:2937–2941

J Mol Model (2012) 18:1401–1408
DOI 10.1007/s00894-011-1170-9

ORIGINAL PAPER

# Theoretical investigation of the radical scavenging activity of shikonin and acylshikonin derivatives

**Ruifa Jin · Yin Bai**

Received: 20 April 2011 / Accepted: 28 June 2011 / Published online: 15 July 2011


**Abstract** The radical scavenging activity of shikonin and acylshikonin derivatives was studied by using density functional theory. The hydrogen bond property of the studied structures was investigated using the atoms in molecules (AIM) theory. It turned out that the hydrogen bond is important for good radical scavenging activity. The hydrogen atom transfer for shikonin and acylshikonin derivatives is difficult to obtain because of the high bond dissociation energy (BDE). However, shikonin and acylshikonin derivatives appear to be good candidates for the one-electron-transfer. The introduction of acyl groups for shikonin decreases the ionization potential (IP) values compared with that of shikonin. The acylshikonin derivatives with 1H-pyrrole, furan, and thiophene groups are expected to be of the highest radical scavenging activity among the compounds investigated in this study. Taking this system as an example, we present an efficient method for the investigation of radical scavenging activity from theoretical point of view.

**Keywords** Acylshikonin derivatives · Hydrogen-atom transfer mechanism · One-electron transfer mechanism · Radical scavenging activity · Shikonin

**Electronic supplementary material** The online version of this article (doi:10.1007/s00894-011-1170-9) contains supplementary material, which is available to authorized users.

R. Jin (✉)
Department of Chemistry, Chifeng University,
Chifeng 024000, China
e-mail: Ruifajin@163.com

Y. Bai
Medical College, Chifeng University,
Chifeng 024000, China

## Introduction

Isohexenylnaphthazarins (IHN), commonly known as shikonins and alkannins, are used as herbal medicines or natural colorants for food, cosmetic, and textile [1]. Monomeric shikonin and alkannin derivatives, mainly esters from the side chain hydroxyl group, have been identified in the outer surface of the roots of at least 150 species belonging to the genera *Alkanna*, *Lithospermum*, *Echium*, *Onosma*, *Anchusa* and *Cynoglossum* of the Boraginaceae family [2]. Recently interest on these two compounds increased because of their pharmaceutical activity. They are potent pharmaceutical substances with a well-established and wide spectrum of anticancer [3–7], antibacterial [8], antifungal [9], antiviral [10], anti-inflammatory [11], cytotoxic [12], enzyme inhibitor [13, 14], antioxidant [15–19], and radical scavenging [20, 21] biological activities.

Phenolic compounds can act as antioxidants by inhibiting biomolecules from undergoing oxidative damage through free radicals mediated reactions [22]. They can act in several ways, including direct quenching of reactive oxygen species, inhibition of enzymes, chelation of transition metals ions (especially iron and copper), H-abstraction tunneling effect, promotion of radical production, and regeneration of membrane-bound antioxidants such as R-tocopherol [23–25]. Phenols are widely used as antioxidants in living organisms. Phenoxyl radicals represent important intermediates in many biological applications [26, 27]. The function of phenolic antioxidants (ArOH) is to intercept and react with free radicals faster than the substrate. There are two generally accepted mechanisms for the radical scavenging processes of chain-breaking antioxidants ArOH, namely hydrogen atom transfer and one-electron transfer followed by proton transfer [27]. Both mecha-

nisms are important for the scavenging activity of reactive species by an ArOH in a certain chemical or biological system and may occur in parallel.

In the H-atom transfer, a free radical $R^{\bullet}$ removes a hydrogen atom from the antioxidant ArOH:

$$R^{\bullet} + ArOH \rightarrow RH + ArO^{\bullet} \quad (1)$$

The efficiency of the antioxidant ArOH depends on the stability of the radical $ArO^{\bullet}$. Furthermore, the conjugation and resonance effects help the electron flow easily from electron-donating moiety to the electron-withdrawing moiety. As a consequence, the stability of the radical $ArO^{\bullet}$ can be increased by conjugation and resonance effects. The bond dissociation energy (BDE) of the O-H bond is an important parameter to evaluate the antioxidant action, because the weaker the O-H bond the easier the reaction of free radical inactivation will be. A high rate of hydrogen atom transfer is expected to be related to a low BDE of phenolic O-H bond.

In the one-electron transfer mechanism, the antioxidant can give an electron to the free radical:

$$R^{\bullet} + ArOH \rightarrow R^{-} + ArOH^{+\bullet}. \quad (2)$$

Again, the radical cation arising from the electron transfer must be stable, so it does not react with substrate molecules. In this case, the ionization potential (IP) is the most significant energetic factor for the scavenging activity evaluation.

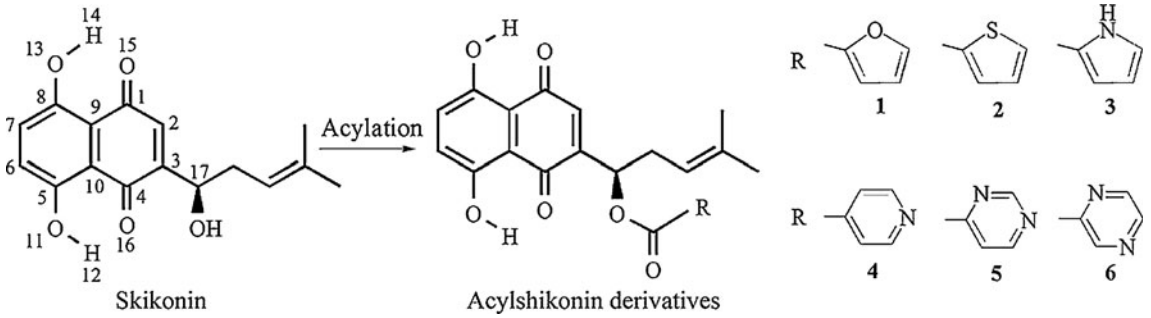
Molecules with the low IP and PDE values are expected to have high activity. Thus, in the present study, BDE and IP values were used as molecular descriptors in an effort to elucidate the radical scavenging activity of compounds under investigation. The antioxidant activity of the phenolic constituents is related to their structures. The intramolecular hydrogen bonds play important roles in stability of the radicals [25, 27–29]. Shikonin and acylshikonin derivatives contain two intramolecular hydrogen bonds between hydroxyl and keto groups. The structure-activity relationships of the shikonin and alkannin derivatives have been investigated in experiments [20]. However, to the best of our knowledge, no calculation about the radical scavenging activity of acylshikonin derivatives has been reported so far.

A acylshikonin derivative, 1-(1,4-dihydro-5,8-dihydroxy-1,4-dioxonaphthalen-2-yl)-4-methylpent-3-enyl-furan-2-caroxylate (1, as shown in Scheme 1), was reported [30]. This acylshikonin derivative has been investigated for topoisomerases I and II (Topo I and II) inhibition and proved to be a potential anticancer agent. Taking this system as an example, the structural and electronic properties of 1 and its radicals were investigated at density functional level. Further in-depth explanations for the experimental results have been discussed by the investigation of the radical scavenging activity and electronic properties of 1. To investigate the substituent effects, several derivatives (2–6) with different acyl groups, as shown in Scheme 1, have been designed to provide a demonstration for the rational design of new candidates for radical scavenging activity compounds. Furthermore, the hydrogen bonds of shikonin and 1–6 were explored using the atoms in molecules (AIM) theory. The aim was to get a further insight on the contribution of the structural features to the radical scavenging potential of compounds. The BDE and IP values calculated for these systems were used as indicators of the ease by which acylshikonin derivatives can deactivate free radicals. The spin distribution gives better insight into delocalization of the unpaired electron and conjugation effects, which is a measure for the stability of the radicals.

## Computational details

All calculations were performed by Gaussian 09 code [31]. The geometries of shikonin and 1–6 were optimized at the B3LYP/6-31+G(d) level without symmetry constraints (UB3LYP was used for the resulting radicals). Harmonic vibrational frequencies calculations using the same level of theory as for the geometry optimizations were performed for both parent (ArOH) molecules and radicals ($ArO^{\bullet}$ and $ArOH^{+\bullet}$) to ascertain the presence of local minima and to evaluate the zero-point energy (ZPE) corrections. Then, single-point electronic energies (SPEs) were obtained by using the 6-311++G(d,p) basis set. The molecular enthalpy in gas phase at 298 K consisted of B3LYP/6-311++G(d,p) calculated SPE values and B3LYP/6-31+G(d) thermal contributions to enthalpy (in which the vibrational contributions include zero-point vibrational energy) were used to determine the BDE values according to the formula BDE = $H_r + H_h - H_p$, where $H_r$ is the enthalpy of the radical generated by H-abstraction, $H_h$ is the enthalpy of the H-atom (−0.499897 Hartree at this level of theory), and $H_p$ is the enthalpy of the parent molecule. The IP values were estimated by using the equation IP = $E_{cr} - E_p$, where p and cr indicate the parent molecule and the corresponding cation radical generated after electron transfer. Accuracy of the energy evaluation for systems involving open-shell species is sensitive to spin contamination. Spin contaminations have been checked for radicals and radical cations, with calculated $< S^2 >$ values very close to the expectation value 0.75. Therefore, spin contamination should not bias obtained values. In order to gain insight into the delocalization possibilities of shikonin and acylshikonin derivatives (which are the reason for their antioxidant activity), spin distributions were calculated for shikonin and acylshikonin derivatives.

AIM theory [32, 33] was applied at the B3LYP/6-311++G(d,p) level to investigate the topology of electron density and the hydrogen bond property of the studied structures. The chemical bonding topology was investigated by confirming




**Scheme 1** Geometries of shikonin and acylshikonin derivatives, along with atom numbering

the presence of relevant bond paths and analysis of bond critical points, with special attention paid to the hydrogen bond. The AIM analysis was carried out using the AIM2000 program [34].

## Results and discussion

### Geometrical parameters

Table 1 presents the most important bond parameters for neutral forms of shikonin and 1–6. According to the suggested cutoff definition for D–H···A hydrogen bond (H···A distances<3.0 Å and D–H···A angles>110º) [35, 36], the interaction between hydroxy-H and oxygen of keto has been considered as hydrogen bond, which has been named as incipient (frozen) proton transfer reaction from D to A [36]. In the conformations of minimum energy of shikonin and 1–6, the hydroxyl groups are oriented in such a way to maximize hydrogen-bond-like interactions. The naphthoquinone rings in equilibrium structures of shikonin and 1–6 are almost planar with $C_5$-$C_{10}$-$C_4$-$O_{16}$ and $C_8$-$C_9$-$C_1$-$O_{15}$ torsional angles of about 5º and 0º, respectively. Inspection of Table 1 reveals clearly that the introduction of the different acyl groups on the side chain hydroxyl group of shikonin does not significantly affect the strength of the interaction of the hydrogen bonds. The values of $R_{O11–H12}$ and $\theta_{O11–H12\cdots O16}$ for 1–6 are greater, while the corresponding values of $R_{H12\cdots O16}$ are lower than those of shikonin, respectively. The values of $R_{H14\cdots O15}$ for 1–6 are greater, while the corresponding values of $R_{O13–H14}$ and $\theta_{O13–H14\cdots O15}$ are almost equal to those of shikonin, respectively. It indicates that the introduction of the acyl groups for shikonin leads to a slight decrease (increase) in the strength of the $H_{14}\cdots O_{15}$ ($H_{12}\cdots O_{16}$) hydrogen bonds for acylshikonin derivatives.

Starting from the absolute minima of each species, two radicals from shikonin and 1–6 were obtained upon H-atom abstraction from every hydroxyl phenolic group, whereas a single radical cation for each parent molecule was obtained by removing one electron. The cartesian coordinates of 5-OH radicals, 8-OH radicals, and radical cations for shikonin and 1–6 are given in Tables SI, SII, and SIII in supporting information, respectively. The naphthoquinone rings in equilibrium structures of 5-OH and 8-OH radicals for shikonin and 1–6 are nonplanar. The $C_5$-$C_{10}$-$C_4$-$O_{16}$ torsional angles of 5-OH radicals for shikonin and 1–6 are about 20º, while the corresponding values of $C_8$-$C_9$-$C_1$-$O_{15}$ torsional angles of 8-OH radicals are about 10º. The two OH groups in the 5- and 8-positions forces the system to assume a planar structure because of the presence of hydrogen-like interactions between 5-OH (8-OH) and $O_{16}$=$C_4$ ($O_{15}$=$C_1$). So, 5-OH and 8-OH radicals lacking of the 5-OH and 8-OH groups should be twisted (about 20° and 10º), respectively. Both 5-OH and 8-OH radicals contain one hydrogen bond. The values of $R_{O13–H14}$ and $\theta_{O13–H14\cdots O15}$ in 5-OH radicals for 1–6 are lower, while the corresponding values of $R_{H14\cdots O15}$ are greater than those of 5-OH radicals for shikonin, respectively. The values of $R_{O11–H12}$ and $\theta_{O11–H12\cdots O16}$ in 8-OH radicals for 1–6 are greater, while the corresponding values of $R_{H12\cdots O16}$ are lower than those

**Table 1** Distances of $R_{O–H}$ and $R_{O\cdots H}$ (in angstroms) and H-bond angles of $\theta_{O–H\cdots O}$ (in degree) for skikonin and 1–6

| Species | $R_{O11–H12}$ | $R_{O13–H14}$ | $R_{H12\cdots O16}$ | $R_{H14\cdots O15}$ | $\theta_{O11–H12\cdots O16}$ | $\theta_{O13–H14\cdots O15}$ |
|---|---|---|---|---|---|---|
| Skikonin | 0.994 | 0.996 | 1.705 | 1.711 | 145.8 | 146.2 |
| 1 | 0.996 | 0.996 | 1.686 | 1.711 | 146.1 | 146.4 |
| 2 | 0.996 | 0.996 | 1.686 | 1.713 | 146.1 | 146.3 |
| 3 | 0.995 | 0.996 | 1.687 | 1.712 | 146.1 | 146.4 |
| 4 | 0.996 | 0.996 | 1.686 | 1.714 | 146.1 | 146.2 |
| 5 | 0.996 | 0.996 | 1.686 | 1.713 | 146.1 | 146.2 |
| 6 | 0.996 | 0.996 | 1.686 | 1.713 | 146.1 | 146.2 |

of 5-OH radicals for shikonin, respectively. Hence, the introduction of the acyl groups for shikonin leads to a slight decrease (increase) in the strength of the $H_{14}\cdots O_{15}$ ($H_{12}\cdots O_{16}$) hydrogen bonds in 5-OH (8-OH) radicals for acylshikonin derivatives. Comparing the results shown in Table 1 with Tables SI and SII, one can find that the values of $R_{O11-H12}$ and $\theta_{O11-H12\cdots O16}$ in 8-OH radicals for shikonin and 1–6 are greater by about 0.01 Å and 2º, while the corresponding values of $R_{H12\cdots O16}$ are lower by about 0.08 Å than those in their neutral forms, respectively. The values of $R_{O13-H14}$ and $\theta_{O13-H14\cdots O15}$ in 5-OH radicals for shikonin and 1–6 are greater by about 0.008 Å and 2º, while the corresponding values of $R_{H14\cdots O15}$ are lower by 0.06 Å than those in their neutral forms, respectively. It suggests that H-atom abstraction from one hydroxyl phenolic group leads to an enhancement of the strength of another hydrogen bond of radicals compared with those in neutral forms. An enhancement of the strength of hydrogen bond in 5-OH and 8-OH radicals confers high stability to 5-OH and 8-OH radicals of acylshikonin derivatives, respectively.

For radical cations, the naphthoquinone rings in equilibrium structures of shikonin and 1–6 are almost planar with $C_5$-$C_{10}$-$C_4$-$O_{16}$ and $C_8$-$C_9$-$C_1$-$O_{15}$ torsional angles of about 5º and 0º as in their neutral forms, respectively. The values of $R_{O11-H12}$ and $\theta_{O11-H12\cdots O16}$ in radical cations for 1–6 are greater, while the corresponding values of $R_{H12\cdots O16}$ are lower than those of radical cation for shikonin, respectively. The values of $R_{H14\cdots O15}$ for 1–6 are greater, while the corresponding values of $R_{O13-H14}$ and $\theta_{O13-H14\cdots O15}$ in radical cations are almost equal to those of radical cation for shikonin, respectively. Thus, the introduction of the acyl groups for shikonin leads to a slight decrease (increase) in the strength of the $H_{14}\cdots O_{15}$ ($H_{12}\cdots O_{16}$) hydrogen bonds in radical cations for acylshikonin derivatives with respect to shikonin. Comparing the results shown in Table 1 with Table SIII in supporting information, the values of both $R_{O11-H12}$ and $R_{O13-H14}$ of radical cations for shikonin and 1–6 are greater by about 0.007 Å, while the corresponding values of $R_{H14\cdots O15}$ are lower by about 0.01 Å than those in their neutral forms, respectively. However, the values of $R_{H12\cdots O16}$ of radical cations for shikonin and 1–6 are almost equal to those in their neutral forms, respectively. The values of both $\theta_{O11-H12\cdots O16}$ and $\theta_{O13-H14\cdots O15}$ of radical cations for shikonin and 1–6 are lower by about 1º than those in their neutral forms, respectively. It indicates that the strength of $H_{14}\cdots O_{15}$ hydrogen bonds in radical cations is stronger, while the corresponding strength of $H_{12}\cdots O_{16}$ hydrogen bonds is almost equal to those in their neutral forms, respectively.

## Hydrogen atom transfer mechanism

The 5-OH and 8-OH radicals of shikonin and 1–6 are found to be practically isoenergetic because of both 5-OH and 8-OH radicals involving the breaking of the one hydrogen bond of the two hydrogen bonds in molecules. The difference between 5-OH and 8-OH radicals are less than 1 kcal $mol^{-1}$, except for shikonin (the difference between 5-OH and 8-OH radical is 1.64 kcal $mol^{-1}$). Figure 1 presents the spin densities of 5-OH and 8-OH radicals for shikonin and 1–3. The spin densities of 5-OH and 8-OH radicals for 4–6 are presented in Fig. SI in supporting information. Hydroxyanthraquinone derivatives are usually considered radical scavengers thanks to their excellent delocalization possibility. As clearly shown in Figs. 1 and SI, the unpaired electron for all 5-OH and 8-OH radicals of shikonin and 1–6 are fairly localized on the naphthoquinone rings without contributions from their

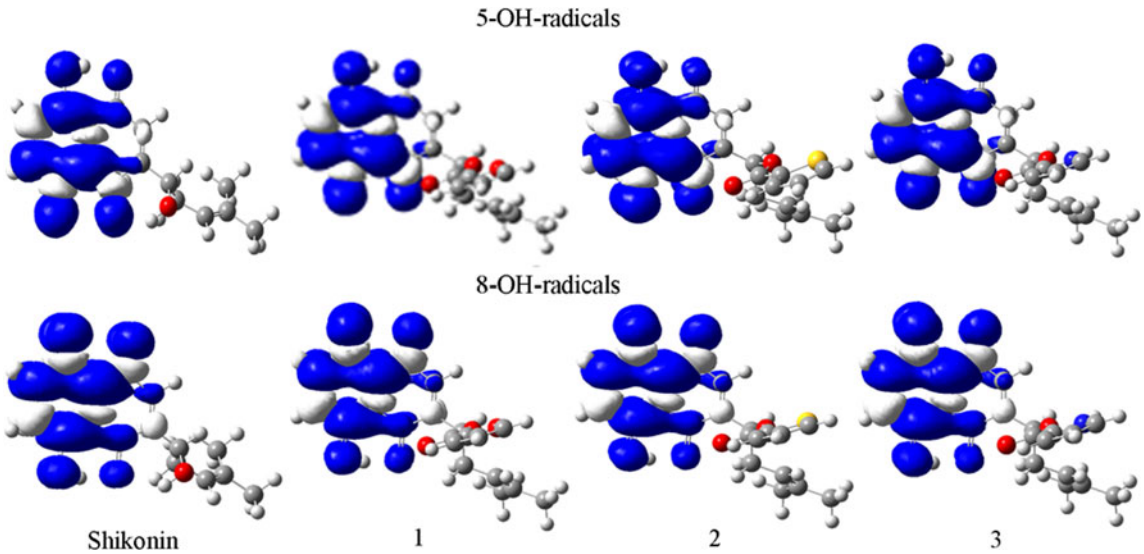



**Fig. 1** The spin densities of 5-OH and 8-OH radicals for shikonin and 1–3

**Table 2** Calculated BDE and IP values (all in kcal $mol^{-1}$) of skikonin and 1–6

| Species | BDE | | IP |
|---|---|---|---|
| | 5-OH | 8-OH | |
| Skikonin | 92.7 | 94.2 | 181.0 |
| 1 | 92.9 | 92.8 | 176.1 |
| 2 | 92.8 | 92.8 | 176.3 |
| 3 | 92.8 | 92.7 | 174.9 |
| 4 | 92.9 | 93.0 | 179.0 |
| 5 | 92.9 | 93.0 | 178.6 |
| 6 | 92.9 | 92.9 | 177.7 |

substituent groups 4-methylpent-3-enyl on 3-position and ester on 17-position. The main contributions of the spin densities come from naphthoquinones because of the planar conformation assumed by these species. The spin distributions of 5-OH radicals indicate the $C_{10}$ and oxygen $O_{11}$ from which the H atom is removed as the most probable radical centers, followed by the atoms $C_5$, $C_6$, $C_7$, $C_8$, and $O_{16}$. For example, the spin densities of $C_{10}$, $O_{11}$, $C_5$, $C_6$, $C_7$, $C_8$, and $O_{16}$ of 5-OH radical for 1 are 0.368, 0.350, -0.106, 0.178, -0.112, 0.189, and 0.114, respectively. For the 8-OH radicals, the unpaired electron center remains on the $C_9$ and oxygen atom $O_{13}$ from which the H atom is removed, followed by the atoms $C_5$, $C_6$, $C_7$, $C_8$, and $O_{15}$. For example, the spin densities of $C_9$, $O_{13}$, $C_5$, $C_6$, $C_7$, $C_8$, and $O_{15}$ of 8-OH radical for 2 are 0.359, 0.329, 0.179, -0.117, 0.164, -0.094, and 0.144, respectively. The spin distributions of 5-OH and 8-OH radicals for other acylshikonin derivatives are similar to those of 1 and 2, respectively.

The calculated BDE and IP values of shikonin and 1–6 are presented in Table 2. The BDE values characterize the hydrogen atom donating ability of shikonin and 1–6 in the gas phase. On the basis of gas phase BDE values, it is evident that the radical scavenging activity of both 5-OH and 8-OH radicals for 2–6 are expected to be equal to that of 1 since the corresponding BDE values are almost equal to that of 1. Such efficiency is related to the further stabilization of the derived formation of an intramolecular hydrogen bond in both 5-OH and 8-OH radicals. Comparing the BDE values of 5-OH and 8-OH radicals for 1–6 with those of shikonin, one can find that the introduction of the different acyl groups for shikonin leads to a slight increase (decrease) in the BDE values of 5-OH (8-OH) radicals. The hydroxyl is the most important site for the H-atom transfer. It suggests that an essential factor for a good activity as radical scavenging is the intramolecular hydrogen bond type structure. The introduction of acyl groups for shikonin decreases (increases) slightly the scavenging activities of 5-OH (8-OH) radicals. However, it can be noted that the BDE values the examined compounds are greater than that of phenol (BDE=82.9 kcal $mol^{-1}$ at this level of theory), which is generally chosen as the zero compound [37]. Abstraction of the H atom from the hydroxyl in positions 5 and 8 implies the breaking of the hydrogen bonds formed with the 4- and 1-keto groups, respectively. Hence, the hydrogen atom transfers for shikonin and 1–6 are difficult to obtain compared with phenol.

## Single electron transfer mechanism

According to the one-electron transfer mechanism, an electron is removed from the HOMO of the parent molecules, giving rise to radical cation [25]. Hydrogen bonds are normally retained as in the parent molecules, contributing to a further stabilization. Figure 2 presents the spin densities for the radical cations arising from shikonin and 1–3. The spin densities of radical cations for 4–6 are presented in Fig. SII in supporting information. All the radical cations of compounds under investigation show a higher degree of conjugation and delocalization with respect to their 5-OH and 8-OH radicals. From Figs. 2 and SII, one can find that the spin densities are fairly localized on the naphthoquinone rings and the substituent groups 4-methylpent-3-enyl on 3-position with only minor but nonzero contributions from ester groups on 17-position, except for 3. The contributions of the spin densities come from naphthoquinone ring, 4-methylpent-3-enyl, and ester groups for shikonin and 1–6, except for 3, are about 65%, 33%, and 2%, respectively. It is very interesting that the spin densities of the radical cation for 3 are delocalized over

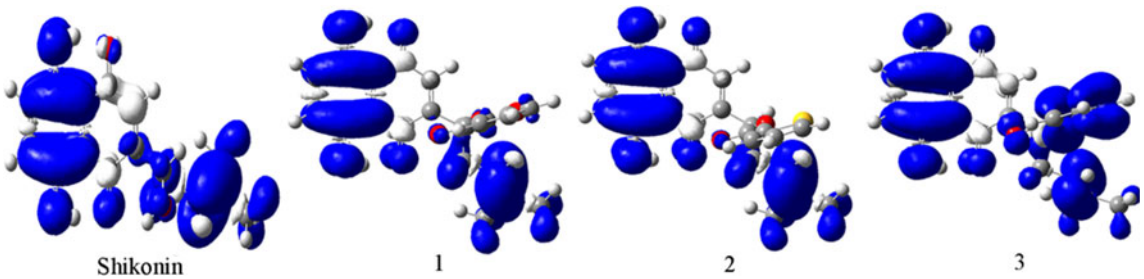



**Fig. 2** The spin densities of radical cations for shikonin and 1–3

**Table 3** Electronic density at BCP $\rho(r)_{bcp}$, the Laplacian $\nabla^2\rho(r)_{bcp}$ (in au), and the hydrogen bond energy $E_{HB}$ (in kcal mol$^{-1}$) of skikonin and 1–6

| | $O_{16}\cdots H_{12}$ | | | $O_{15}\cdots H_{14}$ | | |
|---|---|---|---|---|---|---|
| Species | $\rho(r)_{bcp}$ | $\nabla^2\rho(r)_{bcp}$ | $E_{HB}$ | $\rho(r)_{bcp}$ | $\nabla^2\rho(r)_{bcp}$ | $E_{HB}$ |
| Skikonin | 0.0476 | 0.1400 | 14.4 | 0.0472 | 0.1356 | 14.0 |
| 1 | 0.0500 | 0.1430 | 15.3 | 0.0471 | 0.1358 | 14.0 |
| 2 | 0.0500 | 0.1430 | 15.3 | 0.0470 | 0.1356 | 13.9 |
| 3 | 0.0500 | 0.1430 | 15.3 | 0.0471 | 0.1357 | 14.0 |
| 4 | 0.0500 | 0.1430 | 15.3 | 0.0468 | 0.1354 | 13.9 |
| 5 | 0.0500 | 0.1430 | 15.3 | 0.0469 | 0.1355 | 13.9 |
| 6 | 0.0500 | 0.1430 | 15.3 | 0.0469 | 0.1355 | 13.9 |

the entire molecule. The contributions of the spin densities come from naphthoquinone ring, 4-methylpent-3-enyl on 3-position, and methyl-1H-pyrrole-2-carboxylate groups are 61.7%, 15.1%, and 23.2%, respectively. The contribution of the spin densities come from ester group is greater by about 20% than those of other acylshikonin derivatives. Therefore, 3 shows a higher degree of conjugation and delocalization with respect to other acylshikonin derivatives and is expected to be of higher radical scavenging activity than those of the other acylshikonin derivatives.

The results displayed in Table 2 reveals that the trend for calculated IP values is different from that of BDE values. As expected, the lowest IP value in gas phase is found for 3 (174.9 kcal mol$^{-1}$), which is lower by 17.3 kcal mol$^{-1}$ than that of phenol (IP=192.2 kcal mol$^{-1}$ at this level of theory), followed by 1 (176.1 kcal mol$^{-1}$), 2 (176.3 kcal mol$^{-1}$), 6 (177.7 kcal mol$^{-1}$), 5 (178.6 kcal mol$^{-1}$), and 4 (179.0 kcal mol$^{-1}$). The electron-donating ability of shikonin and 1–6 seems to be related to an extended electronic delocalization over all the molecules. Systems having a high degree of $\pi$-delocalization are the most active among the one-electron-transfer mechanism. Furthermore, the introduction of the acyl groups for shikonin leads to a decrease in the IP values compared with that of shikonin, particularly for 3. The trend of IP values for the examined compounds suggests that an essential factor for good activity as an antioxidant is the hydrogen bond type structure, and it can be further improved by the introduction of acyl groups for the molecules. The IP is dependent on the HOMO energy of the parent molecule. For conjugated system, the IP is raised in energy, so the abstraction of an electron becomes very easy. The IP values of 1–3 are the lowest among the acylshikonin derivatives. It suggests that 1–3 are expected to be of the highest radical scavenging activity among the compounds investigated in this study. The results are in agreement with experimental results that 1 shows inhibitory activity on Topo I and Topo II [30].

## Atoms in molecules analysis

The AIM theory is often applied to study hydrogen bonds. The characteristics of bond critical points (BCPs), especially H···Y BCPs (where Y designates the proton acceptor), are very useful to estimate the strength of hydrogen bonds. To gain a deeper insight into the fundamental nature of OH···O hydrogen bond, it is crucial to obtain reasonable estimates of their relative energies. The electron densities, $\rho(r)_{bcp}$, and their Laplacians, $\nabla^2\rho(r)_{bcp}$, evaluated at BCPs are frequently used as indicators of hydrogen bond. More specifically, the electron energy density $H(r)_{bcp}$ at BCPs, as well as its components, the kinetic $G(r)_{bcp}$ and potential $V(r)_{bcp}$ electron energy densities, are often used to gain additional insight on the strength and nature of a given hydrogen bond. $H(r)_{bcp}$ are the sum of $G(r)_{bcp}$ and $V(r)_{bcp}$ at BCPs. The local kinetic

**Table 4** Electronic density at BCP $\rho(r)_{bcp}$, the Laplacian $\nabla^2\rho(r)_{bcp}$ (in au), and the hydrogen bond energy $E_{HB}$ (in kcal mol$^{-1}$) of radical cations for skikonin and 1–6

| | $O_{16}\cdots H_{12}$ | | | $O_{15}\cdots H_{14}$ | | |
|---|---|---|---|---|---|---|
| Species | $\rho(r)_{bcp}$ | $\nabla^2\rho(r)_{bcp}$ | $E_{HB}$ | $\rho(r)_{bcp}$ | $\nabla^2\rho(r)_{bcp}$ | $E_{HB}$ |
| Skikonin | 0.0456 | 0.1338 | 13.5 | 0.0490 | 0.1344 | 14.6 |
| 1 | 0.0500 | 0.1384 | 15.1 | 0.0484 | 0.1332 | 14.3 |
| 2 | 0.0500 | 0.1384 | 15.1 | 0.0485 | 0.1333 | 14.4 |
| 3 | 0.0499 | 0.1386 | 15.1 | 0.0482 | 0.1334 | 14.3 |
| 4 | 0.0503 | 0.1386 | 15.3 | 0.0485 | 0.1331 | 14.4 |
| 5 | 0.0500 | 0.1387 | 15.1 | 0.0477 | 0.1326 | 14.1 |
| 6 | 0.0499 | 0.1386 | 15.1 | 0.0480 | 0.1329 | 14.2 |

electron energy density can be evaluated from the values of $\rho(r)_{bcp}$ and $\nabla^2\rho(r)_{bcp}$ as [38]

$$G(r)_{bcp} = \frac{3}{10}(3\pi)^{\frac{2}{3}}\rho(r)_{bcp}^{\frac{5}{3}} + \frac{1}{6}\nabla^2\rho(r)_{bcp}. \quad (3)$$

The kinetic energy density $G(r)_{bcp}$ is in turn related to the potential energy density $V(r)_{bcp}$ through the local statement of the virial theorem [39]:

$$V(r)_{bcp} = \frac{h^2}{4m}\nabla^2\rho(r)_{bcp} - 2G(r)_{bcp}. \quad (4)$$

The hydrogen bond energy $E_{HB}$ (defined as -$D$e, where $D$e is the hydrogen bond dissociation energy) in molecules can be estimated within the framework of the AIM analysis using the relationship [40]

$$E_{HB} = -D_e = 0.5V(r)_{bcp}. \quad (5)$$

Table 3 presents the topological parameters and the $E_{HB}$ values for shikonin and 1–6. As a general rule, hydrogen bonds are characterized by positive values of $\nabla^2\rho(r)_{bcp}$ and low $\rho(r)_{bcp}$ values (< 0.1). Inspection of Table 3 reveals clearly that the $\rho(r)_{bcp}$ and $\nabla^2\rho(r)_{bcp}$ values of the $O_{16}\cdots H_{12}$ hydrogen bonds are greater, while the corresponding values of $O_{15}\cdots H_{14}$ in 1–6 are slightly lower than those in skikonin, respectively. It suggests that the $O_{16}\cdots H_{12}$ hydrogen bonds in 1–6 are stronger, while the $O_{15}\cdots H_{14}$ hydrogen bonds are slightly weaker than those in skikonin, respectively. Furthermore, the $E_{HB}$ values of $O_{16}\cdots H_{12}$ hydrogen bonds in 1–6 are greater, while the corresponding values of $O_{15}\cdots H_{14}$ in 1–6 are lower than those in 1, respectively. It suggests that the introduction of the acyl groups on shikonin leads to a slight decrease (increase) in the strength of the $H_{14}\cdots O_{15}$ ($H_{12}\cdots O_{16}$) hydrogen bonds for acylshikonin derivatives. The above results show qualitative agreement with the results based on the geometries.

Table 4 presents the topological parameters and the $E_{HB}$ values of radical cations for skikonin and 1–6. From Table 4, one can find that both the $\rho(r)_{bcp}$ and $\nabla^2\rho(r)_{bcp}$ values of $O_{16}\cdots H_{12}$ and $O_{15}\cdots H_{14}$ hydrogen bonds in 1–6 are slightly greater than those in skikonin, respectively. It suggests that the introduction of acyl groups for skikonin leads to enhancement of the strength of both $O_{16}\cdots H_{12}$ and $H_{14}\cdots O_{15}$ hydrogen bonds. Comparing the $\rho(r)_{bcp}$, their Laplacians $\nabla^2\rho(r)_{bcp}$, and $E_{HB}$ in neutral forms with those in radical cations shown in Tables 3 and 4, one can find that the strength of $H_{14}\cdots O_{15}$ hydrogen bonds in radical cations is stronger, while the corresponding strength of $H_{12}\cdots O_{16}$ hydrogen bonds is equal to those in their neutral forms, respectively. An enhancement of the strength of hydrogen bond in radical cations confers high stability to radical cations of acylshikonin derivatives.

## Conclusions

Our computational results predicted the radical scavenging activity of skikonin and acylshikonin derivatives. It turned out that the hydrogen bond is important for good radical scavenging activity, which confers high stability to the radical species. The introduction of acyl groups for shikonin decreases/increases slightly the radical scavenging activities of 5-OH/8-OH radicals. The hydrogen atom transfer for shikonin and acylshikonin derivatives is difficult to obtain compared with phenol, which is generally chosen as the zero compound. However, skikonin and acylshikonin derivatives appear to be good candidates for the one-electron-transfer, particularly for acylshikonin derivatives with thiophene, furan, and 1H-pyrrole groups (1–3). Their naphthoquinone planar conformation and the extended electronic delocalization between adjacent substituent groups determine low IP values. The introduction of the acyl groups for shikonin decreases the IP values compared with that of shikonin. The IP values of 1–3 are the lowest among the acylshikonin derivatives, suggesting that 1–3 are expected to be of the highest radical scavenging activity among the compounds investigated in this study. The results are in agreement with experimental results. Although the skikonin and acylshikonin derivatives act normally in water solution, our results are obtained in gas phase. However, we think that our calculations results are excellent primary indicators of free radical scavenging activity.

**Acknowledgments** Financial support from the Research Program of Sciences at Universities of Inner Mongolia Autonomous Region (NJzy08148) is gratefully acknowledged.

## References

1. Thomson RH (1971) Naturally occurring quinones. Academic, London, UK
2. Assimopoulou AN, Karapanagiotis I, Vasiliou A, Kokkini S, Papageorgiou VP (2006) Biomed Chromatogr 20:1359–1374
3. Deng R, Tang J, Xie BF, Feng GK, Huang YH, Liu ZC, Zhu XF (2010) Intern J Cancer 127:220–229
4. Ravelo AG, Estevez-Braun A, Chavez-Orellana H, Perez-Sacau E, Mesa-Siverio D (2004) Curr Top Med Chem 4:241–265
5. Garuti L, Roberti M, Pizzirani D (2007) Mini Rev Med Chem 7:481–489
6. Plyta ZF, Li T, Papageorgiou VP, Mellidis AS, Assimopoulou AN, Pitsinos EN, Couladouros EA (1998) Bioorg Med Chem Lett 8:3385–3390
7. Ahn BZ, Baik KU, Kweon GR, Lim K, Hwang BD (1995) J Med Chem 38:1044–1047
8. Kong Y, Zhang L, Yang Z, Han C, Hu L, Jiang H, Shen X (2008) Acta Pharmacol Sin 29:870–876
9. Sasaki K, Abe H, Yoshizaki F (2002) Biol Pharm Bull 25:669–670

10. Kim BH, Yoo J, Park SH, Jung JK, Cho H, Chung Y (2006) Arch Pharm Res 29:123–130
11. Singh B, Sharma MK, Meghwal PR, Sahu PM, Singh S (2003) Phytomedicine 10:375–380
12. Cui XR, Tsukada M, Suzuki N, Shimamura T, Gao L, Koyanagi J, Komada F, Saito S (2008) Eur J Med Chem 43:1206–1215
13. An S, Park YD, Paik YK, Jeonga TS, Lee WS (2007) Bioorg Med Chem Lett 17:1112–1116
14. Kajimoto S, Hori M, Manabe H, Masuda Y, Shibayama-Imazu T, Nakajo S, Gong XF, Obama T, Itabe H, Nakaya K (2008) BBA-Mol Basis Dis 1782:41–50
15. YoshidaL S, Kawada T, Irie K, Yuda Y, Himi T, Ikemoto F, Takano-Ohmuro H (2010) J Pharm Sci 112:343–351
16. Assimopoulou AN, Boskou D, Papageorgiou VP (2004) Food Chem 87:433–438
17. Kourounakisa AP, Assimopoulou AN, Papageorgiou VP, Gavalasa A, Kourounakis PN (2002) Arch Pharm Med Chem 6:262–266
18. Han J, Weng X, Bi K (2008) Food Chem 106:2–10
19. Gao D, Kakuma M, Oka S, Sugino K, Sakurai H (2000) Bioorg Med Chem 8:2561–2569
20. Ordoudi SA, Tsermentseli SK, Nenadis N, Assimopoulou AN, Tsimidou MZ, Papageorgiou VP (2011) Food Chem 124:171–176
21. Assimopoulou AN, Papageorgiou VP (2005) Phytother Res 19:141–147
22. Visioli F, Bellomo G, Galli C (1998) Biochem Biophys Res Commun 247:60–64
23. Leopoldini M, Russo N, Chiodo S, Toscano M (2006) J Agric Food Chem 54:6343–6351
24. Tejero I, González-García N, González-Lafont A, Lluch JM (2007) J Am Chem Soc 129:5846–5854
25. Leopoldini M, Pitarch IP, Russo N, Toscano M (2004) J Phys Chem A 108:92–98
26. Halliwell B, Gutteridge JMC (1989) Free radicals in biology and medicine. Oxford University, Oxford
27. Gugumus F (1990) Oxidation inhibition in organic mMaterials. CRC, Boca Raton
28. Nenadis N, Sigalas MP (2008) J Phys Chem A 112:12196–12202
29. Russo N, Toscano M, Uccella N (2000) J Agric Food Chem 48:3232–3237
30. Yang F, Chen Y, Duan W, Zhang C, Zhu H (2006) J Ding Int J Cancer 119:1184–1193
31. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Mennucci B, Petersson GA, Nakatsuji H, Caricato M, Li X, Hratchian HP, Izmaylov AF, Bloino J, Zheng G, Sonnenberg JL, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Vreven T, Montgomery JA Jr, Peralta JE, Ogliaro F, Bearpark M, Heyd JJ, Brothers E, Kudin KN, Staroverov VN, Kobayashi R, Normand J, Raghavachari K, Rendell A, Burant JC, Iyengar SS, Tomasi J, Cossi M, Rega N, Millam JM, Klene M, Knox JE, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Martin RL, Morokuma K, Zakrzewski VG, Voth GA, Salvador P, Dannenberg JJ, Dapprich S, Daniels AD, Farkas O, Foresman JB, Ortiz JV, Cioslowski J, Fox DJ (2009) Gaussian 09. Gaussian Inc, Wallingford, CT
32. Bader RFW (1991) Chem Res 91:893–928
33. Bader RFW (1998) J Phys Chem A 102:7314–7323
34. AIM2000 designed by Friedrich Biegler-Konig (2000) University of Applied Sciences, Bielefeld, Germany
35. Steiner T, Desiraju GR (1998) Chem Commun 8:891–892
36. Steiner T (2002) Angew Chem Int Edn 41:48–76
37. Wright JS, Johnson ER, Di Labio GA (2001) J Am Chem Soc 123:1173–1183
38. Abramov YA (1997) Acta Crystallogr A 53:264–272
39. Bader RFW (1994) Atoms in Molecules: A Quantum Theory. Clarendon, Oxford, UK
40. Espinosa E, Molins E, Lecomte C (1998) Chem Phys Lett 285:170–173

J Mol Model (2012) 18:1409–1420
DOI 10.1007/s00894-011-1158-5

ORIGINAL PAPER

# Computational synthesis of $C_{60}$ cyano- and azopolyderivatives

**Elena F. Sheka**

Received: 22 October 2010 / Accepted: 21 June 2011 / Published online: 15 July 2011


**Abstract** The cyanation of $C_{60}$ to $C_{60}(CN)_{18}$ and the aziridination of $C_{60}$ to $C_{60}(NH)_9$ were studied by an unrestricted broken spin symmetry Hartree–Fock approach implemented in semiempirical codes based on the AM1 technique. The calculations focused on the successive addition of CN and NH moieties to the fullerene cage following the identification of the target cage atoms as those with the highest atomic chemical susceptibilities calculated at each step. The results obtained were analyzed from the viewpoint of the parallelism between these derivatives as well as $C_{60}$ fluorides and hydrides. The difference between the first-stage $C_{60}$ chlorination and other sterically free processes is discussed.

**Keywords** Cyano and hydrocyano $C_{60}$ fullerenes · $C_{60}$ fulleroaziridines · $C_{60}$ chlorides · Quantum chemistry · Unrestricted broken symmetry approach · Atom chemical susceptibility · Computational synthesis

## Introduction

This study was undertaken to determine the generality of arguments favoring the electronic and structural parallelism of $C_{60}$ fluorides and hydrides, which have been derived both experimentally and computationally and are described in detail in [1]. Indeed, while fluorinated and hydrogenated $C_{60}$ fullerenes do have individual and intriguing peculiarities, they also do have two characteristics in common. The first of these is associated with the absence of steric limitations, which arises due to the monoatomic structures of the addends and their small size. The second characteristic is that atoms are added to the fullerene cage through 1,2-addition. A natural question then arises: how commonly are these two criteria for structural parallelism observed in other cases where the requirements for such parallelism are met. From a computational viewpoint, a crown-like $C_{3v}$ structure for the 18-fold derivatives appears to be the best manifestation of the serial event. It was therefore hard to overcome the temptation to look for other 18-fold derivatives which might be expected to fit the two requirements mentioned above and to look at their structures. CN and NH units looked like the best candidates as possible addends, as their addition to the $C_{60}$ cage is not accompanied by strong steric limitations, and they must be added through 1,2-addition. Recently performed calculations have shown that there is indeed structural parallelism between such derivatives and the fluorinated and hydrogenated $C_{60}$ fullerene families. However, when we turned our attention to chlorine, it soon became empirically apparent that there is no parallelism between chlorides and, say, fluorides. At first glance, the above two criteria do not appear to have been violated drastically in this case. We should therefore ask: is there another, as yet overlooked, criterion for the parallelism of $C_{60}$ polyderivatives? The current paper hopefully suggests an answer to this question.

E. F. Sheka (✉)
Peoples' Friendship University of Russia,
117197 Moscow, Russia
e-mail: sheka@icp.ac.ru

## On the grounds of computational methodology

Analysis of the odd-electron enhanced chemical activity of the $C_{60}$ molecule highlights its "chemical portrait," the atomic chemical susceptibility (ACS) distribution ($N_{DA}$ map) over the atoms of the molecule. As thoroughly discussed in [1], this distribution reveals five groups of atoms consisting of six identical pairs that differ by $N_{DA}$ values which gradually decrease from group 1 to group 5.

The initial step of any addition reaction involving $C_{60}$ fullerene must involve the atoms of group 1, as they are characterized by the highest $N_{DA}$. One may choose any of six pairs to start the reaction involved in attaching any addend R to the fullerene cage. When the first derivative $C_{60}R_1$ is formed, the map of its fullerene cage exhibits new high $N_{DA}$ rank atoms through which the reaction will then proceed. After the formation of $C_{60}R_2$, its $N_{DA}$ map reveals the sites for the next addition step, and so on. When the high-rank $N_{DA}$ data are closely packed, a few isomers are considered and an additional analysis of the formed isomers based on their total energies is performed in order to choose the isomer with the least energy. Following these methodology, a complete list of fluorinated $C_{60}F_{2k}$ [2] and hydrogenated $C_{60}H_{2k}$ [3] fullerenes was computationally synthesized. As in the case of $C_{60}$ fluorides and hydrides, the algorithmic approach based on the precalculated ACS is explored in the current study in order to computationally synthesize two families of $C_{60}$ polyderivatives related to $C_{60}(CN)_n$ and $C_{60}(NH)_m$ compounds. The stepwise successive synthesis is limited to $n$=18 and $m$=9, since linear CN units are added to individual carbon atoms of the fullerene cage, whereas every NH unit breaks one C–C bond, as each nitrogen atom interacts with two carbons. This limitation, however, suggests that we should check whether the numbers $n$=18 and $m$=9 are "magic" numbers, as the $C_{3v}$-crown structure of each polyderivative is expected to be the clearest manifestation of the parallelism associated with the formation of these polyderivatives.

## Polyhydrocyanides $C_{60}H(CN)_{2n-1}$ and polycyanides $C_{60}(CN)_{2n}$

The cyano and hydrocyano $C_{60}$ fullerenes that were first synthesized by Wudl et al. [4, 5] have attracted a great deal of attention as components of proton exchange membrane fuel cells [6, 7] due to their high proton conductivity. Those authors elaborated a multistep route for producing polyhydrocyanofullerenes from $C_{60}$ up to $C_{60}H(CN)_5$ that can be evidently expanded to produce higher members of the family according to the following series ("Series 1"):

$$\begin{aligned} C_{60} &\rightarrow C_{60}H(CN)_1 \rightarrow C_{60}(CN)_2 \rightarrow C_{60}H(CN)_3 \\ &\rightarrow C_{60}(CN)_4 \rightarrow \ldots \rightarrow C_{60}H(CN)_{2n-1} \\ &\rightarrow C_{60}(CN)_{2n}. \end{aligned} \quad (1)$$

This production route has been supported by other chemists (see for example [8, 9]), who have demonstrated the best way to produce hydrocyano $C_{60}H(CN)_{2n-1}$ and cyano $C_{60}(CN)_{2n}$ polyderivatives and their ions. Different molecular and anionic representatives of the two series covering members with $n$=1, 2, and 3 have been obtained so far.

Let us push the production algorithm for Series 1 into the background of the stepwise computational synthesis of these species. Although in practice alkali metal cyanides are the main reagents that are used to obtain anion salts of $C_{60}$ fullerenes, which are then quenched with different acids to get the products of the series [4], it is evident that in a computational experiment the formation of the products can be considered to be the result of intermolecular interactions between the partners of the $C_{60}$–HCN dyad. Calculations were performed within the framework of the unrestricted broken symmetry Hartree–Fock approach using the AM1 version of its semiempirical implementation [10]. As it turns out, the final product depends on the starting distances between the C and H atoms of the molecule and the cage atoms $C_c$; particularly for the hydrogen atom. If HCN is aligned along one of the $C_c$–$C_c$ bonds of the group 1 fullerene atoms with the highest ACS, the molecule is added to the fullerene cage associatively (product I) (see Fig. 1) when the $C_c$–H distance exceeds 1.7 Å. When the distance is less than 1.4 Å, the molecule is attached to the fullerene cage dissociatively (product II). The energy gain is 48.5 kcal $mol^{-1}$ in favor of product II, so the formation of the first product of Series 1 is energetically profitable. In turn, the barrier to the dissociation of the HCN molecule in the vicinity of $C_{60}$ fullerene is located at 1.4–1.7 Å .

Starting from product II and maintaining its equilibrium structure and the orientation of the hydrogen atom, let us now replace the hydrogen atom with the second CN unit and proceed with structural optimization. Thus, the equilibrium structure of biscyanofullerene $C_{60}(CN)_2$ is obtained. Its ACS map is then analyzed to pinpoint the two carbon atoms that are involved in the next step of the dissociative addition of HCN molecules. As it turns out, these two atoms are located in the equatorial space of the fullerene cage, which is also the case for fluorides and hydrides [2, 3]. The addition of the HCN molecule to these atoms results in the formation of $C_{60}H(CN)_3$. Replacing the hydrogen atom of the species by a CN unit, we obtain tetracyanofullerene $C_{60}(CN)_4$, the ACS map of which points to another pair of carbon atoms that are involved in the next HCN addition, resulting in the production of $C_{60}H(CH)_5$, and so forth. It should be noted that, in contrast to the fluorides and hydrides, the difference between the highest rank $N_{DA}$ and the remaining data in the $N_{DA}$ list is very pronounced, so there is no need to analyze isomers by energy in this case. Following this algorithm, the final product of this series, $C_{60}(CN)_{18}$, was obtained.

Figure 2 presents odd members of Series 1: the obtained hydrocyano $C_{60}$ fullerenes. The structures are drawn with a fixed projection of the fullerene cage, thus making it

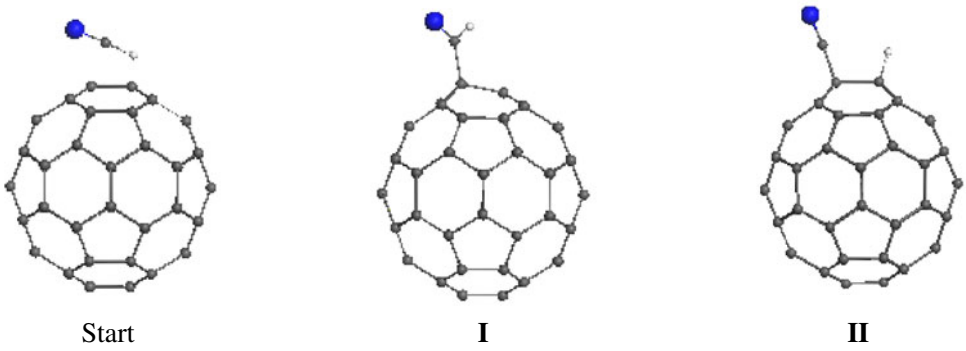



**Fig. 1** The first meeting of an HCN molecule with $C_{60}$ fullerene. The general starting composition and the equilibrium structures of products I and II are shown (see text). *Small black balls* indicate carbon atoms while *small white and big blue balls* show hydrogen and nitrogen atoms, respectively

possible to view successive changes in the structure of the molecule as a whole and the fullerene cage in particular. The even members of the series (the polycyanofullerenes) are shown in Fig. 3. The molecules are presented in a particular projection (we shall call it the "central-hexagon projection") in order to depict the evolution of the molecule as a whole and the fullerene cage structure towards that of 18CN, with its exact $C_{3v}$ symmetry. The molecule has a crown structure similar to those of $C_{60}F_{18}$ and $C_{60}H_{18}$. The two last panels in the lower right corner compare the crown

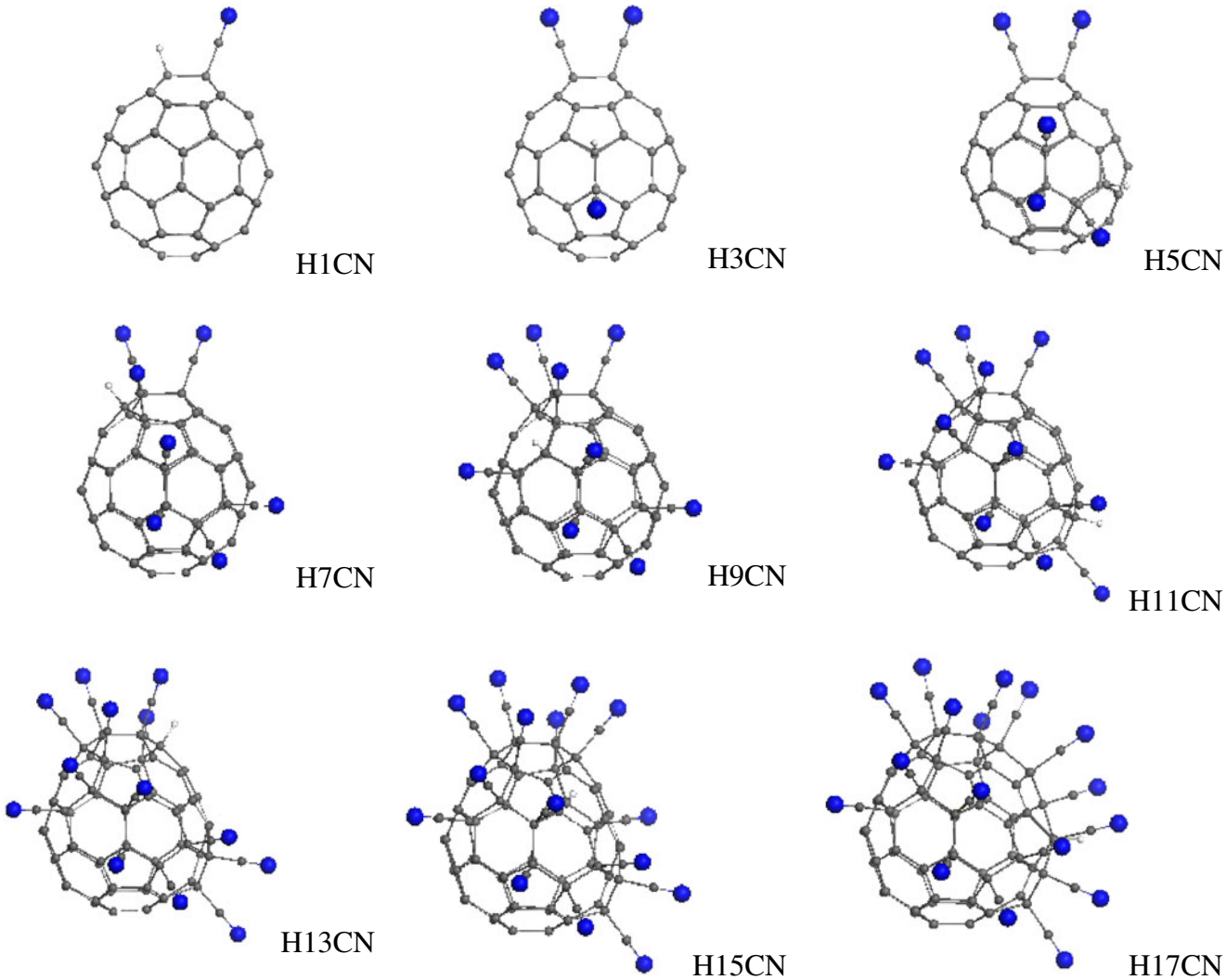



**Fig. 2** Equilibrium structures of the hydrocyanofullerenes $C_{60}H(CN)_{2n-1}$ ($n$=1, …, 9). To simplify the notation, the concise term H($2n$ − 1)CN is used for each structure. For a key to the color labels of atoms, see the caption to Fig. 1

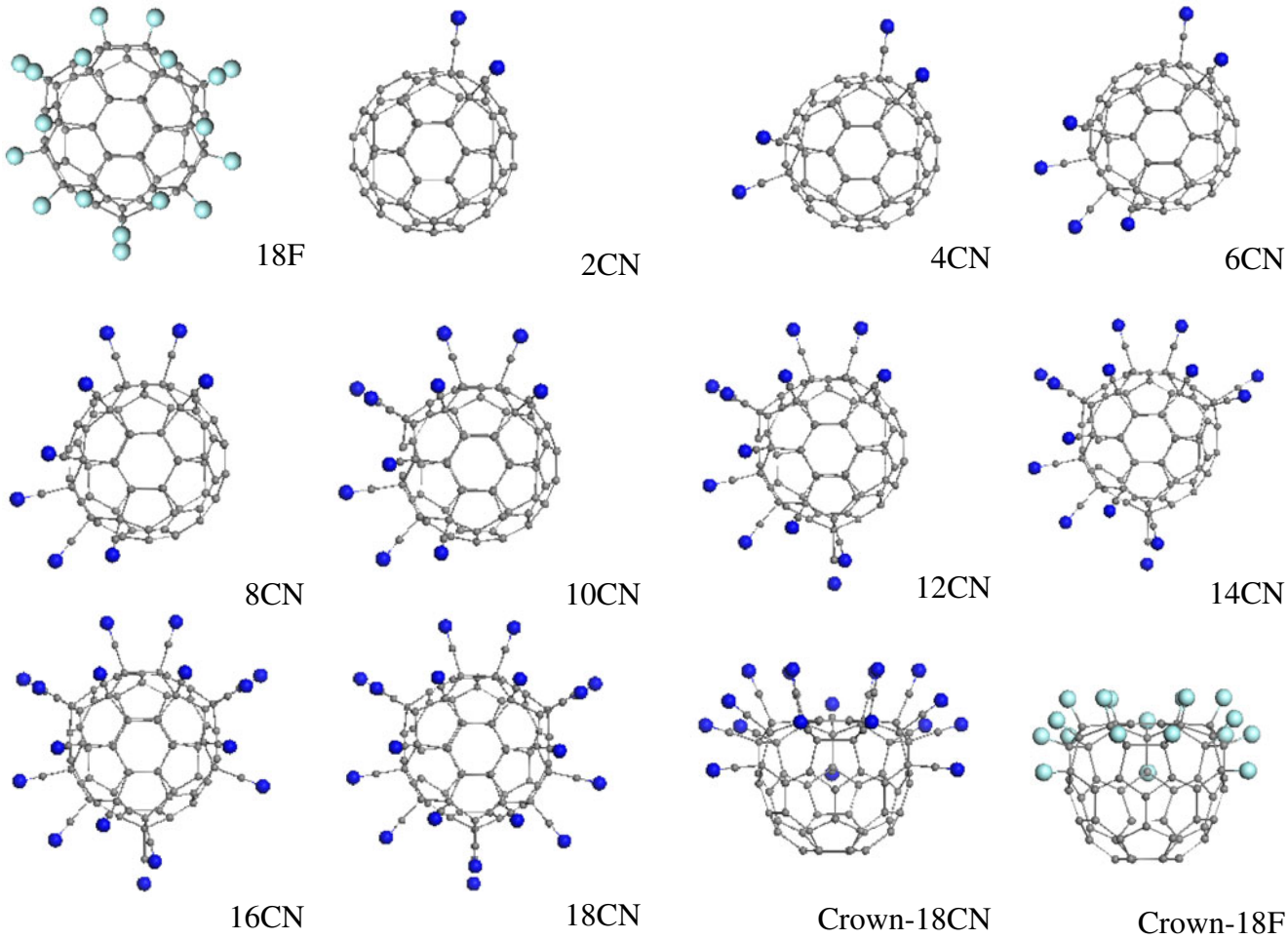



**Fig. 3** Equilibrium structures of the cyanofullerenes $C_{60}(CN)_{2n}$ ($n$=1, …, 9) in the central-hexagon projection. To simplify the notation, the concise term 2$n$CN is used for each structure. For a key to the color labels of atoms, see the caption to Fig. 1

structures of the 18CN and 18F molecules from [2]. As seen from both figures, there are no significant steric complications that accompany the successive cyanation of the fullerene cage when performed as a successive chain of 1,2-additions. The carbon skeleton preserves its closed shape and responds to the deformation caused by the cyanation by simply minimizing angular and bond strengths; a similar response is seen during fluorination and hydrogenation.

The successive changes in the fullerene skeleton structure with cyanation are presented in Fig. 4 for a fixed set of C–C bonds. The main change corresponds to the formation of a long C–C bond between cyanated carbon atoms and neighboring atoms in the uncyanated skeleton. The bond length starts at 1.55 Å for biscyanofullerene but gradually increases until it reaches 1.67 Å for higher members of the series. This considerable increase in the bond length is similar to that observed for $C_{60}$ fluorides [2], and is a measure of the skeletal deformation caused by the $sp^2$ to $sp^3$ transformation of the carbon atom configuration.

The symmetries and total energies of the products are listed in Table 1. Considering that each product of the series is produced in two stages, the coupling energy that is needed to produce a new species at each successive step should be estimated separately. Thus, the production of hydrocyanofullerenes is described by the coupling energy

$$E_{cpl}[H(CN)_{2n-1}] = \Delta H[H(CN)_{2n-1}] - \Delta H[(CN)_{2n-2}] - \Delta H(HCN),$$

where the right-hand side presents the total energies (heats of formation) of hydrocyanofullerene $C_{60}H(CH)_{2n-1}$, cyanofullerene $C_{60}(CN)_{2n-2}$, and the HCN molecule, respectively. In turn, the production of cyanofullerenes is described by the coupling energy

$$E_{cpl}[(CN)_{2n}] = \Delta H[(CN)_{2n}] - \Delta H[H(CN)_{2n-1}] - \Delta H(CN). \quad (2)$$

Here, the right-hand side presents the total energies of cyanofullerene $C_{60}(CN)_{2n}$ hydrocyanofullerene $C_{60}H$

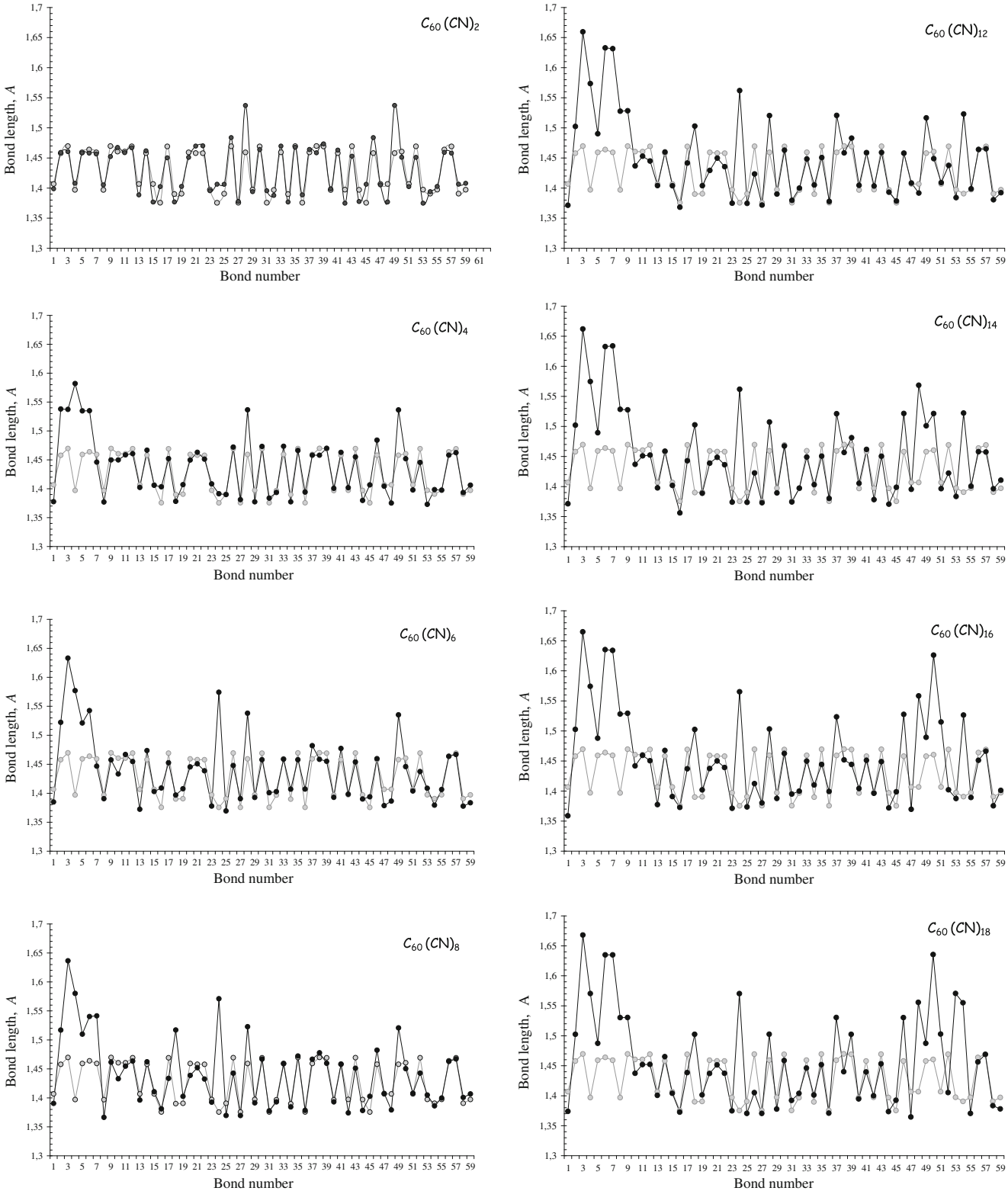



**Fig. 4** $sp^2$ to $sp^3$ transformation of the $C_{60}$ cage structure during the course of successive cyanations. *Lines with gray dots and black dots* correspond to the bond length distributions of the $C_{60}$ cage of pristine fullerene and those of its cyanated derivatives, respectively

$(CH)_{2n-1}$, and the CN molecular unit, respectively. The corresponding energy coupling data are listed in Table 1. As seen from the table, serial production is energetically favorable. $C_s$ symmetry was assigned to compound H1CN and $C_{2v}$ symmetry to compound 2CN on the basis of $^1$H NMR, $^{13}$C NMR, UV-vis, and IR spectra [4]. The latter

**Table 1** Total product and coupling energies and symmetry of polycyanides of fullerenes $C_{60}$

| | Polycyanides $C_{60}(CN)_{2n}$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 2CN | 4CN | 6CN | 8CN | 10CN | 12CN | 14CN | 16CN | 18CN |
| $\Delta H^a$ (kcal/mol) | 996.58 | 1042.44 | 1087.45 | 1146.94 | 1102.22 | 1247.62 | 1298.09 | 1349.46 | 1400.48 |
| $E_{cpl}$ (kcal/mol) | −67.39 | −60.76 | −65.40 | −64.02 | −66.41 | −64.20 | −67.39 | −60.23 | −90.36 |
| Symmetry | $C_{2v}$ | $C_s$ | $C_1$ | $C_s$ | $C_s$ | $C_1$ | $C_s$ | $C_1$ | $C_{3v}$ |
| | Polyhydrocyanides $C_{60}H(CN)_{2n-1}$ | | | | | | | | |
| | H1CN | H3CN | H5CN | H7CN | H9CN | H11CN | H13CN | H15CN | H17CN |
| $\Delta H^a$, (kcal/mol) | 951.70 | 990.93 | 1040.59 | 1098.68 | 1148.24 | 1199.55 | 1246.63 | 1297.47 | 1383.57 |
| $E_{cpl}$ (kcal/mol) | −34.65 | −36.63 | −32.84 | −34.41 | −29.69 | −33.42 | −31.98 | −31.61 | −24.65 |
| Symmetry | $C_1$ | $C_s$ | $C_s$ | $C_1$ | $C_s$ | $C_1$ | $C_1$ | $C_1$ | $C_1$ |

[a] Molecular energies are presented as the heats of formation $\Delta H$, determined as $\Delta H = E_{tot} - \sum_A \left(E^A_{elec} + EHEAT^A\right)$. Here $E_{tot}=E_{elec}+E_{nuc}$, where $E_{elec}$ and $E_{nuc}$ are the electron and core energies. $E^A_{elec}$ and $EHEAT^A$ are the electron energy and the heat of formation of an isolated atom, respectively.

finding is fully consistent with calculations. For the former, calculations attribute $C_1$ symmetry to the H1CN structure. However, a detailed continuous symmetry analysis in the framework of the methodology discussed in [11] assigns 100% $C_s$ symmetry to the $C_1$ H1CN structure, thus removing the inconsistency.

## Polyazoderivatives $C_{60}(NH)_m$

$C_{60}$ fullerene derivatives that contain nitrogen atoms within the attached units represent the largest class of fullerenes currently produced (see the monographs [12, 13]). The structure of the product critically depends on the atomic

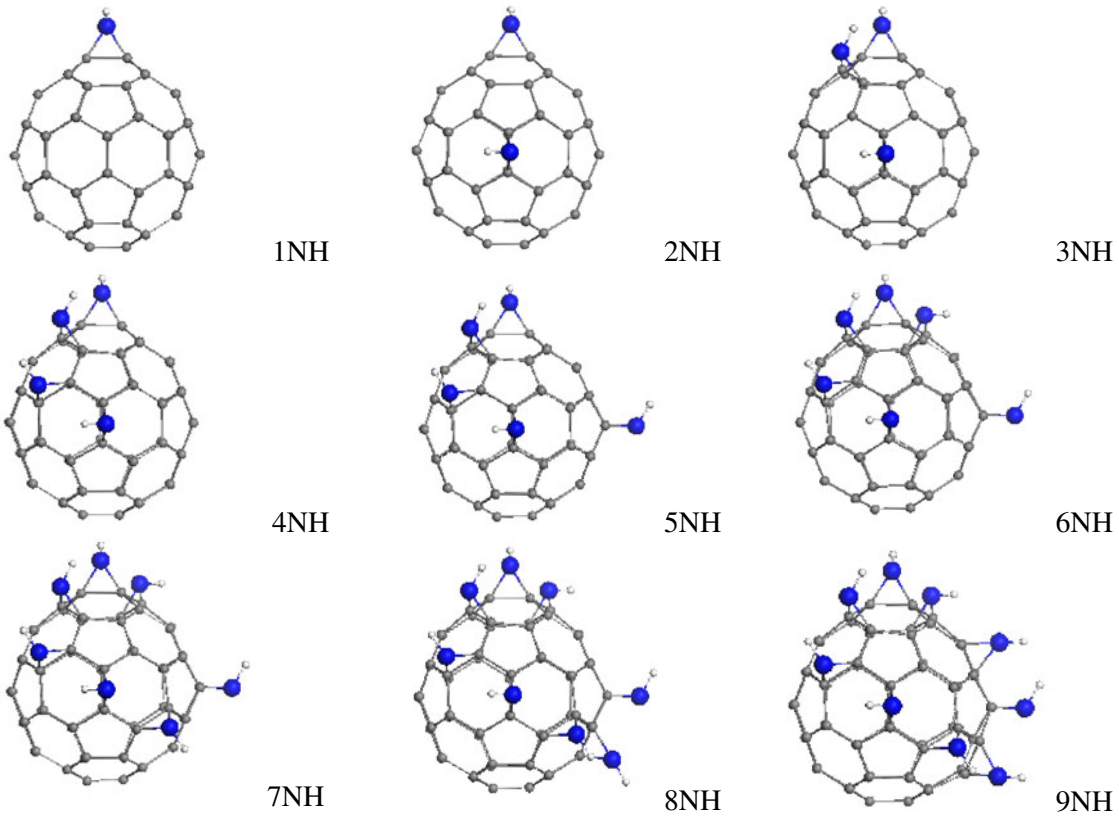



**Fig. 5** Equilibrium structures of the polyaziridines $C_{60}(NH)_m$ ($m$=1, …, 9). To simplify the notation, the concise term $m$NH is used. For a key to the color labels of atoms, see the caption to Fig. 1

composition of the addend, particularly in the case of polyadditions. In the present work, we will consider a particular family of azoderivatives of $C_{60}$ fullerene [14] taken from among the large class of fulleroaziridines. Choosing an addend with only two atoms (NH) makes it interesting from the viewpoint of expected structural parallelism. On the other hand, since an NH unit attaches to the fullerene cage bidentately through the formation of two C–N bonds, 1,2-addition is supported. The computational synthesis discussed here covers a series of polyfulleroaziridines $C_{60}(NH)_m$ with $m$=1, 2,…, 9. The aziridine $C_{60}(NH)_9$ can be compared with the corresponding 18-fold fluoride, hydride, and cyanide derivatives (i.e., each derivative has 18 added units).

The full set of synthesized products is shown in Fig. 5. As previously described, the first product $(NH)_1$ is obtained when the nitrogen atom is attached to a couple of group 1 skeletal carbon atoms that are joined by a short bond. The $N_{DA}$ map of the product $C_{60}(NH)_1$ points to two equivalent pairs of carbon atoms situated in the equatorial region of the cage, which results in the production of the second aziridine $C_{60}(NH)_2$. In turn, its ACS map pinpoints a couple of carbon atoms in the direct vicinity of the first addition, thus causing the formation of aziridine $C_{60}(NH)_3$. Following this algorithm to its ultimate conclusion leads to the final product of the series, $C_{60}(NH)_9$. Just as for the cyanides, the high-rank $N_{DA}$ data are well separated from the others in the list, which is why the presented synthesis does not require an additional isomer energy analysis. Therefore, only the considered products should be expected during a real synthesis.

The structural projection shown in Fig. 5 is similar to that seen in Fig. 2, which makes possible to compare the two cases in terms of both the general patterns and more detailed differences seen with successive additions. As seen in the figure, the successive additions are different in the two cases, even though the final structures are similar in both cases. Another view of the structures of the aziridines

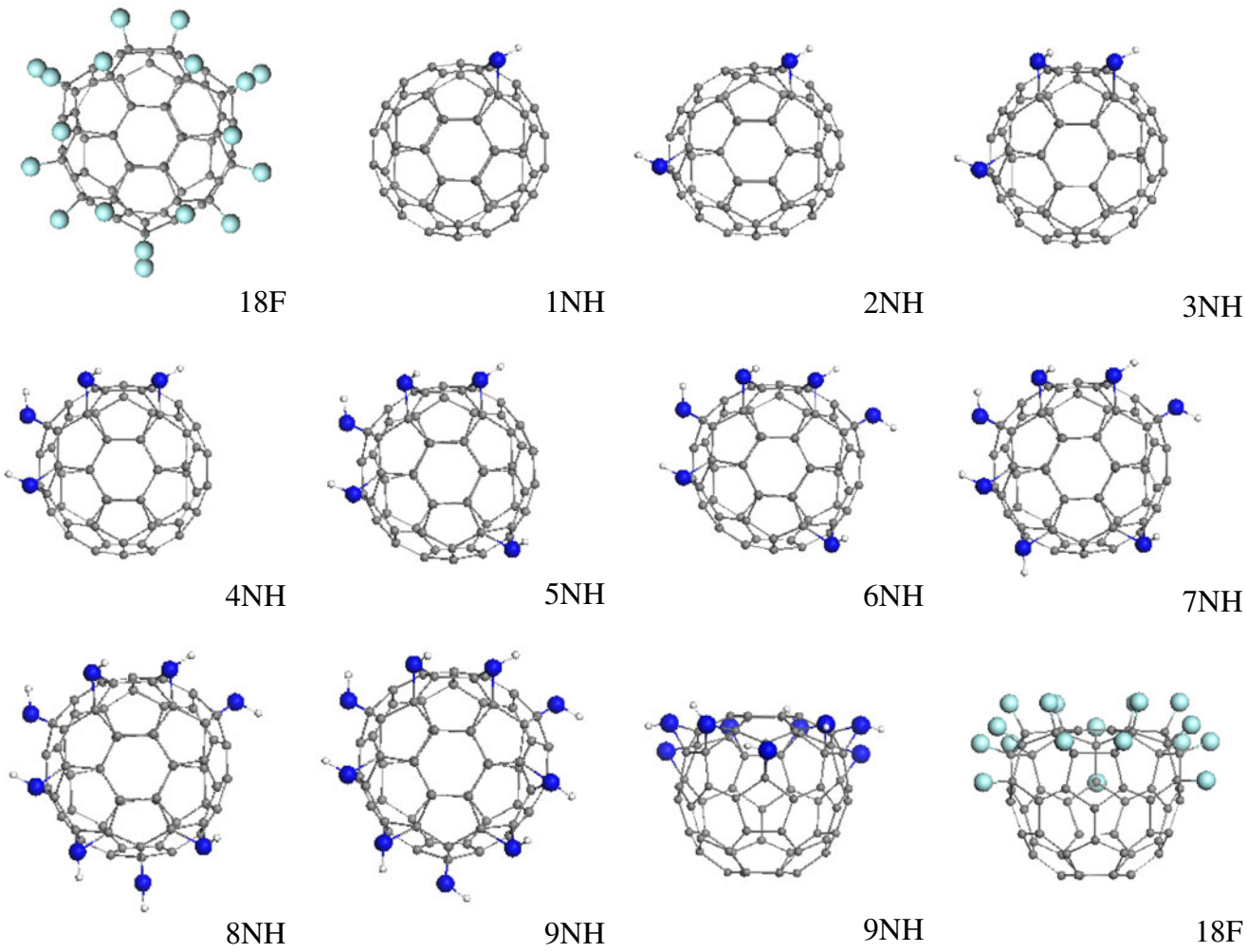



**Fig. 6** Equilibrium structures of the polyaziridines $C_{60}H(CN)_m$ ($m$=1, …, 9) in the central-hexagon projection

is given in Fig. 6. Again, the central-hexagon projection was chosen in order to show the origin and development of the $C_{3v}$ symmetry pattern (if it occurs). As seen in the figure, the nitrogen–carbon carcass of the aziridine

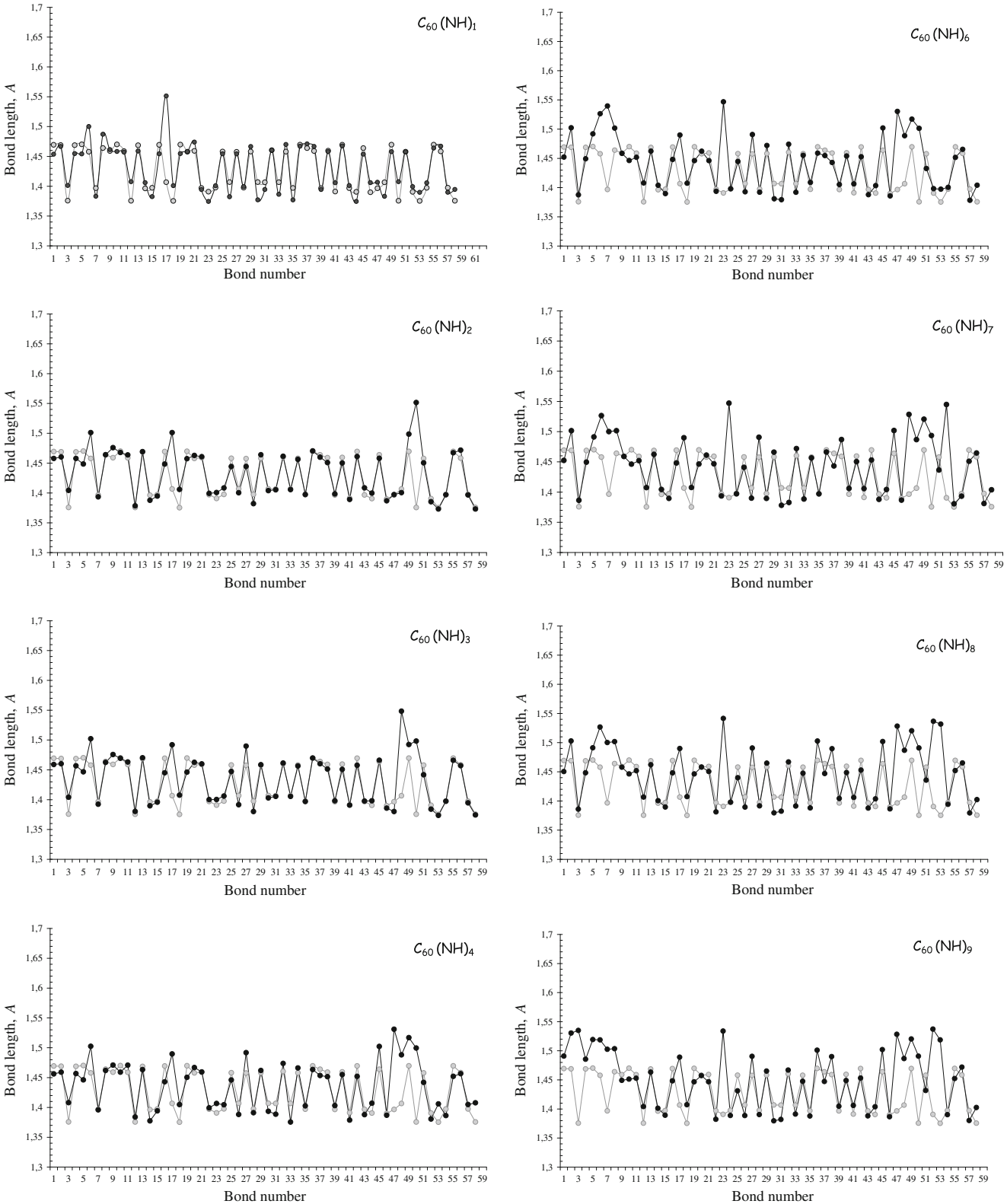



**Fig. 7** $sp^2$ to $sp^3$ transformation of the $C_{60}$ cage structure over the course of successive aziridination. *Lines with gray dots and black dots* correspond to the bond length distributions for the $C_{60}$ cage of pristine fullerene and its fulleroaziridines, respectively

**Table 2** Total product and coupling energies and symmetries of parent polyazidines of $C_{60}$ fullerenes

| | Polyaziridines $C_{60}(NH)_m$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1NH | 2NH | 3NH | 4NH | 5NH | 6NH | 7NH | 8NH | 9NH |
| $\Delta H$[a] (kcal/mol) | 965.24 | 974.24 | 990.05 | 992.57 | 1006.21 | 1011.58 | 1030.02 | 1038.36 | 1051.78 |
| Symmetry | $C_1$ | $C_1$ | $C_1$ | $C_1$ | $C_1$ | $C_1$ | $C_1$ | $C_1$ | $C_1$ |

[a] See footnote to Table 1

$C_{60}(NH)_9$ has a characteristic $C_{3v}$ pattern. However, due to the significant dispersion in N–H bond spatial orientations at practically the same energy, which causes pronounced scattering in the position of the hydrogen atom, the exact symmetry of the aziridine is $C_1$. On the other hand, structural analysis from the viewpoint of continuous symmetry [11] shows that the structure has 99.7% $C_{3v}$ symmetry, so we can convincingly speak of the deep structural parallelism of this aziridine with 18-fold fluoride, hydride, and cyanide derivatives.

The addition of NH units to the fullerene cage, just like all of the other addends considered so far, causes noticeable deformation of the cage. Figure 7 shows deformation of the cage in terms of selected C–C bond lengths during the course of the aziridination. The symmetries and total energies of the products are listed in Table 2. The evolutions of the total product energy and the coupling energy needed for the next addition of NH are shown in Fig. 8. The coupling energy is determined as

$$E_{coupl}[(NH)_m] = \Delta H[(NH)_m] - \Delta H[(NH)_{m-1}] - \Delta H(NH). \quad (3)$$

As seen in the figure, the total energy gradually increases as the aziridination proceeds. Despite of this fact, the coupling energy (determined according to Eq. 3) remains negative and quite large in magnitude. This means that serial production is energetically favorable. The evolution of the total number of

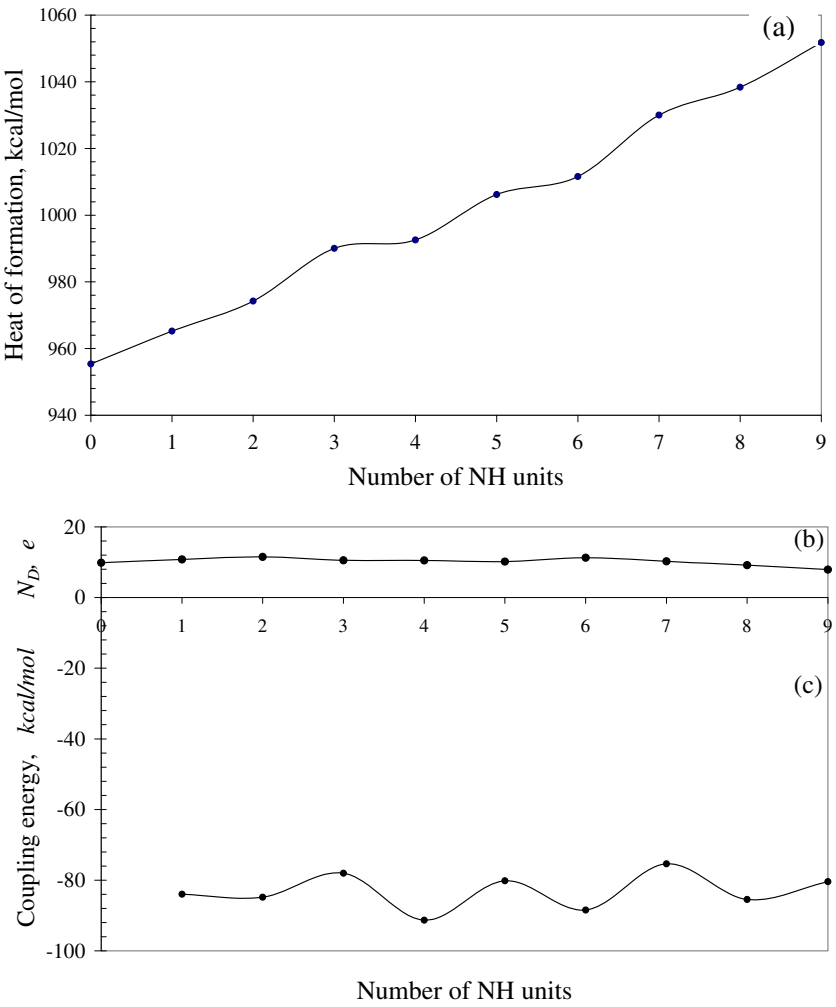



**Fig. 8** **a**–**c** Evolutions of the total energy (**a**) as well as the molecular chemical susceptibility $N_D$ (**b**) and the coupling energy $E_{cpl}$ (**c**) as the number of CN units attached to the $C_{60}$ skeleton increases

effectively unpaired electrons in the family, as shown in the top panel of Fig. 8b, demonstrates that all members of the series possess high chemical reactivity.

So far there has only been one report [14] on the production and characterization of the first member of this series, which assigned its symmetry to the $C_{2v}$ point group. This substance (denoted molecule II) was discussed in [11] in regards to the continuous symmetry of fullerene derivatives and their electronic spectra. According to the analysis performed, the continuous symmetry of the $C_1$ molecule consists of ~94% $I_h$ symmetry, which is supported by the high-symmetry pattern of its optical spectrum. Obviously, a high $I_h$ grade of continuous symmetry implies that symmetric patterns related not only to $I_h$ itself but to its subgroups might be expected. From this viewpoint, attributing the high-symmetry pattern of $^{13}C$ NMR spectra of these species to the $C_{2v}$ group [14] does not seem strange and inconsistent with computational predictions. The continuous symmetry contribution of $C_{3v}$ to the molecular structure of 9NH is 98.7%.

## Concluding remarks: a little about $C_{60}$ chlorination

Our suggestion that there is a strong similarity in the structural patterns of $C_{60}$ fullerene derivatives when the relevant additions occur without steric limitations and via 1,2-additions appears to be true. Additional evidence in support of this statement can be obtained when comparing the structures of the 18-fold derived $C_{3v}$ crown members of all four families in Fig. 9. When looking at that figure, it is important to bear in mind that the bond sets are only the same for fluorination and hydrogenation; the other two sets are grouped in a different manner because the numeration of atoms in the pristine fullerene molecule changes in the latter two cases. Nevertheless, the presented picture quite clearly reveals both the great similarity in the general behavior of the bond distribution and the degree of cage deformation for different series of additions. Obviously, fluorination and cyanation exert the largest effects, while hydrogenation and aziridination disturb the cage to a much weaker and comparable extent.

Evidently, an absence of steric limitations and successive 1,2-additions are necessary to achieve strong parallelism of the relevant derivatives. However, we can also question whether these criteria are sufficient to cover all kinds of possible addition reactions that also meet these requirements. Thus, we cannot ignore $C_{60}$ chlorination, which occurs quite differently in practice to fluorination and hydrogenation [15]. Evidently, the steric limitations are more severe in that case. However, they are not too strong to explain the drastic difference observed from fluorination

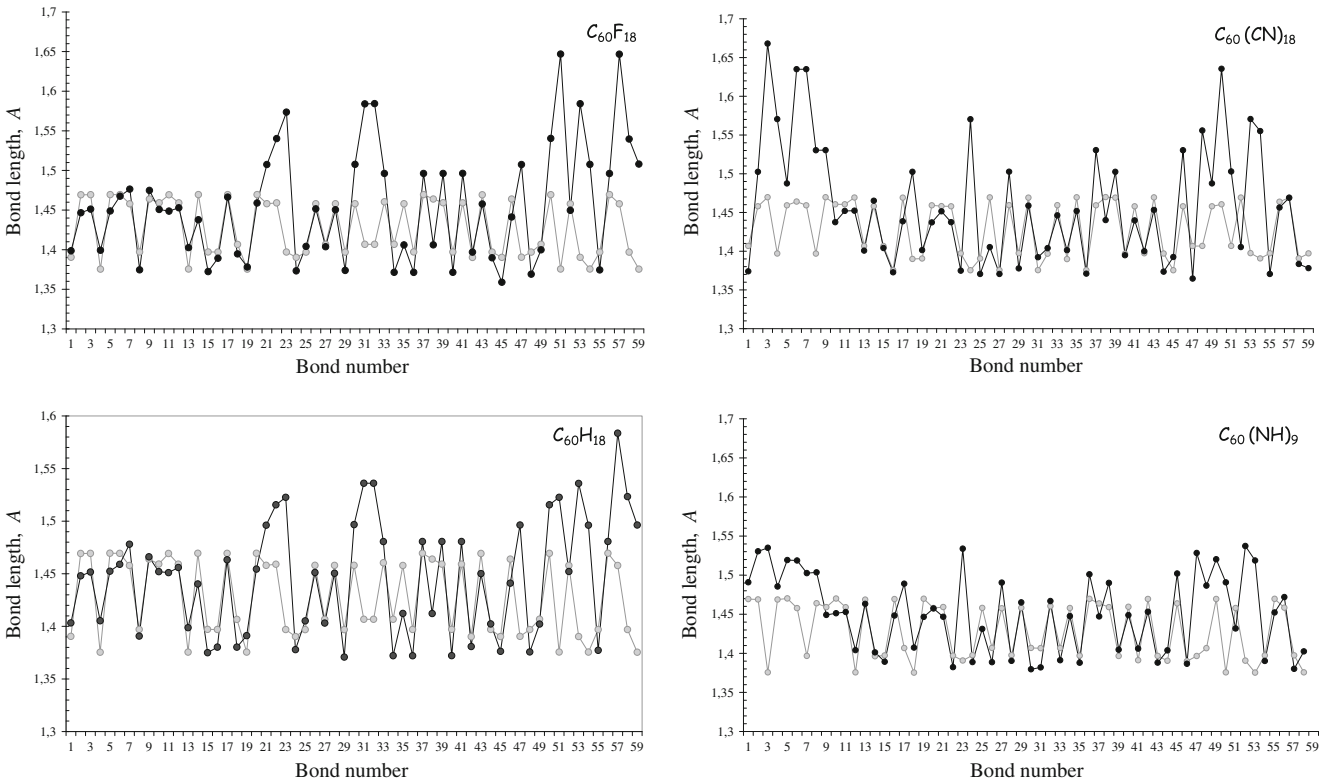



**Fig. 9** Comparison of the $sp^2$ to $sp^3$ transformation of the $C_{60}$ cage structure for 18-fold derivatives obtained through fluorination, hydrogenation, cyanation, and aziridination

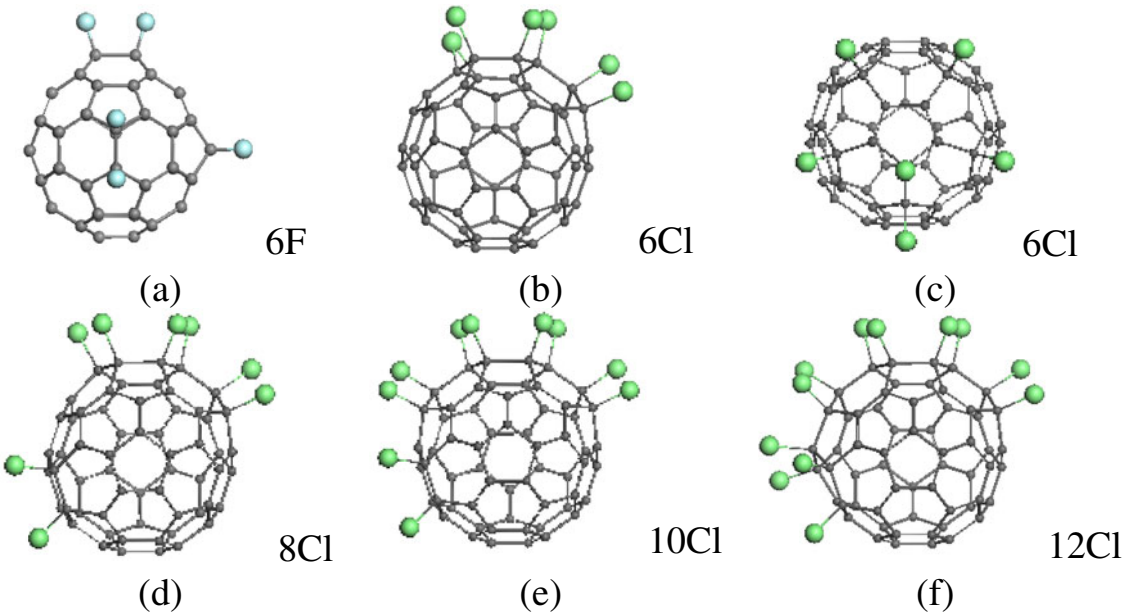



**Fig. 10** **a**–**f** Equilibrium structures of the fluoride $C_{60}F_6$ (**a**) and the chlorides $C_{60}Cl_6$ (chloride **A** in the main text) (**b**), $C_{60}Cl_8$ (**d**), $C_{60}Cl_{10}$ (**f**), and $C_{60}Cl_{12}$ (**e**). The chloride $C_{60}Cl_6$ (chloride **B** in the main text) (**c**) is suggested on the basis of experimental data [19, 20]. *Small black balls* indicate carbon atoms, while *light blue and green balls* show fluorine and chlorine atoms, respectively

and hydrogenation. We will not enter the depths of the chlorination chemistry of $C_{60}$ fullerene here; we will restrict ourselves to only a short excursion into the first step in the computational synthesis of $C_{60}$ chlorides within the framework of the ACS-guided approach.

As it turns out, the stepwise synthesis of chlorides can be broken down into a sequence of 1,2-additions, as in all of the cases considered earlier. A predicted but rather small preference for the 1,4-addition of a chlorine atom to the $C_{60}$ cage over 1,2-addition in the first step [16] is inverted when the distribution of effectively unpaired electrons over the cage atoms is taken into account. However, the chlorination does proceed in a different way to, say, fluorination. This difference is due to the third step of the reaction, when the third chlorine atom attaches to the cage after the first two have been added at two group 1 carbon atoms. In contrast to the third fluorine (hydrogen, cyano) and the second NH units, which were added in the equatorial space of the fullerene cage, the third chlorine atom is attached close to the first two atoms, revealing a tendency for a contiguous sequence of addition steps. It should be noted that the preference for the highest rank $N_{DA}$ data over the remaining ones is the strongest for chlorination.

The contiguous addition occurs until the chloride $C_{60}Cl_6$ is formed; Fig. 10 presents the structures of $C_{60}F_6$ (Fig. 10a) and $C_{60}Cl_6$ (Fig. 10b) for comparison. However, the

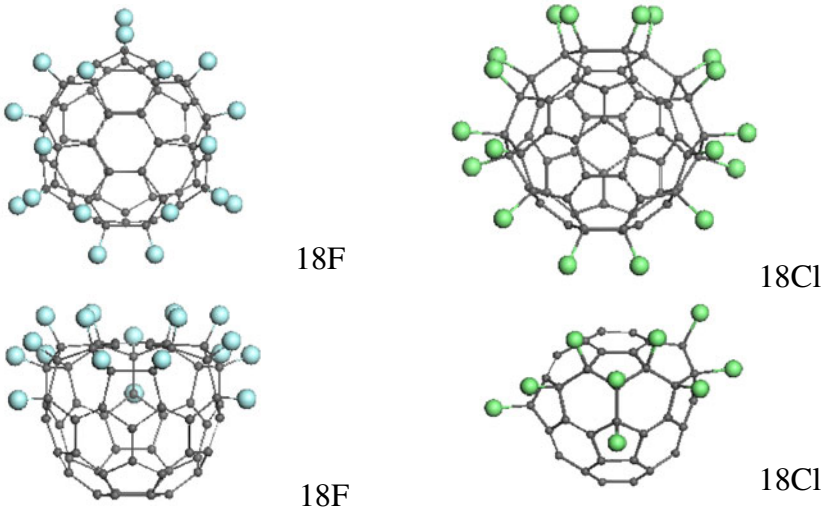



**Fig. 11** Two projections of the equilibrium structures of the fluoride $C_{60}F_{18}$ and the chloride $C_{60}Cl_{18}$. For a key to the color labels of atoms, see the caption to Fig. 10

contiguity of chlorine addition is terminated at the sixth step and subsequent 1,2-additions occur quite far from the first six atoms (see Fig. 10d–f). This leads to an evidently symmetric structure for the chloride $C_{60}Cl_{18}$, as shown in Fig. 11. In spite of the very different structural pattern compared to that of $C_{60}F_{18}$, the added atoms are concentrated in the upper part of the molecules, forming a crown-like "flower bowl" in the first case and a "flower bowl with arms" in the second. The addition of the twentieth chlorine atom causes the destruction of the cage. The resulting chemical bond rupture greatly increases the radicalization of the cage, which, in turn, strengthens the chemical reactivity of the chlorides, thus stimulating their further chlorination.

In 1991, it was shown that $C_{60}$ fullerene readily reacts with both gaseous and liquid chlorine to give a mixture of species with either 12 or 24 chlorine atoms per fullerene molecule in general [15, 17]. This appears to be connected with the fullerene cage destruction prompted by massive chlorination, as mentioned above. The first individual chloride, $C_{60}Cl_6$, was obtained two years later [18], and this remains the most studied representative of the family [19, 20]. $C_s$ symmetry was assigned to this chloride, which is consistent with the symmetry of the $C_{60}Cl_6$ species given in Fig. 10b (chloride **A**). However, the structure of the $C_{60}Cl_6$ molecule presented in Fig. 10c (chloride **B**), which was proposed on the basis of numerous synthetic data where the chloride was used as the initial chemical reactant, differs from the calculated structure quite drastically. Since the energy of molecule **B** is much lower than that of molecule **A**, and since chlorine atoms can easily move over the fullerene surface [21], the initially formed structure **A** may transform into **B** due to "chlorine atom dancing;" in other words, the fullerene cage re-isomerizes to suit new requirements concerning pentagon framing, just as it does for the $C_{60}F_{48}$ fluoride discussed in [1].

Besides $C_{60}Cl_6$, other chlorides (such as $C_{60}Cl_8$, $C_{60}Cl_{12}$, $C_{60}Cl_{28}$, and $C_{60}Cl_{30}$) are produced under severe experimental conditions [13]. These chlorides differ drastically from those of fluorides and hydrides and do not involve 18-, 24-, 36-, and 48-fold derived species, thus indicating that the chlorination process is significantly different from the fluorination and hydrogenation processes for such species. An initial contiguous chlorine addition stage seems to be responsible for the difference. Therefore, to specify the structural evolution of a derivative family, three criteria—steric limitations, tendency for 1,2-additions, and the initial sequence of additions—should be taken into account.

## References


1. Sheka EF (2011) Fullerenes: nanochemistry, nanomagnetism, nanomedicine, nanophotonics. Taylor & Francis, Boca Raton
2. Sheka EF (2010) Step-wise computational synthesis of fullerene $C_{60}$ derivatives. Fluorinated fullerenes $C_{60}F_{2k}$. J Exp Theor Phys 111:395–412
3. Sheka EF (2011) Computational synthesis of hydrogenated fullerenes from $C_{60}$ to $C_{60}H_{60}$. J Mol Mod. doi:10.1007/s00894-010-0895-1
4. Keshavarz KM, Knight B, Srdanov G, Wudl F (1995) Cyanodihydrofullerenes and dicyanodihydrofullerene: the first polar solid based on $C_{60}$. J Am Chem Soc 117:11371–11372
5. Jousselme B, Sonmez G, Wudl FJ (2006) Acidity and electronegativity enhancement of $C_{60}$ derivatives through cyano groups. Mat Chem 16:3478–3482
6. Wang HB, DeSousa R, Gasa J, Tasaki K, Stucky G, Jousselme B, Wudl FJ (2007) Fabrication of new fullerene composite membranes and their application in proton exchange membrane fuel cells. Membrane Sci 289:277–283
7. Tasaki K, Venkatesan A, Wang H, Jousselme B, Stucky G, Wudl FJ (2008) Hydrogen cyano fullerene derivatives as acid for proton conducting membranes. Electrochem Soc B 155:1077–1084
8. Konarev DV, Khasanov SS, Otsuka A, Yoshida Y, Saito G (2003) First ionic multi-component complex of fullerene $C_{60}(CN)_2$ with Co(II)tetraphenylporphyrin and bis(benzene)chromium. Synth Metals 133–134:707–709
9. Troshin PA, Khakina EA, Peregudov AS, Konarev DV, Soulimenkov IV, Peregudova SM, Lyubovskaya RN (2010) $[C_{60}(CN)_5]^-$: a remarkably stable [60]fullerene anion. Eur J Org Chem 17:3265–3268
10. Sheka EF (2007) Chemical susceptibility of fullerenes in view of Hartree–Fock approach. Int J Quantum Chem 107:2803–2816
11. Sheka EF, Razbirin BS, Nelson DK (2011) Continuous symmetry of $C_{60}$ fullerene and its derivatives. J Phys Chem A 115:3480–3490
12. Hirsch A, Brettreich M (2004) Fullerenes—chemistry and reactions. Wiley-VCH, London
13. Troshin PA, Troshina OA, Lyubovskaya RN, Razumov VF (2009) Functional derivatives of fullerenes. synthesis and applications to organic electronics and biomedicine (in Russian). Ivanovo State University, Ivanovo
14. Averdung J, Luftmann H, Mattay J (1995) Synthesis of 1,2-(2,3-dihydro-1H-azirino)-[60]fullerene, the parent fulleroaziridine. Tetrahedron Lett 36:2957–2958
15. Tebbe FN, Becker JY, Chase DB, Firment LE, Holler ER, Malone BS, Krusic PJ, Wasserman E (1991) Multiple, reversible chlorination of $C_{60}$. J Am Chem Soc 113:9900–9901
16. Matsuzawa N, Fukunaga T, Dixon DA (1992) Electronic structures of 1,2- and 1,4-$C_{60}X_{2n}$ derivatives with $n$=1, 2, 4, 6, 8, 10, 12, 18, 24, and 30. J Phys Chem 96:10747–10756
17. Olah GA, Bucsi I, Lambert C, Aniszfeld R, Trivedi NJ, Sensharma DK, Prakash GKS (1991) Chlorination and bromination of fullerenes. Nucleophilic methoxylation of polychlorofullerenes and their aluminum trichloride catalyzed Friedel–Crafts reaction with aromatics to polyarylfullerenes. J Am Chem Soc 113:9385–9387
18. Birkett PR, Avent AG, Darwish AD, Kroto HW, Taylor R, Walton DRM (1993) Preparation and $^{13}$ NMR spectroscopic characterisation of $C_{60}Cl_6$. Chem Commun 1230–1232
19. Troshin PA, Popkov O, Lubovskaya RN (2003) Some new aspects of chlorination of fullerenes. Full Nanotubes Carbon Nanostr 11:163–185
20. Kuvychko IV, Streletskii AV, Popov AA, Kotsiris SG, Drewello T, Strauss SH, Boltalina OV (2005) Seven-minute synthesis of pure Cs-$C_{60}Cl_6$ from [60]fullerene and iodine monochloride: first IR, Raman, and mass spectra of 99 mol% $C_{60}Cl_6$. Chem Eur J 11:5426–5436
21. Troshin PA, Lyubovskaya RN, Ioffe IN, Shustova NB, Kemnitz E, Troyanov SI (2005) Synthesis and structure of the highly chlorinated [60]fullerene $C_{60}Cl_{30}$ with a drum-shaped carbon cage. Angew Chem Int Ed 44:234–237

J Mol Model (2012) 18:1421–1430
DOI 10.1007/s00894-011-1159-4

ORIGINAL PAPER

# The molecular basis of IGF-II/IGF2R recognition: a combined molecular dynamics simulation, free-energy calculation and computational alanine scanning study

**Jingjing Guo · Xiaoting Wang · Huijun Sun · Huanxiang Liu · Xiaojun Yao**

Received: 27 January 2011 / Accepted: 21 June 2011 / Published online: 15 July 2011


**Abstract** Insulin-like growth factor-II (IGF-II) is a key regulator of cell growth, survival, migration and differentiation, and is thus pivotal in many cancers. An individual with a high IGF-II level is at high risk of developing cancer, whereas IGF2R is implicated as being important in tumor suppression. Thus, uncovering the essence of the IGF-II/IGF2R interaction is very important to understanding the origin of the tumor-suppressing effect of IGF2R. In this study, in order to investigate the interaction of the IGF-II/IGF2R complex and to characterize the binding hot spots of this interaction, a 10 ns molecular dynamics simulation combined with MM-PBSA/MM-GBSA computations and computational alanine scanning was performed on the IGF-II/IGF2R complex. From the results of the free-energy decomposition and the computational alanine scanning calculation, we identified the key residues in the IGF-II/IGF-2R interaction. The results from the calculation were consistent with reported experimental mutagenesis studies. The information on the interaction of IGF-II and IGF2R obtained is vital for understanding how the structure of IGF2R influences the function of IGF-II in growth and development. This study will also lead to new opportunities to develop molecular probes that can assist in diagnostic screening, and even novel approaches to controlling tumor development.

**Keywords** IGF-II/IGF2R interaction · Protein–protein interaction · Molecular dynamics simulation · Molecular mechanics generalized born surface area (MM-GBSA) · Computational alanine scanning

J. Guo · X. Wang · H. Sun · X. Yao
State Key Laboratory of Applied Organic Chemistry and Department of Chemistry, Lanzhou University, Lanzhou 730000, China

H. Liu
School of Pharmacy, Lanzhou University, Lanzhou 730000, China

X. Yao (✉)
Key Lab of Preclinical Study for New Drugs of Gansu Province, Lanzhou University, Lanzhou 730000, China
e-mail: xjyao@lzu.edu.cn

## Introduction

The insulin-like growth factor system consists of the IGF ligands (IGF-I and IGF-II), cell-surface receptors that mediate the biological effects of the IGFs [including the IGF-I receptor (IGF1R), the IGF-II receptor (IGF2R), and the insulin receptor (IR)], as well as a family of IGF-binding proteins (IGFBPs) [1, 2]. The IGF system is essential for normal embryonic and postnatal growth, and plays an important role in the function of a healthy immune system, lymphopoiesis, myogenesis, and bone growth, among other physiological functions. Deregulation of the IGF system leads to stimulation of cancer cell growth and survival [2].

Recently, it was demonstrated that IGF-II is a key regulator of cell growth, survival, migration, and differentiation. Individuals with serum IGF-II levels in the upper quartile of the normal range have a relatively high risk for developing breast, prostate, colon and lung cancer [1, 3–6]. Thus, both the expression and activity of IGF-II require tight regulation. It is generally believed that the biological effects of IGF-II are mostly mediated via the IGF-I receptor or isoform A of the insulin receptor, while the IGF-II receptor functions as a "scavenger" receptor for IGF-II [5, 7–11]. By targeting IGF-II to lysosomal degradation, IGF2R plays a key role in the maintenance of correct IGF-

II levels in the circulation and in target tissues [12]. Deregulation of IGF2R expression has therefore been associated with growth-related disease and cancer [13].

No crystal structure of IGF-II interacting with the full-length IGF-II receptor is available as yet, and the only resolved crystal structure is the structure of the complex of human IGF2R domains 11–13 with IGF-II (PDB code 2V5P) [14] (see Fig. 1). However, its resolution (4.1 Å) is insufficient for a detailed interpretation of side-chain placement and explicit bonding information.

Structural and energetic analyses of protein–protein interactions are challenging endeavors, both experimentally and theoretically. As experimental methods are very laborious and time-consuming, efforts have been made to produce fast and accurate computational methodologies. In recent years, molecular dynamics (MD) simulation has been widely used to study protein–protein interactions [15–22]. This simulation approach can provide not only plentiful dynamic structural information on protein complex structures in solution, but also a wealth of energetic information. The most rigorous MD-based approaches to estimating the binding free energy are the free-energy perturbation and thermodynamic integration methods [23–25]. However, because of their long convergence times, they are computationally intensive and prohibitive for large systems. Recently, the MM-PBSA [26, 27] and MM-GBSA [26, 28, 29] approaches have become of interest in studies of biomolecular association such as protein–protein complexes in aqueous solutions. In the MM-PBSA approach, the binding free energy is estimated as the sum of the gas-phase energies, the solvation free energies and the entropic contributions, averaged over a series of snapshots from MD trajectories. The electrostatic contribution to the solvation term is calculated by solving the Poisson–Boltzmann (PB) equation. If the PB model is replaced by a generalized Born (GB) model, we obtain the MM-GBSA method. The difference in the binding free energy can be obtained after calculating the binding free energies of the wild type and alanine mutants based on MM-PBSA/MM-GBSA separately, which is called the computational alanine scanning method [20, 30–32]. However, the GB model is more attractive because it is much faster than the PB model and allows the decomposition of the free energy into atomic contributions in a straightforward manner [16, 18]. Free-energy decomposition and computational alanine scanning have been developed to investigate binding features in detail at the atomic level. These two methods can complement each other during the study of protein–protein interactions.

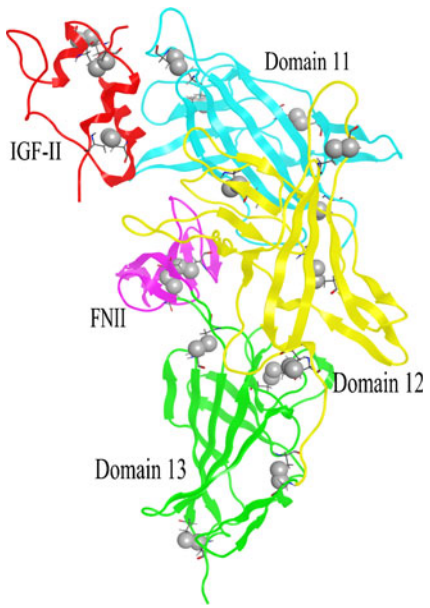



**Fig. 1** Cartoon representation of the "average" structure of the IGF-II/IGF2R complex during a 10 ns MD simulation. Disulfide bonds are shown as *spheres*

In this study, molecular dynamics (MD) simulation was performed on the IGF-II/IGF2R complex in order to describe the features of the protein-protein interaction in much greater detail than can be achieved with just the static crystal structure. Based on the obtained MD trajectory, MM-PBSA and MM-GBSA were applied to calculate the binding free energy and understand the details of the interaction. Furthermore, virtual alanine scanning was used to verify binding hot spots found by pair interaction analysis based on the MM-GBSA method. The detailed interaction profile obtained from the MD simulation and the MM-PBSA/MM-GBSA calculations will help us to understand the critical features determining the IGF-II/IGF2R binding process, and provide some insights that can be utilized in the development of new-generation cancer therapeutic agents exhibiting specificity.

## Material and methods

### Molecular system

From the Protein Data Bank, we obtained the initial structure of the complex of human IGF-II with IGF2R domains 11–13 (PDB ID: 2V5P) [14]. The crystal structure of the complex includes residues 6–63 of IGF-II and residues 1516–1690 of IGF2R domains 11–13 (Fig. 1). The core domain structure is a β-barrel consisting of nine β-strands, termed βA-βI, that forms two crossed β-sheets, the first of which is formed by βA-βD and the second by βE-βI. The N-terminal region preceding βA contains two additional β-strands that form a β-hairpin which caps the β-barrel. Domains 11 and 12 contain four conserved disulfide bonds, while domain 13 has two extra disulfide bonds in the FNII insert (Fig. 1).

## Molecular dynamics simulation

The molecular dynamics simulation was performed with the AMBER 10 software suite [33] and the AMBER ff99SB force field [34]. Fifteen $Na^+$ ions were added to neutralize the system. The system was then solvated using atomistic TIP3P [35] water in a cubic box with a distance of at least 10 Å between the wall of the box and the closest atom of the complex.

The complex was minimized in three consecutive rounds, each of which consisted of 5000 steps (2000 using steepest descent followed by 3000 using the conjugate gradient method), to remove the bad contacts in the crystal structure. Positional restraints were applied to the whole system in the first and second rounds, with force constants of 10 and 2 kcal $mol^{-1}$ $Å^{-2}$, respectively. In the third round the whole system was minimized without restraint. After full relaxation, the system was heated from 0 K to 300 K in 100 ps. Finally, a 10 ns MD simulation was carried out following 500 ps of equilibration at 300 K and 1 atm with the same force constant (2.0 kcal $mol^{-1}$ $Å^{-2}$) on residues 1976–1990 far from the binding surface. Lengths of bonds involving hydrogens were constrained using the SHAKE algorithm [36], and the equations of motion were integrated with a 2 fs time step. The nonbonded cutoff distance was 10 Å ,and the Particle Mesh Ewald (PME) method [37] was used to calculate long-range electrostatics interactions. The temperature of the system was regulated using the Langevin thermostat. All equilibration and subsequent MD stages were carried out in an isothermal isobaric (NPT) ensemble using a Berendsen barostat [38] with a target pressure of 1 bar and a pressure coupling constant of 2.0 ps, recording the coordinate trajectories every 1 ps.

## MM-PBSA/MM-GBSA calculations

Next, we carried out binding free-energy calculations using the conventional MM-PBSA and MM-GBSA approaches. A total of 500 snapshots were extracted every 10 ps from the last 5 ns of the MD trajectory. For each snapshot, the free energy is calculated for each molecular species (complex, IGF-II, and IGF2R), and the binding free energy is computed as the difference

$$\Delta G_{\text{bind}} = G_{\text{complex}} - G_{\text{protein1}} - G_{\text{protein2}}. \quad (1)$$

The free energy, $G$, for each species can be calculated by the following scheme using the MM-PBSA and MM-GBSA methods [26, 27]:

$$G = E_{\text{gas}} + G_{\text{sol}} - TS \quad (2)$$

$$E_{\text{gas}} = E_{\text{int}} + E_{\text{ele}} + E_{\text{vdw}} \quad (3)$$

$$G_{\text{ele,PB(GB)}} = E_{\text{ele}} + G_{\text{PB(GB)}} \quad (4)$$

$$G_{\text{sol}} = G_{\text{PB(GB)}} + G_{\text{sol-np}} \quad (5)$$

$$G_{\text{sol-np}} = \gamma SAS. \quad (6)$$

Here, $E_{\text{gas}}$ is the gas-phase energy; $E_{\text{int}}$ is the internal energy; $E_{\text{ele}}$ and $E_{\text{vdw}}$ are the Coulomb and van der Waals energies, respectively. $E_{\text{gas}}$ was calculated using the ff99SB molecular mechanics force field [34]. $G_{\text{sol}}$ is the solvation free energy and can be decomposed into polar and nonpolar contributions. $G_{\text{PB(GB)}}$ is the polar solvation contribution calculated by solving the PB and GB equations [26, 27]. $G_{\text{ele,PB(GB)}}$ is the polar interaction contribution. Dielectric constants for solute and solvent were set to 1 and 80, respectively. $G_{\text{sol-np}}$ is the nonpolar solvation contribution and was estimated via the solvent-accessible surface area (SAS), which was determined using a water probe radius of 1.4 Å. The surface tension constant $\gamma$ was set to 0.0072 kcal $mol^{-1}$ $Å^{-2}$ [39]. $T$ and $S$ are the temperature and the total solute entropy, respectively. Vibrational entropy contributions can be estimated by either normal mode analysis or the quasi-harmonic approximation. However, it is extremely difficult to handle the Hessian matrix or the covariance fluctuation matrix of an all-atom model system with around 400 residues because it requires a great deal of computing power (our system has 533 residues) and due to the accumulation of round-off errors [31, 40]. Therefore, owing to the prohibitive computational cost, and in accord with previous studies [20, 31], the entropy term was not calculated.

Using the GB model, it was then possible to compute the binding free-energy contribution of each residue at the interface between two interacting proteins [16, 19, 20, 22, 32, 41, 42]. The contribution of a given residue to the binding free energy can be obtained by summing the contribution of each atom of this residue. The separate contributions of backbones and side chains can be determined from the relevant atoms. To obtain a detailed view of the IGF-II/IGF2R interaction, the binding free-energy decomposition was performed using the MM-GBSA method based on the same snapshots as those used in the binding free-energy calculation.

## Computational alanine scanning

Computational alanine scanning consists of replacing a given side chain by an alanine and recalculating the absolute binding free energy for the mutated system. Here, the binding free energy of the alanine mutant was calculated using the MM-GBSA approach [27], and the

difference in the binding free energies of the wild type and alanine, $\Delta\Delta G_{bind}$, was computed as follows:

$$\Delta\Delta G_{bind} = \Delta G_{bind,wild-type} - \Delta G_{bind,mutant}. \quad (7)$$

The binding free energy of the alanine mutant was calculated using the MM-GBSA approach described earlier, based on 500 snapshots extracted from the last 5 ns of the wild-type trajectory every 10 ps. Positive and negative values of $\Delta\Delta G_{bind}$ indicate the unfavorable and favorable contributions, respectively. The alanine mutant structures were generated based on the structures of the collected snapshots by deleting atoms and truncating the mutated residue at Cγ, replacing it with a hydrogen atom. For each frame, the mutant was minimized with the Sander module of Amber before calculating the energy terms. One underlying approximation used in the computational alanine scanning was that the structural perturbation of the system upon such a mutation was small enough that its effect on the binding free energy could be obtained from the trajectory of the wild-type system.

As an extension of the MM-PBSA approach, computational alanine scanning can be used to estimate the contribution of an individual residue to the overall protein–protein binding free energy. The advantage of this method compared with experimental alanine scanning is its high speed, which makes large-scale applications to all protein–protein interfaces with known structures feasible. Computational alanine scanning has been shown to be an effective and reliable method. It can be used as an alternative to experimental mutagenesis in order to identify binding hot spots at protein–protein interfaces [20, 21, 31, 43–46]. Despite its advantages, it cannot be used to add a larger group than those simulated, such as the Gly residue, because of highly unfavorable van der Waals interactions caused by atomic position overlaps. Proline residues were not mutated to alanine in the computational alanine scanning method since they have a severe impact on the secondary structure [30]. In addition, the underlying approximations of the single trajectory mutation protocol are that the mutant and the wild type undergo similar conformational changes to those seen when changing from the unbound to the bound state, and that local side-chain reorganizations are small perturbations relative to the alanine mutation itself, although the conformations of the wild type and the mutant are sometimes quite different.

## Results and discussion

### Structural stabilities from MD simulations

The equilibration of the MD trajectories was monitored based on the convergence of plots of root-mean-square deviations of Cα carbon atoms (Cα-rmsd) at the binding interface from the initial conformation (Fig. 2). As can be seen from Fig. 2, the relative fluctuation in the rmsd of the interface is very small after the initial equilibration. To further test the convergence of the simulation, the time evolutions of various free-energy components for the complex were also monitored. As can be seen from Fig. 3, the fluctuations in the calculated binding free energy obtained using the MM-PBSA and MM-GBSA methods are also very small along the whole trajectory. Although the enthalpy components of Coulomb contributions show large fluctuations between different snapshots, they are compensated for by the polar solvation contribution. As a result, the overall polar interaction ($G_{sol,PB/GB}$) (including the Coulomb contribution and the polar solvation contribution) is stable. From the structural and energetic analysis, we can see that the system is stable during the monitored molecular dynamics simulation.

### Analysis of calculated binding free energy

The calculated binding free energies and their components are presented in Table 1. The mean values of the binding free energies are −62.92 kcal mol$^{-1}$ and −49.11 kcal mol$^{-1}$ based on the MM-PBSA and MM-GBSA methods, respectively. There is therefore a 13.51 kcal mol$^{-1}$ difference between the two methods. This derives from the calculation of the contribution to the polar solvation energy, which is slightly higher in the MM-GBSA calculation (223.36 kcal mol$^{-1}$) than in the MM-PBSA calculation (209.56 kcal mol$^{-1}$).

We now analyze the binding free-energy components in order to gain insights into the molecular driving forces for binding affinity. As can be seen from Table 1, the intermolecular van der Waals interaction and the nonpolar solvation term provide the driving force for binding. The highly favorable nonpolar binding free energy probably

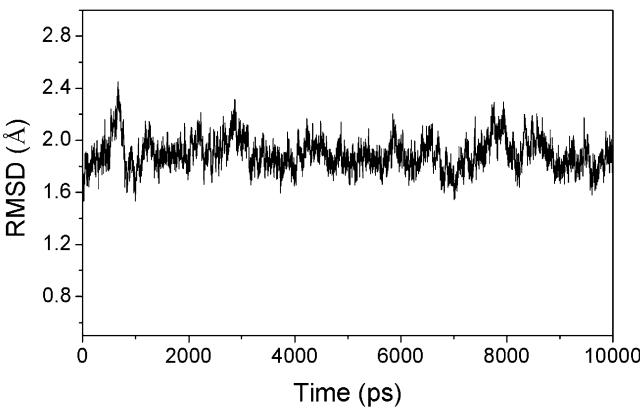



**Fig. 2** Time series of the RMSDs for the CA atoms during 10 ns of MD simulation starting from the initial structure for the complex interface

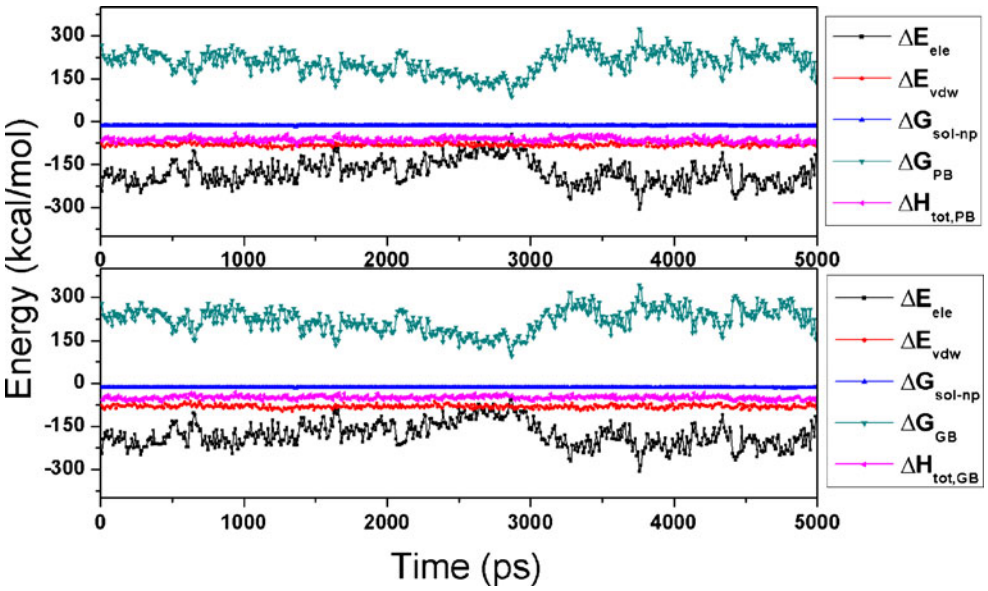



**Fig. 3** Time evolutions of the binding free-energy components for the last 5 ns of the MD simulation of the complex. Figure 3 contains two plots which represent the binding free-energy components calculated by MM-PBSA and MM-GBSA, respectively

accounts for the hydrophobic effects and the solute–water dispersion interaction. The molecular association in a polar medium, the Coulombic interaction $\Delta E_{ele}$, is anti-correlated with the electrostatic desolvation free energy $\Delta G_{PB/GB}$. Although protein–protein Coulombic interactions are always favorable to binding, the contribution from them could not compensate for the large desolvation penalty associated with the binding event. In contrast to the nonpolar component, the electrostatic contribution $\Delta G_{ele,PB/GB}$ (the polar interaction, $\Delta G_{ele,PB/GB}=\Delta E_{ele}+\Delta G_{PB/GB}$) yields an unfavorable contribution. This compensation phenomenon has been previously observed in several studies of protein–protein interactions in solution [40, 47–51].

**Table 1** Binding free energy and its components (kcal mol$^{-1}$) for the IGF-II/IGF2R complex

| Contribution | IGF-II/IGF2R | IGF2R | IGF-II | Delta |
|---|---|---|---|---|
| $E_{ele}$ | −12956.51 | −10757.55 | −2020.96 | −177.99 |
| $E_{vdw}$ | −2039.43 | −1800.34 | −158.03 | −81.06 |
| $E_{int}$ | 11085.96 | 9898.5 | 1187.46 | 0 |
| $E_{gas}$ | −3909.98 | −2659.4 | −991.53 | −259.06 |
| $G_{sol-np}$ | 202.87 | 184.69 | 31.6 | −13.42 |
| $G_{PB}$ | −8620.6 | −7803.36 | −1026.8 | 209.56 |
| $G_{sol,PB}$ | −8417.74 | −7618.67 | −995.2 | 196.14 |
| $G_{ele,PB}$ | −21577.11 | −18560.92 | −3047.76 | 31.57 |
| $H_{tot,PB}$ | −12327.72 | −10278.07 | −1986.73 | −62.92 |
| $G_{sol-np}$ | 202.87 | 184.69 | 31.6 | −13.42 |
| $G_{GB}$ | −8678.77 | −7867.4 | −1034.74 | 223.36 |
| $G_{sol,GB}$ | −8475.91 | −7682.71 | −1003.14 | 209.94 |
| $G_{ele,GB}$ | −21635.28 | −18624.95 | −3055.7 | 45.37 |
| $H_{tot,GB}$ | −12385.89 | −10342.1 | −1994.67 | −49.11 |

## Decomposition of the binding free energy into the per residue contribution

Recently, per residue binding free energies have been widely used to investigate the details of protein–protein interactions at the atomic level, such as in the insulin dimer [18], TCR–p-MHC [20], Ras–Raf, and Ras–RalGDS [16] complexes. The results from these studies demonstrated good correlations between the calculated per residue binding free energy and the experimental binding free-energy differences for the alanine mutants. In order to elucidate the roles of individual amino acid residues in determining the IGF-II/IGF2R association, the binding free energy $\Delta G_{tot,GB}$ was also decomposed into contributions from the residues of the two partners. We identified the hot spot residues that have the greatest impact (per residue contribution >1 kcal mol$^{-1}$) on the protein–protein binding (Fig. 5). The spatial distributions of these hot spots are shown in Fig. 4, and detailed energy information is shown in Table 2.

As shown in Fig. 5, three residues (Glu12, Phe19, Leu53) on the IGF-II surface make large contributions to the binding energy, providing free-energy contributions of >2 kcal mol$^{-1}$. In addition, another three residues (Leu8, Gln18, Ala54) also make considerable contributions to the binding process of IGF-II and IGF2R, each yielding >1 kcal mol$^{-1}$ of free energy. Nine residues in IGF2R domains 11–13, including Tyr1542, Ser1543, Phe1567, Thr1570, Arg1571, Ile1572, Pro1599, Ser1600 and Lys1631, provide free-energy contributions of >2 kcal mol$^{-1}$, and two residues (Leu1629 and Lys1924) yield >1 kcal mol$^{-1}$.

In order to determine the detailed contribution of each important residue, the energy contribution of each important surface residue was further decomposed into several other

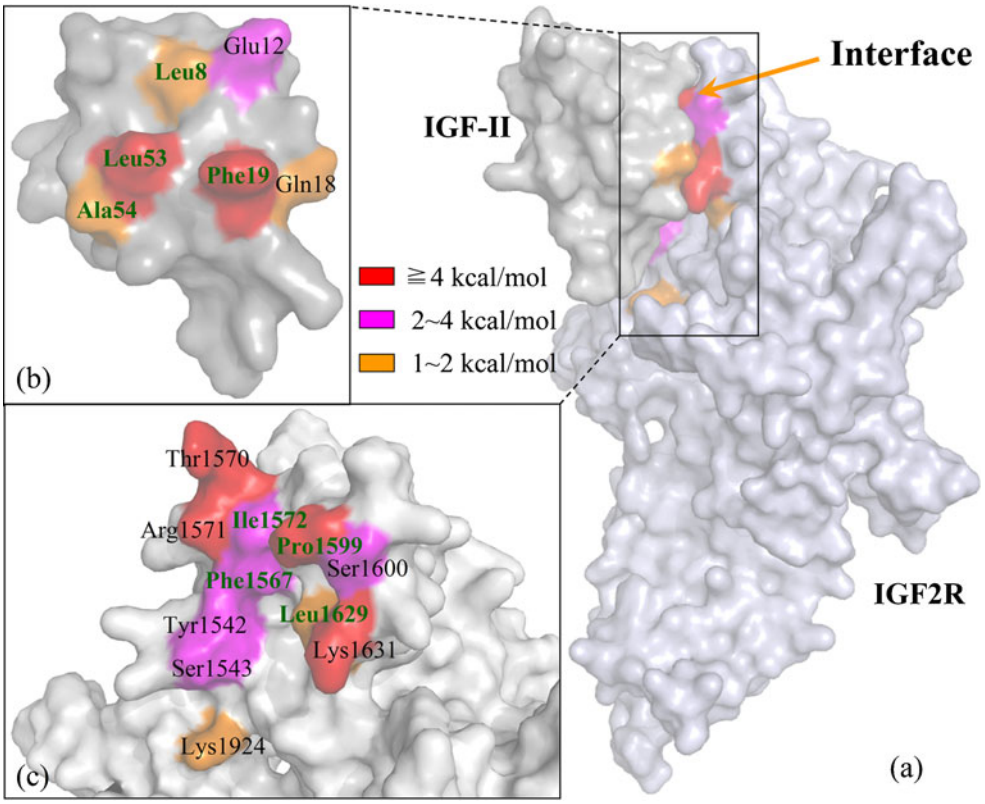



**Fig. 4** **a**–**c** Surface representations of: **a** the IGF-II/IGF2R complex; **b** the key residues of the IGF-II surface; **c** the key residues of the IGF2R surface. The hot spots are colored according to their contributions to IGF-II/IGF2R binding. Hydrophobic residues are shown in *green*

components (Table 2). Our results suggest that most of the important residues for both IGF-II and IGF2R make considerable van der Waals and nonpolar solvation contributions, while the polar interaction ($\Delta G_{i,ele,PB}$) has a relatively small influence, especially for residues from IGF-II (Fig. 6). However, for several residues of IGF2R, polar interactions also clearly contribute; see, for example, Thr1570, Arg1571, Lys1631, and Lys1924. The considerable polar interaction ($\Delta G_{i,sol,GB}$=−7.07 kcal mol$^{-1}$) contributed by Arg1571 is probably due to its unique side chain. Arginine is capable of multiple types of favorable interactions—it forms hydrogen bond complex involving up to five H-bonds, and a

**Table 2** Decomposition of $\Delta G_{bind,GB}$ (kcal mol$^{-1}$) on a per residue basis into contributions from $E_{vdW}$, $E_{ele}$, $G_{sol\text{-}np}$ and $G_{GB}$ (residues: $\Delta G_{i,tot,GB}$>1)

| Contribution | | $\Delta E_{i,vdw}$ | $\Delta E_{i,ele}$ | $\Delta E_{i,gas}$ | $\Delta G_{i,GB}$ | $\Delta G_{i,sol\text{-}np}$ | $\Delta G_{i,sol,GB}$ | $\Delta G_{i,ele,GB}$ | $\Delta H_{i,tot,GB}$ |
|---|---|---|---|---|---|---|---|---|---|
| IGF-II | Leu8 | −1.25 | 1.47 | 0.22 | −1.08 | −0.2 | −1.28 | 0.39 | −1.07 |
| | Glu12 | −2.97 | −29.22 | −32.19 | 29.34 | −0.74 | 28.6 | 0.12 | −3.59 |
| | Gln18 | −3.27 | 3.26 | −0.01 | −1.05 | −0.45 | −1.5 | 2.21 | −1.51 |
| | Phe19 | −6.26 | 0.52 | −5.74 | 2.06 | −1.26 | 0.8 | 2.58 | −4.94 |
| | Leu53 | −6.04 | −0.42 | −6.46 | 1.16 | −0.86 | 0.3 | 0.74 | −6.16 |
| | Ala54 | −1.17 | 1.31 | 0.13 | −1.2 | −0.2 | −1.4 | 0.11 | −1.27 |
| IGF2R | Tyr1542 | −3.32 | −8 | −11.32 | 9.16 | −0.45 | 8.71 | 1.16 | −2.61 |
| | Ser1543 | −2.88 | −1.8 | −4.67 | 3.18 | −0.52 | 2.65 | 1.38 | −2.02 |
| | Phe1567 | −2.31 | −3.33 | −5.64 | 3.69 | −0.28 | 3.41 | 0.36 | −2.23 |
| | Thr1570 | −2.11 | −10.74 | −12.85 | 9.26 | −0.41 | 8.86 | −1.48 | −4 |
| | Arg1571 | −0.69 | −58.07 | −58.76 | 51 | −0.34 | 50.65 | −7.07 | −8.1 |
| | Ile1572 | −2.04 | −0.35 | −2.39 | 0.45 | −0.32 | 0.13 | 0.1 | −2.26 |
| | Pro1599 | −5.13 | −6.1 | −11.23 | 6.69 | −0.96 | 5.73 | 0.59 | −5.51 |
| | Ser1600 | −2.68 | −2 | −4.67 | 2.12 | −0.3 | 1.82 | 0.12 | −2.86 |
| | Leu1629 | −1.98 | 0.78 | −1.2 | −0.18 | −0.4 | −0.59 | 0.6 | −1.79 |
| | Lys1631 | −3.23 | −65.27 | −68.5 | 65.13 | −0.76 | 64.37 | −0.14 | −4.14 |
| | Lys1924 | −0.17 | −43.37 | −43.54 | 42.3 | −0.21 | 42.09 | −1.07 | −1.45 |

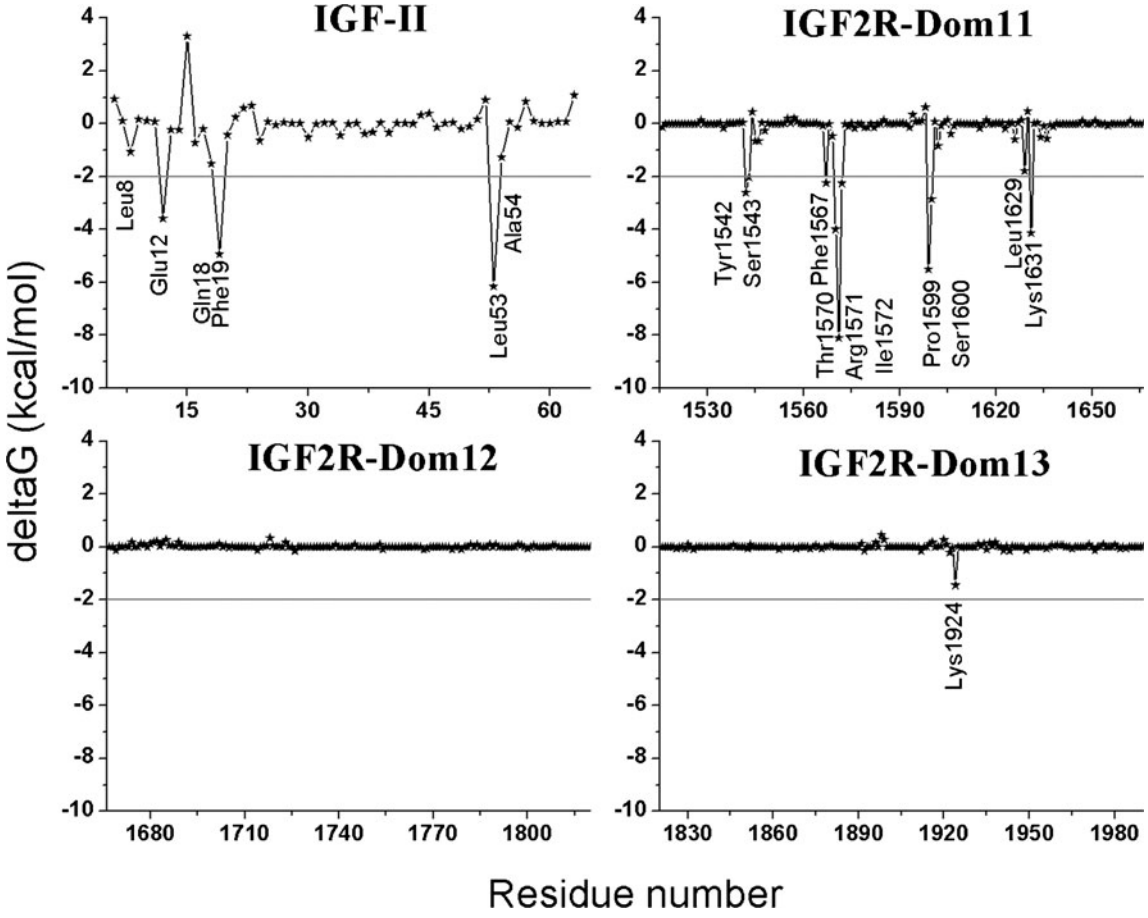



**Fig. 5** Pair interaction energy analysis between IGF-II and IGF2R domains 11–13

salt bridge with the positive charge on its guanidinium motif—so it has a high propensity to be conserved at protein–protein binding sites. Most of these polar interactions originate from the formation of hydrogen bonds, which can be proven by monitoring the probability that these hydrogen bonds exist (their "occupancy") during the MD process (Fig. 7).

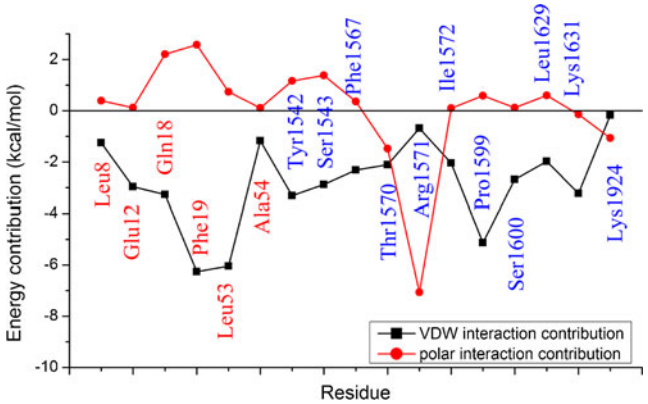



**Fig. 6** The van der Waals and polar interaction contributions for the important residues of IGF-II (labeled in *red*) and IGF2R (labeled in *blue*)

## Computational alanine scanning results

In order to verify the binding hot spots identified by the free-energy decomposition approach, computational alanine scanning was also performed for the residues that contributed more than 1.0 kcal $mol^{-1}$ according to the residue energy decomposition, aside from Pro1599 of IGF2R and Ala54 of IGF-II. This method depends on the assumption that local changes in the protein do not significantly

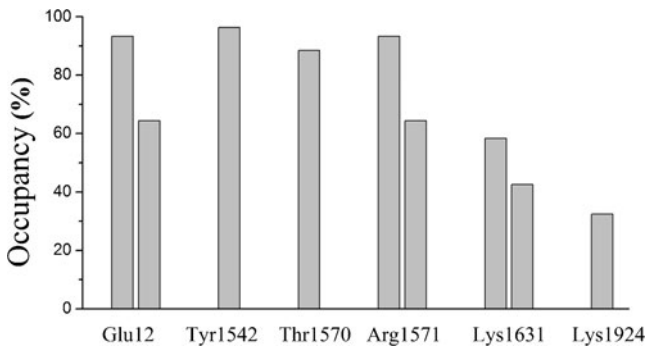



**Fig. 7** Occupancy of intermolecular hydrogen bonds between IGF-II and IGF2R during the molecular dynamics simulation

influence the whole conformation of the complex. The computational alanine scanning results for the selected residues are given in Fig. 6. Because the binding free-energy difference was calculated according to the equation $\Delta\Delta G = \Delta G_{\text{wild-type}} - \Delta G_{\text{mutant}}$, positive and negative values indicate unfavorable and favorable contributions, respectively. By comparing Figs. 5 and 8, we can see that decomposition of the binding free energy and computational alanine scanning give consistent results, which in turn indicates the reliability of our analysis results. The binding affinity drops by more than 2 kcal $mol^{-1}$ when three residues (Glu12, Phe19, Leu53) of IGF-II and eight residues (Tyr1542, Ser1543, Phe1567, Thr1570, Arg1571, Ile1572, Ser1600, Lys1631) of IGF2R are mutated to alanine according to MM-GBSA calculations using both methods. This also proves that these residues are essential for the binding of IGF-II to IGF2R.

## Key features for IGF-II/IGF2R binding

Protein–protein interfaces are frequently hydrophobic and bury a large area of nonpolar surface. Hence, hydrophobic interactions are a leading force in protein–protein interactions, including the interaction of the IGF-II/IGF2R complex. As described for other proteins, an appropriate complement of hydrophobic amino acid residues make significant contributions to the free energy of binding [47]. From the structural information, we can see that the binding interface of IGF-II and IGF2R is full of hydrophobic features (Fig. 4), and there are two hydrophobic pockets. One hydrophobic pocket consists of Tyr1542, Ser1543, Phe1567 and Leu1629, and this is responsible for the binding of Phe19 from IGF-II. The second hydrophobic pocket, formed by Pro1599, Ser1600, Leu1629 and Lys1631, plays an important role in the binding of Leu53 from IGF-II. Thus, for IGF-II, Phe19 and Leu53 are the anchor residues of the interaction. Since the IGF-II/IGF2R binding pocket is full of hydrophobic characteristics and is composed of many hydrophobic residues, the dominance of nonpolar interactions in the binding is not surprising.

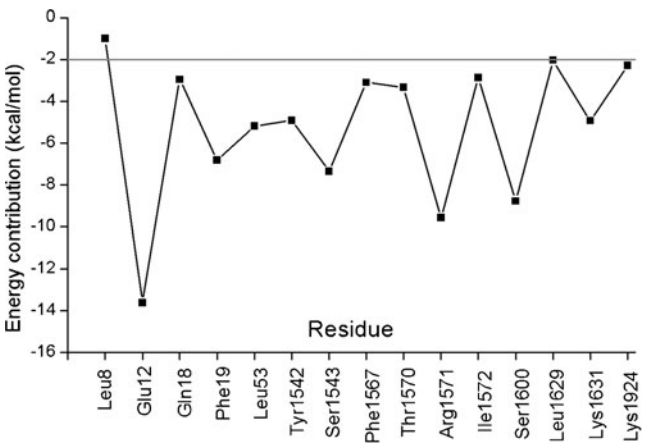



**Fig. 8** Computational alanine scanning results for IGF-II and IGF2R interaction residues

Electrostatic forces are also important in protein–protein interactions because electrostatic complementarity of the interacting protein surfaces promotes complex formation and defines the lifetime of the complex. By monitoring the MD trajectory, we obtained information on the hydrogen bonds between IGF-II and IGF2R (Fig. 7). It can be seen that there are many intermolecular hydrogen bonds. In addition, based on the structural analysis, we can also presume that the formation of these hydrogen bonds means that most of these residues make obvious contributions to the binding process.

Based on the consistent information from free-energy decomposition and computational alanine scanning, three residues (Glu12, Phe19 and Leu53) of IGF-II were found to make large contributions (per residue contributions of >2 kcal $mol^{-1}$) to the binding process, and eight residues of IGF2R (Tyr1542, Ser1543, Phe1567, Thr1570, Arg1571, Ile1572, Ser1600, Lys1631) were also found to do so. Most of these residues are hydrophobic, reiterating the importance of hydrophobic interactions. At the same time, the results of our work are consistent with those of previous mutagenesis studies.

Mutagenesis studies of IGF-II [48] demonstrated that Phe19 and Leu53 are critical for IGF2R binding. In addition, there are 4- and 3.3-fold decreases in binding affinity for Ala54Arg and Leu55Arg, respectively [49]. Both Ala54 and Leu55 are close to the crucial Leu53 side chain, but our results show that Leu55 does not make an obvious contribution. Neither the Thr58Met IGF-II nor the Tyr27Leu IGF-II mutants differed greatly from IGF-II in their IGF2R binding affinities [14]. Single alanine replacement [14] of Phe48, Arg49, Ser50, or Glu57 demonstrates that they are not crucial residues for IGF2R binding, which contradicts previous reports [50, 51]. As can be seen in Fig. 5, the present study also shows that IGF-II residues such as Tyr27, Phe48, Arg49, Ser50, Glu57 and Thr58 do not make favorable contributions to the binding.

The binding hot spots Tyr1542, Ser1543 (AB loop), Phe1567, Thr1570, and Ile1572 (CD loop), as found by experimental mutagenic studies [52], were also identified in our present study. The mutagenic studies [52] also suggested that residues within the FG loop make only minor contributions to the overall affinity, while two proline residues (Pro1597, Pro1599) were not considered because mutations of them to alanine would likely cause structural perturbations of the carbon backbone. However, two proline residues located in the FG loop may contribute to the hydrophobic core formed by Ile1572 and Phe1567. With this experimental result in mind, it is worth noting that

Pro1599 makes a large contribution of 5.51 kcal $mol^{-1}$ while Pro1597 does not have an obvious effect on the binding according to our calculations (Fig. 5).

The importance of IGF2R residues Tyr1542, Thr1570, Arg1571, Lys1631, Pro1599, Lys1631, and Lys1924 can also be demonstrated by monitoring the occupation probabilities of these hydrogen bonds during the MD process. In the binding free-energy decomposition, it was found that Lys1924 (located in FNII) contributes 1.45 kcal $mol^{-1}$. Despite its importance in high-affinity IGF-II binding, some previous research has proposed that FNII contributes indirectly to the formation of the protein–protein complex. However, our studies show that the occupation of the hydrogen bond between Lys1924 (IGF2R domain 13) and Asp23 (IGF-II) is 32.4%. In an experimental study by Zaccheo et al. [52], kinetic analysis showed that $\Delta G$ for wild-type domain 11 interacting with IGF-II was ~9.5 kcal $mol^{-1}$, while Brown et al. [14] and Linnell et al. [53] reported that $\Delta G$ values for the binding of domains 11–13 and IGF-II were ~11 and ~9.94 kcal $mol^{-1}$. Although neglecting the contribution from entropy makes our calculated binding energy much lower than the experimental value, the high correspondence between the hot spots identified by computational free-energy decomposition and experimental alanine scanning of either IGF-Dom11 [52] or IGF-II [14] shows the reliability of our calculated results.

## Conclusions

In this study, MM-PBSA/GBSA calculations, binding free-energy decomposition, and computational alanine scanning were performed to characterize the interaction of IGF-II and IGF2R based on a 10 ns molecular dynamics simulation. The results from the MM-PBSA and MM-GBSA calculations indicate that the intermolecular van der Waals interaction and the nonpolar solvation term provide the driving forces for the binding, while polar interactions do not contribute to any significant degree. In conclusion, three residues (Glu12, Phe19 and Leu53) of IGF-II and eight residues (Tyr1542, Ser1543, Phe1567, Thr1570, Arg1571, Ile1572, Ser1600, Lys1631) of IGF2R make considerable contributions to the IGF-II/IGF2R binding. Although Pro1599 of IGF2R was not explored in the computational alanine scanning, it can also be considered a key residue, with an energy contribution of 5.51 kcal $mol^{-1}$. As IGF-II has been implicated in the growth of various tumors, the details of the interaction between IGF-II and its receptor (IGF2R) obtained in this study could be used in the development of molecular probes to assist in diagnostic screening, and even in novel approaches to control tumor development.

**Acknowledgments** This work was supported by the Program for New Century Excellent Talents in University (grant no. NCET-07-0399) and the National Natural Science Foundation of China (grant no. 20905033). The authors also wish to express their thanks to Zhide Foundation of Lanzhou University for supporting this work.

## References

1. LeRoith D, Roberts CTJ (2003) The insulin-like growth factor system and cancer. Cancer Lett 195:127–137
2. Denley A, Cosgrove LJ, Booker GW, Wallace JC, Forbes BE (2005) Molecular interactions of the IGF system. Cytokine Growth Factor Rev 16:421–439
3. Macaulay VM (1992) Insulin-like growth factors and cancer. Br J Cancer 65:311–320
4. Zhang XH, Yee D (2000) Tyrosine kinase signalling in breast cancer: insulin-like growth factors and their receptors in breast cancer. Breast Cancer Res 2:170–175
5. Fürstenberger G, Senn HJ (2002) Insulin-like growth factors and cancer. Lancet Oncol 3:298–302
6. Gallagher EJ, LeRoith D (2010) The proliferating role of insulin and insulin-like growth factors in cancer. Trends Endocrinol Metab 21:610–618
7. Frasca F, Pandini G, Scalia P, Sciacca L, Mineo R, Costantino A, Goldfine ID, Belfiore A, Vigneri R (1999) Insulin receptor isoform A, a newly recognized, high-affinity insulin-like growth factor II receptor in fetal and cancer cells. Mol Cell Biol 19:3278–3288
8. Dahms NM, Hancock MK (2002) P-type lectins. Biochim Biophys Acta 1572:317–340
9. El-Shewy HM, Luttrell LM (2009) insulin-like growth factor-2/mannose-6 phosphate receptors. Vitam Horm 80:667–697
10. Amritraj A, Rauw G, Baker GB, Kar S (2010) Leu27 insulin-like growth factor-II, an insulin-like growth factor-II analog, attenuates depolarization-evoked GABA release from adult rat hippocampal and cortical slices. Neuroscience 170:722–730
11. Martin-Kleiner I, Troselj KG (2010) Mannose-6-phosphate/insulin-like growth factor 2 receptor (M6P/IGF2R) in carcinogenesis. Cancer Lett 289:11–22
12. Yadavalli S, Kumar AP, Nadimpalli SK (2009) Reptilian MPR 300 is also the IGF-IIR: cloning, sequencing and functional characterization of the IGF-II binding domain. Int J Biol Macromol 44:435–440
13. Brown J, Jones EY, Forbes BE (2009) Interactions of IGF-II with the IGF2R/cation-independent mannose-6-phosphate receptor: mechanism and biological outcomes. Vitam Horm 80:699–719
14. Brown J, Delaine C, Zaccheo OJ, Siebold C, Gilbert RJ, Boxel GV, Denley A, Wallace JC, Hassan AB, Forbes BE, Jones EY (2008) Structure and functional analysis of the IGF-II/IGF2R interaction. EMBO J 27:265–276
15. Stites WE (1997) Protein–protein interactions: interface structure, binding thermodynamics, and mutational analysis. Chem Rev 97:1233–1250
16. Gohlke H, Kiel C, Case DA (2003) Insights into protein–protein binding by binding free energy calculation and free energy decomposition for the Ras–Raf and Ras–RalGDS complexes. J Mol Biol 330:891–913
17. Gohlke H, Case DA (2004) Converging free energy estimates: MM-PB(GB)SA studies on the protein–protein complex Ras–Raf. J Comput Chem 25:238–250
18. Zoete V, Meuwly M, Karplus M (2005) Study of the insulin dimerization: binding free energy calculations and per-residue free energy decomposition. Proteins 61:79–93
19. Michalik L, Zoete V, Krey G, Grosdidier A, Gelman L, Chodanowski P, Feige JN, Desvergne B, Wahli W, Michielin O

(2007) Combined simulation and mutagenesis analyses reveal the involvement of key residues for peroxisome proliferator-activated receptor α helix 12 dynamic behavior. J Biol Chem 282:9666–9677
20. Zoete V, Michielin O (2007) Comparison between computational alanine scanning and per-residue binding free energy decomposition for protein–protein association using MM-GBSA: application to the TCR–p-MHC complex. Proteins 67:1026–1047
21. Cui QZ, Sulea T, Schrag JD, Munger C, Hung MN, Naïm M, Cygler M, Purisima EO (2008) Molecular dynamics–solvated interaction energy studies of protein–protein interactions: the MP1–p14 scaffolding complex. J Mol Biol 379:787–802
22. Lafont V, Schaefer M, Stote RH, Altschuh D, Dejaegere A (2007) Protein–protein recognition and interaction hot spots in an antigen–antibody complex: free energy decomposition identifies "efficient amino acids." Proteins 67:418–434
23. Beveridge DL, DiCapua FM (1989) Free energy via molecular simulation: applications to chemical and biomolecular systems. Annu Rev Biophys Biophys Chem 18:431–492
24. Jorgensen WL (1989) Free energy calculations: a breakthrough for modeling organic chemistry in solution. Acc Chem Res 22:184–189
25. Kollman P (1993) Free energy calculations: applications to chemical and biochemical phenomena. Chem Rev 93:2395–2417
26. Srinivasan J, Cheatham TE III, Cieplak P, Kollman PA, Case DA (1998) Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate-DNA helices. J Am Chem Soc 120:9401–9409
27. Kollman PA, Massova I, Reyes C, Kuhn B, Huo S, Chong L, Lee M, Lee T, Duan Y, Wang W (2000) Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. Acc Chem Res 33:889–897
28. Onufriev A, Bashford D, Case DA (2000) Modification of the generalized Born model suitable for macromolecules. J Phys Chem B 104:3712–3720
29. Tsui V, Case DA (2000) Theory and applications of the generalized Born solvation model in macromolecular simulations. Biopolymers 56:275–291
30. Massova I, Kollman PA (1999) Computational alanine scanning to probe protein–protein interactions: a novel approach to evaluate binding free energies. J Am Chem Soc 121:8133–8143
31. Huo SH, Massova I, Kollman PA (2002) Computational alanine scanning of the 1:1 human growth hormone-receptor complex. J Comput Chem 23:15–27
32. Li T, Froeyen M, Herdewijn P (2008) Computational alanine scanning and free energy decomposition for *E. coli* type I signal peptidase with lipopeptide inhibitor complex. J Mol Graph Model 26:813–823
33. Case D, Darden T, Cheatham T III, Simmerling C, Wang J, Duke R, Luo R, Crowley M, Walker R, Zhang W (2008) AMBER 10. University of California, San Francisco
34. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C (2006) Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins 65:712–725
35. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926–935
36. Ryckaert JP, Ciccotti G, Berendsen HJC (1977) Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of *n*-alkanes. J Comput Phys 23:327–341
37. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG (1995) A smooth particle mesh Ewald method. J Chem Phys 103:8577–8593
38. Berendsen HJC, Postma JPM, Van Gunsteren WF, DiNola A, Haak JR (1984) Molecular dynamics with coupling to an external bath. J Chem Phys 81:3684–3690
39. Sitkoff D, Sharp KA, Honig B (1994) Accurate calculation of hydration free energies using macroscopic solvent models. J Phys Chem 98:1978–1988
40. Brooks BR, Jane D, Karplus M (1995) Harmonic analysis of large systems. I. Methodology. J Comput Chem 16:1522–1542
41. Deng NJ, Cieplak P (2009) Insights into affinity and specificity in the complexes of α-lytic protease and its inhibitor proteins: binding free energy from molecular dynamics simulation. Phys Chem Chem Phys 11:4968–4981
42. Zoete V, Irving MB, Michielin O (2010) MM-GBSA binding free energy decomposition and T cell receptor engineering. J Mol Recognit 23:142–152
43. Kortemme T, Baker D (2002) A simple physical model for binding energy hot spots in protein–protein complexes. Proc Natl Acad Sci USA 99:14116–14121
44. Kortemme T, Kim DE, Baker D (2004) Computational alanine scanning of protein–protein interfaces. Sci STKE 2004:1–8
45. Zoete V, Meuwly M (2006) Importance of individual side chains for the stability of a protein fold: computational alanine scanning of the insulin monomer. J Comput Chem 27:1843–1857
46. Liu HX, Yao XJ (2010) Molecular basis of the interaction for an essential subunit PA-PB1 in influenza virus RNA polymerase: insights from molecular dynamics simulation and free energy calculation. Mol Pharm 7:75–85
47. Li YL, Huang YP, Swaminathan CP, Smith-Gill SJ, Mariuzza RA (2005) Magnitude of the hydrophobic effect at central versus peripheral sites in protein–protein interfaces. Structure 13:297–307
48. Delaine C, Alvino CL, McNeil KA, Mulhern TD, Gauguin L, De Meyts P, Jones EY, Brown J, Wallace JC, Forbes BE (2007) A novel binding site for the human insulin-like growth factor-II (IGF-II)/mannose 6-phosphate receptor on IGF-II. J Biol Chem 282:18886–18894
49. Forbes BE, McNeil KA, Scott CD, Surinya KH, Cosgrove LJ, Wallace JC (2001) Contribution of residues A54 and L55 of the human insulin-like growth factor-II (IGF-II) A domain to type 2 IGF receptor binding specificity. Growth Factors 19:163–173
50. Burgisser DM, Roth BV, Giger R, Luthi C, Weigl S, Zarn J, Humbel RE (1991) Mutants of human insulin-like growth factor II with altered affinities for the type 1 and type 2 insulin-like growth factor receptor. J Biol Chem 266:1029–1033
51. Sakano K, Enjoh T, Numata F, Fujiwara H, Marumoto Y, Higashihashi N, Sato Y, Perdue JF, Fujita-Yamaguchi Y (1991) The design, expression, and characterization of human insulin-like growth factor II (IGF-II) mutants specific for either the IGF-II/cation-independent mannose 6-phosphate receptor or IGF-I receptor. J Biol Chem 266:20626–20635
52. Zaccheo OJ, Prince SN, Miller DM, Williams C, Kemp CF, Brown J, Jones EY, Catto LE, Crump MP, Hassan AB (2006) Kinetics of insulin-like growth factor II (IGF-II) interaction with domain 11 of the human IGF-II/mannose 6-phosphate receptor: function of CD and AB loop solvent-exposed residues. J Mol Biol 359:403–421
53. Linnell J, Groeger G, Hassan AB (2001) Real time kinetics of insulin-like growth factor II (IGF-II) interaction with the IGF-II/mannose 6-phosphate receptor: the effects of domain 13 and pH. J Biol Chem 276:23986–23991

J Mol Model (2012) 18:1431–1445
DOI 10.1007/s00894-011-1169-2

ORIGINAL PAPER

# Mechanism of isoproturon resistance in *Phalaris minor:* in silico design, synthesis and testing of some novel herbicides for regaining sensitivity

**Durg Vijay Singh · Kuruba Adeppa · Krishna Misra**

Received: 25 February 2011 / Accepted: 28 June 2011 / Published online: 15 July 2011


**Abstract** Isoproturon, 3-p-cumenyl-1 dimethylurea was the only herbicide controlling *Phalaris minor*, a major weed growing in wheat fields till the early 1980s. Since it has acquired resistance against isoproturon, like other substituted urea herbicides, where the identified target site for isoproturon is in the photosynthetic apparatus at D1 protein of Photosystem-II (PS-II). Nucleotide sequence of susceptible and resistant *psbA* gene of *P. minor* has been reported to have four point mutations. During the present work D1 protein of both susceptible and resistant biotypes of *P Minor* has been modeled. Transmembrane segments of amino acids were predicted by comparing with the nearest homolog of bacterial D1 protein. Volume and area of active site of both susceptible and resistant biotypes has been simulated. Isoproturon was docked at the active site of both, susceptible and resistant D1 proteins. Modeling and simulation of resistance D1 protein indicates that the resistance is due to alteration in secondary structure near the binding site, resulting in loss in cavity area, volume and change in binding position, loss of hydrogen bonds, hydrophobic interaction and complete loss of hydrophobic sites. To regain sensitivity in resistant biotype new derivatives of isoproturon molecules have been proposed, synthesized and tested. Among the 17 derivatives we found that the N-methyl triazole substituted isoproturon is a potential substitute for isoproturon.

**Keywords** D1 protein · Isoproturon · Modeling · *Phalaris minor* · Pharmacophore · Pot assay · Simulation

**Electronic supplementary material** The online version of this article (doi:10.1007/s00894-011-1169-2) contains supplementary material, which is available to authorized users.

D. V. Singh
Indian Institute of Information Technology,
Deoghat, Jhalwa,
Allahabad 211012, India

K. Adeppa
Uttar Pradesh Technical University,
Lucknow 211026, India

K. Misra (✉)
Center of Biomedical Magnetic Resonance SGPGIMS Campus,
Raebareli Road,
Lucknow, India
e-mail: kmisra@iiita.ac.in
e-mail: krishnamisra@hotmail.com

## Introduction

Isoproturon resistance in *P.minor* is still being perceived as most serious issue by the wheat growers in Indo-gangetic plains of Pakistan, India, Nepal and Bangladesh. In India an area of 10.5 million hectares of land contributes 40% to the food grain production. Among this resistance affected area ranges between 0.8 and 1.0 mha [1]. The herbicide isoproturon was recommended for *P.minor* control in late 1970s and continued to work well for nearly 20 years. Incidence of resistance was first reported by Malik in 1995 [2]. Thereafter laboratory, pot, and field experiments were conducted to quantify level of isoproturon resistance in various biotypes. The recommendation of isoproturon was withdrawn in 1998. However, three new alternate herbicides *viz* fenoxaprop, clodinafop, and sulfosulfuron were introduced in 1997–98 but resistance against these alternate herbicides has been reported in few biotypes [1, 2]. As happens with most of herbicides $G_{50}$ (dose required for 50% growth reduction) consistently increases year by year, like isoproturon this was also true for fenoxaprop, clodinafop, sulfosulfuron and tralkoxydim [2]. Resistance problem is increasing day by day and bringing havoc for wheat

growers since it causes 65-70% loss in yield [3, 4]. Some of *P. minor* biotypes have evolved multiple resistances through three modes of action viz photosynthesis at PS-II, acetyl-CoA carboxylase (ACCase) and acetolactate synthease ALS inhibitor [5].For a developing nation like India food security is a big concern. Hence some strategies are needed to address the above problem urgently.

Isoproturon (trade name Alon, Arelon or Gramion), chemically known as 3-p cumenyl-1, 1-dimethyl urea ($C_{12}H_{18}N_2O$) has molecular weight 206.3. Its solubility in water is only 266 μM at 20 °C, it is a nonpolar hydrophobic molecule and may freely diffuse through the lipid barriers. Isoproturon was largely accepted at 0.75 to 1.5 kg/ha by the farmers. The resistant biotype of *P. minor* grows well at 1.5 kg/ha isoproturon but the susceptible biotype does not survive beyond 45 days of growth. Moreover, the resistance is maintained through generations.

It has been established that the incidence of resistance of isoproturon is the result of over 10–12 years of continuous use. Herbicide resistance is a heritable trait and therefore is difficult to control its spread [6]. Thus, it is essential to address the problem and overcome isoproturon resistance in *P. minor*.

Three possible mechanisms of herbicide resistance can be visualized: (i) modification of target site (ii) detoxification and (iii) reduced uptake [7]. In accordance with the past literature [8–10] it can be stated that mode of action of isoproturon is most likely to be similar to that of other substituted ureas, where the identified target site is in photosynthetic apparatus, especially D1 protein of PS-II. The nucleotide sequence of resistant *psbA* gene encode D1 protein has shown four point mutations [11, 12]. This clearly shows that the *P. minor* populations have developed resistance to isoproturon due to modification of the target site. Isoproturon treatment to these biotype do not affect cytochrome P450 level, however polymorphism has been reported in both biotypes [13]. Hence resistance due to metabolic detoxification and reduced absorption and uptake can be ruled out.

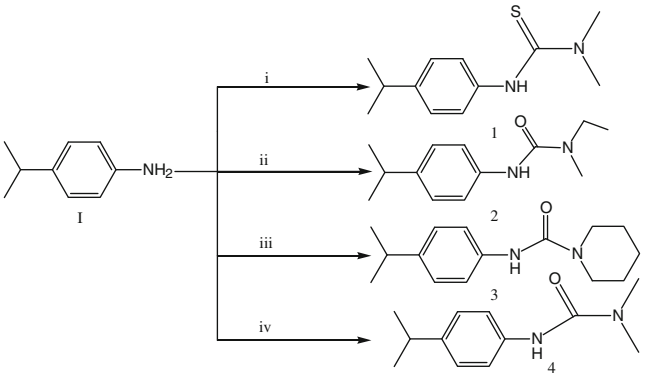
The D1 protein, also referred as the QB binding protein (32 kDa), encoded by a chloroplast *psbA* gene is highly conserved from cyanobacteria to higher plants [14] and has homologies [15–19]. Modification of the QB binding site reduces herbicide affinity to displace plastoquinone and electron flow from QA to QB, thus conferring resistance to urea herbicides [11, 12]. Single site mutation in the *psbA* gene made it herbicide tolerant plant [13]. Single mutation causing a change of either Val-219 to Ile, Gly 256 to Asp, Ser 264 to Thr or Ala, Leu 275 to Phe or double mutations causing changes from Phe 255 to Tyr and Ser 264 to Ala resulted in DCMU (dichlorophenol dimethyl urea also known as diuron) resistance in purple bacteria [20]. Mutation (A117T, P192A, P192S, P192E, P192L, A200V, and W586L) in acetolactate synthase (also known as acetohydroxy acid synthase) has developed several fold resistance against chlorsulfuron, flumetsulam and bispyribac herbicides. Feng et al. have computationally developed conformationaly flexible inhibitors against W586L mutant acetolactate synthase to overcome drug resistance [21].

Point mutation in D1 protein might be cause of resistance in *P. minor* therefore in this study D1 protein of both biotypes, susceptible and resistant, has been modeled, binding site predicted, docked in the cavity of mutated site and receptor based pharmacophore modeling was carried out to explain the cause of resistance. To regain sensitivity structure based isoproturon derivatives were designed synthesized and tested.

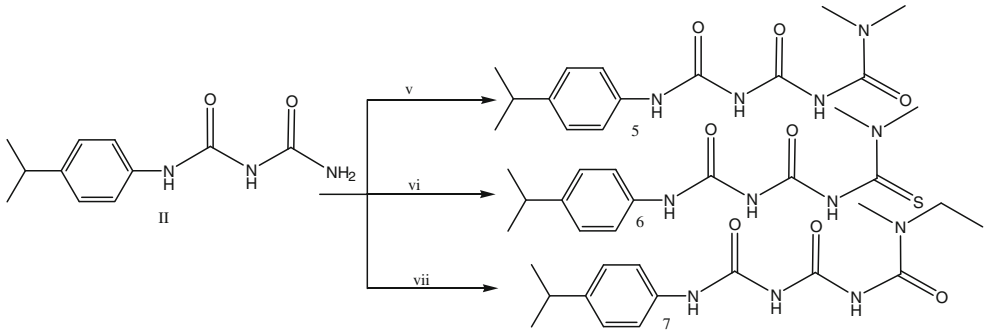
## Methods and materials

### Bioinformatics and computational simulation

Both susceptible [Q85V85|Q85V85_9POAL *psbA*] and resistant D1 proteins [Q7YMJ5|Q7YMJ5_9POA *psbA*] were taken from UniProtKB/TrEMBL database and modeled using homology modeling software modeler [22, 23]. Homology modeling includes 1; template recognition and initial alignment to predict percent identity using BLAST-P (Basic local alignment search tools for protein) from protein structure database. It uses basically two matrices (a) a residue exchange matrix and (b) an alignment matrix to find best score for target sequence 2; backbone generation, it copies the coordinates from template residues that have shown good alignment with target sequence 3; loop modeling: to fill gaps either due to deletion in target or insertion in template




**Scheme 1** Synthesis of N,N-Dimethylthiocarbamoyl (1), N,N-Ethyl-methylcarbamoyl (2), piperidine carbamoyl (3) and N,N-Dimethylcarbamoyl, i.e., isoprutoron (4) derivatives of paracumidine (I) [35, 36]. Reagents & conditions: N,N-dimethylthiocarbamoyl chloride (i), N,N-Ethylmethylcarbamoylchloride(ii), Piperidinyl carbamoyl chloride(iii) and N,N-dimethylcarbamoyl chloride(iv) /triethylamine,20-25 °C,10 h



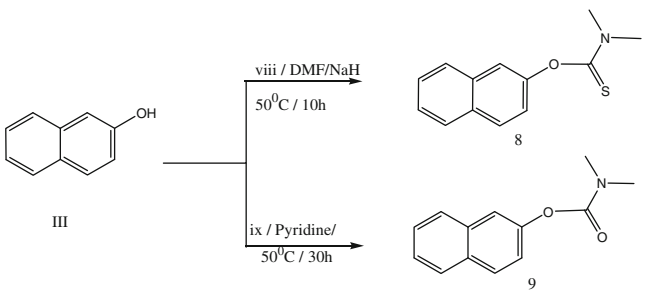

**Scheme 2** Synthesis of N,N-Dimethylcarbamoyl (5),N, N-Dimethylthiocarbamoyl (6) and N,N-Ethyl methyl carbamoyl (7) derivatives of para-cumidinyl biuret [35, 36] ; Reagents & conditions: N,N-dimethylcarbamoyl chloride(v), N-Dimethylthiocarbamoyl(vi) and N,N-Ethylmethylcarbamoylchloride(vii), 40 °C, 12 h

sequences. Side chains were modeled using rotamer library; obtained model was optimized and validated [24].

It is well known that active site prediction in transmembrane protein is tedious; however the topology of D1 protein had to be worked out. This D1 protein has the domain of photoreaction center of PS-II. Since D1 proteins are conserved from cynaobacteria to higher plants, therefore one of them could be used for comparative study for analysis of its transmembrane segments topology.

Since D1 protein is a multipass transmembrane protein and mutated residues occur at the thylakoid lumen therefore we considered all those cavities, which hold mutated amino acids of susceptible D1 protein. Docking was performed using version 3 of the Auto-Dock [25, 26]. The modeled structures of both susceptible as well as resistant D1 proteins were set for docking.

The docking study was performed with isoproturon phenyl urea herbicide at the active site of susceptible D1 protein. Amino acids at 24 and 29 positions were considered as the center of the grid map in both variants of protein. The structures of susceptible and resistant D1 proteins were set up for docking as described for AutoDock utility [26]: The grid maps representing the protein in the actual docking process were calculated with AutoGrid. The grids were chosen to be sufficiently large to include not only the active site residues but also significant portions of the surrounding surface of active sites in both simulations. The dimensions of the grids




**Scheme 3** Synthesis of N,N-Dimethylthiocarbamoyl (8), and N,N-Dimethyl carbamoyl (9) derivatives of 2-Naphthol [36, 37]. Reagents & conditions: N,N-dimethylthiocarbamoyl chloride(viii) and N,N-Dimethyl carbamoyl(ix)

for susceptible D1 protein were 36:36:36 Å, with a spacing of 0.375 Å between the grid points and the center close to the Gly24 and Val29 residues. Similarly the dimensions of the grids for resistant D1 protein were 36:36:36 A°, with a spacing of 0.375 A° between the grid points and the center close to the mutated residues were taken. AutoDock 3 with a genetic search algorithm was chosen for all dockings [27].

The optimized AutoDock run parameters are as follows: the maximum number of energy evaluations, the maximum number of generation in the genetic algorithm and the number of GA per run were 2500000, 2700000 and 50 respectively. All other run parameters were maintained at their default settings. As far as the search methods are concerned, only genetic algorithm was used [28–30]

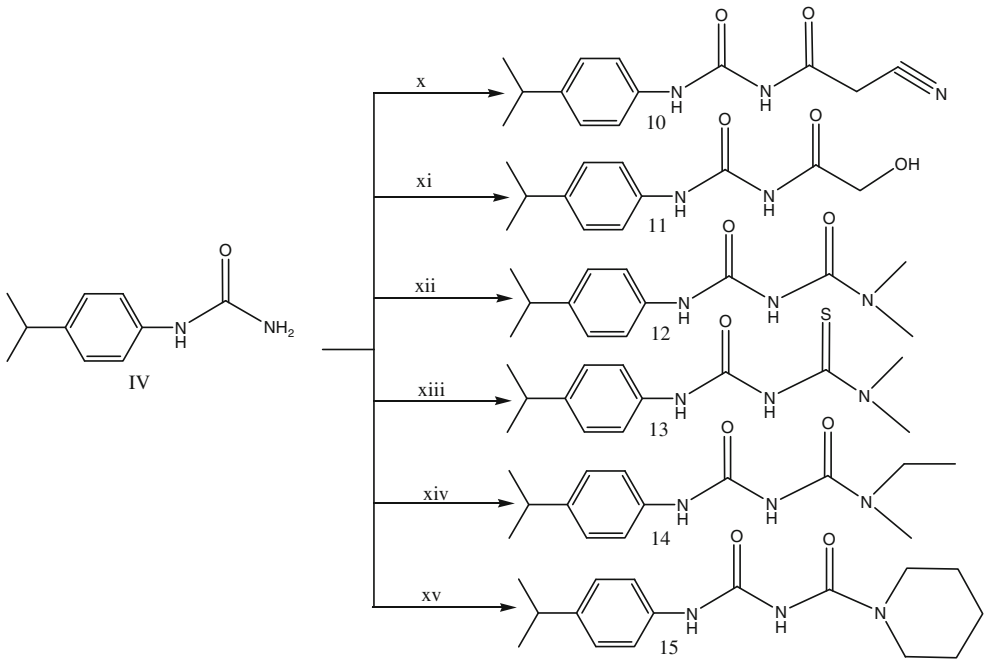
Isoproturon was docked to the active site of both susceptible and resistant D1 protein. Structure of isoproturon were generated by drawing the 2D on Chemsketch [31] and 3-D structures were obtained from CORINA server and optimized with Insight II program (Accelrys Inc.). Atomic charges were added using calibration given in the AutoDock empirical free energy function [29]. Isoproturon was set up for docking with the help of Autotors (root, and no. of rotatable bonds) and all single bonds defined as torsion during the docking process.

To find out hydrogen bond and hydrophobic interaction Ligplot programme [32] was used to simulate hydrogen bonds and nonbonded contacts. HBPLUS program [33] is used by Ligplot. Ligbuilder [34] program was used to design new potential ligand molecules (urea derivatives) as well as for lead optimization, i.e., to improve bioactivity by precalculated key site of interaction and Pharmacophore of the receptor.

## Chemical synthesis

### *I. General procedure for the preparation of para cumidine derivatives (1–4, Scheme 1)*

Para cumidine (18.7 mM) and triethylamine (20 mM) were dissolved in tetrahydrofuran (50 ml) and different carbamoylchloride derivatives [ N,N-dimethylthio carbamoyl chloride (i), N,N-ethylmethyl carbamoylchloride (ii), piper-




**Scheme 4** Synthesis of Cyanoacetyl (10), Glycolyl (11), N,N-Dimethyl carbamoyl (12), N,N-Dimethylthio carbamoyl (13), N, N-Ethyl-methylcarbamoyl (14) and piperidinecarbamoyl (15) derivatives of paracumidyl urea [35, 36]. Reagents & conditions: Cyanoacetic acid (x ), Glycolic acid (xi), N,N-dimethylcarbamoyl chloride (xii), N,N-Dimethyl thiocarbamoyl chloride (xiii) N,N-Ethylmethylcarbamoylchloride (xiv) and, Piperidinyl carbamoylchloride (xv)/ 70 °C, 8 h

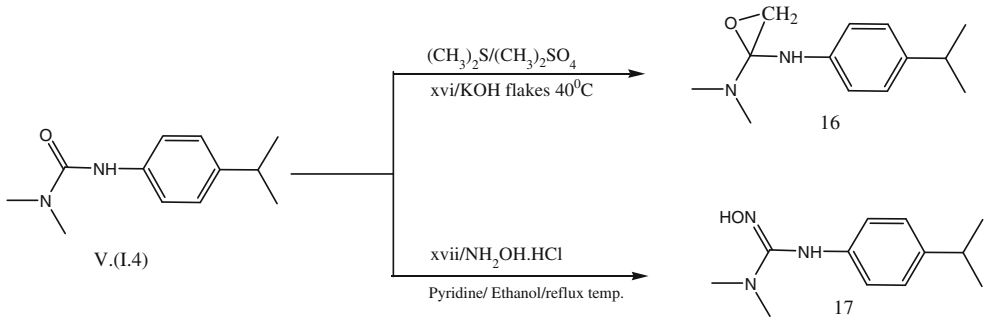
idinyl carbamoyl chloride (iii) and N,N-dimethyl carbamoyl chloride (iv)] were added to the solution. The solution was maintained at 20-25 °C for 10 h, and the reaction mixture was filtered to separate triethylamine hydrochloride. Filtrate was distilled under vacuo, residue was dissolved in ethylacetate and washed with water. Organic portion was collected, dried over sodium sulfate and concentrated under vacuo. Crude product was further purified by column chromatography.

*II. General procedure for the preparation of cumidylbiuret derivatives (5–7, Scheme 2)*

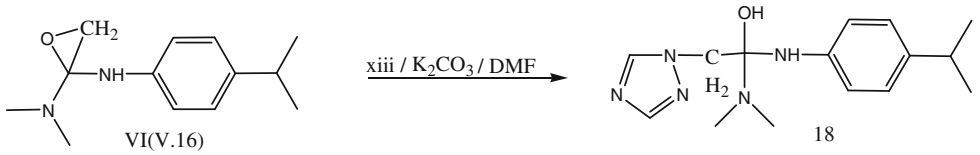
Cumidinyl biuret (13.6 mM) was dissolved in dichloromethane (50 ml) and different carbamoylchloride derivatives (N,N-Dimethyl carbamoyl chloride, N,N-dimethylthiocarbamoyl chloride, N,N-ethylmethylcarbamoyl chloride and piperidinecarbamoyl chloride) were added to the solution. The solution was maintained at 40 °C (reflux temp.) for 12 h, and then the solvent was removed in vacuo, residue was dissolved in ethylacetate and washed with water. Organic portion was collected, dried over sodium sulfate and concentrated under vacuo. Crude product was further purified by column chromatography.

*III. Procedure for the preparation of 2-O-Naphthyl-dimethylcarbamothioate (8, Scheme 3)*

2-Naphthol (0.1 M) was dissolved in dry N,N-dimethylformamide (70 ml) and sodiumhydride of a 60% mineral oil dispersion ( 0.1 M) and N,N-dimethylthio carbamoyl chloride (0.12 M) were added in one portion. The mixture was stirred at 50 °C for 10 h and at room temperature for 2 h. The reaction mixture was poured into cold water (200 ml), followed by extraction with ether (70 ml×3times). The total ethereal layer was washed with sodium bicarbonate and organic portion was collected dried over sodium sulfate and concentrated under vacuo.




**Scheme 5** Synthesis of 2-epoxy derivative (16) and oxime of isoprutoron (17) [38–40] Reagents & conditions: (xvi) $(CH_3)_2S$ / $(CH_3)_2SO_4$ / KOH flakes/ 40 °C; (xvii) $NH_2OH$. HCl /Pyridine /Ethanol /reflux




**Scheme 6** Preparation of 1-dimethylamino-1-(4-Isopropyl-phenylamino)-2-(1,2,4)triazol-1-yl-ethanol(18) [40] Reagents and conditions ; (xiii) N, N-dimethyl-N-(4-(propane-2-yl)phenyl)oxirane-2,2-diamine and :1,2,4-triazole $K_2CO_3$ / DMF /135-140 °C

*III. Procedure for the preparation of 2-O-Naphthyl-dimethylcarbamate (9, Scheme 3)*

2-Naphthol (0.1 M) was dissolved in dry pyridine (70 ml) and N,N-dimethylcarbamoylchloride (0.12 M) was added in one portion. The mixture was stirred at 50 °C for 30 h. The reaction mixture was poured into 75 ml cold water, followed by extraction with ether (50 ml×3times). The total ethereal layer was washed with sodium bi carbonate and organic portion was collected dried over sodium sulfate and concentrated under vacuo.

*IV. General procedure for the preparation of cumidylurea derivatives (10–15, Scheme 4)*

Cumidylurea (3.0 mM) was dissolved in acetic anhydride (20 ml) and different derivatives [Cyanoacetic acid(x), Glycolic acid (xi), N,N-Dimethylcarbamoyl chloride(xii), N,N-dimethylthiocarbamoyl chloride(xiii) and N,N-ethyl-methyl carbamoyl chloride (xiv)] were added to the solution. The solution was maintained at 70 °C for 8 h and then acetic anhydride was removed in vacuo, residue was dissolved in ethylacetate and washed with water. Organic portion was collected dried over sodium sulfate and concentrated under vacuum. Crude product was further purified by column chromatography.

**Table 1** Representing difference in susceptible and resistant biotypes of *Phalaris minor*

| *PsbA* gene in susceptible biotype | | | |
|---|---|---|---|
| Position in *PsbA* gene | Nucleotide | Respective codon | Corresponding amino acid |
| 70 | C | GCT | G(24) |
| 86 | T | GTC | V (29) |
| 291 | C | TTC | F (97) |
| 588 | T | AAT | N (196) |
| *PsbA* gene in resistant biotype | | | |
| 70 | A | AGT | S (24) |
| 86 | C | GCT | A(29) |
| 291 | A | TTA | L(97) |
| 588 | C | AAC | N (196) |

*V. Procedure for the preparation of N,N-dimethyl-N-(4-(propane-2-yl)phenyl)oxirane-2, 2-diamine (16, Scheme 5)*

To a solution of dimethylsulfide ( 0.41 M ) 3-(4-Isopropyl-phenyl)-1,1-dimethylurea (0.07 M), water ( 0.071 M) was added 13gms (0.1 mol) dimethylsulfate at 35 °C for a period of 18 h. Caustic potash flakes (0.2 mol) over a period of 1 hr at 38-40 °C. Maintained this temperature for 12 h and unreacted dimethyl sulfide recovered up to mass temperature 60 °C. Cooled the reaction mixture to room temperature followed by extraction with hexane (50 ml×3times).Combined organic layer was collected dried over sodium sulfate and concentrated under vacuo.

*VI. Procedure for the preparation of 3-(4-Isopropyl-phenyl)-1,1-dimethylurea- Oxime (17, Scheme 5)*

To a solution of 3-(4-Isopropyl-phenyl)-1,1-dimethylurea (0.024 M), pyridine 0.5 ml, ethanol 50 ml added hydroxyl-amine hydrochloride (0.108 M). The reaction was refluxed on water bath for 12 h. After reaction ethanol was removed by vacuum distillation, to the residue 5 ml water was added cooled under stirring until the oxime crystallizes. The crystals was filtered, washed with water and dried.

*VII. Preparation of 1-dimethylamino-1-(4-Isopropyl-phenylamino)-2-(1,2,4)triazol-1-yl-ethanol (18, Scheme 6)*

To a solution of N,N-Dimethylformamide 50 ml, pulverized potassium carbonate (0.048 M) and 1,2,4-triazole(0.048 M), was added N,N-dimethyl-N-(4-(propane-2-yl)phenyl)oxir-ane-2,2-diamine(0.034 M) at 135-140 °C and maintained for 12 hrs. Solvent was removed under vacuo, residue was poured into 50 ml of water, followed by extraction with dichloromethane (50 ml×3times). Organic portion was collected dried over sodium sulfate and concentrated under vacuo.

Formulation

Chemically synthesized isoproturon derivatives were formu-lated with ethyl acetate and surfactants alkylated naphthalene

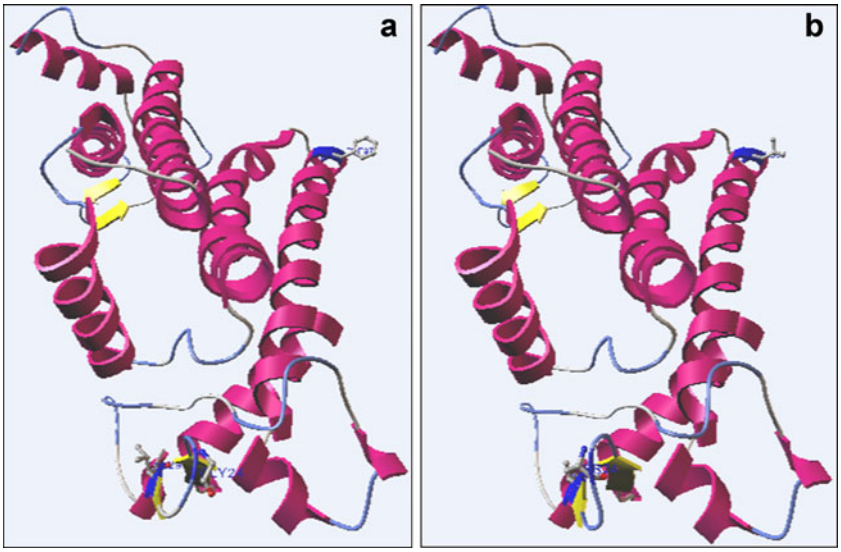



**Fig. 1** Modeled D1 protein showing residues Gly24, Ala29, Phe97 (**a**) Susceptible D1 protein (**b**) Resistant D1 protein

sulfonates either 10% EC (effective concentrations) or 20% EC. All these derivatives were applied for pot assay at the effective concentration of 0.5, 1.0, 1.5 kg/ hectare in 800lit of water.

Pot assay

*P. minor* seeds were grown in earthen pots. All synthesized urea derivatives were tested in three replicates at three leaf stage. For control wheat plants were also tested at three leaf stages. Seeds were procured from CSA Agriculture University and selected for study of the effectiveness of isoproturon derivatives. The pots were filled with soil and well-rotted manure was applied at the surface in ratio of 6:1 v/v volume. Seventy seeds were sown in each pot during the winter season. Plants were thinned to 40 each after 15 days and tested after a month of sowing. Synthesized derivatives were first screened at 1.5 kg/ha. Actives derivatives at 1.5 kg /ha were further tested by spraying at 0.5, 1.0 and 1.5 kg/ha with 800 L/ha of water at three leaf stage. Fresh weight of biomass was measured on 30$^{th}$ day of application and expressed as a percentage of control all data were presented as means of three replicates. Differences amongst treatment means were determined using ANOVA and when F test was significant means were compared with LSD test at 5% level of significance.

Dry weight

Optimized dose of triazole derivative (0.5 kg/ha) was sprayed in wheat and *P.minor* at three leaf stage in triplicates. One control (unsprayed) treatment was also kept for comparison. Dry weight of all treatments was measured on 21$^{st}$ day of treatment. The whole plant of each pot was taken and dry matter accumulation was measured after oven drying at 72 °C for 72 hr.

**Table 2** Variation in secondary structure of D1 protein near the site of mutation

| Mutated residue | Susceptible D1 protein | | Resistant D1 protein | |
|---|---|---|---|---|
| | Secondary structure | Residue range | Secondary structure | Residue range |
| Gly24 & Ala29 | Coil1 | Ala17-Pro19 | Coil 1 | Ala17-Pro19 |
| | Strand1 | Val20 & Asp21 | Strand 1 | Val20 & Asp21 |
| | Turn1 | Ile21-Ile25 | Turn 1 | Ile21-Glu27 |
| | Coil2 | Arg-26 & Glu27 | Loss of coil | ——— |
| | Strand2 | Pro28 & Pro29 | Strand2 | Pro28 & Pro29 |
| | Coil3 | Ser30-Ser32 | Coil2 | Ser30-Ser32 |
| Phe97 | Coil7 | Gly71 | Coil6 | Gly71 |
| | Helix5 | Gly72-Leu99 | Helix5 | Gly72-Leu99 |
| | Coil8 | Gly100-Arg102 | Coil7 | Gly100-Arg102 |

**Table 3** Comparison of transmembrane topology of D1 protein with nearest homologue *Thermosynechococcus elongatus* A-chain of 2AXT.Pdb

| 2AXT/ transmembrane segments | D1Suceptible/ transmembrane segments | Thylakoid lumen/ stromal segments of susceptible D1 protein |
|---|---|---|
| LYVGWFGVIMIPTLLAATICFV | PTLLTATSVFI | IAFIAAPPVDIDGIREPVSGSLLYGNNIISGAIIPTSAAIGLHFY PIWEAASVDEWLYNGGPYELI |
| IFHFLLGASCYMGRQWEL | VLHFLLGVACYMGREWEL | SFRLGMRP |
| WICVAYSAPLASAFAVFL | WIAVAYSAPVAAATAVFL | IYPIGQGSFSDGMPLGISGTFNFMIVFQAEHNILMHP |
| FHQLGVAGVFGGALFCAMHGSL | FHMLGYAGVFGGSLFSAMHGSL | VTSSLIRETTENESANEGYKFGQEEETYNIVAAHGYFGRLIF QYASFNNSRSL |
| HFFLAAWPVVGVWFTALGIS | HFFLAAWPYYGIWFTALGIS | TMAFNLNGST |

## Results

### Sequence analysis and BLAST

Nucleotide sequence of *psbA* gene of both biotypes was taken from the NCBI Genbank database in requisite format. The accession and GI No of susceptible [Accession No AY211527, GI: 30413114] and resistant biotype are [Accession No. AY294643 GI: 31505999] respectively. Nucleotide and amino acids sequence alignment of both biotypes were performed to find out exact location of changes [11]. Four point mutations reported are Gly 24 to Ser 24, Val 29 to Ala 29 and Phe 97 to Leu 97 in resistant D1 protein. However fourth point mutation is a silent mutation (Table 1) [12]. Functional annotation of the sequence has shown that it belongs to SSF81483, PF00124, PS00244 (Pattern[NQH]-x(4)-P-x-H-x(2)-[SAG]-x(11)-[SAGC]-x-H-[SAG](2).), IPB000484, PR00256, SCOP; 2RCR, CATH; 1PRC.

BLAST with PDB structure database of both susceptible and resistant biotype of *P. minor* was performed to find out nearest homologue. All the score greater or equal to>=250 were taken. Both susceptible and resistant D1 proteins have shown maximum score with photosynthetic oxygen evolving center and PS-II of *Thermosynechococcus elongatus*, while both have different scores in terms of e-value, positives and identities. Susceptible and resistant D1 protein produced significant alignment with photosynthetic oxygen evolving center (pdb|1S5L|A and pdb|1S5L|AA), PS-II of *T. elongatus* (pdb|2AXT|A, pdb|2AXT|AA) and crystal structure of PS-II (pdb|1IZL|A and pdb|1IZL|J).

**Table 4** Cavity analysis of D1 protein simulated by CAST-P program

| | Suceptible-D1 protein | | | Resistant-D1 protein | | |
|---|---|---|---|---|---|---|
| | Total no of cavities:33 | | | Total no of cavities:41 | | |
| Cavity ID | 20 | 23 | 26 | 2 | 19 | 32 |
| Area (Å$^2$) | 69.9 | 65.1 | 95.8 | 14.7 | 17.3 | 63.0 |
| Volume (Å$^3$) | 50.2 | 68.6 | 92.5 | 8.9 | 12.3 | 60.7 |
| Mutated residue in Cavity | V29: O | V29:CG2 | V29, CG1 | L97: CD2 | L97: CB,CD1, N | S24: C,O |
| Residues/atoms in the cavity | A17: CB, O | I21: CB, CD1 | V29: CG1 | W93: CZ3 | W93: CD1, NE1, O | D21: CG |
| | P18: O | CG1, N | S30: O | E94: CG | S96: CB | I22: O |
| | P19: CA | D22: CB, N, | G31: CA | L97: CD2 | L97: CB,CD1, N | D23: C,O |
| | V20: CG2 | OD2 | N37: ND2 | R98: NH2 | P103: CB, CG | S24: C,O |
| | V29: O | I25: CB, D1, | S41: O | | | I25: CA, CG2, O |
| | G31: CA | CG2, O | G42: C, CA, O | | | R26: NH1 |
| | I44: CD1 | E26: CB, CG, | A43: N,O | | | R48: OG |
| | I45: CG2, N | O, OE1, CG2 | I44: O | | | A49: CB |
| | | V29: CG2 | I45: CA, CB, CD1 | | | A50: CB,N |
| | | | P135: CB | | | |
| | | | L136: C,N | | | |
| | | | G137: N | | | |
| | | | G140: CA | | | |

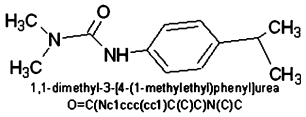



**Fig. 2** 2-D structure of isoproturon

## Homology modeling

Modeling has been performed with modeler [22, 23] and loop regions were refined with swiss PDB viewer with overall quality factor of 96.032 and 94.118 for both susceptible and resistant D1 proteins respectively (Fig. 1a and b). Stereochemical evaluation of backbone Phi ($\varphi$) and Psi ($\psi$) dihedral angle of residues from Pro1 to Gly261 were performed by procheck for both susceptible and resistant D1 proteins [24]. Ramachandran plots of susceptible and resistance D1 has shown that core regions, additionally allowed regions, generously allowed regions and disallowed regions were 89.6%, 9.5%, 0.9%, 0.0% and 86.1%, 13.5%, 0.4%, 0.0% and respectively. Total 99.1% residues of susceptible and 99.6% of resistance D1 has shown the residues are in the most favored and allowed regions. In susceptible protein Arg 26 & Glu27 form a coil which is lost in resistant D1 protein and forms turn from Ile21-Glu27 (Table 2). Thus the change in secondary structure near the mutated residue explains the cause of resistance.

## Prediction of transmembrane segments

To find out transmembrane segments the crystal structure of pdb|2AXT|AA Chain A, of PS-II of *Thermosynechococcus elongatus*, [41] which is a nearest homologue of D1 protein was compared. Finally it is concluded that D1 protein has five transmembrane segments and all three mutated residues lie outside the transmembrane segments. The first thylakoid lumen segment ranging from Ile12 to Ile79 holds mutated residues Ser24 and Ala29 while first stromal segment Ser 96 to Pro103 holds mutated residue Leu97 (Table 3).

## Simulation for binding site

Cast-P program has been used for cavity area and volume simulation in both susceptible and resistant biotype [42]. Cavities which possess mutated residues were considered (Table 4). Cavities 20, 23, 26 hold Val29 of susceptible D1 protein, while in resistant D1 protein only cavity ID: 32 holds mutated residues Ser 24 while the other two cavities are 2 and 19 that hold mutated residue Leu97 with negligible volume and area. Cavity simulation has shown loss of actual size and area of cavities in resistant D1 protein in comparison to susceptible D1 protein that holds Val29 residue, which might be the cause of resistance against isoproturon.

Isoproturon has one phenyl ring, calculated LogP value is 2.813, has one hydrogen bond donor and one acceptor (Fig. 2).

## Docking

Isoproturon molecule was docked at the cavity of susceptible D1 protein and cavity contains mutated residues of resistant D1 protein, i.e., residue 24 and 29 were considered as a center of grid in both cases of energy mapping (Fig. 3a and b). The docking and binding energies of three best conformations are given in Table 5. Three conformations of each simulation were considered. Docking and binding energies in the case of susceptible D1 protein is almost

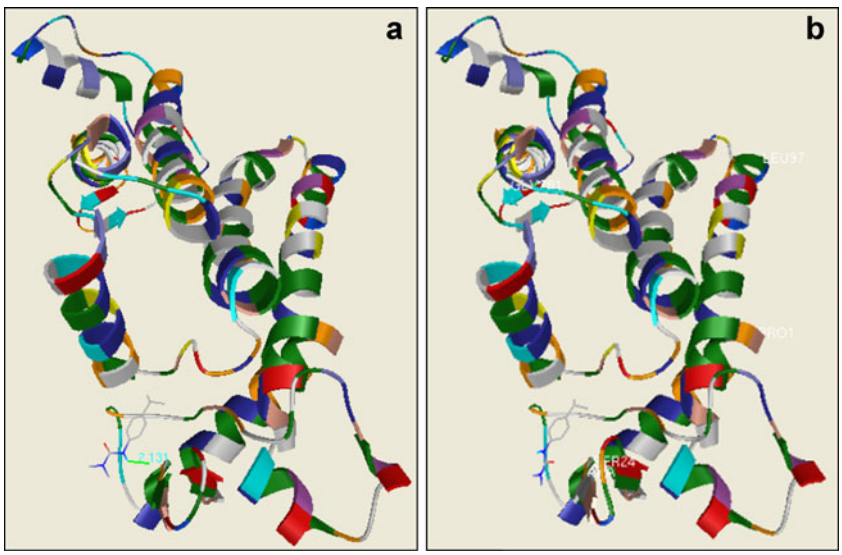



**Fig. 3** The bound conformation of isoproturon at the active site of (**a**) susceptible D1protein (**b**) resistant D1 protein

**Table 5** Summary of docking, binding energy and hydrogen bond interaction with isoproturon

| Susceptible D1 protein | | | Resistant D1 protein | | |
|---|---|---|---|---|---|
| Docking energy (kcal/mol) | Binding energy (kcal/mol) | H-bonding | Docking energy (kcal/mol) | Binding energy (kcal/mol) | H-bonding |
| −6.05 | −5.68 | $OH_{SER30}$…. $N_7$ (Bond length=3.131 Å) | −5.86 | −5.54 | — |
| −6.05 | −5.65 | $OH_{SER30}$…. $N_7$ (Bond length=3.138 Å) | −5.54 | −5.12 | — |
| −6.04 | −5.65 | $OH_{SER30}$…. $N_7$ (Bond length=3.150 Å) | −5.35 | −4.97 | — |

equal while in resistant D1 protein docking energies are −5.86, -5.54, -5.35 kcal $mol^{-1}$ and binding energies are −5.54, -5.12, and −4.97 kcal $mol^{-1}$ respectively. In susceptible D1 protein OH of Ser 30 forms hydrogen bond with $N_7$ of nitrogen atom while hydrogen bonding is lost in resistant D1 protein, explain cause of resistance against isoproturon. Difference in docking energies of three conformers also indicates that isoproturon molecule is not well fitting in the cavity of resistant D1 protein, which also explains loss of cavity volume and area in resistant D1 protein.

We used ligplot to find out nonbonded interaction between ligand and interacting amino acid residues of pockets of each resistant and susceptible D1 protein of *P. minor* (Fig. 4). Resistant D1 protein has shown loss of hydrogen bond while OH of Ser 30 of susceptible D1 protein forms hydrogen bond with N7 atom of isoproturon. C1, C5 and C6 atom of isoproturon in resistant D1 protein have lost hydrophobic interaction with amino acid residues Ser 30, Val 29, Gly 31, Gly 42 and Ile 45 while it remains in the case of susceptible D1 protein Table 6.

## Receptor based pharmacophore

Simulating pharmacophore from of the 3D structure of a target protein provides important insight for analyzing protein-ligand interactions and further improvement of ligand binding affinity. In order to design and synthesis new potent herbicides which can regain sensitivity against resistant variety of *P. minor* pharmacophore for both resistant and susceptible biotypes were simulated. Ligbuilder is a program which simulates a key site of interaction and pharmacophore for a given ligand and receptor.

Pharmacophore of susceptible biotypes has shown two hydrophobic sites, two hydrogen bond donor and one hydrogen bond acceptor sites and maximum spatial distance between two points the hydrophobic site and hydrogen bond acceptor of key interaction, which was 7.25 Å (Fig. 5) (Table 7). Pharmacophore of resistant biotypes has shown one hydrophobic site two hydrogen bonds donors and one hydrogen bond acceptor site and maximum distance among these (hydrophobic site and hydrogen bond acceptor) sites was 7.89 Å.

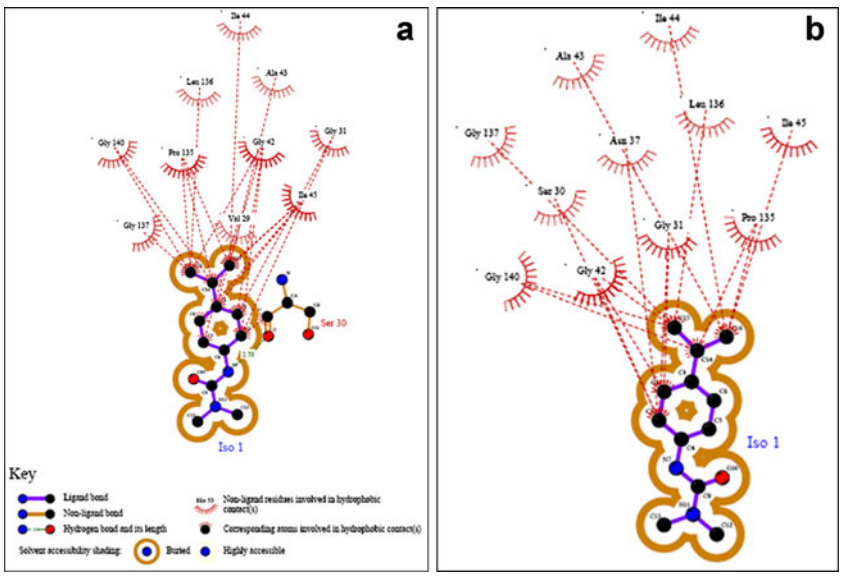



**Fig. 4** 2D representation of docked conformations of (Fig. 3a and b) of Isoproturon showing nonbonded interactions

**Table 6** Interacting amino acids with isoproturon at the active site of susceptible and resistant D1 protein

| Summary of interaction | Hydrogen bonds | Ligand atoms involved in hydrophobic contacts | Amino acids at the active site of D1 protein involved in hydrophobic contacts |
|---|---|---|---|
| Susceptible D1 protein | $ISO_{NH}$——$_{O}$SER30 | C1, C2, C3, C5, C6, C14, C15, C16 | I44, L36, A43, G137, G140, P135, G42, G31, I45, Carbonyl carbon of S30 and V29 |
| Resistant D1 Protein | —— | C2, C3, C14, C15, C16 | I44, L136, A43, G137, G140, P135, G42, G31, I45, S30 and N37 |

Since herbicide of PS-II which binds with D1 protein are derivatives of urea molecule therefore a fragment ($NH_2CONH_2$) of isoproturon molecule was taken as a seed for designing of new ligands at the active site of resistant biotypes. Four growing points (H.spc) were assigned in the seed ($NH_2CONH_2$) molecule. Grow module of ligbuilder was used for addition of a new fragment which uses genetic algorithm for optimization of grown molecule. Process module was used for screening of grown molecule considering the Lipinski rule of five and ADMET (absorption, distribution, metabolism, excretion and toxicity) parameters [43]. De novo design of molecule gives an idea about the organic synthesis of new testable ligands (Fig. 6) (a) 8-(propan-2-yl)-1,4-dihydronaphthalen-2-yl dimethylcarbamate, (b) *N*-(hydroxymethyl)-*N*'-[4-(propan-2-yl)phenyl]dicarbonimidic diamide, (c) -*N*,*N*-dimethyl-*N*'-[4-(propan-2-yl)phenyl] tricarbonodiimidic diamide and (d) 2-{2-methyl-1-[(*Z*)-(methylimino)methyl]hydrazinyl}-*N*-[4-(propan-2-yl)phenyl] acetamide

## Synthesis

As a result of the conclusion drawn from pharmacophore of D1 protein, various isoproturon derivatives were synthesized. The phenyl ring of isoproturon was kept preserved while changes were made at the other end. Isoproturon derivative (1); (thiocarbamoyl), (2); (ethyl-methylcarbamoyl), (3); piperidine carbamoyl) and (4); dimethylcarbamoyl were synthesized using paracumidine as a base molecule. Derivative (5); carbamoyl, (6); thiocarbamoyl, (7) ethyl methyl carbamoyl were synthesized from paracumidinyl biuret. All these derivatives were synthesized with good yields. Napthyl derivatives, 2-O-naphthyl-dimethylcarbamothioate (derivative 8) and 2-O-naphthyl-dimethylcarbamate (derivative 9) [37] were synthesized to add more hydophobicity in the molecule. Derivative 10 to 15 were synthesized from paracumidyl urea and acetic anhydride heated up to 70 °C. Epoxy and oxime derivatives of isoproturon, (16 and 17) were synthesized from 3-(4-Isopropyl-phenyl)-1,1-dimethyl-urea was refluxed with hydroxylamine hydrochloride to synthesize its oxime. Triazole derivative [40] was synthesized from 1,2,4-Triazole and N,N-dimethyl-N-(4-(propane-2-yl)phenyl)oxirane-2,2-diamine [40] and heated up to 135–140 °C, obtained product was an oil. All the derivatives were purified via silica gel column chromatography; yield was in the range of 45% to 92%.

## Pot assay

All synthesized derivatives were screened at 1.5 kg/ha, among them derivatives 2, 5, 9 and 18 were effective. Derivatives 2, 5, 9 and 18 were further tested at 0.5, 1.0 and 1.5 kg/ha. After 30 days of application triazole derivative has shown 100% loss in fresh weight in susceptible and resistant biotypes at all doses in comparison to control. Isoproturon has shown 100% and 32.4% loss in fresh weight in susceptible and resistant biotypes at 1.5 kg/ha in comparison to control. At 0.05 significance level with 4 and 8 degrees of freedom, $F_{.95}$=3.84 (17.66>3.84) for susceptible and (47.0> 3.84) for resistant, i.e., there is significant difference in %loss

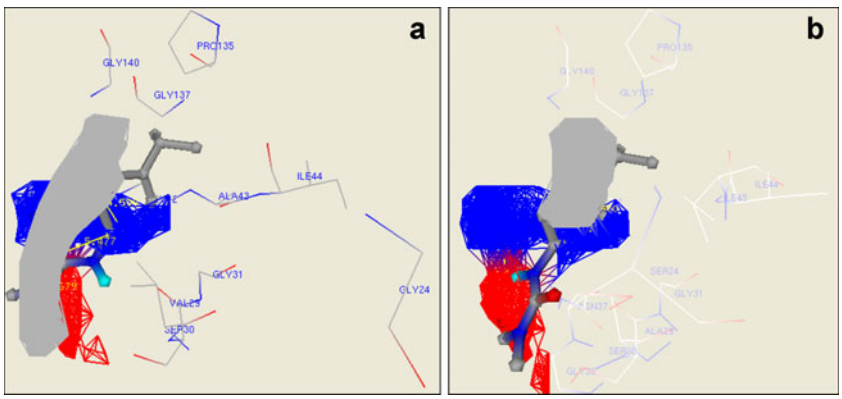



**Fig. 5** Receptor based Pharmacophore of (**a**) susceptible and (**b**) resistant D1 protein of *phalaris minor* has shown loss of hydrophobic sites which explains probable cause of resistance. Color gray, blue and red represents for hydrophobic, H-bond donor and H-bond acceptor sites respectively

**Table 7** Distances in Å shown from each key site of interactions where H, D & A denotes hydrophobic H-bond donor and H-bond acceptor regions of D1 protein pharmacophore

| | Susceptible D1 protein | | | | | Resistant D1 protein | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | H | D | H | D | A | | H | D | D | A |
| H | —— | 4.50 | 3.50 | 7.25 | 4.85 | H | — | 4.50 | 5.85 | 7.89 |
| D | 4.50 | —— | 5.48 | 4.61 | 3.57 | D | 4.50 | — | 3.54 | 4.95 |
| H | 3.50 | 5.48 | —— | 5.68 | 4.97 | D | 5.85 | 3.54 | —— | 4.00 |
| D | 7.25 | 4.61 | 5.68 | —— | 4.06 | A | 7.89 | 4.95 | 4.00 | —— |
| A | 4.85 | 3.57 | 4.97 | 4.06 | —— | | | | | |

of fresh weight due to derivates in both biotypes and with 2 and 8 degrees of freedom, $F_{.95}$=4.46 (14.26>4.46) for susceptible and (9.2>4.46) for resistance that means there is significant difference in % loss in fresh weight due to doses in both biotypes Table 8. Triazole derivative of isoproturon was lethal at 1.5 kg/ha for both wheat and *P. minor*. The optimized dose of triazole derivative for *P. minor* and wheat was 0.5 kg/h. Effect of triazole derivative at 0.5 kg/ha has been shown in both susceptible and resistant biotypes of *P. minor* (Fig. 7).

Effect of isoproturon and triazole derivative at 0.5 kg/ha on dry weight accumulation

Wheat and both biotypes of *P.minor* were exposed to 0.5 kg/ha of triazole derivative and isoproturon at three leaf stage. On 21$^{st}$ day of treatment dry weight was measured. Both R and S biotypes have shown significant loss in dry weight accumulation on triazole derivative treatment. Isoproturon susceptible biotype also has shown remarkable decrease in dry weight while resistant biotype was marginally affected (Table 9).

## Discussion

The two primary sites of herbicide action in photosynthetic electron transport are the inhibition of PS-II electron transport and diversion of electron flow through PS-I. PS-II electron transport inhibitors bind to the D1 protein of the PS-II reaction center, thus blocking electron transfer to plastoquinone, prevents the conversion of absorbed light energy into electrochemical energy and results in the production of triplet chlorophyll and singlet oxygen, which induce peroxidation of membrane lipids [44]. Isoproturon is expected to have a mode of action similar to that of other substituted phenyl urea herbicides, for which the identified target site is in the photosynthetic apparatus especially at D1 protein of PS-II [45–48]. Isoproturon application results in decreased protein and chlorophyll contents and also reduces RuBP Carboxylase activity and promotes disorganization of grana and intergrana in wheat [49].

D1 is a thylakoid membrane protein of PS-II. *P. minor* developed resistance to isoproturon due to mutation in D1 protein. To find out the mechanism of resistance against isoproturon translated *psbA* gene sequence [GI: 31505999 & GI: 30413114] [11] of both susceptible and resistant D1 proteins [Q85V85|Q85V85_9POAL *psbA* & Q7YMJ5|Q7YMJ5_9POA *psbA*] were modeled by modeler. However these sequences are not complete but a fragment therefore it was aligned with refseq of *Triticum aestivum* (>gi|13928185|dbj|BAB47014.1|) of PSII 32 kDa protein which shows 38 residues (MTAILERRESTSLWGRFCNWITSTENR LYIGWFGVLMI) truncated at N-terminal and 52 residues truncated at C-terminal (FNQSVVDSQGRVINTWADIIN

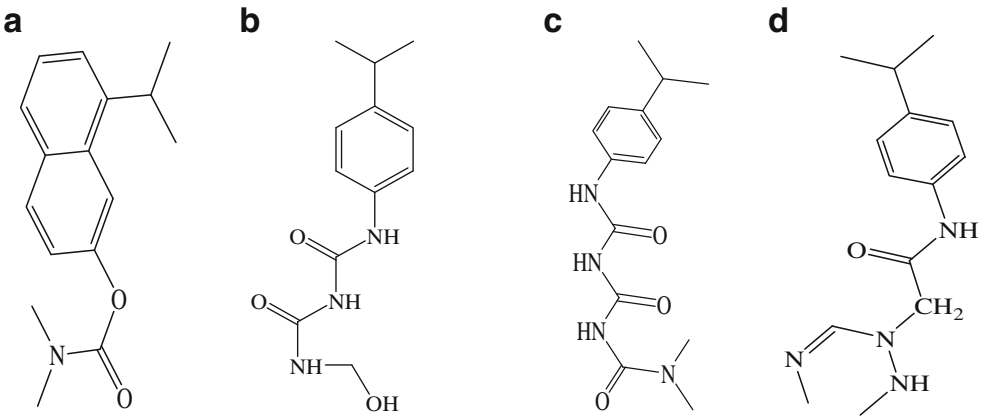



**Fig. 6** Common frame scaffold of ligand molecule drawn from LigBuilder simulation for synthesis of isoproturon analogs

**Table 8** Response in S and R biotypes on application at 0.5, 1.0 and 1.5 kg/ha. (All data were presented in means of three replicates and difference among treatment mean were determined using ANOVA)

| Derivatives | % control S- biotypes | | | Row Total | Row mean | % control R-biotypes | | | Row total | Row mean |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0.5 kg/ha | 1.0 kg/ha | 1.5 kg/ha | | | 0.5 kg/ha | 1.0 kg/ha | 1.5 kg/ha | | |
| Control | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 2 | 17.4 | 32.0 | 67.5 | 116.9 | 38.9 | 0.0 | 0.0 | 22.1 | 22.1 | 7.36 |
| 4* | 62.5 | 82.5 | 100 | 245 | 81.6 | 0.0 | 15.0 | 32.4 | 47.4 | 15.8 |
| 5 | 0.0 | 30.0 | 62.3 | 92.3 | 30.7 | 0.0 | 0.0 | 20.3 | 20.3 | 6.76 |
| 9 | 15.0 | 32.3 | 72.7 | 120 | 40.0 | 0.0 | 17.6 | 52.3 | 69.9 | 23.3 |
| 18 | 97.8 | 100.0 | 100.0 | 297.8 | 99.2 | 97.5 | 100.0 | 100.0 | 297.5 | 99.1 |
| Column total | 192.7 | 276.8 | 402.5 | Grand total=872 | | Column total | 97.5 | 132.6 | 227.1 | Grand total =457.2 |
| | | | | Grand mean =58.1 | | | | | | Grand mean =30.48 |
| Column mean | 38.5 | 55.36 | 80.5 | | | Column mean | 19.5 | 26.52 | 45.42 | |

*Isoproturon

RANLGMEVMHERNAHNFPLDLAAVEVPS). The modeled structure was aligned with 2AXT of A chain to evaluate the impact of the missing structure on the binding of isoproturon. Isoproturon binds apart from the N and C terminal of D1protein. From visual inspection of 3D structure and alignment it was evident that missing part of N and C terminal of D1protein does not affect the binding of isoproturon.

The modeled structure has shown variation in secondary structure in the region adjacent to the site of mutation (Table 2). Since D1 protein is a membrane protein it is necessary to predict membrane topology and for this we compared it with the crystallographic structure of *T. elongatus* [2AXT.pdb] that is nearest homolog of D1 protein. By comparing with chain A of 2AXT.pdb five transmembrane segments were visualized and mutated residues lying outside the domain of transmembrane region. The first two mutated residues [Ser 24 and Ala 29] lie in the thylakoid lumen, third mutated residue Leu 97 lies in the chloroplast stromal region of transmembrane segments while, the fourth point mutation is a silent (Table 3).

To find out the altered binding site of isoproturon in D1 protein, cavities holding mutated residues were considered (Table 4). Susceptible D1 proteins have shown three cavities in the region of Val 29 mutated residue with the same volume and area while no cavity was observed in the region of Phe 97 residue. Therefore it is clear that isoproturon is binding with the first thylakoid lumenal segment of D1 protein region either with Gly 24 or nearby Val 29. Isoproturon is a

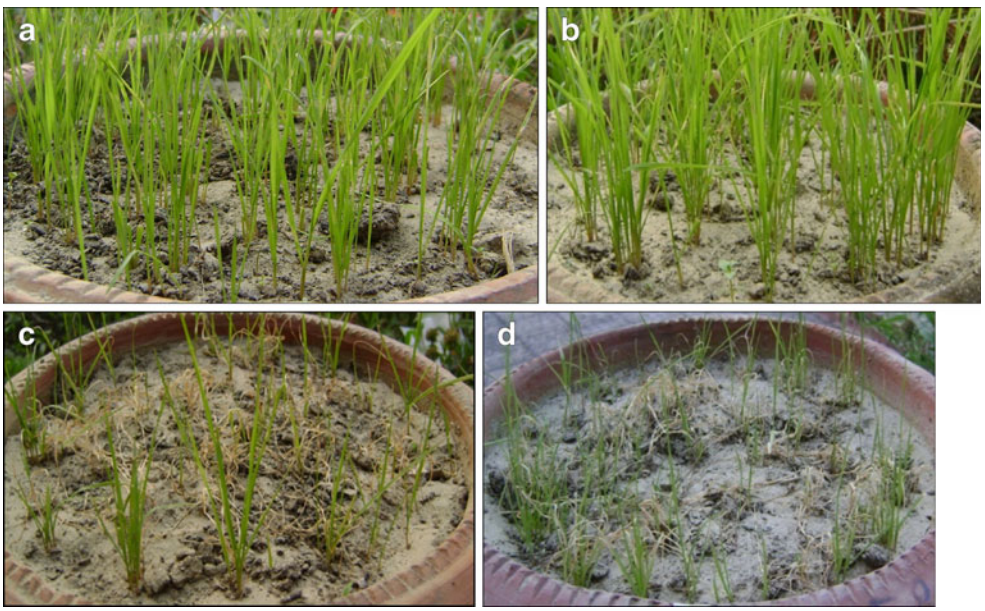



**Fig. 7** Effect of synthetic derivative (Scheme 6, analogue-18) of isoproturon after 7$^{th}$ day of pot assay at 0.5 kg/ha. (**a**) Control susceptible biotype, (**b**) Control resistant biotype, (**c**) Susceptible biotype tested at 0.5 kg/ha, (**d**) Resistant biotype tested at 0.5 kg/ha

**Table 9** Effect of derivative-18 and isoproturon (at 0.5 kg/ha) on dry matter accumulation (g/pot) of *P.minor* (all data are presented from the means of three replicates)

| Wheat/P.minor | Dry weight g/pot | | |
|---|---|---|---|
| | Control | Derivative-18 | Isoproturon |
| Wheat | 15.21 | 13.01 | 14.82 |
| P. minor S | 12.314 | 1.34 | 8.34 |
| P. minor R | 13.01 | 1.54 | 11.64 |

small hydrophobic molecule; its log P value and mol wt indicate that it is quite capable to cross cell membrane.

We docked isoproturon in the same region (near to Val 29 in susceptible and Ala 29 in resistant) with both susceptible and resistant D1 protein. Docking has shown with susceptible D1 protein that it forms a hydrogen bond between N7 of isoproturon and OH of Ser 30, while it is lost in resistant D1 protein (Fig. 3a and b). 2D plot of 3D conformations of isoproturon and resistant D1 protein has shown loss of hydrophobic interaction with C1, C5 and C6 carbon while it is retained in susceptible D1 protein (Fig. 3a and b).

For chemical synthesis of ligand molecule we simulated the key site of interaction and pharmacophore for design and synthesis of a new ligand molecule. Receptor based pharmacophore has shown complete loss of a hydrophobic site at the binding site in resistant D1 protein, thus giving an explicit idea regarding the type of modification to be carried out (Fig. 6). Synthesized derivatives were first tested at only 1.5 kg/ha to screen the effective derivatives of isoproturon. Further effective derivatives (2, 5, 9 and 18) were tested at 0.5, 1.0, 1.5 kg/ha in triplicates. Among these four derivatives, triazole was most effective at 1.5 kg/ha for *P. minor* and also affects on wheat plant. Therefore it was optimized for pot assay as a result triazole derivatives of isoproturon was equally effective in both biotypes at 0.5 kg/ha and no side effect to wheat plant. Therefore triazole derivative (Derivative-18) was redocked at both susceptible and resistant D1 protein. It has shown the same mode of binding at isoproturon docking site and equal docking and binding energies in both D1 proteins. Isoproturon has shown hydrogen bonding with Ser 30 of susceptible D1 protein while it was lost in resistant D1 protein. Triazole derivative has resumed hydrogen bonding in both cases (Table 10). Hence hydrogen bonding with Ser 30 is crucial for binding.

The tolerable dose of triazole derivative to wheat crop is 0.5 kg/ha. Triazole derivative has shown signs of wilting, yellowing and loss of dry weight after 24 hr of treatment and complete wilting were observed on 7$^{th}$ day of application. Dry weight accumulation was negligible in comparison to triazole treatment at same dose in both S and R biotypes measured on 21$^{st}$ day of treatment shows the effectiveness of triazole derivative of isoproturon (Fig. 7).

## Conclusions

*P. minor* is a major weed affecting production of wheat, a stable global food crop. It has developed resistance to isoproturon due to the latter's continued use for a long time. Mutation in herbicide binding protein has displaced the binding site for isoproturon and shifted to resistant variety. The present work describes the docking of herbicide molecules at the active site of susceptible as well as resistant protein D1, suggesting the cause of resistance which in all probability is due to the loss of hydrogen bonding and change in hydrophobic interactions. Based on this data, appropriate modification in the structure of isoproturon molecule has been done and new derivatives synthesized, characterized, and used for spray of both susceptible and resistant varieties. The triazole derivative of the isoproturn is most effective, and it could be used as a substitute of isoproturon.

## Summary

The cause of resistance to the herbicide isoproturon, developed by *P. minor*, a weed affecting growth of wheat, has been found to be due to loss of hydrogen bonding and shifting of hydrophobic sites in protein D1 target site. Urea

**Table 10** Summary of docking, binding energy and hydrogen bond of isoproturon derivative-18

| Susceptible D1 protein | | | Resistant D1 protein | | |
|---|---|---|---|---|---|
| Docking energy (kcal/mol) | Binding energy (kcal/mol) | H-bonding | Docking energy (kcal/mol) | Binding energy (kcal/mol) | H-bonding |
| −4.77 | −5.88 | $H_{28}…OH_{SER30}$ (2.199 Å) $OH_{SER30:HG}…N_9$ (3.138 Å) | −4.38 | −5.77 | SER30:HG…$N_9$ (2.112 Å) |
| −4.75 | −5.77 | ———— | −4.38 | −5.58 | $H_{28}…OH_{SER30:\ O}$ (2.008 Å) $HN_{SER30}$ …$N_7$ (1.617 Å) |
| −4.79 | −5.79 | ———— | −4.29 | −5.46 | ———— |

derivatives having novel structures have been proposed, synthesized, and tested. As a result we report triazole derivative is most effective and could be used as a substitute of isoproturon herbicides.

**Acknowledgments** One of the authors (D. V. Singh) wishes to express his gratefulness to Prof. D. P. Misra, Ex-head, Biochemistry department, G. B. Pant University of Agriculture & Technology, Pantnagar, India for initiating the problem and to Prof. R. K. Malik, Director, Research Extension Education, Chaudhary Charan Singh Haryana Agricultural University, Hissar, Haryana, India for providing seeds of resistant & susceptible biotypes of *P. minor*. D. V. Singh would also like to thank Council of Scientific and Industrial Research (CSIR), Govt. of India for providing him a senior research fellowship. Another author (K.Adeppa) gratefully acknowledges the financial help and laboratory facilities made available to him by India Pesticide Limited, Lucknow, India

## References

1. Yadav A, Malik RK, Chauhan BS, Gill G (2002) In: Malik RK, Balyan RS, Yadav A, Pahwa SK (eds) Present status of herbicides resistance in Haryana. Proceedings of the international workshop on herbicide resistance management and zero tillage in rice-wheat cropping system. CCS Haryana Agricultural University, Hisar, India, pp15-22
2. Yadav A, Malik RK (2005) Herbicide resistant *P. minor* in wheat – a sustainability issue, resource book. Department of Agronomy and Directorate of Extension Education, CCS Haryana Agricultural University, Hisar, India, pp 152–195
3. Chhokar RS, Malik RK (2002) Isoproturon-resistant little seed canarygrass (*Phalaris minor*) and its response to alternate herbicides. Weed Technol 16:116–123
4. Malik RK, Singh S (1995) Little seed canary grass (*Phalaris minor*) resistance to isoproturon in India. Weed Technol 9:419–425
5. Chhokar RS, Sharma RK (2008) Multiple herbicide resistant in little seed canary grass (*Phalaris* minor): A threat production in India. Weed Bio Manag 8:112–123
6. Yaduraju NT (1999) Control of herbicide resistant *Phalaris minor* need for a sound weed management a system. Pestology special issue pp 264–270
7. Singh S, Kirkwood RC, Marshall G (1998) Effect of the monooxygenase inhibitor piperonyl butoxide on the herbicidal activity and metabolism of isoproturon in herbicide resistant and susceptible biotypes *P. minor* and wheat. Pestic Biochem Physiol 59:143–153
8. Gardner G (1989) A stereochemical model for the active site of Photosystem II herbicides. Photochem Photobiol 49:331–336
9. Pfister K, Arntzen CJ (1979) The mode of action of Photosystem II specific inhibitors in herbicide resistant weed biotypes. Z Naturforsch 42c:783–793
10. Van Oorschot JLP (1991) Chloroplastic resistance of weeds to triazines in Europe Herbicide resistance in weeds and crops, Oxford UK Butterworth-Heinmann Ltd
11. Tripathi MK (2003) Biochemical and molecular mechanism of isoproturon resistance in *Phalaris minor*, PhD Thesis GBPUA & T, Pantnagar, India
12. Tripathi MK, Yadav MK, Gaur AK, Mishra DP (2005) PCR based isolation of psbA (herbicide binding protein encoding) gene using chloroplast and genomic DNA from Phalaris minor biotype(s). Physiol Mol Biol Plants 11:161–163
13. Singh DV, Gaur AK, Mishra DP (2004) Biochemical and molecular mechanism of resistance against isoproturon in *Phalaris minor* biotypes: Variation in protein and RAPD profiles of isoproturon resistant and susceptible biotypes. Indian J Weed Sci 36:256–259
14. Mordern CW, Golden SS (1989) *PsbA* gene indicate common ancesstory of chlorophytes and chloroplasts. Nature 337:382–385
15. Michel HOE, Deisenhofer J (1986) Pigment protein interaction in the photosynthetic reaction centre from *Rhodopseudomonas viridis*. EMBO J 5:2445–2451
16. Michel HOE, Deisenhofer (1988) Relevance of the phosynthetic reaction centre from the purple bacteria to the structure of Photosystem II. J Biochem 27:1–7
17. Shigematsu Y, Sato F, Yamada Y (1989) A binding model for phenyl urea herbicides based on analysis of a Thr264 mutation in D1 protein of tobacco. Pestic Biochem Physiol 35:33–41
18. Sinning I, Michel H, Mathis P, Ratherford AW (1989) Characterization of four herbicide resistant mutants of *Rhodopseudomonas viridis* by genetic analysis electron paramagnetic resonance and optical spectroscopy. Biochemistry 28:5544–5553
19. Trebst A (1987) The three dimensional structure of herbicide binding niche on the reaction centre polypeptides of Photosystem II. Z Naturforsch 42C:742–750
20. Trebst A (1991) The molecular basis of resistance of Photosystem II inhibitors Herbicide resistance in weeds and crops. Butterworth-Heinmann Ltd, Oxford UK
21. Feng QJ, Cong WN, Chao NC, Qiong C, Guang FY, Zhen X, Chang GZ (2008) Computational Design and Discovery of Conformationally Flexible Inhibitors of Acetohydroxyacid Synthase to Overcome Drug Resistance Associated with the W586L Mutation. Chem Med Chem 3:1203–1206
22. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234:779–815
23. Lund O, Nielsen M, Lundegaard C, Worning P (2002) CPHmodels 2.0: X3M a Computer Program to Extract 3D models CASP5 conference A102
24. Protein Structure Analysis and Validation server http://nihserver.mbi.ucla.edu/SAVS/
25. Goodsell DS, Olson AJ (1990) Automated docking of substrates to proteins by simulated annealing. Proteins 8:195–202
26. Morris GM, Goodsell DS, Huey R, Olson AJ (1996) Distributed Automated Docking of Flexible Ligands to Proteins: Parallel Applications of AutoDock 2.4. J Comput Aided Mol Des 10:293–304
27. Vajragupta O, Boonchoong P, Morris GM, Olson AJ (2005) Active site binding modes of curcumin in HIV-1 protease and integrase. Bio Med Chem Lett 15:3364–3368
28. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ (1998) Automated docking using Lamarckian genetic algorithm and an empirical binding free energy function. J Comput Chem 19:1639–1662
29. Hetenyi C, Van der Spoel D (2002) Efficient docking of peptides to proteins without prior knowledge of the binding site. Protein Sci 11:1729–1137
30. Sotriffer CA, Ni HH, McCammon JA (2000) Active Site Binding Modes of HIV-1 Integrase Inhibitors. J Med Chem 43:4109–4117
31. ACD/ChemSketch version 8.0 (2006) Advanced Chemistry Development, Inc., Toronto ON, Canada, www.acdlabs.com
32. Wallace AC, Laskowski RA, Thornton JM (1995) LIGPLOT: A program to generate schematic diagrams of protein-ligand interactions. Prot Eng 8:127–134
33. McDonald IK, Thornton JM (1994) Satisfying Hydrogen Bonding Potential in Proteins. J Mol Biol 238:777–793
34. Wang R, Gao Y, Lai L (2000) LigBuilder: A Multiple-Purpose Program for Structure-Based Drug Design. J Mol Model 6:498–516

35. Volford J, Knausz D, Meszticzky A, Horváth L, Csákvári B (1981) New synthesis of carbamate, thiocarbamate and urea type herbicides: Preparation of 14C-labelled diuron and E P T C. J Labelled Compd Rad 18:4555–4561
36. Meessen JH, Petersen H (2002) "Urea" in Ullmann's Encyclopedia of Industrial Chemistry. Wiley-VCH, Weinheim
37. Bruneau C, Dixncuf PH (1987) Catalytic synthesis of O-β-oxoalkylcarbamates. Tetrahedron Lett 28:2005–2008
38. Kutsuma T, Nagayama I, Okazaki T, Sakamoto T (1977) A Convenient Method for a preparation of oxirane. Heterocycles 8:397–401
39. Taj SS, Shah AC, Lee D, Newton G, Soman R (1995) Asymmetric synthesis of oxiranes from carbonyl compounds by methylene transfer reaction using chiral S-neomenthyl and S-exo-2-bornyl sulfoximines. Tetrahedron Asymmetry 6:1731–1740
40. Wolf R, Peter F (1984) Process for the preparation and reactions of oxiranes, USpatent 4929735
41. Loll B, Kern J, Saenger W, Zouni A, Biesiadka J (2005) Towards complete cofactor arrangement in the 3.0 A resolution structure of Photosystem II. Nature 438:1040–1044
42. Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, Liang J (2006) Computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. Nucleic Acids Res 34:116–118
43. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development. Adv Drug Deliv Rev 23:3–26
44. Fuerst EP, Norman MA (1991) Interaction of herbicides with photosynthetic electron transport. Weed Sci 39:458–464
45. Pfister K, Radosevich SR, Gardeney G, Arntzen CJ (1981) Photoaffinity labeling of a herbicides receptor protein in chloroplast membranes. Proc Nat Acad Sci USA 78:981–986
46. Velthuys BR (1981) Studies into the action of some photosynthetic inhibitor herbicides. FEBS Lett 126:277–281
47. Pallett KE, Dodge AD (1980) Studies into the action of some photosynthetic inhibitor herbicides. J Exp Bot 31:1051–1055
48. Barry P, Young AJ, Britton G (1990) Photodestruction of pigments in higher plants by herbicides action: The effect of DCMU (diuron) on isolated chloroplasts. J Exp Bot 91:123–127
49. De Felipe MR, Golvanao MP, Lucas MM, Lang P, Pozuele JM (1989) Differential effects of isoproturon on the photosynthetic apparatus and yield of two varieties of wheat and *L. Rigidum*. Weed Res 28:85–89

J Mol Model (2012) 18:1447–1454
DOI 10.1007/s00894-011-1168-3

ORIGINAL PAPER

# First principles investigation of oxygen adsorptions on hydrogen–terminated ZnO graphene-like nanosheets

**Benjawan Kaewruksa · Vithaya Ruangpornvisuti**

Received: 29 January 2011 / Accepted: 27 June 2011 / Published online: 15 July 2011


**Abstract** The structures of ZnO graphene–like nanosheets (ZnOGLNS), i.e., ZnO aromatic–like (AL–ZnONS), naphthalene–like (NLL–ZnONS), pyrene–like (PRL–ZnONS), coronene–like (CNL–ZnONS) and circumcoronene–like (CCL–ZnONS) and their oxygen adsorptions were obtained using the B3LYP/LanL2DZ method. Adsorption energies of $O_2$ on AL–ZnONS, NLL–ZnONS, PRL–ZnONS, CNL–ZnONS and CCL–ZnONS are reported. The bond strengths of the most inner Zn–O bonds of ZnOGLNSs are in order: CCL–ZnONS > CNL–ZnONS > PRL–ZnONS. It was found that chemisorptions of $O_2$ occur on the hydride atoms of zinc–hydride in the ZnOGLNSs. Physisorptions of $O_2$ only occurring over the plane of ZnOGLNS were found. All the ZnOGLNSs are oxygen sensitive materials and would be developed to be oxygen sensor based on electrical conductivity.

**Keywords** Adsorption energy · DFT · Oxygen adsorption · ZnOGLNS · ZnO–nanosheets

## Introduction

The crystal faces of zinc oxide (ZnO) have been studied by several experimental and theoretical techniques on adsorptions of $H_2$ [1–5], CO [6], $H_2O$ [7–9], $CO_2$ [10], $NH_3$ [11], NO [12], $NO_2$ [12, 13], $SO_2$ [13] and $N_2O$ [14]. ZnO materials have excellent performance in optics, electronics, photoelectronics [1] and piezoelectricity [2]. It can also be used as transducers and sensors due to their strong piezoelectricity. Different surfaces of the wurtzite ZnO [15, 16] and planar graphite–like structure [17] have been studied. A large number of different ZnO nanostructures such as nanorods [18], nanowires [19–21], nanocombs [22], nanorings [23] and nanotubes [24, 25] have been prepared and studied their properties. One dimensional ZnO nanotube, or nanorods, or nanowires ZnO were found that they have shown much higher sensitivity than polycrystalline ZnO at room temperature because of their higher surface–to–volume ratio and stronger dependence of electrical conductance on the amount of adsorbates [26–33]. However, since ultrawide ZnO nanosheets which have high specific surface area were synthesized, [34] they are expected to have the highest surface–to–volume ratio and can be used as gas sensors. Many synthesized ZnO nanosheets have been expected to be wurtzite or planar graphite–like structures [15–17] but they have never been expected to be graphene–like structures. Recently, the single ZnO monolayer with graphene–like structure was theoretically studied and its elastic, piezoelectric, electronic and optical properties were investigated from the first principles calculations [35]. Nevertheless, its chemical properties such adsorption of fundamental gases have almost never been studied.

In this paper, we have therefore investigated the various sizes of ZnO graphene–like nanosheets (ZnOGLNSs) such as aromatic–like (AL–ZnONS, $Zn_3O_3H_6$), naphthalene–like (NLL–ZnONS, $Zn_5O_5H_8$), pyrene–like (PRL–ZnONS, $Zn_8O_8H_{10}$), coronene–like (CNL–ZnONS, $Zn_{12}O_{12}H_{12}$) and circumcoronene–like (CCL–ZnONS, $Zn_{27}O_{27}H_{18}$) and bond strengths of their Zn–O bonds. Adsorptions of oxygen molecule on the ZnOGLNSs and their electronic properties have also been investigated.

B. Kaewruksa · V. Ruangpornvisuti (✉)
Department of Chemistry, Faculty of Science,
Chulalongkorn University,
Bangkok 10330, Thailand
e-mail: vithaya.r@chula.ac.th

## Computational details

Full optimizations of structures of various sizes of hydrogen–terminated ZnO graphene–like nanosheets (ZnOGLNSs) namely AL–ZnONS, NLL–ZnONS, PRL–ZnONS, CNL–ZnONS and CCL–ZnONS were carried out using density functional theory (DFT) method. The calculations have been performed with hybrid density functional B3LYP, the Becke's three–parameter exchange functional [36] with the Lee–Yang–Parr correlation functional [37], using the Los Alamos LanL2DZ split–valence basis set [38–40]. All calculations were performed with the GAUSSIAN 03 program [41].

Initial structures of interaction configurations of oxygen molecule adsorbed on the studied ZnO nanosheets have been set by lining up the oxygen molecule parallel to the Zn–O bond with all possible directions of oxygen molecule above the hexagonal structure of ZnOGLNSs shown in Fig. 1. The highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) energy gaps ($\Delta E_{GAP}$) of the bare ZnO nanosheets and their adsorptions with oxygen molecule were determined at the B3LYP/LanL2DZ level.

The adsorption energy ($\Delta E_{ads}$) has been computed for oxygen molecule adsorbed on clean surface of ZnOGLNS as follows:

$$\Delta E_{ads} = E_{O_2/ZnOGLNS} - (E_{O_2} + E_{ZnOGLNS}), \quad (1)$$

where $E_{O_2/ZnOGLNS}$ is the total energy of oxygen molecule adsorbed on the ZnOGLNS surface, $E_{O_2}$and $E$ZnOGLNS are the total energies of isolated oxygen molecule and clean surface of ZnOGLNS, respectively.

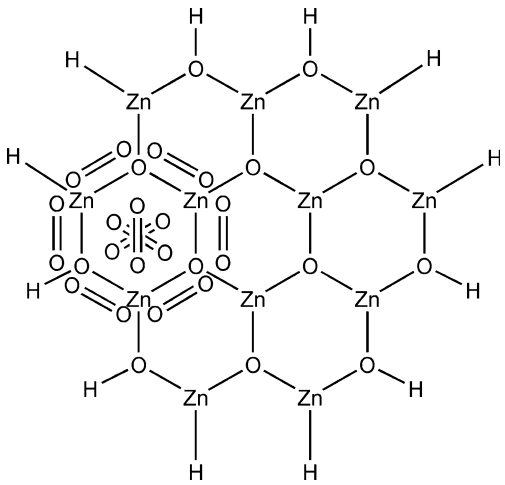


**Fig. 1** Oxygen molecules set as parallel positions over bonds of hexagonal structure of ZnOGLNS and over its center as initial conformations were used in energy optimizations. The CNL–ZnONS ($C_{3h}$ symmetry) is chosen as the representative nanosheet

The adsorption energies of oxygen molecule adsorbed on various sites of the ZnOGLNSs surfaces were obtained at the B3LYP/LanL2DZ level. The basis set superposition error (BSSE) corrections for the adsorption energies were estimated at the B3LYP/LanL2DZ level using the Boys–Bernardi counterpoise (CP) calculations [42–47] and the adsorption energy is denoted as $\Delta E_{ads}^{BSSE}$.

The bond strength (BS in kcal mol$^{-1}$) of Zn–O bond in bare ZnO nanosheets has been computed using the Coulomb's law as following equation:

$$BS = \frac{5.52 \times 10^{-22} \times N_o q_{Zn} q_O}{r_{Zn-O}}, \quad (2)$$

where $N_o$ is an Avogadro number, $q_{Zn}$ and $q_O$ are electronic charges ($e$) of zinc and oxygen atoms, respectively and $r_{Zn\text{-}O}$ (Å) is internuclear distance between zinc and oxygen atoms.

## Results and discussion

### The optimized structures and Zn–O bond strength

The B3LYP/LanL2DZ–optimized structures of the AL–ZnONS, NLL–ZnONS and PRL–ZnONS, CNL–ZnONS and CCL–ZnONS nanosheets obtained by full geometry optimization method are shown in Fig. 2. The molecular symmetries for B3LYP/LanL2DZ–optimized structures of the AL–ZnONS ($C_{3h}$), NLL–ZnONS ($C_{2v}$), PRL–ZnONS ($C_{2v}$), CNL–ZnONS ($C_{3h}$) and CCL–ZnONS ($C_{3h}$) were obtained by assumption of their perfect geometries; their optimized geometries are hardly distorted as shown in Fig. 2. The bond strengths of outer Zn–O bonds of all the ZnOGLNSs are much weaker than inner Zn–O bonds. All the ZnOGLNSs, the bond strengths of their outer Zn–O bonds which zinc are terminated by hydrogen atom, were found within the energy range of −301.88 to −316.65 kcal mol$^{-1}$. The bond strengths of the most inner Zn–O bonds of ZnOGLNSs, computed using Eq. 2 are in order: CCL–ZnONS (−487.10 kcal mol$^{-1}$) > CNL–ZnONS (−485.52 kcal mol$^{-1}$) > PRL–ZnONS (−474.93 kcal mol$^{-1}$).

### Adsorption energies

Geometry configurations of oxygen adsorptions on the rigid structures of the AL–ZnONS, NLL–ZnONS, PRL–ZnONS, CNL–ZnONS and CCL–ZnONS nanosheets were obtained as shown in Fig. 3. Plots of oxygen molecules as minimum energy structures of their adsorptions on ZnOGLNSs as minimum energy structures are shown in Fig. 4. The number of energy minima of oxygen adsorptions on the AL–ZnONS of nine configurations was found for each side

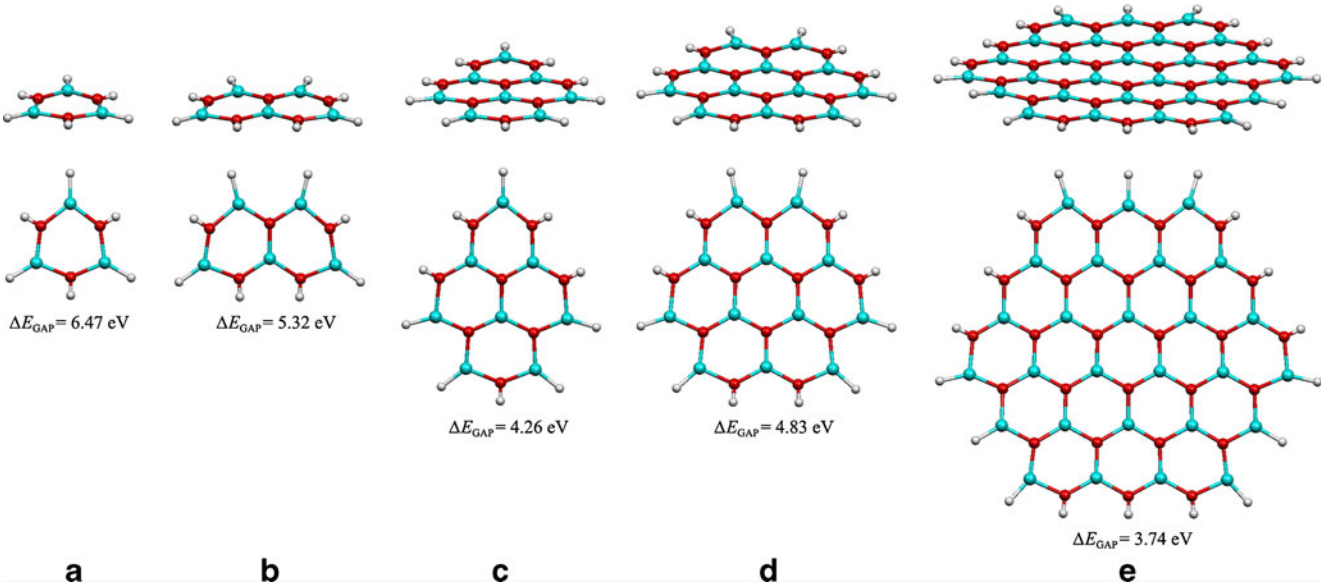



**Fig. 2** The B3LYP/LanL2DZ–optimized structures of ZnO graphene–like nanosheets, (**a**) aromatic–like (AL–ZnONS, $Zn_3O_3H_6$), (**b**) naphthalene–like (NLL–ZnONS, $Zn_5O_5H_8$), (**c**) pyrene–like (PRL–ZnONS, $Zn_8O_8H_{10}$), (**d**) coronene–like (CNL–ZnONS, $Zn_{12}O_{12}H_{12}$) and (**e**) circumcoronene–like (CCL–ZnONS, $Zn_{27}O_{27}H_{18}$). Side and top views are presented at above and below their structures, respectively. Their energy gaps are shown at below their structures

of their molecular planes. These energy minima were obtained from the structure optimizations of interaction configurations within one third of each side of the AL–ZnONS molecular area and oxygen adsorptions over the whole AL–ZnONS molecular area were generated using $C_{3h}$ symmetrical operation. Adsorption energies without ($\Delta E_{ads}$) and with BSSE corrections ($\Delta E_{ads}^{BSSE}$) of oxygen on studied ZnOGLNSs are shown in Table 1. All the adsorption energies mainly discussed in this paper are non-BSSE values unless otherwise indicated. All chemisorptions, adsorption energies with BSSE corrections are decreased within the range of 6.1 to 25.8 %. The most stable configuration of oxygen adsorption on the AL–ZnONS is represented by the configuration of oxygen #1 of which the adsorption energy is −29.87 kcal mol$^{-1}$ ($\Delta E_{ads}^{BSSE} = -25.30$ kcal mol$^{-1}$).

Due to the NLL–ZnONS and PRL–ZnONS nanosheets are in $C_{2v}$ symmetry, the numbers of oxygen adsorptions on each side of their molecular planes are 20 and 26 configurations, respectively, see Fig. 4b and c. Five different types of adsorption sites of oxygen chemisorbed on the NLL–ZnONS are located in vicinity of hydrogen of zinc hydride and the positions of adsorbed oxygen are the oxygen #1 (−23.91 kcal mol$^{-1}$), #2 (−33.09 kcal mol$^{-1}$), #3 (−30.97 kcal mol$^{-1}$), #9 (−29.22 kcal mol$^{-1}$) and #10 (−30.04 kcal mol$^{-1}$). Six different types of adsorption sites of oxygen chemisorbed on the PRL–ZnONS are also located in vicinity of hydrogen of zinc hydride and the positions of adsorbed oxygen are the oxygen #1 (−28.14 kcal mol$^{-1}$), #6 (−27.15 kcal mol$^{-1}$), #7 (−34.53 kcal mol$^{-1}$), #8 (−31.15 kcal mol$^{-1}$), #12 (−33.15 kcal mol$^{-1}$) and #13 (−34.84 kcal mol$^{-1}$).

The CNL–ZnONS and CCL–ZnONS are in $C_{3h}$ symmetry, the numbers of oxygen adsorptions on each side of their molecular planes are 42 and 72 configurations, respectively, see Fig. 4d and e. Three different types of adsorption sites of oxygen chemisorbed on the CNL–ZnONS which are the oxygen #5 (−25.30 kcal mol$^{-1}$), #6 (−34.17 kcal mol$^{-1}$) and #7 (−31.50 kcal mol$^{-1}$). Four different types of adsorption sites of oxygen chemisorbed on the CCL–ZnONS as the positions of oxygen #10 (−9.20 kcal mol$^{-1}$), #11 (−35.26 kcal mol$^{-1}$), #12 (−31.34 kcal mol$^{-1}$) and #13 (−38.83 kcal mol$^{-1}$). It can be concluded that the chemisorption of oxygen on any large ZnOGLNSs occurs on the hydride adsorption site of rim sides of ZnOGLNS. Type number of chemisorption of oxygen on large $C_{3h}$–symmetrical ZnOGLNSs is the number of hydride of zinc hydride ($N_{hydride}$) on each rim side plus two, $N_{hydride}$+2. Figure 4e shows that the oxygen adsorptions on the central ring of the CCL–ZnONS, the adsorbed oxygen are located at positions above the oxygen in CCL–ZnONS hexagonal ring by pointing its oxygen atom toward the nanosheet oxygen with slightly tilt direction. Nevertheless, the oxygen adsorptions occurring on oxygen of the CCL–ZnONS hexagonal rings closed to the central hexagonal ring also show their adsorption position above the nanosheet oxygen, see Fig. 4e. This suggests that oxygen adsorptions on large ZnOGLNSs over the hexagonal ring in central region, adsorbed oxygen molecules located at the position above the nanosheet oxygen as have been expected.

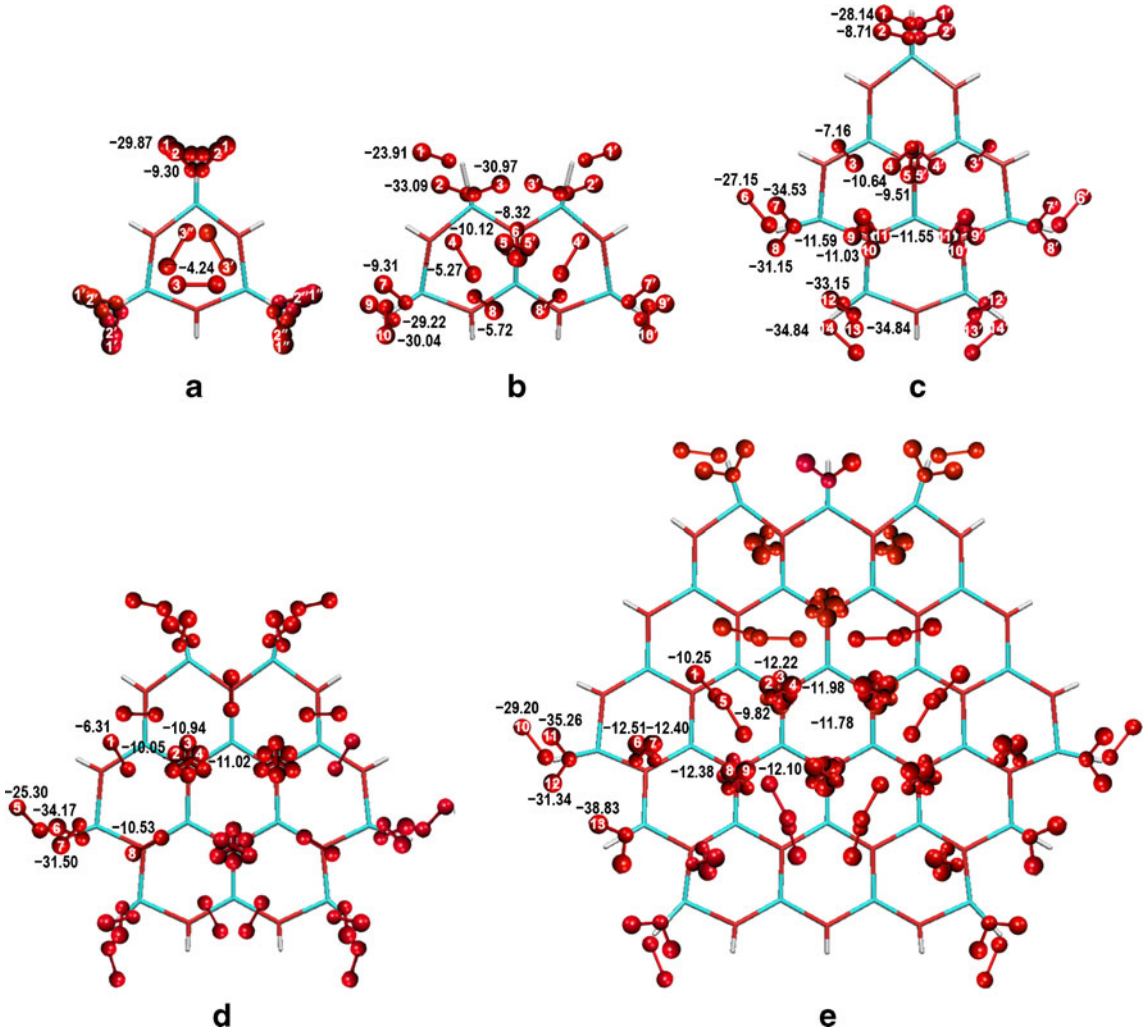



**Fig. 3** The adsorption configurations of oxygen molecule adsorbed on (**a**) the AL–ZnONS ($Zn_3O_3H_6$), (**b**) NLL–ZnONS ($Zn_5O_5H_8$), (**c**) PRL–ZnONS ($Zn_8O_8H_{10}$), (**d**) CNL–ZnONS ($Zn_{12}O_{12}H_{12}$) and (**e**) CCL–ZnONS ($Zn_{27}O_{27}H_{18}$). Bond length and bond distances are in Å

## Atomic charge distribution

NBO charges on oxygen and zinc atoms in the AL–ZnONS, NLL–ZnONS, PRL–ZnONS, CNL–ZnONS and CCL–ZnONS are shown in Fig. 5. It shows that atomic charges on oxygen and zinc atoms in large ZnOGLNSs such as CNL–ZnONS and CCL–ZnONS approach the same (or close to the same) value. Because hydrogen atoms bonded to zinc and to oxygen are hydride and proton, their charges are therefore negative and positive charges, respectively. It also shows that charge distributions over all hydrogen atoms in ZnOGLNSs affect atomic charges of zinc and oxygen atoms of their nanosheets. Atomic charges of zinc and oxygen atoms in large ZnOGLNSs at inner (or central) region with equal charge distribution would be expected. The negative and positive charges on the rim sides of ZnOGLNSs respectively formed due to the charges of hydrides and protons were found.

## Energy gaps

The energy gaps of the clean ZnOGLNSs and their oxygen–adsorption complexes are shown in Table 1. The energy gaps of clean ZnOGLNSs are in order: AL–ZnONS (6.47 eV) > NLL–ZnONS (5.32 eV) > CNL–ZnONS (4.83 eV) > PRL–ZnONS (4.26 eV) > CCL–ZnONS (3.74 eV). In all cases, the energy gaps of clean ZnOGLNSs are higher than those values of their corresponding $O_2$ adsorption complexes, see Table 1. This suggests that the ZnOGLNSs are oxygen sensitive materials and they would be developed to be oxygen sensors based on electrical conductivity. It seems that the energy gap of ZnOGLNSs is an inverse function of its size except PRL–ZnONS which is more active than expectation. However, the energy gap of large ZnO nanosheet should be converted to a single value which can be determined using periodic boundary condition

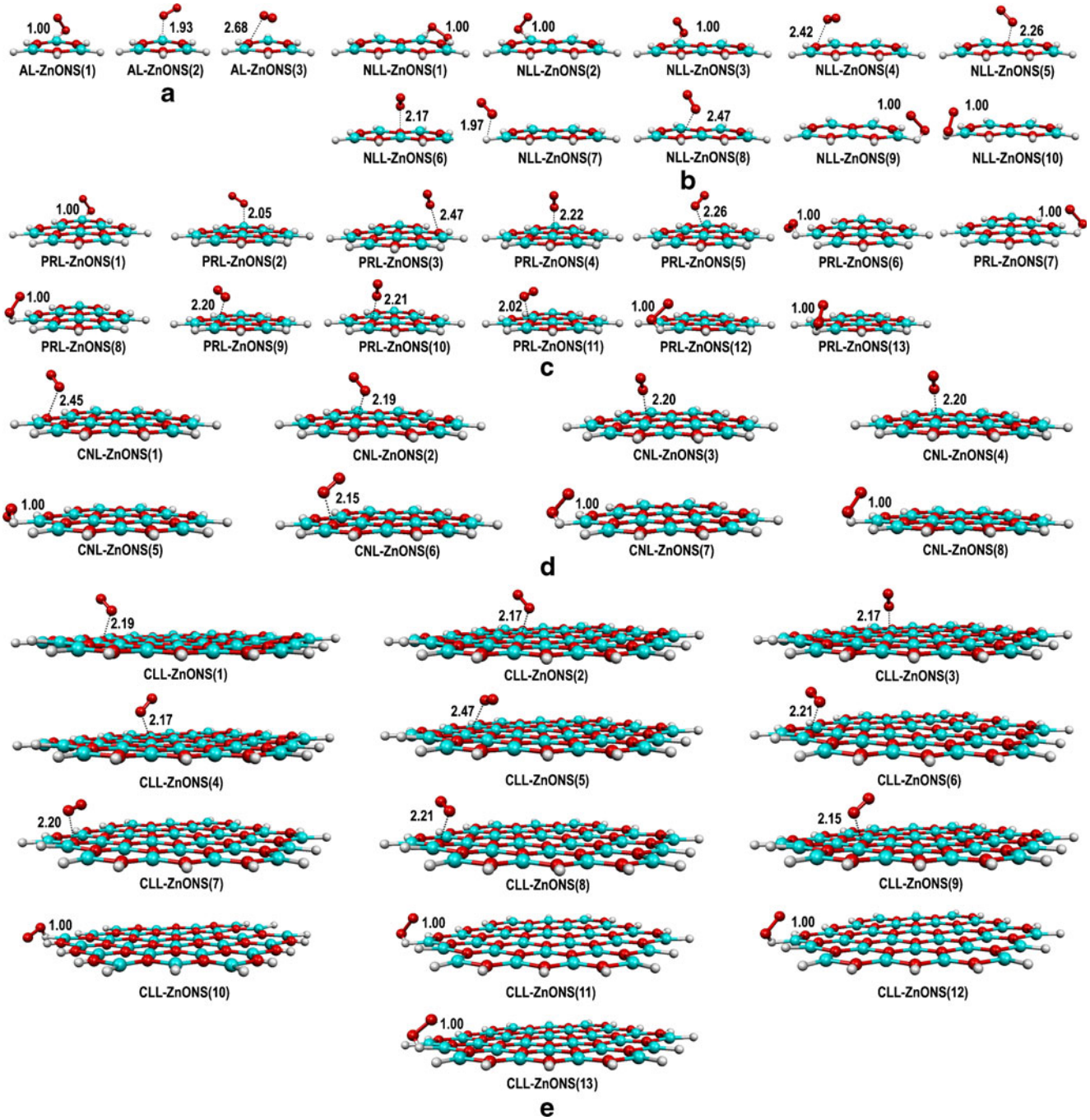



**Fig. 4** Plots of oxygen molecules as minimum energy configurations of their adsorptions on (**a**) the AL–ZnONS, (**b**) NLL–ZnONS, (**c**) PRL–ZnONS, CNL–ZnONS and (**e**) CCL–ZnONS as minimum energy structures. The molecules labeled with numbers represent the oxygen molecule interacting with ZnOGLNSs as representative of molecular symmetry of AL–ZnONS ($C_{3h}$), NLL–ZnONS ($C_{2v}$), PRL–ZnONS ($C_{2v}$), CNL–ZnONS ($C_{3h}$) and CCL–ZnONS ($C_{3h}$). Adsorption energies in kcal mol$^{-1}$ are presented

(PBC) method; band gaps of the wurtzite ZnO of ≈3.4 eV [16] and ≈3.3 eV [48] and the nanocrystalline ZnO of 3.37 eV [49] were found. Nevertheless, the computed band gap of large ZnOGLNS is expected to be a little bit lower than 3.74 eV. It is very rational energy gap for the large hydrogen terminated ZnOGLNS as CCL–ZnONS (3.74 eV) of which the value is a little bit higher than the wurtzite ZnO single crystal because the hydrogen terminated ZnOGLNS, of course for large sheet, must be more stable than the pure ZnO crystal. Inverted values of termination–proton numbers ($N_{TP}$) for AL–ZnONS ($N_{TP}$=6), NLL–ZnONS ($N_{TP}$=8), PRL–ZnONS ($N_{TP}$=10), CNL–ZnONS ($N_{TP}$=12), and CCL–ZnONS ($N_{TP}$=18) plotted against their energy gaps are shown in Fig. 6. It shows that the energy

**Table 1** Adsorption energies (kcal $mol^{-1}$) of oxygen molecule on ZnOGLNSs without ($\Delta E_{ads}$) and with BSSE corrections ($\Delta E_{ads}^{BSSE}$) and energy gaps ($\Delta E_{GAP}$, in eV) of bare ZnOGLNSs and their oxygen adsorption complexes, computed at the B3LYP/LanL2DZ level of theory

| ZnOGLNSs/oxygen adsorption | $\Delta E_{ads}$ | $\Delta E_{ads}^{BSSE}$ | $\Delta E_{GAP}$ |
|---|---|---|---|
| AL–ZnONS: | | | 6.47 |
| $O_2$+AL–ZnONS → $O_2$/AL–ZnONS(1) | −29.87 | −25.30 | 3.50 |
| $O_2$+AL–ZnONS → $O_2$/AL–ZnONS(2) | −9.30 | −5.60 | 2.14 |
| $O_2$+AL–ZnONS → $O_2$/AL–ZnONS(3) | −4.24 | −0.51 | 1.49 |
| NLL–ZnONS: | | | 5.32 |
| $O_2$+NLL–ZnONS → $O_2$/NLL–ZnONS(1) | −23.91 | −19.97 | 2.55 |
| $O_2$+NLL–ZnONS → $O_2$/NLL–ZnONS(2) | −33.09 | −28.61 | 3.41 |
| $O_2$+NLL–ZnONS → $O_2$/NLL–ZnONS(3) | −30.97 | −29.09 | 3.35 |
| $O_2$+NLL–ZnONS → $O_2$/NLL–ZnONS(4) | −5.27 | −1.24 | 1.25 |
| $O_2$+NLL–ZnONS → $O_2$/NLL–ZnONS(5) | −10.12 | −5.94 | 2.01 |
| $O_2$+NLL–ZnONS → $O_2$/NLL–ZnONS(6) | −8.32 | −4.26 | 2.06 |
| $O_2$+NLL–ZnONS → $O_2$/NLL–ZnONS(7) | −9.31 | −5.69 | 2.12 |
| $O_2$+NLL–ZnONS → $O_2$/NLL–ZnONS(8) | −5.72 | −1.99 | 0.87 |
| $O_2$+NLL–ZnONS → $O_2$/NLL–ZnONS(9) | −29.22 | −24.78 | 3.46 |
| $O_2$+NLL–ZnONS → $O_2$/NLL–ZnONS(10) | −30.04 | −25.58 | 3.50 |
| PRL–ZnONS: | | | 4.26 |
| $O_2$+PRL–ZnONS → $O_2$/PRL–ZnONS(1) | −28.14 | −23.67 | 2.73 |
| $O_2$+PRL–ZnONS → $O_2$/PRL–ZnONS(2) | −8.71 | −5.43 | 1.03 |
| $O_2$+PRL–ZnONS → $O_2$/PRL–ZnONS(3) | −7.16 | −7.16 [a] | 0.54 |
| $O_2$+PRL–ZnONS → $O_2$/PRL–ZnONS(4) | −10.64 | −10.64 [a] | 1.16 |
| $O_2$+PRL–ZnONS → $O_2$/PRL–ZnONS(5) | −9.51 | −5.02 | 0.99 |
| $O_2$+PRL–ZnONS → $O_2$/PRL–ZnONS(6) | −27.15 | −22.93 | 2.48 |
| $O_2$+PRL–ZnONS → $O_2$/PRL–ZnONS(7) | −34.53 | −29.79 | 3.33 |
| $O_2$+PRL–ZnONS → $O_2$/PRL–ZnONS(8) | −31.15 | −27.17 | 3.18 |
| $O_2$+PRL–ZnONS → $O_2$/PRL–ZnONS(9) | −11.59 | −6.77 | 1.82 |
| $O_2$+PRL–ZnONS → $O_2$/PRL–ZnONS(10) | −11.03 | −6.98 | 1.92 |
| $O_2$+PRL–ZnONS → $O_2$/PRL–ZnONS(11) | −11.55 | −6.33 | 1.86 |
| $O_2$+PRL–ZnONS → $O_2$/PRL–ZnONS(12) | −33.15 | −28.50 | 3.35 |
| $O_2$+PRL–ZnONS → $O_2$/PRL–ZnONS(13) | −34.84 | −30.19 | 3.32 |
| CNL–ZnONS: | | | 4.83 |
| $O_2$+CNL–ZnONS → $O_2$/CNL–ZnONS(1) | −6.31 | – [b] | 1.06 |
| $O_2$+CNL–ZnONS → $O_2$/CNL–ZnONS(2) | −10.05 | −1.31 | 1.95 |
| $O_2$+CNL–ZnONS → $O_2$/CNL–ZnONS(3) | −10.94 | −4.92 | 1.90 |
| $O_2$+CNL–ZnONS → $O_2$/CNL–ZnONS(4) | −11.02 | −2.46 | 1.92 |
| $O_2$+CNL–ZnONS → $O_2$/CNL–ZnONS(5) | −25.30 | −21.59 | 2.39 |
| $O_2$+CNL–ZnONS → $O_2$/CNL–ZnONS(6) | −34.17 | −25.35 | 3.34 |
| $O_2$+CNL–ZnONS → $O_2$/CNL–ZnONS(7) | −31.50 | −23.89 | 3.27 |
| $O_2$+CNL–ZnONS → $O_2$/CNL–ZnONS(8) | −10.53 | −1.64 | 2.15 |
| CCL–ZnONS: | | | 3.74 |
| $O_2$+CCL–ZnONS → $O_2$/CCL–ZnONS(1) | −10.25 | −10.25 [a] | 1.08 |
| $O_2$+CCL–ZnONS → $O_2$/CCL–ZnONS(2) | −11.78 | −6.19 | 1.41 |
| $O_2$+CCL–ZnONS → $O_2$/CCL–ZnONS(3) | −12.22 | −12.22 [a] | 1.41 |
| $O_2$+CCL–ZnONS → $O_2$/CCL–ZnONS(4) | −11.98 | −7.24 | 1.41 |
| $O_2$+CCL–ZnONS → $O_2$/CCL–ZnONS(5) | −9.82 | −6.52 | 0.90 |
| $O_2$+CCL–ZnONS → $O_2$/CCL–ZnONS(6) | −12.51 | −6.95 | 1.51 |
| $O_2$+CCL–ZnONS → $O_2$/CCL–ZnONS(7) | −12.40 | −6.02 | 1.62 |
| $O_2$+CCL–ZnONS → $O_2$/CCL–ZnONS(8) | −12.38 | −9.30 | 1.46 |
| $O_2$+CCL–ZnONS → $O_2$/CCL–ZnONS(9) | −12.10 | −9.50 | 1.55 |
| $O_2$+CCL–ZnONS → $O_2$/CCL–ZnONS(10) | −29.20 | −26.28 | 2.51 |
| $O_2$+CCL–ZnONS → $O_2$/CCL–ZnONS(11) | −35.26 | −30.18 | 3.22 |
| $O_2$+CCL–ZnONS → $O_2$/CCL–ZnONS(12) | −31.34 | −28.34 | 3.03 |
| $O_2$+CCL–ZnONS → $O_2$/CCL–ZnONS(13) | −38.83 | −35.26 | 2.95 |

[a] BSSE calculation was failed

[b] The high BSSE was found

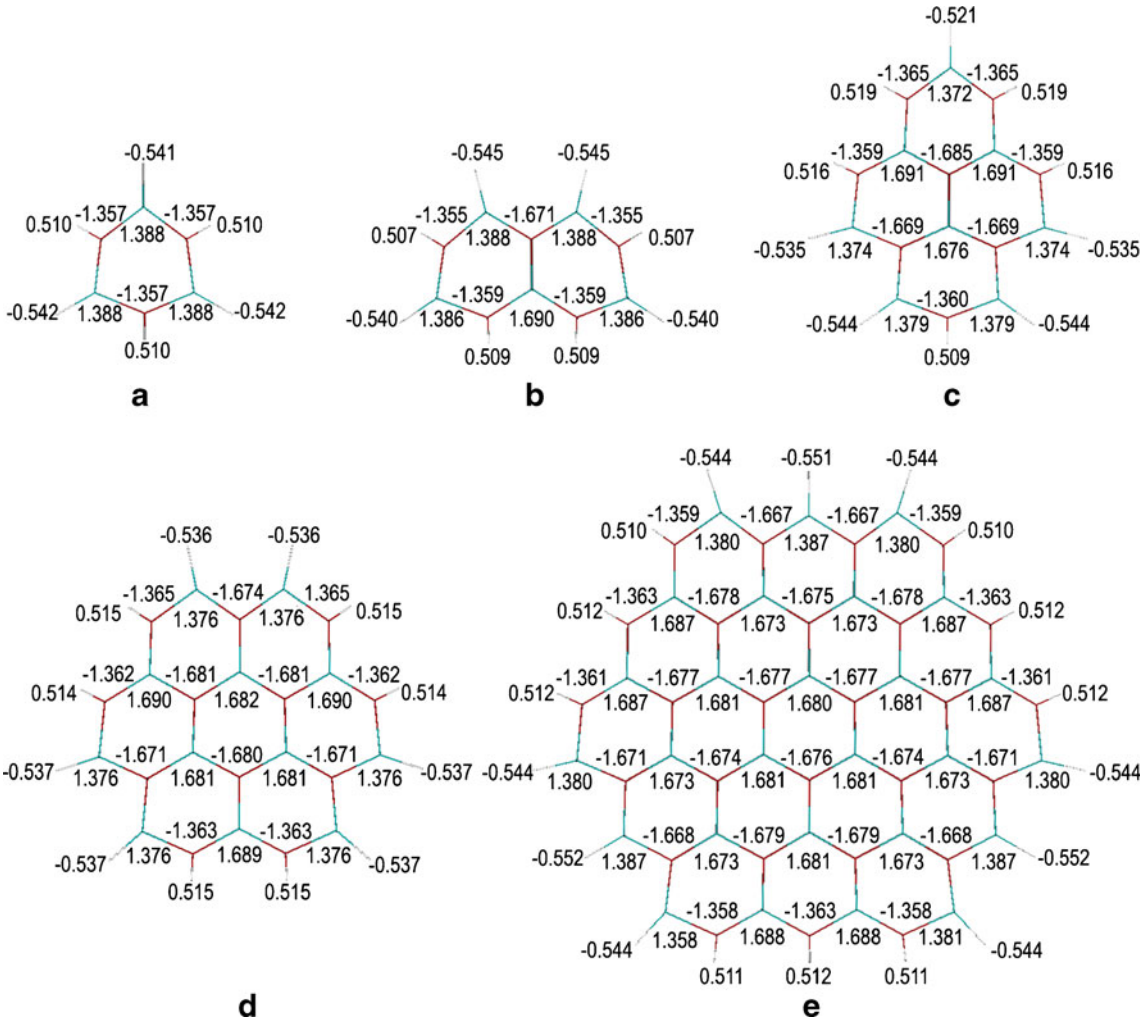



**Fig. 5** NBO charges on oxygen and zinc atoms in the (**a**) AL–ZnONS, (**b**) NLL–ZnONS, (**c**) PRL–ZnONS, (**d**) CNL–ZnONS and (**e**) CCL–ZnONS

gap of large ZnOGLNSs approaches a single value while the inverse value of $N_{TP}$ approaches zero.

## Conclusions

The B3LYP/LanL2DZ–optimized structures of ZnO graphene–like nanosheets of aromatic–like (AL–ZnONS), naphthalene–like (NLL–ZnONS), pyrene–like (PRL–ZnONS), coronene–like (CNL–ZnONS) and circumcoronene–like (CCL–ZnONS) were obtained. The bond strengths of the most inner Zn–O bonds of ZnOGLNSs are in order: CCL–ZnONS (−487.10 kcal mol$^{-1}$) > CNL–ZnONS (−485.52 kcal mol$^{-1}$) > PRL–ZnONS (−474.93 kcal mol$^{-1}$). Potential energies surfaces of oxygen molecule interaction with AL–ZnONS, NLL–ZnONS, PRL–ZnONS, CNL–ZnONS and CCL–ZnONS were

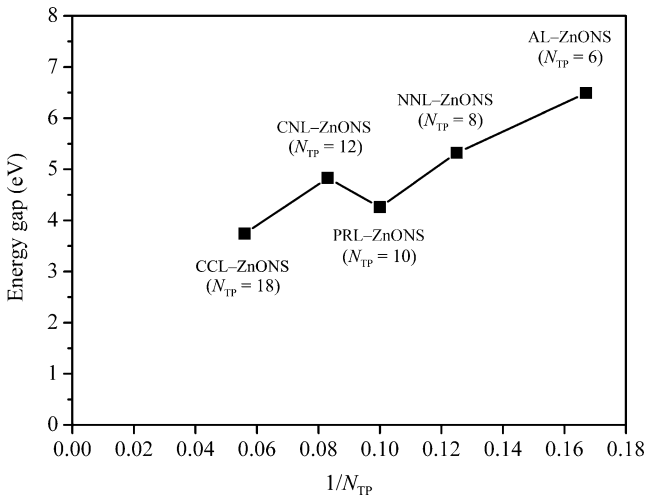



**Fig. 6** Plot of energy gaps of ZnO nanosheets against invert values of their termination-proton numbers ($N_{TP}$)

obtained at the B3LYP/LanL2DZ level of theory. It can be concluded as follows:

- The chemisorption of oxygen on any large ZnOGLNSs occurs on the hydride adsorption site of rim sides of ZnOGLNS.
- The physisorptions of oxygen only occur over the plane of ZnOGLNS.
- Type number of chemisorption of oxygen on large $C_{3h}$–symmetrical ZnOGLNSs is the number of hydride of zinc hydride on each rim side plus two.
- For the oxygen adsorptions in central region of large ZnOGLNSs, the adsorbed oxygen are located at the position above the oxygen in ZnOGLNS hexagonal ring by pointing its oxygen atom toward the nanosheets oxygen with slightly tilt direction.
- The energy gaps of clean ZnOGLNSs are higher than those values of their corresponding $O_2$ adsorption complexes, in all cases.

All the ZnOGLNSs are oxygen sensitive materials and would be developed to be oxygen sensors based on electrical conductivity.

**Acknowledgments** The authors gratefully acknowledge the financial support from the Research, Development and Engineering (RD&E) Fund through the National Nanotechnology Center (NANOTEC), the National Science and Technology Development Agency (NSTDA), Thailand (Project No. P–11–00409) to Chulalongkorn University.

## References

1. Wander A, Harrison NM (2001) J Phys Chem B 105:6191–6193
2. Martins JBL, Andres J, Longo E, Taft CA (1996) J Quantum Chem 57:861–870
3. Martins JBL, Taft CA, Longo E, Andres J (1997) J Mol Struct Theochem 398:457–466
4. Martins JBL, Taft CA, Lie SK, Longo E (2000) J Mol Struct Theochem 528:161–170
5. Becker T, Hovel S, Kunat M, Boas C, Burghaus U, Woll C (2001) Surf Sci 486:L502–L506
6. Gay RR, Nodine MH, Henrich VE, Zeiger HJ, Solomon EI (1980) J Am Chem Soc 102:6752–6761
7. Nakazawa M, Somorjai GA (1995) Appl Surf Sci 84:309–323
8. Nagao M, Kumashiro R, Matsuda T, Kuroda Y (1995) Thermochim Acta 253:221–233
9. Calzolari A, Catellani A (2009) J Phys Chem C 113:2896–2902
10. Hotan W, Gopel W, Haul R (1979) Surf Sci 83:162–180
11. Martins JBL, Longo E, Salmon ODR, Espinoza VAA, Taft CA (2004) Chem Phys Lett 400:481–486
12. Breedon M, Spencer MJS, Yarovsky I (2009) Surf Sci 603:3389–3399
13. Prades JD, Cirera A, Morante JR (2009) Sens Actuators B 142:179–184
14. Spencer MJS, Wong KWJ, Yarovsky I (2010) Mater Chem Phys 119:505–514
15. Na SH, Park CH (2009) J Korean Phys Soc 54:867–872
16. Tang Q, Li Y, Zhou Z, Chen Y, Chen Z (2010) ACS Appl Mater Inter 8:2442–2447
17. Tu ZC, Hu X (2006) Phys Rev B 74:035434–035440
18. Yi GC, Wang C, Park WI (2005) Semicond Sci Technol 20:S22–S34
19. Wang N, Cai Y, Zhang RQ (2008) Mater Sci Eng R 60:1–51
20. Lieber CM, Wang ZL (2007) MRS Bull 32:99–104
21. Jiang ZY, Xu T, Xie ZX, Lin ZW, Zhou X, Xu X, Huang RB, Zheng LS (2005) J Phys Chem B 109:23269–23273
22. Lao CS, Gao PX, Yang RS, Zhang Y, Dai Y, Wang ZL (2006) Chem Phys Lett 417:358–362
23. Peng Y, Bao L (2008) Front Chem China 4:458–463
24. She GW, Zhang XH, Shi WS, Fan X, Chang JC (2007) Electrochem Commun 9:2784–2788
25. She GW, Zhang XH, Shi WS, Fan X, Chang JC, Lee CS, Lee ST, Liu CH (2008) Appl Phys Lett 92:053111/1–3
26. Li QH, Liang YX, Wan Q, Wang TH (2004) Appl Phys Lett 85:6389–6391
27. Fan Z, Wang D, Chang PC, Tseng WY, Lu JG (2004) Appl Phys Lett 85:5923–5925
28. Wang HT, Kang BS, Ren F, Tien LC, Sadik PW, Norton DP, Pearton SJ, Lin J (2005) Appl Phys Lett 86:243503–243505
29. Kang BS, Heo YW, Tien LC, Norton DP, Ren F, Gila BP, Pearton SJ (2005) Appl Phys A 80:1029–1032
30. Rout CS, Kulkarni GU, Rao CNR (2007) J Phys D Appl Phys 40:2777–2782
31. Xu JQ, Chen Y, Chen DY, Shen JN (2006) Sens Actuators B 113:526–531
32. Xu JQ, Chen YP, Li YD, Shen JN (2005) J Mater Sci 40:2919–2921
33. Wang JX, Sun XW, Yang Y, Huang H, Lee YC, Tan OK, Vayssieres L (2006) Nanotechnology 17:4995–4998
34. Park JH, Choi HJ, Choi YJ, Sohn SH, Park JG (2004) J Mater Chem 14:35–36
35. Tu ZC (2010) J Comput Theor Nanosci 7:1182–1186
36. Becke AD (1993) J Chem Phys 98:5648–5652
37. Lee C, Yang W, Parr RG (1988) Phys Rev B 37:785–789
38. Hay PJ, Wadt WR (1985) J Chem Phys 82:270–283
39. Wadt WR, Hay PJ (1985) J Chem Phys 82:284–298
40. Hay PJ, Wadt WR (1985) J Chem Phys 82:299–310
41. Frisch MJ et al (2004) Gaussian03, Revision D.02. Gaussian Inc, CT, Wallingford
42. Boys SF, Bernadi F (1970) Mol Phys 19:553–566
43. Mayer I, Surjan PR (1992) Chem Phys Lett 191:497–499
44. Turi L, Dannenberg JJ (1993) J Phys Chem 97:2488–2490
45. Simon S, Duran M, Dannenberg JJ (1999) J Phys Chem A 103:1640–1643
46. Hamza A, Vibok A, Halasz GJ, Mayer I (2000) J Mol Struct Theochem 427:501–502
47. Makowski M, Raczyska ED, Chmurzyski L (2001) J Phys Chem A 105:869–874
48. Wang ZL, Kong XY, Ding Y, Gao P, Huges WL, Yang R, Zhang Y (2004) Adv Funct Mater 14:943–956
49. Tsukazaki A, Ohtomo A, Onuma T, Ohtani M, Makino T, Sumiya M, Ohtani K, Chichibu SF, Fuke S, Segawa Y, Ohno H, Koinuma H, Kawasaki M (2005) Nat Mater 4:42–46

J Mol Model (2012) 18:1455–1462
DOI 10.1007/s00894-011-1160-y

ORIGINAL PAPER

# UV-spectroscopy, electronic structure and ozonolytic reactivity of sesquiterpenes: a theoretical study

**Shu-Xian Hu · Jian-Guo Yu · Eddy Yongping Zeng**

Received: 17 April 2011 / Accepted: 21 June 2011 / Published online: 16 July 2011


**Abstract** Sesquiterpenes, one of the most important classes of biogenic volatile organic compounds, are potentially significant precursors to secondary organic aerosols (SOAs) in nature. The electronic structure of sesquiterpenes and their reactivity in the ozonolysis reaction were investigated by density functional theory. Results from the CIS calculations combined with an analysis of transition intensities show that the first peaks in the ultraviolet (UV) spectra for saturated and unsaturated isomers are σ–σ* and π–π* transitions, respectively. The UV absorption wavelength and absorbency are dictated by the electronic structures of these compounds. An increase in the number of double bonds and formation of a conjugated system expand the range of absorption in the UV region. An isomer with an endocyclic C = C bond presents weaker UV transition intensity than its corresponding exocyclic isomer. Results from conceptual DFT chemical reactivity indices of isomers suggest that no quantitative linear relationships between the structural changes and their reactivity, such as different degrees of unsaturated C = C double bonds, or the number of substituents attached to the C = C bond were discovered. In the ozonolysis reaction of sesquiterpenes, isomers with larger steric hindrance of substituents or endocyclic C = C bond possess higher chemical reactivity compared to isomers with smaller steric hindrandce or with an exocyclic C = C bond. These results are imperative to a better understanding of SOAs production mechanisms in the troposphere.

**Keywords** Conceptual DFT · Electronic structure · Ozonolytic reactivity · Sesquiterpenes · Ultraviolet spectroscopy

**Electronic supplementary material** The online version of this article (doi:10.1007/s00894-011-1160-y) contains supplementary material, which is available to authorized users.

S.-X. Hu · E. Y. Zeng (✉)
State Key Laboratory of Organic Geochemistry, Guangzhou Institute of Geochemistry, Chinese Academy of Sciences, Guangzhou 510640, China
e-mail: eddyzeng@gig.ac.cn

J.-G. Yu
College of Chemistry, Beijing Normal University, Beijing 100875, China

S.-X. Hu
Graduate School of the Chinese Academy of Sciences, Beijing 100049, China

## Introduction

Secondary organic aerosols (SOAs) are ubiquitous components of atmospheric particulate matter which have attracted much attention [8, 12, 33, 34]. Formation of SOAs involves atmospheric reaction of volatile organic compounds and condensation of corresponding products such as polar multifunctional groups over pre-existing particles via gas-to-particle process. Numerous data indicate that biogenic volatile organic compounds play an important role in the formation of SOAs and show much higher reactivity toward the three electrophilic oxidants, i.e., hydroxyl radical (·OH), ozone ($O_3$) and nitrate radical ($NO_3$), which are present in trace amounts in the troposphere than that of anthropogenic volatile organic compounds [2, 4, 7, 29, 31, 38]. Accumulated laboratory evidence on oxidation of terpenes, a large group of more than 5000 components such as monoterpenes, sesquiterpenes and diterpenes, indicates that reactions with ozone are the most important pathways for formation of aerosols [11]. Sesquiterpenes ($C_{15}H_{24}$), with approximately 100

representative components found in various vegetative species [5], have been demonstrated to be the major and potential precursors for SOAs formation [9, 15, 32].

Ozone monitoring is negatively impacted by large uncertainties with the chemical compositions and characteristics, especially with the highly variable physical properties due to the short tropospheric lifetimes of sesquiterpenes, resulting in great difficulty to quantify the contributions to SOAs. Previous studies using ultraviolet photo electron spectroscopy (UPS) combined with theoretical calculation examined the ionization cross sections and first ionization energies for some sesquiterpenes, but failed to obtain the complete picture of a potentially significant part of atmospheric chemistry and environmental effects of these materials [22–24]. Experimental studies of ozonolysis of sesquiterpenes, which showed that semi-volatile products were produced in gas-phase reactions, were limited to individual configurations [13, 21]. Currently, no laboratory measurements and simulation have been performed to explain the field-based results that sesquiterpenes exerted negative effects on climate change because of their absorption in UV region [14, 17, 32]. In addition, information concerning the relationship between the electronic states and atmospheric reactivity of sesquiterpenes is also scarce.

The main objectives of the present study were to simulate the ultraviolet (UV) spectra and obtain the optical properties and structures of 28 sesquiterpene isomers (Fig. 1) which have been recently identified in the environment, using configuration interaction with single replacement from reference configurations (CIS) [25], time-dependent density functional theory (TDDFT) [30], Zerner's intermediate neglect of differential overlap (ZINDO) [39] methodology. In addition, the isomeric reactivity was predicted based on the conceptual DFT calculated reactivity indices.

## Conceptual density functional theory and reactivity indices

Parr et al. and other researchers [19, 26–28, 37] suggested that DFT methods and reactivity indices are conceptually insightful and practically convenient to predict the chemical reactivity of molecules. It was anticipated that some reactivity indices are equally important for understanding structural changes. Electronic chemical potential ($\mu$), one of the reactivity indices, refers to the escaping tendency of electrons from equilibrium and is identified as the negative of electronegativity ($\chi = -\mu$). Chemical hardness ($\eta$) measures the resistance of $\mu$ to change in total number of electrons. Several studies have employed these parameters to describe the molecular reactivity by obtaining $\eta$ and $\chi$ values from semi-empirical data, the frontier molecular orbital (FMO) energies.

The theoretical definitions of $\mu$ and $\eta$ are provided by the DFT as the first and second derivatives of the electronic energy with respect to the number of electrons ($N$) for a constant external potential $V(\mathrm{r})$:

$$\mu = -\chi = \left(\frac{\partial E}{\partial N}\right)_v, \eta = \left(\frac{\partial^2 E}{\partial N^2}\right) = \left(\frac{\partial \mu}{\partial N}\right)_v. \quad (1)$$

The electrophilicity index ($\omega$) derived by Parr et al. [28] takes the following form:

$$\omega = \frac{\mu^2}{2\eta} = \frac{\chi^2}{2\eta}. \quad (2)$$

Based on the finite difference approach method, the working equations for calculating $\mu$ and $\eta$ are given by

$$\mu = -\frac{I + A}{2} \quad (3)$$

and

$$\eta = I - A \quad (4)$$

where $I$ and $A$ are the first ionization potential and the electron affinity, respectively. Under Koopmans' theorem, the above equations can be expressed as

$$\mu = -\frac{\varepsilon_{\mathrm{HOMO}} + \varepsilon_{\mathrm{LUMO}}}{2} \quad (5)$$

and

$$\eta = \varepsilon_{\mathrm{LUMO}} - \varepsilon_{\mathrm{HOMO}} \quad (6)$$

where $\varepsilon_{\mathrm{HOMO}}$ and $\varepsilon_{\mathrm{LUMO}}$ designate the energies of the highest occupied and lowest unoccupied molecular orbits, respectively. In addition, two new reactivity indices related to electrophilicity and nucleophilicity, as well as electrofugality and nucleofugality in terms of the reactant's first ionization potential ($I$) and electron affinity ($A$), have recently been introduced by Ayers et al. [1]. Electrofugality $\Delta E_{\mathrm{e}}$ is defined as

$$\Delta E_e \equiv I + \omega = \frac{(\mu - \eta)^2}{2\eta}. \quad (7)$$

Nucleofugality $\Delta E_{\mathrm{e}}$ is defined as

$$\Delta E_{\mathrm{n}} \equiv -A + \omega = \frac{(\mu + \eta)^2}{2\eta} \quad (8)$$

These equations assess the quality of electron-fleeing and electron-accepting, respectively. These parameters allow us to measure the inherent quality of an isomer, instead of specific effects in a particular reaction environment [37].

## Computational details

To obtain information on electronic conformations and energies for the sesquiterpene congeners (Fig. 1), the B3LYP hybrid DFT method with 6-31 G (d, p) basis set was used to optimize molecular geometries of all 28 compounds. In these calculations, fully relaxed minimization was used (symmetry restrictions were excluded during the optimization). Afterward, harmonic vibrational frequencies were calculated to affirm that the optimized structures

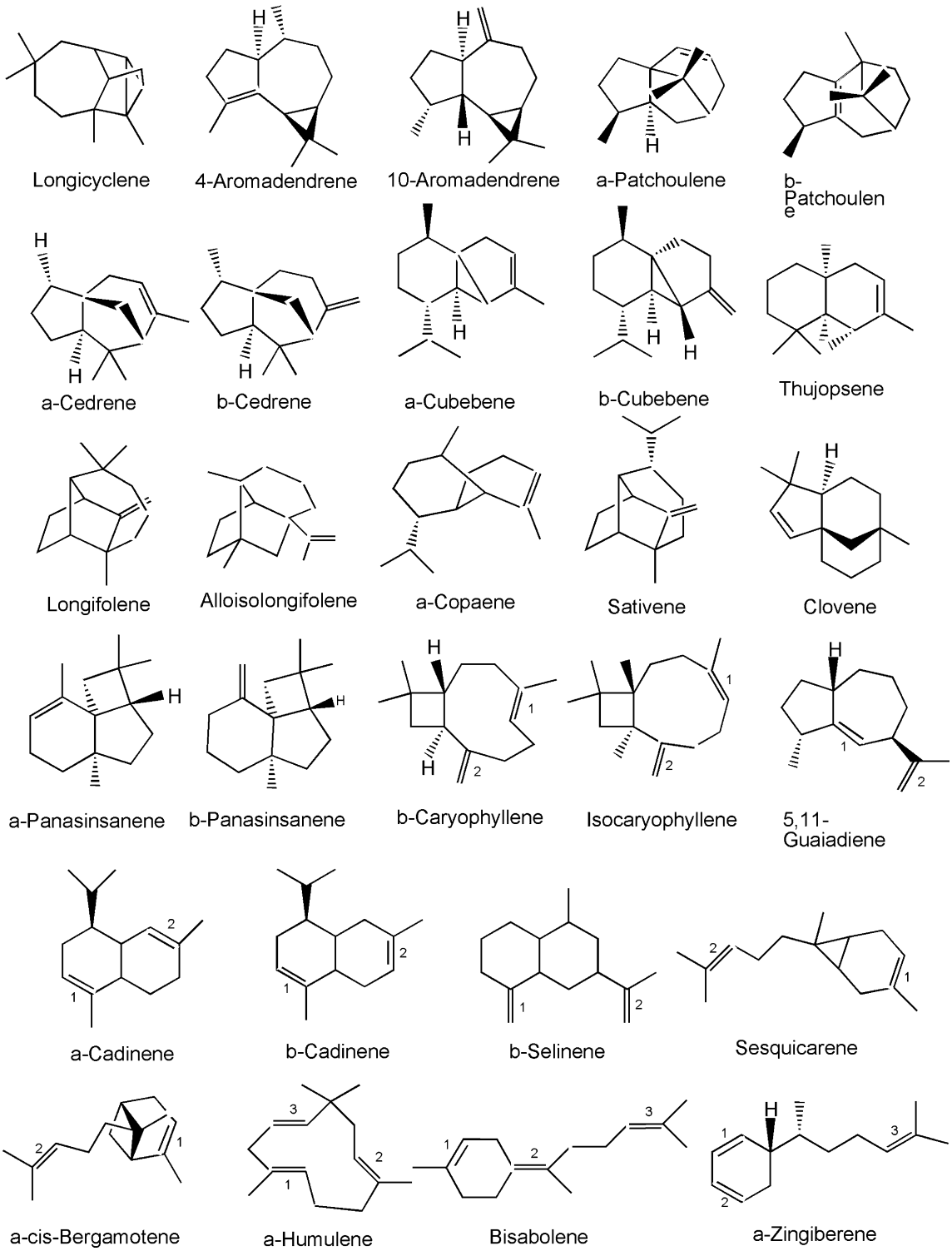



**Fig. 1** The target sesquiterpenes investigated in the present study

are energetic minimums on the potential energy surface (using tight criteria). Spectral calculations were performed using the geometry of a neutral molecule. Excitation energies and oscillator strengths for UV spectra of all target molecules were calculated using CIS, ZINDO and TDDFT within GAUSSIAN03 program [16]. All calculations were performed assuming the target molecules were in the gas phase and without solvent effects.

## Results and discussion

### Comparison of different methods to describe the UV spectra of sesquiterpenes

Although rigorous theoretical interpretation of UV spectra for molecules as large as sesquiterpenes remains quite challenging [4], reproducing a few lower-energy transitions using computational chemical methods is possible. In the present study, the UV spectra of 28 sesquiterpene isomers have been qualitatively simulated using the TDDFT, CIS and ZINDO methods. All these methods are known to be able to produce semi-quantitative agreement with experiments for the transition energies of one-electron excitation.

ZINDO, a semi-empirical method that deals with some aspects of the Hartree-Fock calculation, underestimated the transition energies by more than 20 nm relative to those of isolated double bonds [35]. The transition energies by TD-DFT calculations are systematically overestimated by 1 to 40 nm compared to the standard value for isolated double bonds and by wider margins compared to the values from CIS calculations. Furthermore, the excitation energies of the valence states are relatively well described, but excitation energies with a significant Rydberg character are systematically underestimated by TDDFT calculations [3]. In comparison with ZINDO and TDDFT, CIS provides the best possible description, with almost all values within the standard range [35], of the first peak for these 11 sesquiterpenes for comparison (Table 1). Moreover, CIS calculations obtained reliable characteristic UV spectra for an additional 17 sesquiterpenes (Table 2), and peak assignments were also confirmed through a comparison to available experimental data [20, 22, 24]. This indicates that the lowest energy peak, usually well separated from the rest, originates from the excitation of HOMO-orbitals.

### Theoretical UV spectra of sesquiterpenes

In the spectra of sesquiterpene series with different numbers of C = C bond, the peaks in the UV spectra (Figs. S1, S2, S3 and S4; "S" designates tables and figures in the Supplementary data afterward) are associated with $\pi$-$\pi$* transitions because the electronic state with the strongest oscillator strength has a $\pi$* character, except for long-icyclene which lacks $\pi$ double bond in its electronic structure. Furthermore, several additional weaker transitions both from HOMO to higher virtual orbitals and from lower occupied molecular orbitals to LUMO (in the range of 4–11 eV) were obtained by our calculations (Table S1), and these transitions may account for the fine structures observed around the maximum peak in the sesquiterpenes spectra.

**Table 1** Comparison of the UV-excitation energies (nm) of sesquiterpenes based on the TD-B3LYP/6-31 G(d,p) and ZINDO levels and the standard value ($\lambda_{max}$=180 nm) [35]

| Congeners | $\lambda_{TD}$ | $\lambda_{ZINDO}$ | $\Delta\lambda_{T\text{-}180}$ | $\Delta\lambda_{Z\text{-}180}$ |
|---|---|---|---|---|
| 4-Aromadendrene | 202 | 232 | 22 | 52 |
| Longifolene | 181 | 200 | 1 | 20 |
| $\alpha$-Cedrene | 184 | 212 | 4 | 32 |
| $\alpha$-cis-Bergamotene | 195 | 219 | 15 | 39 |
| $\beta$-Caryophyllene | 184 | 214 | 4 | 34 |
| Isocaryophyllene | 220 | 231 | 40 | 51 |
| Guaiadiene | 203 | 215 | 23 | 35 |
| $\alpha$-Humulene | 222 | 232 | 42 | 52 |
| $\alpha$-Copaene | 192 | 217 | 12 | 37 |
| $\alpha$-Panasinsanene | 182 | 211 | 2 | 31 |
| $\beta$-Panasinsanene | 183 | 199 | 3 | 19 |

Previous UPS studies on methyl-substituted ethenes and terpenoids revealed that $\pi$-excitation energy is reduced with increasing number of substituents attached to C = C double bond [24]. It should be noted that, however, not all isomers with higher-degree of substitution of C = C bonds have lower values of $\pi$-excitation energy than those with lower degree of substitution. For example, the difference between $\beta$-cedrene and $\beta$-cubebene, both containing a two-substituted C = C bond, is 0.28 eV (Table 2). This suggests that there is no simple relationship between the degree of substitution of the C = C bond and the $\pi$-transition energy. Therefore, in the discussions of the relative stabilization of $\pi$-orbital, it is more accurate to compare a pair of isomers with the same basic structural skeleton rather than with different positions of the C = C double bonds. In other words, comparison of the first excitation energy and UV band intensities is made between isomer pairs: 4-aromadendrene (7.28 eV) and 10-aromadendrene (7.98 eV); $\alpha$-cedrene (7.67 eV) and $\beta$-cedrene (7.99 eV); and $\alpha$-cubebene (7.33 eV) and $\beta$-cubebene (7.71 eV)} (Table 2). The evidence in support of the alkylation as the source of $\pi$-orbital destabilization can only be found in these pairs. The $\pi$-orbitals have higher transition energy in exo isomers than in corresponding endo isomers. The positive deviation of $\pi$-transition energy to the endo isomers ranges from 0.70 eV in aromadendrene pairs to 0.14 eV in

**Table 2** UV- excitation energy ($E_t$; eV), assignments, UV-spectra maxima ($\lambda$; nm) and transition intensity (f) of sesquiterpenes

| Species | $E_t$ | Assignment | $\lambda$ | f |
|---|---|---|---|---|
| Longicyclene | 10.51 | σ-σ* | 117.98 | 0.0017 |
| 4-Aromadendrene | 7.28 | π-π* | 170.33 | 0.5853 |
| 10-Aromadendrene | 7.98 | π-π* | 155.45 | 0.6495 |
| α-Patchoulene | 7.53 | π-π* | 164.65 | 0.4196 |
| β-Patchoulene | 7.34 | π-π* | 168.81 | 0.3987 |
| α-Cedrene | 7.66 | π-π* | 161.79 | 0.4348 |
| β-Cedrene | 7.99 | π-π* | 155.22 | 0.8103 |
| α-Cubebene | 7.33 | π-π* | 169.13 | 0.3212 |
| β-Cubebene | 7.71 | π-π* | 160.91 | 0.9 |
| Longifolene | 7.79 | π-π* | 159.23 | 0.5459 |
| Alloisolongifolene | 7.98 | π-π* | 155.33 | 0.55 |
| α-Panasinsanene | 7.73 | π-π* | 160.29 | 0.6861 |
| β-Panasinsanene | 7.88 | π-π* | 157.33 | 0.6278 |
| Sativene | 7.87 | π-π* | 157.62 | 0.6033 |
| Clovene | 7.9 | π-π* | 156.99 | 0.5505 |
| α-Copaene | 7.39 | π-π* | 167.69 | 0.3354 |
| Thujepsene | 7.44 | π-π* | 166.72 | 0.3585 |
| α-cis-Bergamotene | 7.4 | $\pi$-$\pi^*_1$ | 167.61 | 0.3481 |
| | 7.97 | $\pi$-$\pi^*_2$ | 155.53 | 1.0178 |
| Sesquicarene | 7.95 | $\pi$-$\pi^*_1$ | 156.03 | 0.4299 |
| | 7.99 | $\pi$-$\pi^*_2$ | 155.08 | 1.1384 |
| β-Selinene | 8.09 | $\pi$-$\pi^*_1$ | 153.33 | 0.7242 |
| | 8.17 | $\pi$-$\pi^*_2$ | 151.71 | 0.7555 |
| β-Caryophyllene | 7.7 | $\pi$-$\pi^*_1$ | 160.93 | 0.7876 |
| | 7.89 | $\pi$-$\pi^*_2$ | 157.09 | 0.4271 |
| Isocaryophyllene | 7.33 | $\pi$-$\pi^*_1$ | 169.24 | 0.4041 |
| | 7.76 | $\pi$-$\pi^*_2$ | 159.84 | 0.5794 |
| Guaiadiene | 7.27 | $\pi$-$\pi^*_1$ | 170.55 | 0.7022 |
| | 8.24 | $\pi$-$\pi^*_2$ | 150.46 | 0.6329 |
| α-Cadinene | 7.64 | $\pi$-$\pi^*_1$ | 162.19 | 0.0935 |
| | 7.83 | $\pi$-$\pi^*_2$ | 158.34 | 1.1315 |
| β-Cadinene | 7.68 | $\pi$-$\pi^*_1$ | 161.44 | 0.3459 |
| | 8.04 | $\pi$-$\pi^*_2$ | 154.23 | 0.7628 |
| α-Humulene | 7.22 | $\pi$-$\pi^*_1$ | 171.7 | 0.1897 |
| | 7.79 | $\pi$-$\pi^*_2$ | 159.11 | 0.6806 |
| | 7.83 | $\pi$-$\pi^*_3$ | 158.29 | 0.8969 |
| Bisabolene | 7.47 | $\pi$-$\pi^*_2$ | 165.97 | 1.2055 |
| | 7.94 | $\pi$-$\pi^*_1$ | 156.19 | 0.294 |
| | 8.08 | $\pi$-$\pi^*_3$ | 153.39 | 0.9521 |
| α-Zingiberene | 5.6 | $\pi$-$\pi^*_{1,2}$ | 221.36 | 0.1838 |
| | 7.97 | $\pi$-$\pi^*_3$ | 155.57 | 1.1995 |
| | 8.87 | $\pi$-$\pi^*_{1,2}$ | 139.77 | 0.5712 |

panasinanene pairs (Table 2). Based on the comparison of the structures and transition energies for all unsaturated isomers of sesquiterpene, the number of substituents and the type of C = C bonds have effects on the excitation energy within the same molecule or within the same skeletal structure.

Some conclusions can be drawn from the above results: i) endo C = C double bonds have higher reactivity than exo double bonds, which was confirmed in previous studies [6, 22], and thus their inductive contribution to SOAs is higher than that by the exo double bonds; ii) increasing number of substituents within the same basic structural skeleton lowers the electronic stability and broadens the UV absorption of isomers; and iii) an increase in the number of double bonds and formation of a conjugated system expand the range of absorption region, thus indirectly posing positive impact on the absorption of radiation.

## Implications for transition intensity

The transition intensities for the structural pairs with the same skeletons, such as 4-aromadendrene and 10-aromadendrene, α-cedrene and β-cedrene, and α-cubebene and β-cubebene at 0.5853 and 0.6495, 0.4348 and 0.8103, and 0.3212 and 0.9000, respectively (Table 2) based on our calculations, are worth further notice. These results indicate that the exo-double bond in each pair has relatively stronger transition intensity than the endo-double bond. However, the pair of α-panasinsanene (0.6861) and β- panasinsanene (0.6278) (Table 2) is an exception, i.e., the peak of exo-one is weaker than that of endo-one. These data suggest that different intensities may be related to the rigidity of the ring systems, particularly because the endo-double bond somewhat increases the rigidity of a ring. Further evidence supporting this explanation can be found in the transition intensities of sesquicarene which have two peaks corresponding to two types of double bonds. For example, the transition of exo- $\pi_2$-bond (1.1384) is more intense than that of the endo-$\pi_1$-bond (0.4299) (Table 2), although both have three-alkyl groups attached to the C = C bond. Therefore, an isomer with an endocyclic C = C bond presents weak UV transition intensity compared to its corresponding exocyclic one, disregarding any substitution effect.

## Conceptual DFT descriptors of sesquiterpene isomers

The rate for ozonolysis of sesquiterpenes, a major process to produce SOAs, is governed by the LUMO of ozone and the HOMO of the sesquiterpenes [3, 10, 18, 36]. Conceptual DFT indexes ($\mu$, $\eta$, $\omega$, $\Delta E_e$ and $\Delta E_n$) derived from basic density functional theory have been used to investigate static or dynamic inorganic and organic systems. Because $\eta$ is resistant to deformation or change in donating or withdrawing electrons, $\mu$ measures the escaping tendency of electrons from regions with higher $\mu$ to areas with lower $\mu$ and $\omega$ is served as a measure of the stabilization energy when atoms or molecules drawing maximal electrons, it is

clear that if a molecule has lower $\eta$ or has a higher absolute value of $\mu$ or $\omega$, then the molecule will have higher reactivity and nucleophilicity.

As electron donors in electron-transfer interactions with ozone in tropospheric reactions, sesquiterpenes act as reductants with strong nucleophilicity. Initially, higher unsaturated isomers were expected to possess lower chemical hardness and higher electrophilicity index. However, the results from the present study show that the values of all electronic property descriptors have weak correlation to their degree of unsaturation from the calculation at the B3LYP level (Table 3). For example, isocaryophyllene with two C = C bonds has lower $\eta$ and higher $\omega$ values than bisabolene with three C = C bonds (the deviations are −0.13 and 0.10 eV, respectively). It should be noted that, however, $\alpha$-zingiberene ($\eta$, $|\mu|$ and $\omega$ are 5.06, 2.88 and 0.82 eV, respectively), with three unsaturated bonds, is the most reactive among all isomers under investigation largely due to its high number of double bonds. Also note that saturated isomers are less reactive in ozonolysis reaction than unsaturated isomers. For example, longicyclene has the largest $\eta$ value (8.63 eV) and smallest $\omega$ (0.32 eV) among all isomers, indicating its weak reactive ability in electron-transfer reaction. Overall, isomer with higher degree of unsaturation possesses higher reactivity in ozonation, which is consistent with the results of a previous experimental study [24].

Moreover, an interesting result arising from our calculations as well as from previous work [24] is that the intramolecular position of the double bond within the sesquiterpene skeleton does not govern chemical hardness ($\eta$). For example, $\beta$-caryophyllene and sesquicarene have different molecular skeletons, although their $\mu$, $\eta$, $\omega$, $\Delta E_n$ and $\Delta E_e$ values are similar. In addition, no direct correlation has been observed between the values of reactive descriptors and the number of alkyl substitutions of unsaturated bonds or the substitution pattern, which is contradictory to the findings of several previous studies

**Table 3** Calculated conceptual DFT reactivity indices (eV)[a] for sesquiterpenes at the B3LYP/6-31 G (d, p) level

| Species | $E_{HOMO}$ | $E_{LUMO}$ | A | I | μ | η | ω | $\Delta E_n$ | $\Delta E_e$ |
|---|---|---|---|---|---|---|---|---|---|
| Longicyclene | −6.67 | 1.96 | −1.96 | 6.67 | −2.35 | 8.63 | 0.32 | 2.28 | 6.99 |
| $\alpha$-Panasinsanene | −5.96 | 0.9 | −0.9 | 5.96 | −2.53 | 6.86 | 0.47 | 1.36 | 6.43 |
| $\beta$-Panasinsanene | −6.23 | 0.71 | −0.71 | 6.23 | −2.76 | 6.94 | 0.55 | 1.26 | 6.78 |
| $\alpha$-Cubebene | −5.69 | 0.9 | −0.9 | 5.69 | −2.39 | 6.58 | 0.44 | 1.33 | 6.12 |
| $\beta$-Cubebene | −5.93 | 0.73 | −0.73 | 5.93 | −2.6 | 6.67 | 0.51 | 1.24 | 6.44 |
| $\alpha$-Patchoulene | −5.85 | 0.9 | −0.9 | 5.85 | −2.48 | 6.75 | 0.45 | 1.35 | 6.3 |
| $\beta$-Patchoulene | −5.71 | 0.93 | −0.93 | 5.71 | −2.39 | 6.64 | 0.43 | 1.36 | 6.15 |
| $\alpha$-Cedrene | −6.01 | 0.87 | −0.87 | 6.01 | −2.57 | 6.88 | 0.48 | 1.35 | 6.49 |
| $\beta$-Cedrene | −6.34 | 0.71 | −0.71 | 6.34 | −2.82 | 7.05 | 0.56 | 1.27 | 6.9 |
| 10-Aromadendrene | −6.2 | 0.79 | −0.79 | 6.2 | −2.71 | 6.99 | 0.52 | 1.31 | 6.73 |
| 4-Aromadendrene | −5.69 | 0.71 | −0.71 | 5.69 | −2.49 | 6.39 | 0.48 | 1.19 | 6.17 |
| Clovene | −6.29 | 0.82 | −0.82 | 6.29 | −2.73 | 7.1 | 0.53 | 1.34 | 6.81 |
| $\alpha$-Copaene | −5.88 | 0.79 | −0.79 | 5.88 | −2.54 | 6.67 | 0.49 | 1.27 | 6.36 |
| Sativene | −6.26 | 0.71 | −0.71 | 6.26 | −2.78 | 6.97 | 0.55 | 1.26 | 6.81 |
| Thujepsene | −5.9 | 0.73 | −0.73 | 5.9 | −2.58 | 6.64 | 0.5 | 1.24 | 6.41 |
| Alloisolongifolene | −6.34 | 0.76 | −0.76 | 6.34 | −2.79 | 7.1 | 0.55 | 1.31 | 6.89 |
| Longifolene | −6.23 | 0.65 | −0.65 | 6.23 | −2.79 | 6.88 | 0.56 | 1.22 | 6.8 |
| $\beta$-Caryophyllene | −6.34 | 0.71 | −0.71 | 6.34 | −2.82 | 7.05 | 0.56 | 1.27 | 6.9 |
| Isocaryophyllene | −5.93 | 0.44 | −0.44 | 5.93 | −2.75 | 6.37 | 0.59 | 1.03 | 6.52 |
| $\alpha$-Cadinene | −5.99 | 0.73 | −0.73 | 5.99 | −2.63 | 6.72 | 0.51 | 1.25 | 6.5 |
| $\beta$-Cadinene | −6.1 | 0.84 | −0.84 | 6.1 | −2.63 | 6.94 | 0.5 | 1.34 | 6.59 |
| Guaiadiene | −6.07 | 0.49 | −0.49 | 6.07 | −2.79 | 6.56 | 0.59 | 1.08 | 6.66 |
| $\beta$-Selinene | −6.45 | 0.73 | −0.73 | 6.45 | −2.86 | 7.18 | 0.57 | 1.3 | 7.02 |
| Sesquicarene | −6.07 | 0.93 | −0.93 | 6.07 | −2.57 | 6.99 | 0.47 | 1.4 | 6.54 |
| $\alpha$-cis-Bergamotene | −5.9 | 0.76 | −0.76 | 5.9 | −2.57 | 6.67 | 0.5 | 1.26 | 6.4 |
| $\alpha$-Humulene | −5.8 | 0.6 | −0.6 | 5.8 | −2.6 | 6.39 | 0.53 | 1.13 | 6.32 |
| Bisabolene | −5.77 | 0.73 | −0.73 | 5.77 | −2.52 | 6.5 | 0.49 | 1.22 | 6.26 |
| $\alpha$-Zingiberene | −5.41 | −0.35 | 0.35 | 5.41 | −2.88 | 5.06 | 0.82 | 0.47 | 6.24 |

[a] $E_{HOMO}$: the highest occupied molecular orbital (HOMO) energy; $E_{LUMO}$: the lowest unoccupied molecular orbital (LUMO) energy; $I$: first ionization energy; $A$: electron affinity; $\mu$: Electronic chemical potential; $\eta$: chemical hardness; $\omega$: electrophilicity index; $\Delta E_n$: nucleofugality; $\Delta E_e$: electrofugality

[22–24]. Successive alkyl substitution results in no gradual and systematic increases in all electronic property indexes, and the magnitude of increase or decrease depends largely on the specific position of a substituent. For instance, $\alpha$-patchoulene, which has three alkyl groups attached to the C = C bond, has higher $\eta$ and $\omega$ values than $\beta$-patchoulene (the positive deviations are 0.11 and 0.02 eV, respectively). Although $\beta$-caryophyllene and isocaryophyllene are cis-trans isomers with the same alkyl substituents, their $\mu$, $\eta$, $\omega$, $\Delta E_n$ and $\Delta E_e$ values are quite different (the deviations are −0.09, 0.86, -0.03 and 0.24 eV, respectively). Rather, the type of double bonds determines the values of $\mu$, $\eta$, $\omega$, $\Delta E_n$ and $\Delta E_e$. Congeners with an endo C = C bond possess higher chemical reactivity than their corresponding isomers with an exo C = C bond. This can be demonstrated by comparing the chemical hardness values of $\alpha$-panasinsanene ($\eta$=6.86 eV) with $\beta$-panasinsanene ($\eta$=6.94 eV) and $\alpha$-cubebene ($\eta$= 6.58 eV) with $\beta$-cubebene ($\eta$=6.67 eV).

Additionally, the size of a substituent dictates the values of $\mu$, $\eta$, $\omega$, $\Delta E_n$ and $\Delta E_e$. A molecule with a double bond group attached with a large-size alkyl substituent is more reactive than other isomers. For example, the steric hindrance in longifolene, where two large substituents (i.e., 1,4,8,8-tetramethyldecahydroazulenyl and 4,4,8,8-tetramethyldecahydroazulenyl; Fig. 1) attach to the C = C bond, is larger than that in alloisolongifolene; therefore, longifolene has a smaller $\eta$ value (the negative deviation is 0.22 eV), i.e., it is less reactive than alloisolongifolene. To examine the efficiency of conceptual DFT on the computational predictions, the relative enthalpies of formation ($\Delta H_{rel}$), in addition to rate constants ($k$) of sesquiterpene isomers, were referenced to comparison with chemical hardness ($\eta$) in our work. The results (Tables S2) indicate the efficiency of conceptual DFT and their consistency with available data in the literature (see the Supplementary material for detailed discussions).

## Conclusions

The present theoretical study obtains the relationship between the electronic states and atmospheric reactivity of sesquiterpenes by examining their UV spectra. A congener with an exo C = C bond or containing a large-size alkyl substituent attached to any of the C = C bonds (if more than one) and its degradation product in the atmosphere contribute more to the formation of SOAs than other isomers. Because the reaction of sesquiterpenes with $O_3$ is the dominant loss process in the atmosphere, the results obtained from the present study are useful for predicting intermediate or final products of secondary and subsequent reactions such as SOAs formation from sesquiterpenes.

**Acknowledgments** Financial support from the Earmarked Fund of the State Key Laboratory of Organic Geochemistry (No. SKLOG2009A04) and the National Natural Science Foundation of China (Nos. 40821003, 20733002 and 20873008) is greatly appreciated. The authors gratefully thank Dr. Liang Peng of the Department of Chemistry and Environment, South China Normal University for useful suggestions and stimulated discussions. This is contribution No. IS-1357 from GIGCAS.

## References

1. Ayers PW, Anderson JSM, Rodriguez JI, Jawed Z (2005) Indices for predicting the quality of leaving groups. Phys Chem Chem Phys 7:1918–1925
2. Blanco MB, Bejan I, Barnes I, Wiesen P, Teruel MA (2009) The Cl-initiated oxidation of CH3C(O)OCH = CH2, CH3C(O) OCH2CH = CH2, and CH2 = CHC(O)O(CH2)3CH3 in the troposphere. Environ Sci Pollut Res 16:641–648
3. Casida ME, Jamorski C, Casida KC, Salahub DR (1998) Molecular excitation energies to high-lying bound states from time-dependent density-functional response theory: characterization and correction of the time-dependent local density approximation ionization threshold. J Chem Phys 108:4439–4450
4. Chuong B, Davis M, Edwards M, Stevens PS (2002) Measurements of the kinetics of the OH + a-Pinene and OH + b-Pinene reactions at low pressure. Int J Chem Kinet 34:300–308
5. Fraga BM (2002) Natural sesquiterpenoids. Nat Prod Rep 19:650–672
6. Fu PQ, Kawamura K, Chen J, Barrie LA (2009) Isoprene, monoterpene, and sesquiterpene oxidation products in the high arctic aerosols during late winter to early summer. Environ Sci Technol 43:4022–4028
7. Grayson DH (1998) Monoterpenoids. Nat Prod Rep 15:439–475
8. Griffin RJ, Cocker DR III, Flagan RC, Seinfeld JH (1999) Organic aerosol formation from the oxidation of biogenic hydrocarbons. J Geophys Res 104:3555–3567
9. Griffin RJ, Cocker DR, Seinfeld JH (1999) Estimate of global atmospheric organic aerosol from oxidation of biogenic hydrocarbons. Geophys Res Lett 26:2721–2724
10. Griffin RJ, Cocker DR III, Seinfeld JH (1999) Incremental aerosol reactivity: application to aromatic and biogenic hydrocarbons. Environ Sci Technol 33:2403–2408
11. Heiden AC, Hoffmann T, Kahl J, Kley D, Klockow D, Langebartels C, Mehlhorn H, Sandermann H, Schraudner M, Schuh G, Wildt J (1999) Emission of volatile prganic compounds from ozone-exposed plants. Ecol Appl 9:1160–1167
12. Jacobson MC, Hansson H-C, Noone KJ, Charlson RJ (2000) Orangic atmospheric aerosols: review and state of the science. Rev Geophys 38:267–294
13. Jaoui M, Sexton KG, Kamens RM (2004) Reaction of $\alpha$-cedrene with ozone: mechanism, gas and particulate products distribution. Atmos Environ 38:2709–2725
14. Kanakidou M, Seinfeld JH, Pandis SN, Barnes I, Dentener FJ, Facchini MC, Van Dingenen R, Ervens B, Nenes A, Nielsen CJ, Swietlicki E, Putaud JP, Balkanski Y, Fuzzi S, Horth J, Moortgat GK, Winterhalter R, Myhre CEL, Tsigaridis K, Vignati E, Stephanou EG, Wilson J (2005) Organic aerosol and global climate modelling: a review. Atmos Chem Phys 5:1053–1123
15. Kasali AA, Ekundayob O, Paula C, Konig WA (2002) epi-Cubebanes from Solidago canadensis. Phytochemistry 59:805–810

16. Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C, Pople JA (2004) Gaussian 03, Revision D.01. Gaussian Inc, Wallingford, CT
17. Kubala D, Drage EA, Al-Faydhi AME, Kocísek J, Papp P, Matejcík V, Mach P, Urban J, Limao-Vieira P, Hoffmann SV, Matejcík S, Mason NJ (2009) Electron impact ionisation and UV absorption study of α- and β-pinene. Int J Massspectrom 280:169–173
18. Lee A, Goldstein AH, Keywood MD, Gao S, Varutbangkul V, Bahreini R, Ng NL, Flagan RC, Seinfeld JH (2006) Gas-phase products and secondary aerosol yields from the ozonolysis of ten different terpenes. J Geophys Res 111(D07):D07302
19. Liu SB (2009) Conceptual density functional theory and some recent developments. Acta Phys Chim Sin 25:590–600
20. Nauduri D, Greenberg A (2009) Calculated ionization energies for a series of sesquiterpenes: comparisons with experimental vertical ionization energies and comments on related structure–activity relationships (SARs). Struct Chem 20:417–421. doi:10.1007/s1122400994312
21. Nguyen TL, Winterhalter R, Moortgat G, Kanawati B, Peeters J, Vereecken L (2009) The gas-phase ozonolysis of β-caryophyllene (C15H24). Part II: a theoretical study. Phys Chem Chem Phys 11:4173–4183
22. Novak I (2001) Electronic structure of terpenoids. J Org Chem 66:4728–4731
23. Novak I, Kovac B (2005) Photoelectron spectroscopy of natural products: terpenes. Spectrochim Acta A 61:277–280
24. Novak I, Kovac B, Kovacevic G (2002) Structure and stability of common sesquiterpenes. Spectrochim Acta, Part A 58:2223–2226
25. Pariser R, Parr RG (1953) A semi-empirical theory of the electronic spectra and electronic structure of complex unsaturated molecules. I J Chem Phys 21:466–471
26. Parr RG, Donnelly RA, Levy M, Palke WE (1978) Electronegativity: the density functional viewpoint. J Chem Phys 68:3801–3807
27. Parr RG, Pearson RG (1983) Absolute hardness: companion parameter to absolute electronegativity. J Am Chem Soc 105:7512–7516
28. Parr RG, Szentpaly LV, Liu SB (1999) Electrophilicity index. J Am Chem Soc 121:1922–1924
29. Pfrang C, King M, Canosa-Mas CE, Wayne RP (2006) Gas-phase rate coefficients for the reactions of O(3P), S(3P), Se(3P), and Te (3P) with alkenes: application of perturbation frontier molecular orbital theory, correlations, and structure–activity relations (SARs). Int J Chem Kinet 38:351–356
30. Runge E, Gross EKU (1984) Density-functional theory for time-dependent systems. Phys Chem Chem Phys 52:997–1000
31. Shu YH, Atkinson R (1995) Atmospheric lifetimes and fates of a series of sesquiterpenes. J Geophys Res 100:7275–7281
32. Vizuete W, Junquera V, Allen DT (2004) Sesquiterpene emissions and secondary organic aerosol fomation potentials for southeast Texas. Aerosol Sci Technol 38(S1):167–181
33. Wang W, Wang SL, Li L, Zhang DP, Wang YJ, Sheng GY, Fu JM (2008) Advances in biogenic secondary organic aerosols. Geochimica 37:77–86
34. Went FW (1960) Blue hazes in the atmosphere. Nature 187:641–643
35. Williams DH, Fleming I (eds) (2009) Spectroscopic methods in organic chemistry. McGraw-Hill, New York
36. Winterhalter R, Herrmann F, Kanawati B, Nguyen TL, Peeters J, Vereeckenb L, Moortgat GK (2009) The gas-phase ozonolysis of b-caryophyllene (C15H24). Part I: an experimental study. Phys Chem Chem Phys 11:4152–4172
37. Xia Y, Yin DL, Rong CY, Xu Q, Yin DH, Liu SB (2008) Impact of lewis acids on Diels-Alder reaction reactivity: a conceptual density functional theory study. J Phys Chem A 112:9970–9977
38. Yao L, Ge MF, Qiao ZM, Sun Z, Wang DX (2006) Progresses of tropospheric chemistry of volatile organic compounds. Chemistry Bulletin 69(W049):1–7 (in Chinese)
39. Zerner MC (1991) Reviews in computational chemistry, vol 2. Semiempirical molecular orbital methods. Wiley-VCH, New York

J Mol Model (2012) 18:1463–1472
DOI 10.1007/s00894-011-1181-6

ORIGINAL PAPER

# OH-functionalized open-ended armchair single-wall carbon nanotubes (SWCNT) studied by density functional theory

**Elżbieta Chełmecka · Karol Pasterny · Teobald Kupka · Leszek Stobiński**

Received: 7 April 2011 / Accepted: 5 July 2011 / Published online: 23 July 2011


**Abstract** The structures of ideal armchair (5,5) single-wall carbon nanotubes (SWCNTs) of different lengths (3.7, 8.8, and 16.0 Å for $C_{40}H_{20}$, $C_{80}H_{20}$, and $C_{140}H_{20}$) and with 1–10 hydroxyl groups at the end of the nanotube were fully optimized at the B3LYP/3-21G level, and in some cases at the B3LYP/6-31G* level, and the energy associated with the attachment of the OH substituent was determined. The OH-group attachment energy was compared with the OH functionalization of phenanthrene and picene models and with previous results for zigzag (9.0) SWCNT systems. In comparison to zigzag SWCNTs, the armchair form is more (by about 5 to 10 kcal $mol^{-1}$) reactive toward hydroxylation.

E. Chełmecka
Division of Statistics, Department of Instrumental Analysis,
Medical University of Silesia,
30, Ostrogórska Street,
41-200 Sosnowiec, Poland

K. Pasterny
A. Chełkowski Institute of Physics, University of Silesia,
4, Uniwersytecka Street,
40-007 Katowice, Poland

T. Kupka (✉)
Faculty of Chemistry, University of Opole,
48, Oleska Street,
45-052 Opole, Poland
e-mail: teobaldk@yahoo.com

L. Stobiński
Institute of Physical Chemistry, Polish Academy of Sciences,
44/52, Kasprzaka 44/52,
01-224 Warsaw, Poland

L. Stobiński
Faculty of Materials Science and Engineering,
Warsaw University of Technology,
Wołoska 141,
02-507 Warszawa, Poland

**Keywords** Armchair SWCNT · DFT · OH functionalization · End substitution · Hydroxylation energy

## Introduction

Armchair-type SWCNTs, along with zigzag and chiral carbon nanotube structures [1–6], are important for both basic science studies and practical applications (e.g., in modern material science). Single-wall carbon nanotubes are frequently referred to as one-dimensional structures due to their nanoscale dimensions and quantum properties. Typical SWCNT diameters are from 0.4 to 4 nm, and their lengths range from 100 nm to several micrometers. Their unique mechanical, electro-optical and physicochemical properties make them potentially promising building blocks, components, and molecular devices in both engineering and nanomedicine [2, 7, 8]. The presence of flexible SWCNT additives with significant tensile strength to a typical polymer matrix allows the fabrication of materials with enhanced mechanical properties. However, due to the "smooth" surfaces of SWCNTs, their "anchoring" to and interactions with the surrounding matrix are relatively poor.

The rational design of advanced composite materials requires some kind of interaction between the SWCNTs and the surrounding matrix. Anchoring hydrophobic pristine SWCNTs, used as an additive, becomes much easier when various substituents are introduced (e.g., through end, side, and end-side functionalization of SWCNT), preferably with polar functional groups. On the other hand, acids are used to clean and remove the remains of the metal catalyst applied to the raw carbon nanotube material during the process of SWCNT synthesis. As result, partial damage, chemical oxidation, and the introduction of both –OH and –COOH substituents occur in the purified material [9].

It is known that pristine SWCNTs are long tubes with high length/diameter ratios, and that their surfaces consist of neighboring carbon hexagons (or, more formally, rolled graphene planes); these surfaces are closed at their ends by hemispheres made from carbon pentagons and hexagons. On the other hand, the deviation from planarity (and the instability of the ideally planar benzene-like ring) of the rolled graphene plane decreases as the tube diameter increases, so the tips or ends of the tubes are more reactive (unstable). This means that the tips of SWCNTs are more prone to damage and scission during acid cleaning of the raw material. As a consequence, SWCNTs that are functionalized at their ends (rims) by COOH or OH groups are formed [10, 11]. More chemically reactive media are able to covalently functionalize the surfaces of SWCNTs (mainly at surface defects and by disrupting aromatic ring bond systems). Two types of SWCNT functionalization—sidewall and end functionalization—are known and have been studied experimentally and theoretically [10–15].

The functionalization of pristine SWCNTs dramatically changes their chemical, electronic and transport properties [16]. The initially relatively chemically inert SWCNTs, after functionalization with –OH and –COOH, can form chemical links with more complex reagents such as aminoacids, DNA components, various drugs, and can be dissolved in common solvents. Unfortunately, experimentally derived information on the molecular structures of SWCNTs (from X-ray, SEM, TEM, IR, Raman spectroscopy, etc.) are incomplete, and severe limitations exist [10, 15, 17].

Theoretical calculations [18] are only able to produce results with high chemical accuracy (±1 kcal $mol^{-1}$) for very small and simple molecular systems. Such calculations are performed at a very high level of theoretical sophistication [coupled clusters with singles, doublets and perturbative treatment of triple excitations—CCSD(T)], and using very large basis sets. In order to gain valuable insights into systems containing several or more atoms, several theoretical approximations must be made.

Accurate theoretical calculations are very expensive, and a common practice is to "cut" a small piece of a larger structure and saturate the dangling bonds with hydrogen atoms, and often to ignore intermolecular interactions (H-bonds or van der Waals), the presence of solvent, etc. The RHF method does not take into account electron correlation and is not recommended for systems with multiple bonds (aromatic rings) and lone electron pairs (OH). DFT takes into account electron correlation to some extent, and is comparable in terms of computational expense to the former method. B3LYP is the most popular and versatile density functional, and was deliberately used in our studies [19]. Thus, among the current trends, density functional theory (DFT) is one of the best compromises between accuracy and computational expense. Other approaches, including semi-empirical molecular modeling, are cheaper but less reliable. The results obtained from these theoretical tools must therefore be compared with experimental data and the results of accurate ab initio calculations.

In materials science, molecular modeling—in particular at the density functional theory level (DFT)—of structures and energies [18] provides a fairly efficient and inexpensive way of supporting future laboratory studies. For example, Morokuma and coworkers [20] studied the structures and Raman spectra of model pristine and oxidized SWCNTs, and Leszczynski and coworkers [21] conducted calculations of hydrogen atoms that were chemisorbed by armchair SWCNTs. Basiuk [22] reported on the higher reactivity of carboxylic groups toward methanol exhibited by armchair vs. zigzag functionalized SWCNTs. Denis [23] reported on the stability and the electronic properties of SH-functionalized SWCNTs. Recently, Hashemi and coworkers [24] studied the influence of end OH substitution on armchair SWCNT diameter and the subsequent atomic charges in detail. Unfortunately, they used an inferior Hartree–Fock method instead of density functional theory, which incorporates some amount of electron correlation and enables results of comparable quality to those obtained with post-HF methods (e.g., MP2 or better) to be obtained. In addition, they used a very small and unreliable basis set (STO-3G).

Raman spectroscopy is often used in studies of SWCNTs, and, in particular, the low-frequency radial breathing mode (RBM) at about 140–400 $cm^{-1}$ is inversely correlated with nanotube diameter [3, 25–27].

In our recent paper, we reported on end-substituted zigzag SWCNTs with OH groups, and calculated the respective hydroxylation energies [19]. From one to nine OH substituents were considered, and the functionalized SWCNT obtained, their geometries and their energetics were analyzed.

In this paper, we report the results of our study of the structures of three models of pristine armchair (5,5) SWCNTs 3.7, 8.8, and 16.0 Å in length that were optimized at the level of B3LYP/3-21G or better, and their corresponding end-functionalized structures. From one to ten OH groups were added at one end of the nanotube, and the hydroxylation energies were calculated and compared with previously reported zigzag SWCNTs and selected small molecular systems (phenanthrene and picene). The RBM fragment of the measured Raman spectra of pristine SWCNTs was used as the reference for the theoretically predicted wavenumbers.

The paper provides insights into the OH functionalization of armchair single-wall carbon nanotubes, and indicates the possibility of functionalizing one end of the tube with 1–10 OH substituents.

## Computational methods

All calculations were performed using the Gaussian 03 [28] and 09 [29] programs. Reliable B3LYP density functional and basis sets of a relatively small size (3-21G and 6-31G*) were selected, enabling fairly large-scale calculations to be performed. To facilitate these fairly large-scale calculations, we selected a model of an armchair (5,5) SWCNT with a diameter of about 6.9 Å (similar to a previously reported zigzag (9,0) SWCNT with a diameter of 7.1 Å). Full structure optimization of the unsubstituted and OH-modified SWCNTs was performed. IR and Raman harmonic frequencies were calculated. All positive frequencies indicated the ground-state structure of the optimized system. The energy of the formation of one OH group ($\Delta E$ in kcal mol$^{-1}$) at the edge of the nanotube, which was initially terminated with H atoms, was calculated by considering the hypothetical reaction [19]

$$\mathrm{SWCNT-H}+\frac{1}{2}\,\mathrm{O_2}\longrightarrow \mathrm{SWCNT-OH}+\Delta E. \quad (1)$$

The energies required to add subsequent groups were assumed to be calculated as follows:

$$\Delta E=\left(E_{\mathrm{SWCNT(OH)_{n-1}}}+\frac{1}{2}E_{\mathrm{O_2}}\right)-E_{\mathrm{SWCNT(OH)_n}}, \quad (2)$$

where $n$=1, 2 …10. There are many paths to go from one to ten hydroxyl groups on the rim of the tube, corresponding to "full hydroxylation." The addition of a subsequent OH group close to the previous one would be energetically favored due to the formation of H-bonding. Therefore, we continued to replace neighboring hydrogen atoms at the nanotube rim in the calculations, leading to the estimation of the hydroxylation energy from Eq. 2.

We also considered the hydroxylation energies for model systems (picene and phenanthrene) with fully optimized planar and curved structures, cut from the SWCNTs. Despite the large energy loss due to delocalization upon ring bending (about 123 kcal mol$^{-1}$ for picene calculated at B3LYP/6-31G* level), the corresponding hydroxylation energies were similar (31.7 and 36.6 kcal mol$^{-1}$ for curved and planar picene models). Thus, in the subsequent text, we prefer to focus on the results obtained for fully optimized (planar) structures.

## Raman measurements

A purified commercial sample of a mixture of HiPco SWCNTs [30] was used for the Raman studies. The HiPco samples were high-purity SWCNTs produced using a gas-phase CVD (chemical vapor deposition) process in the so-called HiPco process.

Micro-Raman spectra with resolution of 1 cm$^{-1}$ were measured with a Dilor (Lille, France) XY 800 spectrometer at room temperature using the 514.5 nm line from an Ar laser (Spectra Physics, model 2017). More details on the experimental methods used are described in our previous work [31].

## Results and discussion

In Fig. 1, selected ideal armchair SWCNTs with three different lengths (3.7, 8.8, and 16.0 Å) are shown, all with a diameter of 6.9 Å. A similar diameter has been reported previously for model armchair SWCNTs [6, 32]. The free dangling bonds on carbon atoms are saturated with hydrogen atoms, and the corresponding molecules are labeled armchair structures A, B, and C, respectively. It is worth noticing that the corresponding zigzag tubes are significantly shorter (2.8, 7.1, and 13.5 Å).

Substitution of the hydrogen atoms at one end of the SWCNT with 1–10 hydroxyl groups led to stable OH-functionalized structures. Unlike the zigzag SWCNT [19], the currently analyzed armchair has several possibilities for the arrangement of substituents at the rim. For example, there are three possible functionalization patterns for two neighboring OH groups (Fig. 2). Thus, two substituents can be positioned in two neighboring benzene rings and form a larger ring due to H-bond stabilization (seven-membered structure **I**), or they can be positioned ortho on the same benzene ring, forming a smaller H-bonding ring (five-membered structure **II**), or they can be positioned relatively

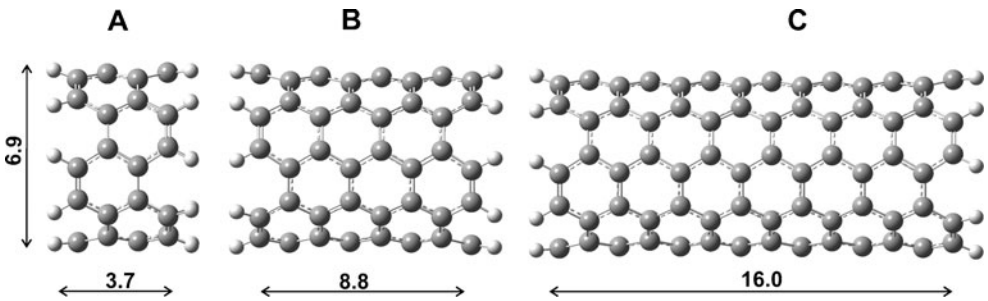



**Fig. 1** Schematic structures (side views) and dimensions (in Å) of B3LYP/3-21 G optimized ideal armchair (5,5) SWCNTs with one (*structure A*), three (*structure B*) and six (*structure C*) strings of attached hexagon rings

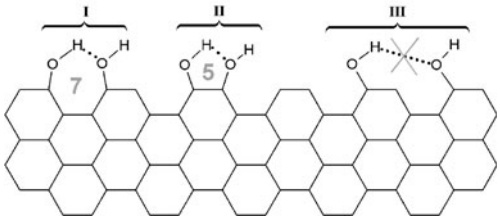



**Fig. 2** Schematic definitions of various localizations of two substituents on an armchair (5,5) SWCNT rim. H-bonding leads to the formation of seven- and five-membered rings

far from each other on two neighboring benzene rings (structure **III**). The latter case is unfavorable for the formation of H-bonding. Obviously, separating the OH substituents by a significant distance will also exclude the formation of H-bonding and so, for simplicity, this situation will not be considered here. To summarize the possible H-bond structures, Table 1 provides the geometric parameters of structures **I** and **II**. According to Desiraju [33], the formation of the O–H…O bonds shown in Fig. 2 indicates fairly strong stabilization of the rings formed (Table 1).

Three neighboring OH substituents would form adjacent larger and smaller H-bond rings (due to symmetry, the opposite case is equivalent). However, four, six and eight adjacent OH groups lead to more complex situations. Thus, for H-bonding between four substituents, three rings are formed (seven-, five-, seven- or five-, seven-, five-membered). We will arbitrarily designate the structures in which larger or smaller H-rings prevail as "L" and "S," respectively. These structures contain two, four, six, and eight OH substituents, respectively. In addition, the label "L" corresponds to cases with equal numbers of five- and seven-membered rings where the latter rings were initially formed. Similarly, the initial formation of five-membered rings would lead to structures labeled "S." These structures contain three, five, seven, and ten OH groups, respectively. Table 2 summarizes the number of possible variants of OH-functionalized armchair SWCNTs as a function of substituent number.

Figure 3 shows B3LYP/3-21G OH-functionalized armchair nanotubes (structures C and B) with ten, four and five substituents, respectively. The calculated diameters

**Table 1** Geometric parameters of OH bonds formed (see Fig. 2 for definitions of structures **I** and **II**)

| Parameter | Structure **I** (seven-membered ring) | Structure **II** (five-membered ring) | Strong H-bond (according to Desiraju [33]) |
|---|---|---|---|
| O…H (in Å) | 1.565 | 1.946 | 1.5–2.2 |
| …H–O (in Å) | 0.986 | 0.980 | |
| O…O (in Å) | 2.607 | 2.554 | 2.5–3.2 |
| O–H...O angle (°) | 160.3 | 117.8 | 130–180 |

**Table 2** Classification of OH-functionalized armchair structures

| No. of OH groups | L | | S | |
|---|---|---|---|---|
| | No. of structures of type **I** (seven-membered ring) | No. of structures of type **II** (five-membered ring) | No. of structures of type **I** (seven-membered ring) | No. of structures of type **II** (five-membered ring) |
| 0 | – | – | – | – |
| 1 | – | – | – | – |
| 2 | 1 | 0 | 0 | 1 |
| 3 | 1 | 1 | 1 | 1 |
| 4 | 2 | 1 | 1 | 2 |
| 5 | 2 | 2 | 2 | 2 |
| 6 | 3 | 2 | 2 | 3 |
| 7 | 3 | 3 | 3 | 3 |
| 8 | 4 | 3 | 3 | 4 |
| 9 | 4 | 4 | 4 | 4 |
| 10 | 5 | 5 | 5 | 5 |

of structures A, B and C with one end completely substituted (ten OH groups) are 7.22, 7.23, and 7.20 Å, respectively—slightly larger than the corresponding unsubstituted analogs (6.94, 6.98, and 6.93 Å). Similar diameters for pristine and end-OH-substituted armchair (5,5) SWCNTs were calculated by Hashemi and coworkers [24]. We can speculate that this small increase in diameter is caused by steric destabilization and structural rearrangement, which leads to the formation of a chain of consecutive hydrogen bonds upon the addition of 10 OH groups. It is also clear that a kind of funnel-shaped deformation occurs upon end functionalization with OH groups (Fig. 3a). The last OH group in Fig. 3b, which belongs to five-membered ring, is nearly collinear with the tube radius and is significantly less tilted in Fig. 3c. In the latter case, the last OH substituent forms a seven-membered ring. The two different patterns shown in Fig. 3b and c illustrate general geometric differences at the rims of multiply substituted armchair (5,5) SWCNTs (see also Table 2). As reported previously [24], the change in SWCNT diameter prompted by OH functionalization would change the atomic charge distribution, thus changing the reactivity of the side of the nanotube toward electrophilic and nucleophilic attack. As a result, the reactivities of functionalized and pristine SWCNTs would differ. However, we would not discuss this issue in the present study.

As seen in Figs. 2 and 3, when one to several OH groups are attached one by one to the nanotube rim, the last substituent is tilted outwards (i.e., it protrudes outwards), and thus resembles the zigzag SWCNT [19]. The tilt angle

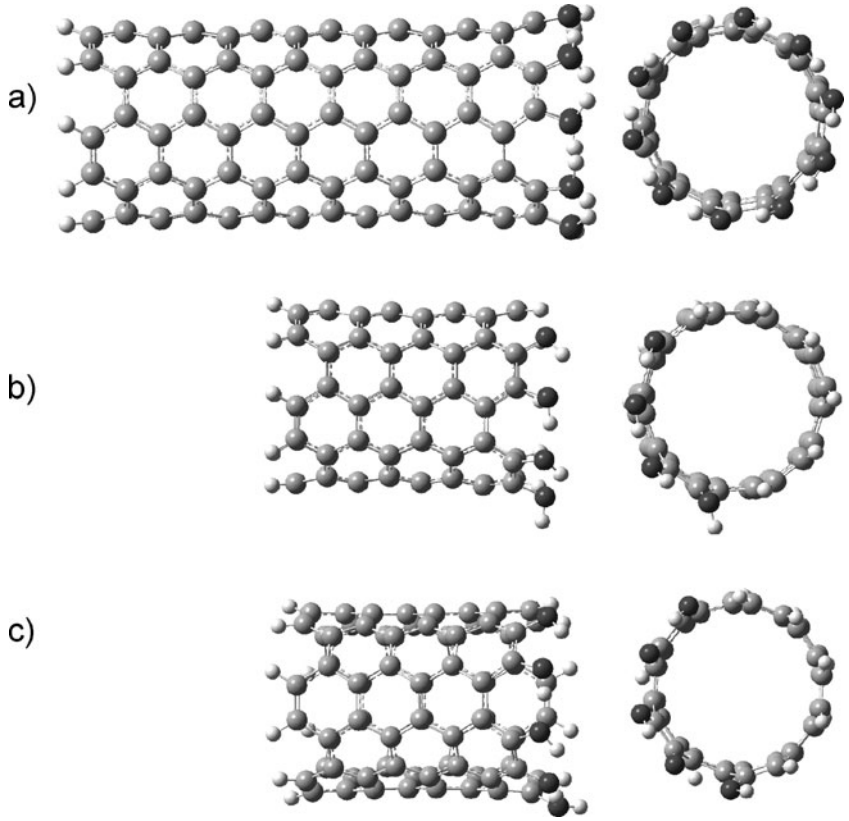



**Fig. 3** Side and top views of OH-functionalized armchair nanotubes: **a** $C_{140}H_{10}(OH)_{10}$ with ten OH substituents, **b** $C_{80}H_{16}(OH)_4$ with four OH substituents, and **c** $C_{80}H_{15}(OH)_5$ with five OH substituents. All tubes show a funnel-type deformation at the rim

is larger when two OH groups are attached to the same benzene ring (ortho substitution; see Figs. 2, 3).

Two types of C–C bonds are present on the "surface" of an armchair SWCNT: those perpendicular to and those tilted at some angle with respect to the tube axis (Fig. 4). Their lengths oscillate according to their positions with respect to the axis and rim. This behavior is similar to that reported previously by Zhou and coworkers [6] and Tanaka and coworkers [34]. It is evident from Fig. 4 that a model formed from three strips of carbon hexagon rings (structure B) reproduces the C–C bond lengths of larger (longer) CNTs fairly well.

The structures of armchair SWCNTs differ from those of zigzag ones, and their rims show different shapes. In the latter, each hydrogen atom is attached to a different hexagon (benzene) ring [19]. In the former, every two hydrogen atoms are at ortho positions on one hexagon ring. Thus, in principle, up to nine substituents—each of them attached to a different ring—could functionalize a rim of zigzag SWCNTs. Similarly, for armchair SWCNTs, one could expect ten substituents "sitting" on five hexagon rings.

To gain some insight into the energetics and substitution pattern at the nanotube rim, we will consider two smaller models first: phenanthrene and its derivative picene, as shown in Fig. 5a and b, as well as their singly and doubly substituted structures, which resemble the functionalization patterns at the SWCNT rim. Obviously, these model compounds are planar, their structures are fully optimized and do not take into account the curvature of the SWCNT. On the other hand, such models do provide some semi-quantitative information about the geometry and energetics of H-bonds formed. The other solution would be artificial bending of their structures, which we prefer to avoid. Such bending would make these models unbalanced (nanotubes with small diameters form saturated and balanced geometrical structures).

The hydroxylation energies per single OH group for phenanthrene and picene, calculated at the B3LYP/3-21 G and 6-31G* levels, are given in Table 3. The energies calculated using a larger and more flexible basis set (6-31G*) are about 10% smaller, but the absolute trends are reproduced by the 3-21G basis set and are similar to those previously calculated for zigzag SWCNTs [19]. It is worth mentioning that adding two OH groups that form a seven-membered H-bonding pattern to phenanthrene causes a distortion (twisting) of the initially planar molecule due to the formation of H-bonds, and is energetically less favorable than the five-membered motif in which the planarity of all aromatic rings is preserved. A similar situation is observed for picene (Table 3), for which molecular planarity is even more significant.

Replacing the first hydrogen atom with an OH group in an armchair SWCNT requires a considerable amount of energy (about 37 kcal $mol^{-1}$ for structures A and B when calculated

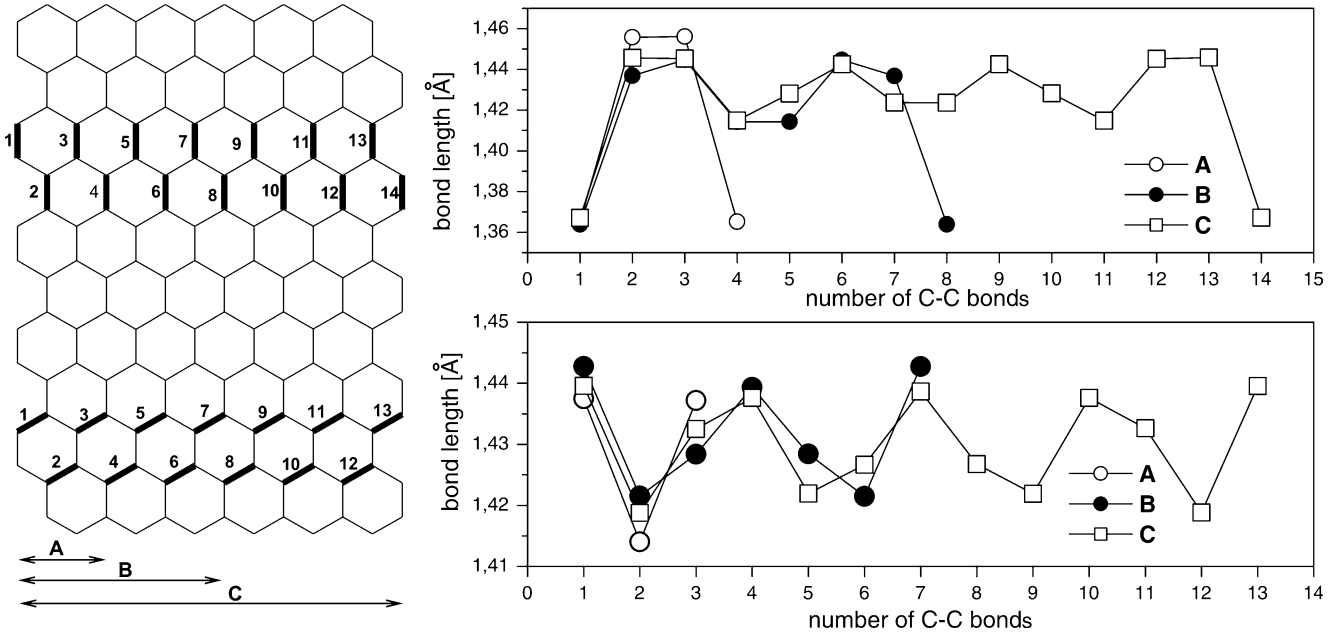



**Fig. 4** The dependence of the length of the C–C bond on its position within structures A, B, and C. Selected bonds that are directed perpendicularly, and at an angle to the nanotube axis, are marked in *bold*

at the B3LYP/3-21G level), but adding the next group requires only slightly more energy (about 4 kcal $mol^{-1}$, see Fig. 6). However, in the case of the zigzag structure, significantly more energy is required [19]. We can assume that the destabilization of the armchair nanotube rim is efficiently compensated for by the formation of a hydrogen bond that stabilizes the structure. As result, upon adding 2–9 OH groups, we observe roughly the same energy cost (a kind of plateau is formed; see Fig. 6). However, closing a loop of substituents requires more energy. Besides, adding substituents to a shorter nanotube (structure A, Fig. 6a) produces a less scattered energy pattern than adding them to a longer nanotube (structure B, Fig. 6b). This kind of energy pattern also supports the importance of H-bond formation. In addition, in contrast to planar model molecules (see Fig. 6 for picene), the initial shape of a SWCNT is curved, and no energy is needed for the molecular plane twisting required to produce efficient H-bonding. Picene was taken as a very crude model of an armchair SWCNT, and for the corresponding S structure we placed 1–4 OH groups at

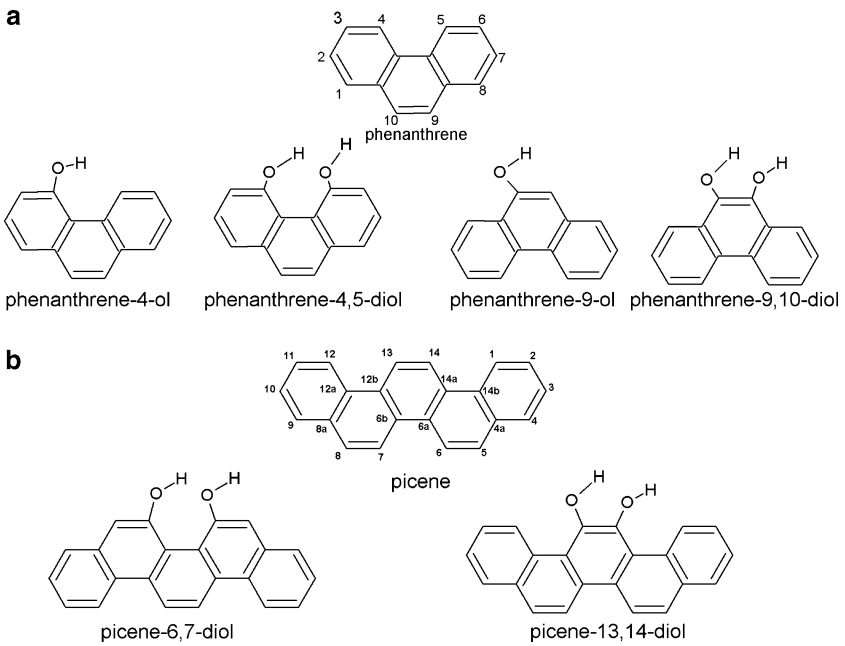



**Fig. 5** **a** Phenanthrene and its mono- and disubstituted OH derivatives. **b** Picene and its disubstituted OH derivatives

**Table 3** B3LYP-calculated hydroxylation energies for the model molecules (in kcal mol$^{-1}$ per substitution)

| Model system | 3-21G | 6-31G* |
|---|---|---|
| Phenanthrene-9-ol | 38.42 | 34.83 |
| Phenanthrene-9,10-diol | 38.21 | 34.24 |
| Phenanthrene-4-ol | 36.96 | 32.38 |
| Phenanthrene-4,5-diol | 38.00 | 33.36 |
| Picene-13-ol | 36.58 | 31.79 |
| Picene-13,14-diol | 40.36 | 33.54 |
| Picene-6-ol | 35.92 | 31.06 |
| Picene-6,7-diol | 36.43 | 32.07 |

carbons 5, 6, 7 and 8, respectively. In a similar manner, OH substituents were placed at carbons 1, 14, 13, and 12 for structure L. We want to stress that in Fig. 6, the differences in energy according to Eq. 2, or the relative values, are shown. Thus, each point $n$ shows the amount of energy needed to introduce one OH group into a structure containing $(n - 1)$ OH substituents. Note that for ten OH groups, there is only one structure (and total energy), but there will be two different numbers due to the different energies of the L and S structures containing nine OH substituents.

We also examined the effect of the basis set size on the absolute value of hydroxylation energy. It is evident from Fig. 7 that the absolute value of hydroxylation energy per single OH group calculated using the larger basis set (6-31G*) is about 5–8 kcal mol$^{-1}$ lower than that obtained using the 3-21G basis set. Obviously, the $\Delta E$ values obtained with the smaller basis set (3-21G) differ, and are generally considered less accurate (and describe the energy landscape more roughly), but the overall qualitative pattern observed for both basis sets is the same. The above result is very important, as it shows that reliable calculations can be conducted on larger molecular systems.

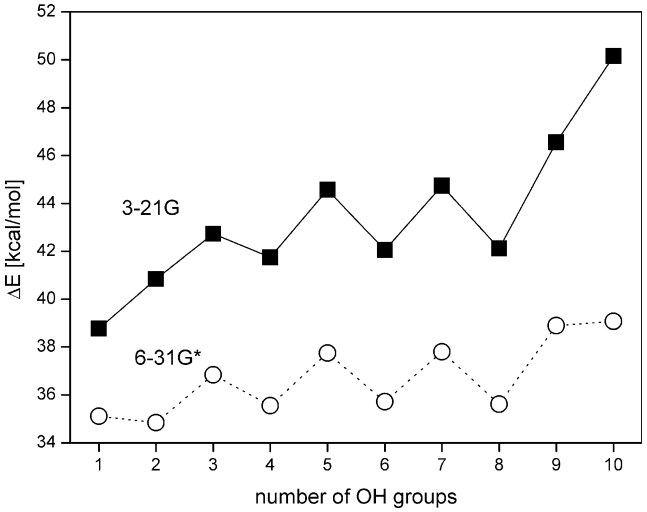



**Fig. 7** Dependence of the B3LYP-predicted absolute hydroxylation energy for armchair SWCNT OH functionalization (structure B, in kcal mol$^{-1}$ per single OH group) on the number of OH groups added, when small and medium-sized basis sets (3-21G and 6-31G*) are employed for the calculations. The data points are connected in order to improve visualization

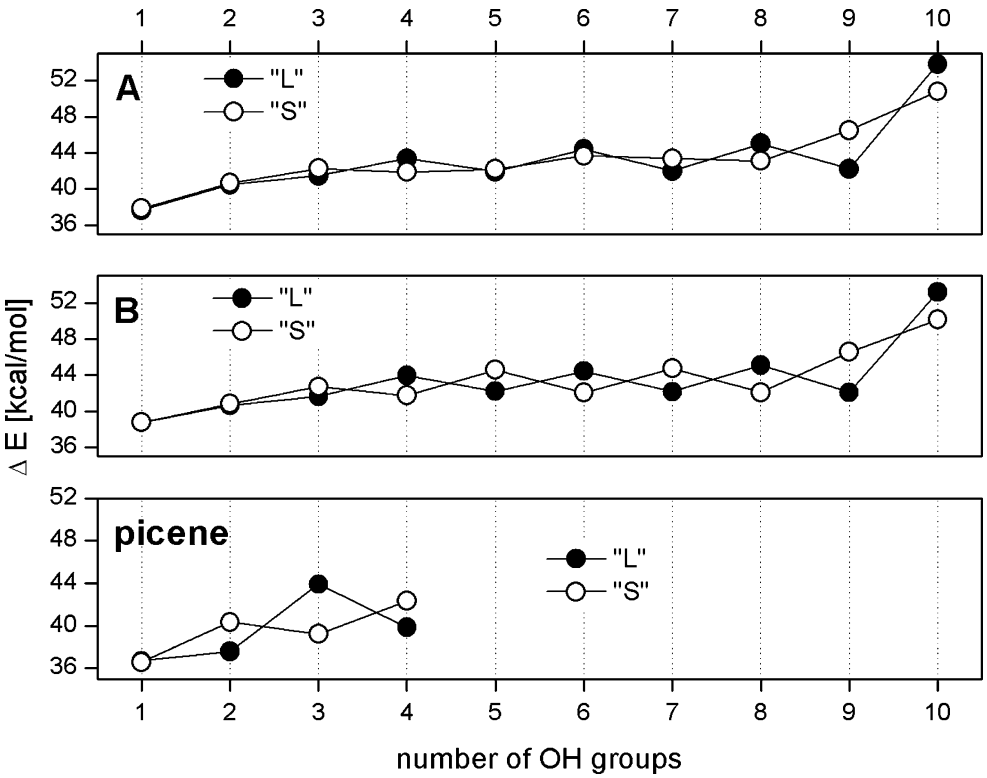



**Fig. 6** Dependence of the hydroxylation energy for structures A and B of armchair SWCNTs (in kcal mol$^{-1}$ per single OH group, calculated at the B3LYP/3-21G level) on the number of substituents for L/S functionalization (to improve visualization, the data points are connected). The energy values represent relative energies according to Eq. 2. Results for picene are included for comparison

Figure 8a and b compare the hydroxylation energies calculated via B3LYP/3-21G per OH substituent for zigzag [19] and armchair SWCNTs. L and S H-bonding patterns were considered for armchair functionalization, and both are notably below the corresponding zigzag energies. As mentioned before, the scatter in the energy difference ($\Delta E$) for armchair SWCNTs is due to the formation of structures **I** or **II** (see Fig. 2).

In the last stage of our study, we checked the quality of our DFT calculations by comparing those results with some experimental results. A comparison of DFT-calculated and measured Raman spectra of SWCNTs provides an indirect test of the proposed theoretical strategy, including structural and vibrational studies of models A, B and C. Thus, below we briefly analyze the selected harmonic vibrations and compare them with measured Raman spectra of pristine SWCNTs.

The most characteristic feature of the Raman spectra of SWCNTs is a collective skeleton vibration called the radial breathing mode (RBM), which is assumed to be inversely proportional to their radius [3, 25–27, 35]. This kind of vibrational mode originates from the symmetric movement of all carbon atoms along the radius, and can be visualized as a form of tube breathing. Thus, Fig. 9 presents a fragment of the Raman spectrum of a commercial sample containing SWCNTs with different diameters ($d$), showing several peaks in the RBM region. The commercial HiPco sample is a mixture of SWCNTs with different diameters (and lengths), and it is apparent that the contents of the tubes (as characterized by their diameters) are reflected in the fitted intensities (dotted lines in Fig. 9). The most intense peaks are 3 and 1. On the other hand, peak intensities systematically decrease, forming a long tail toward higher wavelengths.

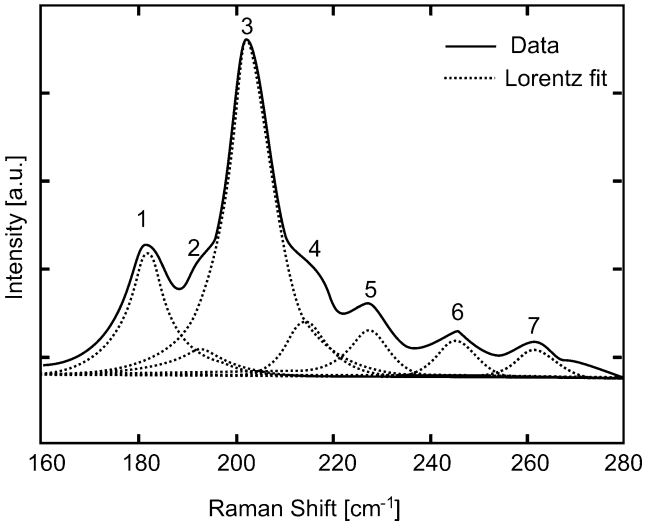



**Fig. 9** Raman spectrum of purified SWCNTs in the RBM region. Individual peaks correspond to nanotubes with different diameters

Several empirical formulae for estimating the radius of an SWCNT from its observed RBM have been reported [3, 25–27, 35]. For example,

$$\nu(\mathrm{RBM}) = A/\mathrm{d}, \tag{3}$$

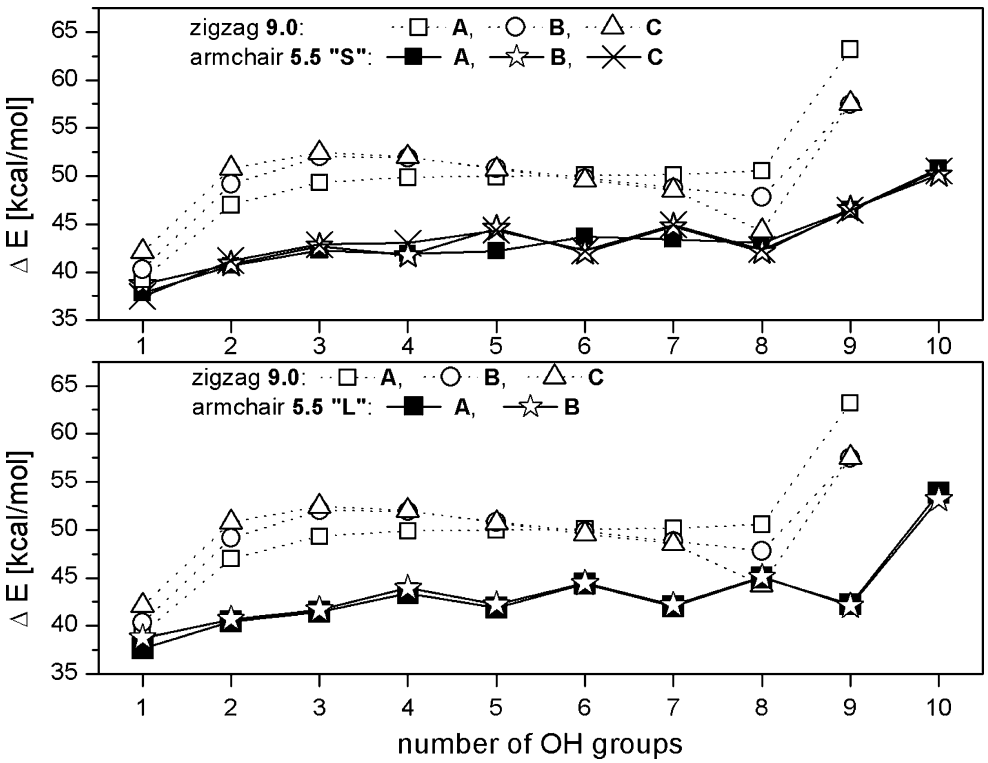



**Fig. 8** Comparison of the absolute hydroxylation energies for zigzag [19] and armchair SWCNT OH functionalization (in kcal mol$^{-1}$ per single OH group) predicted using B3LYP/3-21G. SWCNTs of different lengths were selected, and both L and S patterns of H-bonding were considered. The data points are connected in order to improve visualization

**Table 4** SWCNT diameters ($d$) estimated from the Raman radial breathing mode (RBM; in $cm^{-1}$) for peaks observed in Fig. 9

| Peak no. | $d$ (nm) | RBM [36] | B3LYP |
|---|---|---|---|
| 1 | 1.25 | 182.11 | |
| 2 | 1.16 | 194.78 | |
| 3 | 1.11 | 203.64 | |
| 4 | 1.05 | 215.38 | |
| 5 | 1.00 | 226.26 | |
| 6 | 0.92 | 246.15 | |
| 7 | 0.86 | 263.53 | |
| Theoretical model | | | |
| Armchair (5,5) | 0.69 | 328.99 | 318.06[a] |
| Zigzag (9,0) | 0.71 | 319.72 | 327.6 [19] |

[a] This work

where $A$=227, was suggested in [36]. For small-diameter SWCNTs (<1.4 nm), a slightly different formula was reported [26]:

$$\nu(\mathrm{RBM}) = A/\mathrm{d} + 15, \quad (4)$$

where $A$=219.

The estimated SWCNT diameters determined from the positions of the peaks (Raman frequencies) in Fig. 9 are gathered in Table 4. The SWCNT diameters were calculated according to Eq. 3.

We now discuss in detail the RBM position calculated at the B3LYP/6-31G* level for initial structure B (a medium-length SWCNT without any OH groups) and its derivatives containing one, two and ten OH substituents. The frequencies of this characteristic vibration are 318.06, 318.50, 319.26, and 344.27 $cm^{-1}$ for structure B and its derivatives, respectively, and their Raman activities are 26.03, 16.36, 12.58 and 110.61 $A^4$/AMU. According to Eq. 3, for our pristine SWCNT (0.69 nm in diameter), we should observe a Raman peak at 329 $cm^{-1}$ (see Table 4). Thus, the DFT-calculated position of this vibration is fairly accurate, and differs by only 11 $cm^{-1}$ from the empirical results. Similar agreement between the RBM values obtained from B3LYP calculations for zigzag (9,0) SWCNTs [19] and estimated from Eq. 3 is evident from Table 4. It is worth noting that, starting from 1.05 nm, as the carbon nanotubes decrease in diameter their contributions become fairly small. Thus, our experimental RBM pattern cannot prove the presence of nanotubes with diameters of about 0.7 nm, as used in our molecular modeling.

Also, the observed Raman spectrum in Fig. 9 originates from a sample of pristine SWCNTs. On the other hand, on the basis of our DFT calculations presented above, it is evident that the positions of RBM peaks should not shift significantly upon hydroxylation. This can be explained by the fact that the presence of OH groups at the ends of (long) nanotubes only distorts the rim geometry; their presence does not affect the rest of the tube (the interior part further from the end).

We also note that upon increasing the number of substitutions, the intensity of the RBM in the experimental Raman spectrum initially decreases due to a decrease in symmetry. However, full substitution (with ten OH groups) produces significantly more intense vibrations than the pristine nanotube.

## Conclusions

The present density functional theory studies using B3LYP/3-21G and B3LYP/6-31G* calculations indicate that it is possible to OH functionalize one end of an armchair SWCNT with 1–10 substituents. The hydroxylation energy per OH group is about 10 kcal $mol^{-1}$ lower than for a zigzag SWCNT. This result suggests a significantly higher reactivity toward OH functionalization of armchair vs. zigzag SWCNTs. Functionalization with ten OH groups increases the rim diameter from 6.9 (in pristine CNT) to 7.2 Å, and shifts the RBM mode from 318 to 344 $cm^{-1}$. The intensity of this diagnostic Raman vibration increases about fivefold.

**Acknowledgments** This work was supported by the Polish Council for Science through development grants for the years 2008–2011 (NR 15-0011-04/2008, KB/72/13447/IT1-B/U/08). All of the calculations were performed with the aid of hardware and software at the WCSS (Wrocławskie Centrum Sieciowo-Superkomputerowe) Wrocław.

We also acknowledge the support of the Medical University of Silesia, the University of Silesia and the University of Opole.



## References

1. Iijima S, Ichihashi T (1993) Single-shell carbon nanotubes of 1-nm diameter. Nature 363:603–605
2. Saito R, Dresselhaus MS, Dresselhaus G (1998) Physical properties of carbon nanotubes. Imperial College Press, London
3. Osswald S, Flahaut E, Gogotsi Y (2006) In situ Raman spectroscopy study of oxidation of double- and single-wall carbon nanotubes. Chem Mat 18:1525–1533
4. Iijima S (1991) Helical microtubules of graphitic carbon. Nature 354:56–58
5. Kar T, Akdim B, Duan X, Pachter R (2006) Open-ended modified single-wall carbon nanotubes: a theoretical study of the effects of purification. Chem Phys Lett 423:126–130
6. Zhou Z, Steigerwald M, Hybertsen M, Brus L, Friesner RA (2004) Electronic structure of tubular aromatic molecules derived

from the metallic (5,5) armchair single wall carbon nanotube. J Am Chem Soc 126:3597–3607
7. Khabashesku VN, Margrave JL, Barrera EV (2005) Functionalized carbon nanotubes and nanodiamonds for engineering and biomedical applications. Diamond Rel Mat 14:859–866
8. Calvert P (1999) A recipe for strength. Nature 399:210–211
9. Veloso MV, Filho AGS, Filho JM, Fagan SB, Mota R (2006) Ab initio study of covalently functionalyzed carbon nanotubes. Chem Phys Lett 430:71–74
10. Kuzmany H, Kukovecz A, Simon F, Holzweber M, Kramberger C, Pichler T (2004) Functionalization of carbon nanotubes. Synth Met 114:113–122
11. Kar T, Scheiner S, Roy AK (2008) The effect on acidity of size and shape of carboxylated single-wall carbon nanotubes. A DFT-SLDB study. Chem Phys Lett 460:225–229
12. Wongchoosuk C, Udomvech A, Kerdcharoen T (2009) The geometrical and electronic structures of open-end fully functionalized single-walled carbon nanotubes. Curr Appl Phys 9:352–358
13. Kar T, Scheiner S, Patnaik SS, Bettinger HF, Roy AK (2010) IR characterization of tip-functionalized single-wall carbon nanotubes. J Phys Chem C 114:20955–20961
14. Kar T, Adkim B, Duan X, Pachter R (2006) Open-ended modified single-wall carbon nanotubes: a theoretical study of the effects of purification. Chem Phys Lett 423:126–130
15. Salzmann CG, Llewellyn SA, Tobias G, Ward MAH, Huh Y, Green MLH (2007) The role of carboxylated carbonaceous fragments in the functionalization and spectroscopy of a single-walled carbon-nanotube material. Adv Mater 19:883–887
16. Zhao J, Lu JP, Han J, Yang CK (2003) Noncovalent functionalization of carbon nanotubes by aromatic organic molecules. Appl Phys Lett 82:3746–3751
17. Vigoloa B, Mamane V, Valsaque F, Le TNH, Thabit J, Ghanbaja J, Aranda L, Fort Y, McRae E (2009) Evidence of sidewall covalent functionalization of single-walled carbon nanotubes and its advantages for composite processing. Carbon 47:411–419
18. Foresman JB, Frisch A (eds) (1996) Exploring chemistry with electronic structure methods. Gaussian Inc., Pittsburgh
19. Chełmecka E, Pasterny K, Kupka T, Stobiński L (2010) Density functional theory studies of OH-modified open-ended single-wall zigzag carbon nanotubes (SWCNTs). J Mol Struct Theochem 948:93–98
20. Irle S, Mews A, Morokuma K (2002) Theoretical study of structure and Raman spectra for models of carbon nanotubes in their pristine and oxidized forms. J Phys Chem A 106:11973–11980
21. Dinadayalane TC, Kaczmarek A, Łukaszewicz J, Leszczyński L (2007) Chemisorption of hydrogen atoms on the sidewalls of armchair single-walled carbon nanotubes. J Phys Chem C 111:7376–7383
22. Basiuk VA (2002) Reactivity of carboxylic groups on armchair and zigzag carbon nanotube tips: a theoretical study of esterification with methanol. Nano Lett 2:835–839
23. Denis PA (2009) Theoretical investigation of the stability, electronic and magnetic properties of thiolated single-wall carbon nanotubes. Int J Quantum Chem 109:772–781
24. Hashemi SF, Keshvari K, Moradi S (2009) A theoretical ab initio study on functionalized single-walled carbon nanotubes as a molecular absorbent. Fullerenes Nanotubes Carbon Nanostruct 17:390–400
25. Kuzmany H, Plank W, Hulman M, Kramberger C, Gruneis A, Pichler T, Peterlik H, Kataura H, Achiba Y (2001) Determination of SWCNT diameters from the Raman response of the radial breathing mode. Eur Phys J B 22:307–320
26. Dresselhaus MS, Dresselhaus G, Jorio A (2007) Raman spectroscopy of carbon nanotubes in 1997 and 2007. J Phys Chem C 111:17887–17893
27. Kürti J, Zólyomi V, Kertesz M, Sun G (2003) The geometry and the radial breathing mode of carbon nanotubes: beyond the ideal behaviour. New J Phys 5:125.1–125.21)
28. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JA Jr, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann, RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C, Pople JA (2004) Gaussian03, revision E01. Gaussian, Wallingford
29. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JA Jr, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin R. L, Fox DJ (2009) Gaussian 09, revision A.02. Gaussian, Wallingford
30. Bendjemil B (2009) Electronic and optical properties of the express purified SWCNTs produced by HiPco process. Int J Nanoelectron Mater 2:173–182
31. Stobiński L, Tomasik P, Lii CY, Chan HH, Lin HM, Liu HL, Kao CT, Lu KS (2003) Single-walled carbon nanotube-amylopectin complexes. Carbohydr Polym 51:311–316
32. Bertoni G, Calmels L (2006) First-principles calculation of the electronic structure and energy loss near edge spectra of chiral carbon nanotubes. Micron 7:486–491
33. Desiraju GR, Steiner T (1999) The weak hydrogen bond in structural chemistry and biology. Oxford University Press, New York
34. Tanaka K, Ago H, Yamabe T, Okahara K, Okada M (1997) Bond alternation in carbon nanotubes including σ-electrons. J Quantum Chem 63:637–644
35. Chang T (2007) Explicit solution of the radial breathing mode frequency of single-walled carbon nanotubes. Acta Mech Sin 23:159–162
36. Jorio A, Fantini C, Pimenta MA, Capaz RB, Samsonidze GG, Dresselhaus G, Dresselhaus MS, Jiang J, Kobayashi N, Grűneis A, Saito R (2004) Resonance Raman spectroscopy (n,m) dependent effects in small diameter single-wall carbon nanotubes. Phys Rev B 71:075401(1–11)

J Mol Model (2012) 18:1473–1479
DOI 10.1007/s00894-011-1172-7

ORIGINAL PAPER

# Conformation analysis of a surface loop that controls active site access in the GH11 xylanase A from *Bacillus subtilis*

**Davi Serradella Vieira · Richard John Ward**

Received: 26 April 2011 / Accepted: 29 June 2011 / Published online: 23 July 2011


**Abstract** Xylanases (EC 3.2.1.8 endo-1,4-glycosyl hydrolase) catalyze the hydrolysis of xylan, an abundant hemicellulose of plant cell walls. Access to the catalytic site of GH11 xylanases is regulated by movement of a short β-hairpin, the so-called thumb region, which can adopt open or closed conformations. A crystallographic study has shown that the D11F/R122D mutant of the GH11 xylanase A from *Bacillus subtilis* (BsXA) displays a stable "open" conformation, and here we report a molecular dynamics simulation study comparing this mutant with the native enzyme over a range of temperatures. The mutant open conformation was stable at 300 and 328 K, however it showed a transition to the closed state at 338 K. Analysis of dihedral angles identified thumb region residues Y113 and T123 as key hinge points which determine the open-closed transition at 338 K. Although the D11F/R122D mutations result in a reduction in local inter-intramolecular hydrogen bonding, the global energies of the open and closed conformations in the native enzyme are equivalent, suggesting that the two conformations are equally accessible. These results indicate that the thumb region shows a broader degree of energetically permissible conformations which regulate the access to the active site region. The R122D mutation contributes to the stability of the open conformation, but is not essential for thumb dynamics, i.e., the wild type enzyme can also adapt to the open conformation.

D. S. Vieira (✉) · R. J. Ward
Departamento de Química, Faculdade de Filosofia Ciências e Letras de Ribeirão Preto - Universidade de São Paulo,
Avenida dos Bandeirantes,
3900 – 14040-90 Ribeirão Preto, SP, Brazil
e-mail: davisv@usp.br

**Keywords** Thumb conformation transitions · Active-site movement · Conformations energetically equivalents · Hinge points · Intramolecular mobility · Molecular dynamics

## Introduction

Intramolecular mobility and conformational changes of enzymes play a crucial role in biocatalysis, and enzyme activity is therefore dependent on protein flexibility. Conformation changes and mobility in enzymes are involved in substrate binding and product release from the active site together with stabilization of reaction intermediates in the catalytic process. Investigation of enzyme dynamics can provide information about functionally relevant movements [1, 2], identification of residues for mutagenesis [3] and structural determinants of thermostability [4].

Xylanases (EC 3.2.1.8) are produced by many microorganisms and degrade xylan to xylo-oligosaccharides and xylose which are used as carbon sources for cell growth [5]. Xylanases have been grouped into families GH10 and GH11 based on the three-dimensional structure and amino acid sequence similarity [6]. Although the structures of these enzymes have been studied extensively by x-ray crystallography, a complete understanding of how xylanase structure dynamics contributes to the catalytic mechanism must arise from investigation of the internal motions exhibited by the enzyme, which allows distinguishing functionally important conformational flexibility.

The structures of GH11 xylanases are comprised of a single α-helix and two twisted β-sheets forming a jelly-roll fold, in which three regions have been recognized using an analogy to a right hand. A fingers region and a

palm region comprised of β-sheet structures form the floor and a wall of the active site cleft, respectively. The thumb region is comprised of a short two stranded β-sheet, which forms the opposing wall of the active site cleft, and has been shown to be highly flexible [1–3, 7–9]. Recently the role of conformational flexibility of the thumb loops of family GH11 xylanases has been the focus of debate in the literature [2–7]. The primary structure of the thumb region is highly conserved throughout the family GH11 xylanases, and comparative studies of the crystal structures and molecular dynamics (MD) simulations have suggested that the thumb sequence is the most flexible region in the molecule [7, 8]. It has been suggested that substrate access to the active site is controlled by coordinated and opposing movements of the thumb and the fingers region in a so-called "hinge-bending" motion [9]. More recently, we have suggested that substrate access to the active-site region may be regulated by temperature dependent open-close movement performed exclusively by the thumb region [2], and that these movements do not change the overall structure of the thumb [2]. Furthermore it is known that the thumb region can adopt different conformations and that it is closed in the presence of substrate [2, 10, 11]. The major challenge is to characterize the nature of these conformational differences and motions in order to understand their contribution to enzyme function.

Experimental evaluation of polypeptide flexibility is generally difficult [3], and molecular simulations have become a useful alternative to evaluate protein dynamics, especially in cases where a wide sampling of conformation space is required. Although microsecond time-scales for molecular simulations are now accessible, and allow the study of biological events such as peptide folding [12], dynamics information involving conformational changes and side-chain movements that occur in the nanosecond range are more frequently used to evaluate properties such as flexibility [4, 7, 13].

The structural basis of catalysis has been investigated by a variety of experimental methods, including MD simulation [7, 12, 13], and several questions correlating conformational changes and dynamics of GH11 xylanases with their biological function remain unresolved, including the degree of thumb mobility in relation to catalysis. Site-directed mutagenesis of the residues situated at the tip and the base of the thumb loop has shown that the thumb is important both for the correct positioning of the substrate in the active-site and for assisting product release after hydrolysis [14, 15]. Recently, it has been demonstrated that the D11F/R122D mutant of xylanase A, a GH11 xylanase from *Bacillus subtilis* (BsXA) displays a thumb region with a stable extended open conformation not previously observed in studies of GH11 xylanases [3]. Here we present a comparative MD study between the D11F/R122D mutant from *Bacillus subtilis* xylanase A [3] and the native enzyme (BsXA) [7] over a range of temperatures. The goal has been to elucidate the basis for the stable open thumb conformation based on previously published data and the dynamic/steric parameters of the enzyme obtained from the MD calculations.

## Materials and methods

### MD simulations

All systems were simulated in the NpT ensemble at pH 7.0 and 300, 328 and 338 K, where the temperatures were controlled by using the V-rescale thermostat [16]. A pressure of 1.0 atm was controlled isotropically by the Berendsen barostat [17], with a coupling time constant of 1.0 ps. The hydrogen covalent bonds and SPC [18] water molecules were constrained by LINCS [19] and SETTLE [20] algorithms, respectively. The leap-frog algorithm [21] was employed to integrate the equations of motion with a time step of 2.0 fs. The initial velocities were sampled with the aid of the Maxwell-Boltzmann distribution. The long-range interactions were treated using particle-mesh Ewald sum (PME) method [22], with a cutoff equal to 1.2 nm. The long-range contributions of the electrostatic interactions were updated every 10 time steps.

The initial atomic coordinates of the native BsXA and the D11F/R122D mutant for building the dodecahedron simulation boxes were obtained from the Worldwide Protein Data Bank (www.rcsb.org/pdb). Relevant information about the crystallographic structures of BsXA and the mutant can be obtained from references [7] and [3], respectively. After initial energy minimization of the BsXA and D11F/R122D mutant structures using a steepest descent algorithm [23], SPC water molecules were introduced into each system avoiding overlaps, and the water concentration obtained in this way was approximately 53.0 $mol.L^{-1}$. Prior to the simulation, the protonation states of the ionizable residues in the BsXA and D11F/R122D mutant structures at pH=7.0 were determined by use of the $H^{++}$ program (Poisson-Boltzmann based pKa calculations) [24]. Counter ions were introduced to neutralize the overall charge of the system, in which 10 $Na^{+}$ and 12 $Cl^{-}$ ions were inserted in the most electrostatically favorable positions. The initial atomic coordinates used for the MD simulations at 328 and 338 K were derived from the final minimum energy configurations obtained at 300 and 328 K, respectively.

All simulations were equilibrated for approximately 150 ps by MD runs at the chosen temperature with restrained protein atom positions, both to allow solvent

relaxation and to avoid undesirable protein-water interactions. The total time of each simulation was varied in accord with the root mean square deviation (RMSD) behavior, resulting in total simulation times at all temperatures of around 15 ns for the D11F/R122D mutant and 10 ns for BsXA. Intramolecular hydrogen bonds were identified on the basis of geometric criteria, where the donor-receptor distance was shorter than 0.24 nm, the (A-H  O) angle larger than 90° (A ≡ N or O), and fractional occurrence (Fr) was greater than 10% [25]. All MD runs and analyses were performed with the GROMACS 4.0 software package [26] using the GROMOS-96(43a1) force field [27].

An estimate of the conformational energies ($E_{conf}$) of the proteins were obtained as the sum of all interactions energies between all the protein atoms, $E_{conf} = \sum_{i}^{N} E_{i,i+1}$, where $N$ is the total number of atoms and $E_{i,i+1}$ is the energy between atoms $i$ and $i + 1$.

## Results and discussion

### Overall stability and thumb conformational changes

Figure 1a presents the superposed 3-D structures of BsXA and D11F/R122D mutant showing the canonical family GH11 xylanase jellyroll fold in which two curved β-sheets form the active site and substrate-binding cleft. The GH11 jellyroll structure resembles the shape of a right hand, in which the regions have been accordingly named as the fingers, palm and thumb regions. Figure 1a also shows the thumb loop regions in both the open and closed conformations. The active-site is formed by a surface cleft whose floor and walls are defined by residues in the fingers and palm regions, in which the two catalytic Glu residues are located [1, 8, 10, 11, 28]. Substrate access to the active-site cleft is controlled by the thumb region, a β-hairpin structure which may adopt a variety of conformations and which is a key structural element in the regulation of catalytic activity.

The influence of temperature on the intramolecular mobility of the BsXA and the D11F/R122D mutant can be analyzed by calculating the RMSD(t) profiles from the atomic MD trajectories using the Cα atoms positions of the initial structure at 300 K as the reference. The calculated average RMSD values are shown in Table 1, and show increases with temperature for both proteins. The average RMSD values are considerably higher for the mutant than for the native BsXA, mainly at 300 and 328 K. This indicates that the ensemble of structures generated by the MD simulations differ more from the initial structure in the case of D11F/F122D mutant than the native BsXA. These results suggest that the mutant undergoes a greater conformational change as a consequence of increased temperature. Figure 1b show the

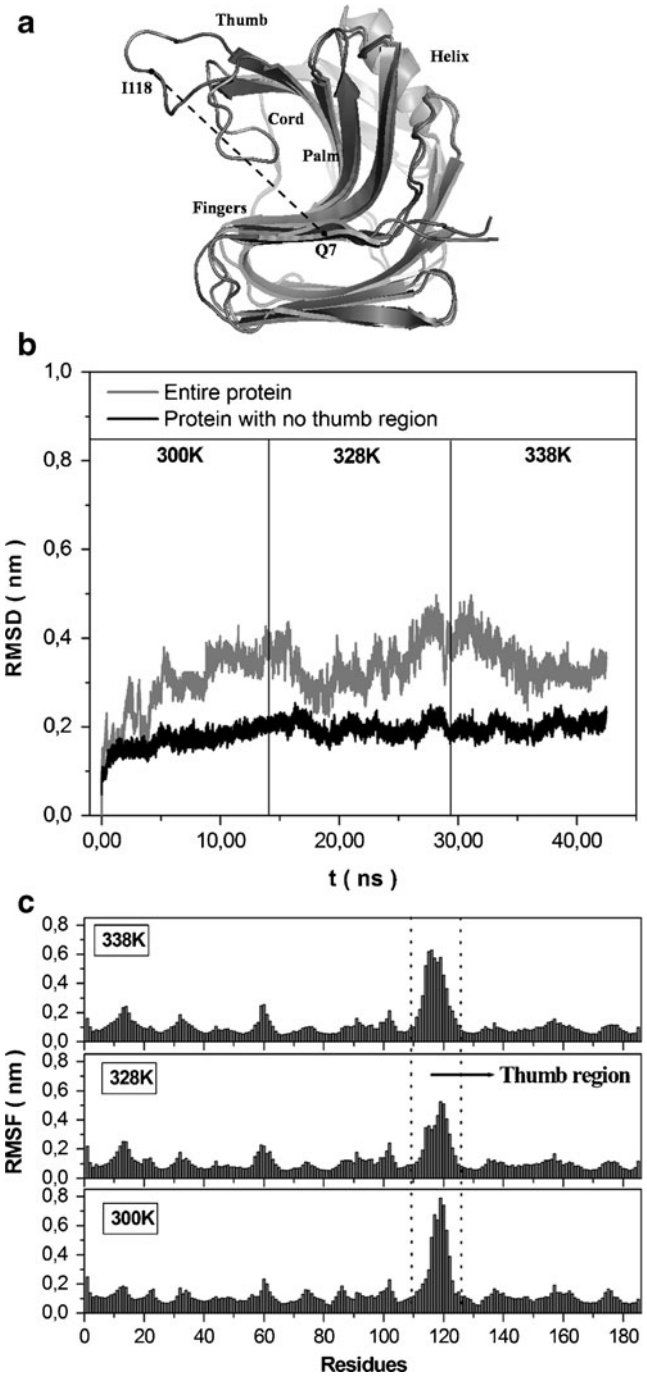



**Fig. 1** Typical three dimension structures of the open and closed thumb conformations of BsXA. (**a**) The superposed 3-D structures of BsXA (PDB ID: 1XXN) and D11F/R122D (PDB ID: 1EXU) mutant showing the jellyroll fold in which two curved β-sheets form the active site and substrate-binding cleft. The dashed line connects the alpha carbons of I118 and Q7 residues. Global dynamics of the BsXA and mutant proteins. (**b**) RMSD(t) profile including and excluding the thumb region in the calculation and (**c**) RMSF per residue calculated for the mutant D11F/R122D at temperatures of 300 K, 328 K and 338 K

**Table 1** Mean RMSD and standard deviations for BsX and D11F/R122D mutant proteins at different temperatures

| T(K) | BsX/RMSD (nm) | Mutant/RMSD (nm) |
|---|---|---|
| 300 | 0.19±0.024 | 0.30±0.070 |
| 328 | 0.27±0.028 | 0.34±0.050 |
| 338 | 0.33±0.028 | 0.35±0.042 |

RMSD(t) profile of the mutant D11F/R122D including and excluding the thumb region in the calculation. Excluding the thumb region considerably reduces the RMSD(t) values, indicating that the thumb region accounts for up to half the calculated RMSD(t) values. The exclusion of the thumb region from the RMSD(t) calculations results in an average value of <0.20 nm at all temperatures, indicating a low degree of intramolecular mobility of the other regions of the enzyme which is characteristic of a well equilibrated and highly convergent biomolecular system.

Figure 1c presents the root mean square fluctuations (RMSF) per amino acid residue, which clearly shows the high flexibility of the thumb region at all temperatures. The RMSF values show that the thumb region can be up to four times more flexible than the other regions of the protein. In the case of native enzyme the flexibility of the thumb region is at most two times that of the other regions of the protein [7]. Previous studies have reported that in the absence of substrate the GH11 xylanases exhibited open thumb conformations at higher temperatures [2, 7, 8]. Furthermore, it was previously observed that the thumb loop adopts a closed conformation below the optimum catalytic temperature ($T_{opt}$), however at the $T_{opt}$ the thumb assumes an open conformation facilitating substrate access to the catalytic site pocket [7].

## Evaluation of functional movements of family GH11 xylanases

Functional movements are specific to each enzyme, and due to their conformational flexibility surface loops frequently play an important role in catalysis. Loop motions are generally involved in substrate binding and or product release in the active site or in the stabilization of the reaction intermediates. These functionalities have been described for several enzymes [29–31], including the GH11 xylanases [2, 3, 10, 14].

The crystal structure of the BsXA presents the closed conformation which is typical of family GH11 xylanases reported in the literature. In contrast, the 3-D structure of the dimeric D11F/R122D mutant has recently been reported [3], in which one of the monomers presents an extended open thumb conformation. The authors proposed that this extended open thumb conformation was due to intermolecular interactions with the symmetry-related protein in the crystal lattice.

In order to further investigate the flexibility of the thumb region of the D11F/R122D mutant, the distance between the $C_\alpha$ atoms of residues I118 at the tip of the thumb loop and residue Q7 in the fingers region was calculated during the MD trajectories. The result is presented in Fig. 2, where an open thumb conformation results in a greater distance between the two residues. Figure 2 shows that the open conformation is stable during the simulations at 300 K, 328 K and for the initial 5.0 ns at 338 K. The mutant protein maintained a distance of approximately 2.7 nm up to 338 K, as compared to the native BsXA which has previously been shown to have the greatest value at 328 K, reaching a distance of 2.3 nm [2, 7]. It is well established in the literature that the closed to open transition occurs at 328 K, which coincides with the optimum catalytic temperature of the native GH11 xylanase [1, 2, 7, 9]. Figure 2 clearly shows that mutant protein undergoes a similar transition, however at a higher temperature, where an open to closed transition was observed at 338 K, which probably is the cause of the inactivation of the enzyme under these conditions [3]. Since the MD simulations were performed using a protein monomer it can be concluded that intermolecular interactions with a symmetry-related molecule such as may be formed in a dimer are not crucial to maintain the open thumb conformation.

To investigate the thumb movement in substrate binding and product release, R112, N114 and T126 have been suggested as potential targets for site-directed mutagenesis [3]. Among these residues, R112 and N114 are highly conserved in the GH11 xylanases, whereas T126 is not [32]. In a previous study [3], R112, N114 and T126 were

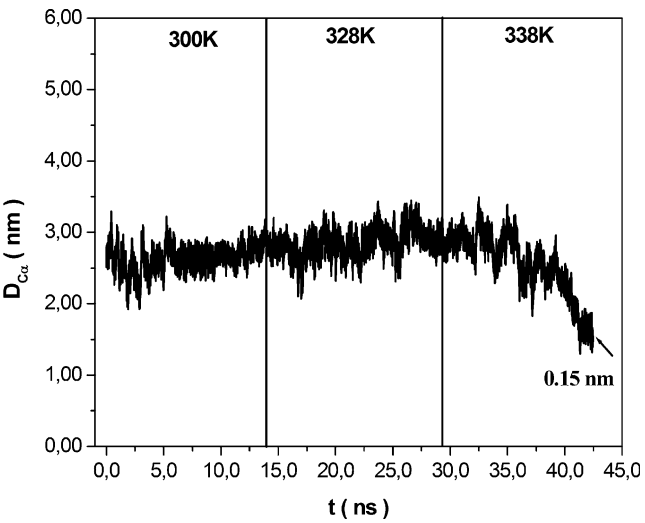



**Fig. 2** The distance between the tip of thumb loop and the fingers region ($D_{C\alpha}$) at increasing temperatures for D11F/R122D mutant

modified by either glycine or proline to introduce more flexibility or more rigidity, respectively. The R112G mutation reduced the specific activity by 86% and the R112P decreased the specific activity fivefold, indicating that glycine or proline cannot fulfill the role of these residues [3]. In the case of T126, the modifications only slightly changed the specific activity [3].

The β-hairpin thumb region is comprised of two β-strands, one formed by residues 111 to 114 (shown in black in Fig. 3) and the other by residues 123–126 (gray in Fig. 3). The two β-strands are linked by an 8-residue turn comprised of residues 115–122. The Ramachandran dihedral angles (Φ and ψ) for all thumb residues were computed from MD trajectories for the D11F/R122D mutant. Figure 4a and b show the Ramachandran plots at all the points in the MD simulation for residues T111, R112, Y113 and N114 (Fig. 4a), and T123, T124, F125 and T126 (Fig. 4b), respectively. Analysis of the Ramachandran angles reveal that most points lie in favored regions of the plot, however in some cases; R112, N114, Y113 and T123 also populate "prohibited" regions (see Fig. 4a) particularly under conditions of greater thumb mobility above 328 K when the thumb begins the closing process. Since the Ramachandran angles provide a direct measure of main chain flexibility, the high degree of freedom associated with conformational changes is a characteristic that is consistent with hinge point residues. Therefore, at 300 and 328 K, residues R112 and N114 are the only hinge points in the thumb region.

We suggest that two factors contribute to the reduced specific activity observed in the R112G mutation: *(i)* an extended positively charged side chain is replaced by a short neutral side chain, resulting an altered intermolecular hydrogen bonding between the enzyme and substrate[2], and *(ii)* the flexibility conferred by the glycine residue generates a high percentage of unfavorable conformations. At 338 K,

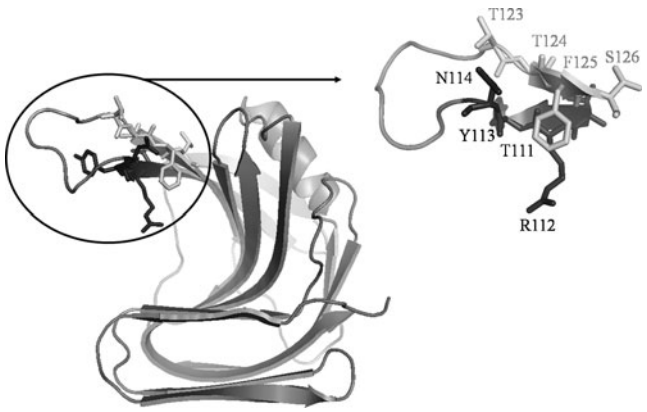



**Fig. 3** Structure of the thumb region showing residues chosen for Ramachandran angle analysis: T111, R112, Y113 and N114 (black), and T123, T124, F125 and S126 (gray)

the temperature where closing of the thumb region was detected, residues Y113 and T123 also show hinge point characteristics, therefore hinge point residues on both β-strands in the thumb region seem to be crucial to the open/closed conformation transition. Since, T124 and F125 show flexibility characteristic of hinge points only at 300 K, and as the open to closed conformation transition occurs only at elevated temperatures and requires mobility of residues in both strands of the thumb region, we suggest that these residues are unlikely to have a hinge function.

## Intra and intermolecular interactions – local analysis

Hydrogen bonds (HB) and salt-bridges make important contributions to maintain the protein tertiary structure, and perturbations in these interactions result in changes in the 3D-structure of the enzyme. The interactions and local environments of the mutated residues in the D11F/R122D were analyzed during the MD simulation. The D11 residue is located in the fingers region in a favorable position for interaction with the tip of the thumb region. In the native BsXA enzyme the side-chain atoms of D11 and R122 do not form intramolecular HB, but participate in around 4.5  5.0 intermolecular HB with solvent molecules. The distance between the side chain atoms of D11 and R122 varies between 0.5 to 0.8 nm during the simulation, thus although these residues do not form HB, they do have a stabilizing influence due to the long-range nature of the Coulomb forces, and the non-bonded interaction energies of D11 and R122 are approximately −14.4 and −24.0 kcal $mol^{-1}$, respectively.

The mutations D11F and R122D in the BsXA affect these interactions, since the physical-chemistry properties of the mutated residues and their environments are significantly changed. The D11F mutation substitutes a negatively charged for a hydrophobic residue, while the R122D mutation inverts the electrostatic charge of the residue.

The F11 in the mutant enzyme does not form intra- or intermolecular HB, however the D122 can form intermolecular HB with solvent molecules with an energy of around −6.7 kcal $mol^{-1}$, and can form intramolecular HB with Y113 and S117 for more than 60% of the time at 300 and 328 K. The non-bonded interaction energies of F11 and D122 for the mutant protein are approximately −5.1 and −18.4 kcal $mol^{-1}$, respectively. The D11F and R122D mutations generate a significant decrease of 64.6% and 23.3%, respectively, in the energy of the non-bonded interactions in relation to native BsXA.

Analysis of the R112 residue reveals that the mutations D11F and R122D increase the number of intramolecular HB without loss of its intermolecular HB. In the native BsXA, R112 residue forms no intramolecular HB and

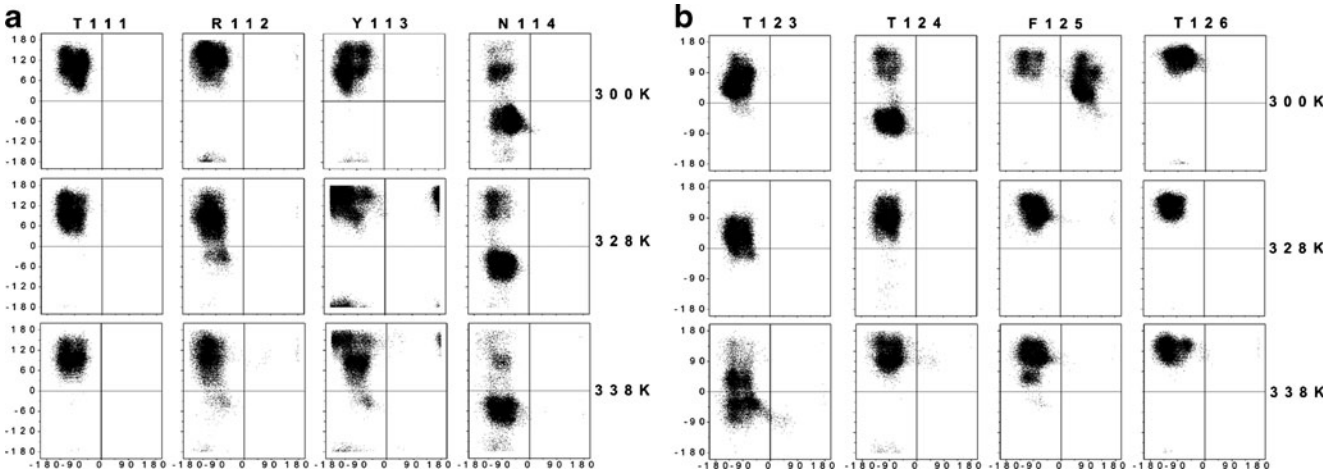



**Fig. 4** Hinge point analysis for residues in the thumb region. Ramachandran graphs for residues T111, R112, Y113 and N114 (**a**), and T123, T124, F125 and S126 (**b**)

approximately 4.0 intermolecular HB with solvent molecules. However, in the D11F/R122D mutant, R112 participates in approximately 2.0 intramolecular HB with E78 and Y80, whilst maintaining the intermolecular interactions at all temperatures investigated. The importance of the R112 residue has been reported in the literature and explained in terms of its interaction with substrate and water molecules [1, 2]. The present results indicate that in addition to the proposed role in enzyme/substrate complex formation, the R112 can form stable intramolecular HB with residues located in the substrate binding cleft (E78 and Y80), thereby stabilizing the open thumb conformation.

The energy barrier between different conformation states were evaluated by comparing $E_{conf}$ values, which are presented in Table 2 for all temperatures studied. It is observed that $E_{conf}$ is essentially equivalent for both proteins at all temperatures, indicating that the open and closed conformations are stable relative to each other. Furthermore the energy fluctuation during the simulation shows that both conformation states are energetically accessible, and that the choice of either conformation is probably determined by a delicate local energy balance. Overall, our results indicate that the D11F mutation is not involved in the opening and closing mechanism of the thumb region, in contrast to the R122D mutation which confers local stability to the enzyme by potentiating the HB formed by R112 that stabilize the open conformation.

**Table 2** Conformational energies ($E_{CONF}$) calculated for all atoms of the proteins during the simulation at different temperatures

| T(K) | BsXA/ $E_{CONF}$ (kcal/mol) | Mutant/$E_{CONF}$ (kcal/mol) |
|---|---|---|
| 300 | −3810.3±34.6 | −3784.8±50.3 |
| 328 | −3794.5±40.6 | −3744.5±43.4 |
| 338 | −3807.7±54.3 | −3725.1±41.1 |

## Conclusions

The present work provides insights as to the thumb region mobility in different conformational states. The thumb region in native BsXA assumes a closed conformation which is characteristic of the crystal structures of family GH11 xylanases, whereas the D11F/R112D mutant exhibits a recently observed open thumb conformation. Here we have shown that the high flexibility/mobility of the thumb region is independent of its initial open or closed conformation, and that the open to closed conformation transition performed by the thumb region is fully reversible. At higher temperatures, the open thumb conformation of the mutant closes and adopts a conformation similar to the native BsXA. The open or closed thumb conformations are energetically equivalent and therefore are both accessible to these enzymes, thus even though the crystal structure of the mutant shows a dimer, dimerization is not necessary to maintain the open thumb conformation. Furthermore, the thumb can adopt several open or closed conformations with similar energies and with equal probabilities of occurrence. From these results we propose that the transition between the two conformations occurs at high temperature and is a key determinant of access to the active site, and therefore influences substrate binding and product release. The R122D mutation contributes to the stability of the open conformation, but is not essential for thumb dynamics, i.e., the wild type enzyme can also adapt to the open conformation. Finally, our results indicate that the thumb conformation can be selected by understanding the factors that drive the functional movement, suggesting that the

physicochemical properties of the residues in the thumb region can be modulated to optimize local conformation changes in order to increase the catalytic activity at different temperatures. This work supports the idea that the thumb region of the GH11 xylanases makes a major contribution to the overall flexibility of the protein. This flexibility is responsible for the variability of the minimum energy conformations that populate the conformational space for this enzyme, and is functionally important to their biological activity.

**Acknowledgments** This work was founded by Fundação de Amparo à Pesquisa do Estado de São Paulo (2009/50838-5) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (152669/2007-8) fellowships.

## References

1. Wakarchuk WW, Campbell RL, Sung WL, Voodi J, Yaguchi M (1994) Mutational and crystallographic analyses of the active site residues of the Bacillus circulans xylanase. Protein Sci 3:467–475
2. Vieira DS, Degrève L, Ward RJ (2009) Characterization of temperature dependence and substrate-binding cleft movements in Bacillus circulans family 11 xylanase: a molecular dynamics investigation. Biochim Biophys Act 1790:1301–1306
3. Pollet A, Vandermarliere E, Lammertyn J, Strelkov SV, Delcour JA, Courtin CM (2009) Crystallographic and activity-based evidence for thumb flexibility and its relevance in glycoside hydrolase family 11 xylanases. Proteins 77:395–403
4. Vieira DS, Degrève L (2009) An insight into the thermostability of a pair of xylanase: the role of hydrogen bonds. Mol Phys 1:59–69
5. Kulkarni N, Shendy A, Rao M (1999) Molecular and biotechnological aspects of xylanases. FEMS Microbiol Rev 23:411–456
6. Henrissat B, Bairoch A (1996) Updating the sequence-based classification of glycosyl hydrolases. Biochem J 316:695–696
7. Murakami MT, Arni RK, Vieira DS, Degrève L, Ruller R, Ward RJ (2005) Correlation of temperature induced conformation change with optimum catalytic activity in the recombinant GH11 xylanase A from strain 168 (1A1). FEBS Lett 579:6505–6510
8. Gruber K, Klintschar G, Hayn M, Schlacher A, Steiner W, Kratky C (1998) Thermophilic xylanase from Thermomyces lanuginosus: high-resolution X-ray structure and modeling studies. Biochemistry 37:13475–13485
9. Muilu J, Törronen A, Peräkyä M, Rouvinen J (1998) Functional conformational changes of Endo-1,4-xylanase II from trichoderma reesei: a molecular dynamics study. Proteins 31:434–444
10. Törrönen A, Rouvinen J (1995) Structural comparison of 2 major endo-1,4-xylanases from Trichoderma reesei. Biochemistry 34:847–856
11. Havukainen R, Törrönen A, Laitinen T, Rouvinen J (1996) Covalent binding of three epoxyalkyl xylosides to the active site of endo-1,4-xylanase II from Trichoderma reesei. Biochemistry 35:9617–9624
12. Karplus M, McCammon A (2002) Molecular dynamics simulations of biomolecules. Nat Struct Biol 9:646–652
13. Klepeis JL, Lindorff-Larsen K, Dror RO, Shaw DE (2009) Long-timescale molecular dynamics simulations of protein structure and function. Curr Opin Struct Biol 19:120–127
14. Paës G, Tran V, Takahashi M, Boukari I, O'Donohue MJ (2007) New insights into the role of thumb-like loop in GH-II xylanases. Protein Eng Des Selec 20:15–23
15. Cortés J, Simeon T, de Ângulo VR, Guieysse AD, Remaud-Simeon M, Tran V (2005) A path planning approach for computing large-amplitude motions of flexible molecules. Bioinformatics 21:I116–I125
16. Bussi G, Donadio D, Parrinello M (2006) Canonical sampling through velocity rescaling. J Chem Phys 126:14101–14108
17. Berendsen HJC, Postma JPM, DiNola A, Haak JR (1984) Constant temperature molecular dynamics simulations of energetic particle-solid collisions: comparison of temperature control methods. J Chem Phys 81:3684–3690
18. Berendsen HJC, Postma JPM, van Gunsteren WF, Hermans J (1981) Interactions models for water in relation to protein hydration. In: Pullman B (ed) Intermolecular forces. Reidel , Dordrecht
19. Hess B, Becker H, Berendsen HJC, Fraaije JGEM (1997) LINCS: a linear constraint solver for molecular simulations. J Comput Chem 18:1463–1472
20. Miyamo S, Kollman PA (1992) SETTLE: an analytical version of the SHAKE and RATTLE algorithm for rigid water models. J Comput Chem 13:952–962
21. van Gunsteren WF, Berendsen HJC (1988) A leap-frog algorithm for stochastic dynamics. Mol Sim 1:173–185
22. Darden T, York D, Pedersen L (1993) Particle mesh Ewald-an N. Log(N) method for Ewald sums in large systems. J Chem Phys 98:10089–10092
23. Arfken G (1985) The method of steepest descents in mathematical methods for physicists, 3rd edn. Academic Press, Orlando, p 1985
24. Gordon JC, Myers JB, Folta T, Shoja V, Health LS, Onufriev A (2005) A server for estimating pKas and adding missing hydrogens to macromolecules. Nucleic Acids Res 33:W368–W371
25. McDonald IK, Thornton JM (1994) Satisfying hydrogen bonding potential in proteins. J Mol Biol 238:777–793
26. Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) GROMACS 4: algorithms for highly efficient, load-balanced and scalable molecular simulation. J Chem Theor Comput 4:435–447
27. van Gunsteren WF, Billeter SR, Eising AA, Hünenberger PH, Krüger P, Mark AE, Scott WRP, Tironi IG (1996) In Biomolecular Simulation: the GROMOS96 Manual and User Guide, Biomos, Groningen
28. Wakarchuk WW, Sung WL, Campbell RL, Cunningham A, Watson DC, Yaguchi M (1994) Thermostabilization of the *Bacillus circulans* xylanase by the introduction of disulfide bonds. Protein Eng 7:1379–1386
29. Derreumaux P, Schilk T (1998) The loop opening/closing motion of the enzyme triosephosphate isomerase. Biophys J 74:72–81
30. Ramasubu N, Ragunath C, Mishra PJ (2003) Probing the role of a mobile loop in substrate binding and enzyme activity of human salivary amylase. J Mol Biol 325:1061–1076
31. Gulotta M, Deng H, Deng H, Dyer RB, Callender RH (2002) Toward an understanding of the role of dynamics on enzymatic catalysis in lactate dehydrogenase. Biochemistry 41:3353–3363
32. Vandermarliere E, Bourgois TM, Rombouts S, Van Campenhout S, Volckaert G, Strelkov SV, Delcour JA, Rabijins A, Courtin CM (2008) Crystallographic analysis shows substrate binding at the -3 to +1 active-site subsites and at the surface of glycoside hydrolase family 11 endo-1,4-beta-xylanases. Biochemistry J 410:71–79

J Mol Model (2012) 18:1481–1493
DOI 10.1007/s00894-011-1177-2

ORIGINAL PAPER

# De novo design of 7-aminocoumarin derivatives as novel falcipain-3 inhibitors

**Anand S. Chintakrindi · Mushtaque S. Shaikh · Evans C. Coutinho**

Received: 13 April 2011 / Accepted: 1 July 2011 / Published online: 23 July 2011


**Abstract** The availability of the crystal structure of falcipain-3, knowledge of the peptides carrying the 7-aminocoumarin moiety as falcipain-3 ligands/substrates, and a need for new antimalarial agents stimulated us to look at the possibility of finding some novel falcipain-3 inhibitors. In this paper, we report the effect of substitution at the 7-amino position of the coumarin nucleus on the inhibition of falcipain-3, which is a well-validated antimalarial target. The de novo drug design was assisted by QSAR studies that shed light on the binding patterns of known and the newly designed inhibitors, thus taking this discovery process to the next level.

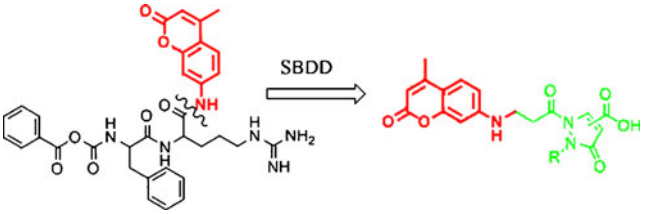
**Keywords** Falcipain · *P. falciparum* · 7-Aminocoumarins · Antimalarials · Ludi · CoMFA · CoMSIA · CoRIA

## Introduction

Malaria has been a major cause of human suffering for thousands of years, and despite significant advancements in our understanding of the disease, it continues to be one of the greatest causes of serious illness and death in the world. According to a recent report, malaria accounted for an estimated one million deaths in 2008, of which 89% were in the African region, followed by the Eastern Mediterranean (6%) and the South-East Asian regions (5%) [1]. The ever-increasing resistance to existing antimalarials makes the development of new, safe and effective antimalarial drugs an important goal of malaria control. This goal could be achieved in two ways: by focusing on validated targets in order to generate new drug candidates, or by identifying new potential targets for malaria chemotherapy. Among the potential new targets for antimalarial chemotherapy are cysteine proteases—enzymes which cleave small peptide substrates by the action of a nucleophilic thiol group of a cysteine residue, and thus play various important physiological roles [2]. The differences in the nature of cysteine proteases in parasites and in humans have provoked interest in this class of proteases. Also, it has been shown that cysteine protease inhibitors arrest the erythrocytic life cycle of *Plasmodium falciparum*, the most virulent human malaria parasite, by blocking the hydrolysis of the host hemoglobin [3]. *P. falciparum* contains four typical cysteine proteases belonging to the papain family that are known as falcipains: falcipain-1, falcipain-2, falcipain-2′, and falcipain-3. Incidentally, the role of falcipain-1 (FP1) is yet not clear, while the biochemical roles of falcipain-2 (FP2) and its isotype falcipain-2′ (FP2′) are very similar to each other. The two proteases falcipain-2 and falcipain-3 (FP3) have been studied comprehensively in the past, and genes coding for them have even been expressed in trophozoites. They have been found to be localized in the food vacuole, and are able to degrade hemoglobin [3]. So far, several peptidic, peptidomimetic, as well as nonpeptidic cysteine protease inhibitors have been synthesized and tested against malaria parasites [4].

One of the strategies employed in the development of new inhibitors against any target is the modification of a known ligand using de novo techniques. This strategy is the basis of the present work. Benzyloxycarbonyl-Phe-Arg-7-amino-4-methylcoumarin (Scheme 1) had been previously developed as a ligand for falcipain-3 [2]. In an earlier study, Sabins et al. reported homology models of both falcipain-2

A. S. Chintakrindi · M. S. Shaikh · E. C. Coutinho (✉)
Department of Pharmaceutical Chemistry,
Bombay College of Pharmacy,
Kalina, Santacruz (E), Mumbai 400 098, India
e-mail: evans@bcpindia.org




**Scheme 1** Logic of designing 7-amino coumarin derivatives from benzyloxycarbonyl-Phe-Arg-7-amino-4-methylcoumarin

and falcipain-3 and docking studies of amino-4-methylcoumarin and other derivatives [5]. In the present work, a set of novel nonpeptidic inhibitors has been developed by deleting the dipeptide (Phe-Arg) portion from this ligand and appending groups complementary to residues in the active site around the amino function of 7-amino-4-methylcoumarin. Further, we have also built 3D-QSAR models to predict the activities of these molecules and to help us fine tune their structures so as to improve their activity.

## Computational details

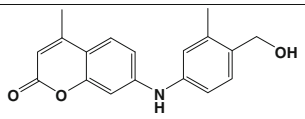
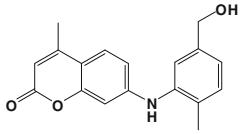
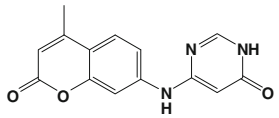
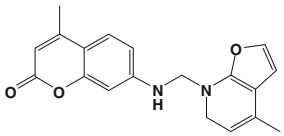
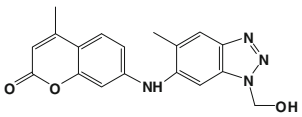
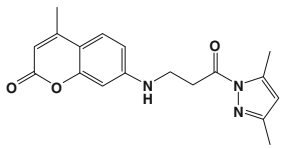
The computational studies were carried on a PC with an AMD 3.0 GHz processor and 1.5 GB RAM operating under the RedHat Linux Enterprise WS4 OS. Structures of molecules were built and their geometries were optimized with the suite of programs in Sybyl (v.7.1, Tripos Inc. St. Louis, MO, USA) [6]. The molecular docking was carried out using the software GOLD (v.3.1.1, CCDC, Cambridge, UK) [7]. De novo design of inhibitors was done using the LUDI module of Insight II (Accelrys Inc., San Diego, CA, USA) [8]. The 3D-QSAR techniques CoMFA (comparative molecular field analysis) and CoMSIA (comparative molecular similarity indices analysis) were carried out with Sybyl running on the abovementioned computer, while CoRIA (comparative residue interaction analysis) studies were carried out using Cerius 2 (v.4.11, Accelrys Inc.) [9] running on a SGI Fuel machine with the IRIX 6.4 OS.

### Investigating the ligand–protein interaction by docking

#### *Preparation of the ligand for docking*

All molecules (Tables 1, 2, 3, and 4 were built using Sybyl. The falcipain-3 inhibitors used in the QSAR studies were taken from a recent series of papers [4, 10, 11]. The geometries of ligands were optimized by the conjugate gradient method of energy minimization with the Tripos force field and Gasteiger–Hückel charges for all atoms until a gradient of 0.01 kcal $mol^{-1}$ $Å^{-1}$ was reached. The dielectric constant was set to 1.0.

**Table 1** Some inhibitors of falcipain-3 designed using the de novo approach of Ludi

| Sr. no | Structure | Energy estimate_2 |
|---|---|---|
| 01 |  | 560 |
| 02 |  | 553 |
| 03 |  | 531 |
| 04 |  | 489 |
| 05 |  | 482 |
| 06 |  | 416 |

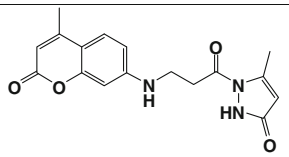
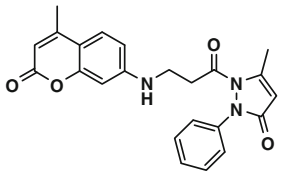
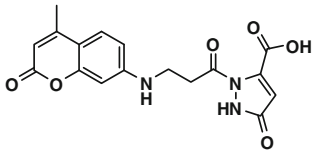
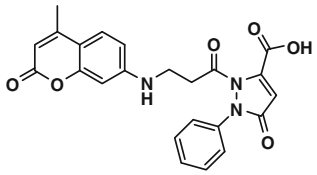
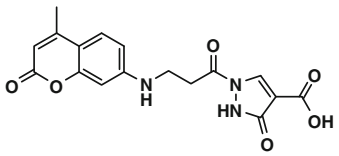
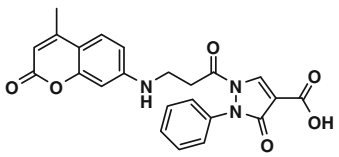
#### *Receptor preparation for docking*

The crystal structure of falcipain-3 with the PDB code 3BPM [12] was used for the docking studies. The enzyme exists as a dimer in the crystal. The crystal structure was "cleaned" by deleting the ligand and water molecules, and then hydrogens were added to all heavy atoms corresponding to a pH of 4.5 using the Biopolymer module in Insight II. To facilitate compatibility with the GOLD program, Sybyl atom and bond types were assigned to the protein atoms.

#### *Docking protocol*

With any docking program, it is necessary to determine its effectiveness for the protein under study. The crystal structure of falcipain-3 with the PDB code 3BPM [12] was used for the docking studies. The GOLD docking parameters were optimized so as to reproduce the X-ray pose of the ligand E64 (Leupeptin) in the PDB structure. Residues within a 15 Å radius from Cys51 were defined as the active site for the docking study; they included Gln45, Ala46, Leu47, Cys48, Gly49, Ser50, Cys51, Cys89, Tyr90,

**Table 2** Structures of molecules optimized with docking studies

| Sr. no. | Structure | Gold score |
|---|---|---|
| 7 |  | 46.85 |
| 8 |  | 45.80 |
| 9 |  | 46.64 |
| 10 |  | 43.33 |
| 11 |  | 51.18 |
| 12 |  | 51.32 |

Gly91, Gly92, Ile159, Ala161, Ala180, Pro181, Asn182, His183, Asn213, Trp215, and Gly216. The program was set to terminate when twenty solutions within an RMSD of 1.5 Å were identified. Options such as flip ring corners, flip amide bonds, and flip pyramidal nitrogen were activated, while unionized carboxylates were allowed to rotate so as to form favorable H-bonds. Docking was carried out with the default GA settings, and the number of GA runs was set to 20. The fitness function GoldScore was used to score the poses.

## De novo design of inhibitors

### *Protein and ligand preparation*

The ligand benzyloxycarbonyl-Phe-Arg-7-amino-4-methylcoumarin was docked into the enzyme falcipain-3 using GOLD. The benzyloxycarbonyl-Phe-Arg portion of the ligand was deleted, leaving the 7-amino-4-methylcoumarin moiety at the active site of the enzyme. An assembly of falcipain-3 with 7-amino-4-methylcoumarin was made from this pose with the Biopolymer module of Insight II.

### *Ludi protocol*

The active site of the enzyme defined in the docking study was considered for de novo design using the program Ludi [8]. Ludi was run in the standard mode with the Zinc database [13]. The Ligand Spec parameter was toggled on; this specifies the ligand on which substitutions are to be made. The Linkage parameter was set to "Specified," and both hydrogens of the amino group of 7-amino-4-methylcoumarin were specified as link sites. The Cys51 residue of falcipain-3 was defined as the center of the search, and the radius of the search was set to 15 Å. A conformational search of the fragments being added was done, considering attachment through one rotatable bond. The remaining parameters were considered at their default settings. The Ludi scoring function (Energy Estimate_2) [8] was used to prioritize the hits from the Zinc database.

## QSAR studies of falcipain-3 inhibitors

We used CoMFA, CoMSIA and CoRIA to build 3D-QSAR models. These models were used to predict the activities of molecules developed as potential inhibitors, and also to fine tune their structures to improve their activities.

### *Selection of the training and test sets*

The best conformations of the molecules found by docking were used for subsequent QSAR studies. These best docked conformations were aligned with respect to the receptor (i.e., "receptor-based" alignment) to form the test and training sets for QSAR. The 32 molecules (Tables 3 and 4) were split into a training set (24 molecules) and a test set (8 molecules) using the Library analysis/score option in the CombiChem-I module in Cerius 2. The selection was done on the basis of structural, chemical and biological diversity using a similarity search technique. The dataset was split into training and test sets containing diverse molecules based on the D-Optimal design [14]. The remaining parameters (normalization criteria, number of Monte Carlo steps, Monte Carlo temperature, Monte Carlo seed, Gaussian alpha, bucket solution for $K_d$ tree, and the number of termination steps for the D-Optimal design) were set to their default values. Other parameters that were considered for the diverse set selection were the default molecular properties defined in Cerius 2, like molecular weight, *A*log*P*, H donor, H acceptor, accessible surface area, and the experimental $pIC_{50}$ values.

**Table 3** Structures and $IC_{50}$ values of training set molecules

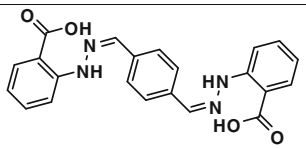
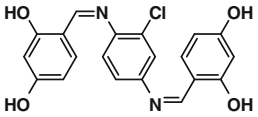
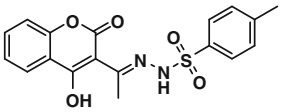
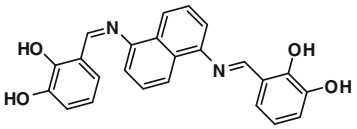
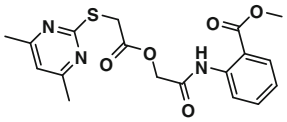
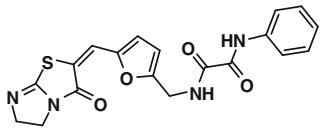
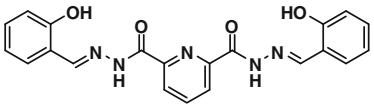
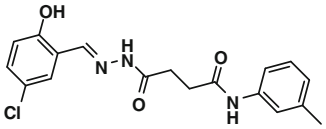
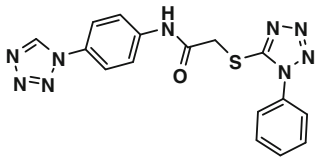
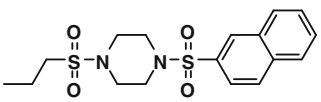
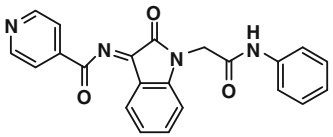
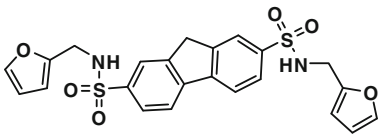
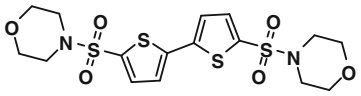
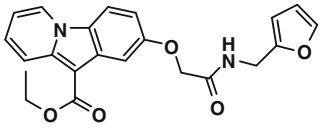
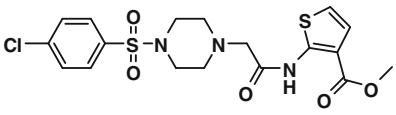
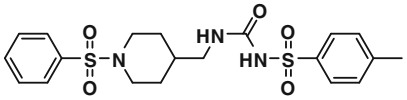
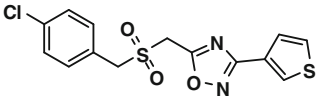
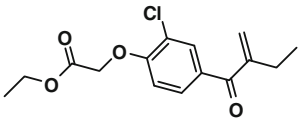
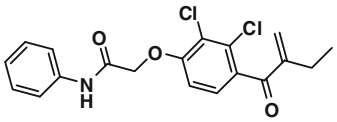
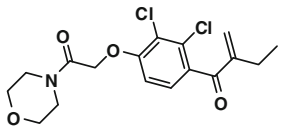
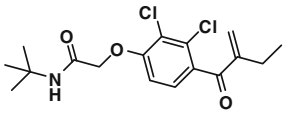
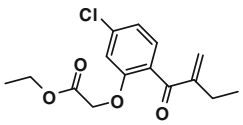
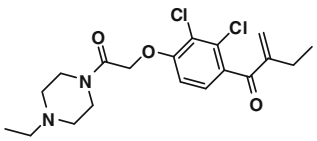
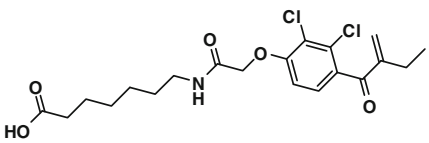
| Sr. no. | Structure | $IC_{50}$ (μM) | $pIC_{50}$ |
|---|---|---|---|
| 13 |  | 4.9 | 5.31 |
| 14 |  | 7.5 | 5.12 |
| 15 |  | 13.9 | 4.86 |
| 16 |  | 5.8 | 5.24 |
| 17 |  | 12 | 4.92 |
| 18 |  | 11.8 | 4.93 |
| 19 |  | 10.5 | 4.98 |
| 20 |  | 5.4 | 5.27 |
| 21 |  | 5.4 | 5.27 |
| 22 |  | 29.6 | 4.53 |
| 23 |  | 62.2 | 4.21 |
| 24 |  | 11.4 | 4.94 |

**Table 3** (continued)

| Sr. no. | Structure | $IC_{50}$ (μM) | $pIC_{50}$ |
|---|---|---|---|
| 25 |  | 37.7 | 4.42 |
| 26 |  | 31.2 | 4.51 |
| 27 |  | 49.2 | 4.31 |
| 28 |  | 18.8 | 4.73 |
| 29 |  | 56.6 | 4.25 |
| 30 |  | 346 | 3.46 |
| 31 |  | 87.2 | 4.06 |
| 32 |  | 479 | 3.32 |
| 33 |  | 153 | 3.82 |
| 34 |  | 381 | 3.42 |
| 35 |  | 557 | 3.25 |
| 36 |  | 96.5 | 4.02 |

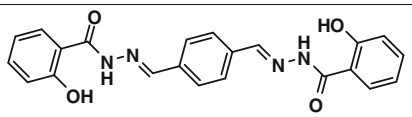
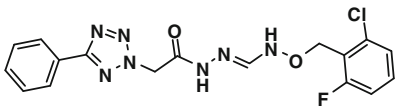
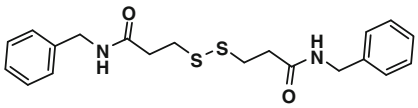
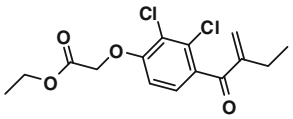
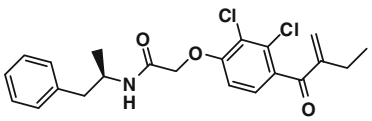
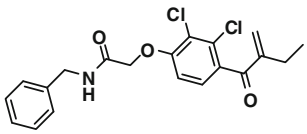
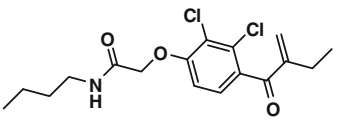
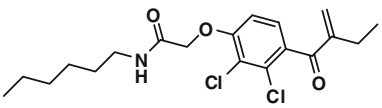
*Biological data*

For the QSAR study, the $IC_{50}$ values were converted to $pIC_{50}$ values by taking negative logarithms of the $IC_{50}$ values.

*Docking as a method of alignment*

The molecules in the training and test sets were docked into the active site of falcipain-3 and aligned by overlaying the

**Table 4** Structures and $IC_{50}$ values of test set molecules

| Sr. no | Structure | $IC_{50}$ (μM) | $pIC_{50}$ |
|---|---|---|---|
| 37 |  | 27.8 | 4.56 |
| 38 |  | 38.5 | 4.41 |
| 39 |  | 34 | 4.47 |
| 40 |  | 163 | 3.79 |
| 41 |  | 56.7 | 4.25 |
| 42 |  | 158 | 3.80 |
| 43 |  | 175 | 3.76 |
| 44 |  | 123 | 3.91 |

active site residues. This technique is often referred to as "receptor-based alignment" or "docking-based alignment" [15], and is known to produce better models than the standard procedure of atom-based alignment.

*CoMFA and CoMSIA studies*

Comparative molecular field analysis (CoMFA) [16] and comparative molecular similarity indices analysis (CoMSIA) [17–19] studies were carried out with the default settings for the 3D cubic lattice, the grid spacing, the probe atom, and the cut-off for the interaction energy. For CoMSIA, four physicochemical properties (steric, electrostatic, hydrophobic and hydrogen-bond donor) were explored, and the attenuation factor was set to the default value of 0.3. The CoMFA and CoMSIA descriptors were used as independent variables and the $pIC_{50}$ values as the dependent variable in a partial least squares (PLS) regression analysis to derive 3D-QSAR models. The models were internally evaluated by leave-one-out (LOO) cross-validation. The optimal number of components was determined by the SAMPLS [20] method, which were subsequently used to derive the final QSAR models. In addition to the $q^2$, the conventional regression coefficient $r^2$, the standard error (SE), and the $F$ value were also computed. The robustness of each model was gauged by cross-validation using leave-group-out (LGO) in a group of ten molecules and bootstrap [21, 22] carried out with 100 runs.

*CoRIA studies*

In CoRIA [23–26], the thermodynamics contained in the ligand receptor complex is incorporated into the QSAR formalism. The ligand–receptor complexes were obtained by docking with the GOLD program. The geometries of the ligands were optimized within the receptor complexes using molecular dynamics with the CFF91 [27] force field. At the end of molecular dynamics simulation, the structures were optimized by energy minimization with the steepest descent and conjugate gradient methods until a gradient of 0.01 kcal $mol^{-1}$ $Å^{-1}$ was reached. These refined complexes were used to calculate the CoRIA descriptors, as described below.

*Nonbonded interaction energies*

The van der Waals (vdW) and Coulombic interaction energies for the enzyme–ligand complexes were calculated with the CFF91 force field using the Discover 3 module of Insight II. The functional forms of the two nonbonded interaction energies are expressed as

$$E_{vdW} = \left(A_{ij}/r_{ij}^{12}\right) - \left(B_{ij}/r_{ij}^{6}\right) \quad (1)$$

$$E_{ele} = q_i q_j/\varepsilon r_{ij}. \quad (2)$$

Here, $A_{ij}$ and $B_{ij}$ are the repulsive and attractive term coefficients between atoms $i$ and $j$, respectively, $r_{ij}$ is the distance between atoms $i$ and $j$, $q_i$ and $q_j$ are the atomic charges of the interacting atoms $i$ and $j$, respectively, and $\varepsilon$ is the dielectric constant.

*Strain energy*

The strain energies were computed as the difference between the energies of the ligands in complex with the protein and the energies of the ligands after free minimization in vacuo. The ligands were extracted from their respective enzyme complexes and minimized in vacuo to relieve the strain due to constraints imposed by the protein environment. This was done with a combination of 5000 steepest descent steps followed by 50,000 conjugate gradient steps to an energy derivative of 0.01 kcal $mol^{-1}$ $Å^{-1}$.

*Solvent-accessible surface area*

Solvent-accessible surface area (SASA) is the residual surface of the ligand that is still accessible to the solvent after it has bonded to the receptor. It correlates with the tightness and more or less with the strength and number of binding interactions with the pocket. SASA was calculated with the POPS program [28].

*Solvation free energy*

The free energy of solvation of the ligand under physiological conditions is the hydration free energy, which is the difference between the hydration of the free and the bound states. It corresponds to the energy required to strip the solvent molecules off the ligand when changing from an aqueous environment to a hydrophobic receptor cavity. In many complexes, the conformation of the receptor does not change much from the uncomplexed structure, so the net solvation free energy for the receptor (free and bound) is negligible compared to that of the ligand. The electrostatic contribution to the solvation free energy was calculated with the Jaguar [29] module in the Schrödinger suite of programs.

*Lipophilicity*

The octanol/water partition coefficient (log$P$) is related to the hydrophobic characteristics of the molecule. It was calculated with the Ghose and Crippens [30] atom-based approach as implemented in Cerius 2 [9], where each atom of the molecule is assigned to a particular class. The total value of $A$log$P$ is the sum of the contributions of all of the atoms, which is calculated by the following equation:

$$A\log P = \sum_i n_i a_i, \quad (3)$$

where $n_i$ is the number of atoms of type $i$ and $a_i$ is the contribution from an atom of type $i$.

*Statistical analysis*

The statistical tool used to develop meaningful QSAR models was G/PLS. The G/PLS methodology combines a genetic function approximation (GFA) [31] that selects the appropriate descriptors for the QSAR model, while partial least squares (PLS) [32, 33] is the regression technique that weighs the relative contribution of each descriptor in the final model. Only linear terms were used to develop the QSAR equation. To preserve the information contained in the descriptor dataset, pretreatment based on the correlation matrix was avoided. The descriptors were scaled according to their mean and standard deviations, where each value in the column was subtracted from the column mean and then divided by the standard deviation of that column such that all scaled descriptors had a mean of zero and a standard deviation of unity. This scaling assigns equal weight to all descriptors and puts them on the same platform for a meaningful statistical analysis. The optimal number of components was selected for which the cross-validated $r^2$ ($q^2$) was found to be highest; the length of the equation was fixed, and the number of generations was limited to 10,000 with a population size of 500. Crossover and mutation probabilities of 50% (default setting) were used. The models developed were validated internally using leave-one-out (LOO), leave-group-out (LGO, group of 5), and by bootstrap procedures. Externally, the model was evaluated for its predictive power on the test set.

## Results and discussion

### De novo design of inhibitors

The design of new and selective inhibitors of an enzyme is one of the most important applications in contemporary rational drug design. When information on the binding site of the target protein is available, novel inhibitors can be

designed using de novo approaches. In an earlier study, Sabnis et al. studied benzyloxycarbonyl-Phe-Arg-7-amino-4-methylcoumarin as a substrate/ligand of falcipain-3. Sijwali et al. actually synthesized this series of compounds with the idea that attaching the coumarin ring to the dipeptide would help to elucidate its binding to the target via fluorescence experiments [5]. Eventually, that molecule turned out to be a good ligand itself. We used this molecule as a starting point for designing novel inhibitors of falcipain-3.

The structures of novel inhibitors suggested by the de novo approach are shown in Table 1, along with their Ludi scores. These molecules have the properties of lead-like molecules, thus they have enough scope for optimization and making them to follow Lipinski's rule of five [34]. These molecules offer a basic lead structure for recognition within the falcipain-3 active site, and can be further optimized for potency and selectivity.

These molecules were docked into the active site of falcipain-3 to confirm their binding poses. Molecule **6** was selected for further development based on its docking score (GoldScore). Analogs of molecule **6** (i.e., molecules **7**–**12**) were designed by modifying the substituents on the pyrazole ring so as to strengthen the interactions with the active site residues (Gln45, Ser50, Cys51, Asn182, Trp215). The structures of these molecules are listed in Table 2.

## Docking studies

As mentioned earlier, Sabnis et al. also reported docking studies of coumarin-tagged dipeptides. The present study is an extension of and improvement on that study in two aspects. Firstly, a homology model was used in the earlier study, while the more accurate crystal structure of the protein–ligand complex has been used in the present study. Secondly, unlike Sabnis et al., who performed preliminary studies, our intention was to design nonpeptidic derivatives of 7-aminocoumarins as antimalarial agents. The initial validation studies with the docking protocol could satisfactorily reproduce the binding conformation of the ligand leupeptin in the crystal structure. Molecules **7**–**12** (Table 2) were docked into the active site using this protocol. The interactions between the ligand and the residues in the active site are presented in Fig. 1.

The coumarin ring of these molecules π-stacks with Trp215 except for molecules **11** and **12**. These two molecules dock in exactly the opposite orientation compared to the other molecules. The amide group of molecules **7** and **8** is directed towards Cys51, which is suitably poised to form hydrogen bonds. The carboxyl group on the pyrazole ring of molecules **9** and **10** is seen to form hydrogen bonds with the thiol group in Cys51. It should be noted that Cys51 is crucial for inhibition, since it is involved in nucleophilic attack on the peptide bond of the substrate that results in hydrolysis of the substrate [35]. The amino group on the coumarin ring of **11** and **12** is located near Cys51, which is disposed to form hydrogen bonds. Although the orientations of these last two molecules at the active site are different, their docking scores are higher than the other molecules.

**Fig. 1** Docking poses of molecules **7**–**12** along with their GoldScores: **a** molecule **7**, **b** molecule **8**, **c** molecule **9**, **d** molecule **10**, **e** molecule **11**, **f** molecule **12** ▶

## QSAR studies

We have made use of the CoMFA, CoMSIA and CoRIA models to predict the activities of molecules **6**–**12**, and also to help us to fine tune their structures so as to improve their activities. The 3D-QSAR models were generated from diverse molecules that formed the training set (Table 3) and were externally validated by predicting the activities of molecules in the test set (Table 4).

### *CoMFA and CoMSIA models*

A nonpredictive pharmacophore-based model for the dataset has been reported by Desai et al. [11]. As seen in Tables 3 and 4, the molecules are quite diverse, which makes the dataset ideal for QSAR studies. Since the ligands are structurally diverse, it is difficult to align them by the atom-based technique. As a result, the molecules in Tables 3 and 4 were aligned by the receptor based technique. The statistical parameters for the 3D-QSAR models are robust, as is evident from Table 5. The predictive $r^2$ ($r^2_{\text{pred}}$) values for these 3D-QSAR models are given in Table 5.

Moderate values of $r^2_{\text{pred}}$ for models generated using receptor-based alignment is an indication that the receptor-based alignment can adequately account for the interactions that these molecules make with the enzyme active site. The $r^2_{\text{pred}}$ values show that the models can be used confidently to predict the activities of new molecules outside the training set but within the physicochemical space of the training set.

The CoMFA and CoMSIA field contours are very useful for understanding the SAR and designing new molecules. The CoMFA and CoMSIA steric and electrostatic contours around molecule **7** are shown in Figs. 2 and 3. The CoMFA steric contours suggest that the addition of bulk around the pyrazole ring would lead to an increase in potency. This can be achieved if the 5-methyl group in the pyrazole ring is replaced with bulkier groups. Some yellow contours are observed in the vicinity of molecule **7**, which indicates that steric bulk is disfavored. Simultaneous comparison of the receptor topology and the contours also reveals that

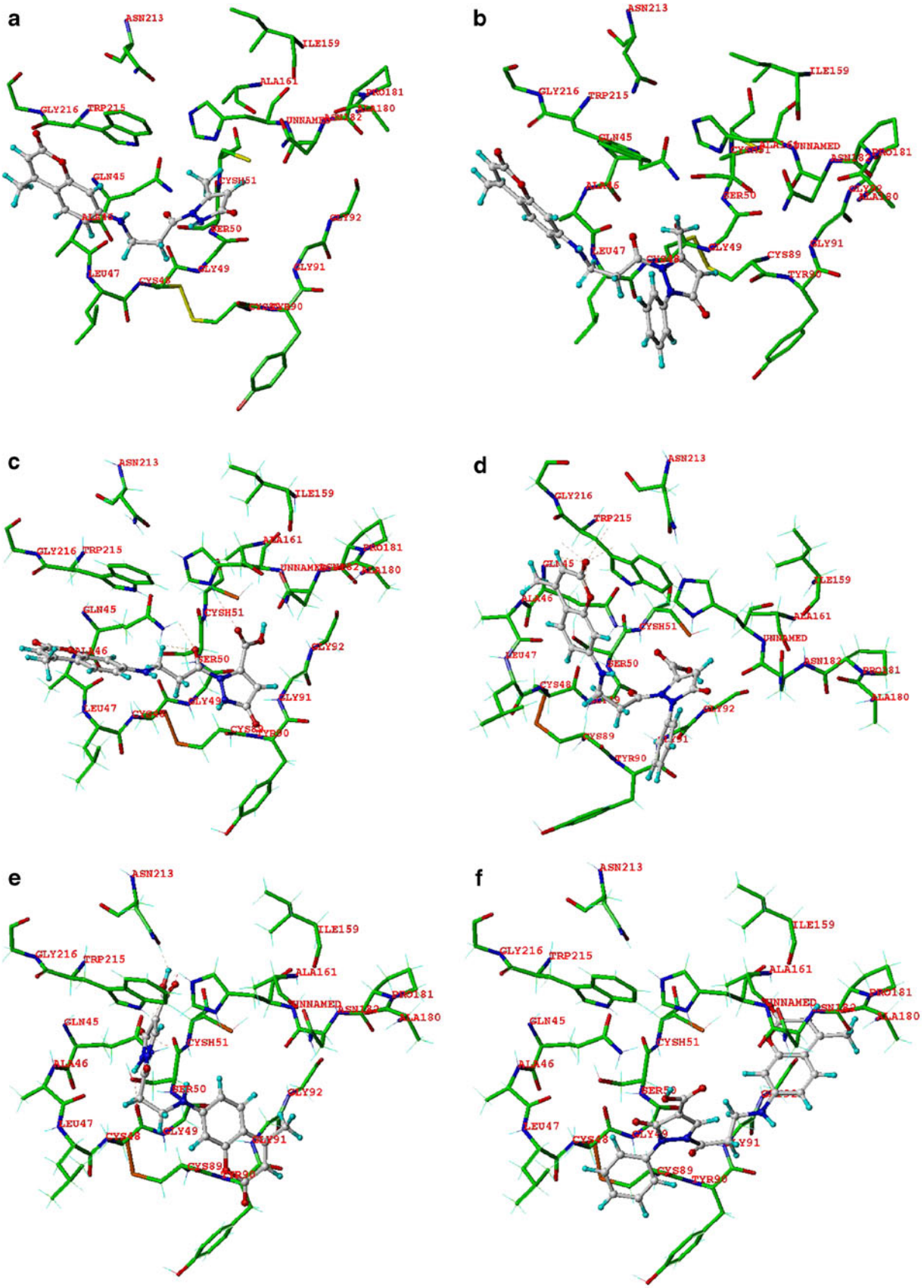

a
ASN213
ILE159
ALA161
PRO181
GLY216
TRP215
ALA180
GLN45
CYSH51
ALA46
GLY92
SER50
LEU47
GLY49
GLY91
b
ASN213
ILE159
GLY216
TRP215
GLN45
UNNAMED
PRO181
ALA46
SER50
LEU47
GLY49
CYS89
GLY91
TYR90
c
ASN213
ILE159
ALA161
GLY216
TRP215
PRO181
ALA180
GLN45
CYSH51
GLY92
SER50
LEU47
GLY49
GLY91
d
ASN213
GLY216
TRP215
GLN45
ILE159
ALA46
ALA161
CYSH51
UNNAMED
LEU47
SER50
ASN182
PRO181
CYS48
ALA180
GLY92
CYS89
TYR90
e
ASN213
ILE159
GLY216
TRP215
ALA161
UNNAMED
PRO181
GLN45
ALA180
CYSH51
ALA46
SER50
GLY92
LEU47
GLY49
GLY91
CYS89
f
ASN213
ILE159
GLY216
TRP215
ALA161
UNNAMED
PRO181
GLN45
ALA180
CYSH51
ALA46
LEU47
CYS89
TYR90

**Table 5** PLS statistics of CoMFA and CoMSIA models derived by receptor-based alignment

| PLS statistics | Receptor-based alignment | |
|---|---|---|
| | CoMFA | CoMSIA |
| $n$ | 24 | 24 |
| $r^2$ | 0.98 | 0.97 |
| $q^2$ (LOO) | 0.429 | 0.461 |
| $q^2$ (LGO) | 0.427 | 0.427 |
| $r_{BS}{}^2$ | 0.99 | 0.98 |
| $r^2{}_{pred}$ | 0.43 | 0.46 |
| $F$ | 331 | 283 |
| PLS components | 4 | 3 |
| Field contributions | | |
| Steric | 0.662 | 0.188 |
| Electrostatic | 0.338 | 0.158 |
| Hydrophobic | – | 0.295 |
| H donor | – | 0.358 |
| H acceptor | – | – |

unfavorable regions are superimposable on the backbone of the protein (Fig. 3a). Electropositive groups are favored on the pyrazole ring; the 3-keto group on the pyrazole ring satisfies the need for electronegative groups (Fig. 3b).

The CoMSIA steric contours (Fig. 3a) are in accordance with CoMFA contours, although they also suggest that the activity can be enhanced by adding bulky substituents to the phenyl part of the coumarin ring. The CoMSIA electrostatic contours (Fig. 3b) provide additional information over the CoMFA electrostatic fields. Electropositive groups are favored on the phenyl moiety of the coumarin ring, as well as on the 3-oxopropylamino portion of the molecule. The hydrophobic yellow contours (Fig. 3c) suggest that hydrophobic groups are favored at positions 4 and 5 of the pyrazole ring. However, some white contours could be observed around both nitrogens of the pyrazole ring, which indicates that hydrophobicity is disfavored around this space. Though the H-bond donor contours dominate in the CoMSIA model, the favored and disfavored contours are adjacent to each other and limit their application. Docking results also suggest that the active pocket is largely hydrophobic, with H-bonding groups being rather rare.

*CoRIA models*

CoRIA is a receptor-based 3D-QSAR method, so it is able to provide a greater understanding of ligand–receptor binding than the ligand-based 3D-QSAR methods CoMFA and CoMSIA. The CoRIA models, along with their statistical data, are reported in Table 6. The best model is

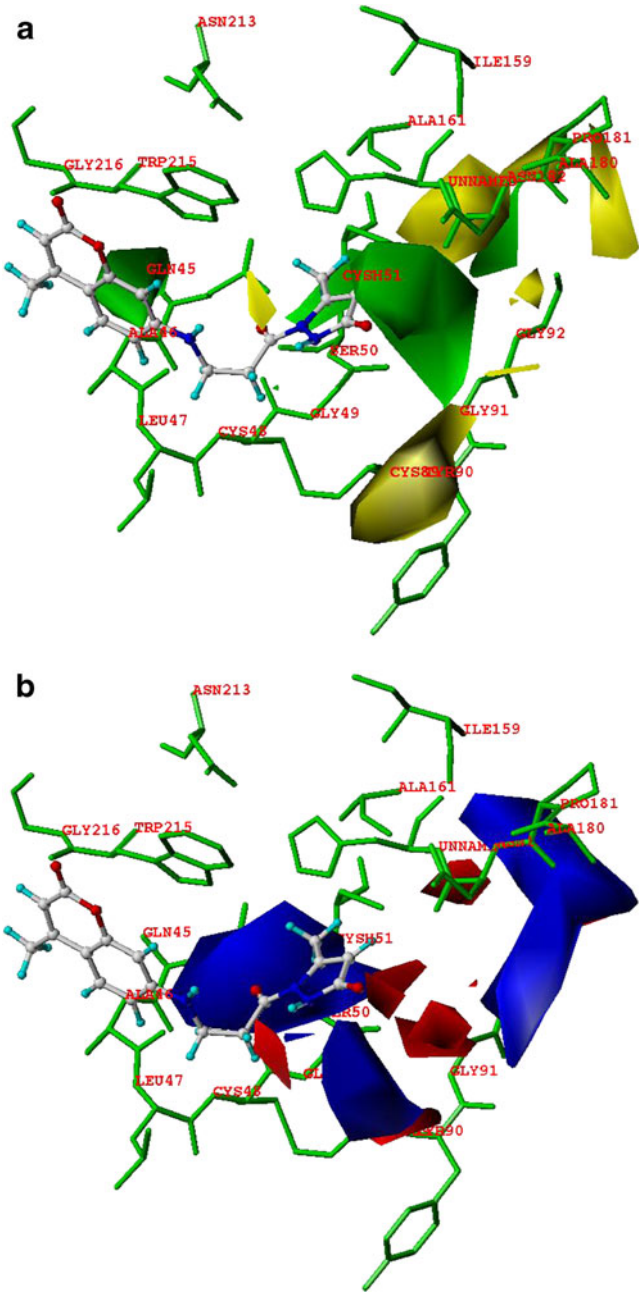
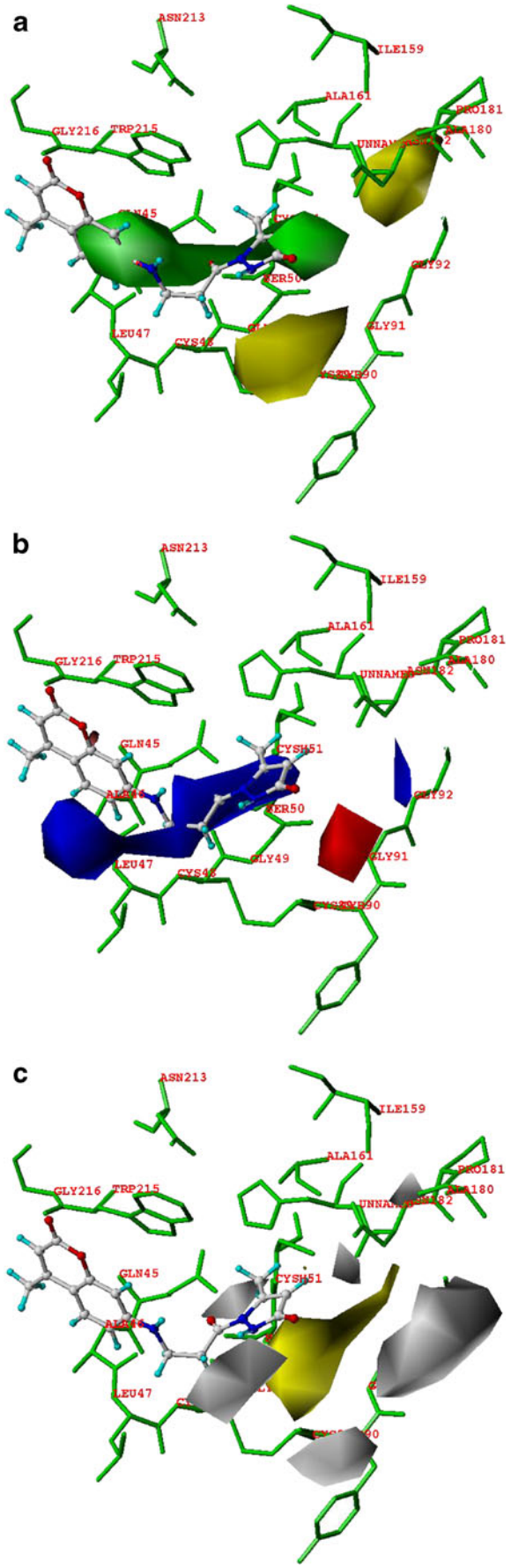



**Fig. 2** The CoMFA contours around molecule **7**. **a** Steric contours: favored (*green*), disfavored (*yellow*). **b** Electrostatic contours: electropositive (*blue*), electronegative (*red*)

statistically significant, with a correlation coefficient ($r^2$) of more than 0.9. The cross-validation correlation coefficients ($r_{CV}{}^2$) obtained by the leave-one-out (LOO) and leave-five-out (LGO) procedures and the bootstrap values are statistically significant. The predictive $r^2$ ($r^2{}_{pred}$) for the test set was also more than 0.6, indicating good predictive power of the model for molecules outside the training set.

In the CoRIA approach, the central elements are the interaction energies of the entire ligand with individual residues in the active site of the receptor. Analysis of the




◀ **Fig. 3** The CoMSIA contours around molecule **7**. **a** Steric contours: favored (*green*), disfavored (*yellow*). **b** Electrostatic contours: electropositive (*blue*), electronegative (*red*). **c** Hydrophobic contours: favored (*yellow*), disfavored (*white*)

equations in Table 6 reveals that van der Waals interactions are the major forces driving the binding of the ligands with the receptor. *A*log*P* has a positive coefficient, suggesting that enhancing hydrophobic interactions of the ligands with the receptor may significantly improve their binding with the receptor. On the other hand, the van der Waals interaction energies of residues Ile157 and Ser156 with the ligands appear as negative coefficients in the equation. This indicates that an increase in the biological activity can be gained by strengthening the van der Waals interactions of the ligands with these residues. The Coulombic interaction energy of Gly58 has a negative coefficient, indicating that strengthening the Coulombic interaction of the ligand with this residue will increase binding affinity. The plots of experimental vs. predicted $p\mathrm{IC}_{50}$ values of the molecules in the training and test sets for the CoRIA model are shown in Figs. 4a and b. As far as *A*log*P* is concerned, it is in concordance with the steric field contributions in the CoMFA and CoMSIA studies.

### *Prediction of the activities of molecules 7–12 based on QSAR models*

All three QSAR models were generated from a structurally diverse set of molecules, and these models are based on the receptor-based alignment for CoMFA and CoMSIA and ligand–receptor interactions for CoRIA. Thus, these models should be useful for predicting the activities of molecules outside the training set, provided they fit well in the receptor pocket and satisfy the pharmacophoric/SAR requirement. The predicted $p\mathrm{IC}_{50}$ values for these molecules based on the three QSAR approaches are listed in Table 7. The CoMFA and CoMSIA models predict activities for all molecules at the microgram level, with molecule **7**

**Table 6** CoRIA equation and statistical analysis of the model

| | |
|---|---|
| Equation | $p\mathrm{IC}_{50} = 4.41 - 0.39 \times V_{\mathrm{ILE157}} + 1.09 \times Nb_{\mathrm{GLU243}} + 0.95 \times Nb_{\mathrm{SER156}} - 0.16 \times V_{\mathrm{SER156}} - 0.65 \times C_{\mathrm{GLY58}} - 0.25 \times G_{\mathrm{solv}} + 0.17 \times A\log P$ |
| $r^2$ | 0.935 |
| $r_{\mathrm{CV}}^2$ (LOO) | 0.874 |
| $r_{\mathrm{CV}}^2$ (LGO) | 0.861 |
| $r_{\mathrm{BS}}^2$ | 0.760 |
| $r^2_{\mathrm{pred}}$ | 0.619 |

*C* Coulombic, *V* van der Waals interactions, respectively

*Nb* total nonbonded interaction energies (*C*+*V*)

$G_{solv}$ free energy of solvation

*A*log*P* Ghose and Crippens' octanol/water partition coefficient

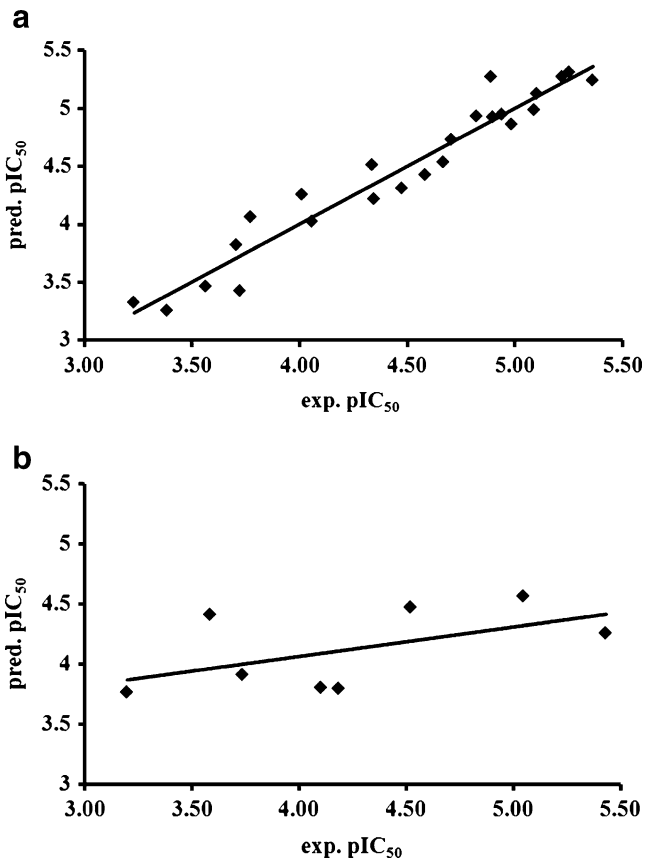



**Fig. 4** The plots of predicted vs. experimental activity from the model for **a** the training set and **b** the test set

predicted to be the most active. In slight contrast, the CoRIA model predicts that molecules **11** and **12** are the most active. This finding is in harmony with docking studies, where the docking scores for these two molecules are higher than those for the others (Fig. 1). This justifies the receptor-dependent nature of the CoRIA approach as a better abstraction of the ligand–receptor interaction. The predictive $r^2$ ($r^2_{pred}$) value obtained by CoRIA for all of the molecules is slightly higher than those calculated by the CoMFA and CoMSIA approaches (Tables 5 and 6), indicating that the CoRIA model is more reliable for predicting the activities of the newly designed molecules.

**Table 7** The predicted $pIC_{50}$ values for the newly designed molecules

| Molecule | CoMFA | CoMSIA | CoRIA |
|---|---|---|---|
| **6** | 4.64 | 4.47 | 5.17 |
| **7** | 5.02 | 4.77 | 3.51 |
| **8** | 4.69 | 4.48 | 4.61 |
| **9** | 4.84 | 4.75 | 3.40 |
| **10** | 4.75 | 4.62 | 1.92 |
| **11** | 4.35 | 4.76 | 8.03 |
| **12** | 4.50 | 4.05 | 5.21 |

## Conclusions

The present work was initiated with the intention of developing new falcipain-3 inhibitors. The well-known substrate benzyloxycarbonyl-Phe-Arg-7-amino-4-methyl-coumarin was docked into the falcipain-3 binding site, and the benzyloxycarbonyl-Phe-Arg portion was deleted from this pose, leaving the 7-amino-4-methylcoumarin moiety in the active site pocket of the enzyme. The assembly of falcipain-3 with 7-amino-4-methylcoumarin was used as a template for de novo design using Ludi. Among the several suggestions made by Ludi, one molecule (**6**) was selected for optimization, from which additional molecules (**7**–**12**) were designed. Various 3D-QSAR models were built using known inhibitors of falcipain-3, and the activities of these newly designed molecules **6**–**12** were predicted. Among various QSAR models that were constructed, the CoRIA model was found to be most reliable for predicting the activities of newly designed molecules.

**Acknowledgments** The authors are thankful to Dr. Prashant Desai, Eli Lilly, USA, for valuable input. The computational facilities at BCP (Bombay College of Pharmacy) were provided by the All India Council for Technical Education through grant f. no. 8022/RID/NPROJ/RPS-5/2003-04/. The Department of Science and Technology is also thanked for infrastructure facilities through the FIST program (SR/FST/LS1-163/2003).

## References

1. WHO (2009) World malaria report 2009. World Health Organization (WHO), Geneva
2. Sijwali PS, Shenai BR, Gut J et al (2001) Expression and characterization of the *Plasmodium falciparum* haemoglobinase falcipain-3. Biochem J 360:481–489
3. Verissimo E, Berry N, Gibbons P et al (2008) Design and synthesis of novel 2-pyridone peptidomimetic falcipain 2/3 inhibitors. Bioorg Med Chem Lett 18:4210–4214
4. Dude MA, Kaeppler U, Herb M et al (2009) Synthesis and evaluation of non-peptidic cysteine protease inhibitors of *P. falciparum* derived from ethacrynic acid. Molecules (Basel, Switzerland) 14:19–35
5. Sabnis YA, Desai PV, Rosenthal PJ et al (2003) Probing the structure of falcipain 3, a cysteine protease from *Plasmodium falciparum*: comparative protein modeling and docking studies. Protein Sci 12:501–509
6. Tripos Associates Inc. (2005) Sybyl, version 7.1. Tripos Associates Inc., St. Louis
7. CCDC (2006) GOLD, version 3.1. Cambridge Crystallographic Data Centre (CCDC), Cambridge
8. Accelrys, Inc. (2005) Insight II, version 2005L. Accelrys, Inc., San Diego
9. Accelrys, Inc. (2003) Cerius 2, version 4.8. Accelrys, Inc., San Diego
10. Desai PV, Patny A, Sabnis Y et al (2004) Identification of novel parasitic cysteine protease inhibitors using virtual screening. 1. The ChemBridge database. J Med Chem 47:6609–6615

11. Desai PV, Patny A, Gut J et al (2006) Identification of novel parasitic cysteine protease inhibitors by use of virtual screening. 2. The available chemical directory. J Med Chem 49:1576–1584
12. Kerr ID, Lee JH, Pandey KC et al (2009) Structures of falcipain-2 and falcipain-3 bound to small molecule inhibitors: implications for substrate specificity. J Med Chem 52:852–857
13. Irwin JJ, Shoichet BK (2005) ZINC—a free database of commercially available compounds for virtual screening. J Chem Inf Model 45:177–182
14. Baroni M, Clementi S, Cruciani G et al (1993) D-optimal designs in QSAR. Quant Struct Act Relat 12:225–231
15. Datar PA, Coutinho EC (2004) A CoMFA study of COX-2 inhibitors with receptor based alignment. J Mol Graphics Modell 23:239–251
16. Cramer RD III, Patterson DE, Bunce JD (1988) Comparative molecular field analysis (CoMFA) I. Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc 110:5959–5967
17. Klebe G (1998) Comparative molecular similarity indices analysis: CoMSIA. Perspect Drug Discovery Des 12:87–104
18. Klebe G, Abraham U, Mietzner T (1994) Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J Med Chem 37:4130–4146
19. B hm M, Sturzebecher J, Klebe G (1999) Three-dimensional quantitative structure–activity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin, and factor Xa. J Med Chem 42:458–477
20. Bush BL, Nachbar RB (1993) Sample-distance partial least squares: PLS optimized for many variables, with application to CoMFA. J Comput Aided Mol Des 7:587–619
21. Cramer RD III, Bunce JD, Patterson DE et al (1988) Cross-validation, bootstrapping, and partial least squares compared with multiple regression in conventional QSAR studies. Quant Struct Act Relat 7:18–25
22. Richard D, Cramer RD III, Bunce JD et al (1988) Crossvalidation, bootstrapping, and partial least squares compared with multiple regression in conventional QSAR studies. Quant Struct Act Relat 7:18–25
23. Datar PA, Khedkar SA, Malde AK et al (2006) Comparative residue interaction analysis (CoRIA): a 3D-QSAR approach to explore the binding contributions of active site residues with ligands. J Comput Aided Mol Des 20:343–360
24. Khedkar SA, Malde AK, Coutinho EC (2007) Design of inhibitors of the MurF enzyme of *Streptococcus pneumoniae* using docking, 3D-QSAR, and de novo design. J Chem Inf Model 47:1839–1846
25. Verma J, Khedkar VM, Prabhu AS et al (2008) A comprehensive analysis of the thermodynamic events involved in ligand-receptor binding using CoRIA and its variants. J Comput Aided Mol Des 22:91–104
26. Dhaked DK, Verma J, Saran A et al (2009) Exploring the binding of HIV-1 integrase inhibitors by comparative residue interaction analysis (CoRIA). J Mol Model 15:233–245
27. Maple JR, Hwang MJ, Stockfisch TP et al (1994) Derivation of class II force fields I. Methodology and quantum force field for the alkyl functional group and alkane molecules. J Comput Chem 15:162–182
28. Cavallo L, Kleinjung J, Fraternali F (2003) POPS: a fast algorithm for solvent accessible surface areas at atomic and residue level. Nucleic Acids Res 31:3364–3366
29. Schrodinger LLC (2007) Jaguar version 7.0: user manual, Schrodinger LLC, New York (see http://www.schrodinger.com)
30. Ghose AK, Crippen GM (1986) Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure–activity relationships. I. Partition coefficients as a measure of hydrophobicity. J Comput Chem 7:565–577
31. Rogers D, Hopfinger AJ (1994) Application of genetic function approximation to quantitative structure–activity relationships and quantitative structure–property relationships. J Chem Inf Model 34:854–866
32. Wold S, Johansson E, Cocchi M (1993) PLS: partial least squares projections to latent structures. In: Kubinyi H (ed) 3D QSAR in drug design: theory, methods and applications. ESCOM, Leiden, pp 523–550
33. Leach AR, Gillet VJ (2007) Computational Models. In: Leach AR, Gillet VJ (eds) Molecular modelling: principles and applications. Springer, Dordrecht, pp 75–97
34. Lipinski CA, Lombardo F, Dominy BW et al (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Del Rev 23:3–25
35. Rzychon M, Chmiel D, Stec-Niemczyk J (2004) Modes of inhibition of cysteine proteases. Acta Biochim Pol 51:861–873

J Mol Model (2012) 18:1495–1502
DOI 10.1007/s00894-011-1176-3

ORIGINAL PAPER

# Atomistic modeling of water diffusion in hydrolytic biomaterials

**Alfonso Gautieri · Andrea Mezzanzanica · Alberto Motta · Alberto Redealli · Simone Vesentini**

Received: 17 February 2011 / Accepted: 1 July 2011 / Published online: 23 July 2011


**Abstract** One of the most promising applications of hydrolytically degrading biomaterials is their use as drug release carriers. These uses, however, require that the degradation and diffusion of drug are reliably predicted, which is complex to achieve through present experimental methods. Atomistic modeling can help in the knowledge-based design of degrading biomaterials with tuned drug delivery properties, giving insights on the small molecules diffusivity at intermediate states of the degradation process. We present here an atomistic-based approach to investigate the diffusion of water (through which hydrolytic degradation occurs) in degrading bulk models of poly(lactic acid) or PLA. We determine the water diffusion coefficient for different swelling states of the polymeric matrix (from almost dry to pure water) and for different degrees of degradation. We show that water diffusivity is highly influenced by the swelling degree, while little or not influenced by the degradation state. This approach, giving water diffusivity for different states of the matrix, can be combined with diffusion-reaction analytical methods in order to predict the degradation path on longer time scales. Furthermore, atomistic approach can be used to investigate diffusion of other relevant small molecules, eventually leading to the *a priori* knowledge of degradable biomaterials transport properties, helping the design of the drug delivery systems.

A. Gautieri · A. Mezzanzanica · A. Motta · A. Redealli · S. Vesentini (✉)
Biomechanics Group,
Department of Bioengineering, Politecnico di Milano,
Via Golgi 39,
20133 Milan, Italy
e-mail: simone.vesentini@polimi.it

**Keywords** Atomistic modeling · Biodegradation · Diffusion · Drug delivery · Hydrolytic degradation

## Introduction

Biodegradable polymers are currently used in biomedical applications and regenerative medicine solutions. Significant applications include the development of biodegradable shape-memory polymeric networks [1], the design of drug delivery [2], absorbable sutures [3], orthopedic implants [4], scaffolds for tissue engineered constructs [5] and medicated and biodegradable stents [6].

For these systems, a detailed knowledge of the degradation behavior is of particular importance: the polymeric matrix should degrade in a controlled way since the degradation behavior directly determines the drug release kinetics. Within the class of biodegradable polymers, aliphatic poly(α-hydroxy esters) derived from lactic acid (i.e., PLA) have often been utilized as synthetic degradable biomaterials [7, 8]. However, a reliable model that predicts the degradation rate and kinetics for this class of biomaterials is not yet available, thus the development of models to obtain this knowledge is essential in order to attain effective control over their degradation behavior [6].

The prevailing mechanism of erosion of synthetic aliphatic polyesters is random scission of the hydrolytically unstable backbone chain by passive hydrolysis. The rate of swelling of aliphatic polyesters is usually much higher than the rate of hydrolysis, hence the reaction occurs extensively through the swollen polymer bulk due to water presence and the resulting mode of erosion observed in this class of polymers is therefore bulk erosion [9]. The degradation ultimately leads to monomers (i.e., lactic acid in PLA), which can be metabolized into carbon dioxide and water or

excreted. The transport of monomers and oligomers through the material is responsible for polymer erosion which leads to biomaterial mass loss

In order to use biodegradable polyesters for biomedical applications there is a strong need for reliable methods to characterize the kinetics of hydrolytic degradation. Nowadays, the relationship between the composition of a material and its degradation dynamics is still not sufficiently understood and biomaterial design is performed by time-consuming trial and error experimental procedures. Indeed, the degradation behavior of materials is mainly studied *in vitro* in buffer solution, in the presence of enzymes, cells, and tissues, and in animal models *in vivo* [10]. In this context, a major problem is the considerable scatter of degradation rates that is observed when different methods are employed for the same material. Recently, a theoretical model able to accomplish the degradation behavior of PLA has been developed [11]. This study provides a multi-scale model that describes both degradation (chain scission) and erosion (mass loss). It allows the investigation of both bulk and surface erosion modes, by unifying both behaviors with a non-dimensional parameter that relates the characteristic times of reaction (degradation) and diffusion (erosion) and is able to describe the entire spectrum of mixed behavior in between the two extreme cases. The water diffusion within a bulk leads to two different macroscopic mechanisms: surface or bulk erosion. The surface mechanism takes place when diffusing water molecules are absorbed quickly by hydrolysis and are limited from penetrating deep into the polymer bulk. In this case, erosion is restricted to the surface of the polymer, and accordingly this phenomenon is referred to as surface erosion. On the contrary bulk erosion takes place if hydrolysis is slower than the rate of diffusion of water molecules through the polymer. In this case penetrant molecules cannot be absorbed quickly enough to be hindered from reaching deep into the polymer and the reaction takes place through its entire swollen bulk. This multi-scale model requires as input the diffusion coefficient of water and the diffusion coefficients of each polymer sub-fraction that constitute the current state of the diffusing medium atomistic modeling has been successfully used to predict diffusivity of small molecules into polymeric materials [12–16]. Knowledge-based modeling techniques can therefore provide a more systematic and efficient method to predict the evolution of biodegradable polymers over the course of degradation and thus would enhance the biodegradable drug eluting implant design process [10, 17].

In this work, we present a systematic analysis of the diffusivity of water molecules, i.e. for different swelling states and for different degrees of degradation of the polymeric matrix. The motivations of this research are twofold. On one hand, we aim to provide quantitative data that serves as input for higher scale models, such as diffusion-reaction models proposed in ref [11]. In this way, the degradation kinetics of the polymeric matrix could be predicted by micro-scale model, which in turn rely on quantitative data on diffusivity provided by atomistic simulations. On the other hand, we aim to advance the field of molecular modeling, showing that this tool can be of great help in investigating not only static materials, but also biodegradable polymeric matrices as those considered in this study.

## Methods

### Models generation

In order to study how the water transport properties of PLA matrices change during the degradation process we generate several atomistic models of PLA, with increasing level of water content ($\phi$, modeling the swelling process) and with decreasing chain length ($i$, resulting from the hydrolysis process) (Fig. 1a). Indeed, within the classical molecular dynamics (MD) framework, it is not possible to model the covalent bond break, thus in the present work different bulk models with decreasing chain length account for the hydrolysis process. The molecular models construction and subsequent MD simulations are performed using the Materials Studio 4.0 software (Accelrys, Inc [18]) and the COMPASS force field [19]. The polymeric chains with the desired length are generated starting from the repeating units (i.e., lactic acid, available as standard model within Materials Studio) using the Build Polymers tool of Materials Studio. Finally, we generate solvated amorphous bulk models containing PLA and water using the Amorphous Cell tool of Materials Studio. The polymeric chains are solvated in a periodic box with a total atom number varying from≈6,000 atoms (for quasi-dry systems) to≈35,000 atoms (for highly solvated systems) (see Fig. 1b-c).

### Models equilibration

The initial geometries of the bulk models are refined following a procedure which consists in a series of MD calculations at different temperature and ensembles in order to obtain chain redistribution within the periodic cell [13, 17, 20, 21]. Specifically, we perform a preliminary minimization of the systems (using the steepest descent algorithm) followed by a sequence of nine MD simulations (see Table 1). In the first part of equilibration (simulations #1 to #4) the systems are subjected to initial NVT (constant number of particles, constant volume and constant temperature) simulations and a NPT (constant number of particles, constant pressure and constant temperature) molecular dynamics run in order to allow preliminary chain redistribution and

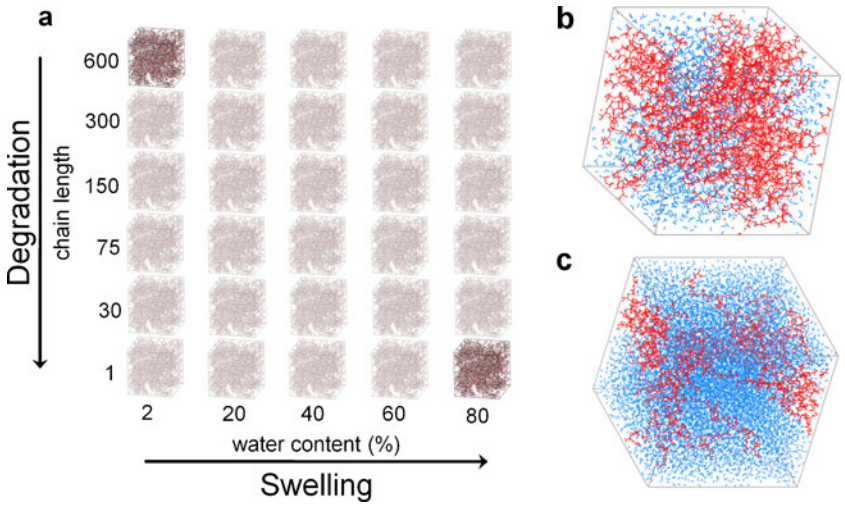



**Fig. 1** Schematics of the PLA bulk models considered in this study. A total of 30 PLA bulk models are considered in order to account for the different states of the polymer matrix during the degradation process (panel **a**). The upper left model represents the initial, dry and undegraded system, while the bottom right model represents a highly degraded and swollen matrix. The hydrolysis process considers decreasing PLA chains length (from 600 monomers per chain to single monomers), while the swelling process is modeled introducing a higher amount of water, from a water weight ratio of 2% to a weight ratio of 80%. Panel **b** shows a partially swollen unit cell (20% water content), whereas panel c represents a highly solvated system (80% water content)

density equilibration. In the second part of the equilibration (simulations #5 to #9) the packing models are subjected to simulated annealing in NVT ensemble (to allow chain redistribution) starting at a temperature of 750 K and cooling down to 300 K. A final NPT run is performed to allow for box volume (i.e., cell density) equilibration. MD molecular dynamics simulations are carried out using the Discover module implemented in Materials Studio. Nonbonding interactions are computed using a cutoff for neighbor list at 1 nm, with a switching function between 0.85 and 0.95 nm for van der Waals and Coulomb interactions. We carefully check that the potential energy, temperature, pressure and density reach a stable value after each step of the equilibration procedure. The latest simulation (#9) is used to validate the models by comparing the average density during the NPT dynamics with the experimental one.

## Fickian (*normal*) diffusion and random walk

In one of his ground-breaking papers that Einstein published in 1905 [22], he offered a statistical mechanics means for theoretical calculations involving Brownian motion. Einstein realized that the quantity involving Brownian motion which can be best observed under a microscope in an experiment is the "diffusivity":

$$D = \lim_{t\to\infty} \frac{\left\langle |r(t) - r(t_0)|^2 \right\rangle}{6t} \tag{1}$$

where $r(t)$ denotes the observed displacement of the Brownian particle at time $t$ and $\left\langle |r(t) - r(t_0)|^2 \right\rangle$ represent the mean square displacement (MSD). Einstein employed a random walk model for his analysis and showed that the diffusivity defined in Eq. 1 is identical to the diffusion

**Table 1** Summary of the simulation protocol applied to obtain equilibrated bulk model of PLA matrices

| Simulation | Ensemble | Time | Time step | Temperature | Pressure |
|---|---|---|---|---|---|
| #1 | NVT | 0.2 ps | 0.1 fs | 300 K | - |
| #2 | NVT | 2 ps | 1 fs | 600 K | - |
| #3 | NVT | 100 ps | 1 fs | 300 K | - |
| #4 | NPT | 60 ps | 1 fs | 300 K | 0 GPa |
| #5 | NVT | 20 ps | 1 fs | 750 K | - |
| #6 | NVT | 20 ps | 1 fs | 600 K | - |
| #7 | NVT | 20 ps | 1 fs | 450 K | - |
| #8 | NVT | 100 ps | 1 fs | 300 K | - |
| #9 | NPT | 100 ps | 1 fs | 300 K | 0 GPa |

constant that describes the macroscopic evolution of the concentration density $n(x, t)$ of a large number of Brownian particles:

$$\frac{\partial n(x,t)}{\partial t} = D\nabla^2 n(x,t). \quad (2)$$

Thus, the macroscopic Fickian (sometimes called "normal") diffusion can be seen at the microscopic scale as a random walk of a particles.

In the molecular dynamics approach, the diffusion coefficient is calculated by applying the Eq. 1 and thus assuming that the diffusing particles undergo random walk. It can be shown that in the limit $t << \tau$ (where $\tau$ is the collision time, the time between two collision of the diffusing particle), the mean square displacement is proportional to $t^2$, which is called "ballistic" diffusion and means that for short time particles are not hindered by collision yet and move very fast. On the other hand, in the other limit, $t >> \tau$, the mean square displacement is proportional to $t$, the motion is Brownian and Eq. 1 holds. Therefore, in order to properly apply Eq. 1 the simulation time should be sufficiently long to reach normal (fickian) diffusion regime, that is when MSD~$t$.

### Atomistic investigation of diffusion

In order to obtain water and polymer diffusivity we run NPT production simulation runs for a simulated time of 7 ns. We then select an ensemble of $M$ (water or polymer) molecules in the model and we compute their mean square displacement $MSD(t)$. Denoting with:

$$\left\langle r(t)^2 \right\rangle = \frac{1}{M}\sum_{m=1}^{M} r_m(t)^2 \quad (3)$$

the averaging over all the particles, the mean square displacement is defined as

$$MSD(t) = \left\langle |r(t) - r(t_0)|^2 \right\rangle. \quad (4)$$

The diffusion coefficient $D$ of water/polymer molecule is then calculated using Einstein's relation, Eq. 1. Manipulating Einstein's formula one can easily obtain that for sufficiently large times $\log(6D)+\log(t)=\log(MSD(t))$. The realm of normal diffusion (also known as Fickian diffusion) is reached when $\log(MSD(t))$ is a linear function of time with unit slope.

## Results

As preliminary validation of the equilibration procedure, we compare the density of the quasi-dry systems with the experimental values. The predicted value for PLA, averaged over the final 100-ps NPT simulations, is 1.18 g/cm$^3$, close to the experimental density (1.24 g/cm$^3$) [23]. In general, for the systems with intermediate water content the density decreases approximately linearly from the density of quasi-dry PLA to the density of pure water (see Fig. 2a). This behavior is expected and suggests that the models are reliable and properly equilibrated. The density, on the other hand is not affected by the polymer chain length. An exception to this behavior is the case of PLA monomers, for which a significant density drop to 1.05 g/cm$^3$ is observed, whereas a≈1.20 g/cm$^3$ density for monomeric PLA is expected. This suggests that the results for PLA monomers must be considered with care. However, we note that this condition (monomeric PLA) is the less relevant

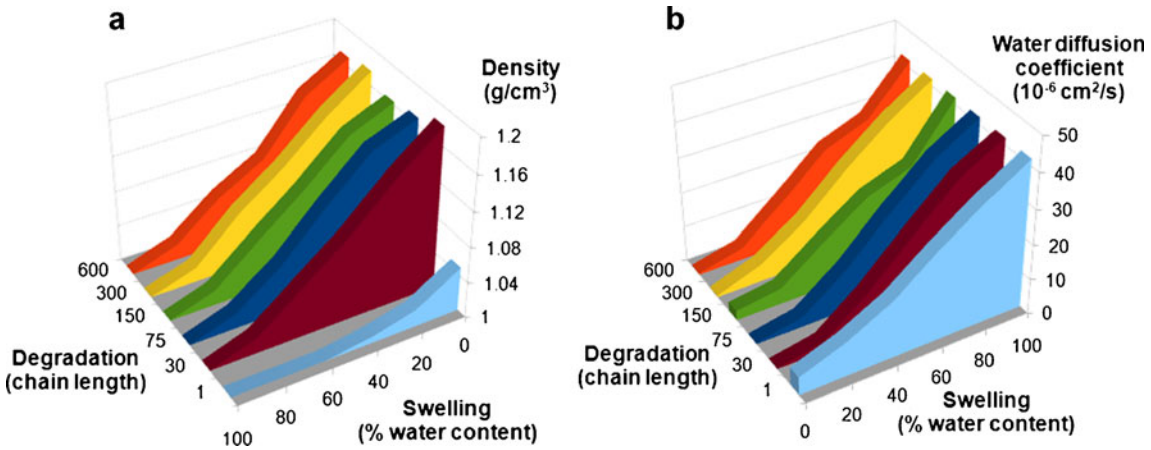



**Fig. 2** Density and water diffusion coefficients trends in degrading PLA matrices. The density of the systems (panel **a**) shows an almost linear dependence with respect to the water content, whereas it is not influenced by the average length of the polymer chains in the matrix, with the notable exception of systems where the polymeric matrix is reduced to single monomers. A similar behavior is observed for water diffusivity (panel **b**), which increases in swollen polymeric matrices but is not affected by the PLA chain length

from the application point of view, since a material reaching this state would no longer be structurally whole.

The equilibrated systems represented the starting configurations for subsequent diffusion analysis. The diffusion of water within the polymeric bulks in the dry and wet models is evaluated during 7 ns NPT MD simulations according to Eq. 1. Water diffusivity in PLA (see Table 2) spans two orders of magnitude (from $10^{-7}$ to $10^{-5}$ $cm^2/s$) and it is mainly affected by the water content. As a validation of the approach we calculated the self-diffusion coefficient of water, obtaining a value of $42.0 \cdot 10^{-6}$ $cm^2/s$, in the same order of magnitude of those obtained experimentally ($22.7 \cdot 10^{-6}$ $cm^2/s$) [24]. The self-diffusion coefficient for water is calculated in a similar way with respect to the method used to assess diffusivity of water in PLA matrices, the only difference is that the system, that is the simulated box, is made of only water molecules. We then calculate the diffusivity of polymeric chains, which show diffusion coefficients in the order of $10^{-8}$ $cm^2/s$ (see Table 3), thus several orders of magnitude below those found for water molecules. However, we note that the normal diffusion regime is not reached for polymer chains (with the exclusion of single monomers), and thus these values can be considered only as trend indicators.

## Discussion

In the present work we study how the diffusion of water changes during the degradation process of a biodegradable material, such as PLA. We assume that a hydrolytically degrading medium can be characterized by two main features:

- The swelling degree, referred to as $\phi$. The swelling is defined as the water mass fraction per unit volume. We assume that system starts with a dry, intact material (0% water) and, as the polymer erosion proceeds, the water-to-polymer mass ratio increases until only water remains (100% water).

**Table 2** Diffusion coefficient of water ($10^{-6}$ $cm^2/s$) as a function of degree of polymerization (columns) and % of water content (rows)

| | | $\phi$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | 2% | 20% | 40% | 60% | 80% | 100% |
| *i* | 600 | 0.41 | 3.39 | 13.5 | 24.2 | 28.9 | 42.0 |
| | 300 | 0.42 | 4.97 | 13.2 | 24.3 | 34.0 | 42.0 |
| | 150 | 0.27 | 4.56 | 13.8 | 22.2 | 16.6 | 42.0 |
| | 75 | 0.32 | 2.79 | 12.0 | 23.9 | 34.7 | 42.0 |
| | 30 | 0.24 | 2.59 | 12.3 | 24.2 | 35.3 | 42.0 |
| | 1 | 5.20 | 9.84 | 16.4 | 26.0 | 34.5 | 42.0 |

**Table 3** Diffusion coefficient ($10^{-8}$ $cm^2/s$) of PLA chains of length *i* into themselves (columns) with respect to % of water content (rows)

| | | $\phi$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | 2% | 20% | 40% | 60% | 80% | 100% |
| *i* | 600 | 0.33 | 1.37 | 3.03 | 3.67 | 29.59 | - |
| | 300 | 0.50 | 1.56 | 5.17 | 1.08 | 35.17 | - |
| | 150 | 0.45 | 0.83 | 8.36 | 6.82 | 36.21 | - |
| | 75 | 0.17 | 1.17 | 5.06 | 6.83 | 24.05 | - |
| | 30 | 0.53 | 1.04 | 3.83 | 6.94 | 24.33 | - |
| | 1 | 287 | 418 | 536 | 593 | 903 | - |

- The degree of polymerization (i.e., the average length of the polymeric chains), denoted by *i*. The degree of polymerization takes into account polymer degradation (i.e., chain scission). We assume that the system starts with long polymeric chains (here 600 monomer per chain) and, as degradation proceed, the chain length is reduced, eventually leading to just monomers.

Thus, in our work we study the water diffusion coefficient ($D_w$) as a function of the medium conditions ($\phi$, *i*). We also provide an estimate of the polymer chain diffusion ($D_i$, where *i* denotes the length of the chains).

The analysis of the water diffusivity in PLA shows that the diffusion coefficient increases almost linearly with the degree of swelling, whereas it is little or not affected by the degree of polymerization (see Fig. 2b). In evaluating the results of diffusivity from atomistic simulations, it is important to observe that the Einstein relation holds only if the regime of normal diffusion is reached. Indeed, Eq. 3 relies on the assumption of a random walk for each simulated water molecules through the polymer matrix, which happens only when the regime of normal diffusion is reached, that is when $\log(MSD(t))$ is an affine linear function with respect to $\log(t)$ with unit slope (see "Methods" section). We observe that the normal diffusion of water is reliably reached during the data production runs, since the aforementioned slope ranges from 0.87 to 1.00 (see Fig. 3a). Consequently, the diffusivity coefficients obtained for water molecules in the polymer matrix can be assumed to be reliable for all the systems. In the past, a similar problem (water diffusivity in PLA matrix) has been studied with atomistic simulations [17] and our results qualitatively agree with the diffusivity value reported by the authors. Nonetheless, our work represents a significant advance over previous works. In their pioneering work, Castano et al. [17] explored two cases, *i.e.*, water content of 2% and 7%., whereas in the present work we conducted a systematic investigation of the effect of a wide range of hydration degrees and polymer chain lengths on water

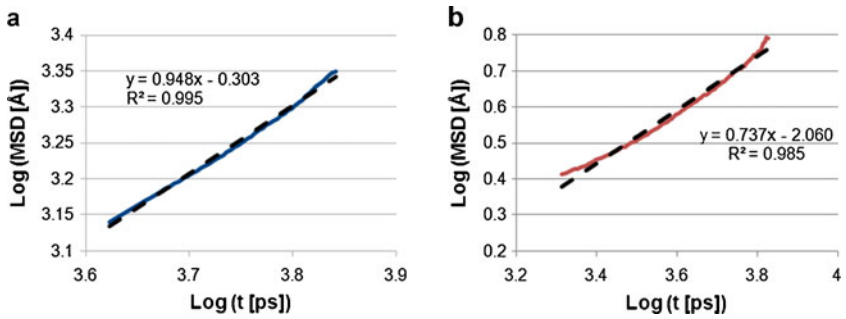



**Fig. 3** Normal diffusion regime. Log(*MSD*(*t*)) vs. log(*t*) plots (straight lines) and linear interpolation (dashed lines) for the diffusion of water (panel **a**) and for the diffusion of PLA chains (panel **b**) obtained from MD simulations. The plots show that in the case of water the normal diffusion regime is reached, as indicated by the slope very close to 1, whereas in the case of PLA chains (panel **b**) a longer simulation time would be required since the normal diffusion regime is not reached within the 7 ns ($\approx 10^{3.8}$ ps) considered in the production MD simulations

diffusivity. Furthermore, we were able to reach the normal diffusion regime by running longer (7 ns vs. 1 ns) MD simulations and thus providing more reliable estimates of diffusivity.

In Fig. 4 we track the position assumed by one water molecule during the 7 ns simulation, for systems with different swelling degree. The water molecule trajectories show that, depending on water content, there is a high difference in diffusive paths. As reflected in the diffusivity coefficients, in quasi-dry systems water molecules explore a small portion of the unit cell, whereas in high solvates systems, the diffusing molecules explore a much higher volume.

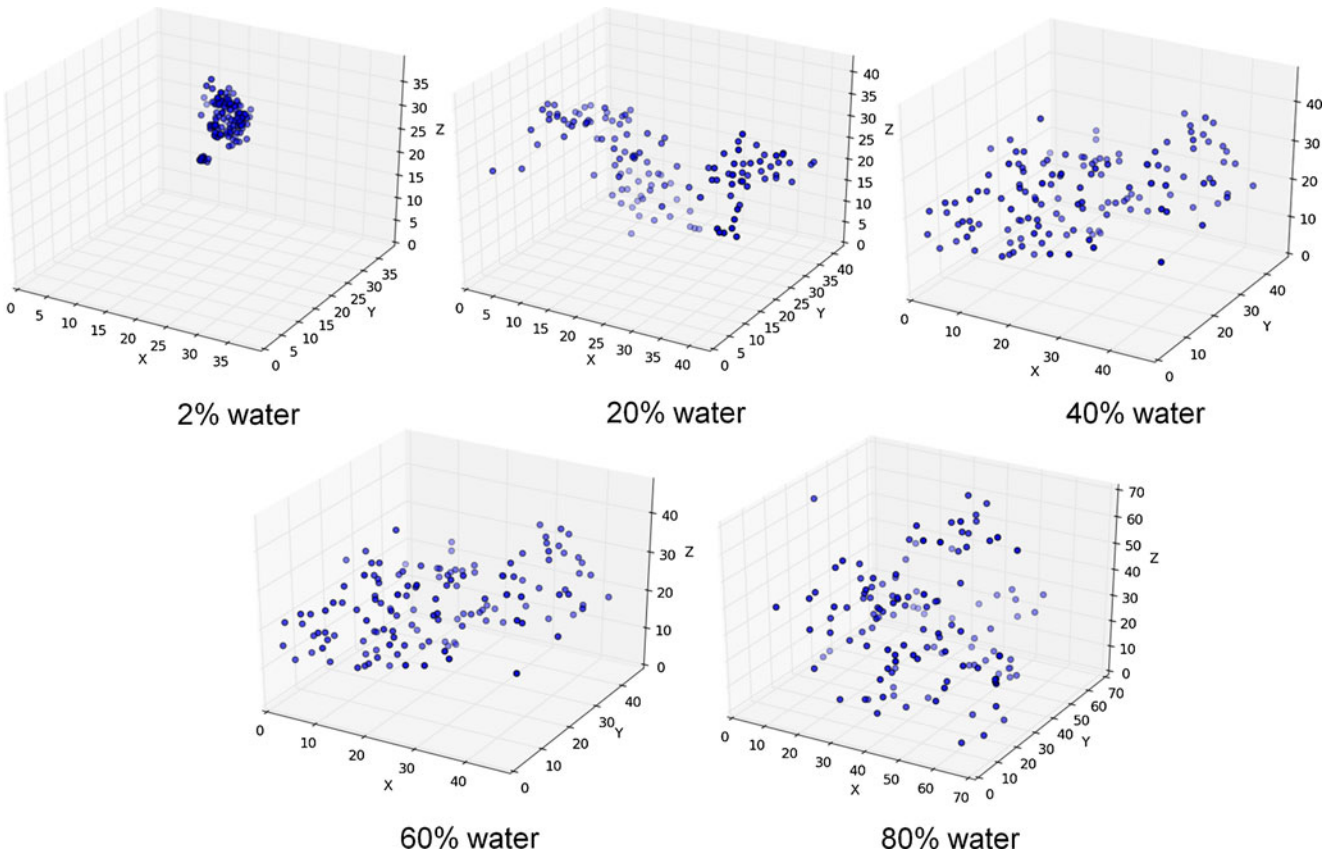



**Fig. 4** Water molecules trajectories. The trajectory of a single water molecule during the 7 ns MD simulation is represented for systems with different swelling degree (from 2% to 80% water content) but constant PLA chain length (150 monomers per chain). Cubic boxes represent the periodic unit cells. The plots show that in the quasi-dry systems (2% water) water molecules are confined in a small region of the polymeric matrix whereas in solvated systems water molecules are able to explore a larger volume within the unit cell

We note that in order to get a good estimate of the diffusivity, a large number of averaging diffusing particles (water molecules in our case) is desired. In the case of 2% water simulations the limited number of water molecules (10) in the systems can thus limit the precision of the diffusivity estimates. However, in order to include a higher number of diffusive molecules and keep the same water/polymer weight ratio we would need to increase also the number of polymeric chains. This, in turn, would significantly increase the computational cost involved since it scales with $\sim N^2$, where $N$ is the number of atoms of the system. Thus, water diffusivities in the first column of Table 2 must be considered with care, as they can be influenced by the limited averaging. However, we note that water diffusivity shows a linear dependence with respect to water content and that it is insensitive to polymeric chain length. These two aspects are reflected also in the case of 2% water content, which suggests that the obtained values fall in a reasonable range.

In addition to water diffusivity, we estimated the diffusion coefficients of the polymeric chains. In opposition to the case of water molecules, the normal diffusion regime is not reached for polymer chains (with the exclusion of single monomers), as shown by the slopes of the log(*MSD*(*t*)) vs. log(*t*) functions which are below the unit value (see Fig. 3b). Thus, the calculated diffusivities cannot be considered as Fickian diffusivity coefficient values. However, they can be useful as coarse approximations and as trend indicators. The polymer diffusivity results suggest that, irrespectively of the chain length, the polymer diffuses much less than water and this is likely due to the polymer's larger molecular weight. The only deviation from this behavior is observed for the systems with single PLA monomers (highly degraded matrices, bottom row in Table 3) for which a much higher PLA diffusion constant is observed, approaching the diffusivity obtained for water molecules.

A limitation of the work is related to the assessment of polymer chains diffusivity. Although a coarse estimate can be provided by atomistic simulations, we were not able to reach normal diffusion regime and thus provide reliable values. In order to reach Fickian diffusion also for polymer it would be necessary to simulate larger systems and much longer times, which is however very challenging due to the computational burden. A possible way to obtain a more precise assessment of polymer diffusivity is the use of coarse-grain simulation as those described in earlier works [25, 26]. This approach could also provide the diffusivity coefficient of molecules larger than water (e.g., drugs).

## Conclusions

The hydrolytic degradation in polymer is a transport-reaction mechanism, encompassing the transport of water

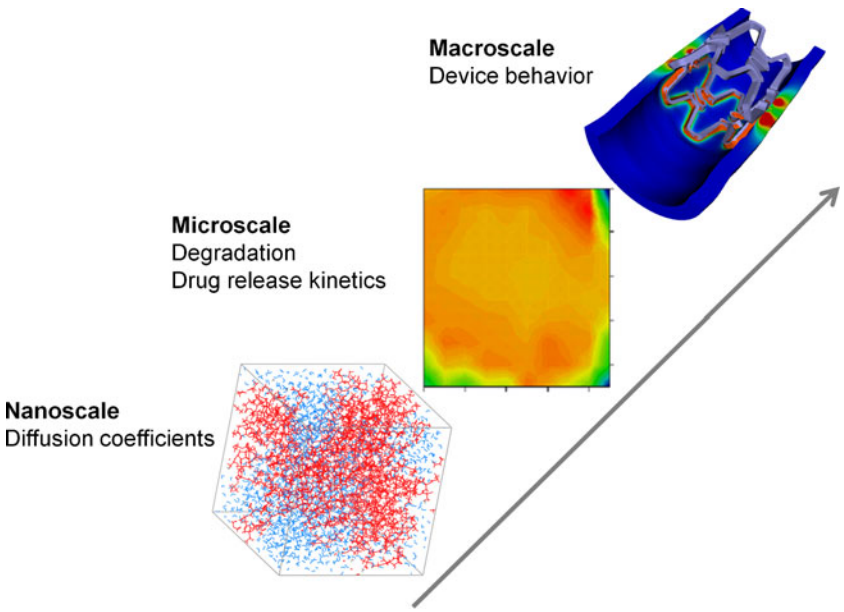



**Fig. 5** Multiscale modeling framework. A multi-scale approach is envisioned for the design of biodegradable devices in which it is essential for precise control over the features of the polymeric matrix (e.g., drug release rates) during degradation process. Within this approach, atomistic simulations can provide the diffusivity coefficients of small molecules (e.g, water, gases and drugs) for different states of the degrading matrix. Microscale analytical models [11, 27], using these data as an input, can then predict the degradation kinetics and thus how mechanical properties or drug release rates change with time. Finally, macroscale modeling would incorporate the information from the micro-scale to investigate the behavior of the final device

in the polymer, the hydrolytic cleavage of the polymeric bonds, the transport of degradation products and their subsequent absorption in the surrounding media. Over the last decades experimental work has delivered important information to elucidate the basic phenomena involved in this process. However, an even more detailed understanding of the mechanism of degradation can be achieved by combining this information with insight into the basic processes at an atomistic level. This insight is difficult to access via experimental techniques, but can be improved using simulation techniques, such as atomistic molecular modeling. In this work, we show that atomistic simulations allow the investigation of the diffusion of individual water molecules into a biodegrading material. This information can then be coupled with diffusion-reaction models that describe the degradation process [11]. This approach could allow predicting the evolution with time of several properties of biodegradable matrices (e.g., the transport properties), according to the change of relevant indicators such as the extent of degradation and swelling of the matrix. Furthermore, by using atomistic simulations to assess the diffusivity of drugs, the approach presented here could be useful within a multi-scale framework for the design of biodegradable devices with predictable properties (e.g., drug delivery rate or mechanical properties) during the degradation process (Fig. 5).

**Acknowledgments** The authors thank the anonymous reviewers of the manuscript for their very helpful comments which greatly helped to improve the manuscript. This research was supported by the Italian Institute of Technology (IIT) (Grant "Models and methods for degradable materials"). High-performance computing resources have been provided by Regione Lombardia and CILEA (Consorzio Interuniversitario Lombardo per L'Elaborazione Automatica) through a LISA initiative 2010 grant and by Cineca under the ISCRA initiative. The authors declare no conflict of interest of any sort.

## References

1. Lendlein A, Langer R (2002) Biodegradable, elastic shape-memory polymers for potential biomedical applications. Science 296:1673–1576
2. Langer R (1998) Drug delivery and targeting. Nature 392:5–10
3. Laufman H, Rubel T (1977) Synthetic absorable sutures. Surg Gynecol Obstet 145:597–608
4. Pietrzak WS, Sarver DR, Verstynen ML (1997) Bioabsorbable polymer science for the practicing surgeon. J Craniofac Surg 8:87–91
5. Agrawal CM, Ray RB (2001) Biodegradable polymeric scaffolds for musculoskeletal tissue engineering. J Biomed Mater Res 55:141–150
6. Soares JS (2009) Bioabsorbable polymeric drug-eluting endovascular stents A clinical review. Minerva Biotecnol 21:217–230
7. Hayashi T (1994) Biodegradable polymers for biomedical uses. Prog Polym Sci 19:663–702
8. Hutmacher DW (2000) Scaffolds in tissue engineering bone and cartilage. Biomaterials 21:2529–2543
9. Tamada JA, Langer R (1993) Erosion kinetics of hydrolytically degradable polymers. Proc Natl Acad Sci USA 90:552–556
10. Hofmann D, Entrialgo-Castano M, Kratz K et al. (2009) Knowledge-based approach towards hydrolytic degradation of polymer-based biomaterials. Adv Mater 21:3237–3245
11. Soares JS, Zunino P (2010) A mixture model for water uptake, degradation, erosion and drug release from polydisperse polymeric networks. Biomaterials 31:3032–3042
12. Tocci E, Hofmann D, Paul D et al. (2001) A molecular simulation study on gas diffusion in a dense poly(ether-ether-ketone) membrane. Polymer 42:521–533
13. Gautieri A, Ionita M, Silvestri D et al. (2010) Computer-aided molecular modeling and experimental validation of water permeability properties biosynthetic materials. J Comput Theor Nanosci 7:1–7
14. Hofmann D, Fritz L, Ulbrich J et al. (2000) Detailed-atomistic molecular modeling of small molecule diffusion and solution processes in polymeric membrane materials. Macromol Theor Simul 9:293–327
15. Marque G, Verdu J, Prunier V et al. (2010) A molecular dynamics simulation study of three polysulfones in Dry and hydrated states. J Polym Sci Part B Polym Phys 48:2312–2336
16. Horstermann H, Hentsehke R, Amkreutz M et al. (2010) Predicting water sorption and volume swelling in dense polymer systems via computer simulation. J Phys Chem B 114:17013–17024
17. Entrialgo-Castaño M, Salvucci AE, Lendlein A et al. (2008) An atomistic modeling and quantum mechanical approach to the hydrolytic degradation of aliphatic polyesters. Macromol Symp 269:47–64
18. Polymer User Guide, Amorphous Cell Section, Version 4.0.0 p+. Molecular Simulations Inc, San Diego, CA, 1999
19. Sun H (1998) COMPASS: an ab initio force-field optimized for condensed-phase applications - overview with details on alkane and benzene compounds. J Phys Chem B 102:7338–7364
20. Entrialgo-Castano M, Lendlein A, Hofmann D (2006) Molecular modeling investigations of dry and two water-swollen states of biodegradable polymers. Adv Engineering Mater 8:434–439
21. Ionita M, Silvestri D, Gautieri A et al. (2006) Diffusion of small molecules in bioartificial membranes for clinical use: molecular modelling and laboratory investigation. Desalination 200:157–159
22. Einstein A (1905) Von der molekulärkinetischen Theorie der Wärme gefordete Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen (The motion of elements suspended in static liquids as claimed in the molecular kinetic theory of heat). Ann Phys, Berlin 17:549–560
23. Van de Velde K, Kiekens P (2002) Biopolymers: overview of several properties and consequences on their applications. Polym Test 21:433–442
24. Tanaka K (1978) Self-diffusion coefficients of water in pure water and in aqueous solutions of several electrolytes with 18O and 2H as tracers. Journal of the Chemical Society, Faraday Transactions 1: Phys Chem Condens Phases 74:1879–1881
25. Gautieri A, Vesentini S, Redaelli A (2010) How to predict diffusion of medium-sized molecules in polymer matrices. From atomistic to coarse grain simulations. J Mol Model 16:1845–1851
26. Gautieri A, Russo A, Vesentini S et al. (2010) Coarse-grained model of collagen molecules using an extended MARTINI force field. J Chem Theor Comput 6:1210–1218
27. Soares JS, Moore JE, Rajagopal KR (2008) Constitutive framework for biodegradable polymers with applications to biodegradable stents. ASAIO J 54:295–301

J Mol Model (2012) 18:1503–1512
DOI 10.1007/s00894-011-1174-5

ORIGINAL PAPER

# Understanding the desensitizing mechanism of olefin in explosives: shear slide of mixed HMX-olefin systems

**Chaoyang Zhang · Xia Cao · Bin Xiang**

Received: 13 May 2011 / Accepted: 30 June 2011 / Published online: 23 July 2011


**Abstract** We simulated the shear slide behavior of typical mixed HMX-olefin systems and the effect of thickness of olefin layers (4–22 Å) on the behavior at a molecular level by considering two cases: bulk shear and interfacial shear. The results show that: (1) the addition of olefin into HMX can reduce greatly the shear sliding barriers relative to the pure HMX in the two cases, suggesting that the desensitizing mechanism of olefin is controlled dominantly by its good lubricating property; (2) the change of interaction energy in both systoles of shear slide is strongly dominated by van der Waals interaction; and (3) the thickness of olefin layers in the mixed explosives can influence its desensitizing efficiency. That is, the excessive thinness of olefin layers in the mixed explosive systems, for example, several angstroms, can lead to very high sliding barriers.

**Keywords** Desensitizing mechanism · HMX · Molecular forcefield method · Olefin · Shear slide

## Introduction

The explosive safety versus external stimuli such as impact, friction, shock, electrostatic spark, high temperature, flame, and so on, as well as the explosive power, is one of the two most important properties of explosives. Because the power and the safety of explosives determine directly the offensive abilities and the security of armaments, respectively, people hopefully seek some more powerful explosive compounds with acceptable, even lower sensitivities. Seldom can they succeed, due to a formidable intrinsic energy-sensitivity conflict of CHNO explosives. Namely, the higher energy usually goes with the higher sensitivity [1]. As a matter of fact, some more powerful compounds than 1,3,5,7-tetranitro-1,3,5,7-tetrazocine (HMX) have been synthesized for a long time, such as hexanitrobenzene (HNB), octuplenitrocubane (ONC) and 3,4-dinitrofuroxan (DNOF) with detonation velocities above 10000 m/s, much higher than that of HMX, about 8900 m/s when its load density is 1.851 g/cm$^3$. But they are still not applicable in practice due to their dissatisfactory security despite their high energy densities [2]. Actually, in most cases the security may be more concerned than the power, especially in this people-foremost-society.

It is a common and important means to add desensitizers into explosives to form applicable mixed explosive formulations with acceptable sensitivities [3]. Olefin, an amorphous mixture of *n*-alkane with 18–30 carbons, is a famous and extensively used desensitizer which can distinctly decrease the mechanical sensitivity in contrast to the neat explosives contained in formulations [3]. It is widely accepted that its desensitizing characteristics are attributed to its low melting point and good lubricating property, which are helpful to disperse the internal stress or the stress potential caused by external mechanical stimuli to decrease or avoid the formation of hot spots [3].

Very recently, a molecular simulation showed that the desensitizing mechanism of olefin is dominantly attributed to its good lubricating property and little to its compressibility. That is, for example, by the addition of 3 wt% olefin

C. Zhang (✉) · X. Cao
Laboratory of Material Chemistry, Institute of Chemical Materials,
China Academy of Engineering Physics (CAEP),
P.O.Box 919–327, Mianyang, Sichuan,
People's Republic of China 621900
e-mail: zcy19710915@yahoo.com.cn

X. Cao · B. Xiang
College of Chemistry and Chemical Engineering,
Chongqing University,
Chongqing, People's Republic of China 400044

into HMX, the shear sliding barrier is remarkably lowered several tens of times relative to neat HMX, from 414 to 13 MJ/m$^3$ under the condition of bulk shear slide without any volume change of the related system [4]. It suggests that the shear slide behavior of mixed explosive-olefin systems is an important factor influencing the desensitizing mechanism of olefin and should be paid more attention to. It therefore becomes the topic of this paper.

These shear slide without any volume change induced by external mechanical force can include two typical styles: bulk shear slide and interfacial shear slide. These two styles of shear slide can both appear in practice and accordingly have been considered in our simulations. Also, a deeper insight into the mechanism requires in general an intrinsical description of the interactions of explosive-olefin systems during shear sliding. For these reasons, a molecular force-field in which the total interaction energy is partitioned into the valence bonded and the nonbonded interaction items (van der Waals and electrostatic) [5], and some typical HMX+olefin systems, were employed to make the shear slide behavior and its physical nature clear. Moreover, the effect of the olefin dispersion in explosives on the behavior was discussed too.

CHNO explosives are still the main energy source for offensive abilities of weapons even though some conceptual energetic materials with much higher energy densities have been proposed for a long time, such as all nitrogen molecules, metal hydrogen, anti-matter, and so forth [2]. That is, there is nevertheless a very long way for them to actual applications. It is efficient to find good desensitizers to partly overcome the said intrinsic conflict between energy and sensitivity of CHNO explosives. To expose the structure-property relationship of explosive compounds and desensitizers applied in practical formulations is helpful to find new more appropriate ones [6–22]. We have discussed the mechanism of some applicable desensitizers versus external mechanical stimuli including crystalline graphite and TATB [23, 24], and the reason for low mechanical sensitivities of some explosives from a viewpoint of molecular stacking [25]. This study of olefin, an amorphous desensitizer, provides another further understanding on desensitizing mechanism and is hopeful to give some theoretical support for molecular design of desensitizers.

## Methodologies

### Theory of desensitizing mechanism

According to the theory of hot spot, an explosion occurs after the temperature and the heat of one or some spots in explosives increase to a certain extent. The factors causing the enhanced temperature and heat may be various, but at least, external energy has been definitely accumulated on the spot before explosion. That is, the external energy can lead to the strong molecular vibrations, then the chemical decomposition, and the final explosion. Therefore, the desensitizing mechanism in explosives should be to disperse the external energy to avoid the formation of hot spots and the effect of desensitizers can be regarded as a kind of dispersing ability against external stimuli.

However, it is now difficult to directly simulate the dispersing ability of desensitizers at a molecular level. An alternative method is to calculate the energy change of unit volume or unit area to denote the dispersing ability: the fewer the energy change of unit volume or unit area, the better the dispersing ability, that is, the better ability to disperse the external energy to avoid the formation of hot spots, and the higher ability to decrease sensitivity. This can be simply understood as that when the same amount of energy is respectively exerted on two objects with different volumes, the smaller one will possess the higher energy density and the lower energy dispersing ability. By this way, the unit energy change can be related with the desensitizing mechanism. Also, it has been confirmed that the increased internal stress (actually it can be understood as the energy density) can induce the lowered decomposition barrier of explosives [26, 27]. Therefore, the shear behavior in this paper can be described by the unit energy change during shear slide.

Two typical cases of shear slide of the mixed HMX-olefin systems were considered: bulk shear slide and interfacial shear slide (Fig. 1). Both cases assume that there is no vacuum between olefin and HMX. That is to say, olefin and HMX interact with each other compactly. As mentioned above, the shear slide behavior can be described by the energy change of unit volume or unit area during shear slide. Because bulk shear slide and interfacial shear

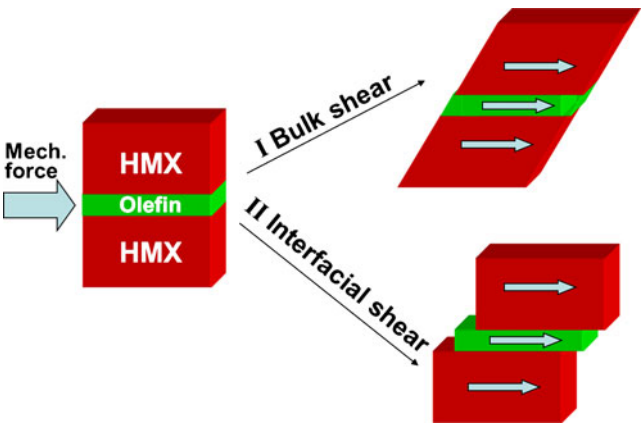



**Fig. 1** Bulk shear slide and interfacial shear slide of the mixed olefin-HMX systems. The decreased sizes of arrows denote the dispersed mechanical energy due to the shear slide

slide are related to the volume of total bulk and the area of interaction interface, respectively, the shear behavior can be respectively described by the potential of bulk shear slide ($P_{BS}$) and the potential of interfacial shear slide ($P_{IS}$) as Eqs. 1 and 2

$$P_{BS,i} = \frac{\Delta E_i}{V} = \frac{E_i - E_{\min}}{V} \quad (1)$$

$$P_{IS,i} = \frac{\Delta E_i}{A} = \frac{E_i - E_{\min}}{A}. \quad (2)$$

In the above two equations, $E$, $V$, and $A$ represent the internal energy, the interfacial area, and the volume of interested systems, respectively; the subscript $i$ denotes the value of $i$th step of shear slide; $E_{min}$, the minimal energy during the total shear sliding, is adopted because $E_0$, the energy before slide, is not always $E_{min}$, owing to the amorphous characteristics of olefin.

## Modeling

### *Modeling for bulk shear slide*

Combining the experimental data and molecular forcefield calculations, we established models for shear slide calculations. As illustrated in Fig. 2, the original structures of both HMX and olefin for constructing some mixed HMX-olefin systems were derived from their experimental data [28]. That is, the initial molecular stacking structure of HMX was from its X-ray diffraction results, and the initial composition and density of amorphous olefin was from its experimental physical-chemical parameters. Then, these structures were dynamically balanced and relaxed using a molecular forcefield named COMPASS.

Considering that the most probable shear slide of mixed HMX-olefin systems takes place along the dominant faces of HMX crystal and the (010) face is a dominant one in its practical crystal habit [29], we built the models for shear slide calculations by adding amorphous olefin (for example, Fig. 2f) onto the (010) face of crystal HMX (Fig. 2d) to form the mixed HMX+olefin systems (for example, Fig. 2g) along the orientation of $b$ axis.

At the same time, the effect of the layer thickness of olefin in the mixed explosives was also accounted in modeling. That is, three mixed HMX+olefin systems with the same components (91 wt% HMX and 9 wt% olefin, or 320 HMX molecules and 30 $C_{22}H_{46}$ molecules) but different layer thickness of olefin were built as Fig. 3. From Fig. 3a to b and c, the thickness decreases. Obviously, the effect of the thickness can be studied through these models. By decreasing the lattice parameter $\gamma$ of the built cells in Fig. 3 from 90 to 50 degree with a step of 2 degree, and elongating $b$ (*OB*) axis correspondingly and synchronously to keep the lattice volume fixed, we can simulate the bulk shear slide along *OA* axis in the plane of the (010) face of HMX crystal.

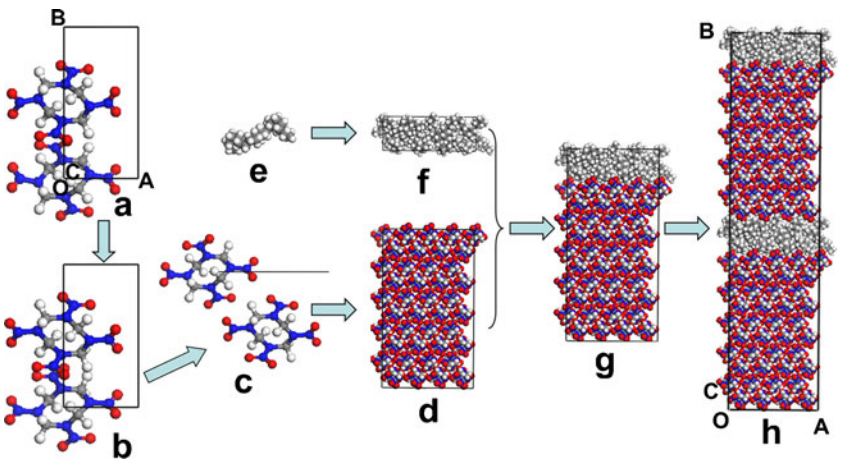



**Fig. 2** Modeling for the bulk shear slide calculations. **a** is the experimental unit cell of HMX cited from Ref [28], whose lattice parameters are $a$=6.540 Å, $b$=11.050 Å, $c$=8.700 Å, $\alpha$=90$^0$, $\beta$=124.30$^0$ and $\gamma$=90$^0$, and density is 1.894 g/cm$^3$; **b** is a relaxed **a** using COMPASS forcefield, its parameters change to $a$=6.482 Å, $b$=10.365 Å, $c$=8.954 Å, $\alpha$=90$^0$, $\beta$=123.443$^0$ and $\gamma$=90$^0$, and its density increases to 1.959 g/cm$^3$; **c** is a cleaved face along (010) face of **b**, that is, it shows only one (010) face layer of HMX; **d** is a relaxed 4×5×4 supercell by the forcefield whose unit cell is built by a vacuum slab crystal without vacuum thickness from **c**, and whose lattice parameters are $a$=30.53 Å, $b$=51.805 Å, $c$=25.92 Å, $\alpha$=$\gamma$=90$^0$, and $\beta$=78.3$^0$; **e** is an optimized olefin alkane chain with 22 carbon atoms; **f** is a relaxed structure from a 100 ps *NVT* dynamically equilibrated amorphous cell composed of 15 **e** and possessing the same lattice parameters with **d** except $b$; because **d** and **f** have five same lattice parameters excluding $b$, a cell combining them along $b$ axis can be obtained. After first a 100 ns *NVT* MD and then a relaxation of the MD equilibrium structure of the cell, **g**, is obtained; and **h** is a relaxed 1×2×1 supercell of **g** and is Fig. 3b. The cells with different dispersion of olefin in mixed systems in Fig. 3 are constructed by the similar methods

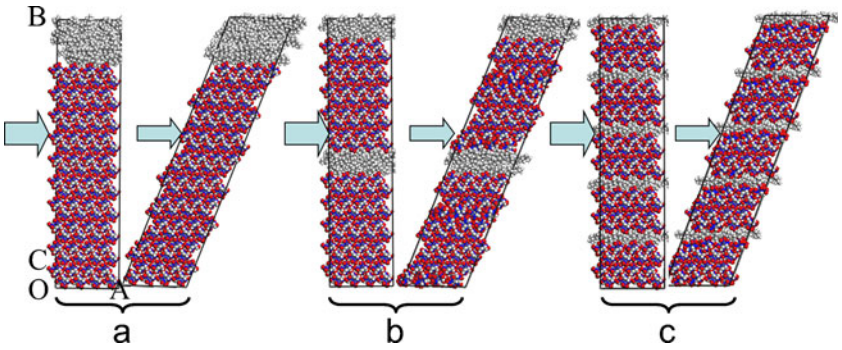



**Fig. 3** Bulk shear slide of the mixed 91 wt% HMX+9 wt% olefin (320 HMX+30 $C_{22}H_{46}$) systems: the right and the left in each group denote the structure before and after bulk shear slide, respectively. **a** with one part of olefin and one part of HMX is built by combining a 4×10×4 supercell with the same unit cell similar to Fig. 2d and 30 olefin chains similar to Fig. 2f. **b** with two equal parts of mixed olefin+HMX is Fig. 2h, and **c** with five equal parts of mixed olefin+HMX is a 1×5×1 supercell whose unit cell is built by combining a 4×2×4 supercell with the same unit cell similar to Figs. 2d and 6 olefin chains similar to Fig. 2f. The decreased arrows after slide show the dispersed mechanical energy. The thickness of the olefin layers in a, b and c is 22.08, 11.04 and 4.42 Å, respectively

*Modeling for interfacial shear slide*

Different from in above bulk shear slide, the interfacial interactions instead of total bulk interactions are concerned in interfacial shear slide. That is to say, by changing interfaces instead of bulks we can establish some models for interfacial shear slide calculations. Also, interfacial shear slide along (010) plane were taken into account. It assumes that this slide takes place between the HMX molecules on (010) face and the molecules in olefin neighboring them. The molecules far from the interface can be accordingly kept fixed, as the constrained top and bottom layers in Fig. 4.

By translating the top block of HMX we can simulate the interfacial shear slide between HMX and olefin. Also, the effect of the layer thickness of olefin on the slide was investigated through changing the thickness, or changing the molecular number of olefin, between the two HMX blocks (Fig. 5).

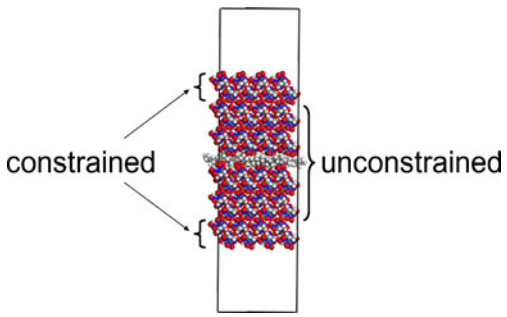



**Fig. 4** Modeling for interfacial shear slide calculations. Combining two 4×3×4 supercells like Fig. 2d and an amorphous cell of olefin like Fig. 2f, we obtained a sandwich. After a 100 ns *NVT* MD on the sandwich, we relaxed it. Then make a vacuum with a thickness of 100 Å over the relaxed sandwich and evaluate it 50 Å to get the model. In the following shear slide calculations, the first two molecular layers neighboring the surface were constrained to keep the related systems in a given space, and the remaining atoms were unconstrained

*Calculating methods*

A COMPASS forcefield based on first principle calculations and experiments whose validity to HMX and the HMX+olefin system has been already established [4, 30], was adopted for a series of geometry relaxations and molecular dynamics (MD) of above models. As indicated in Eqs. 3 and 4, total energy is divided into bonded energy (valence energy) and nonbonded energy including van der Waals (vdW) and electrostatic terms. The physical nature of above shear slide can be accordingly described by these terms with definite physical meanings.

$$E_{total} = E_{bonded}(E_{valence}) + E_{nonbonded} \tag{3}$$

$$E_{nonbonded} = E_{vdW} + E_{Electrostatic} \tag{4}$$

where

$$E_{vdW} = \sum_{\substack{pairs(i,j)\\ i<j}} \varepsilon_{ij}\left[2\left(\frac{r_{ij}^0}{r_{ij}}\right)^9 - 3\left(\frac{r_{ij}^0}{r_{ij}}\right)^6\right] \tag{5}$$

$$r_{ij}^0 = \left(\frac{(r_i^0)^6 - (r_j^0)^6}{2}\right)^{1/6} \tag{6}$$

$$\varepsilon_{ij} = 2\sqrt{\varepsilon_i \varepsilon_j}\left(\frac{(r_i^0)^3 (r_j^0)^3}{(r_i^0)^6 (r_j^0)^6}\right) \tag{7}$$

$$E_{electrostactic} = \sum_{\substack{pairs(i,j)\\ i<j}} \frac{q_i q_j}{\varepsilon r_{ij}} \tag{8}$$

A series of step-by-step relaxations were carried out to simulate the mixed HMX+olefin systems against external

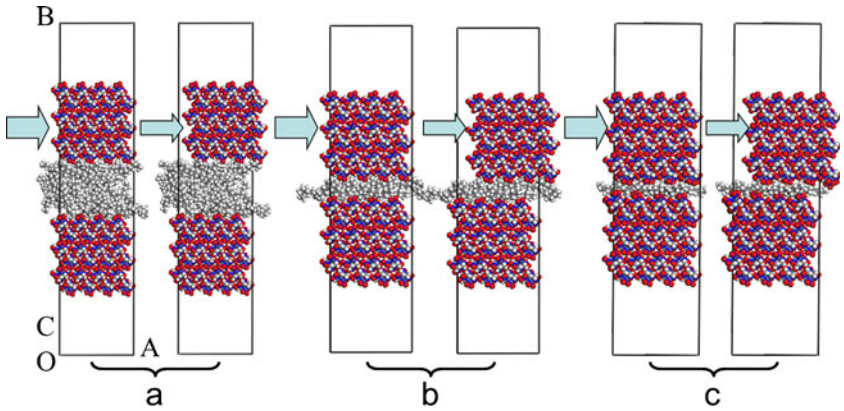



**Fig. 5** Interfacical shear slide of the HMX+olefin mixed systems. There are 30, 10, and 5 alkane chains containing 22 carbon atoms representing olefin in **a**, **b**, and **c**, respectively. Only the top HMX block slides uniformly in all cases. The decreased arrows after slide show the dispersed mechanical energy

mechanical stimuli, namely, the bulk shear slide and the interfacial shear slide, as well as compression for a complement to the previous study [4]. This step-by-step static process can be to a certain extent close to a dynamic process, or a quasi-dynamic process, which may be more near practice. That is to say, in these successive relaxation calculations, the relaxed structure of a loaded (slid or compressed) model in one step (*ith* step) was used as the initial structure for the next step (*(i+1)th* step) in which the shear slide or the compression was loaded again. In all relaxations, the lattice parameters kept fixed. Moreover, the *constraining* method was introduced in the simulation of interfacial shear slide (Fig. 4). Namely, the molecules of the top and bottom layers were kept fixed during geometry relaxation and molecular dynamics, and the rest were not fixed and kept moveable. In addition, a canonical ensemble dynamic method (*NVT*) was used in modeling for obtaining dynamically balanced structures and building initial structures for simulations. In all energy summation calculations, Ewald and atom based methods were employed for electrostatic and van de Waals interactions, respectively. Andersen thermostat method was used for *NVT* dynamic simulations. All simulations were carried out with use of Accelrys' Material Studio code [31].

In studying the interfacial shear slide, a series of static point calculations were carried out. As illustrated in Fig. 6, two cases were accounted for: 21 spot calculations for all three cases shown in Fig. 5 and 121 spot calculations only for the case shown in Fig. 5b.

The sliding potentials (SP) including the highest SP, the sliding barriers (SB), and the physical nature can be obtained by these calculations.

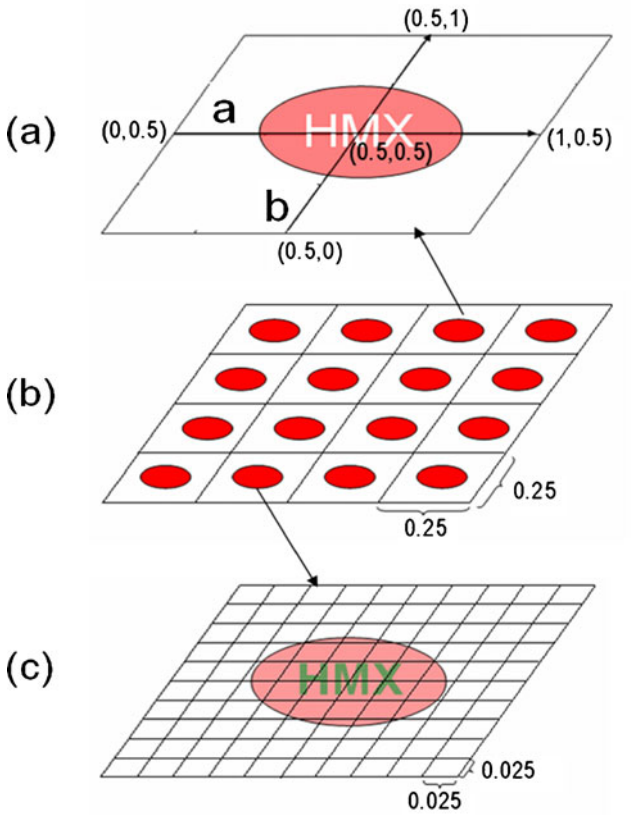



**Fig. 6** Plot showing static point calculations for interfacial shear slide. **(b)** shows a (010) surface of HMX indicated in the cells in Fig. 5 with 16 HMX molecules. **(a)** and **(c)** show one HMX molecule and the 1/16 total surface area occupied, and the relaxation calculations on 21 points on two lines *a* parallel to *OA* and *b* parallel to *OC* axes in Fig. 4, and 121 points on the total plane, respectively. Owing to the few changes of HMX molecules during sliding, it can be considered that 16 molecules are arrayed periodically on the surface. That is, only 1/16 of the total surface area, the area of one HMX molecule occupied on the surface (a and c), is accounted for calculations. As illustrated in **a**, sliding with a step of 0.025 (0.1×0.25) fractional coordinates along two lines' orientations (*a* and *b*) is considered. Namely, there are 21 (11×2-1) spots for calculation for one case. And the initial spot for calculations is (0.5, 0.5) for this case. And as illustrated in **c**, sliding on total plane with the same step is taken into account. That is, there are 121 (11×11) lattice spots for calculations for one case. And the initial spot for calculations is the middle spot in the lattice for this case

## Results and discussion

### Bulk shear slide

Figure 7 indicates the potentials of bulk shear slide of five systems: neat HMX, neat olefin, and three mixed systems

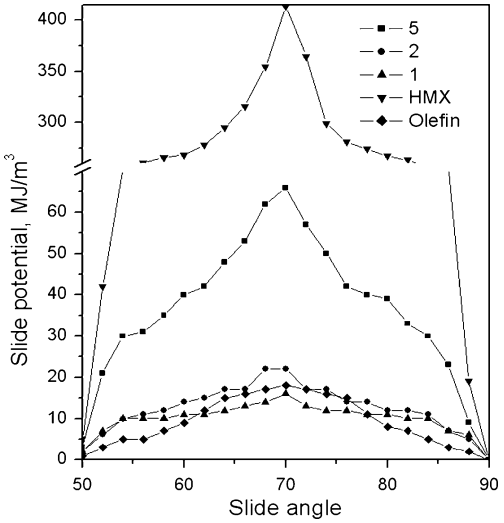



**Fig. 7** Potentials of bulk shear slide of five systems. 1, 2, 5, HMX, and Olefin correspond to Fig. 3a, b and c, neat HMX, and neat olefin, respectively

of 91 wt% HMX and 9 wt% olefin, which are denoted by 1, 2, and 5, and contain one, two, and five olefin monolayer in the cell, respectively. Usually, SB is the most concerning during shear sliding due to its direct relation to energy dispersion. Comparing the SB of five systems we can find that: (1) SB of neat HMX, about 420 MJ/m$^3$, is much higher than that of the other four systems containing olefin, within 13–68 MJ/m$^3$. As reported previously, the addition of olefin into HMX can sharply reduce SB from over 400 to several tens of MJ/m$^3$. It is therefore confirmed again that the improvement of shear slide of olefin in explosives is predominant in its desensitizing mechanism versus external mechanical stimuli; (2) SB of three mixed HMX-olefin systems with the same components are different from one another. In particular, SB of 5, 68 MJ/m$^3$, is obviously higher than that of 1 and 2, 13 and 22 MJ/m$^3$, respectively. It shows that the layer thickness of olefin in mixed explosives can influence the bulk shear slide behavior: the smaller the thickness, the higher SB. The thickness of 5 corresponding to the structure shown in Fig. 3c is only close to 5 Å, only one-fifth of the thickness of 1 (25 Å) corresponding to Fig. 3a and two-fifths of the thickness of 2 (12.5 Å) corresponding to Fig. 3b. Rationally, the excessive thinness of olefin monolayer, for example, several angstroms of 5, is disadvantageous to relax the slid HMX molecules or molecules of olefin to a point with a much lower local potential. It suggests that the desensitizing effect of olefin can be evident only if there is enough thickness for olefin monolayer. In practical formulae, we may be only concerned about the uniformity of all components contained in mixed explosives, but not the thickness. According to the experimental examinations, the thickness of olefin monolayer is usually in excess of tens of nm, and is then usually not a concerned object. That is to say, the olefin in explosives with good uniformity and not excessively thin monolayer has a low SB and then a good desensitizing ability.

We examined the effect of olefin dispersion on the compressibility of mixed explosives (Fig. 8). Comparing the internal stress ($P = \Delta E/\Delta V$) of the three systems with the same components but the different olefin dispersion, we cannot find an obvious difference among them. But their stress is higher than that of neat olefin and a little lower than that of neat HMX. This result indicates that the compressibility has no relation on the dispersion, but on the components. Different from the crystalline desensitizers like graphite and TATB with anisotropic compressibility (the easiest compression along their *c* axes), olefin with amorphous structures does not show any anisotropy of compressibility. It is one of the differences of desensitizing mechanism between crystalline and amorphous desensitizers.

To investigate the physical nature of the bulk shear slide of HMX-olefin systems, we took the system in Fig. 3b as a sample and the Eqs. 3 and 4 to do a series of calculations. As illustrated in Fig. 9, the total energy is strongly dominated by the vdW contribution, and very little by valence energy and electrostatic energy. It suggests that the sliding behavior of bulk shear is mostly controlled by vdW interactions and there are a few changes of molecular conformations of the system in Fig. 3b. Also as a complement, the physical nature of the compression of the mixed HMX-olefin system shown in Fig. 3b were described by the total energy and three partitioned terms in Fig. 10. Similar to the above shear nature, the increased total energy

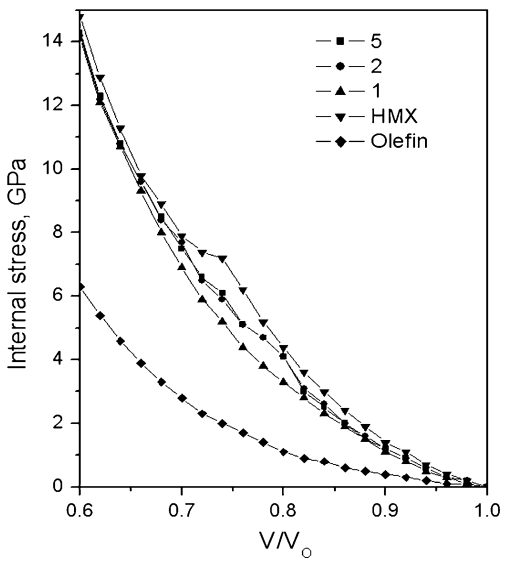



**Fig. 8** Internal stress of five systems induced by compression along *b* axis as a function of $V/V_0$. 1, 2, 5, HMX, and Olefin correspond to Fig. 3a, b and c, neat HMX, and neat olefin, respectively

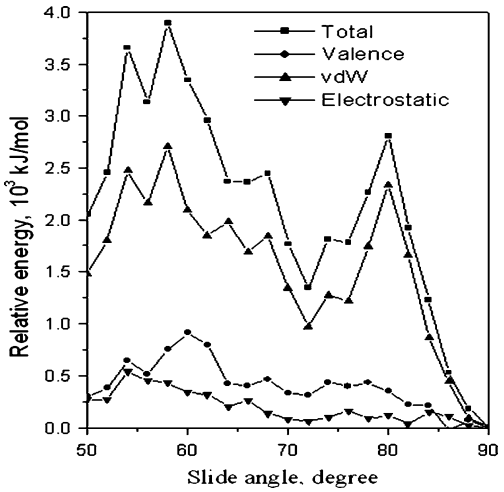



**Fig. 9** Change of energy terms during the bulk shear slide of the mixed HMX+olefin system in Fig. 3b. Relative energy is the energy of the relaxed structure of the *ith* step relative to that of the first step ($\beta$=90$^0$)

during compression is also dominantly due to the increasing of vdW interactions. In compression, the valence energy is enhanced to a certain extent by compressing the molecules in the system, and electrostatic attraction is strengthened a little by shortening the interatomic distances. Summarily, the vdW interactions are the main contributors to the total interaction energy in both the bulk shear slide and the compression in the given range here.

## Interfacial shear slide

The interfacial shear slide of a mixed HMX-olefin system takes place on the interface between HMX and olefin. Three systems in Fig. 5 containing olefin with different dispersion, or different thickness of olefin monolayer, were designed to investigate the effect of olefin dispersion on the sliding behavior.

Firstly, comparing SB of neat HMX, 1776 J·mol$^{-1}$·Å$^{-2}$, with those of the mixed HMX+olefin systems, 190, 307, and 337 J·mol$^{-1}$·Å$^{-2}$, respectively, we can find that the introduction of olefin into HMX can obviously decrease SB to disperse external mechanical energy more readily in the cases of interfacial shear slide. It shows that the good lubricating property of olefin can be apparent during both bulk and interfacial shear slide.

Secondly, as illustrated in Fig. 11, there is few differences of SB, only about 10 J·mol$^{-1}$·Å$^{-2}$, when the interfacial shear slide takes place along the designated orientation *a* or *b*. It can be understood that the sliding potential is determined by the obstructive ability of interfacial molecules against sliding, which can be denoted by the salient parts of HMX molecule on the sliding plane as Fig. 12, like some hills. That of molecules in olefin is not considered because of their good lucubrating property or fluidity. Apparently, all the hills on the sliding plane should be overcome along whichever orientation after a complete slide finished.

Finally, as indicated in Fig. 11, the mixed HMX-olefin systems, a, b, and c, respectively representing the systems in Fig. 5a, b and c with 30, 10 and 5 $C_{22}H_{46}$ chains of olefin, have an increasing order of SB by decreasing the thickness of olefin monolayer: 190, 307, and 337 J·mol$^{-1}$·Å$^{-2}$. Similar to the case of bulk shear slide, excessively thin monolayer of olefin can forbid both HMX molecules and molecules in olefin moving and

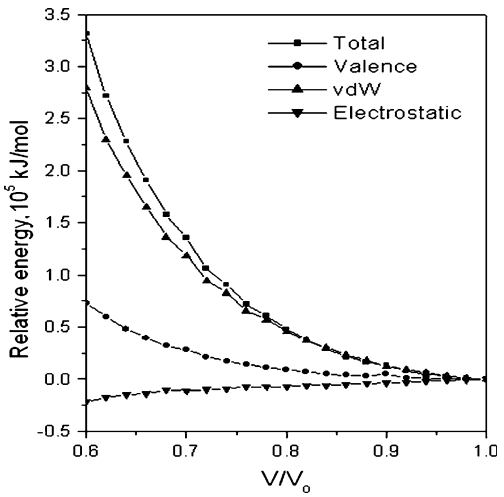



**Fig. 10** Change of energy terms during the unaxial compression along *b* axis of the mixed HMX-olefin system in Fig. 3b. Relative energy is the energy of the relaxed structure of the *ith* step relative to that of the first step ($V/V_0$=1)

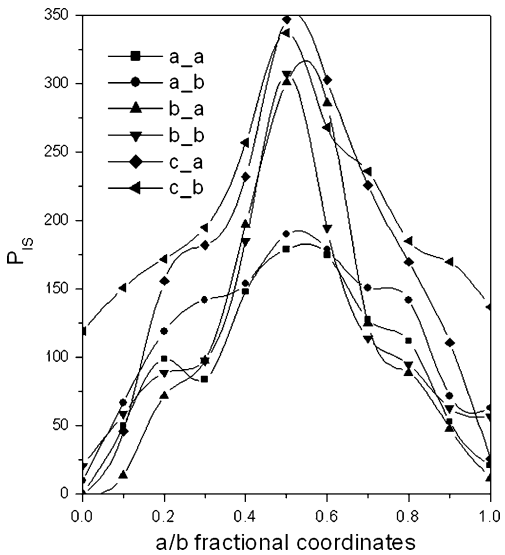



**Fig. 11** Sliding potentials of interfacial shear, PIS (unit in J·mol$^{-1}$·Å$^{-2}$) of the three systems shown in Fig. 5. The first letters of icon a, b and c correspond to the systems in Fig. 5a, b, and c, respectively, the second letter a or b denote the sliding along a or b orientation shown in Fig. 3a

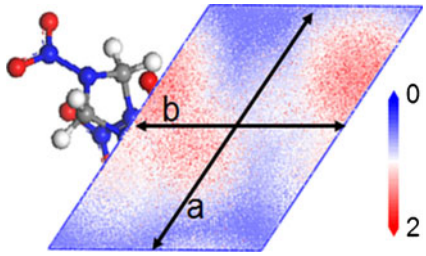



**Fig. 12** Plot showing the height of the salient parts of HMX molecule on the sliding plane, the (010) face

being relaxed enough to lower total energy much. In other words, it is such a thin monolayer that olefin in it has little freedom to display its good lubricating property or fluidity enough. At the same time, too short a distance between the top and the bottom blocks of HMX in the mixed HMX-olefin systems can lead to much sliding potential close to that of neat HMX (or, the distance is zero). For example, the distance of c corresponding to Fig. 5c is only about 5 Å and makes some interactions between the two HMX blocks and further a little anisotropy of sliding. This is the reason that group c has the most difference between two curves (c_a and c_b) in it, in contrast to groups a (a_a and a_b) and b (b_a and b_b).

It can be concluded from above discussion that both in bulk shear sliding and in interfacial shear sliding, excessively thin monolayer of olefin is helpless to decrease the mechanical sensitivity of the mixed explosives containing it.

To expose the physical nature during the interfacial shear sliding, we carried out a series of lattice calculations on 121 spots shown in Fig. 3c for the system in Fig. 5b. As illustrated in Fig. 13, the total energy (Tot) change is also strongly dominated by van der Waals (vdW) force: (1) Tot

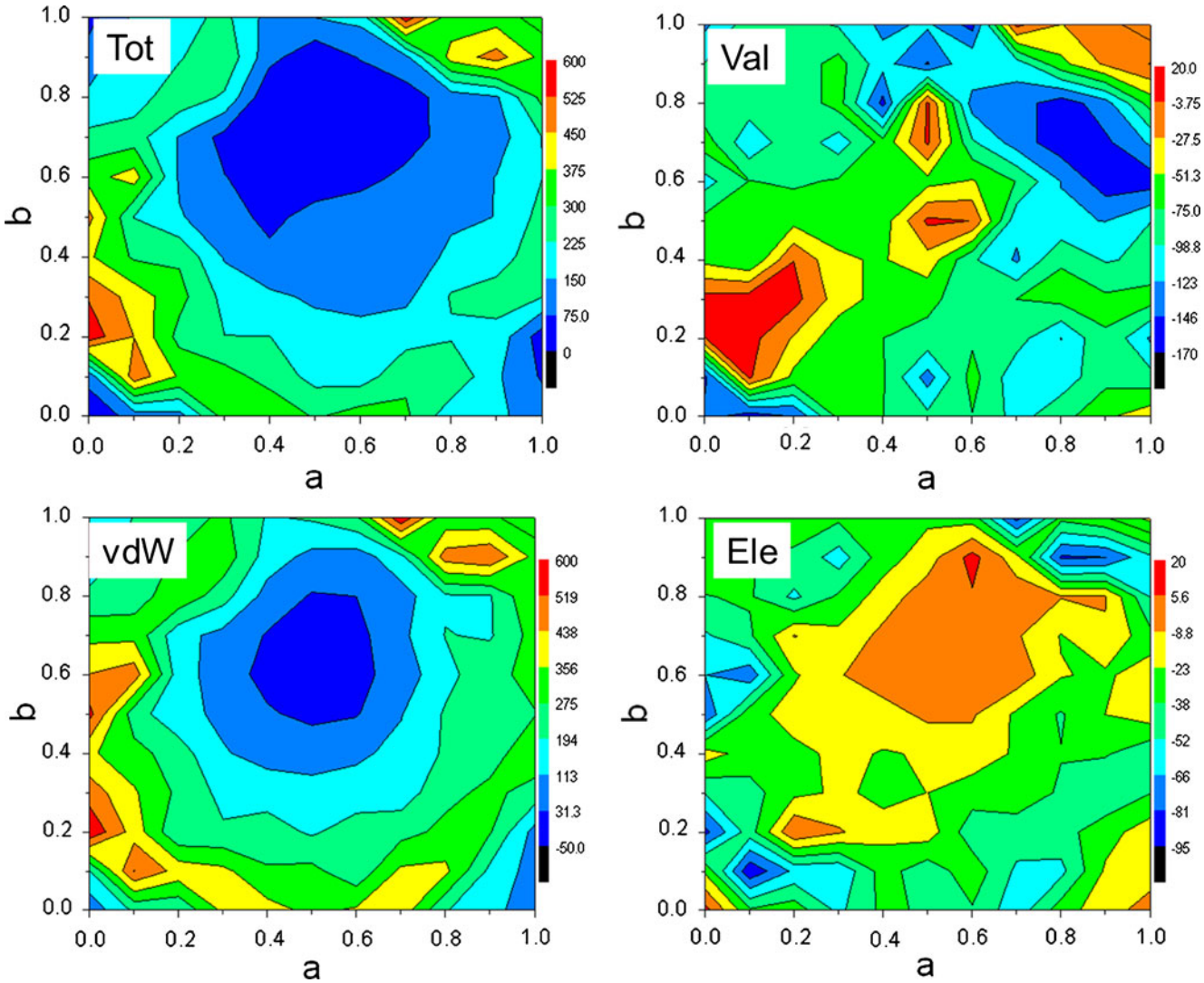



**Fig. 13** Change of energy terms during the interfacial shear slide of the mixed HMX+olefin system of Fig. 9b. Each energy term is relative to that of spot (0.5, 07) with the lowest total energy

and vdW have close energy change ranges, within 0 to 600 $J \cdot mol^{-1} \cdot Å^{-2}$ for Tot and within −50 to 600 $J \cdot mol^{-1} \cdot Å^{-2}$ for vdW, which are much different from those of valence interactions (Val) and electrostatic interactions (Ele), within −170 to 20 and within −95 to 20 $J \cdot mol^{-1} \cdot Å^{-2}$, respectively (2) Tot and vdW have similar contours of energy. Combining this result and that of the case of bulk shear slide, we can draw a conclusion that in both cases of bulk and interfacial shear slide, the energy change is predominantly attributed to the vdW interaction change.

Different from the interfacial shear slide of π-stacked TATB along its *AOB* plane where the change of total interaction energy is mainly caused by the change of electrostatic interaction energy [25], this shear slide involved in mixed HMX-olefin systems is dominated by vdW interactions. We think this difference may be determined by the freedom of interfacial atoms during shear sliding: the more freedom, the more vdW contribution; the less freedom, the more electrostatic contribution. In the interfacial shear slide of TATB along its molecular plane, all atoms almost keep rigid due to the strong conjugated effect and inter-/intra-molecular hydrogen bonding effect, suggesting that the total interatomic distances change a little even though each designated interatomic distance may change much in the systems with periodic structures. That is to say, the total shear slide should not be controlled by vdW interactions, which are only determined by interatomic distances for a given system and change a little. Accordingly the electrostatic interactions become dominant, which are determined by not only interatomic distances but also the sign and the quantity of atomic charges. In the interfacial shear slide between HMX and olefin, the atoms on and near the interface have much freedom to be relaxed to a local energy minimum. It means that vdW interactions may dominate the total interactions.

## Conclusions

Taking some mixed explosives containing a typical mechanical desensitizer olefin and a typical explosive HMX (olefin-HMX) as examples, and using a COMPASS force-field method based on first principle calculations and experimental results, we studied the desensitizing mechanism of olefin versus mechanical stimuli to understand its sliding behaviors including two cases, the bulk shear slide and the interfacial shear slide. As results, the addition of olefin into HMX can obviously decrease the sliding barrier in contrast to the neat HMX, and the total energy change during the shear is dominantly attributed to the change of vdW interaction energy. Moreover, it is found that the thickness of olefin monolayer contained in the mixed explosives can influence its sliding behavior. That is, the excessively thin olefin monolayer can cause a high sliding barrier and is helpless to decrease the mechanical sensitivity of the mixed explosives.

**Acknowledgments** Financial support from the funds of China Academy of Engineering Physics (CAEP) (2009B0302032 and 2011A0302014), the State Key Development Program for Basic Research of China (973–61383), and the National Natural Science Foundation of China (10972025/A020601) is greatly appreciated.

## References

1. Politzer P, Alper HE (1999) In: Leszczynski J (ed) Computational chemistry: reviews of current trends. World Scientific River Edge, NJ, pp 271–286 and references therein
2. Dong HS (2004) Strategy for developing energetic materials. Chin J Energ Mater 12:1–12, and references therein
3. Dong HS, Zhou FF (1994) (ed) High energetic explosives and relatives. Science Press, Beijing
4. Zhang C (2010) Understanding the desensitizing mechanism of olefin in explosives versus external mechanical stimuli. J Phys Chem C 114:5068–5072
5. Sun H (1998) COMPASS: An ab initio force-field optimized for condensed-phase applications – overview with details on alkane and benzene compounds. J Phys Chem B 102:7338–7364
6. Sorescu DC, Rice BM (2010) Theoretical predictions of energetic molecular crystals at ambient and hydrostatic compression conditions using dispersion corrections to conventional density functionals (DFT-D). J Phys Chem C 114:6734–6748
7. Li J (2010) Relationships for the impact sensitivities of energetic C-nitro compounds based on bond dissociation energy. J Phys Chem B 114:2198–2202
8. Landerville AC, Conroy MW, Oleynik II, White CT (2010) Prediction of isothermal equation of state of an explosive nitrate ester by van der Waals density functional theory. J Phys Chem Lett 1:346–348
9. Zeman S, Atalar T, Friedl Z, Ju XH (2009) Accounts of the New Aspects of Nitromethane Initiation Reactivity. Central Eur J Energ Mater 6:119–119
10. Murray JS, Lane P, Göbel M, Klapötke TM, Politzer P (2009) Intra- and intermolecular electrostatic interactions and their significance for the structure, acidity, and tautomerization behavior of trinitromethane. J Chem Phys 130:104304
11. Dawes R, Siavosh-Haghighi A, Sewell TD, Thompson DL (2009) Shock-induced melting of (100)-oriented nitromethane: Energy partitioning and vibrational mode heating. J Chem Phys 131:224513
12. Isayev O, Gorb L, Qasim M, Leszczynski J (2008) Ab Initio Molecular Dynamics Study on the Initial Chemical Events in Nitramines: Thermal Decomposition of CL-20. J Phys Chem B 112:11005–11013
13. Citroni M, Datch F, Bini R, Vaira MD, Pruzan P, Canny B, Schettino V (2008) Crystal Structure of Nitromethane up to the Reaction Threshold Pressure. J Phys Chem B 112:1095–1103
14. Swadley MJ, Li T (2007) Reaction mechanism of 1,3,5-Trinitro-s-triazine (RDX) deciphered by density functional theory. J Chem Theor Comput 3:505–513
15. Ye S, Koshi M (2006) Theoretical Studies of Energy Transfer Rates of Secondary Explosives. J Phys Chem B 110:18515–18520
16. Zhang C, Shu Y, Huang Y, Zhao X, Dong H (2005) Investigation of correlation between impact sensitivities and nitro group charges in nitro compounds. J Phys Chem B 109:8978–8982

17. Rice BM, Hare JJ (2002) A Quantum Mechanical Investigation of the Relation between impact sensitivity and the charge distribution in energetic molecules. J Phys Chem A 106:1770–1783
18. Zhang MX, Eaton PE, Gilardi R (2000) Hepta- and Octanitrocubanes. Angew Chem Int Edit 39:401–404
19. Politzer P, Murray J, Concha M, Lane P (2007) Effects of electric fields upon energetic molecules: nitromethane and dimethylnitramine. Central Eur J Energ Mater 4:3–21
20. Zhang C (2009) Review of the establishment of nitro group charge method and its applications. J Hazard Mater 161:21–28
21. Yang L (2008) Surface polarity of β-HMX crystal and the related adhesive forces with estane binder. Langmuir 24:13477–13482
22. Reed EJ, Joannopoulos JD, Fried LE (2000) Electronic excitations in shocked nitromethane. Phys Rev B 62:16500–16509
23. Zhang C (2007) Investigation of the slide of the single layer of the 1,3,5-Triamino-2,4,6-trinitrobenzene crystal: sliding potential and orientation. J Phys Chem B 111:14295–14298
24. Zhang C (2007) Computational investigation on the desensitizing mechanism of graphite in explosives versus mechanical stimuli: compression and glide. J Phys Chem B 111:6208–6213
25. Zhang C, Wang X, Huang H (2008) π-Stacked interactions in explosive crystals: buffers against external mechanical stimuli. J Am Chem Soc 130:8359–8365
26. Kuklja MM, Rashkeev SN (2007) Shear-strain-induced chemical reactivity of layered molecular crystals. Appl Phys Lett 90:151913
27. Kuklja MM, Rashkeev SN (2007) Shear-strain-induced structural and electronic modifications of the molecular crystal 1,1-diamino-2,2-dinitroethylene: Slip-plane flow and band gap relaxation. Phys Rev B 75:104111
28. Cady HH, Larson AC, Cromer DT (1963) The crystal structure of α-HMX and a refinement of the structure of β-HMX. Acta Cryst 16:617–623
29. Gruzdkov YA, Gupta YM (1998) Mechanism of amine sensitization in shocked nitromethane. J Phys Chem A 102:2322–2331
30. Xiao J, Fang G, Ji G, Xiao H (2005) Simulation investigations in the binding energy and mechanical properties of HMX-based polymer-bonded explosives. Chin Sci Bull 50:21–26
31. Material Studio 4.0 (2005) Acceryls Inc, San Diego

J Mol Model (2012) 18:1513–1521
DOI 10.1007/s00894-011-1171-8

ORIGINAL PAPER

# Structure and electronic absorption spectra of nematogenic alkoxycinnamic acids – a comparative study based on semiempirical and DFT methods

**Pogula Lakshmi Praveen · Durga Prasad Ojha**

Received: 15 March 2011 / Accepted: 29 June 2011 / Published online: 23 July 2011


**Abstract** Structure of nematogenic *p*-*n*-Alkoxy cinnamic acids (*n*OCAC) with various alkyl chain carbon atoms ($n$=2, 4, 6, 8) has been optimized using density functional B3LYP with 6-31+G (d) basis set using crystallographic geometry as input. Using the optimized geometry, electronic structure of the molecules has been evaluated using the semiempirical methods and DFT calculations. Molecular charge distribution and phase stability of these systems have been analyzed based on Mulliken and Löwdin population analysis. The electronic absorption spectra of *n*OCAC molecules have been simulated by employing DFT method, semiempirical CNDO/S and INDO/S parameterizations. Two types of calculations have been performed for model systems containing single and double molecules of *n*OCAC. UV-Visible spectra have been calculated for all single molecules. The UV stability of the molecules has been discussed in light of the electronic transition oscillator strength (*f*). The dimer complexes of higher homologues ($n$=6, 8) have also been reported to enable the comparison between single and double molecules.

**Keywords** Nematogen · Oscillator strength · Phase stability

## Introduction

A variety of liquid crystal displays (LCDs) lead the market for flat panel displays, especially for handy applications. The common facets of these devices are low weight, thin planar construction, low operating voltages and power consumption, acceptable contrast and viewing angles [1]. The design and synthesis of new organic compounds, which exhibit a nematic phase with a specific spectrum of properties, have essential contributions in establishing and expanding the multibillion dollar LCD industry. The need to develop new liquid crystalline (LC) compounds, for desired applications and technology provides a new motivation to develop a deeper understanding in molecular electronic structure and bonding [2, 3]. Even though advances in synthesis and characterization explore many achievements [4] regarding nematic materials, it is often desirable; to have a sense of material behavior prior to synthesis [5].

LC behavior is a genuine supramolecular phenomenon based on the existence of extended weak interactions (dipole-dipole, dispersion forces) between the molecules. It is usually necessary that these molecules have anisotropic shapes [6] able to pack efficiently, so that these weak interactions can be high in number, co-operate, and become strong as to keep the molecules associated in a preferred orientation. The differences between the nematic behavior of simple conventional mesogens (monomers) and dimers are thought to stem entirely from the configurational correlations [7] that the spacers impose on the mesogenic units. Thus, the inclusion of such correlations constitutes the major step in the formulation of a molecular theory that goes from the description of aggregation behavior of single mesogen to dimer, and to multi-mesogen phases. Besides the common display applications, it can be expected that LCs will also be used as a carrier material in the examination of non-linear optical properties of organic molecules.

The simulation of UV-visible spectra by computational chemistry tools is particularly appealing since a large

P. L. Praveen · D. P. Ojha (✉)
Liquid Crystal Research Laboratory,
Post-Graduate Department of Physics, Andhra Loyola College,
Vijayawada 520 008 AP, India
e-mail: durga_ojha@hotmail.com

number of methods have been elaborated and employed to calculate the absorption wavelength and oscillator strength of electronic transitions. In this sense, methods based on time dependent density functional theory (TDDFT) applied to small and middle sized systems [8, 9] provide rather good accuracy at low computational cost [10, 11]. These methods still remain of rather limited application for establishing realistic molecular models and biological systems. Hence, the alternate use of semiempirical schemes has an extensive use to determine absorption spectra as well as electronic transition oscillator strength. Such approaches allow for the calculation of electronic transitions between the ground state and the different excited states, which gives the energies of the corresponding radiations. Each transition can be enlarged with a Gaussian shape due to several factors such as thermal excitement, natural line width etc.

The present theoretical contribution aims at providing a comparative picture of structural and optical properties of nematogenic *p-n*-Alkoxycinnamic acids (*n*OCAC, *n*=2, 4, 6, 8) using semiempirical CNDO/S (complete neglect of differential overlap/ spectroscopy) [12, 13], INDO/S (intermediate neglect of differential overlap/ spectroscopy) [14, 15] schemes and DFT method [10]. In particular, it concentrates on the alteration of these properties with respect to homologue number. The highest occupied molecular orbital (HOMO)/ lowest unoccupied molecular orbital (LUMO) energies, oscillator strength (*f*) have also been reported for all single molecules using these three methods. The dimer complexes of higher homologues (*n*=6, 8) during the different modes of interactions and their properties have been reported based on DFT calculations. An attempt has been made to find out the most energetically favorable configuration in each mode of interaction. An examination of thermodynamic data has revealed that *n*OCAC molecules exhibits nematic-isotropic transition temperature as follows [16]: 2OCAC at 471 K; 4OCAC at 460.5 K; 6OCAC at 452 K; and 8OCAC at 445 K.

## Computational methods

The present model is comprised of the monomer and dimer assemblies of *p-n*-Alkoxy cinnamic acids (*n*OCAC, *n*=2, 4, 6, 8). The geometry optimizations have been performed using the density functional theory (DFT) approach [10] and in particular the Becke3-Lee-Yang-Parr hybrid functional (B3LYP) exchange-correlation functional and the 6-31G+(d) basis set. The DFT approach was originally developed by Hohenberg and Kohn [17], Kohn and Sham [18, 19] to provide an efficient method of handling many-electron system. The theory allows us to reduce the problem of an interacting many-electron system to an effective single-electron problem. On the basis of the DFT geometries, the electronic structures, and excitation energies and excited state wave functions have been calculated coupled with the configuration interaction (CI) single level of approximation including all $\pi\rightarrow\pi^*$ single excitations. This has been found adequate to determine the UV-visible absorption spectra [20, 21] provided that suitable parameterizations are used. In the present work, a comparative analysis has been carried out by employing CNDO/S, INDO/S, and DFT methods to simulate the electronic absorption spectrum. The DFT calculations have been performed by a spectroscopy oriented configuration interaction procedure (SORCI) [22], whereas, a revised version of QCPE 174 by Jeff Reimers, University of Sydney and coworkers have been used for the semiempirical calculations [23].

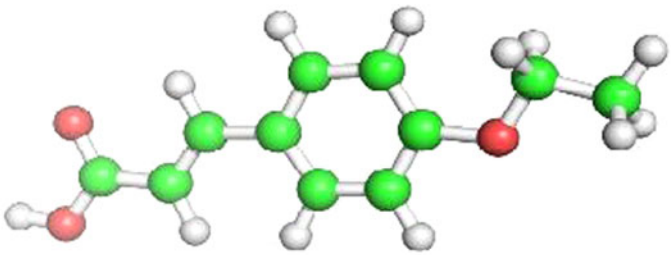
The general structural parameters of the systems such as bond lengths and bond angles have been taken from published crystallographic data [16]. The end chains have the all-*trans* extended conformation, and the molecules exist in the crystal as planar hydrogen-bonded dimers. The dimers are arranged in end-to-end fashion in parallel rows. Similarly, side-to-side packing pairs of dimers are found to give a good fit between adjacent aromatic cores. Charge distributions of the molecules have been calculated by performing Mulliken and Loewdin population analysis. The mesomorphic behavior and the nematic phase stability have been predicted through the calculated local charge distributions.

## Results and discussion

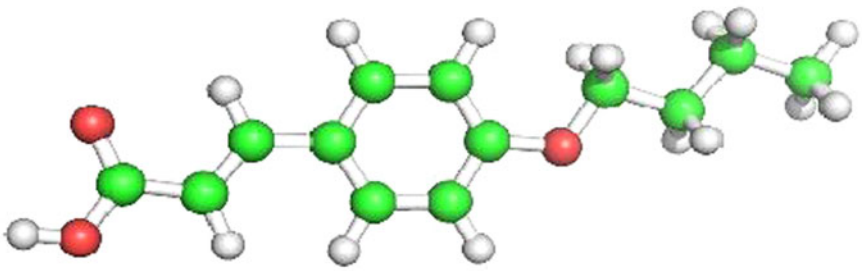
The electronic structures of *n*OCAC (*n*=2, 4, 6, 8) have been shown in Fig. 1. The structures have been constructed on the basis of published crystallographic data with standard values of bond lengths and bond angles [16]. Generally, the *trans* isomers are linear in shape and tend to form mesophases, while the *cis* isomers are not linear, and tend not to form mesophases. The structures of the systems have the all-*trans* extended conformation, and the molecules exist in the crystals as planar hydrogen-bonded dimers. The molecular charge distribution, and phase stability of the molecules has been analyzed as given below:

### Molecular charge distribution

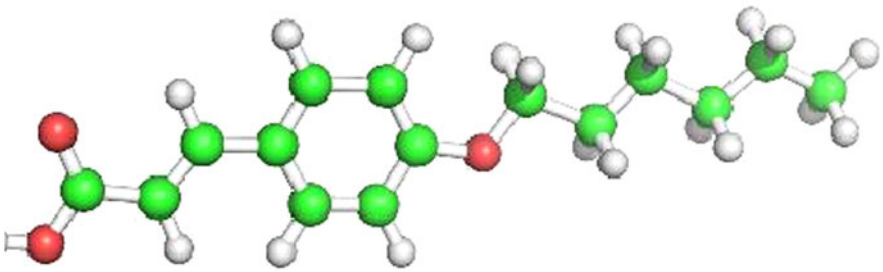
It is expected that the specific charge distributions and electrostatic interactions in LC molecules play an influential role in the formation of various mesophases. An appropriate modeling of this fundamental molecular feature relies on the possibility of assigning a partial charge to all atomic



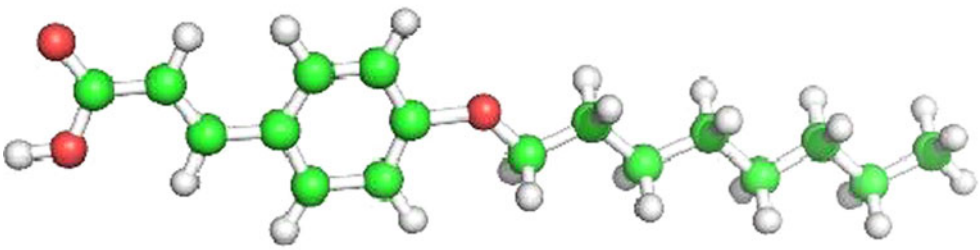
**(a) 2OCAC**



**(b) 4OCAC**



**(c) 6OCAC**



**(d) 8OCAC**

**Fig. 1** The electronic structures of (**a**) 2OCAC, (**b**) 4OCAC, (**c**) 6OCAC, and (**d**) 8OCAC molecules

centers [24]. Analysis of molecular charge distribution can deliver good information about local electrostatic interactions, which is not possible from an experimental point of view. To parameterize the molecular interactions for computational studies atom positioned partial charges are helpful [25]. Quantum chemical computations offer the possibility to take a detailed look at the electronic structure of the molecules. This can be done, by determining atom-based partial charges, which are not quantum mechanical observables. These charges represent the electrostatic molecular interactions very well, but they do not show the real charge distribution in the molecule.

**Table 1** Mulliken (M), Loewdin (L) group charges and nematic-isotropic transition temperatures for *n*OCAC (*n*=2, 4, 6, 8) molecules

| Molecule | Side group | | Core | | Alkyl | | $T_{N-I}$ [16] |
|---|---|---|---|---|---|---|---|
| | M | L | M | L | M | L | |
| 2OCAC | −0.39 | −0.34 | −0.25 | −0.18 | 0.63 | 0.52 | 471 |
| 4OCAC | −0.40 | −0.35 | −0.21 | −0.17 | 0.61 | 0.52 | 460.5 |
| 6OCAC | −0.40 | −0.35 | −0.21 | −0.16 | 0.62 | 0.52 | 452 |
| 8OCAC | −0.38 | −0.34 | −0.29 | −0.21 | 0.67 | 0.55 | 445 |

Since group charges are needed to explain the behavior of mesogens, Mulliken population analysis, which partitions the total charge among the atoms in the molecule, has been performed, and the results have been compared with those obtained from Löwdin population analysis. However, there is much agreement among the methods while it comes to the group charges of each molecule. It is evident from Table 1 that the positively charged alkyl chains of 2OCAC will be strongly attracted by the negatively charged side group as well as the core, causing the formation of longer units in the nematic phase. Hence, the nematic phase stability is expected to be high for 2OCAC. Further, the thermal vibration amplitude of the chain carbon atoms increase markedly with the increase of chain length, indicating a low packing efficiency for higher homologues. This leads to the drastic decrease in nematic phase stability. This is in agreement with the nematic-isotropic transition temperature reported by the crystallographer (Table 1).

## Electronic absorption spectrum

One of the most exciting developments in liquid crystal science and technology is the possibility of using light instead of electricity to control the behavior of a material [26]. Increasing the number of carbon atoms in the end chain is the widely used technique to alter the physical properties of LC molecules. The description of molecular quantities by quantum chemical methods underlies some principle restrictions, *i.e.*, there exists a compromise between the complexity of the systems studied and the accessible theoretical accuracy. For the calculation of electronic spectra, the configuration interaction (CI) method is widely employed. Using a CI method in combination with a semi-empirical model Hamiltonian, an evaluation of absorption spectra of large organic molecules and LCs becomes possible [27–29]. It is expected that the absorption spectrum of a LC molecule containing a single phenyl ring would exhibit similarities to an isolate benzene molecule. The principal absorption bands in the molecules are due to the $\pi \rightarrow \pi^*$ transitions in the benzene analogous part of the molecule. In general, these benzene-like transitions are roughly conserved in the model systems studied, but they are influenced by the conjugation length, degree of conjugation, and the different substituents.

The experimental studies show that cinnamic acid exhibits two strong absorptions at 220 nm and 267 nm ($\lambda_{max}$) [30]. The present calculations have been carried out on cinnamic acid derivatives (*n*OCAC, *n*=2, 4, 6, 8) to study their ultraviolet stability in the light of the shift of absorption wavelength with respect to the different substitution of various alkoxy groups. The variations in the absorption spectra have been observed due to the different substitutions. The detailed picture of electronic absorption spectra of isolated molecules based on TDDFT calculations have been given below. Further, the electronic spectral data based on semiempirical schemes have been reported in Tables 2 and 3.

(a) p-n-Ethoxy cinnamic acid (2OCAC):

The electronic absorption spectrum of 2OCAC is shown in Fig. 2. In the UV region, three strong absorptions at 203.52 nm ($\lambda_1$), 224.02 nm ($\lambda_2$), and 295.51 nm ($\lambda_3$) have

**Table 2** The vertical excited energy ($E_V$), and oscillator strength (*f*) of *n*OCAC molecules corresponding to absorption maxima ($\lambda_{max}$) at TDDFT, CNDO/S, and INDO/S levels

| Molecule | $\lambda_{max}$/nm | | | $E_V$/eV | | | *f* | | |
|---|---|---|---|---|---|---|---|---|---|
| | TDDFT | CNDO/S | INDO/S | TDDFT | CNDO/S | INDO/S | TDDFT | CNDO/S | INDO/S |
| 2OCAC | 295.51 | 261.52 | 275.00 | 4.19 | 4.74 | 4.50 | 0.69 | 0.93 | 0.95 |
| 4OCAC | 310.74 | 271.48 | 288.47 | 3.99 | 4.57 | 4.29 | 0.56 | 0.87 | 0.95 |
| 6OCAC | 291.99 | 256.83 | 273.24 | 4.25 | 4.82 | 4.54 | 0.65 | 0.90 | 0.90 |
| 8OCAC | 298.44 | 260.93 | 276.17 | 4.15 | 4.76 | 4.51 | 0.69 | 0.88 | 0.82 |

**Table 3** The HOMO, LUMO energies, and energy gap ($E_g$) of *n*OCAC molecules at TDDFT, CNDO/S, and INDO/S levels

| Molecule | HOMO/eV | | | LUMO/eV | | | $E_g=E_{LUMO}-E_{HOMO}$/eV | | |
|---|---|---|---|---|---|---|---|---|---|
| | TDDFT | CNDO/S | INDO/S | TDDFT | CNDO/S | INDO/S | TDDFT | CNDO/S | INDO/S |
| 2OCAC | −8.76 | −9.30 | −8.19 | −0.95 | −1.62 | −0.83 | 7.81 | 7.68 | 7.36 |
| 4OCAC | −8.54 | −9.07 | −7.97 | −0.97 | −1.60 | −0.84 | 7.57 | 7.47 | 7.13 |
| 6OCAC | −8.75 | −9.28 | −8.18 | −0.88 | −1.49 | −0.75 | 7.87 | 7.79 | 7.43 |
| 8OCAC | −8.90 | −9.40 | −8.30 | −1.14 | −1.74 | −0.98 | 7.76 | 7.66 | 7.32 |

been observed. However, no absorption has been observed in the visible region. The strongest band appears in a region of 276.17 nm to 320.12 nm with absorption maxima ($\lambda_{max}$) at 295.51 nm ($\lambda_3$). This band arises from the HOMO→LUMO transition and is assigned as π→π* transitions in the molecule. The oscillator strength (f) values corresponding to three wavelengths are 0.39, 0.32, and 0.69 respectively. Further, the calculation also predicts π→π* transitions corresponding to weak absorption bands at the remaining two wavelengths ($\lambda_1$ and $\lambda_2$).

(b) p-n-Butoxy cinnamic acid (4OCAC):

Figure 2 shows the three prominent bands in the UV region with absorption maxima at 202.93 nm ($\lambda_1$), 226.37 nm ($\lambda_2$), and 310.74 nm ($\lambda_3$). The strongest band has been observed from 275 nm to 325.39 nm with absorption maxima at 310.74 nm. This band arises due to the HOMO→LUMO transition and is assigned as π→π* transitions in the molecule. The oscillator strength (f) values corresponding to three wavelengths are 0.36, 0.32, and 0.56 respectively. However, the other weak absorption bands at $\lambda_1$ and $\lambda_2$ also indicate the possibility of π→π* transitions.

(c) p-n-Hexyloxy cinnamic acid (6OCAC):

Three strong absorption bands have been noticed for 6OCAC in the UV region with absorption maxima at 208.79 nm ($\lambda_1$), 228.71 nm ($\lambda_2$), and 291.99 nm ($\lambda_3$) (Fig. 2). The strongest band has been observed in a region from 271.48 nm to 313.09 nm with absorption maxima at 291.99 nm, which arises due to the transition between HOMO→LUMO energy states. This can be assigned as π→π* transitions in the molecule. The oscillator strength (f) values corresponding to three wavelengths are 0.29, 0.28, and 0.65 respectively. However, the appearance of other two bands indicates the possibility of π→π* transitions.

(d) p-n-Octyloxy cinnamic acid (8OCAC):

The absorption spectrum of 8OCAC molecule (Fig. 2) exhibits three strong absorption bands have been noticed in the UV region with absorption maxima at 208.79 nm ($\lambda_1$), 233.4 nm ($\lambda_2$), and 298.44 nm ($\lambda_3$). The strongest UV band has been noticed in a region from 269.73 nm to 321.29 nm with absorption maxima at 298.44 nm with one shoulder at 280.27 nm. This arises due to the transition between

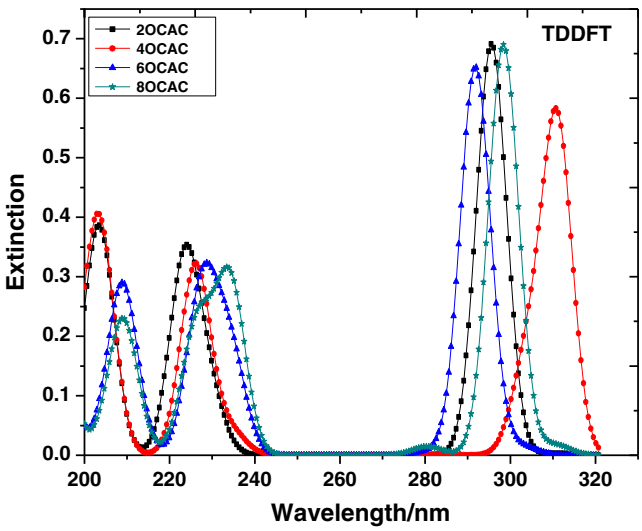



**Fig. 2** Electronic absorption spectra of *n*OCAC molecules using TDDFT approximation. Extinction unit: $10^4$ $dm^3$ $mol^{-1}$ $cm^{-1}$

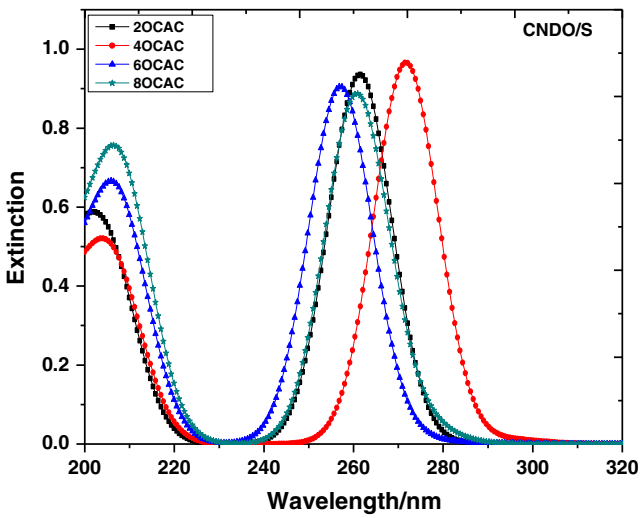



**Fig. 3** Electronic absorption spectra of *n*OCAC molecules using CNDO/S approximation. Extinction unit: $10^4$ $dm^3$ $mol^{-1}$ $cm^{-1}$

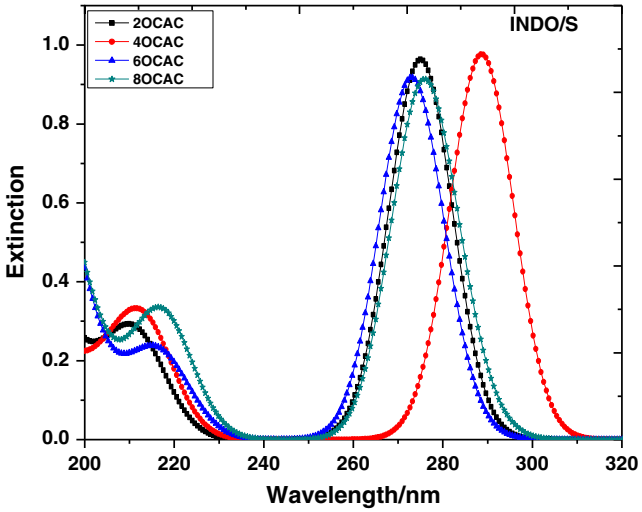



**Fig. 4** Electronic absorption spectra of *n*OCAC molecules using INDO/S approximation. Extinction unit: $10^4$ $dm^3$ $mol^{-1}$ $cm^{-1}$

HOMO→LUMO energy states, and assigned as π→π* transitions in the molecule. The oscillator strength (f) values corresponding to three wavelengths are 0.23, 0.29, and 0.69 respectively. However, the appearance of other two bands indicate the possibility of additional π→π* transitions.

Thus, for 2OCAC, substitution of two additional alkyl groups (forming 4OCAC), and six additional alkyl groups (forming 8OCAC) leads to a bathochromic shift (the shift of absorption maxima to a longer wavelength), while addition of four alkyl groups (forming 6OCAC) causes a hypsochromic shift (the shift of absorption maxima to a shorter wavelength) respectively. Further, increment in the alkyl chain length does not show any hyperchromic effect (increment in absorbance). However, the substitution of two and four additional alkyl groups in 2OCAC (forming 4OCAC, and 6OCAC respectively) causes a hypochromic effect (decrement in absorbance) as evident from Table 3. The electronic absorption spectra of *n*OCAC (*n*=2, 4, 6, 8) molecules have been reported in Figs. 3 and 4 using the CNDO/S and INDO/S parameterizations respectively. Evidently, both the DFT and INDO/S methods show a good agreement in exhibiting the shift of absorption wavelength. The tendency of increment/ decrement of absorption wavelength with respect to increment in homologue number are the same in both methods. However, a little deviation in shift of absorption maxima has been observed for 8OCAC molecule using the CNDO/S method.

A comparative picture of the vertical excited energy ($E_V$), and the oscillator strength (*f*) corresponding to absorption maxima ($\lambda_{max}$) using the CNDO/S, INDO/S, and TDDFT methods have been reported in Table 2. Further, a comparison of HOMO, LUMO energies, and the energy gap ($E_g$) values of isolated molecules have been made as reported in Table 3. Evidently, the molecules do not show any preference for increment in homologue number to exhibit HOMO/ LUMO energies. However, the energy gap ($E_g$=$E_{LUMO}$ − $E_{HOMO}$) shows a preference with an increase in homologue number. A bathochromic shift and hypsochromic shift in the series of molecules causes a decrement, and an increment in the energy gap respectively. The molecule 2OCAC exhibits higher oscillator strength (*f*) among the four molecules. Further, a comparative analysis of electronic transitions among the methods (CNDO/S, INDO/S, and DFT) reveals that the maximum oscillator strength is exhibited by 4OCAC molecule in UV range at longer wavelength side (Table 2). Hence, the UV stability of 4OCAC is high among *n*OCAC molecules, which may be exploited for several optoelectronic applications.

## Intermolecular interactions between pair of molecules

The results obtained through these computations suggest that the optical properties of isolated molecules are influenced by their electronic structures. Therefore, in order to understand the self-organizing ability of mesogens, the different modes of molecular interactions namely; stacking, in-plane, and terminal interactions have been considered between a pair of *n*OCAC (*n*=2, 4, 6, 8) molecules. The interaction energies of dimer complexes have been taken to investigate the most energetically stable configuration using the DFT method. Further, the optical properties of dimer complexes have been reported only for the higher homologues (*n*=6, 8) as shown in Table 4. The basic idea

**Table 4** Electronic spectral data of *n*OCAC (*n*=6, 8) dimer molecules at TDDFT level during various modes of interactions

| Molecule | λ/ nm | Extinction* | HOMO/eV | LUMO/eV | $E_g$/eV | *f* |
|---|---|---|---|---|---|---|
| 6OCAC | | | | | | |
| Stacking | 296.68 | 1.22 | −8.68 | −0.96 | 7.72 | 1.20 |
| In-plane | 292.58 | 1.33 | −8.83 | −0.95 | 7.88 | 1.33 |
| Terminal | 292.58 | 1.15 | −8.50 | −0.78 | 7.72 | 1.13 |
| 8OCAC | | | | | | |
| Stacking | 304.30 | 1.12 | −8.85 | −1.21 | 7.64 | 1.01 |
| In-plane | 299.02 | 1.42 | −8.99 | −1.22 | 7.77 | 1.41 |
| Terminal | 300.20 | 1.26 | −8.61 | −1.02 | 7.59 | 1.25 |

* Extinction unit: $10^4$ $dm^3$ $mol^{-1}$ $cm^{-1}$

underlying the study of molecular conformations is to study the physical and chemical properties of compounds that are closely related to the preferred conformations.

The conformational behavior of LCs displays a large variation in intermolecular effects that depends on the nature and magnitude of interactions. Each conformation may exhibit the distinct energy, and lower energy conformations will be populated in preference to those of higher energy. The most energetically stable configurations of 6OCAC and 8OCAC molecules have been shown in Figs. 5 and 6 respectively. A comparison of stacked dimers between both the molecules suggests that the extension of the chain length to eight carbon atoms, a recognizable segregation of the dimers into a highly tilted layer structure has been obtained. The mutual interaction between the dimers in the structure is, however, quite weak, in particular to chain atoms. Hence, the end chains provide enough disorder to the crystal to pass on to nematic rather than smectic phase. The chains have the all-*trans* extended conformation, and molecules exist in

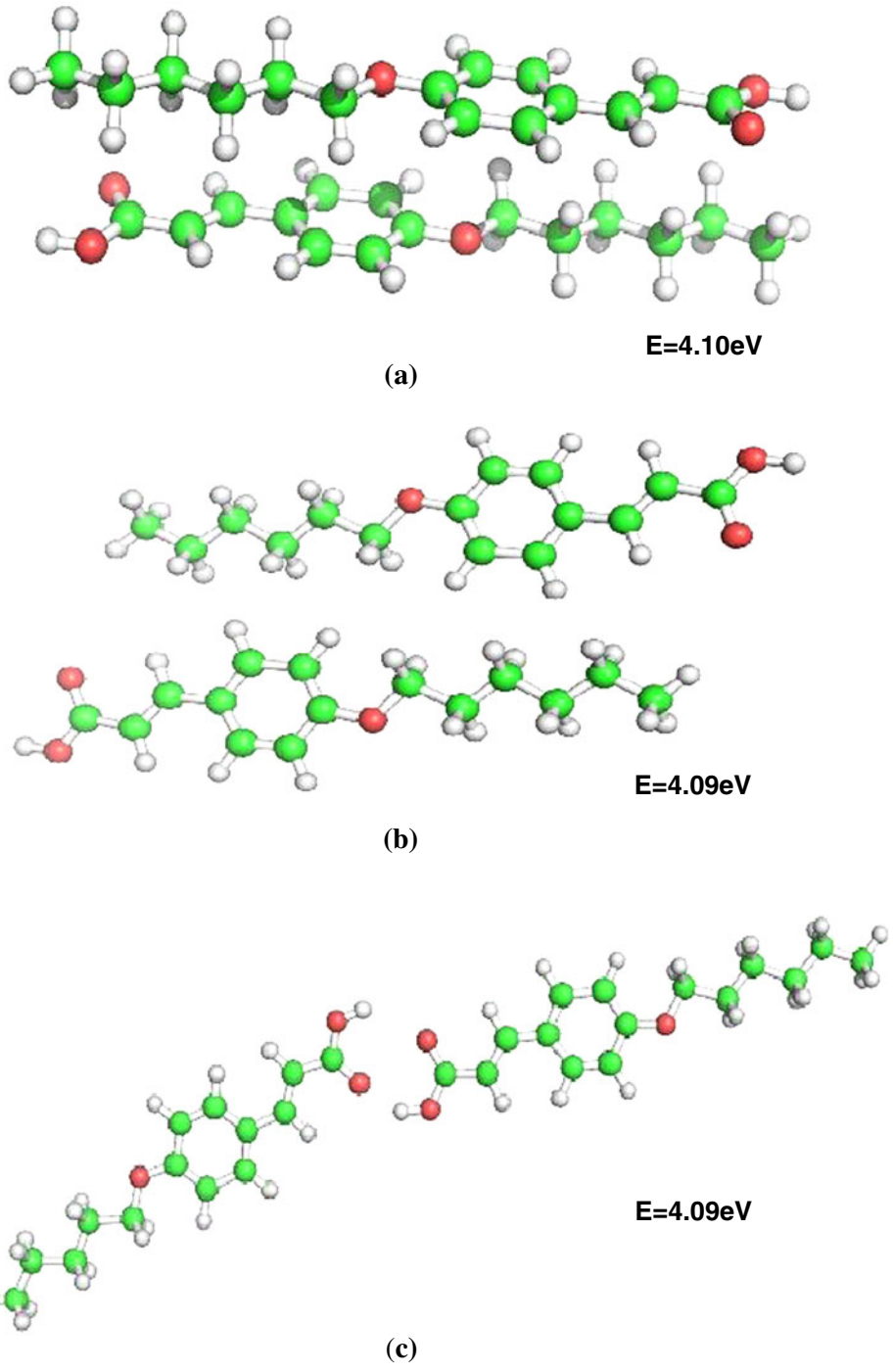



**Fig. 5** Energetically favorable structures of 6OCAC dimer in (**a**) stacking, (**b**) in-plane, and (**c**) terminal interactions

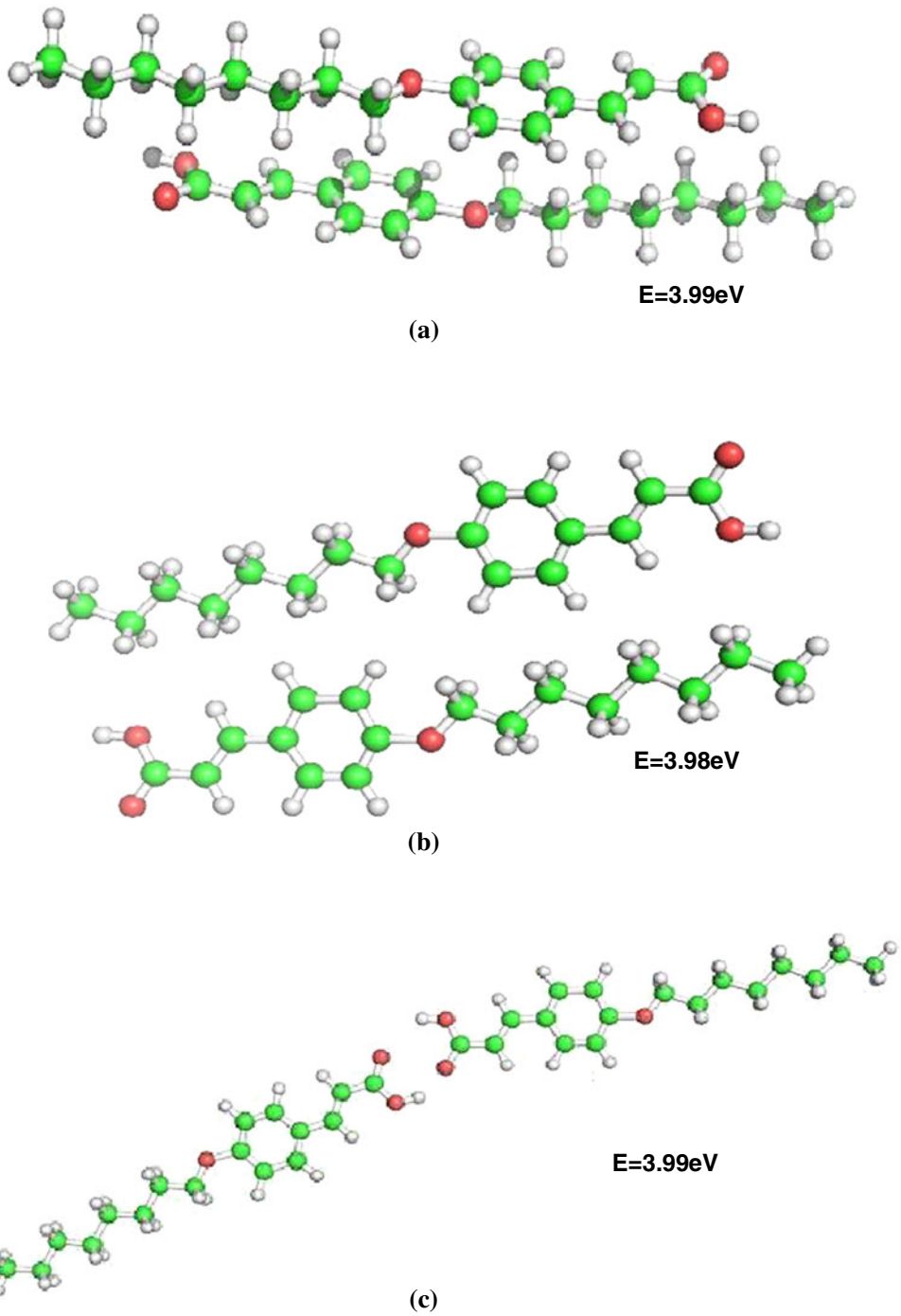



**Fig. 6** Energetically favorable structures of 8OCAC dimer in (**a**) stacking, (**b**) in-plane, and (**c**) terminal interactions

the crystal as planar hydrogen-bonded dimers. The dimers are arranged in end-to-end fashion in parallel rows. Similarly, side-to-side packing of dimers (Figs. 5b and 6b) is found in each molecule in a good fit between adjacent aromatic cores. But the amplitudes of thermal vibration of the carbon chain atoms increase markedly with the increase of chain length, indicating a low packing efficiency, and low nematic-isotropic transition temperatures that is in agreement with charge distribution analysis reported in this article.

## Conclusions

- A comparative analysis of the electronic absorption spectra of nematogenic alkoxy cinnamic acids has been carried out by using the DFT, CNDO/S, and INDO/S methods that show strong absorption bands in the UV region. The maximum absorption wavelength ($\lambda_{max}$) has been obtained for 4OCAC molecule. Further, the molecules show strongest absorption band due to the transition between HOMO→LUMO energy states that

can be assigned as π→π* transitions in the molecules. However, the appearance of other bands indicate the possibility of weak absorptions exhibiting additional π→π* transitions.

- The substitution of two, and six additional alkyl groups in 2OCAC (forming 4OCAC, and 8OCAC respectively) leads to a bathochromic shift. An addition of four alkyl groups (forming 6OCAC) causes a hypsochromic shift. Further, the increment in the alkyl chain length does not create any hyperchromic effect, however, the substitution of two, and four additional alkyl groups in 2OCAC (forming 4OCAC, and 6OCAC respectively) causes a hypochromic effect. A bathochromic shift in the series of molecules causes a decrement in energy gap, while the hypsochromic shift causes an increment in the energy gap.
- A good agreement among the methods has been observed in exhibiting the spectral parameters with respect to increase in the homologue number.
- A comparison of stacked dimers of *n*OCAC molecules suggests that the extension of the chain length to eight carbon atoms, a recognizable segregation of the dimers into a highly tilted layer structure has been observed. The mutual interaction between the dimers in the structure is quite weak, in particular to chain atoms. Hence, the end chains provide enough disorder to the crystal to pass on to nematic rather than smectic phase.

**Acknowledgments** The financial support rendered by the Council of Scientific & Industrial Research (CSIR), and University Grants Commission (UGC), New Delhi, India is gratefully acknowledged.

## References

1. Ramamoorthy A (2007) Thermotropic liquid crystals: recent advances. Springer, Berlin
2. Zhang Z, Guo H (2010) J Chem Phys 133:144911–144923
3. Gorkunov MV, Osipov MA, Kocot A, Vij JK (2010) Phys Rev E 81(061702):1–10
4. Druzbicki K, Mikuli E, Minoslawa D, Chrusciel O (2010) Vib Spectroscopy 52:54–62
5. Castari M, Bosco A, Ferrarini A (2009) J Chem Phys 131:054104–054120
6. Peroukidis SD, Vanakaras AG, Photinos DJ (2008) J Phys Chem B 112:12761–12767
7. Ojha DP (2006) J Mol Model 12:161–167
8. Bringmann G, Maksimenka K, Comar JM, Knauer M, Bruhn T (2007) Tetrahedron 63:9810–9824
9. Pecul M, Ruud K, Helgaker T (2004) Chem Phys Lett 388:110–119
10. Koch W, Holthausen MC (2000) A chemist's guide to density functional theory. Wiley-VCH, Weinheim
11. Peng HL, Payton JL, Protasiewicz JD, Simpson MC (2009) J Phys Chem A 113:7054–7063
12. Bouman T, Lightner DA (1976) J Am Chem Soc 98:3145–3154
13. Volosov AP, Zublov VA (1977) Theor Chim Acta 44:375–384
14. Komori H, Inai Y (2006) J Phys Chem A 110:9099–9107
15. Kawmura SI, Morita T, Kimura S (2006) Sci Technol Adv Mater 7:544–549
16. Bryan RF, Hartley P (1981) Mol Cryst Liq Cryst 69:47–69
17. Hohenberg P, Kohn W (1965) Phys Rev 136:B864–B871
18. Kohn W, Sham LJ (1965) Phys Rev 140:A1133–A1188
19. Jones RO, Gunnarsson O (1989) Rev Mod Phys 61:689–746
20. Botek E, Champagne B (2007) J Chem Phys 127:204101–201110
21. Danel KS, Gasiorski P, Matusiewicz M, Calus S, Uchacz T, Kityk AV (2010) Spectro Chim Acta 77:16–23
22. Neese FA (2003) J Chem Phys 119:9428–9443
23. Tejerina B, Reimers J (2007) CNDO/INDO. doi:10254/nanohub-r3352.9
24. Lakshmi Praveen P, Ojha DP (2011) Mat Chem Phys 126:248–252
25. Eikelschulte F, Yakovenko SY, Paschek D, Geiger A (2007) Liq Cryst 27:1137–1146
26. Sakagami S, Koga T, Takase A (2001) Liq Cryst 28:1199–1202
27. Hanemann T, Bohm MC, Haase W, Wu ST (1992) Liq Cryst 11:917–927
28. Markovic Z, Manojlovic N, Zlatanovic S (2008) J Ser Soc Comput Mech 2:73–79
29. Quotschalla U, Hanemann T, Bohm MC, Haase W (1991) Mol Cryst Liq Cryst 207:103–116
30. Wawzonek S, Lyons P (1950) J Org Chem 15:593–599

J Mol Model (2012) 18:1523–1540
DOI 10.1007/s00894-011-1180-7

ORIGINAL PAPER

# Computer-assisted design for atenolol prodrugs for the use in aqueous formulations

**Rafik Karaman · Khuloud Dajani · Hussein Hallak**

Received: 10 May 2011 / Accepted: 4 July 2011 / Published online: 23 July 2011


**Abstract** Based on stability studies on the drugs atenolol and propranolol and some of their derivatives it is believed that increasing the lipophilicity of the drug will lead to an increase in the stability of its aqueous solutions and will provide a prodrug system with the potential for releasing atenolol in a controlled manner. Using DFT theoretical calculations we have calculated an intramolecular acid catalyzed hydrolysis in nine maleamic (4-amino-4-oxo-2butenoic) acids (Kirby's N-alkylmaleamic acids), **1**–**9**. The DFT calculations confirmed that the acid-catalyzed hydrolysis mechanism in these systems involves: (1) a proton transfer from the hydroxyl of the carboxyl group to the adjacent amide carbonyl carbon, (2) an approach of the carboxylate anion toward the protonated amide carbonyl carbon to form a tetrahedral intermediate; and (3) a collapse of the tetrahedral intermediate into products. Furthermore, DFT calculations in different media revealed that the reaction rate-limiting step depends on the reaction medium. In aqueous medium the rate-limiting step is the collapse of the tetrahedral intermediate whereas in the gas phase the formation of the tetrahedral intermediate is the rate-limiting step. Furthermore, the calculations establish that the acid-catalyzed hydrolysis efficiency is largely sensitive to the pattern of substitution on the carbon-carbon double bond. Based on the experimental $t_{1/2}$ (the time needed for the conversion of 50% of the reactants to products) and EM (effective molarity) values for processes **1**–**9** we have calculated the $t_{1/2}$ values for the conversion of the two prodrugs to the parental drug, atenolol. The calculated $t_{1/2}$ values for **ProD 1**–**2** are predicted to be 65.3 hours and 11.8 minutes, respectively. Thus, the rate by which atenolol prodrug undergoes cleavage to release atenolol can be determined according to the nature of the linker of the prodrug (Kirby's N-alkylmaleamic acids **1**–**9**).

**Keywords** Amide hydrolysis · Anti-hypertensive agents · Atenolol prodrugs · DFT calculations · Enzyme catalysis · Intramolecular acid-catalyzed hydrolysis · Maleamic acid amide derivatives · Strain effects

## Introduction

Atenolol [4-(2-hydroxy-isopropylaminopropoxy)-phenylacetamide], is a cardioselective beta1-adrenergic receptor blocking agent used to lower high blood pressure (hypertension), prevent angina pectoris, treat arrhythmia and to reduce the risk of heart complications following a heart attack. Atenolol works by affecting the response to nerve impulses in certain parts of the body, like the heart. As a result, a slower heart beat and lower blood pressure can be expected. When the blood pressure is lowered, a decrease in the heart oxygen demand is observed [1–8].

Atenolol is available as 25, 50 and 100 mg tablets for oral administration. However, most of these medicines are not formulated for easy or accurate administration to children for the migraine indication or in elderly patients who may have difficulty swallowing tablets. Attempts to prepare a liquid formulation was challenging because

**Electronic supplementary material** The online version of this article (doi:10.1007/s00894-011-1180-7) contains supplementary material, which is available to authorized users.

R. Karaman (✉) · H. Hallak
Faculty of Pharmacy, Al-Quds University,
Box 20002, Jerusalem, Palestine
e-mail: dr_karaman@yahoo.com

K. Dajani
Faculty of Public Health Sciences, Al-Quds University,
Box 20002, Jerusalem, Palestine

atenolol is unstable in solutions. Studies showed that the degradation rate of atenolol is dependent on the temperature, indicating higher stability at 4 °C. Atenolol syrup is stable only for 9 days. Furthermore, oral doses of atenolol are incompletely absorbed (range 46-62%), even when formulated as a solution. Thus the development of more lipophilic prodrug that is stable in aqueous medium is a significant challenge [1–8].

Pharmacologically inactive chemical derivative that could be used to alter the physicochemical properties of atenolol, in a temporary manner and to increase its usefulness should be lipophilic linker that is covalently linked to the parent drug and can be converted in vivo to the active drug molecule, enzymatically or nonenzymatically, to exert a therapeutic effect. Ideally, the prodrug should be converted to the original drug as soon as the goal is achieved, followed by the subsequent rapid elimination of the released derivatizing group [9–11].

Improvement of atenolol pharmacokinetic absorption properties and hence its effectiveness may increase the absorption of the drug via a variety of administration routes.

To achieve this goal, atenolol prodrugs must possess the following properties: (i) to be readily soluble in a physiological environment (ii) to have a moderate hydrophilic lipophilic balance (HLB) value (iii) to provide upon chemical cleavage the active drug in a controlled manner, and (iv) to furnish upon cleavage a safe and non-toxic by-products. By complying with the four requirements described above the following objectives may be fulfilled: (1) a relatively high absorption of the prodrug into the body tissues. (2) The capability to use the drug in different dosage forms, especially in oral solutions form. (3) A chemically driven sustained release system that releases atenolol in a controlled manner once the prodrug reaches the human blood circulation system; and (4) a drug with physico-chemical properties such that leading to a high bioavailability and efficient pharmacokinetic properties.

Since atenolol is a very hydrophilic molecule with a pKa of 9.6, it undergoes ionization in the stomach and intestine thus its oral bioavailability is low due to inefficient absorption through membranes. Approximately 50% of an oral dose is absorbed from the gastrointestinal tract, the remainder being excreted unchanged in the feces. Unlike propranolol or metoprolol, but like nadolol, atenolol undergoes little or no metabolism by the liver, and the absorbed portion is eliminated primarily by renal excretion. Note the log P (partition coefficient) for atenolol is 0.5. Hence, design and synthesis of a relatively more lipophilic prodrugs that can provide the parental drug in a sustained release manner might result in a better clinical outcome, more convenient dosing regimens and potentially less side effects [1–8].

Recently we have been studying the mechanisms for a number of intramolecular processes that have been utilized as enzyme models and lately as prodrug linkers [12–30]. Using different molecular orbital methods such as DFT and ab initio at various levels, we have researched: (a) $S_N2$-based-cyclization reactions of di-carboxylic semi-esters as studied by Bruice [31, 32] (b) acid-catalyzed lactonization of hydroxy-acids as investigated by Cohen [33–35] and Menger [36–43], (c) intramolecular $S_N2$-based cyclization reactions as explored by Brown's group [44] and Mandolini's group [45], (d) proton transfer between two oxygens in Kirby's acetals [46–55], and proton transfer between nitrogen and oxygen in Kirby's enzyme models [46–55], (e) proton transfer between two oxygens in rigid systems as investigated by Menger [36–43], and (f) proton transfer from oxygen to carbon in some of Kirby's enol ethers [46–55]. These studies revealed the following: (1) rate enhancement in intramolecular processes is due to both entropy and enthalpy effects. In the cases by which enthalpic effects were predominant such as in cyclization reactions, steric effects was the cause for the accelerations, whereas proximity orientation was the dominant factor in the cases of proton transfer reactions. (2) The reaction type intermolecular or intramolecular is determined on the distance between the two reacting centers. In the cases by which the distance between the two reactive centers exceeds 3 Å an intermolecular reaction is preferred due to the feasibility of engagement of a water molecule (solvent) whereas, an intramolecular process is preferable when the distance between the electrophile and nucleophile is less than 3 Å. (3) The *gem*-dialkyl effect in $S_N2$-based ring-closing reactions leading to three-, four- and five-membered rings is dominant in those involving the formation of five-membered ring, and the need for directional flexibility decreases as the size of the ring being formed increases. (4) The efficiency of proton transfer between two oxygens and between nitrogen and oxygen in Kirby's acetal systems is due to a strong hydrogen bonding developed in the products and the corresponding transition states leading to them [12–33].

We have concluded from our recent studies on intramolecularity that there is a necessity to explore the reaction mechanism in order to assign the factors affecting the reaction rate for a design of an efficient chemical device to be used as prodrug linker capable of librating the parental drug in a controlled manner [12–33]. The prodrugs are designed such that they have the potential to undergo cleavage reactions in physiological environments (stomach at pH 1.5, intestine at pH 6.5 or/and blood circulation at pH 7.4) with rates that are dependent on the structural features of the pharmacologically inactive linker. Different linkers could be used to obtain a number of prodrugs with the capability to release the parental drug in

various rates that are depending on the nature or the structural features of the linker.

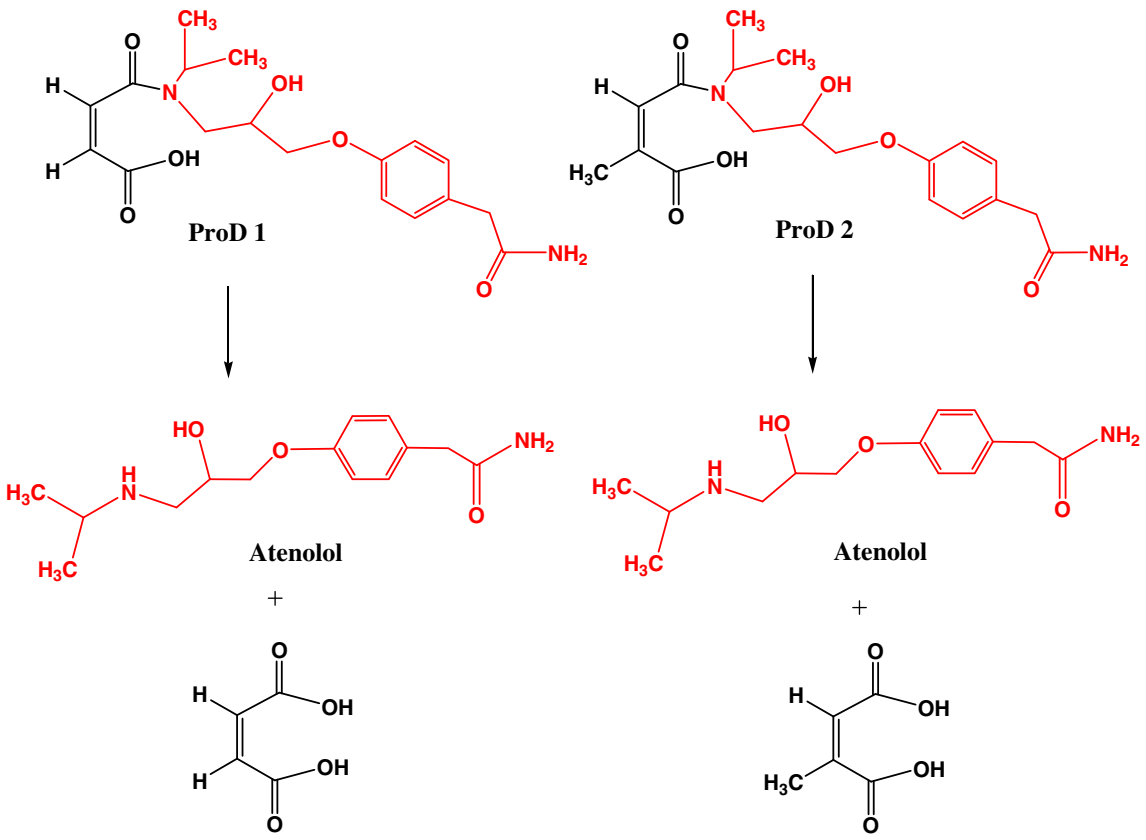
Continuing our study for utilizing intramolecularity to design potential linkers for amine drugs, we sought to study the mechanism and driving forces affecting the acid-catalyzed hydrolysis rate in some of Kirby's acid amides (prodrugs linkers) [56, 57]. It is expected that such linkers might have a potential to be good carriers to the anti-hypertensive agent, atenolol.

Our proposed atenolol prodrugs based on acid-catalyzed cyclization reactions of amide acids are depicted in Scheme 1.

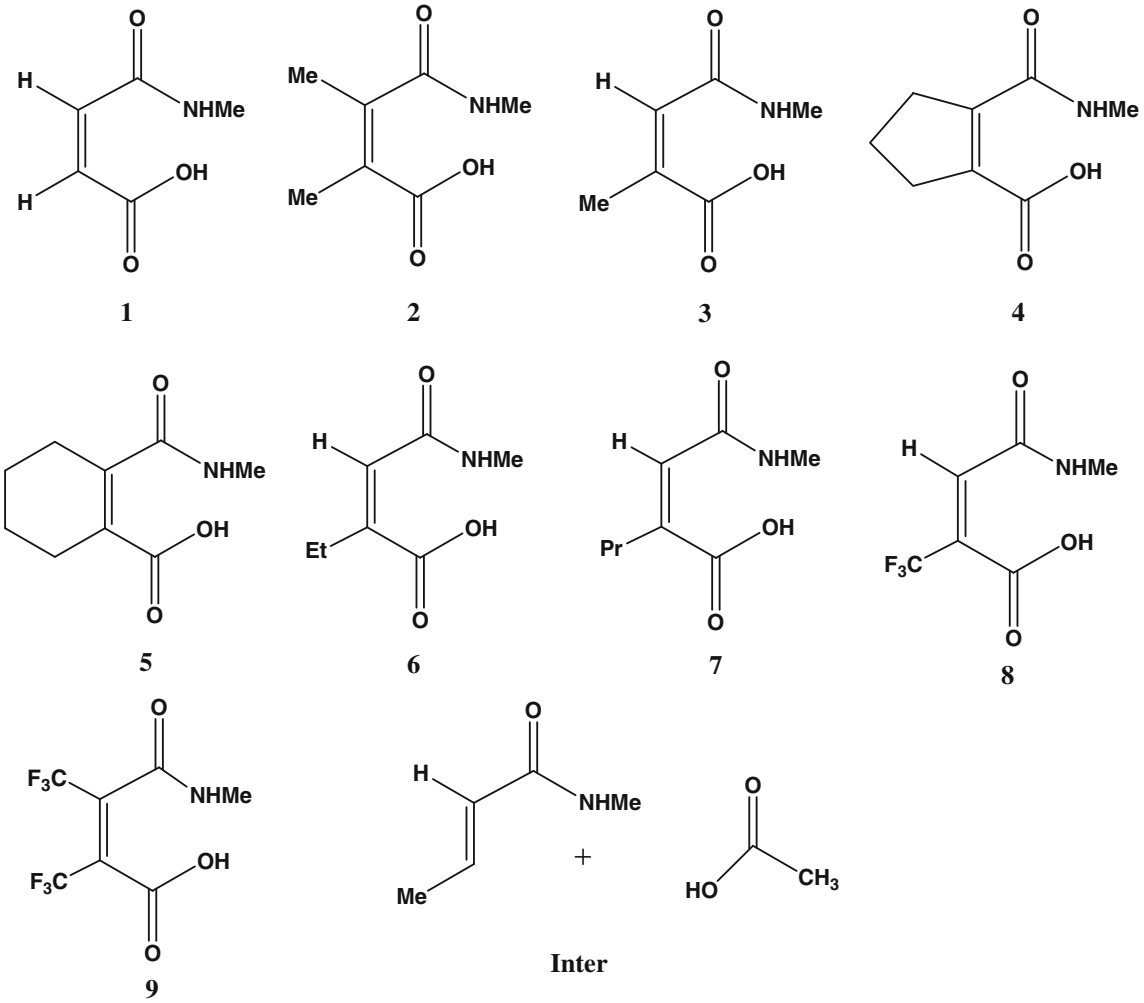
As shown in Scheme 1, the only difference between atenolol prodrugs and the parental drug, atenolol, is that the former has an amide moiety instead of the free amine group in the latter. By doing so the alcohol derivative thus formed (prodrug) will be much more stable than the corresponding amino alcohol, atenolol, due to general chemical stability for tertiary alcohols when compared to amine alcohols. Further, stability studies on ester derivatives of atenolol have showed that the ester derivatives are much more stable than the corresponding alcohol, atenolol, when they are formulated in aqueous solutions. In addition, kinetic studies on the drugs atenolol and propranolol revealed that increasing the lipophilicity of the drug leads to an increase in the stability of its aqueous solutions. Based on the above it is expected that atenolol prodrugs shown in Scheme 1 will be more resistant to heat or/oxidation when standing in aqueous solutions [1–8].

Based on the DFT calculations on the acid-catalyzed hydrolysis of amide acids **1**–**9** reported herein (Scheme 2), two atenolol prodrugs are proposed. As shown in Scheme 1, the atenolol prodrugs, **ProD 1** and **ProD 2**, have a carboxylic acid group (hydrophilic moiety) and a lipophilic moiety (the rest of the molecule), where the combination of both groups ensures a moderate HLB. It should be noted that the HLB value of the prodrug moiety will be determined upon the physiologic environment by which the prodrug is dissolved. For example, in the stomach, the atenolol prodrugs will primarily exist in the carboxylic acid form whereas in the blood circulation the carboxylate anion form will be predominant. It is planned that **ProD 1–2** (Scheme 1) will be obtained as sodium or potassium carboxylate salts since this form is expected to be stable in aqueous medium.

In this manuscript, we report a computational study which revealed the mechanism for intramolecular acid catalyzed hydrolysis of maleamic (4-amino-4-oxo-2butenoic) acids (Kirby's N-alkylmaleamic acids) **1**–**9** (Scheme 2). In addition, based on the rates for processes **1**–**9**, the calculated $t_{1/2}$ values for **ProD 1–2** were predicted.




**Scheme 1** Chemical structures for atenolol prodrugs **ProD1-2**




**Scheme 2** Chemical structures for maleamic acid derivatives 1–9 and Inter

The goals of this work were to: (1) calculate all possible mechanistic routes for the reactions of **1**–**9** and to unravel the nature of the force(s) affecting the rate as a function of substitution on the carbon-carbon double bond, (b) design of atenolol prodrugs that are relatively stable in aqueous solutions and have the capability to undergo cleavage reaction in physiological environment to furnish atenolol in a controlled manner.

## Calculation methods

The Becke three-parameter, hybrid functional combined with the Lee, Yang, and Parr correlation functional, denoted B3LYP, were employed in the calculations using density functional theory (DFT). Calculations were carried out based on the restricted Hartree-Fock method and were run using the quantum chemical package Gaussian-98 [58]. The starting geometries of all the molecules were obtained using the Argus Lab program [59] and were initially optimized at the HF/6-31 G level in the presence of one molecule of water, followed by optimization at the B3LYP/6-31 G(d,p) level [58]. Total geometry optimizations included all internal rotations. Second derivatives were estimated for all three N-6 geometrical parameters during optimization. The search for the global minimum structure in each of the systems studied was accomplished by 360° rotation of the carboxylic group about the bond C6-C7 (i.e., variation of the dihedral angle O1C7C6C5, Chart 1), and 360° rotation of the carbonyl amide group about the bond C4-C5 (i.e., variation of the dihedral angle O3C4C5C6) in increments of 10° and calculation of the conformational energies (see Chart 1). In the DFT calculations for **1**–**9** and **ProD 1**–**2**, two types of conformations in particular were considered: one in which the amide carbonyl is *syn* to the carboxyl group and another in which it is *anti*. An energy minimum (a stable compound or a reactive intermediate) has no negative vibrational force constant. A transition state is a saddle point which has only one negative vibrational force constant [60]. Transition states were located first by the normal reaction coordinate method [61] where the enthalpy

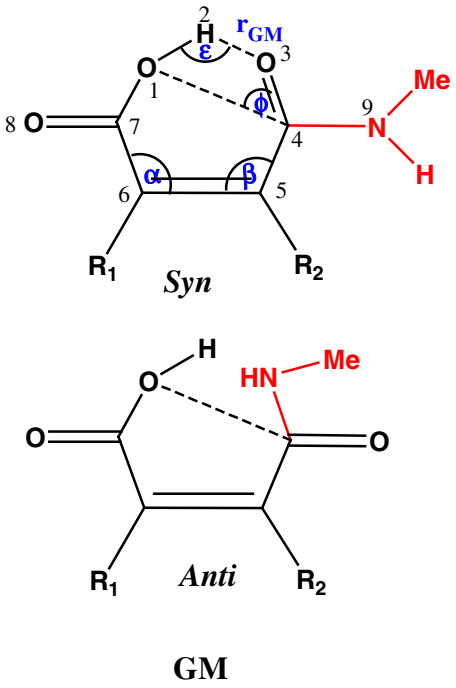



**Chart 1** Representation of an acid-catalyzed hydrolysis in Kirby's acid amides **1**–**9**. GM is the global minimum structure. $r_{GM}$ and $r_{AP}$ are the hydrogen bond distance and the approach distance, respectively. α, β, ε and ρ are bond angles

changes was monitored by stepwise changing the interatomic distance between two specific atoms. The geometry at the highest point on the energy profile was re-optimized by using the energy gradient method at the B3LYP/6-31 G (d,p) level of theory [58]. The "reaction coordinate method" [61] was used to calculate the activation energy in maleamic (4-amino-4-oxo-2butenoic) acids (Kirby's N-alkylmaleamic acids), **1**–**9**, **ProD 1–2** and **Inter** (Schemes 1 and 2). In this method, one bond length is constrained for the appropriate degree of freedom while all other variables are freely optimized. The activation energy values for the approach processes (the approach of O1 toward C4, Chart 1) were calculated from the difference in energies of the global minimum structures (GM) and the derived transition states ($TS_2$ in Scheme 3). Similarly, the activation energies of the dissociation processes (the breakdown of C4-N9 bond, Chart 1) were calculated from the difference in energies of the global minimum structures (GM) and the corresponding transition states ($TS_4$ in Scheme 3). Verification of the desired reactants and products was accomplished using the "intrinsic coordinate method" [61]. The transition state structures were verified by their one negative frequency. Full optimization of the transition states was accomplished after removing any constrains imposed while executing the energy profile. The activation energies obtained from DFT at B3LYP/6-31 G (d,p) level of theory for **1**–**9**, **ProD 1–2** and **Inter** were calculated with and without the inclusion of solvent (water and ether). The calculations with the incorporation of a solvent were performed using the integral equation formalism model of the polarizable continuum model (PCM) [62–65].

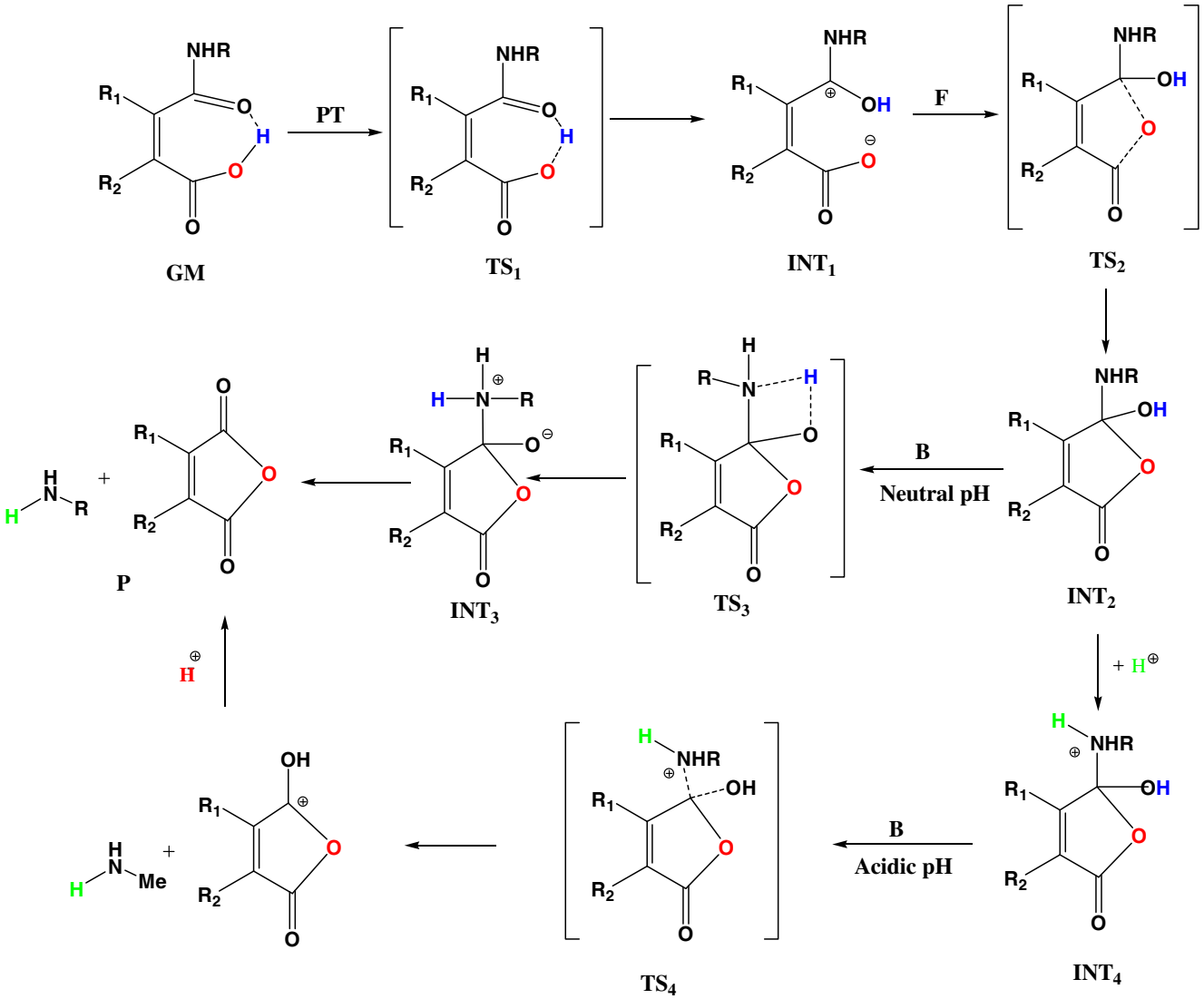
## Results and discussion

Kirby and coworkers have investigated the acid-catalyzed hydrolysis reactions of **1**–**9** (Scheme 2) and have concluded that the amide bond cleavage is due to intramolecular nucleophilic catalysis by the adjacent carboxylic acid group. Based on their kinetic results it was suggested that the rate-limiting step is the one by which dissociation of the tetrahedral intermediate occurs [56]. In 1990, Katagi has theoretically investigated the hydrolysis reaction using AM1 semiempirical method. Katagi's AM1 study revealed that the rate- limiting step is the formation of the tetrahedral intermediate and not its collapse [66]. Since Katagi's theoretical calculations were run in the presence of only two water molecules which resembles somewhat the gas phase environment we sought to study this reaction using Becke three-parameter hybrid functional combined with the Lee, Yang, and Parr correlation functional (B3LYP) methods, which are superior to semiempirical methods, and to run the calculations in the gas phase as well as in other solvents such as ether and a cluster of water due to the importance of the solvent in stabilizing or destabilizing the chemical entities involved in the hydrolysis reaction.

Replacing the N-methyl amide group in **1**–**9** (Scheme 2) with atenolol drug, as shown for **ProD 1** and **ProD 2** in Scheme 1, is not expected to have any significant effect on the relative rates of these processes. Therefore, computational calculations of the kinetic and thermodynamic properties for processes **1**–**9** will shed some light on the rates for the chemical cleavage of prodrugs **ProD 1** and **ProD 2** to the anti-hypertensive drug, atenolol.

In order to utilize systems **1**–**9** as prodrug carriers for atenolol as shown in the cases of **ProD 1**–**2** their mechanism should be unraveled. The suggested mechanism for the acid-catalyzed hydrolysis of **1**–**9** involves proton transfer from the carboxylic group to the adjacent amino carbonyl carbon followed by a nucleophilic attack of the carboxylate anion onto the protonated carbonyl carbon to form tetrahedral intermediate which dissociates to provide products (Scheme 3). In order to determine whether the rate-limiting step in the hydrolysis process is the formation of the tetrahedral intermediate as was suggested by Katagi [66] or its collapse as was proposed by Kirby [56] DFT calculations at the B3LYP/6-31 G (d,p) level were done in the gas phase as well as in ether and water.

Computations were directed toward elucidation of the transition and ground state structures (global minimum,




**Scheme 3** Possible mechanistic routes for the acid-catalyzed hydrolysis of 1–9 and ProD 1–2. R, R1 and R2 are alkyl groups. PT, F and B refer to proton transfer, tetrahedral intermediate formation and tetrahedral intermediate breakdown, respectively. GM, INT, TS and P refer to global minimum, intermediates, transition states and products, respectively

intermediates and products) for the acid-catalyzed hydrolysis of **1**–**9** and **ProD 1**–**2** in the presence of one molecule of water, a cluster of water, in the gas phase and in ether. It is expected that the stability of the chemical entities (GM, TS and etc.) will be different in the gas phase or solvent having low dielectric constant compared to that in water (a relatively high dielectric constant).

## General consideration

Because the energy of a carboxylic acid amide molecule is strongly dependent on its conformation and the latter determines its ability to be engaged in intramolecular hydrogen bonding, we were concerned with the identification of the most stable conformation (global minimum) for each of Kirby's acid amides **1**–**9** and prodrugs **ProD 1**–**2** calculated in this study. This was accomplished by 360° rotation of the carboxylic group about the bond C6-C7 (i.e., variation of the dihedral angle O1C7C6C5, Chart 1), and 360° rotation of the carbonyl amide group about the bond C4-C5 (i.e., variation of the dihedral angle O3C4C5C6) in increments of 10° and calculation of the conformational energies (see Chart 1).

In the DFT calculations for **1**–**9** and **ProD 1**–**2**, two types of conformations in particular were considered: one in which the amide carbonyl is *syn* to the carboxyl group and another in which it is *anti*. It was found that the global minimum structures for **1**–**9** and **ProD 1**–**2** all reside in the *syn* conformation (see Fig. 1a).

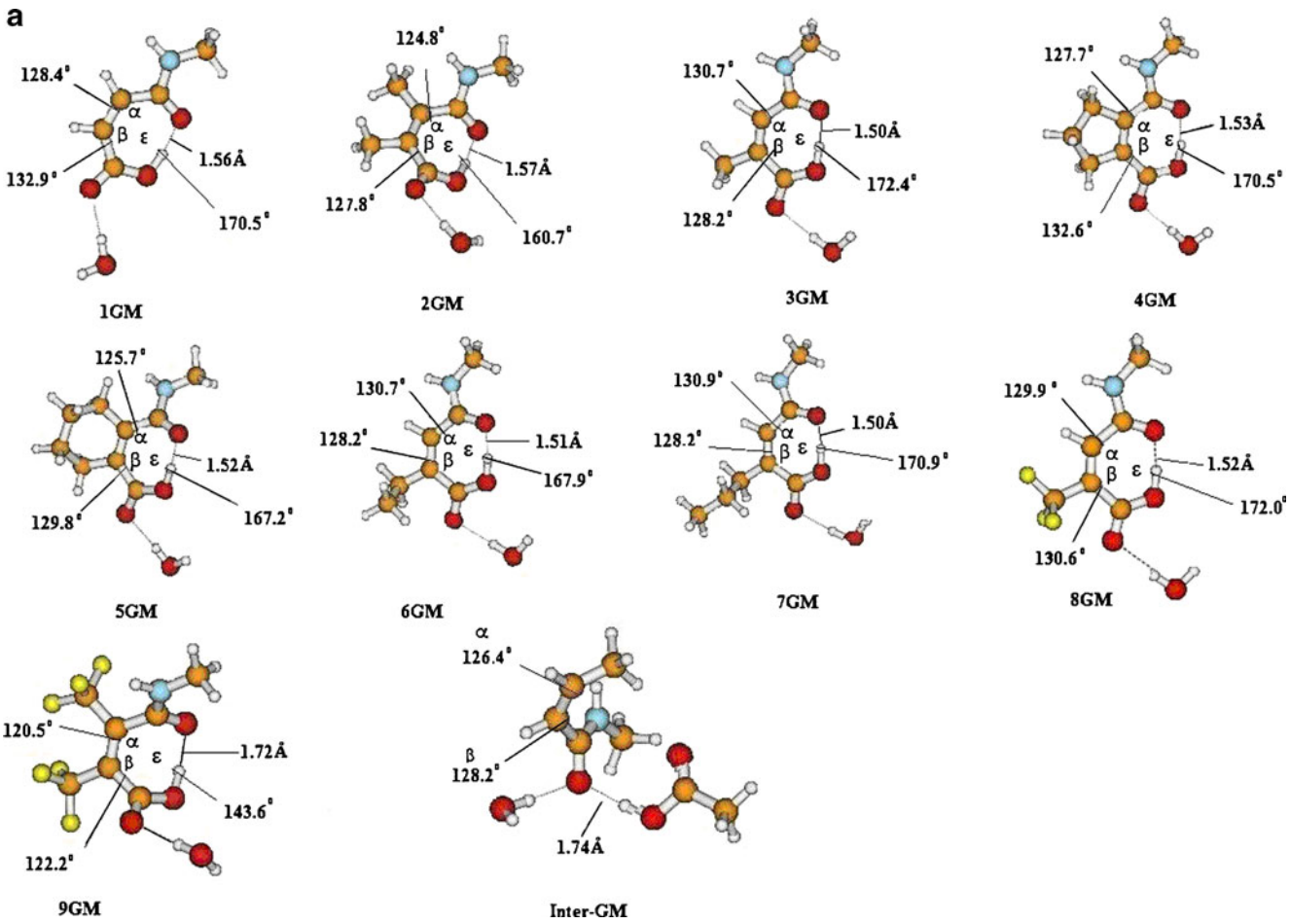



**Fig. 1** (**a**) DFT optimized structures for the global minimum (GM) in processes **1**–**9** and **Inter** (**b**) DFT optimized structures for the intermediate ($INT_2$) in processes **1**–**9** and **Inter** (**c**) DFT optimized structures for the transition state ($TS_4$) in processes **1**–**9** and **Inter**

Conformational analysis for the chemical entities involved in the acid-catalyzed hydrolysis of Kirby's acid amides 1–9 and ProD 1–2

*Starting geometries (GM)*

Since the acid-catalyzed hydrolysis reactions of **1**–**9** were conducted in aqueous medium, we have computed the geometries of the chemical entities involved in these reactions in the presence of one molecule of water in order to account for any specific intermolecular hydrogen bonding. The calculated B3LYP/6-31 G (d,p) geometries along with selected bond distances and bond angles for the global minimum structures of **1**–**9** (**1GM-9GM**) and **ProD 1GM-2GM** are illustrated in Figs. 1a and 2, and listed in Table 1. Examination of the calculated geometries of **1GM-9GM** (Fig. 1a and Table 1) and **ProD 1GM-2GM** (Fig. 2) indicates that all of them exhibit conformation by which the carboxyl group is engaged intramolecular in a hydrogen bond with the neighboring amide oxygen. This engagement results in the formation of seven-membered ring in all the global minimum structures (see Figs. 1a and 2). The calculated B3LYP/6-31 G (d,p) intramolecular hydrogen bonding length ($r_{GM}$ in Chart 1) in **1GM-9GM** and **ProD 1GM-2GM** was found in the range of 1.50 Å −1.87 Å and that for the hydrogen bond angle ε (the hydrogen bond angle, O1H2O3 in Chart 1) in the range of 143.6°-172.4°. Furthermore, the calculated DFT values for angles α and β were in the range of 120.5°- 130.7° and 122.2°- 133.8°, respectively. Inspection of Table 1 indicates that the calculated DFT values for the interatomic distance between the carboxyl hydroxyl oxygen and the amide carbonyl carbon range between 3.05 Å and 3.16 Å, while the angle of attack (C(O)OH/C(O)/C(O), angle θ in Chart 1) was found in the range 49.8°-56.8°.

*Tetrahedral intermediate geometries (INT2)*

The calculated DFT properties for the tetrahedral intermediate geometries of **1**–**9** (**1INT2-9INT2**) and **ProD 1$INT_2$-2$INT_2$** are shown in Figs. 1b and 2 and are tabulated in Table 1. Examination of the calculated structures for these

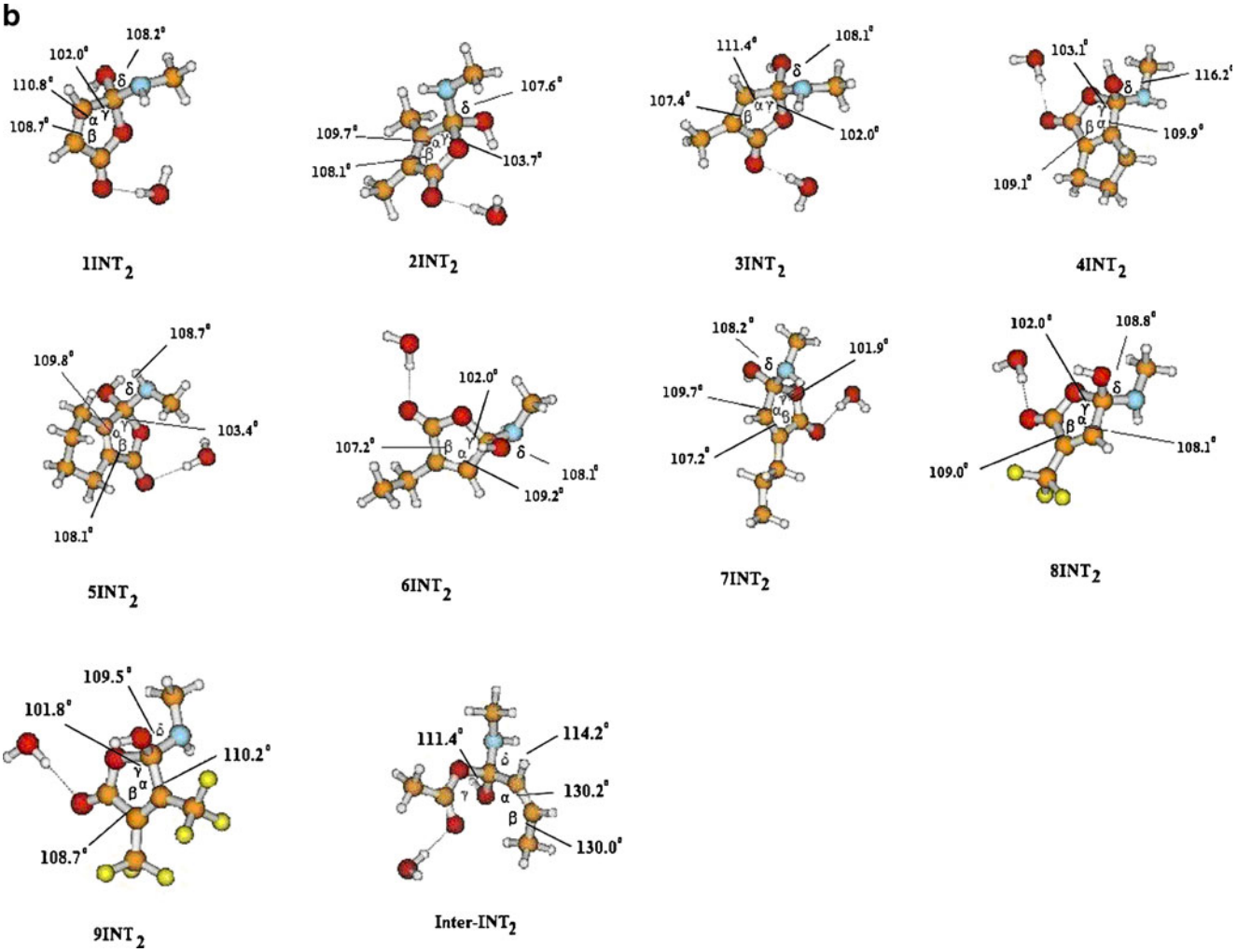



**Fig. 1** continued

intermediates revealed that angles α and β values in the intermediates are reduced when compared to that in the corresponding global minimum structures. The α values range was 107.2°-109.1° whereas for β values the range was 108.1°-111.4°. Furthermore, the angles γ and δ were in the range for similar regular tetrahedral intermediates. The range for γ was 102.0°- 103.7° and for δ was 107.6°- 116.2°.

*Transition state geometries for the tetrahedral intermediate dissociation (TS4)*

The calculated properties for the transition state geometries of the tetrahedral intermediate collapse in **1**–**9** and **ProD 1**–**2** (**1TS$_4$-9TS$_4$**, **ProD 1TS$_4$** and **ProD 2TS$_4$**) are summarized in Table 1 and illustrated in Figs. 1c and 2. Inspection of the optimized structures revealed that the geometries of the transition states resemble that of the corresponding tetrahedral intermediates. The angle α and β values are quite identical in both the tetrahedral intermediates and their corresponding transition states. The range of the α values were found in the range 107.0°-108.4° and that of β was in the range 108.3°-109.2°.

## Mechanistic investigation

Using the quantum chemical package Gaussian-98 [58] we calculated the DFT at B3LYP/6-31 G (d,p) level of theory kinetic and thermodynamic properties for all possible routes for the acid-catalyzed reactions of Kirby's acid amides **1**–**9** and **ProD 1**–**2**. The proposed routes are illustrated in Scheme 3.

The enthalpy and entropy energy values in the gas phase, ether and in the presence of a cluster of water for the global minimum structure (GM), the products (P) and the corresponding transition states, $TS_1$, $TS_2$, $TS_3$ and $TS_4$, and corresponding intermediates $INT_1$, $INT_2$, $INT_3$ and $INT_4$ for the suggested routes in the acid-catalyzed hydrolysis of **1**–**9** and **ProD 1**–**2** were calculated

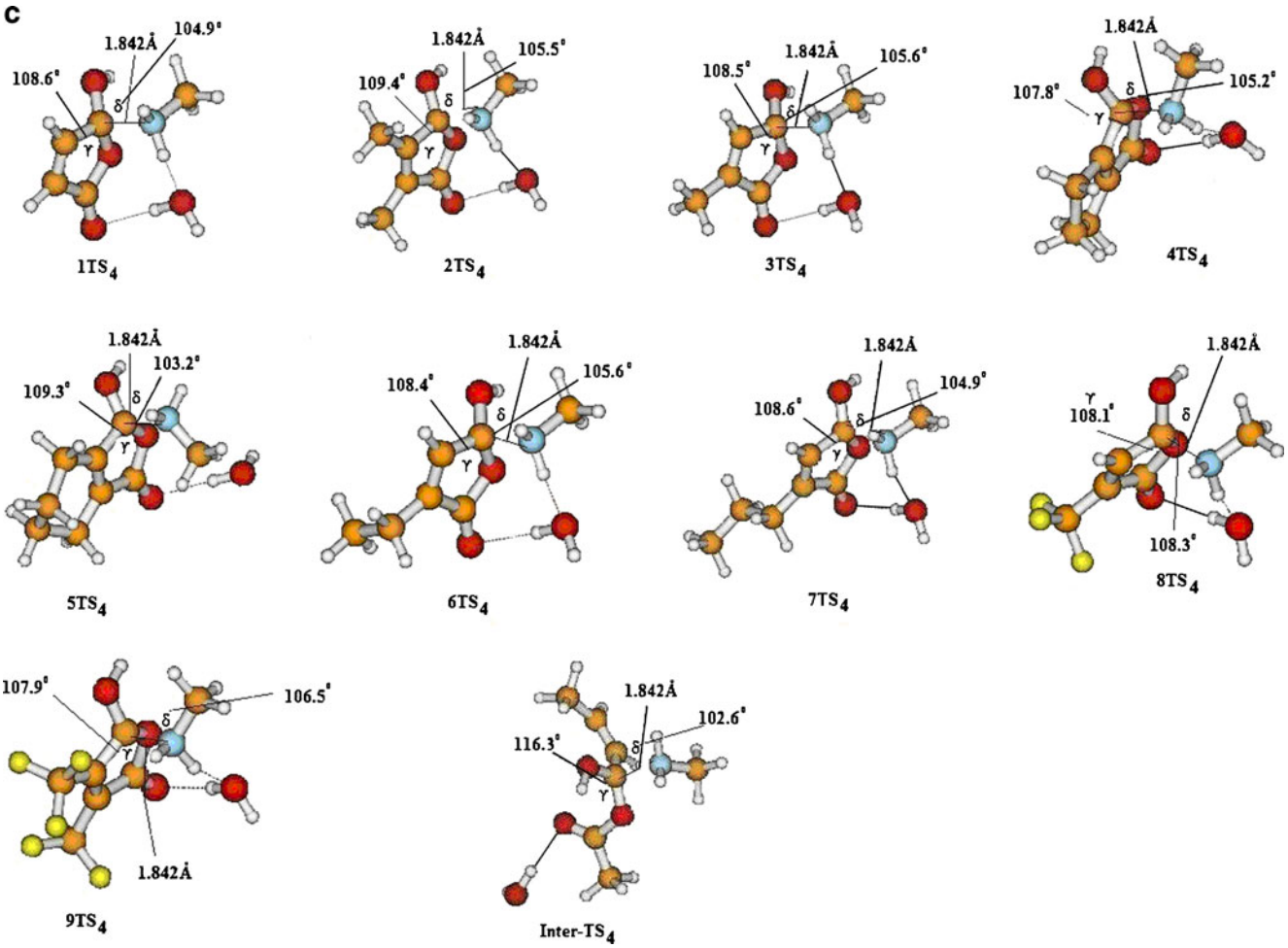



**Fig. 1** continued

(Scheme 3). Table 2 lists the energy values for **1GM-9GM**, **ProD 1GM**, **ProD 2GM**, **1INT$_2$-9INT$_2$**, **ProD 1INT$_2$**, **ProD 2INT$_2$**, **1TS$_4$-9TS$_4$**, **ProD 1TS$_4$** and **ProD 2TS$_4$**. Figs. 1a, b, c and 2 depict the DFT optimized structures for GM, INT$_2$ and TS$_4$ in **1**–**9** and **ProD 1**–**2**.

Using the calculated DFT values for the enthalpy and entropy of the GM and the four transition states in the hydrolysis of **1**–**9** and **ProD 1**–**2** the enthalpy activation energies ($\Delta H^{\ddagger}$), entropy activation energies ($T\Delta S^{\ddagger}$), and the free activation energies in the gas phase, ether and water ($\Delta G^{\ddagger}$) for the various routes in the mentioned processes (Scheme 3) were calculated and are summarized in Tables 3 and 4.

Inspection of the calculated B3LYP/6-31 G (d,p) activation energy values listed in Tables 3 and 4 revealed that the rate-limiting step for the acid-catalyzed hydrolysis of **1**–**9** and **ProD 1**–**2** is dependent on the medium by which the calculations were done. When the calculations were run in the gas phase the formation of the tetrahedral intermediate, step F in Scheme 3, was the rate limiting step, whereas when the calculations were executed in the presence of a cluster of water the step by which the tetrahedral intermediate breaks down was the rate-limiting (step B in Scheme 3). On the other hand, the energy calculations in the presence of ether indicate that while in some processes such as **2**, **5** and **ProD 2** the rate-limiting step is the one by which the carboxyl group approaches the amide carbonyl carbon to form the tetrahedral intermediate (step F in Scheme 3) the rate in processes **1**, **3**–**4**, **6**–**9** and **ProD 1** is determined on the step by which the tetrahedral intermediate collapses to yield products (step B in Scheme 3).

The gas phase energy barrier for the proton transfer process (step PT in Scheme 3), tetrahedral intermediate formation (F) and tetrahedral intermediate breakdown (step B in Scheme 3) were found in the range 17.77-19.33 kcal mol$^{-1}$, 25.94-45.37 kcal mol$^{-1}$ and 15.25-36.77 kcal mol$^{-1}$, respectively (Table 3). Furthermore, the gas phase free activation energy for the rate-limiting step (F) is about 5–15 kcal higher than that for the breakdown process. On the other hand, the activation energies difference as calculated in water between the

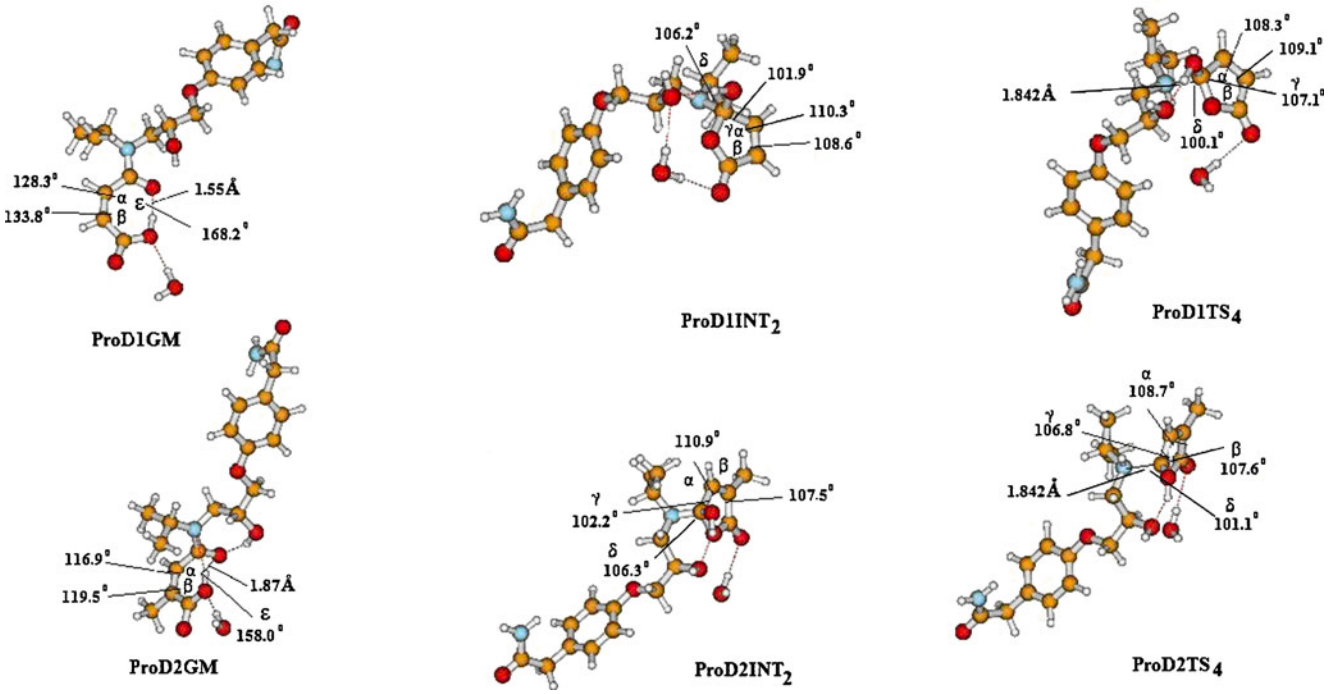



**Fig. 2** DFT optimized structures for the global minimum (GM), intermediate ($INT_2$) and transition state ($TS_4$) in processes **ProD 1–2**

intermediate collapse and intermediate formation, is about 2–9 kcal $mol^{-1}$ (Table 4).

*The factors affecting the hydrolysis rate*

Careful inspection of Tables 3 and 4 revealed that both the substituent on the double bond and the reaction medium has a profound effect on the reaction rate.

1. **The substituent effect**: to examine the structural factors associated with high reactivity such as in systems **2**, **5** and **ProD 2**, analysis of the conformation of the rigid structures of the ground states should be evaluated. A careful examination of Tables 1 and 4 revealed the existence of two significant trends in the reactivity versus the structural features of the ground state structures. The closest distance of approach of the carboxyl oxygen and the amide oxygen in the ground state ($r_{AP}$ in Chart 1) is less for more reactive compounds. For example, the distance for the unsubstituted maleamic acid amide **1** is 3.16 Å whereas for the dimethylsubstituted derivative **2** is 3.05 Å. In fact,

**Table 1** DFT (B3LYP) calculated geometries for GM, $INT_2$ and $TS_4$ in **1–9** and **ProD 1-2**

| System | α in GM | β in GM | α in $INT_2$ | β in $INT_2$ | γ in $INT_2$ | δ in $INT_2$ | α in $TS_4$ | β in $TS_4$ | γ in $TS_4$ | δ in $TS_4$ | θ in GM | $r_{AP}$ in GM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 128.4 | 132.9 | 108.7 | 110.8 | 102.0 | 108.2 | 107.6 | 108.5 | 108.6 | 104.9 | 50.3 | 3.16 |
| 2 | 124.8 | 127.8 | 108.1 | 109.7 | 103.7 | 107.6 | 107.0 | 108.3 | 109.4 | 105.5 | 53.0 | 3.05 |
| 3 | 130.7 | 128.2 | 107.4 | 111.4 | 102.0 | 108.1 | 107.3 | 108.7 | 108.5 | 105.6 | 51.0 | 3.10 |
| 4 | 127.7 | 132.6 | 109.1 | 109.9 | 103.1 | 116.2 | 108.2 | 109.8 | 107.8 | 105.2 | 49.8 | 3.15 |
| 5 | 125.7 | 129.8 | 108.1 | 109.8 | 103.4 | 108.7 | 107.2 | 108.8 | 109.3 | 103.2 | 51.6 | 3.09 |
| 6 | 130.7 | 128.2 | 107.2 | 109.2 | 102.0 | 108.1 | 107.1 | 108.8 | 108.4 | 105.6 | 51.4 | 3.09 |
| 7 | 130.9 | 128.2 | 107.2 | 109.7 | 101.9 | 108.2 | 107.2 | 108.8 | 108.6 | 104.9 | 51.1 | 3.09 |
| 8 | 129.9 | 130.6 | 109.0 | 108.1 | 102.0 | 108.8 | 107.6 | 109.2 | 108.1 | 108.3 | 51.0 | 3.11 |
| 9 | 120.5 | 122.2 | 108.7 | 110.2 | 101.8 | 109.5 | 107.6 | 108.6 | 107.9 | 106.5 | 56.8 | 3.05 |
| ProD 1 | 128.3 | 133.4 | 108.8 | 110.3 | 101.9 | 115.6 | 108.4 | 108.7 | 107.1 | 106.3 | 48.7 | 3.19 |
| ProD 2 | 126.2 | 129.6 | 107.5 | 110.9 | 102.2 | 115.2 | 107.6 | 108.7 | 106.8 | 103.0 | 55.1 | 3.32 |

B3LYP refers to values calculated by B3LYP/6-31 G (d, p) method. GM, $INT_2$ and $TS_4$ refer to global minimum, intermediate and transition state structures, respectively (see Scheme 3). For the definition of the bond angles α, β, γ and δ see Fig. 1a, b and c. For the definition of rAP see Chart 1.The angle unit is degree and the $r_{AP}$ unit is Å

**Table 2** DFT (B3LYP/6-31 G (d,p)) calculated properties for the acid catalyzed hydrolysis of acid amides **1–9**, **ProD 1–2** and **Inter**

| Compound | Enthalpy, H (gas phase) In Hartree | (gas phase) Entropy, S, Cal/Mol-Kelvin | Frequency $Cm^{-1}$ | MM2Es kcal/mol |
|---|---|---|---|---|
| 1GM | −551.6209552 | 115.71 | ——— | 10.16 |
| $1INT_2$ | −551.5899873 | 113.12 | ——— | 20.55 |
| $1TS_4$ | −551.9510979 | 100.58 | 29.18i | ——— |
| 2GM | −630.2556684 | 133.45 | ——— | 10.82 |
| $2INT_2$ | −630.2424070 | 125.08 | ——— | 16.16 |
| $2TS_4$ | −630.6125865 | 117.89 | 57.26i | ——— |
| 3GM | −590.9452364 | 126.16 | ——— | 9.40 |
| $3INT_2$ | −590.9162845 | 124.52 | ——— | 17.32 |
| $3TS_4$ | −591.284155 | 109.50 | 45.91i | ——— |
| 4GM | −668.3716758 | 131.90 | ——— | 12.30 |
| $4INT_2$ | −668.3272626 | 125.09 | ——— | 27.89 |
| $4TS_4$ | −668.7001822 | 116.95 | 73.72i | ——— |
| 5GM | −707.6805570 | 136.69 | ——— | 9.18 |
| $5INT_2$ | −707.6689275 | 122.73 | ——— | 19.25 |
| $5TS_4$ | −707.6321442 | 122.88 | 118.05i | ——— |
| 6GM | −630.2606708 | 132.74 | ——— | 5.12 |
| $6INT_2$ | −630.2326178 | 132.45 | ——— | 17.59 |
| $6TS_4$ | −630.60156591 | 116.93 | 52.77i | ——— |
| 7GM | −669.5780833 | 138.31 | ——— | 6.20 |
| $7INT_2$ | −669.5495044 | 137.75 | ——— | 18.55 |
| $7TS_4$ | −669.9185072 | 124.96 | 68.64i | ——— |
| 8GM | −888.6478787 | 140.60 | ——— | 12.86 |
| $8INT_2$ | −888.6201859 | 137.62 | ——— | 22.34 |
| $8TS_4$ | −888.9742696 | 120.84 | 37.75i | ——— |
| 9GM | −1225.6575172 | 145.62 | ——— | 28.29 |
| $9INT_2$ | −1225.6422023 | 147.77 | ——— | 26.92 |
| $9TS_4$ | −1225.9938255 | 140.05 | 33.82i | ——— |
| ProD 1GM | −1337.6974501 | 214.38 | ——— | 20.97 |
| ProD $1INT_2$ | −1337.6654432 | 188.26 | ——— | 41.97 |
| ProD $1TS_4$ | −1338.0399174 | 158.59 | 70.33i | ——— |
| ProD 2GM | −1377.0175828 | 210.68 | ——— | 17.31 |
| ProD $2INT_2$ | −1376.9882479 | 211.30 | ——— | 38.42 |
| ProD $2TS_4$ | −1377.3516015 | 193.30 | 123.35i | ——— |
| InterGM | −631.4579281 | 141.61 | ——— | 2.98 |
| $InterINT_2$ | −631.4282861 | 129.50 | ——— | 13.64 |
| $InterTS_4$ | −631.7933134 | 130.70 | 57.27i | ——— |

GM, INT and TS are global minimum, intermediate and transition state structures, respectively. MM2Es is the strain energy calculated by Allinger's MM2 method [67]

when the values of $r_{AP}$ were examined for correlation with the experimental and water calculated free activation energies ($\Delta G_B^{\ddagger}$) strong correlations with correlation coefficient of R=0.92 -0.96 were obtained (Fig. 3a). The second structural factor that might have a significant effect on the reactivity of an acid amide is the angle by which the carboxyl group approaches the amide carbonyl carbon (the attack angle ρ, see Chart 1). The calculated θ values for the reactive derivatives were found to be larger than those for the less reactive acid amides. For example the calculated θ value for system **2** was 53° whereas for **4** was 49.8°. Linear correlation with strong correlation coefficients (R= 0.92-0.96) were found between the DFT calculated ρ values and the experimental and water calculated free activation energy values (Fig. 3b) to shed some light on the factors responsible for the unusual accelerations in rate and to examine whether the discrepancy in rates for processes **1–9** stems from steric effects (strain energy) or other effects we have calculated, using Allinger's MM2 method [67], the strain energy values for the reactants (GM), intermediates (INT) and products (P) in **1–9** and **ProD 1–2**. The MM2 strain energies (Es) for these geometries are listed in Table 4. The MM2

**Table 3** DFT (B3LYP) calculated kinetic and thermodynamic properties in the gas phase for the acid catalyzed hydrolysis of **1**–**9**, **ProD 1**–**2** and **Inter**

| System | $\Delta H_{BGP}^{\ddagger}$ (kcal/mol) | $T\Delta S_{BGP}^{\ddagger}$ (kcal/mol) | $\Delta G_{BGP}^{\ddagger}$ (kcal/mol) | $\Delta H_{FGP}^{\ddagger}$ (kcal/mol) | $T\Delta S_{FGP}^{\ddagger}$ (kcal/mol) | $\Delta G_{FGP}^{\ddagger}$ (kcal/mol) | $\Delta G_{BEth}^{\ddagger}$ (kcal/mol) | $\Delta G_{FEth}^{\ddagger}$ (kcal/mol) | $\Delta G_{BGP}^{\ddagger}$ (kcal/mol) In high pH |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 27.31 | −0.77 | 28.08 | 32.46 | −1.09 | 33.53 | 34.27 | 31.35 | 45.55 |
| 2 | 13.93 | −2.49 | 16.42 | 25.67 | −1.41 | 27.08 | 16.32 | 21.29 | 33.59 |
| 3 | 24.41 | −0.49 | 24.90 | 30.68 | −1.89 | 32.57 | 29.53 | 28.23 | 44.68 |
| 4 | 34.42 | −2.35 | 36.77 | 41.88 | −3.49 | 45.37 | 40.23 | 39.27 | 53.5 |
| 5 | 13.25 | −4.16 | 17.41 | 24.55 | −2.32 | 26.87 | 18.51 | 22.34 | 34.52 |
| 6 | 23.83 | −0.09 | 23.92 | 30.11 | −2.01 | 32.12 | 29.53 | 27.50 | 45.72 |
| 7 | 24.86 | −0.17 | 25.03 | 30.76 | −1.54 | 32.30 | 28.35 | 27.76 | ——— |
| 8 | 24.08 | −0.89 | 24.87 | 29.79 | −2.58 | 32.37 | 29.76 | 28.20 | ——— |
| 9 | 17.88 | 0.64 | 17.24 | 24.17 | −1.77 | 25.94 | 24.46 | 18.69 | ——— |
| ProD 1 | 20.08 | −7.78 | 27.86 | 29.96 | −5.99 | 35.95 | 38.76 | 39.49 | ——— |
| ProD 2 | 15.43 | 0.18 | 15.25 | 31.94 | 1.19 | 30.75 | 21.15 | 24.35 | 43.53 |
| Inter | 39.90 | −3.35 | 43.25 | ——— | ——— | ——— | ——— | ——— | ——— |

B3LYP refers to values calculated by B3LYP/6-31 G (d, p) method. $\Delta H^{\ddagger}$ is the calculated activation enthalpic energy (kcal/mol). $T\Delta S^{\ddagger}$ is the calculated activation entropic energy (kcal/mol). $\Delta G^{\ddagger}$ is the calculated activation free energy (kcal/mol). B and F refer to tetrahedral intermediate breakdown and tetrahedral intermediate formation. GP and Eth refer to calculated in the gas phase and in ether, respectively

calculated Es values were examined for correlation with the experimental relative rate values, log $k_{rel}$ (Table 4). Strong correlations were obtained with a correlation coefficient R=0.94 for the correlation with the intermediates values and R=0.97 for the values of the products (Fig. 3c). Furthermore, good correlations were obtained between the experimental and calculated free activation energy values on one hand and the strain energy difference values between the intermediates and reactants ($Es_{INT-GM}$) on the other hand (R=0.85-0.94) (Fig. 3d). The results shown in Fig. 3c and d indicate that the rate of the reaction for systems having less-strained intermediates or products such as **2** and **5** are higher than that having more strained intermediates or products such as **1** and **4**. This might be attributed to the fact that the transition state structures in **1**–**9** resemble that of the corresponding intermediates or products (Figs. 2b and c).

In order to test whether the substituent on the double bond has the same effect on the tetrahedral intermediate

**Table 4** DFT (B3LYP) calculated kinetic and thermodynamic properties for the acid catalyzed hydrolysis of **1**–**9**, **ProD 1**–**2** and **Inter**

| System | $\Delta H_{BW}^{\ddagger}$ (kcal/mol) | $\Delta G_{BW}^{\ddagger}$ (kcal/mol) | log $k_{rel}$ [56] | log EM (Exp) [56] | log EM (Calc) | Es ($INT_2$) (kcal/mol) | Es (P) (kcal/mol) | Es (GM) (kcal/mol) | $\Delta G_{FW}^{\ddagger}$ (kcal/mol) | Exp $\Delta G^{\ddagger}$ [12](kcal/mol) |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 32.29 | 33.06 | 0 | 7.724 | 8.52 | 20.55 | 25.08 | 10.16 | 26.10 | 23.70 |
| 2 | 17.56 | 20.05 | 4.371 | 15.86 | 18.08 | 16.16 | 18.93 | 10.82 | 17.90 | 17.30 |
| 3 | 27.93 | 28.42 | 1.494 | 7.742 | 11.93 | 17.32 | 21.70 | 9.40 | 24.80 | 21.14 |
| 4 | 35.76 | 38.11 | −4.377 | 1.255 | 4.81 | 27.89 | 32.75 | 12.30 | 32.16 | 30.70 |
| 5 | 18.96 | 23.12 | 2.732 | 15.190 | 15.82 | 19.25 | 23.13 | 9.18 | 17.89 | 19.75 |
| 6 | 27.19 | 27.28 | 1.516 | 6.962 | 12.76 | 17.59 | 22.95 | 5.12 | 23.87 | ——— |
| 7 | 27.38 | 27.55 | 1.648 | 8.568 | 12.57 | 18.55 | 24.00 | 6.20 | 24.40 | ——— |
| 8 | 29.23 | 30.12 | ——— | ——— | 6.36 | 22.34 | 27.77 | 12.86 | 23.66 | |
| 9 | 15.79 | 15.15 | ——— | ——— | 21.68 | 26.92 | 35.64 | 28.29 | 11.97 | |
| ProD 1 | 31.37 | 39.12 | ——— | ——— | 4.06 | 41.97 | 25.08 | 20.97 | 35.76 | |
| ProD 2 | 26.98 | 26.8 | ——— | ——— | 13.11 | 38.42 | 18.98 | 17.31 | 17.85 | |
| Inter | 41.30 | 44.65 | ——— | ——— | ——— | 13.64 | 12.10 | ——— | ——— | ——— |

B3LYP refers to values calculated by B3LYP/6-31 G (d, p) method. $\Delta H^{\ddagger}$ is the calculated activation enthalpic energy (kcal/mol). $T\Delta S^{\ddagger}$ is the calculated activation entropic energy (kcal/mol). $\Delta G^{\ddagger}$ is the calculated activation free energy (kcal/mol). Es refers to strain energy calculated by Allinger's MM2 method [67]. $INT_2$ and P refer to intermediate two and product, respectively. $EM=e^{-(\Delta G^{\ddagger}inter - \Delta G^{\ddagger}intra)/RT}$. BW and FW refer to tetrahedral intermediate breakdown and tetrahedral intermediate formation calculated in water, respectively. Exp refers to experimental value. Calc refers to DFT calculated values

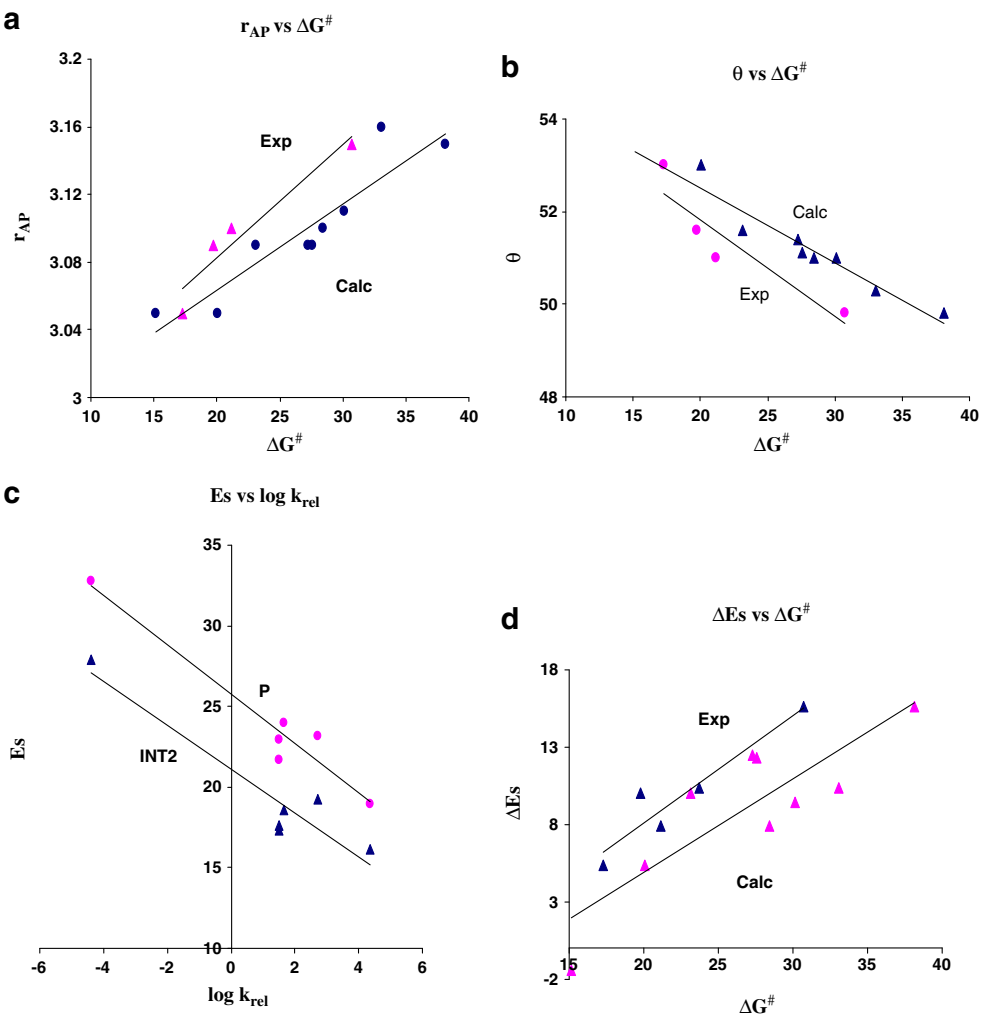



**Fig. 3** (**a**) Plot of $r_{AP}$ vs. $\Delta G^{\ddagger}$ in **1**–**9**, where the blue points for the DFT calculated values (in water) and the pink points for the experimental values. $r_{AP}$ is the distance between the hydroxyl oxygen of the carboxylic group and the carbon of the amide carbonyl in the reactant. $\Delta G^{\ddagger}$ is the free activation energy. (**b**) Plot of ρ vs. $\Delta G^{\ddagger}$ in **1**–**9**, where the blue points for the DFT calculated values (in water) and the pink points for the experimental values. ρ is the angle of attach of the approach of the hydroxyl of the carboxylic group toward the carbon of the amide carbonyl in the reactant. $\Delta G^{\ddagger}$ is the free activation energy. (**c**) Plot of Es vs. log $k_{rel}$ in **1**–**9**, where the blue points for the intermediates (INT2, see Scheme 3) and the pink points for the products (P, see Scheme 3). log $k_{rel}$ is the relative rate. (**d**) Plot of ΔEs vs. $\Delta G^{\ddagger}$ in **1**–**9**, where the blue points for the experimental values and the pink points for the DFT calculated values (in water). $\Delta G^{\ddagger}$ is the free activation energy

formation (step F in Scheme 3) and the tetrahedral intermediate collapse (B in Scheme 3), the DFT calculated activation energies in the gas phase and in water for the formation of the tetrahedral intermediate ($\Delta G_F^{\ddagger}$) were correlated with the energy values needed for its collapse ($\Delta G_B^{\ddagger}$). The correlation results illustrated in Fig. 4a indicate strong correlations with a correlation coefficient R=0.92 and 0.95, respectively. This indicates that the driving force for the formation and collapse of the tetrahedral intermediate is the same and the rate of the hydrolysis reaction is dependent on strain effects.

The combined results might be able to test some theories that were invoked to account for rate accelerations in intramolecular and enzymatic processes. Such theories are the orbital steering theory of Koshland [68] that describes the effect of critical angle of approach on reactivity, and Bruice's [31, 32] and Menger's proximity orientation [36–43].

2. **The solvent effect**: Inspection of Scheme 3 revealed that the step in the intramolecular hydrolysis reaction most powerfully affected by structural changes and solvent effects are the ring-closure of $INT_1$ which is involved in the tetrahedral intermediate formation. Examination of the geometries of the corresponding transition states for $INT_2$ ($TS_2$) indicates that $TS_2$ is in the ionic form. It is expected that the effect of polar solvents on the stability of $TS_2$ will be different from those with less polarity. The free activation energies for the approach process ($\Delta G_F^{\ddagger}$) and the dissociation process ($\Delta G_B^{\ddagger}$) calculated in the gas phase, ether and in presence of a cluster of water listed in Tables 3 and 4 indicate that solvents with low dielectric constant such in the case of the gas phase tend to shift the equilibrium to the side of the reactants by destabilizing $TS_2$ and thus making the approach step as the rate-limiting whereas solvents with high dielectric constants such as water interact strongly with the ionic transition state. These interactions stabilize the approach transition state ($TS_2$) and consequently lower the approach barrier, hence making the reaction rate dependent on the dissociation step. A representation of the energy profiles

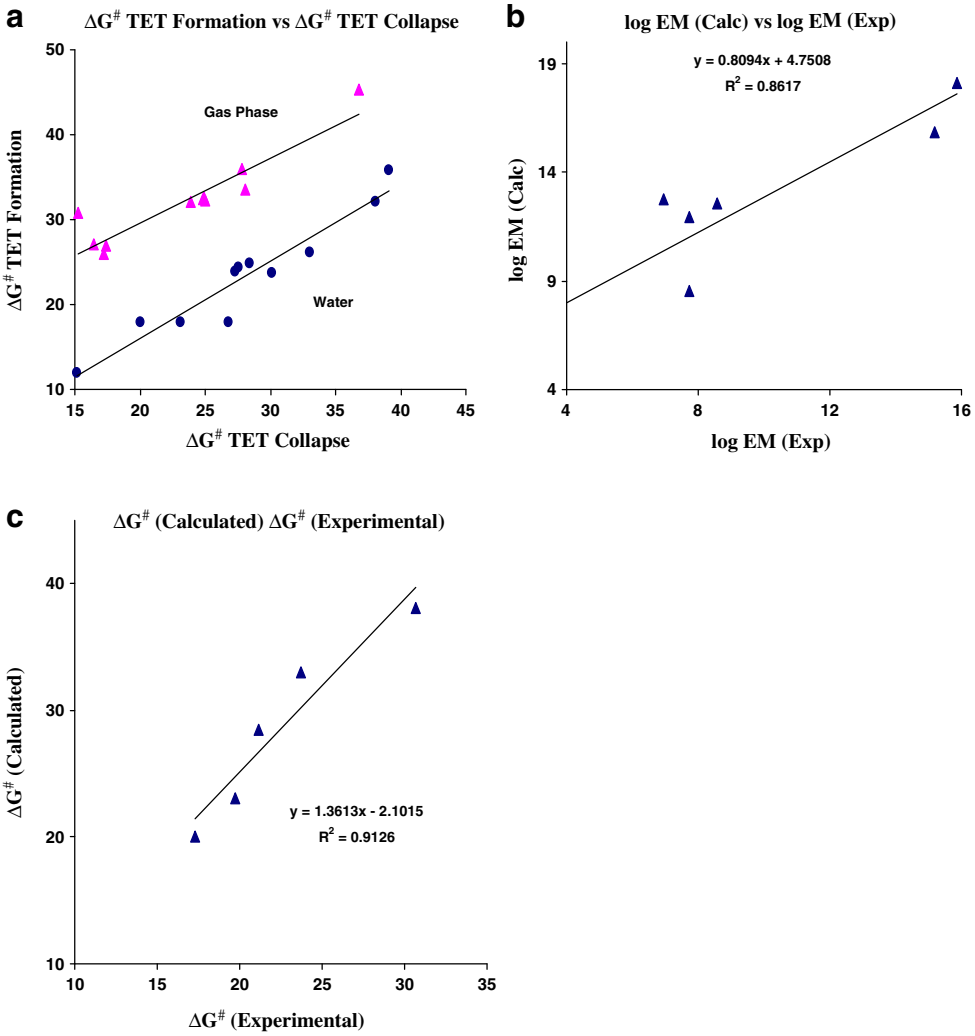



**Fig. 4** (**a**) Plot of $\Delta G^{\ddagger}$ (Formation) vs. $\Delta G^{\ddagger}$ (Collapse) in **1**–**9**, where the blue points for the DFT calculated values (in water) and the pink points for that calculated in the gas phase. $\Delta G^{\ddagger}$ (Formation) and $\Delta G^{\ddagger}$ (Collapse) are the free activation energy needed for the tetrahedral intermediate formation and collapse, respectively. (**b**) log EM (Calc) vs. log EM (Exp) in **1**–**9**. EM refers to effective molarity. Calc and Exp refer to calculated by DFT and experimental, respectively. (**c**) Plot of $\Delta G^{\ddagger}$ (Calculated) vs. $\Delta G^{\ddagger}$ (Experimental) in **1**–**9**. $\Delta G^{\ddagger}$ (Calculated) and $\Delta G^{\ddagger}$ (Experimental) are the DFT calculated (in water) free activation energy and experimental free activation energy, respectively

for the hydrolysis of **ProD 1** as calculated in the gas phase and in presence of a cluster of water is illustrated in Fig. 5.

On the other hand, the DFT calculations in presence of ether having dielectric constant of 4.5 indicate that in systems **2, 5** and **ProD 1–2** the rate-limiting step is the formation of the tetrahedral intermediate whereas in the other processes the rate-limiting step is the dissociation of the tetrahedral intermediate (see Tables 3 and 4).

## Calculations of the effective molarity (EM) values for processes 1–9 and ProD 1–2

It is widely accepted that the measure for intramolecular efficiency is the effective molarity (EM). The effective molarity is defined as a ratio of the intramolecular rate and its corresponding intermolecular where both processes are driven by identical mechanisms. Ring size, solvent and reaction type are the main factors affecting the EM parameter. Ring-closing reactions via intramolecular nucleophilic addition are much more efficient than intramolecular proton transfers. Values in the order of $10^{9}$-$10^{13}$ M have been measured for the EM in intramolecular processes occurring through nucleophilic addition. Whereas for proton transfer processes EM values of less than 10 M were obtained [69].

To obtain credibility for our calculation results we introduce our computation rational for calculating the EM values for processes **1**–**9** and **ProD 1–2** based on the DFT calculated activation energies ($\Delta G^{\ddagger}$) of **1**–**9** and **ProD 1–2**, and the corresponding intermolecular process **Inter** (Scheme 2).

The intermolecular process **Inter** (Scheme 2) was calculated to be used in the calculation of the effective molarity values (EM) for the corresponding intramolecular processes **1**–**9** and **ProD 1–2**.

Utilizing Eqs. 1–4, we derived Eq. 5 which describes the EM parameter as a function of the difference in the

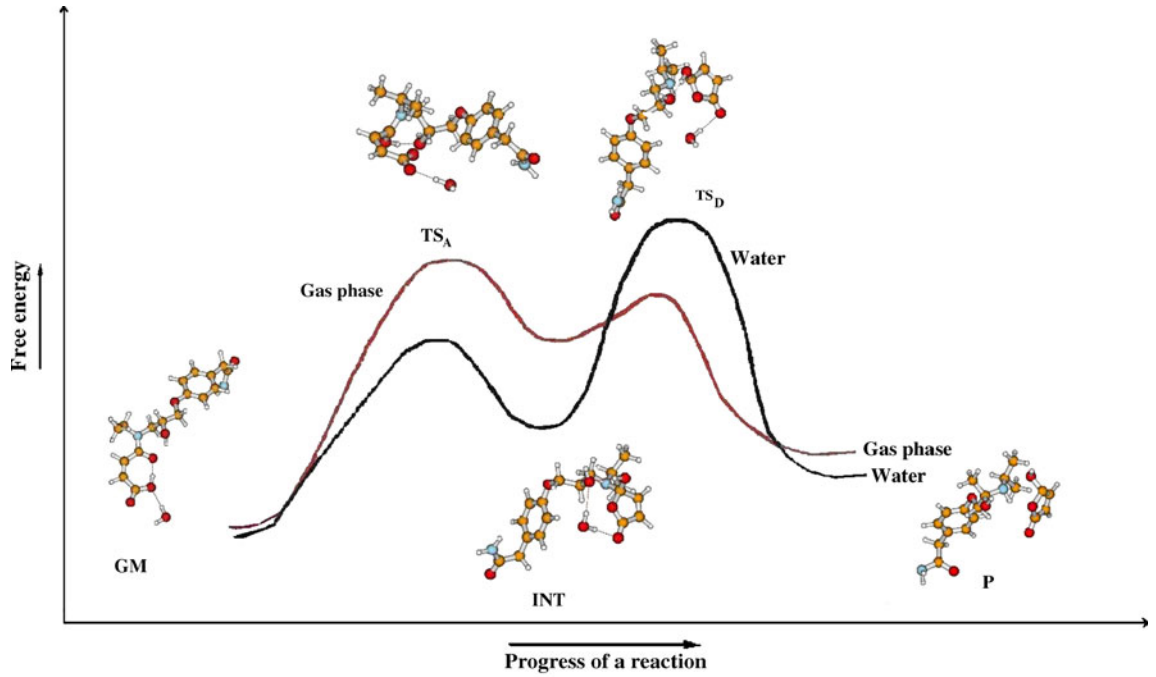



**Fig. 5** A representation of an energy profile for acid-catalyzed hydrolysis of ProD 1 as calculated in the gas phase (GP) and in water (W). GM. TS, INT and P refer to global minimum, transition state, intermediate and product structures, respectively. $TS_A$ and $TS_D$ are approach and dissociation transition state structures, respectively

activation energies of the intra- and the corresponding intermolecular processes. Table 4 lists the calculated EM values for processes **1**–**9** and **ProD 1–2** as calculated by Eq. 5,

$$EM = k_{intra}/k_{inter} \quad (1)$$

$$\Delta G^{\ddagger}_{inter} = -RT \ln k_{inter} \quad (2)$$

$$\Delta G^{\ddagger}_{intra} = -RT \ln k_{intra} \quad (3)$$

$$\Delta G^{\ddagger}_{intra} - \Delta G^{\ddagger}_{inter} = -RT \ln k_{intra}/k_{inter} \quad (4)$$

$$\ln EM = -\left(\Delta G^{\ddagger}_{intra} - \Delta G^{\ddagger}_{inter}\right)/RT \quad (5)$$

where T is the temperature in Kelvin and R is the gas constant.

The calculated log EM values listed in Table 4 were examined for correlation with the corresponding log EM experimental values [56]. The correlation results along with the correlation coefficients are illustrated in Fig. 4b. Inspection of the log EM values listed in Table 4 and Fig. 4b revealed that **2** and **5** were the most efficient processes among **1**–**9**, whereas process **4** was the least. The discrepancy in rates between **2** and **5** on one hand and **4** on the other hand is attributed to strain effects.

Although the calculated and experimental EM values are comparable there is a discrepancy in their absolute values. This is due to the fact that the experimental measurement of the EM values in **1**–**9** was conducted in the presence of aqueous acid whereas the DFT calculations were done in plain water. The dielectric constant value for a mixture of acid/water is expected to be different from that of water (78.39) and hence the discrepancy in the calculated and experimental EM values [70].

In order to draw more credibility to our DFT calculations the calculated DFT free activation energies in water ($\Delta G_{BW}^{\ddagger}$) were correlated with the corresponding experimental free activation energies (Exp $\Delta G^{\ddagger}$) [56]. Strong correlations were found with R value of 0.96 (Fig. 4c).

## Calculation of the $t_{1/2}$ values for the cleavage reactions of prodrugs ProD 1–2

The calculated DFT properties for processes **ProD 1–2** including the geometries of global minimum (GM), intermediate (INT2) and transition state (TS4) structures are shown in Fig. 2 and listed in Tables 1, 2, 3, 4. Comparison of these structures to that calculated for processes **1**–**9** revealed a close similarity.

Using Eq. 6 obtained from the correlation of log $k_{rel}$ vs. $\Delta G^{\ddagger}$ (Fig. 4d) and the $t_{1/2}$ value for process 2 ($t_{1/2}$=1 second) [56], we have calculated the $t_{1/2}$ values for **ProD**

**1** and **ProD 2**. The calculated $t_{1/2}$ at pH 2 for **ProD 1** and **ProD 2** are 65.3 hours and 11.8 minutes, respectively.

$$\log k_{rel} = -0.4421\Delta G^{\ddagger} + 13.534 \quad (6)$$

## Conclusions and future directions

The DFT calculations confirmed that the acid-catalyzed hydrolysis mechanism in Kirby's acid amides **1**–**9** involves: (1) a proton transfer from the hydroxyl of the carboxyl group to the adjacent amide carbonyl carbon, (2) an approach of the carboxylate anion toward the protonated amide carbonyl carbon to form tetrahedral intermediate; and (3) collapse of the tetrahedral intermediate into products. Furthermore, DFT calculations in different media revealed that the reaction rate-limiting step depends on the reaction medium. In aqueous medium the rate-limiting step is a collapse of the tetrahedral intermediate whereas in the gas phase the formation of the tetrahedral intermediate is the rate-limiting step. Furthermore, the calculations establish that the acid-catalyzed hydrolysis efficiency is largely sensitive to the pattern of substitution on the carbon-carbon double bond. The hydrolysis rate was found to be linearly correlated with the following: (a) the strain energy of the tetrahedral intermediate and product and the strain energy difference between the intermediate and the reactant. (b) The distance between the hydroxyl oxygen of the carboxyl group and the amide carbonyl carbon, and (c) the attack angle by which the approach step commences. Further, a linear correlation between the calculated DFT EM values and the experimental EM values demonstrates the credibility of using DFT methods in predicting energies as well as rates for reactions of the type described herein.

Utilizing the linear correlation of the experimental log $k_{rel}$ and EM values the predicted $t_{1/2}$ values of proposed two atenolol prodrugs, **ProD 1**–**2** were predicted.

The strategy to achieve desirable prodrugs of atenolol that are capable of being stable in aqueous solutions and to have the potential for releasing atenolol in a slow release manner is: (1) synthesis of the linker and coupling of the parent drug to the linker using Kirby's synthetic procedure [36]; (2) kinetic studies (in vitro) of the synthesized prodrugs (**ProD 1**–**2**) will be performed in physiological environment (37 °C, pH =2.0 and 6.0 in aqueous medium) and (3) for the prodrugs that show desirable aqueous stability and slow release in the in vitro studies, in vivo pharmacokinetic studies will be conducted in order to determine the bioavailability and the duration of action of the tested prodrugs. In the light of the in vivo pharmacokinetics new prodrugs will be designed and synthesized.

**Acknowledgments** The Karaman Co. is thanked for support of our computational facilities. Special thanks are also given to Angi Karaman, Donia Karaman, Rowan Karaman and Nardene Karaman for technical assistance.

## References

1. Foppa T, Murakami FS, Silva MA (2007) Development, validation and stability study of pediatric atenolol syrup. Pharmazie 62:519–521
2. Anroop B, Ghosh B, Parcha V, Khanam J (2009) Transdermal delivery of atenolol: effect of prodrugs and iontophoresis. Curr Drug Delivery 6:280–290
3. Mcainsh J, Simpson WT, Holmes BF, Young J, Ellis SH (1980) Bioavailability of atenolol formulations. Biopharm Drug Dispos 1:223–232
4. Vergin H, Nitsche V (1989) Oral bioavailability of atenolol. J Int Med Res 17:417–425
5. Shin SC, Choi JS (2003) Enhanced bioavailability of atenolol by transdermal administration of the ethylene-vinyl acetate matrix in rabbits. Eur J Pharm Biopharm 56:439–443
6. http://en.wikipedia.org/wiki/atenolol
7. http://homepage.ntlworld.com/bhandari/Imperial/Atenolol/Synthesis.htm
8. Chan K, Swenden J (2007) Pilot study of the short-term physico-chemical stability of atenolol tablets stored in a multi-compartment compliance aid. Eur J Hosp Pharm Sci 13:60–66
9. Testa B, Mayer J (2003) Hydrolysis in drug and prodrug metabolism-Chemistry Biochemistry and Enzymology. Wiley-VHCA, Zürich
10. Testa B, Mayer JM (2001) Concepts in prodrug design to overcome pharmacokinetic problems. In: Testa B, van de Waterbeemd H, Folkers G, Guy R (eds) Pharmacokinetic optimization in drug research: biological, physiochemical and computational strategies. Wiley-VHCA, Zurich, pp 85–95
11. Wang W, Jiang J, Ballard CE, Wang B (1999) Prodrug approaches in the improved delivery of peptide drugs. Curr Pharm Des 5:265–287
12. Karaman R (2008) Analysis of Menger's spatiotemporal hypothesis. Tetrahedron Lett 49:5998–6002
13. Karaman R (2009) A new mathematical equation relating activation energy to bond angle and distance: a key for understanding the role of acceleration in the lactonization of the trimethyl lock system. Bioorg Chem 37:11–25
14. Karaman R (2009) Reevaluation of Bruice's proximity orientation. Tetrahedron Lett 50:452–456
15. Karaman R (2009) Accelerations in the lactonization of trimethyl lock systems is due to proximity orientation and not to strain effects. Res Lett Org Chem. doi:10.1155/2009/240253
16. Karaman R (2009) The effective molarity (EM) puzzle in proton transfer reactions. Bioorg Chem 37:106–110
17. Karaman R (2009) Cleavage of Menger's aliphatic amide: a model for peptidase enzyme solely explained by proximity orientation in intramolecular proton transfer. J Mol Struct THEOCHEM 910:27–33
18. Karaman R (2009) The *gem*-disubstituent effect-computational study that exposes the relevance of existing theoretical models. Tetrahedron Lett 50:6083–6087
19. Karaman R (2010) Effects of substitution on the effective molarity (EM) for five membered ring-closure reactions- a computational approach. J Mol Struct THEOCHEM 939:69–74
20. Karaman R (2009) Analyzing Kirby's amine olefin – a model for amino-acid ammonia lyases. Tetrahedron Lett 50:73047309

21. Karaman R (2010) The effective molarity (EM) puzzle in intramolecular ring-closing reactions. J Mol Struct THEOCHEM 940:70–75
22. Karaman R (2010) The efficiency of proton transfer in Kirby's enzyme model, a computational approach. Tetrahedron Lett 51:2130–2135
23. Karaman R (2010) Proximity vs Strain in ring-closing reactions of bifunctional chain molecules- a computational approach. J Mol Phys 108:1723–1730
24. Karaman R (2010) The effective molarity (EM) – a computational approach. Bioorg Chem 38:165–172
25. Karaman R (2010) A general equation correlating intramolecular rates with “attack" parameters distance and angle. Tetrahedron Lett 51:5185–5190
26. Karaman R, Alfalah S (2010) Multi transition states in SN2 intramolecular reactions. Int Rev Biophys Chem 1:14–23
27. Karaman R, Pascal R (2010) A computational analysis of intramolecularity in proton transfer reactions. Org Biomol Chem 8:5174–5178
28. Karaman R, Hallak H (2010) Anti-malarial pro-drugs- a computational aided design. Chem Biol Drug Design 76:350–360
29. Karaman R (2010) Prodrugs of aza nucleosides based on proton transfer reactions. J Comput Mol Des 24:961–970
30. Hejaz H, Karaman R, Khamis M (2011) Computer-assisted design for paracetamol masking bitter taste prodrugs. J Mol Model. doi:10.1007/s00894-011-1040-5
31. Bruice TC, Pandit UK (1960) The effect of geminal substitution ring size and rotamer distribution on the intramolecular nucleophilic catalysis of the hydrolysis of monophenyl esters of dibasic acids and the solvolysis of the intermediate anhydrides. J Am Chem Soc 82:5858–5865
32. Bruice TC, Pandit UK (1960) Intramolecular models depicting the kinetic importance of “Fit” in enzymatic catalysis. Proc Natl Acad Sci USA 46:402–404
33. Milstein S, Cohen LA (1970) Concurrent general-acid and general-base catalysis of esterification. J Am Chem Soc 92:4377–4382
34. Milstein S, Cohen LA (1970) Rate acceleration by stereopopulation control: models for enzyme action. Proc Natl Acad Sci USA 67:1143–1147
35. Milstein S, Cohen LA (1972) Stereopopulation control I. Rate enhancement in the lactonizations of o-hydroxyhydrocinnamic acids. J Am Chem Soc 94:9158–9165
36. Menger FM, Ladika M (1990) Remote enzyme-coupled amine release. J Org Chem 35:3006–3007
37. Menger FM, Ladika M (1988) Fast hydrolysis of an aliphatic amide at neutral pH and ambient temperature. A peptidase model. J Am Chem Soc 110:6794–6796
38. Menger FM (1985) On the source of intramolecular and enzymatic reactivity. Acc Chem Res 18:128–134
39. Menger FM, Chow JF, Kaiserman H, Vasquez PC (1983) Directionality of proton transfer in solution. Three systems of known angularity. J Am Chem Soc 105:4996–5002
40. Menger FM (1983) Directionality of organic reactions in solution. Tetrahedron 39:1013–1040
41. Menger FM, Grosssman J, Liotta DC (1983) Transition-state pliability in nitrogen-to-nitrogen proton transfer. J Org Chem 48:905–907
42. Menger FM, Galloway AL, Musaev DG (2003) Relationship between rate and distance. Chem Commun 2370–2371
43. Menger FM (2005) An alternative view of enzyme catalysis. Pure Appl Chem 77:1873–1876, and references therein
44. Brown RF, van Gulick NM (1956) The geminal alkyl effect on the rates of ring closure of bromobutylamines. J Org Chem 21:1046–1049
45. Galli C, Mandolini L (2000) The role of ring strain on the ease of ring closure of bifunctional chain molecules. Eur J Org Chem 3117–3125 and references therein
46. Kirby A J, Parkinson A (1994) Most efficient intramolecular general acid catalysis of acetal hydrolysis by the carboxyl group. J Chem Soc Chem Commun 707–708.
47. Brown CJ, Kirby AJ (1997) Efficiency of proton transfer catalysis Intramolecular general acid catalysis of the hydrolysis of dialkyl acetals of benzaldehyde. J Chem Soc Perkin Trans 2:1081–1093
48. Craze GA, Kirby AJ (1974) The hydrolysis of substituted 2-methoxymethoxybenzoic acids. J Chem Soc Perkin Trans 2:61–66
49. Barber SE, Dean KES, Kirby AJ (1999) A mechanism for efficient proton-transfer catalysis. Intramolecular general acid catalysis of the hydrolysis of 1-arylethyl ethers of salicylic acid. Can J Chem 792–801
50. Kirby AJ, de Silva MF, Lima D, Roussev CD, Nome F (2006) Efficient intramolecular general acid catalysis of nucleophilic attack on a phosphodiester. J Am Chem Soc 128:16944–16952
51. Kirby AJ, Williams NH (1994) Efficient intramolecular general acid catalysis of enol ether hydrolysis. Hydrogen-bonding stabilization of the transition state for proton transfer to carbon. J Chem Soc Perkin Trans 2:643–648
52. Kirby AJ, Williams NH (1991) Efficient intramolecular general acid catalysis of vinyl ether hydrolysis by the neighbouring carboxylic acid group. J Chem Soc Chem Commun 1643–1644
53. Hartwell E, Hodgson DRW, Kirby AJ (2000) Exploring the limits of efficiency of proton-transfer catalysis in models and enzymes. J Am Chem Soc 122:9326–9327
54. Kirby AJ (1997) Efficiency of proton transfer catalysis in models and enzymes. Acc Chem Res 30:290–296
55. Asaad N, Davies JE, Hodgson DRW, Kirby AJ (2005) The search for efficient intramolecular proton transfer from carbon: the kinetically silent intramolecular general base-catalysed elimination reaction of o-phenyl 8-dimethylamino-1-naphthaldoximes. J Phys Org Chem 18:101–109
56. Kirby AJ, Lancaster PW (1972) Structure and efficiency in intramolecular and enzymatic catalysis Catalysis of amide hydrolysis by the carboxy-group of substituted maleamic acids. J Chem Soc Perkin Trans 2:1206–1214
57. Kirby AJ, Hollfelder F (2009) From Enzyme Models to Model Enzymes. R Soc Chem, 1st edn, pp 1–273
58. http://www.gaussian.com.
59. Casewit CJ, Colwell KS, Rappé AK (1992) Application of a universal force field to main group compounds. J Am Chem Soc 114:10046–10053
60. Murrell JN, Laidler KJ (1968) Symmetries of activated complexes. Trans Faraday Soc 64:371–377
61. Muller K (1980) Reaction paths on multidimensional energy hypersurfaces. Angew Chem Int Edn Engl 19:1–13
62. Cancès MT, Mennucci B, Tomasi J (1997) A new integral equation formalism for the polarizable continuum model: theoretical background and applications to isotropic and anisotropic dielectrics. J Chem Phys 107:3032–3041
63. Mennucci B, Tomasi J (1997) Continuum solvation models: A new approach to the problem of solute’s charge distribution and cavity boundaries. J Chem Phys 106:5151–5158
64. Mennucci B, Cancès MT, Tomasi J (1997) Evaluation of solvent effects in isotropic and anisotropic dielectrics and in ionic solutions with a unified integral equation method: Theoretical bases, computational implementation, and numerical applications. J Phys Chem B 101:10506–10517
65. Tomasi J, Mennucci B, Cancès MT (1997) The IEF version of the PCM solvation method: an overview of a new method addressed

to study molecular solutes at the QM ab initio level. J Mol Struct THEOCHEM 464:211–226
66. Katagi T (1990) AM1 study of acid-catalyzed hydrolysis of maleamic (4-amino-4-oxo-2-butenoic) acids. J Comput Chem 11:1094–1100
67. Burker U, Allinger NL (1982) Molecular Mechanics. American Chemical Society, Washington DC, USA
68. Dafforn A, Koshland DE Jr (1971) Theoretical aspects of orbital steering. Proc Natl Acad Sci USA 68:129–139
69. Kirby AJ (2005) Effective molarities for intramolecular reactions. J Phys Org Chem 18:101–278
70. It should be noted that DFT calculations in the presence of a mixture of acid and water are not feasible

J Mol Model (2012) 18:1541–1552
DOI 10.1007/s00894-011-1182-5

ORIGINAL PAPER

# Molecular dynamics simulations of EMI-$BF_4$ in nanoporous carbon actuators

**Endel Soolo · Daniel Brandell · Anti Liivat · Heiki Kasemägi · Tarmo Tamm · Alvo Aabloo**

Received: 12 April 2011 / Accepted: 7 July 2011 / Published online: 30 July 2011


**Abstract** An artificial muscle composite material consisting of carbide derived carbon (CDC) and 1-ethyl-3-methylimidazolium tetrafluoroborate (EMI-$BF_4$) ionic liquid was modeled using molecular dynamics (MD) simulations, in order to determine the molecular structural rearrangements causing actuation. CDC was represented as separate curved graphene-like flakes with charges of +2, 0 or −2 on each flake, with 24–27 aromatic rings each. The charge distribution in the flakes was determined by PM6 semi-empirical optimization. The pore size distribution of CDC and the density of the material were comparable to experimental data. Molecular structure analysis revealed a preferential parallel orientation for the cations over the negatively charged CDC surfaces, while cationic rotations and reorientations could be observed for positively charged CDC. Changes in the pore occupancy for each ionic type were observed for pore sizes between 4 and 7 Å, which, together with the replacement of large cations with smaller anions, could explain the volume decrease in the anodes (and, vice versa, the volume increase in the cathodes) in this type of actuator.

**Keywords** Artificial muscle · 1-Ethyl-3-methyl-imidazolium tetraborofluorate · Carbide derived carbon · Molecular dynamics

**Electronic supplementary material** The online version of this article (doi:10.1007/s00894-011-1182-5) contains supplementary material, which is available to authorized users.

E. Soolo · H. Kasemägi · T. Tamm · A. Aabloo (✉)
IMS Lab, Institute of Technology, University of Tartu,
Nooruse 1,
50411 Tartu, Estonia
e-mail: alvo@ut.ee

D. Brandell · A. Liivat
Department of Materials Chemistry, Uppsala University,
Box 538, 75121 Uppsala, Sweden

## Introduction

Electroactive polymers (EAP) are materials that change their shape in response to electric stimulation, and are therefore useful as lightweight actuators in the space industry [1], medical applications [2], or in biomimetic robotics [3, 4]. EAPs can generally be divided into two basic categories: ionic and electronic EAPs. Ionic EAPs act by ion/ion-pair relocation during actuation—a process similar to supercapacitors. They are better suited for miniature devices, since they can be operated under low voltage (1–10 V) [5]. However, most ionic EAPs offer only a small force, display low electromechanical efficiency, and have a short lifetime compared to traditional electromechanical actuators. Several practical solutions have been proposed to improve the total actuator force; e.g., designing specific shapes [6] or bundling multiple actuators [7]. In order to increase efficiency, however, it is vital to understand the details of the actuation process and the ion migration at the atomistic level [5].

Ionic EAP materials are composed of three basic components: a porous electrode material (often carbon), an ionic transfer medium, and a polymer binder to add toughness and flexibility. Several types of microporous carbon composite materials have been used in supercapacitors, where charges accumulate in a double layer on a large surface area. Carbide derived carbon (CDC) capacitor capacitance is determined mainly by two components [8]: electrostatic/double layer capacitance [9] and faradaic/pseudocapacitive components [10]. Pore diameters smaller than 1 nm result in higher capacitance due to larger surface area and confinement effects [11]. CDCs are one of the best known nanoporous materials for ion storage, since the pore-size is uniform and can be fine-tuned over a wide range [12]. For EAP applications, ionic liquids (ILs) [13] are frequently used as ion transfer media. ILs, or molten

salts, consist of charged particles without additional solvent. Due to their low vapor pressure, ILs avoid evaporation of the electrolyte when functioning in a dry environment. IL electrolytes can also improve the cycle lifetime of the material compared to water-based electrolytes [14]. The EAP actuator comprises two electrode sheets of the above-mentioned components laminated together, with an ion-permeable membrane, impregnated with electrolyte, in between.

Different kinds of ion storage, ion transport medium, binder and membrane materials offer countless possibilities for EAP construction. For example, IL electrolytes in Nafion membranes have been tested in combination with metal [15], carbon nanotubes [16], CDC [17] or $RuO_2$ [16, 17] electrodes. CDC in a polytetrafluoroethylene (PTFE) matrix has also been used as electrodes, in combination with an IL electrolyte and an ion permeable paper membrane [18]. Asaka et al. [19, 20] have studied carbon bucky gel actuators, where the electrodes contain carbon nanotubes in different ILs, supported by polyvinylidene fluoride (PVDF)-based binders [19]. The largest strain was achieved by using 1-ethyl-3-methylimidazolium tetrafluoroborate (EMI-$BF_4$) as the IL [20]. Carbon nanotube sheets have also been demonstrated to work as actuators in water-based electrolytes [21].

The EAP bending motion is caused either by electrode or membrane expansion at one side of the device and/or contraction at the other side upon actuation. The mechanisms causing the expansions and contractions are still to a large degree debated [22]. Suggestions include bond length changes due to repulsive Coulomb forces accompanying double layer charging [21], insertion of counterions [23], electro-osmosis [24], electrostatic repulsion of the charged molecules in the material [21], decreased interfacial tension at the pore surfaces due to change in double-layer charges [25], etc. A correlation has been found between increased accumulation of ions on the electrode surface (i.e., the capacity) and improved peak strain [16], but that result was seen to be dependent on electrolyte type [18]. Considering the large influence of the choice of IL or the type of carbon on EAP performance [17, 18], it is likely that the primary factor for strain ratio and maximum strain is the ionic transport and structure rather than changes in electronic configuration. For example; when comparing EAPs with EMI-trifluoromethanesulfonate(EMI-Tf)-based electrolyte with similar materials using tetraethyl ammonium (TEA) $BF_4$-based electrolyte, which has a larger effective ion size due to ion pairing, the EMI-Tf can generate larger strains and has larger capacitance [18]. Increasing the ion mobility could, therefore, help significantly in increasing both the strain ratio and the maximum strain of EAPs [14].

Several approaches can be used to increase ionic conductivity. It is well known from the field of battery polymer electrolytes that changing the monomeric sequence [26], adding plasticizers [27] or nano-particle fillers [28] have all proven to be successful strategies. In this context, molecular dynamics (MD) simulations have helped generate a detailed view of the ion movement processes, and have been used frequently to study electrolyte systems, primarily for batteries [29–35] but also for EAPs [36, 37].

In this study, the structure-dynamic properties of ionic motion in EAPs based on CDC carbon and EMI-$BF_4$ ILs were investigated using MD simulation techniques. Changes in the microstructure when the charge accumulated in the carbon electrodes is varied were investigated, in an attempt to understand fundamentally the volume change causing actuation, and to search for strategies for enlarging peak strain and ion mobility. To our knowledge, this is the first MD simulation of a CDC-based actuator material.

## Methodology

The MD method is used for computer simulations to estimate coordination numbers, diffusion coefficients, transport mechanisms and other macroscopic characteristics [38].

### Simulated systems

Nanoporous amorphous CDC consists of slightly curved graphene-like sheets, for which the amount of long-range ordering depends on the quenching rate [39]. This structure was approximated here by slightly curved graphene flakes of similar sizes, with 24–27 aromatic rings each. The number of flakes (24 per MD box) was chosen based on the consideration that the simulation box should contain enough flakes for the carbon structure to be amorphous. In CDC, the carbon flake structures are connected to each other via strong bonds, but this has been neglected here, rendering an "exfoliated" system where a liquid-like behavior could be expected. This approximation is valid, however, considering the limited time frame over which the system is simulated—only local relaxation effects such as reorientation and absorbance of IL molecules are expected, and are indeed the focus of this study. The pore sizes in our model are defined as the shortest distances between carbon flakes. In that way, simulation results can be compared with the experimental pore size distribution of the material.

To gain insight into the structural changes associated with the ion conduction mechanism during actuation, the charge of the CDC carbon was varied between +2, 0 and −2 unit charges per carbon flake for all flakes in the MD box, whereafter the anion:cation ratio was adjusted to maintain electroneutrality. The carbon to IL (here EMI-$BF_4$; molecular structure and atom labels used are presented in Fig. 1)

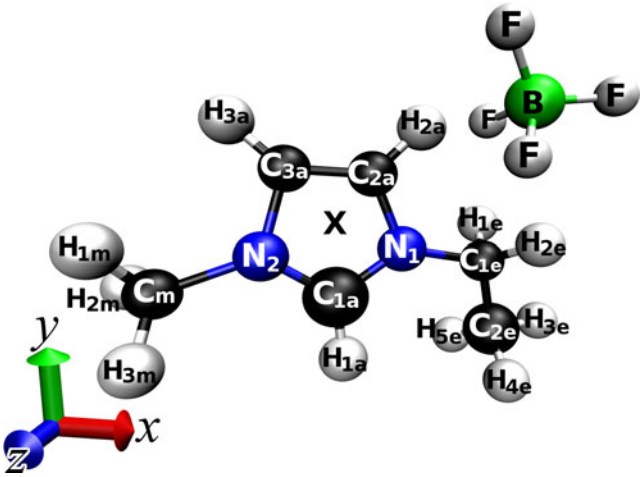



**Fig. 1** The ionic liquid EMI-$BF_4$ with atom labels

mass ratio was fixed at 1.89, which falls within the range of experimentally estimated values [18, 40]. Due to its chemical inertness and low concentration, the polymer binder was excluded from the simulations. The three different components—carbon flakes, $EMI^+$ cations and $BF_4^-$ anions—were inserted into four MD boxes as indicated in Table 1.

## Construction of simulation cells

Separate slightly curved graphene flakes were used to create a structure resembling CDC. The individual shapes of the flakes were created in planar form initially by putting together six-membered carbon rings with occasional five-membered rings (defects). The energy of the flakes was minimized by the molecular mechanics routine in DLPOLY [41] (1 ns long simulations in vacuum with 0.1 fs timestep using energy minimization). The defects caused the flakes to bend. To allow some variation of the carbon flakes, eight different types of flakes were created with 24–27 aromatic rings in each flake, including 0–3 five-membered rings. In the following MD simulations, three flakes of each type were included, thus resulting in 24 carbon flakes per system. The detailed parameters of each flake are listed in the Electronic supplementary material. A typical CDC supercapacitor shows specific capacities of >100 F/g, equivalent to >200 C/g or ~200 C/cm$^3$. The resulting charge accumulation is then ~1 e/nm$^3$. Since the MD box used is ~64 nm$^3$, the CDC charges simulated are realistic.

The amorphous structure of the simulated material was generated with Monte Carlo inspired methods, using in-house software [42]. The carbon flakes were rotated in rigid body manner and inserted sequentially in random locations. Anions and cations were thereafter inserted sequentially into the carbon matrix until the box was filled, generating a density of 1.46 g/cm$^3$. The models were created in cubic simulation boxes measuring 40×40×40 Å with periodic boundary conditions. The model with neutral carbons was created first, while for the systems with charged carbon the configuration of the carbon flakes was left unchanged, but their partial atomic charges were modified according to the quantum chemical calculations described below. All flakes in a system were given equal charges. To maintain both mass and electroneutrality, 27 anions were removed and 21 cations were inserted into the system when charge on carbon flakes was changed from 0 to +2 and vice versa for charge −2. The compositions of all studied systems are listed in Table 1.

## Force field

The force field used was a combination of parameters derived from quantum chemical calculations and force fields published in the literature. The partial charges on carbon and hydrogen atoms in the graphene flakes for charged flakes (+2 or −2 charge) were calculated for each type of carbon flake (with different number and location of defects) using the PM6 [43] semi-empirical optimization method with open shells (UHF). For neutral carbon flakes, the average partial charges of hydrogen atoms for the flake type without defects were calculated using the PM6 method, and the same charges were used for all hydrogen atoms in all flake types. The charges on carbon atoms were chosen to maintain electroneutrality over each flake type with different numbers of carbon atoms. A complete list of partial charges in each type of carbon flakes can be found in the Electronic supplementary material. Parameters for bonded interactions between carbon atoms within the carbon flakes were taken from Walther et al. [44], while

**Table 1** Simulated systems. *IL* Ionic liquid

| System no. | IL:carbon mass ratio | Carbon | | | $EMI^+$ cations | | $BF_4^-$ anions | | Cell size | Density |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Mass% | No. of flakes | Flake charge | Mass% | No. of cations | Mass% | No. of anions | (Å) | (g/cm$^3$) |
| 1 | 1.89 | 34.6 | 24 | +2 | 32.6 | 165 | 32.8 | 213 | 41.12 | 1.34 |
| 2 | 1.89 | 34.6 | 24 | 0 | 36.7 | 186 | 28.7 | 186 | 41.47 | 1.31 |
| 3 | 1.89 | 34.6 | 24 | −2 | 40.9 | 207 | 24.5 | 159 | 41.75 | 1.28 |

parameters from AMBER [45] were used for hydrogen atoms at the edges of the carbon flakes. The description of the EMI-$BF_4$ interactions was taken from Lopes et al. [46]. van der Waals forces between atoms with force fields originating from separate sources were calculated using standard combination rules (Eq. 1) for the Lennard-Jones potential (Eq. 2) parameters for each atom pair (subscripts 1 and 2 denote Lennard-Jones parameters of the atom types):

$$\varepsilon = \sqrt{\varepsilon_1 \varepsilon_2}, \sigma = \frac{\sigma_1 + \sigma_2}{2} \quad (1)$$

$$V(r) = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \quad (2)$$

All force field parameters are presented in the Electronic supplementary material.

### Simulation details

Each system was first equilibrated by a 2-ns-long MD simulation in the NVT ensemble at 293 K temperature in the original 40 Å cubic cell to relieve internal stress originating from generation. Thereafter, a 10-ns-long simulation followed in the NPT ensemble at 293 K temperature, using a Nose-Hoover barostat with 1 atm pressure and a relaxation time of 0.3 ps. The simulations used a Verlet leap-frog algorithm with 1 fs timestep. The temperature was kept at 293 K by a Nose-Hoover thermostat with 0.1 ps relaxation time. Ewald summation with $10^{-5}$ precision was used for all electrostatic interactions. The cubic symmetry of the simulation cell was retained in all simulations.

## Results and discussion

### System description

After equilibration, the box volume increased slightly, generating stable densities for all simulated systems at ~1.3 g/cm$^3$ (Table 1). This result is comparable to values of the measured density (1.0–1.3 g/cm$^3$) of an actuator based on these materials measured by Janno Torop [18], but the latter is also a system consisting of other components such as polymer binders, and where the density varies substantially due to inhomogeneous results of the synthesis process. The decreasing densities also indicate that the initial box generation was satisfactory. The carbon flakes and ions were distributed evenly in the simulation cell after equilibration, without any detectable phase separation (Fig. 2).

### Pore size and content distribution

The pore size distribution was estimated from the spaces between the carbon flakes; modeled as an array of non-overlapping spherical pores. The pores were found by starting with a uniform grid covering the whole simulation cell. A sphere was created around each grid point with a radius equivalent to the distance to the nearest CDC atom (minus its van der Waals radius). The spheres were thereafter sorted by size and overlapping spheres of smaller

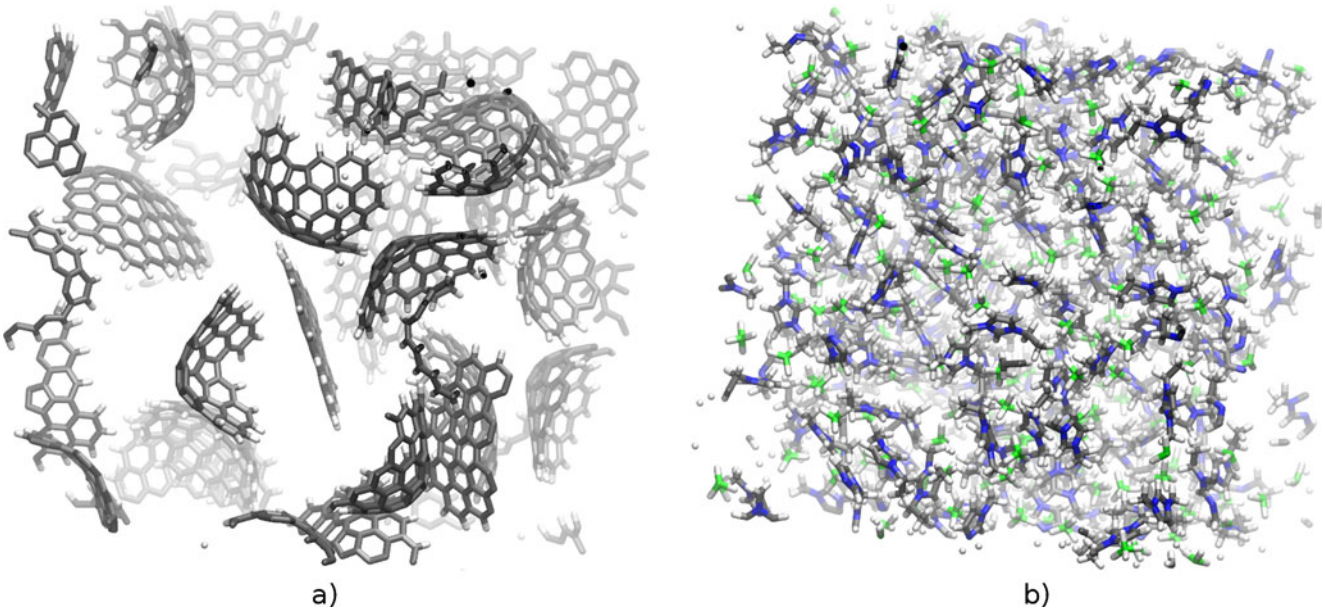



**Fig. 2** The molecular dynamics (MD) simulation box, separated into **a** carbon flake and **b** ionic liquid content

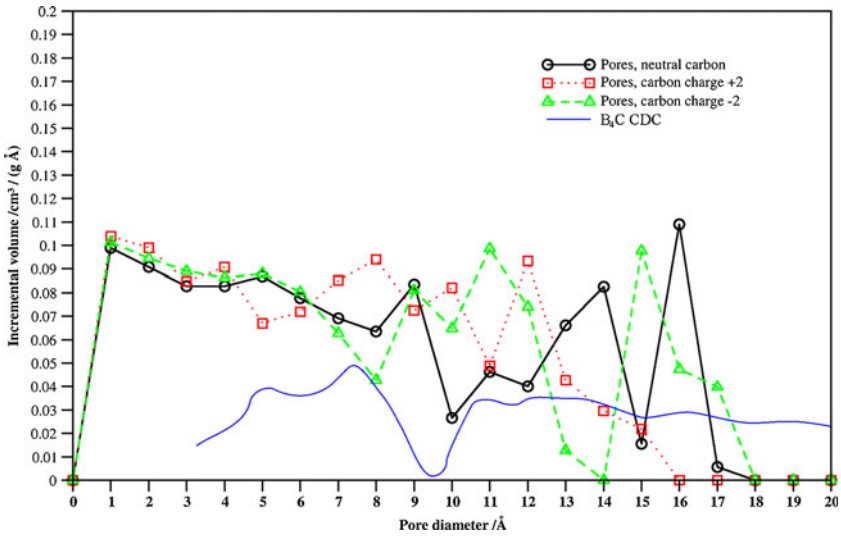



**Fig. 3** Pore size distribution in the simulated systems

size were deleted, thus generating a picture of the overall pore size distribution; see Fig. 3. It should be noted that this pore size distribution is estimated for an "exfoliated" CDC structure, which might display some differences as compared to the material used in experimental structures.

It is clear from Fig. 3 that, although the number of very small pores is large, the total volume of the different pore sizes is rather uniform up to 8 Å, with a decrease in volume (and number) visible thereafter. A few pores with substantial pore diameter (e.g., 1.3 and 1.6 nm for 0 charge) also exist. The lack of any larger pores can be explained by the size and density of the simulation box, which effectively reduces the possibility of forming large distances between the CDC flakes.

The pore size distribution in CDC can be estimated experimentally by BET nitrogen absorption. However, the experimental data in the literature vary substantially, and are very dependent on the synthesis technique. For example, in one sample [47], also shown in Fig. 3, the pore size had a peak at 7 Å, a minimum at 9.5 Å and ua niform distribution of pores with sizes ranging from 11 to 20 Å. These data display a similar pattern to the simulated pore size distribution here, although the numbers are somewhat different. Generally, the pore volume is also larger in the estimates from the simulation data. The discrepancies can be explained by the fact that the simulated carbon was not generated from a crystalline precursor, and that BET measurement generates only approximate values and tends to neglect very small pores (<5 Å).

When the carbon flakes are charged positively or negatively, the distribution of pores larger than 7 Å changes, while the pore size distribution of pores smaller than 6 Å is unchanged. This fis in agreement with the experimental observation [18] that the largest EAP strain come from CDCs with large amounts of pores larger than 10 Å. This, in turn, can be due to the ion distribution inside the pores. Due to the different sizes of cations and anions, they occupy differently sized pores in the CDC, with small pores left unoccupied, and therefore less affected by the charging of the carbon flakes. This is illustrated in Fig. 4: most pores are empty up to 6 Å diameter, and $BF_4^-$ anions can occupy much smaller pores than $EMI^+$. Larger pores can contain clusters with larger numbers of cations and anions (clusters with two anions and two cations can be found most often in pores of 9 Å diameter), and this can give rise to structural reorganization during charging.

Figure 5 illustrates the shift in pore occupancy depending on actuator charge. This is due to both relocation of the ions upon charging, and changing pore sizes. When the carbon flakes are positively charged, a small proportion of the cations move into slightly larger pores, due to the unfavorable electrostatic interaction with carbon atoms; the mean size of a pore containing a single $EMI^+$ cation

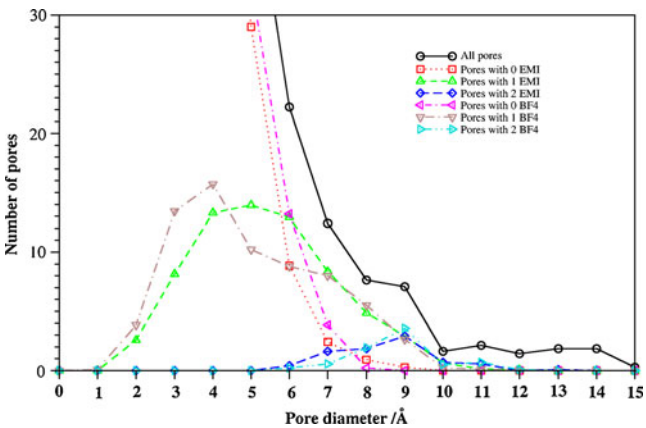



**Fig. 4** Ion distribution in pores in the simulated system with uncharged carbon flakes

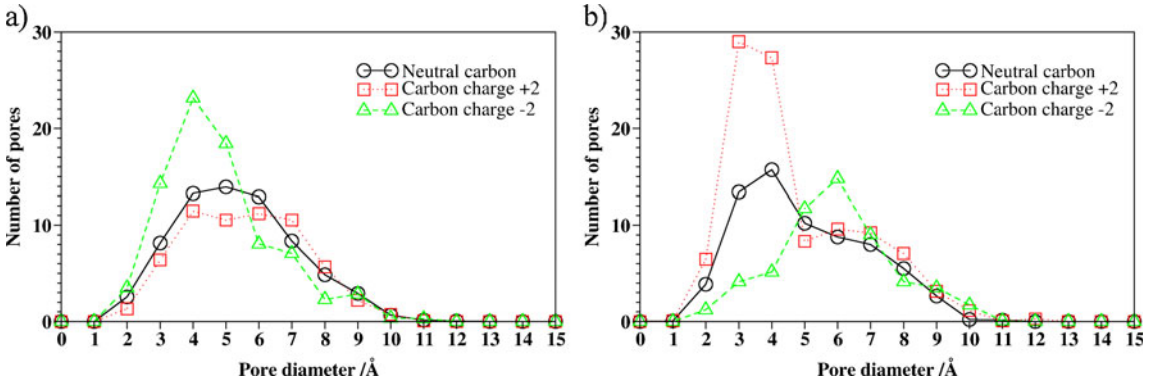



**Fig. 5** Distribution of pores containing either one cation (**a**) or one anion (**b**) depending on the charge of the carbon flakes

increases from 6 Å to 7 Å. On the other hand, when the carbon flakes are negatively charged, a significant amount of cations move into smaller pores with only 4 Å diameter—almost the smallest pore size which can host an $EMI^+$ ion. The $EMI^+$ ions also attract the negatively charged carbon flakes towards it, thereby effectively reducing the pore size. Naturally, this effect is reversed for $BF_4^-$ cations, which move into larger pores during negative charging, where they can pair up with $EMI^+$. During positive charging, the most frequent pore size containing a single anion stays at 3–4 Å (due to steric hindrance, the anions have difficulties moving to smaller pores), but the number of ions occupying this pore size increases.

## Carbon flake interactions with ions

Structural properties were determined by the radial distribution function (RDF), whose integral is equivalent to the coordination number (CN) function. By calculating the RDF for different directions, the spatial distribution function (SDF) can be obtained.

According to quantum chemical calculations, the average charge of a carbon atom in a neutral carbon flake is −0.05 e, while the average charge of terminating hydrogen atoms is +0.16 e. However, the charge is not distributed uniformly in the structure of the flake, but comprises a span of a charge difference of 0.25 e, which is more than the difference between the average carbon atom charge in a positively charged flake and that of negatively charged flake. Each carbon flake in fact contains both positively and negatively charged carbon atoms. Nevertheless, the—on average positively charged—CDC carbon atoms in a neutral flake are coordinated more closely to cations than anions, which can be seen in the RDFs in Fig. 6. For example, at a distance of 5.5 Å from the CDC carbons, the CN=1.24 for the cation, and 0.84 for the anion. It is also obvious from Fig. 6 that there exists a first coordination sphere for cations around CDC at ~4.3 Å, although CN reaches 1 first at 5.2 Å distance.

Different atoms of $EMI^+$ cations can coordinate directly to the CDC carbons, which in turn generates a picture of the preferred orientation of the cations towards the carbon flakes. Figure 7 shows the RDFs between CDC and different cation atoms. In the first coordination sphere, the preferred cation orientation is planar and parallel to the carbon surface, indicated by equal 4.2–4.3 Å distances to all aromatic ring atoms in the first coordination sphere of the CDC carbon. An example of such a configuration is also

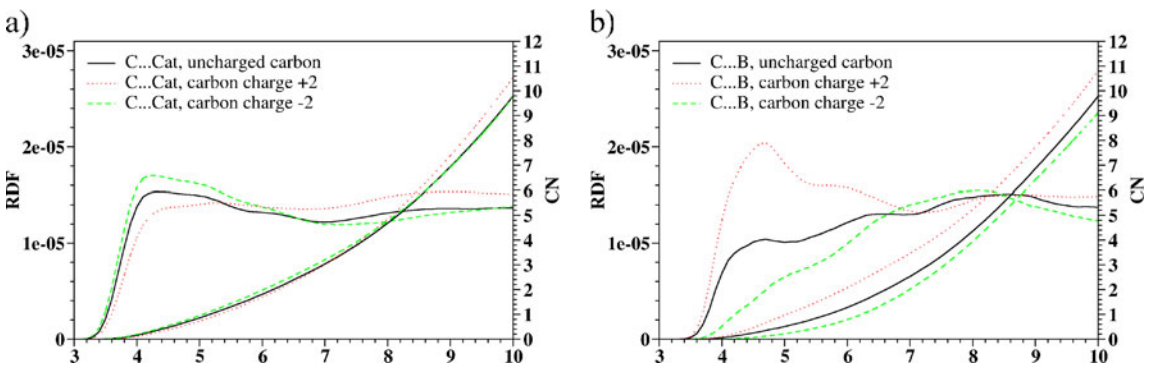



**Fig. 6** Radial distribution functions between carbide derived carbon (CDC) and cation atoms averaged for all coordinating atoms (**a**) and B in $BF_4^-$ (**b**)

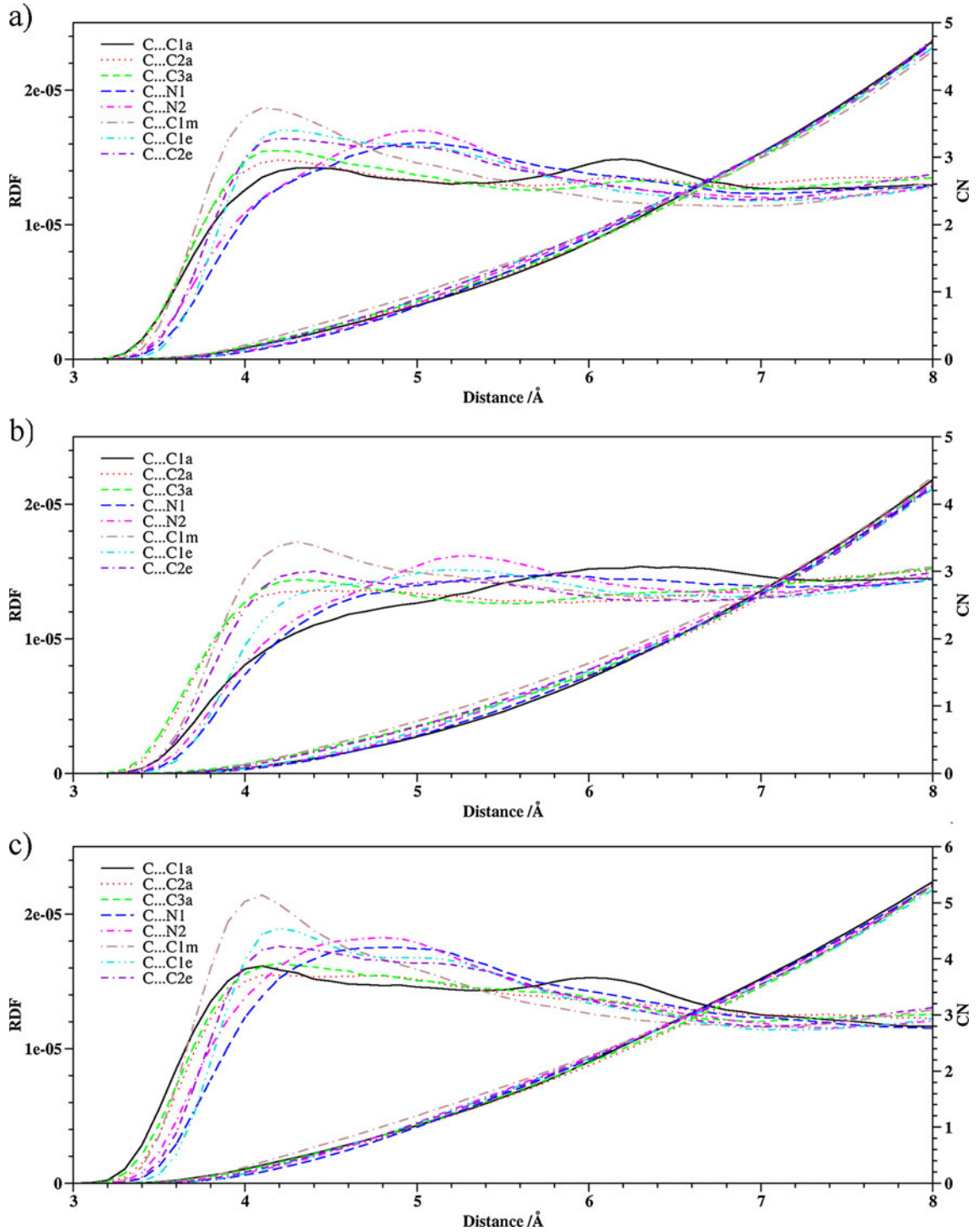



**Fig. 7** Radial distribution functions between CDC carbon and different cation atoms. Charge on carbon flakes: **a** 0, **b** +2, **c** −2

depicted in Fig. 8a. Also, the non-aromatic $EMI^+$ carbon atoms ($C_{1e}$, $C_{2e}$ and $C_m$) are generally found at 4.2–4.3 Å distance values, which is not surprising considering that these units carry positive partial charges.

Another configuration that can be found in the systems is where the $EMI^+$ ion is coordinated perpendicularly to the CDC surface, i.e., where the CDC carbons are coordinated directly only to the $C_{2a}$ and $C_{3a}$ carbons on the aromatic ring (see Fig 8b). This is also indicted by the 5.0 Å peaks in the RDFs for $N_1$…C and $N_2$…C, and the >6 Å peak for the $C_{1a}$…C RDF (see Fig. 7a). The parallel configuration is more common, and can be found for ~80 % of all cations on

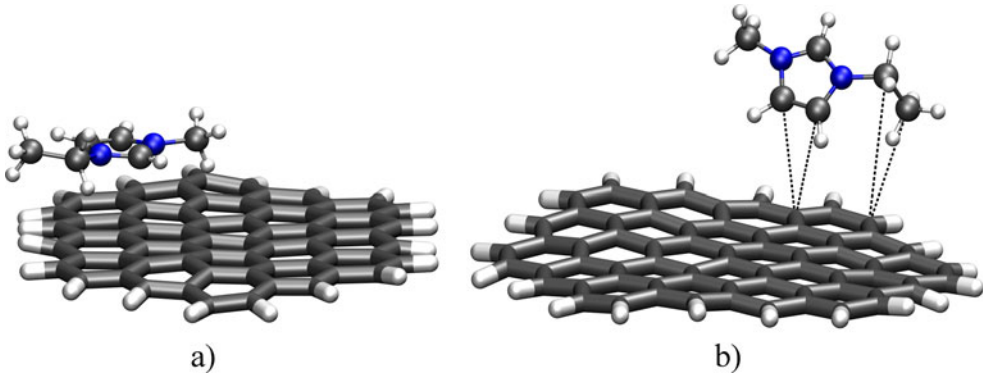



**Fig. 8** Typical cation coordinations to a CDC flake coordination. **a** Parallel, **b** perpendicular

the CDC surface, which can be determined from the ratio of coordination numbers for $C_{CDC}…C_{2a}$ (CN=0.48) and $C_{CDC}…N_1$ (CN=0.38) at the same 4.5 Å distance. There are also rare cases where the CDC carbons are coordinated to either the methyl or ethyl end groups of the $EMI^+$ ions.

When the carbon flakes are positively charged, the electrostatic repulsion causes ~15 % of the cations to turn 90° around the $C_{2a}$–$C_{3a}$ bond (i.e., around the *x*-axis in Fig. 1), thereby generating a perpendicular orientation to the carbon surface. This can be seen as a small shift in the $C_{CDC}…C_{1a}$ coordination from a coordination distance of 4.3 Å to 6.3 Å (Fig. 7a,b), while coordination to $C_{2a}$ and $C_{3a}$ remain virtually unchanged. The reason for this rotation is that the partial positive charges of the $EMI^+$ ions are concentrated more to the $C_{1a}$ part of the aromatic ring (+0.40e as compared to +0.16e for the $C_{2a}/C_{3a}$ side), and is therefore more strongly repelled. During the rotation, the outer ethyl group atom $C_{2e}$ remains coordinated to the carbon surface at 4.2 Å distance, while the inner ethyl group atom $C_{1e}$ is moving away from carbon surface, from a coordination distance of 4.2 Å to 5.2 Å (Fig. 7b), which is slightly surprising considering the free rotation of the $C_{2e}$ atom and its positive partial charge. It is thus probable that this is a steric effect.

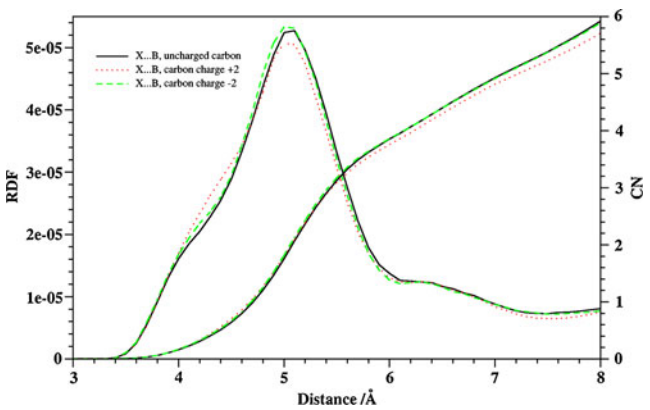



**Fig. 9** Radial distribution functions between the center of the cation aromatic ring (X) and B in $BF_4^-$

The $EMI^+$ cation also rotates around the axis perpendicular to the aromatic ring plane (the *z*-axis in Fig. 1) when the carbon is positively charged. On average, the ion rotates 11° due to a stronger repulsion of the ethyl than the methyl group. This is indicated by the $C_{CDC}…N_1$ coordination distance increasing from 5.0 Å to 5.7 Å, while the $C_{CDC}…N_2$ coordination increases only from 5.0 Å to 5.3 Å (Fig. 7a, b). A tendency for some cations to rotate 90° around the *y*-axis in Fig. 1 can also be noticed, resulting in an increased proportion of cations oriented with either the methyl or ethyl $EMI^+$ end groups towards the carbon flake instead of being parallel to the surface.

When the carbon flakes are charged negatively with charge −2e on each flake, the cations, to some degree, rotate in the opposite way as compared to the positive charge, i.e., with stronger coordination to the more positively charged part of the ion. This is visible through a preferred coordination to the CDC carbon surface from the $C_{1a}$ carbon atom (see Fig. 7c). This configuration is 7% more common than coordination to $C_{2a}$ or $C_{3a}$, according to the ratio between $C_{CDC}…C_{1a}$ and $C_{CDC}…C_{2a}$ CNs at 5 Å distance. The cation does not rotate around the axis perpendicular to the aromatic ring plane (*z*-axis in Fig. 1) during negative charging, which is indicated by the equal shifts of the $C_{CDC}…N_1$ and $C_{CDC}…N_2$ RDFs. The reason for the cation rotations being less pronounced for negative than for positive charging of the carbon flakes is due to the positive electrostatic attraction between negative CDC and $EMI^+$, which renders as more pronounced only the preferred coordination configurations from the neutral state. This is seen as a general shift of the RDFs to lower values during negative charging (Fig. 7a,c).

The anions coordinate to the CDC carbon within two less well-defined coordination spheres with radii of 4.6 Å and 6.4 Å; the corresponding accumulated coordination numbers are 0.6 and 2.8, respectively (see Fig. 6). The first coordination sphere corresponds to a direct coordination, while the second represents anions coordinated primarily to cations in the vicinity of CDC carbon. Due to the small size of the anion, which makes it easier for it to

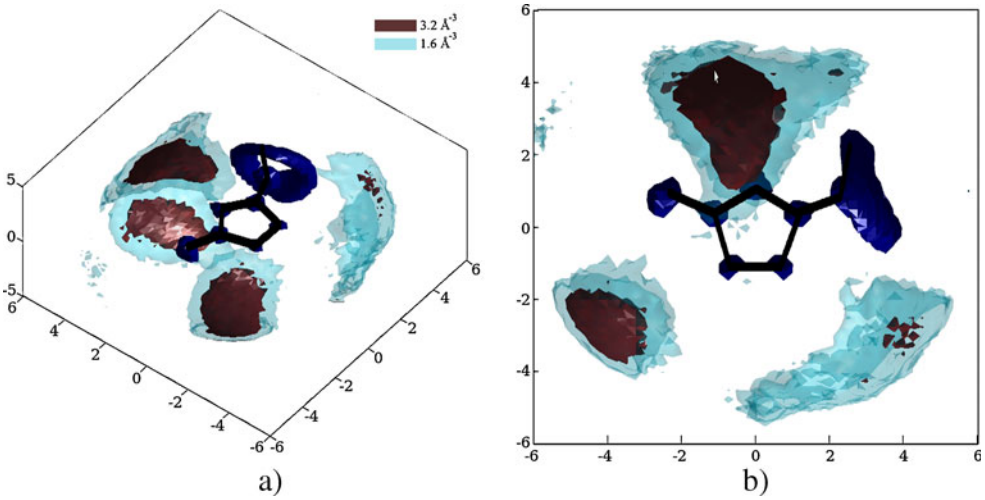



**Fig. 10** Spatial distribution functions for B in $BF_4^-$ around the cations. The probability density of finding an anion at a specific location is illustrated in *light blue* and *red*, while the $EMI^+$ atom mobility is represented in *dark blue*. **a** Angle view, **b** top view. All coordinates are in Ångströms, and hydrogens are omitted

move and reorient, its coordination is more sensitive to charging of the carbon flakes than that of the cation. Generally, the anions move between the three coordination spheres of the carbon flakes during charging, with the first preferred at positive charging and the third preferred at negative. When the carbon flakes are positively charged, the coordination number for $C_{CDC}$…B at 5.2 Å (corresponding to the first coordination sphere) increases from 0.6 to 1.3, while it drops to only 0.3 for negative charging.

## Ion-ion interaction

When the pore size so allows, a cation is always surrounded by one or more anions, which can be located in several locations around the cation. If the center of the aromatic ring is denoted X, a strong maximum in the X…B RDF can be noted at 4–6 Å distance (see Fig. 9). In this first coordination sphere, the $EMI^+$ cations are typically coordinated by 4 $BF_4^-$ anions. This coordination is rather independent of the CDC charging, although the CN increases slightly for negative charge and decreases for positive charge. One explanation for this behavior could be that the positive charging stimulates the anions to move into smaller pores (see Fig. 4), where they are inaccessible for cation for coordination, which has less correspondence for negative charging due to the immobility of the bulky $EMI^+$ cations.

In order to get the full picture of the ion–ion coordination, the SDFs are plotted in Fig. 10 with the cation in two different planes. It is obvious that there are three more strongly preferred location sites for the anions: two near the $C_{1a}$ atom, almost symmetrically on each side of its mirror plane. These anions generally coordinate to the $C_{1a}$, $N_1$ and $C_{1e}$ atoms or to the $C_m$, $C_{1a}$ atoms and one of the ethyl carbons, respectively; both ions having a tridentate coordination with one F atom coordinating one C or N atom on $EMI^+$. The third site is located close to the $C_m$ and $C_{1a}$, also in a tridentate coordination with the third F atom coordinating to $N_2$. These locations also represent the areas around the $EMI^+$ ions where the accumulated positive partial charge is the largest (the most positive partial charges can be found on the aromatic hydrogen, followed by $N_2$), and is therefore energetically favorable for the $BF_4^-$ anion. It is also visible in Fig. 10 that there is a high probability of finding a fourth anion around the $C_{2e}$ group, which also carries a substantial positive partial charge. However, the free rotation around the

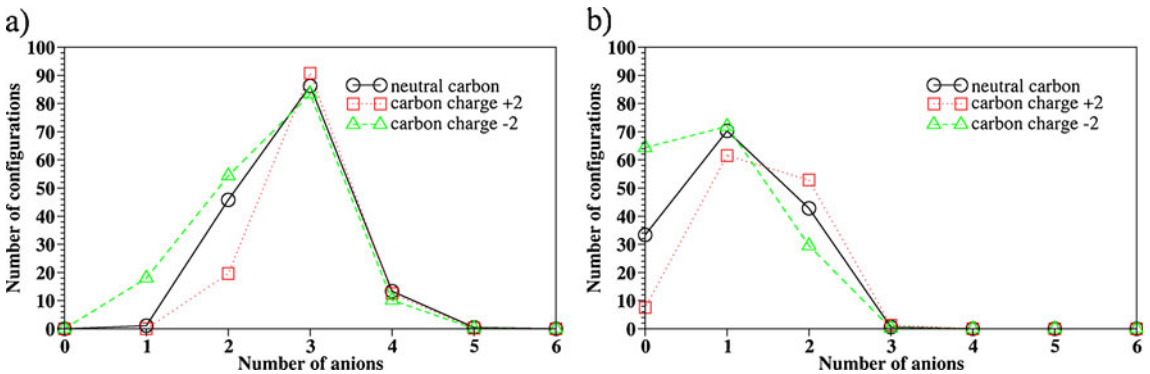



**Fig. 11** Distribution of anions on either side of the $EMI^+$ cation when the numbers are unequal. Distribution on the side with **a** more or **b** less anions

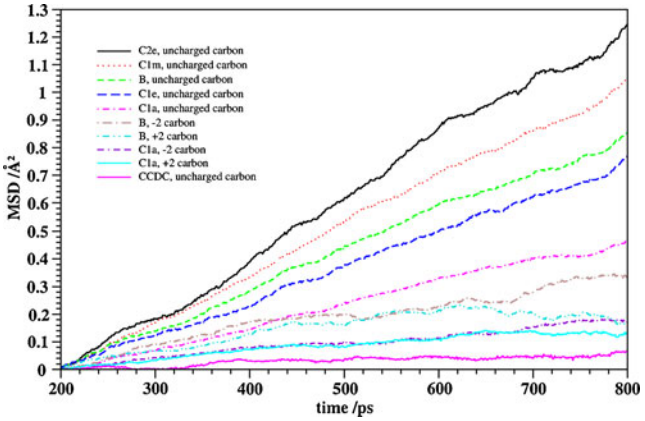



**Fig. 12** Mean-square-displacement functions for different species in the simulation boxes

$N_1$–$C_{1e}$ axis results in many energetically favorable positions for $C_{2e}$ group (visible as a dark blue ring in Fig. 10a), which, in turn, results in a less exact position for the fourth coordinating anion.

## Mirror symmetry of the cation

The force field of the cations is completely symmetrical with respect to the aromatic ring mirror plane, so the anions can be expected to coordinate equally to both sides of the cations. The side of the cation can be distinguished mathematically by the sign of the $C_{1a}$–$N_1$…$N_2$…B dihedral angles, so that positive numbers correspond to one side, and negative to the other. The distribution of dihedral angles is indeed symmetrical, and independent of the charge on carbon flakes. The total number of anions on each side of the cation is on average 2 on each side irrespective of charge.

However, this does not mean that any particular cation should have two anions on each side, since the local environment can be less symmetrical. Figure 11 illustrates the anionic distribution on either side of the cationic symmetry plane exclusively for situations where the number of ions on different sides is unequal. It is then visible that the most common uneven distribution of anions is to 3:1, with the more populated side having 2–4 anions, and the less populated having 0–2 anions. Charging the carbon flakes does not significantly affect the number of anions on the more populated side, but induces a significant redistribution of anions on the less populated side. In the positively charged system, the distribution becomes more symmetrical with less systems displaying unequal distribution, while the negatively charged system displays a less symmetrical distribution. For negative charging, there is also a clear increase in configurations with only one or none anions at one side of the cation. This asymmetry can be explained by the cations having one side against the carbon surface when it is negatively charged (see Fig. 8a), which means that less anions can coordinate on this side due to steric hindrance and unfavorable carbon-anion interactions.

## Dynamic properties

The mobility of the different species in the simulation boxes can be estimated by the mean-square-displacement function (see Fig. 12); the slopes of these curves are in principle proportional to the diffusion coefficients if the simulations are allowed to run into the diffusive range. It is clear that, in the system with neutral carbon flakes, the ethyl and methyl groups of the cations are the most mobile ionic components, followed by the anions. This is somewhat surprising considering the much smaller size of the anions, and must be explained by the rotations of part of the $EMI^+$ ions. The cation aromatic ring atoms display only half the mobility of the anions. The carbon flakes, in turn, show very low mobility and can be considered more or less static. This validates the approximation of neglecting the bonds between the CDC flakes in our model.

Charging the flakes by either −2e or +2e lowers the mobility of both cations and anions approximately three-fold. This is interesting, as there is little correspondence in the structural data discussed above. One reason for this observation might be that ions become more strongly coordinated to the immobile carbon species, but this is obviously not compensated by larger anion mobility for positive charge or larger cation mobility for negative charge. Rather, introducing charge on the carbon flakes leads to somewhat of a structural rearrangement according to the discussion above, but then tends to stabilize the entire IL matrix by the more dominating electrostatic interactions in the systems. This apparently affects all ions in the system, which could contribute to the small pore size. If the pore size were larger, the charged carbon would probably have less effect on the ionic movement in the middle of the pores, and thereby perhaps prolong the actuating response time.

## Actuator movement

The simulation cell volume increased by 1.5 $nm^3$ upon negative charging and decreased by 1.8 $nm^3$ upon positive charging, corresponding to +2 vol% per −2e charge on each carbon flake (Table 1). This volume change can be largely explained by the charge compensation of introducing more (large) $EMI^+$ ions during negative charge, and vice versa, considering the volumes of the two ionic types (87 $Å^3$ for $EMI^+$ and 10 $Å^3$ for $BF_4^-$). This indicates that the driving force for the actuation motion is mainly replacement of ions of different sizes during charging of the electrodes; however, it can also, to a minor degree, be considered that the volume change can be due to the movement of anions

into smaller pores (Fig. 5) during positive charging and vice versa during negative charging. This is indicated by the fact that it is primarily anions that rearrange during charging (see the $C_{CDC}$…B RDF in Fig. 6).

A typical actuator consists of two electrodes with $D_e$= 100 μm thickness, separated by a separator with D=10 μm thickness. If the same 2% volume change occurred in both electrodes of such an actuator, and without considering the elastic response of the separator, the bending radius would be:

$$R = \frac{D + D_e}{2 \times \frac{0.02}{3}} = 8.25mm \quad (3)$$

A volume change of this nature and size has also been recorded experimentally by Janno Torop [18].

## Conclusions

As the first MD study of a CDC-EMI-$BF_4$ artificial muscle material, this study aimed to describe and qualitatively assess the atomic-level processes responsible for the movement of the actuator. The proposed carbon flake model for CDC resulted in a realistic density value for the actuator (1.3 g/cm$^3$), although its pore size distribution below 8 Å diameter was more uniform than previous results for BET measurements of CDC. The modeled system was able to actuate when the carbon was charged, which could be observed as a volume change of the MD box. The volume expansion of the cathode and contraction of anode seems to be caused mostly by the replacement of smaller anions with significantly larger cations in the cathode, and vice versa in the anode. Rotation of cations from a parallel to a perpendicular configuration on the surface of the CDC flakes, and the redistribution of anions and cations between different pore sizes in the 3–8 Å region may also contribute to actuator movement. In the cation-anion clusters, the tridentate coordination of anions to the most concentrated positive charge locations at the cations is largely unaffected by CDC charging. In the cathode, however, the cation-anion coordination becomes asymmetric around the cation plane since only one side of a significant number of cations is accessible for the anions; the other cation side is being pulled tight against the CDC surface. This configuration of the EMI-CDC coordination upon negative charging could be significant, since it will lead to little steric hindrance for other ions to enter the larger pores, and thereby to possibilities for larger degrees of charge accumulation and subsequent actuation.

**Acknowledgments** This research has been supported by the European Structural Fund through the programme "DoRa", Estonian Science Foundation grant #8553 and financing from Ministry of Science of Education #SF0180008s08.

## References

1. Bar-Cohen Y, Leary SP, Shahinpoor M, Harrison JS, Smith JG (1999) Electroactive polymer (EAP) actuators for planetary applications. Proc SPIE 3669:57. doi:10.1117/12.349708
2. Bar-Cohen Y (2001) Transition of EAP material from novelty to practical applications: are we there yet? Proc SPIE Int Soc Opt Eng 4329:1–6. doi:10.1117/12.432639
3. Anton M, Punning A, Aabloo A, Listak M, Kruusmaa M (2004) Towards a biomimetic EAP robot. TAROS 2004:1–7. http://www.essex.ac.uk/csee/research/publications/technicalreports/2004/csm415/Anton_etal.pdf
4. Liu Y, Shi L, Liu L, Zhang Z, Leng J (2008) Inflated dielectric elastomer actuator for eyeball's movements: fabrication, analysis and experiments. Proc SPIE 6927:69271A. doi:10.1117/12.775946
5. Bar-Cohen Y et al (2004) Electroactive polymer (EAP) actuators as artificial muscles: reality, potential, and challenges, 2nd edn. SPIE Press Book, Bellingham, WA
6. Wang H, Li C (2009) A linear dielectric EAP actuator with large displacement output. ICMTMA'09 1:73–76. doi:10.1109/ICMTMA.2009.114
7. Fernández D, Moreno L, Baselga J (2005) A bio-inspired EAP actuator design methodology. Proc SPIE Int Soc Opt Eng 5759:361. doi:10.1117/12.599108
8. Hoefer M, Bandaru PR (2009) Determination and enhancement of the capacitance contributions in carbon nanotube based electrode systems. Appl Phys Lett 95:183108. doi:10.1063/1.3258353
9. Brett CMA, Brett AMO (1993) Electrochemistry: principles, methods, and applications. Oxford University Press, Oxford
10. Conway BE (1991) Transition from "supercapacitor" to "battery" behavior in electrochemical energy storage. J Electrochem Soc 138:1539–1548. doi:10.1149/1.2085829
11. Huang J, Sumpter BG, Meunier V (2008) A universal model for nanoporous carbon supercapacitors applicable to diverse pore regimes, carbon materials, and electrolytes. Chem Eur J 14:6614–6626. doi:10.1002/chem.200800639
12. Gogotsi Y, Nikitin A, Ye H, Zhou W, Fischer JE, Yi B, Foley HC, Barsoum MW (2003) Nanoporous carbide-derived carbon with tunable pore size. Nat Mater 2:591–594. doi:10.1038/nmat957
13. Armand M, Endres F, MacFarlane DR, Ohno H, Scrosati B (2009) Ionic-liquid materials for the electrochemical challenges of the future. Nat Mater 8:621–629. doi:10.1038/nmat2448
14. Carpi F, Smela E et al (2009) Biomedical applications of electroactive polymer actuators. Wiley, Chichester
15. Bennett MD, Leo DJ (2004) Ionic liquids as stable solvents for ionic polymer transducers. Sens Actuators A 115:79–90. doi:10.1016/j.sna.2004.03.043
16. Akle BJ, Bennett MD, Leo DJ, Wiles KB, Mcgrath JE (2007) Direct assembly process: a novel fabrication technique for large strain ionic polymer transducers. J Mater Sci 42:7031–7041. doi:10.1007/s10853-006-0632-4
17. Palmre V, Brandell D, Mäeorg U, Torop J, Volobujeva O, Punning A, Johanson U, Kruusmaa M, Aabloo A (2009) Nanoporous carbon-based electrodes for high strain ionomeric bending actuators. Smart Mater Struct 18:095028. doi:10.1088/0964-1726/18/9/095028
18. Torop J, Arulepp M, Leis J, Punning A, Johanson U, Palmre V, Aabloo A (2010) Nanoporous carbide-derived carbon material-based linear actuators. Materials 3:9–25. doi:10.3390/ma3010009
19. Fukushima T, Asaka K, Kosaka A, Aida T (2005) Fully plastic actuator through layer-by-layer casting with ionic-liquid-based bucky gel. Angew Chem Int Edn 44:2410–2413. doi:10.1002/anie.200462318
20. Asaka K, Mukai K, Takeuchi I, Kiyohara K, Sugino T, Terasawa N, Hata K (2008) Fast fully plastic actuator based on ionic-liquid-based bucky gel. Proc SPIE 7037:703710. doi:10.1117/12.794503

21. Baughman RH, Cui C, Zakhidov AA, Iqbal Z, Barisci JN, Spinks GM, Wallace GG, Mazzoldi A, De Rossi D, Rinzler AG, Jaschinski O, Roth S, Kertesz M (1999) Carbon nanotube actuators. Sci 284:1340–1344. doi:10.1126/science.284.5418.1340
22. Kiyohara K, Sugino T, Takeuchi I, Mukai K, Asaka K (2009) Expansion and contraction of polymer electrodes under applied voltage. J Appl Phys 105:063506. doi:10.1063/1.3078031
23. Baughman RH (1996) Conducting polymer artificial muscles. Synth Met 78:339–353. doi:10.1016/0379-6779(96)80158-5
24. Asaka K, Oguro K, Nishimura Y, Mizuhata M, Takenaka H (1995) Bending of polyelectrolyte membrane–platinum composites by electric stimuli I. Response characteristics to various waveforms. Polym J 27:436–440. doi:10.1295/polymj.27.436
25. Oren Y, Glatt I, Livnat A, Kafari O, Soffer A (1985) The electrical double layer charge and associated dimensional changes of high surface area electrodes as detected by more deflectometry. J Electroanal Chem Interfacial Electrochem 187:59–71. doi:10.1016/0368-1874(85)85575-1
26. Ghosh A, Wang CS, Kofinas P (2010) Block copolymer solid battery electrolyte with high Li-ion transference number. J Electrochem Soc 157:A846–A849. doi:10.1149/1.3428710
27. Dias FB, Plomp L (2000) Trends in polymer electrolytes for secondary lithium batteries. J Power Sources 88:169–191. doi:10.1016/S0378-7753(99)00529-7
28. Ahmad S, Agnihotry SA (2009) Effect of nano $\gamma$-$Al_2O_3$ addition on ion dynamics in polymer electrolytes. Curr Appl Phys 9:108–114. doi:10.1016/j.cap.2007.12.003
29. Karo J, Brandell D (2009) A molecular dynamics study of the influence of side-chain length and spacing on lithium mobility in non-crystalline $LiPF_6{\cdot}PEO_x$; x=10 and 30. Solid State Ionics 180:1272–1284. doi:10.1016/j.ssi.2009.07.009
30. Kasemägi H, Klintenberg M, Aabloo A, Thomas JO (2002) Molecular dynamics simulation of the $LiBF_4$–PEO system containing $Al_2O_3$ nanoparticles. Solid State Ionics 147:367–375. doi:10.1016/S0167-2738(02)00013-9
31. Kasemägi H, Klintenberg M, Aabloo A, Thomas JO (2003) Molecular dynamics simulation of temperature and concentration dependence of the 'filler' effect for the LiCl/PEO/$Al_2O_3$-nanoparticle system. Electrochimica Acta 48:2273–2278. doi:10.1016/S0013-4686(03)00214-7
32. Borodin O, Smith GD (1998) Molecular dynamics simulations of poly(ethylene oxide)/LiI melts. 1. Structural and conformational properties. Macromolecules 31:8396–8406. doi:10.1021/ma980838v
33. Borodin O, Smith GD (2000) Molecular dynamics simulations of poly(ethylene oxide)/LiI melts. 2. Dynamic properties. Macromolecules 33:2273–2283. doi:10.1021/ma991429h
34. Brandell D, Liivat A, Aabloo A, Thomas JO (2005) Conduction mechanisms in crystalline $LiPF_6{\cdot}PEO_6$ doped with $SiF_6^{2-}$ and $SF_6$. Chem Mater 17:3673–3680. doi:10.1021/cm0505401
35. Liivat A, Brandell D, Aabloo A, Thomas JO (2007) A molecular dynamics study of short-chain ordering in crystalline $LiPF_6{\cdot}PEO_6$. Polymer 48:6448–6456. doi:10.1016/j.polymer.2007.08.051
36. Soolo E, Brandell D, Liivat A, Kasemägi H, Tamm T, Aabloo A (2010) Force field generation and molecular dynamics simulations of $Li^+$–Nafion. Electrochim Acta 55:2587–2591. doi:10.1016/j.electacta.2009.03.039
37. Brandell D, Kasemägi H, Aabloo A (2010) Poly(ethylene oxide)–poly(butadiene) interpenetrated networks as electroactive polymers for actuators: a molecular dynamics study. Electrochim Acta 55:1333–1337. doi:10.1016/j.electacta.2009.04.070
38. Rapaport DC (2004) The art of molecular dynamics simulation, 2nd ed. Cambridge University Press. ISBN: 0521825687
39. Palmer JC, Llobet A, Yeon S-H, Fischer JE, Shi Y, Gogotsi Y, Gubbins KE (2010) Modeling the structural evolution of carbide-derived carbons using quenched molecular dynamics. Carbon 48:1116–1123. doi:10.1016/j.carbon.2009.11.033
40. Arulepp M, Leis J, Lätt M, Miller F, Rumma K, Lust E, Burke AF (2006) The advanced carbide-derived carbon based supercapacitor. J Power Sources 162:1460–1466. doi:10.1016/j.jpowsour.2006.08.014
41. Smith W, Forester T, The DL_POLY project. Daresbury Laboratory, Daresbury, Warrington, UK
42. Soolo E, Karo J, Kasemägi H, Kruusmaa M, Aabloo A (2006) Application of the Monte Carlo method for creation of initial models of EAP molecules for molecular dynamics simulation. Proc SPIE 6168:61682A. doi:10.1117/12.658664
43. Stewart JJP (2007) Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. J Mol Model 13:1173–1213. doi:10.1007/s00894-007-0233-4
44. Walther JH, Jaffe R, Halicioglu T, Koumoutsakos P (2001) Carbonnanotubes in water: structural characteristics and energetics. J Phys Chem B 105:9980–9987. doi:10.1021/jp011344u
45. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM Jr, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 117:5179–5197. doi:10.1021/ja00124a002
46. Lopes JNC, Deschamps J, Pádua AAH (2004) Modeling ionic liquids using a systematic all-atom force field. J Phys Chem B 108:2038–2047. doi:10.1021/jp0362133
47. Yushin G, Dash R, Jagiello J, Fischer JE, Gogotsi Y (2006) Carbide-derived carbons: effect of pore size on hydrogen uptake and heat of adsorption. Adv Funct Mater 16:2288–2293. doi:10.1002/adfm.200500830

J Mol Model (2012) 18:1553–1560
DOI 10.1007/s00894-011-1189-y

ORIGINAL PAPER

# Molecular modeling of cytochrome $b_5$ with a single cytochrome *c*-like thioether linkage

**Ying-Wu Lin · Yi-Mou Wu · Li-Fu Liao · Chang-Ming Nie**

Received: 8 June 2011 / Accepted: 14 July 2011 / Published online: 30 July 2011


**Abstract** Bovine liver cytochrome $b_5$ (cyt $b_5$), with heme bound noncovalently, has been converted into a cyt *c*-like protein (cyt $b_5$ N57C) by constructing a thioether linkage between the heme and the engineered cysteine residue. With no X-ray or NMR structure available, we herein performed a molecular modeling study of cyt $b_5$ N57C. On the other hand, using amino acid sequence information for a newly discovered member of the cyt $b_5$ family, domestic silkworm cyt $b_5$ (DS cyt $b_5$), we predicted the protein structure by homology modeling in combination with MD simulation. The modeling structure shows that both Cys57 in cyt $b_5$ N57C, and Cys56, a naturally occurring cysteine in DS cyt $b_5$, have suitable orientations to form a thioether bond with the heme 4-vinyl group, as the heme is in orientation A. In addition to providing structural information that was not previously obtained experimentally, these modeling studies provide insight into the formation of cyt *c*-like thioether linkages in cytochromes, and suggest that *c*-type cyt $b_5$ maturation involves a *b*-type intermediate.

**Keywords** Heme proteins · Covalent bond · Homology modeling · Molecular dynamics simulation · Intermediate

**Electronic supplementary material** The online version of this article (doi:10.1007/s00894-011-1189-y) contains supplementary material, which is available to authorized users.

Y.-W. Lin (✉) · L.-F. Liao · C.-M. Nie
School of Chemistry and Chemical Engineering,
University of South China,
Hengyang 421001, China
e-mail: linlinying@hotmail.com

Y.-W. Lin
State Key Laboratory of Coordination Chemistry,
Nanjing University,
Nanjing 210093, China

Y.-M. Wu
Institute of Pathogenic Biology, University of South China,
Hengyang 421001, China

## Introduction

Both *b*- and *c*-type cytochromes contain the same heme group, but the heme group is associated with the protein noncovalently in *b*-type but covalently in *c*-type cytochromes. For example, the heme group is coordinated to two axial histidine ligands, His39 and His63, in cytochrome $b_5$ (cyt $b_5$) (Fig. 1a, PDB entry 1cyo [1]), and is covalently attached to the polypeptide in cytochrome *c* (cyt *c*) through two thioether linkages between the heme vinyl groups and the thiols of cysteine residues in a typical Cys-Xaa-Xaa-His (CXXCH) motif (Fig. 1b, PDB entry 1hrc [2]). Recently, the structure of *c*-type cytochromes with the heme attached to an (A/F) XXCH motif through only one thioether linkage was elucidated (Fig. 1c, PDB entry 2w9k [3]). Moreover, an atypical heme was discovered in the cytochrome $b_6f$ complex, which is covalently bound by one thioether linkage and has no axial amino acid ligand (Fig. 1d and e, PDB entries 1q90 [4] and 1vf5 [5], respectively).

The characteristic that distinguishes *b*-type and *c*-type cytochromes has attracted much attention over the last two decades in relation to the mechanisms of cyt *c* biogenesis [6–11]. Several groups [12–15], including ours [16], showed that *b*-type cytochromes such as cyt $b_5$ and cyt $b_{562}$ can be converted into cyt *c*-like proteins by introducing cysteine residues close to the heme vinyl groups in cyt $b_5$ (Asn57 and/or Ser71 to Cys mutation) [12, 16], and by creating an artificial CXXCH motif in cyt $b_{562}$ (Arg98 and Tyr101 to Cys mutation) [13–15]. The X-ray crystal structure of *c*-type cyt $b_{562}$, R98C/Y101C cyt $b_{562}$, was resolved recently by Gray and co-workers [15]. In addition,

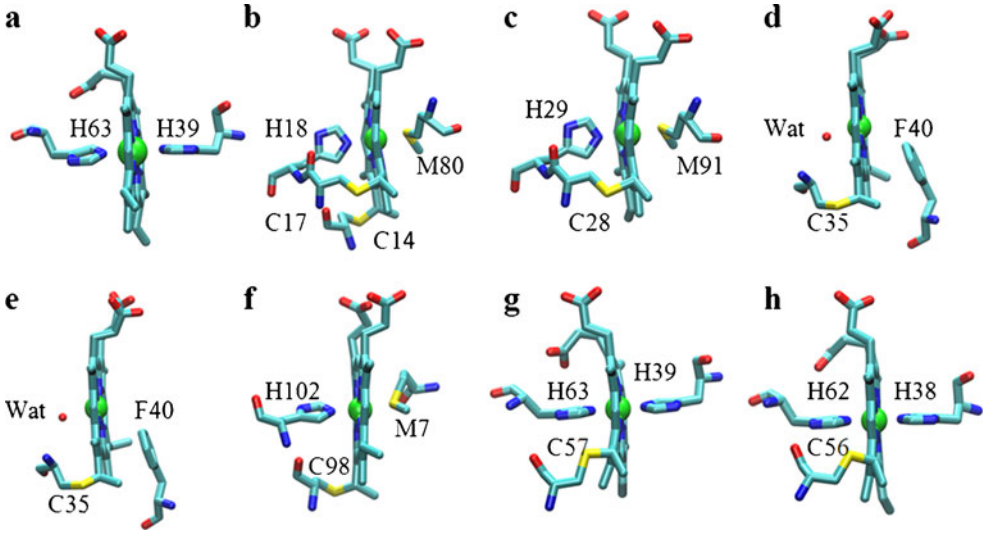



**Fig. 1** Heme active sites in cytochromes investigated by molecular modeling in this study, showing coordination interactions and/or covalent thioether bond(s): **a** bovine liver cyt $b_5$ (PDB entry 1cyo), **b** horse heart cyt *c* (PDB entry 1hrc), **c** *C. fasciculata* mitochondrial cyt *c* (PDB entry 2w9k), **d** cyt $b_6f$ complex (PDB entry 1q90, **e** cyt $b_6f$ complex (PDB entry 1vf5), **f** cyt $b_{562}$ R98C (PDB entry 1qq3), **g** bovine cyt $b_5$ N57C, and **h** silkworm cyt $b_5$

a solution structure of *c*-type cyt $b_{562}$ with a single thioether linkage, R98C cyt $b_{562}$, was also well resolved (Fig. 1f, PDB entry 1qq3 [14]). On the other hand, no X-ray or NMR structure is currently available for cyt $b_5$ with a cyt *c*-like thioether linkage.

Moreover, although cyt $b_5$ from vertebrates, such as bovine microsomal cyt $b_5$ (Mc cyt $b_5$) [1, 17] and rat outer mitochondrial cyt $b_5$ (OM cyt $b_5$) [18], have been studied in depth from a structural perspective, their counterparts from insects are rather poorly investigated. To date, only one cyt $b_5$ from an insect—house fly cyt $b_5$ (HF cyt $b_5$)—has been characterized structurally, by Benson and co-workers [19]. For other insect cyt $b_5$, such as domestic silkworm cyt $b_5$ (DS cyt $b_5$), the amino acid sequence was only recently identified (GenBank ABF51280.1). There was an interesting finding from the sequence alignment, which showed that DS cyt $b_5$ has a naturally occurring cysteine residue at position 56, corresponding to Asn57 in bovine Mc cyt $b_5$ (Fig. 2). Hence, Cys56 is likely to form a cyt *c*-like thioether bond with the heme vinyl group if DS cyt $b_5$ has a protein fold similar to that of bovine cyt $b_5$ N57C.

Due to the recent strong growth in available computing power, molecular modeling and molecular dynamics (MD) simulations have become powerful tools for obtaining atomic structural information on proteins that is difficult to obtain experimentally [20–23]. To date, plenty of theoretical studies have been performed for cyt $b_5$ [24–32], cyt $b_{562}$ [33–36], and cyt *c* [37–39]; these studies have aimed to gain structural insights, as well as to elucidate the mechanisms of holoprotein maturation. Using molecular modeling in combination with MD simulations, we recently studied the structural consequences of introducing a histidine tag, the $(His)_6$ tag, to the N-terminus of bovine Mc cyt $b_5$ [32].

In this study, in order to gain insight into *b*- to *c*-type heme conversion and obtain structural information that has not been obtained experimentally, we use molecular modeling to investigate the structural consequences of covalently attaching the heme to the protein polypeptide chain via just one thioether linkage, as engineered in the bovine cyt $b_5$ variant N57C [12, 16]. Additionally, we predict the structure of the newly discovered member of the cyt $b_5$ cytochromes, DS cyt $b_5$, based on homology modeling and MD simulation in which a single cyt *c*-like thioether is modeled. The implications of the modeling studies are also discussed.

```
Bovine Cyt b5      AEESSKAVKYYTLEEIQKHNNSKSTWLILHYKVYDLTKFLEEHPGGEEVLREQAGGDATEN 61
House Fly Cyt b5   --MSSEDVKYFTRAEVAKNNTKDKNWFIIHNNVYDVTAFLNEHPGGEEVLIEQAGKDATEH 59
Silkworm Cyt b5    -----MSVHKYTRAEVAARDNNKQNLIIIDNVVYDVAAFLEDHPGGTEVLVDNAGSDASEC 56
                         *: :*  *:  .:..... :*:.  ***:: **::**** *** ::** **:**

Bovine Cyt b5      FEDVGHSTDARELSKTFIIGELHPDDRSKITKPSESIITTIDS---------------------------- 104
House Fly Cyt b5   FEDVGHSSDAREMMKQYKVGELVAEERSNVPEK------------------------------------- 92
Silkworm Cyt b5    FHEVGHSEIAIEWRNTFKVGEIVDEEKLEVKCKQPSAAESAEPLTLGGLLAVYGPPVAMAVLAYLLYTFLFG 128
                   .:****  * *  : : :**:  ::: ::
```

**Fig. 2** Sequence alignment of bovine, house fly and silkworm cyt $b_5$ by the program Clustal. *Asterisks*, *colons*, and *dots* indicate highly, moderately, and poorly conserved residues, respectively. The *italicized residues* at the N- and C-termini of bovine and house fly cyt $b_5$ were not observed in the X-ray diffraction data. The *italicized residues* at the C-terminus of silkworm cyt $b_5$ were not modeled in this study

## Methods

### Molecular modeling of bovine cyt $b_5$ N57C

The initial structure of bovine liver microsomal cyt $b_5$ was taken from the NMR structure (PDB entry 1hko [17]; model 1, a fragment of Ser6–Ser98 corresponding to lipase proteolysis with 93 residues, was selected). The cyt $b_5$ variant N57C was generated using the program VMD 1.8.7 (visual molecular dynamics) [40] by replacing Asn57 (i.e., Asn61 in full sequence numbering, see Fig. 2) with a Cys residue. A cyt *c*-like thioether linkage between Cys57 and the heme 4-vinyl group was constructed by defining a new residue of HOBC for the oxidized heme with bis–His coordination, as well as one thioether bond as in cyt *c* (see the "Electronic supplementary material," ESM). The classical force field of CHARMM27 was used for MD simulation. psfgen of the program NAMD 2.7 (nanoscale molecular dynamics) [41] was used to add hydrogen atoms and assign charges to the cyt $b_5$ N57C at pH 7.0. The protein was then solvated in a cubic box (size: about 67 Å×65 Å×60 Å) of TIP3 water s with periodic boundary conditions. The box was at least 15 Å from any given protein atom. The protein was then neutralized by adding a total of 21 counterions (14 $Na^+$ and 7 $Cl^-$) using the autoionize plug-in of VMD 1.8.7, resulting in a physiologic ionic strength of 0.15 M.

The resulting system was minimized with NAMD using 5,000 minimization steps with the conjugate gradient method, and equilibrated for 10,000 molecular dynamics steps (1 fs per step) via an NVT ensemble (where the number of particles *N*, the volume *V*, and the temperature *T* of the system were kept constant) at 300 K, a procedure similar to modeling a rationally designed nitric oxide reductase based on myoglobin [42]. The system was further minimized for 5,000 steps before analyzing it with VMD 1.8.7. An atom-based cutoff of 14 Å with switching at 12 Å was used for nonbonded van der Waals interactions. No constraints were employed for the whole system during the simulation. The modeling structure was evaluated by performing Verify3D [43, 44] and Ramachandran plot [45] analysis.

### Molecular modeling of domestic silkworm cyt $b_5$

The initial structure of domestic silkworm cyt $b_5$, DS cyt $b_5$, was generated with SWISS-MODEL Workspace [46, 47], a web-based environment for protein structure homology modeling, using the NMR structure of bovine cyt $b_5$ (PDB entry 1hko [17] model 1) as a template with a sequence identity of 42.35%. The modeled DS cyt $b_5$ is an apoprotein that contains only the polypeptide chain from 2 to 86. The heme group, the first Met residue, and the C-terminal polypeptide of KCKQPS were then added using VMD 1.8.7 based on the coordination of bovine cyt $b_5$. The resultant holoprotein of DS cyt $b_5$ contains 92 residues, with a C-terminus truncated at a conserved serine residue, Ser92, corresponding to Ser98 in cyt $b_5$ N57C. A thioether linkage between Cys56 and the heme 4-vinyl group was then constructed in DS cyt $b_5$ using the same method as used for the cyt $b_5$ N57C.

After adding hydrogen, TIP3 water, and 20 counterions (15 $Na^+$ and 5 $Cl^-$), in order to achieve complete relaxation of the entire protein, the resultant system was equilibrated using the procedure involving ten time steps used for cyt $b_5$ N57C: 50,000 minimization steps, followed by 100,000 molecular dynamics steps (1 fs per step) via an NVT ensemble at 300 K, and a further 50,000 minimization steps, as described recently for modeling a low-spin myoglobin variant with a bis–His heme coordination [23]. For further model refinement, the resulting DS cyt $b_5$ system was subjected to a MD simulation with the temperature held at 300 K for 2 ns, as done when modeling a histidine-tagged cyt $b_5$ [32], with Langevin dynamics used for constant temperature control. Coordinates were saved every 1,000 steps (corresponding to 1 ps) to produce trajectory files containing 2,000 configurations of DS cyt $b_5$. Visualization and data analysis were done with VMD 1.8.7. Sequence alignment was performed using the Clustal 2.1 program server (http://www.ebi.ac.uk/Tools/clustalw2/) [48].

## Results and discussion

### Modeling the structure of bovine cyt $b_5$ N57C

The overall modeling structure of the cyt $b_5$ variant N57C was based on an initial NMR structure of bovine cyt $b_5$ in solution, and is shown in Fig. 3a. The heme group with two axial histidine ligands, His39 and His63, as well as one thioether linkage between Cys57 and the heme 4-vinyl group are highlighted (also shown in Fig. 1g). The structure shows that the helical structure of α4 is maintained near the mutation of Cys57 without any distortion (Fig. 3a), as in R98C cyt $b_{562}$ [15]. In addition, cyt $b_5$ N57C has the same overall structure as cyt $b_5$, and overlaps well with the initial NMR structure (Fig. 3b), which agrees with the experimental observation that cyt $b_5$ N57C has a similar circular dichroism (CD) spectrum to that of cyt $b_5$ (Fig. S1). These observations suggest that Cys57 is suitable for thioether bond formation in cyt $b_5$, and the mutation from Asn to Cys at position 57 is a rational design [12, 16].

It is interesting to observe that the two histidine imidazole rings are rotated such that they are nearly parallel to each other in cyt $b_5$ N57C, as observed previously in the modeling of a histidine-tagged cyt $b_5$, $(His)_6$-cyt $b_5$, as well as in a control simulation of cyt $b_5$ [32]. The resulting conformation overlaps

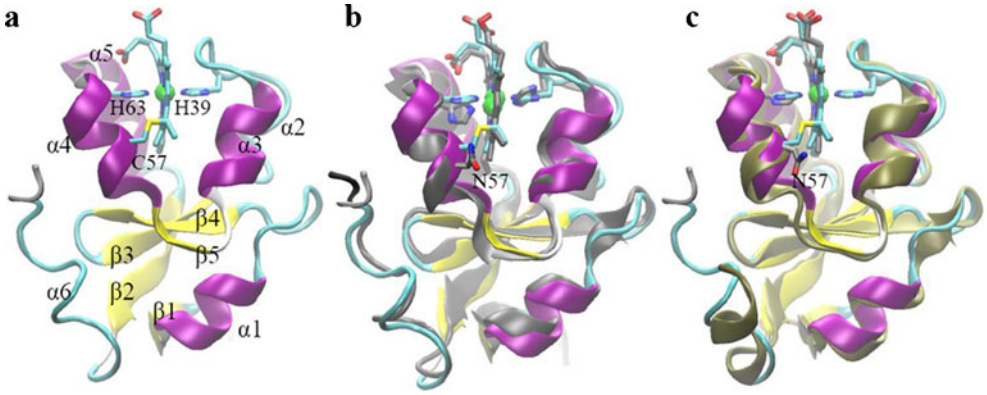



**Fig. 3** Modeling structure of bovine cyt $b_5$ N57C (**a**), and the overlay between the initial NMR structure (PDB entry 1hko, *gray*) (**b**) and the X-ray structure (PDB entry 1cyo, *ochre*) (**c**) of bovine cyt $b_5$, respectively. The heme group with two axial histidine ligands, His39 and His63, the thioether bond with Cys57, and the native Asn57 are highlighted. The elements of secondary structure, α1–α6 and β1–β5, are also labeled for clarity

well with the crystal structure of bovine cyt $b_5$ (PDB entry 1cyo) (Fig. 1c). This is very interesting, since although molecular modeling was performed in solution using the NMR structure as an initial structure, the resulting structure is closer to the crystal structure for the heme active site.

The quality of the modeling structure of cyt $b_5$ N57C was evaluated using Verify3D, a tool designed to analyze the compatibility of an atomic model (3D) with its own amino acid sequence (1D) [43, 44]. As shown in Fig. 4, all residues of cyt $b_5$ N57C have scores higher than 0.3, with 81% of residues having scores higher than 0.4. The profile is similar to that of both X-ray and NMR structures, indicating that cyt $b_5$ N57C is a valid, high-quality model. In addition, residues after position 60 in cyt $b_5$ N57C have scores higher than that of the X-ray structure, suggesting that cyt $b_5$ N57C adopts a more favored conformation in this region, as determined by NMR in the solution state.

The Ramachandran plot shows that ~78% of the residues of cyt $b_5$ N57C are in the favored region (shaded in blue) and allowed regions (shaded in green) (Fig. S2a), a similar value to those observed for the X-ray structure (~83%) (Fig. S2b) and NMR structure (~80%) (Fig. S2c). Specifically, Cys57 is found in the favored region, as indicated in Fig. 5. In addition, it has a more favored conformation in comparison to other cysteines that form thioether bonds with heme vinyl groups (Fig. 1b–f), such as Cys14 and Cys17 in horse heart cyt *c* (PDB entry 1hrc) [2], Cys28 in *C. fasciculata* mitochondrial cyt *c* (PDB entry 2w9k [3]), Cys35 in the cyt $b_6f$ complex (PDB entries 1q90 [4] and 1vf5 [5]), and Cys98 in cyt $b_{562}$ R98C (PDB entry 1qq3 [14]) (Fig. 5).

The calculated thioether bond length for Cys57–heme in cyt $b_5$ N57C was found to be 1.83 Å, within the range of C–S bond lengths (1.75–1.98 Å) in cyt *c* with one or two thioether linkages, as determined by NMR and X-ray crystallography (Table 1). Moreover, the C–S–C angle of the thioether bond in cyt $b_5$ N57C (108.6°) is close to that determined for *c*-type cytochromes (100.8–106.9°). These observations suggest that Cys57 has an ideal orientation in cyt $b_5$ N57C that allows for the spontaneous formation of a cyt *c*-like thioether bond without requiring the classic CXXCH heme-binding motif of cyt *c* [12, 16].

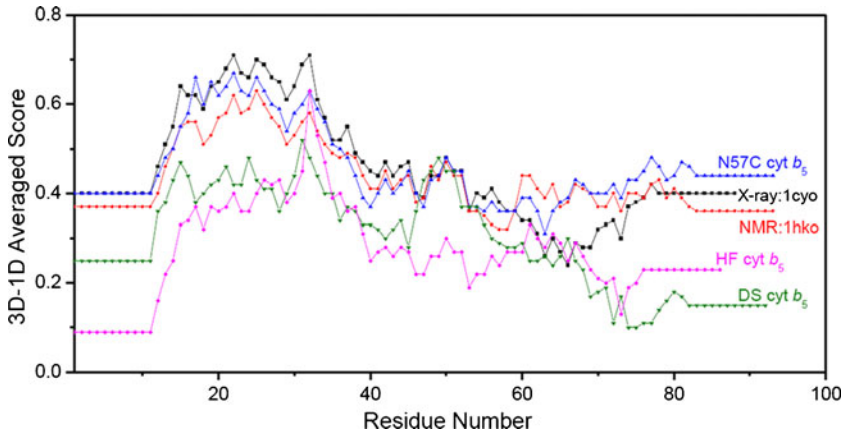



**Fig. 4** The verified 3D profiles obtained from the modeling of bovine cyt $b_5$ N57C and DS cyt $b_5$, with the X-ray and NMR structures of bovine cyt $b_5$ and HF cyt $b_5$ shown for comparison. The *vertical axis* represents the average 3D–-1D profile score for residues in a 21-residue sliding window (from −10 to +10). A quality score of greater than 0.1 indicates a valid structural model with a correct fold, whereas a higher score indicates a more accurate model. Scores for the first nine and the final nine sequence positions have no meaning [43, 44]

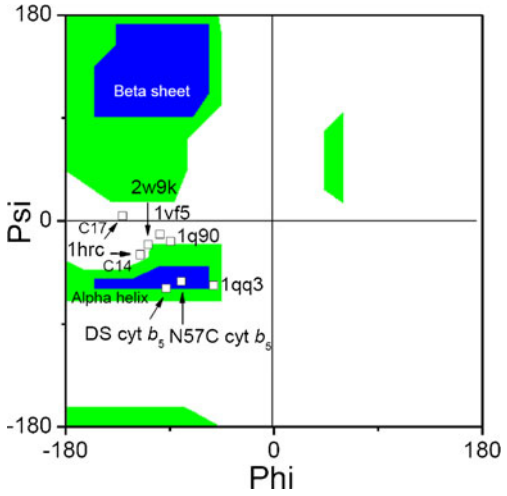



**Fig. 5** Ramachandran plots of the cysteine residues that form a thioether bond with the heme vinyl group in the modeling structures of bovine cyt $b_5$ N57C and DS cyt $b_5$, as indicated by *arrows*. Plots of horse heart cyt *c* (PDB entry 1hrc), *C. fasciculata* mitochondrial cyt *c* (PDB entry 2w9k), the cyt $b_6f$ complex (PDB entries 1q90 and vf5), and cyt $b_{562}$ R98C (PDB entry 1qq3) are also shown for comparison

## Modeling the structure of DS cyt $b_5$

Using homology modeling in combination with MD simulation, the structure of DS cyt $b_5$ containing 92 residues was predicted, and is shown in Fig. 6a. The heme group with two axial histidine ligands and the modeled thioether bond between Cys56 and the heme 4-vinyl group are highlighted (also shown in Fig. 1h). DS cyt $b_5$ adopts a structure with six helices (α1–α6) and five strands (β1–β5); in particular, the C-terminal α6 folds very well (Fig. 6a). Also, the overall structure of DS cyt $b_5$ overlaps well with those of bovine cyt $b_5$ N57C (Fig. 6b) and HF cyt $b_5$ (Fig. 6c), respectively. Helix α4, where Cys56 is located, shifts slightly towards the heme group with respect to that in cyt $b_5$ N57C. As a result, the heme group is flat in DS cyt $b_5$ but slightly bent in cyt $b_5$ N57C. Interestingly, the loop connecting α1 and β4 adopts an open conformation in DS cyt $b_5$, whereas a closed conformation is seen in cyt $b_5$ N57C and HF cyt $b_5$ (Fig. 6b and c). Previous studies also showed that an open–closed cleft transition occurs in this region for bovine and other cyt $b_5$ species [24, 26, 28].

Note that although sequence alignment shows that there is another cysteine, Cys88, in the C-terminus of DS cyt $b_5$ (Fig. 2), the modeling structure shows that Cys88 is located at the end of α6 and is remote from Cys56, with a separation of 20.4 Å between the two sulfur atoms (Fig. 6a). There is therefore little chance that these two cysteine residues form a disulfide bond. On the other hand, Cys88 likely functions as an α6 $C_{cap}$ in DS cyt $b_5$, a role assigned to Cys84 in frog OM cyt $b_5$ species [19]. A previous study also found that Cys, a nonpolar residue, is the fifth most overrepresented residue at the $C_{cap}$ position in short $3_{10}$ helices [49].

When compared to the X-ray and NMR structures of cyt $b_5$ N57C and bovine cyt $b_5$, Verify3D analysis of DS cyt $b_5$ showed that all of the residues had lower but satisfactory scores, with 74% of the residues having scores of >0.2 (Fig. 4). Despite this observation, it is interesting to find that DS cyt $b_5$ has similar scores to that of the X-ray structure of HF cyt $b_5$. This implies that both DS cyt $b_5$ and HF cyt $b_5$ from insects show similar folding behavior, thus resulting in a good overlap between them through spatial alignment (Fig. 6c). Moreover, the Ramachandran plot shows that ~81% of the residues of DS cyt $b_5$ are in the favored and allowed regions (Fig. S2d), similar to the results seen in the plot of cyt $b_5$ N57C (~78%). Among these residues, Cys56 was found between the favored and allowed regions (Fig. 5), indicating that it has a suitable conformation for the spontaneous formation of a thioether bond with the heme 4-vinyl group. Detailed analysis further shows that the modeled thioether bond has a length of 1.82 Å with a C–S–C angle of 107.7°, in agreement with experimental observations (Table 1).

In order to evaluate the structural stability of DS cyt $b_5$, the time evolutions of both the residue $C_\alpha$ root mean square deviation (RMSD) and the secondary structure were monitored during a 2 ns MD simulation at 300 K. As shown in Fig. 7a, the RMSD fluctuates around 1.2 Å after 1 ns of simulation, indicating that stable conformation equilibration has been achieved for this small system. Figure 7b shows that both the six helices (α1–α6) and five

**Table 1** Comparison of the C–S bond lengths and C–S–C angles in cyt *c* species reported in the literature and the *c*-type cyt $b_5$ species simulated in this study

| Cytochromes | C–S bond length (Å) | C–S–C angle (°) | PDB entry | Reference |
|---|---|---|---|---|
| Horse heart cyt *c* | C-S(Cys14) 1.86<br>C-S(Cys17) 1.75 | 106.9<br>102.4 | 1hrc | [2] |
| Cyt $b_6f$ complex | 1.80 | 101.6 | 1q90 | [3] |
| Cyt $b_6f$ complex | 1.79 | 100.8 | 1vf5 | [4] |
| *C. fasciculata* mitochondrial cyt *c* | 1.98 | 105.8 | 2w9k | [5] |
| Cyt $b_{562}$ R98C | 1.78 | 101.1 | 1qq3 | [14] |
| Cyt $b_5$ N57C | 1.83 | 108.6 | / | This work |
| DS cyt $b_5$ | 1.82 | 107.7 | / | This work |

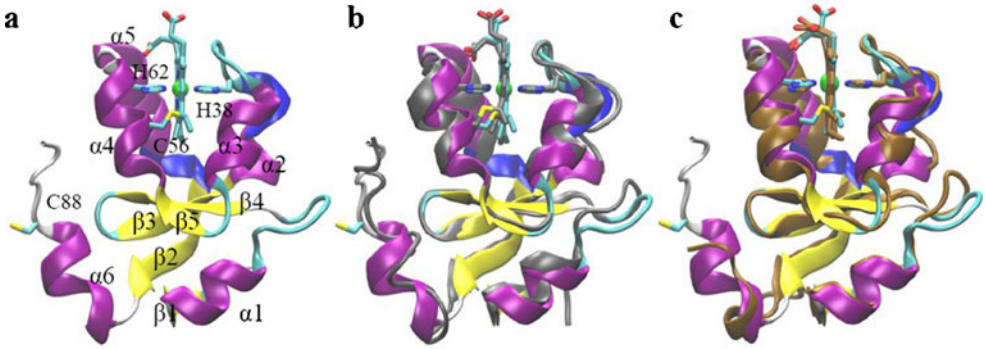



**Fig. 6** Modeling structure of DS cyt $b_5$ (**a**), and the overlay between the modeling structure of bovine cyt $b_5$ N57C (**b**, *gray*) and the X-ray structure of HF cyt $b_5$ (PDB entry 2ibj) (**c**, *orange*). The heme group with two axial histidine ligands, His38 and His62, the thioether bond with Cys56, and another Cys88 in the C-terminus are highlighted. The elements of secondary structure, α1–α6 and β1–β5, are also labeled for clarity

strands (β1–β5) are well maintained throughout the time course of the simulation, indicating that no dramatic structural changes occurred, which in turn suggests the acceptance of the predicted structure of DS cyt $b_5$. In addition, two $3_{10}$ helices adjacent to the C-terminus, α2 and α5, as shown in blue, also largely retain their conformations throughout the MD simulation. Note that the $3_{10}$-helix conformation is not observed for the modeling structure of cyt $b_5$ N57C (Fig. 3a). Meanwhile, in the X-ray structures of bovine and HF cyt $b_5$, the $3_{10}$ helix observed is in the C-terminus, and is defined as α6 [1, 19]. On the other hand, several $3_{10}$ helices were transitionally formed in the simulation of apo-cyt $b_5$ as well as its variants, and these were observed in the regions adjacent to the C-terminus of α2 and α5 [30], just as in DS cyt $b_5$.

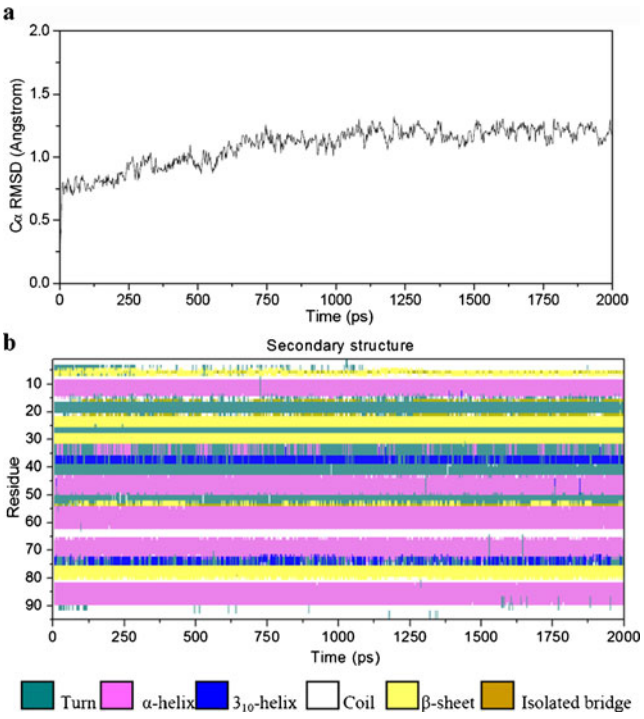



**Fig. 7** Time evolutions of **a** the $C_\alpha$ RMSD and **b** the secondary structure at 300 K of DS cyt $b_5$ for 2 ns. The different secondary elements are presented in a color-coded format, and the key to this format is shown at the bottom of **b**

## Implications for the formation of cyt *c*-like thioether linkages

As reviewed very recently by Stevens [50], the covalent attachment of heme to *c*-type cytochromes is performed by four known post-translational modification protein systems. Nevertheless, it is also possible to form holo-cyt *c* species via *b*-type intermediates spontaneously by reconstituting heme to apo-cyt *c* species in vitro in the absence of any biosynthesis proteins [51, 52]. Additionally, the spontaneous formation of a single thioether bond can occur in vitro, as apo-cyt *c* species contain either the AXXCH or CXXAH motif [53]. This method has also been successfully applied to the design and synthesis of de novo cyt *c* species [54].

On the other hand, *b*-type cytochromes can be converted into *c*-type cytochromes, as exemplified by cyt $b_5$ and cyt $b_{562}$ [12–16]. Unlike *c*-type cyt $b_{562}$, in which an artificial CXXCH (R98C/Y101C/H102) or CXXYH (R98C/H102) motif was created, there is no conserved heme-binding motif in *c*-type cyt $b_5$ species [12, 16]. The modeling structure of cyt $b_5$ N57C shows that the introduced Cys57 has an orientation that is suitable for forming a thioether bond with the heme 4-vinyl group (Figs. 1g and 3a), whereby the typical CXXCH motif is not essential. It has been suggested that the formation of cyt *c*-like cyt $b_5$ involves a *b*-type cytochrome intermediate [16], as proposed for cyt *c* maturation in vitro [51–53]. The present simulation provides further evidence for this suggestion. As shown in Fig. 3b and c by spatial alignment, cyt $b_5$ N57C overlaps well with cyt $b_5$, and the latter can be considered a *b*-type intermediate in

the formation of *c*-type cyt $b_5$. A previous study showed that *c*-type R98C cyt $b_{562}$ has a very similar solution structure to that of the *b*-type cyt $b_{562}$ [15]. In addition, during the formation of cyt *c* species in vitro, the *b*-type intermediates were found to have similar spectroscopic properties to those of the *c*-type products [54].

Note that when bound to cyt $b_5$, the heme group has two orientations (A and B), which differ by a 180° rotation around the α-γ-meso axis of the heme. In the case of bovine cyt $b_5$, the ratio A:B was found to be 9:1 [55], and the major orientation (A) was modeled in the present simulation for cyt $b_5$ N57C, where Cys57 forms a thioether bond with the heme 4-vinyl group (Figs. 1g and 3a). Comparatively, HF cyt $b_5$ has a lower ratio of A:B of 1.2:1 [56], with both orientations resolved in the X-ray structure (PDB entry 2ibj, Fig. 6c, orange) [19]. For the modeling of DS cyt $b_5$ performed herein, only the A orientation was used, because C56 was found to be unsuitable for forming a thioether bond with the 2-vinyl group, as heme is in the B orientation. This was also observed for cyt $b_5$ N57C when an attempt was made to model with heme in the B orientation. These results suggest that the heme has to be in the correct stereochemical orientation for the spontaneous formation of cyt *c*-like cyt $b_5$ species, in agreement with previous experimental observations which showed that single thioether bond formation in cyt *c* can be selective with respect to the heme orientation [53].

## Conclusions

Experimental studies [12, 16] showed that the heme group in bovine liver cyt $b_5$ N57C is covalently attached to the protein polypeptide chain via a single cyt *c*-like thioether linkage, although there is no X-ray or NMR structure available for this species. To provide structural information that is complementary to experimental observations, we performed a molecular modeling study of cyt $b_5$ N57C. At the same time, based on sequence alignment, homology modeling and MD simulation, we predicted the structure of DS cyt $b_5$, a newly discovered member of the cyt $b_5$ family, in which a thioether bond was modeled between the heme 4-vinyl group and a naturally occurring Cys56 residue. The modeling structures provide insight into the spontaneous formation of a *c*-type cyt $b_5$ that requires no typical CXXCH heme-binding motif of cyt *c*; instead, it depends on the appropriate orientations of cysteine and the heme. The structural similarities between the modeling structure of the *c*-type cyt $b_5$ and the native *b*-type protein suggest that *c*-type cyt $b_5$ maturation involves the *b*-type intermediate. More direct structural information on cyt $b_5$ N57C and DS cyt $b_5$ obtained from experiments is currently under investigation.

**Acknowledgments** We gratefully thank Prof. Zhong-Xian Huang at Fudan University, Shanghai, China, for providing the cytochrome b5 gene, and Dr. Tianlei Ying at National Institutes of Health (NIH), USA, for helpful discussions. NAMD and VMD were developed by the Theoretical Biophysics Group in the Beckman Institute for Advanced Science and Technology at the University of Illinois at Urbana-Champaign, USA. This work is supported by Hunan Provincial Natural Science Foundation of China, No. 11JJ4017.

## References

1. Durley RC, Mathews FS (1996) Acta Crystallogr D 52:65–76
2. Bushnell GW, Louie GV, Brayer GD (1990) J Mol Biol 214:585–595
3. Stroebel D, Choquet Y, Popot JL, Picot D (2003) Nature 426:413–418
4. Kurisu G, Zhang H, Smith JL, Cramer WA (2003) Science 302:1009–1014
5. Fülöp V, Sam KA, Ferguson SJ, Ginger ML, Allen JW (2009) FEBS J 276:2822–2832
6. Stevens JM, Daltrop O, Allen JW, Ferguson SJ (2004) Acc Chem Res 37:999–1007
7. Stevens JM, Uchida T, Daltrop O, Ferguson SJ (2005) Biochem Soc Trans 33:792–795
8. Allen JW, Ginger ML, Ferguson SJ (2005) Biochem Soc Trans 33:145–146
9. Giegé P, Grienenberger JM, Bonnard G (2008) Mitochondrion 8:61–73
10. Kranz RG, Richard-Fogal C, Taylor JS, Frawley ER (2009) Microbiol Mol Biol Rev 73:510–528
11. Sanders C, Turkarslan S, Lee DW, Daldal F (2010) Trends Microbiol 18:266–274
12. Barker PD, Ferrer JC, Mylrajan M, Loehr TM, Feng R, Konishi Y, Funk WD, MacGillivray RT, Mauk AG (1993) Proc Natl Acad Sci USA 90:6542–6546
13. Barker PD, Nerou EP, Freund SM, Fearnley IM (1995) Biochemistry 34:15191–1203
14. Arnesano F, Banci L, Bertini I, Ciofi-Baffoni S, Woodyear TL, Johnson CM, Barker PD (2000) Biochemistry 39:1499–1514
15. Faraone-Mennella J, Tezcan FA, Gray HB, Winkler JR (2006) Biochemistry 45:10504–10511
16. Lin YW, Wang WH, Zhang Q, Lu HJ, Yang PY, Xie Y, Huang ZX, Wu HM (2005) Chembiochem 6:1356–1359
17. Muskett FW, Kelly GP, Whitford D (1996) J Mol Biol 258:172–189
18. Altuve A, Silchenko S, Lee KH, Kuczera K, Terzyan S, Zhang X, Benson DR, Rivera M (2001) Biochemistry 40:9469–9483
19. Wang L, Cowley AB, Terzyan S, Zhang X, Benson DR (2007) Proteins 67:293–304
20. Karplus M, McCammon JA (2002) Nat Struct Biol 9:646–652
21. Daggett V (2006) Chem Rev 106:1898–1916
22. Sotomayor M, Schulten K (2007) Science 316:1144–1148
23. Lin YW (2011) Proteins 79:679–684
24. Storch EM, Daggett V (1995) Biochemistry 34:9682–9693
25. Storch EM, Daggett V (1996) Biochemistry 35:11596–11604
26. Lee KH, Kuczera K (2003) Biopolymers 69:260–269
27. Giachetti A, La Penna GL, Perico A, Banci L (2004) Biophys J 87:498–512
28. Cheng Q, Benson DR, Rivera M, Kuczera K (2006) Biopolymers 83:297–312
29. Lin YW, Wang ZH, Ni FY, Huang ZX (2008) Protein J 27:197–203
30. Lin YW, Nie CM, Liao LF (2009) Mol Struct (Theochem) 910:154–162
31. Lin YW, Ying TL, Liao LF (2009) Chin Chem Lett 20:631–634
32. Lin YW, Ying TL, Liao LF (2011) J Mol Model 17:971–978
33. Thompson TB, Chou KC, Zheng C (1995) J Protein Chem 14:559–566

34. Laidig KE, Daggett V (1996) Fold Des 1:335–346
35. Qin S, Zhou HX (2009) Biophys J 97:12–1
36. Lin YW, Ni FY, Ying TL (2009) J Mol Struct THEOCHEM 898:82–89
37. Daidone I, Amadei A, Roccatano D, Nola AD (2003) Biophys J 85:2865–2871
38. Wang ZH, Lin YW, Rosell FI, Ni FY, Lu HJ, Yang PY, Tan XS, Li XY, Huang ZX, Mauk AG (2007) Chembiochem 8:607–609
39. Singh SR, Prakash S, Vasu V, Karunakaran C (2009) J Mol Graph Model 28:270–277
40. Humphrey W, Dalke A, Schulten K (1996) J Mol Graph 14:33–38
41. Kalé L, Skeel R, Bhandarkar M, Brunner R, Gursoy A, Krawetz N, Phillips J, Shinozaki A, Varadarajan K, Schulten K (1999) J Comput Phys 151:283–312
42. Yeung N, Lin YW, Gao YG, Zhao X, Russell BS, Lei L, Miner KD, Robinson H, Lu Y (2009) Nature 462:1079–1082
43. Bowie JU, Lüthy R, Eisenberg D (1991) Science 253:164–170
44. Lüthy R, Bowie JU, Eisenberg D (1992) Nature 356:83–85
45. Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) J Mol Biol 7:95–99
46. Arnold K, Bordoli L, Kopp J, Schwede T (2006) Bioinformatics 22:195–201
47. Bordoli L, Kiefer F, Arnold K, Benkert P, Battey J, Schwede T (2009) Nat Protocols 4:1–13
48. Thompson JD, Higgins DG, Gibson TJ (1994) Nucl Acids Res 22:4673–4680
49. Pal L, Chakrabarti P, Basu G (2003) J Mol Biol 326:273–291
50. Stevens JM (2011) Metallomics 3:319–322
51. Daltrop O, Allen JW, Willis AC, Ferguson SJ (2002) Proc Natl Acad Sci USA 99:7872–7876
52. Daltrop O, Ferguson SJ (2003) J Biol Chem 278:4404–4409
53. Daltrop O, Smith KM, Ferguson SJ (2003) J Biol Chem 278:24308–24313
54. Ishida M, Dohmae N, Shiro Y, Oku T, Iizuka T, Isogai Y (2004) Biochemistry 4:9823–9833
55. La Mar GN, Burns PD, Jackson JT, Smith KM, Langry KC, Strittmatter P (1981) J Biol Chem 256:6075–6079
56. Cowley AB, Rivera M, Benson DR (2004) Protein Sci 13:2316–2329

J Mol Model (2012) 18:1561–1572
DOI 10.1007/s00894-011-1175-4

ORIGINAL PAPER

# Theoretical studies on 2-diazo-4,6-dinitrophenol derivatives aimed at finding superior propellants

**Yan Liu · Lianjun Wang · Guixiang Wang · Hongchen Du · Xuedong Gong**

Received: 5 May 2011 / Accepted: 1 July 2011 / Published online: 30 July 2011


**Abstract** In an attempt to find superior propellants, 2-diazo-4,6-dinitrophenol (DDNP) and its $–NO_2$, $–NH_2$, –CN, –NC, $–ONO_2$, and $–NF_2$ derivatives were studied at the B3LYP/6-311++G** level of density functional theory (DFT). Sensitivity was evaluated using bond dissociation enthalpies (BDEs) and molecular surface electrostatic potentials. The C–$NO_2$ bond appears to be the trigger bond during the thermolysis process for these compounds, except for the $–ONO_2$ and $–NF_2$ derivatives. Electrostatic potential results show that electron-withdrawing substituents make the charge imbalance more anomalous, which may change the strength of the bond, especially the weakest trigger bond. Most of the DDNP derivatives have the impact sensitivities that are higher than that of DDNP, making them favorable for use as solid propellants in micro-rockets. The theoretical densities ($\rho$), heats of formation (HOFs), detonation energies ($Q$), detonation pressures ($P$), and detonation velocities ($D$) of the compounds were estimated. The effects of various substituent groups on $\rho$, HOF, $Q$, $D$, and $P$ were investigated. Some derivatives exhibit perfect detonation properties. The calculated relative specific impulses ($I_{r,sp}$) of all compounds except for $–NH_2$ derivatives were higher than that of DDNP, and also meet the requirements of propellants.

**Keywords** 2-Diazo-4,6-dinitrophenol (DDNP) · Propellant · Detonation property · Sensitivity · Specific impulse

Y. Liu · L. Wang · G. Wang · H. Du · X. Gong (✉)
School of Chemical Engineering, Nanjing University of Science and Technology,
Nanjing 210094, Jiangsu, PR China
e-mail: gongxd325@mail.njust.edu.cn

## Introduction

The miniaturization of spacecraft and satellites has been attracting increasing interest, as miniaturization reduces costs, time, and the risk of launch failure as compared with conventional systems. Micro rockets have also been developed to maneuver these miniaturized spacecraft and satellites in space [1–5]. Three types of micro rockets—those employing turbine engines, gaseous propellants, and solid propellants—have been explored [6], but solid propellant micro rockets are preferred as they are more stable, do not leak, and have a relatively high energy density [7–10]. For solid propellant micro rockets, the propellant used is the most important issue. Many kinds of conventional propellants and explosives, such as glycidyl azide polymer (GAP), nitrocellulose (NC), lead styphnate, hexahydro-1,3,5-trinitro-1,3,5-*s*-triazine (RDX), 1,3,5,7-tetranitro-1,3,5,7-tetraazacyclooctane (HMX), and pentaerythritol tetranitrate (PETN), have been tested and used for micro rockets [11]. The results of these tests show that conventional solid propellants do not work in tanks with dimensions on the order of 100 μm when the ignition energy is limited, while explosives with relatively high sensitivity do work and thus can be used for micro rockets with limited ignition power [12, 13]. However, the explosive powder often causes unintentional ignition and handling difficulties [14]. In recent years, researchers have proposed that 2-diazo-4,6-dinitrophenol (DDNP) should be used in micro rockets [15, 16]. DDNP, which was first prepared by Griess in 1858 [17], is generally used as a primary explosive [18]. The sensitivity of DDNP is pretty low in comparison with other primary explosives, such as lead styphnate, which makes it favorable for use in micro rocket arrays with high number densities. Its ignition point is also low (180 °C), which is another advantage for space-

based applications [19]. In their study, DDNP was fabricated into round shapes 80–500 microns in diameter by a unique synthesis technique that largely circumvents handling issues [15, 16]. A micro rocket using DDNP was successfully fired under atmospheric pressure, but not in a vacuum, due to the high diffusion velocity or activation energy of DDNP [20]. Such miniaturization also causes quenching due to the increased surface-to-volume ratio [13]. Hence, it is important to find new and superior propellants for use in micro rockets.

Incorporating some energy-rich functional groups into a molecule is a widely used and effective means to change the detonation performances and sensitivities of explosives. Our primary interest in this work is to look for potential solid propellants for micro rockets among a series of DDNP derivatives of energy-rich functional groups, such as nitro ($-NO_2$), amino ($-NH_2$), cyan ($-CN$), isocyan ($-NC$), nitrate ($-ONO_2$), and difluoroamino ($-NF_2$) groups (see Fig. 1 for the structures of these compounds). Density functional theory (DFT) was used to study their geometries, densities, detonation properties, sensitivities, and specific impulses, and the roles of various substituents in the design of efficient propellants were investigated.

## Computational methods

The geometries of DDNP derivatives were fully optimized without any symmetry restriction using density functional theory at the B3LYP level with the 6-311++G** basis set in the Gaussian 03 program package [21]. Vibrational frequency analyses were performed at the same level. The optimized structures were confirmed to be local minima without imaginary frequencies. Natural bond orbital (NBO) analyses [22] were performed at the same level.

The gas-phase bond dissociation enthalpy (BDE) for a bond R1–R2, defined as the enthalpy change during bond homolysis at 298 K in a vacuum, was calculated from the sum of the enthalpies ($H_{298}$) of the products (radicals) minus the enthalpy of the reactant (parent molecule):

$$\mathrm{R1-R2(g)} \rightarrow \mathrm{R1\cdot(g)} + \mathrm{R2\cdot(g)} \tag{1}$$

$$\mathrm{BDE(R1-R2)} = H_{298,\mathrm{R1}\cdot} + H_{298,\mathrm{R2}\cdot} - H_{298,\mathrm{R1-R2}} \tag{2}$$

where R1–R2 denotes the neutral molecule, and R1· and R2· are the corresponding radical products after bond dissociation; BDE(R1–R2) is the bond dissociation enthalpy of the bond R1–R2; and $H_{298,\mathrm{R1-R2}}$, $H_{298,\mathrm{R1}\cdot}$, and $H_{298,\mathrm{R2}\cdot}$ are the enthalpies of the parent compound and the corresponding radicals, respectively. The enthalpy at 298 K for each species was calculated using the following equation:

$$H_{298} = E_0 + \mathrm{ZPE} + H_{\mathrm{trans}} + H_{\mathrm{rot}} + H_{\mathrm{vib}} + RT \tag{3}$$

where $E_0$ is the total energy; ZPE is the zero point energy; $H_{\mathrm{trans}}$, $H_{\mathrm{rot}}$, and $H_{\mathrm{vib}}$ are the standard thermal correction terms calculated in terms of equilibrium statistical mechanics under the harmonic oscillator and rigid rotor approximations; and $RT$ ($PV$ work term) is the conversion factor from energy to enthalpy.

The electrostatic potentials $V_s(r)$ on the 0.001 au molecular surfaces were calculated for these geometries [23]. The positive and negative extrema of $V_s(r)$ ($V_{s,\max}$ and $V_{s,\min}$) as well as the positive and negative averages ($\overline{V}_s^+$ and $\overline{V}_s^-$) were obtained. The average deviations ($\Pi$), the positive, negative, and total variances ($\sigma_+^2$,$\sigma_-^2$,$\sigma_{\mathrm{tot}}^2$), and the balance parameter $\nu$ were calculated according to the following equations [24–26] and using a self-compiled program [27]:

$$\prod = \frac{1}{t}\sum_{k=1}^{t}\left|V_s(r_k) - \overline{V}_s\right| \tag{4}$$

$$\sigma_+^2 = \frac{1}{m}\sum_{i=1}^{m}\left[V_s^+(r_i) - \overline{V}_s^+\right]^2 \tag{5}$$

$$\sigma_-^2 = \frac{1}{n}\sum_{j=1}^{n}\left[V_s^-(r_j) - \overline{V}_s^-\right]^2 \tag{6}$$

$$\sigma_{\mathrm{tot}}^2 = \sigma_+^2 + \sigma_-^2 \tag{7}$$

$$\nu = \frac{\sigma_+^2\sigma_-^2}{\left(\sigma_+^2 + \sigma_-^2\right)^2} \tag{8}$$

The theoretical density ($\rho$) of each title compound was calculated using the following equation:

$$\rho = \frac{M}{V_m} \tag{9}$$

where $M$ is the molar weight of the compound, and $V_m$ is the average molar volume, which was obtained from the arithmetic average of 100 molar volumes: the volume inside the 0.001 electrons·Bohr$^{-3}$ molecular surface computed by Monte Carlo integration on the basis of the optimized geometry [28–31]. The improved procedure proposed by Politzer et al. [32] was also employed in order to calculate the density of DDNP. Moreover, the crystal density of DDNP was predicted using rigorous molecular

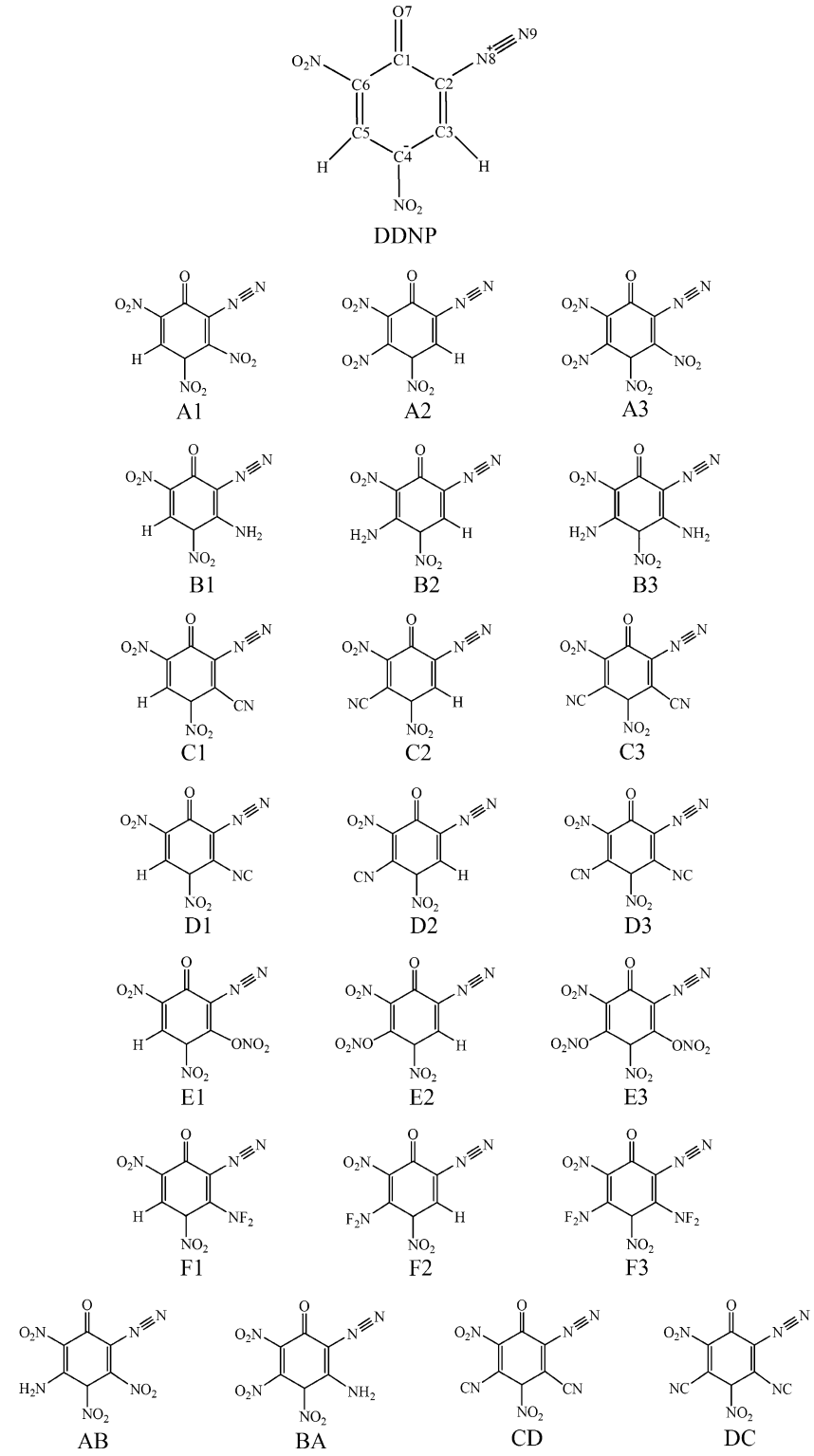



**Fig. 1** Molecular frameworks of DDNP derivatives

packing calculations under the space groups of $P2_1$, $P2_1/c$, $P\text{-}1$, $P2_12_12_1$, $C2/c$, $Pbca$, $Pna2_1$, $Pbcn$, $Cc$, and $C2$ using the Compass, Universal, and Dreiding force fields [28] in the polymorph module of Materials Studio [33].

For each compound, the explosion reaction was determined in terms of the maximal exothermal principle; i.e., all of the N atoms combine into $N_2$ molecules, halogen atoms form hydrogen halide with hydrogen atoms, and the O atoms initially react with the H atoms to give $H_2O$ and then form $CO_2$ with the C atom. If the number of O atoms is more than is needed to oxidize the H and C atoms, the redundant O atoms will combine into $O_2$ molecules. If the

O content is insufficient to allow full oxidation of the H and C atoms, the remaining H atoms will combine into $H_2$ molecules, and the C atoms will remain as solid-state C. The detonation energy $Q$ was then deduced from the explosion reaction, and the empirical Kamlet–Jacobs (K-J) equations were used to estimate the detonation velocity $D$ and detonation pressure $P$ [34]:

$$D = 1.01\left(N\overline{M}^{0.5}Q^{0.5}\right)^{0.5}(1 + 1.30\rho) \quad (10)$$

$$P = 1.558\rho^2 N\overline{M}^{0.5}\mathrm{Q}^{0.5} \quad (11)$$

where $\rho$ is the density in g cm$^{-3}$, $N$ is the number of moles of gas produced by per gram of explosive, and $\overline{M}$ is the mean molecular weight of the gaseous products.

The specific impulse ($I_{sp}$) is a performance measure that is used for propellants, and it is defined as the ratio of the thrust produced by a rocket engine to the rate of fuel consumption. It can be estimated by the following formula [35–37]:

$$I_{sp} \approx T_C{}^{1/2}N^{1/2} \quad (12)$$

where $T_C$ is the absolute temperature in the combustion chamber, and this in turn can be estimated by the following equation based on the maximal exothermal principle:

$$T_C = T_0 - \frac{\Delta H_{comb}}{C_{p,products}} \quad (13)$$

where $T_0$ is the initial temperature in the combustion chamber, $\Delta H_{comb}$ is the enthalpy of combustion for the detonation reaction, and $C_{p,products}$ is the total heat capacity of the gaseous products. Here it is assumed that the formation and discharge rates of gaseous products are equal, so $\Delta H_{comb}$ can be calculated from the heats of formation of the propellant $\Delta_f H_{m,i}$ and the gaseous products $\Delta_f H_{m,propellant}$ as follows:

$$\Delta H_{comb} = \sum_{i}^{products} N_i \Delta_f H_{m,i} - \Delta_f H_{m,propellant} \quad (14)$$

Since the calculated $I_{sp}$ is not very precise, the relative specific impulse ($I_{r,sp}$), which can be used to compare the relative magnitudes of $I_{sp}$ for compounds with similar structures, is used in the present work:

$$I_{r,sp} = \frac{I_{sp}}{I_{sp,DDNP}} \quad (15)$$

where $I_{sp}$ and $I_{sp,\ DDNP}$ represent the specific impulses for the DDNP derivatives and DDNP, respectively.

## Results and discussion

### Validation of the structure

The structure and bonding of 2-diazo-4,6-dinitrophenol is somewhat special in that it has two tautomers and several resonance structures. Though X-ray diffraction experiments have shown that the DDNP crystal exists in quinoid form, to obtain more information on the bonding of DDNP, NBO analysis was carried out based on the optimized geometry at the B3LYP/6-311++G** level. Table 1 presents the Wiberg bond indices (WBI) and bond lengths for the DDNP molecule. For comparison, the computed and experimental results of Holl et al. [38] are also listed in the table.

The Wiberg bond index and bond length indicate that there is a double bond between C1 and O7, because the WBI for the bond (1.706) is close to 2.0, and the bond length (1.219 Å) is very close to that of a normal C=O double bond (1.20 Å) [39]. Similarly, a triple bond exists between N8 and N9. C1–C2, C6–C1, C4–$NO_2$, C6–$NO_2$, C3–H, and C5–H are single bonds. For other bonds, electronic delocalization causes their WBI to be slightly higher than 1.0 and their bond lengths to be shorter than that of the corresponding single bond. Taking the C2–N8 bond as an example, its WBI (1.194) is higher than 1.0 and its length (1.335 Å) lies between that of the normal double bond (1.28 Å) and the single bond (1.47 Å) [39].

### Sensitivity

For propellants, besides good detonation performance, sensitivity is another prerequisite. They should be safe, stable and reliable enough to detonate under specific conditions [40]. Sensitivity refers to the propellant's

**Table 1** Wiberg bond indices and bond lengths for the DDNP molecule

| Bond | WBI | Bond length | | |
|---|---|---|---|---|
| | | Calc.[a] | Calc.[b] | Exp.[b] |
| C1–C2 | 1.002 | 1.504 | 1.490 | 1.460 |
| C2–C3 | 1.277 | 1.410 | 1.400 | 1.376 |
| C3–C4 | 1.484 | 1.370 | 1.360 | 1.365 |
| C4–C5 | 1.215 | 1.422 | 1.410 | 1.401 |
| C5–C6 | 1.590 | 1.363 | 1.360 | 1.355 |
| C6–C1 | 1.081 | 1.473 | 1.460 | 1.460 |
| C1–O7 | 1.706 | 1.219 | 1.215 | 1.220 |
| C2–N8 | 1.194 | 1.335 | 1.330 | 1.379 |
| N8–N9 | 2.538 | 1.127 | 1.120 | 1.081 |

[a] This work

[b] The calculated and experimental results of Holl et al. [38]

stability to various external stimuli such as heat, impact, shock, or an electrical spark [41]. Predicting the sensitivity, which can be done by examining the molecular structure, is a very important step when developing new energetic materials.

*Bond dissociation enthalpy*

Previous investigations have shown that for many energetic compounds, the rupture of the weakest bond is the first step in detonation initiation, and the properties of this bond (such as length, strength and electrostatic potential) are related to the stability and sensitivity of the compound [42, 43]. The strength of the weakest bond is most widely used to index the relative sensitivity at present. Usually, the stronger the weakest bond, the more stable the energetic material; that is, the sensitivity and stability of the energetic compound is related to the strength of the weakest bond to a certain extent, and this can be evaluated via the BDE [44]. A bond with a smaller BDE is usually weaker and more likely to be the trigger bond. Correspondingly, a compound with a low trigger BDE is generally less stable and has a higher sensitivity. Therefore, studies of BDEs are important and essential in order to understand the stabilities and decomposition processes of energetic materials. However, experimental values of BDE are not easy to access, because the thermochemistry of a short-lived species is not amenable to traditional calorimetric techniques. On the other hand, computational chemistry can be a more convenient and reliable method in this context.

BDEs were calculated for the following bonds: C–$NO_2$, C–NN, C–$NH_2$, C–CN, C–NC, C–$ONO_2$, C–$NF_2$, and O–$NO_2$. Since there may be several bonds of the same kind in a molecule, the WBIs obtained from NBO analysis were used to ascertain the weakest bond. Usually, a smaller WBI implies a weaker bond, so the bond with the smallest WBI among all bonds of the same type was considered. Table 2 gives the WBIs for the title compounds, and Table 3 shows the BDEs corrected for the zero point energy (ZPE).

From Table 2, the WBIs of C–NN in the title compounds, except for the –$NH_2$ derivatives, are the largest. Generally, the larger the WBI, the more stable the bond, so this kind of bond will not rupture first. This is consistent with the result that the BDEs of C–NN are the highest (they are all about 800 kJ $mol^{-1}$ in Table 3). On the other hand, the WBI for C–$NO_2$ is the lowest in each compound, other than in –$ONO_2$ derivatives. Similarly, C–$NO_2$ bonds have much smaller BDEs than other bonds, except for O–$NO_2$ and C–$NF_2$. This suggests that the C–$NO_2$ bond is the trigger bond for the thermolysis process in most of these compounds, and that the nitro group has an activating effect. Although C–$NO_2$ bonds in different

**Table 2** Wiberg bond indices for various bonds in the title compounds

| Compound | C–NN | C4–$NO_2$ | C6–$NO_2$ | C3 subst. | C5 subst. | (C3–)O–$NO_2$ | (C5–)O–$NO_2$ |
|---|---|---|---|---|---|---|---|
| DDNP | 1.1942 | 0.9452 | 0.9253 | | | | |
| A1 | 1.1876 | 0.9397 | 0.9001 | 0.8757 | | | |
| A2 | 1.1840 | 0.9369 | 0.9214 | | 0.8867 | | |
| A3 | 1.1781 | 0.8939 | 0.8926 | 0.8825 | 0.9103 | | |
| B1 | 1.2435 | 1.0097 | 0.9442 | 1.3204 | | | |
| B2 | 1.2057 | 0.9799 | 0.9960 | | 1.3753 | | |
| B3 | 1.2524 | 1.0747 | 1.0073 | 1.3268 | 1.3624 | | |
| C1 | 1.1876 | 0.9447 | 0.9202 | 1.0869 | | | |
| C2 | 1.1872 | 0.9448 | 0.8893 | | 1.0788 | | |
| C3 | 1.1837 | 0.9398 | 0.8879 | 1.0839 | 1.0756 | | |
| D1 | 1.1979 | 0.9482 | 0.9236 | 1.0742 | | | |
| D2 | 1.1910 | 0.9430 | 0.8984 | | 1.0702 | | |
| D3 | 1.2020 | 0.9229 | 0.9008 | 1.0647 | 1.0664 | | |
| E1 | 1.1977 | 0.9622 | 0.9261 | 1.0597 | | 0.6201 | |
| E2 | 1.1920 | 0.9474 | 0.9052 | | 1.0162 | | 0.6850 |
| E3 | 1.1999 | 0.9475 | 0.9076 | 1.0518 | 1.0249 | 0.6429 | 0.6597 |
| F1 | 1.1880 | 0.9291 | 0.9192 | 1.0058 | | | |
| F2 | 1.1910 | 0.9297 | 0.8857 | | 1.0036 | | |
| F3 | 1.1929 | 0.8870 | 0.8855 | 1.0120 | 1.0038 | | |
| AB | 1.2325 | 0.9951 | 0.9086 | 1.3154 | 0.8708 | | |
| BA | 1.1974 | 0.9614 | 0.9872 | 0.8812 | 1.3663 | | |
| CD | 1.1971 | 0.9292 | 0.8933 | 1.0669 | 1.0751 | | |
| DC | 1.1901 | 0.9276 | 0.8969 | 1.0827 | 1.0668 | | |

**Table 3** BDEs (in kJ $mol^{-1}$) for various kinds of bonds in the title compounds

| Compound | C–NN | C–$NO_2$ | C–$NH_2$ | C–CN | C–NC | C–$ONO_2$ | O–$NO_2$ | C–$NF_2$ |
|---|---|---|---|---|---|---|---|---|
| DDNP | 802.36 | 272.34 | | | | | | |
| A1 | 797.55 | 225.43 | | | | | | |
| A2 | 801.76 | 250.45 | | | | | | |
| A3 | 793.41 | 227.29 | | | | | | |
| B1 | 818.34 | 284.15 | 489.37 | | | | | |
| B2 | 818.76 | 303.47 | 495.69 | | | | | |
| B3 | 817.50 | 300.31 | 469.38 | | | | | |
| C1 | 807.41 | 280.69 | | 531.08 | | | | |
| C2 | 800.57 | 279.05 | | 517.56 | | | | |
| C3 | 801.40 | 255.88 | | 526.90 | | | | |
| D1 | 808.47 | 278.98 | | | 440.41 | | | |
| D2 | 800.81 | 275.94 | | | 428.27 | | | |
| D3 | 801.37 | 256.16 | | | 439.14 | | | |
| E1 | 810.59 | 288.64 | | | | 360.25 | 33.71 | |
| E2 | 800.53 | 285.65 | | | | 352.27 | 65.78 | |
| E3 | 803.84 | 267.33 | | | | 343.50 | 16.28 | |
| F1 | 800.61 | 267.16 | | | | | | 247.97 |
| F2 | 795.25 | 263.40 | | | | | | 231.15 |
| F3 | 797.61 | 242.02 | | | | | | 260.18 |
| AB | 809.71 | 213.83 | 492.81 | | | | | |
| BA | 815.38 | 255.00 | 484.75 | | | | | |
| CD | 800.71 | 254.36 | | 527.13 | 424.85 | | | |
| DC | 800.74 | 255.64 | | 511.53 | 437.44 | | | |

compounds have slightly different WBIs and BDEs due to the incorporation of different substituent groups, the effects of the various substituents are very small.

For the –$NH_2$ derivatives (B1–B3) only, the WBI and the BDE of C-$NO_2$ are higher than those for C–$NO_2$ in DDNP; this is in accord with the results of previous studies [22–25, 36] which showed that introducing –$NH_2$ is an effective way to increase the thermal stability of an explosive due to the formation of a hydrogen bond between –$NO_2$ and –$NH_2$. For the –$ONO_2$ derivatives (E1–E3), the WBI and the BDE of O–$NO_2$ are much smaller than those of other bonds, hence the O–$NO_2$ bond may pyrolyze first. As for the –$NF_2$ derivatives (F1–F3), C–$NF_2$ may be the trigger bond for the compound with only one –$NF_2$, whereas C–$NO_2$ may be the trigger bond when the compounds have more –$NF_2$ groups.

*Molecular surface electrostatic potentials*

Politzer's studies [45–50] showed that the impact sensitivities of energetic compounds correlate with the anomalous charge imbalance, which is a feature of their molecular surface electrostatic potentials. The positive and negative extrema ($V_{s,\,max}$ and $V_{s,\,min}$), the average ($\overline{V}_S$), the positive and negative averages ($\overline{V}_S^+$ and $\overline{V}_S^-$) of $V_s(r)$, the positive, negative and total variances ($\sigma_+^2, \sigma_-^2, \sigma_{tot}^2$), the balance parameter $\nu$, and the average deviations ($\Pi$) for the title compounds were calculated and are given in Table 4.

The molecular surface electrostatic potentials for the title compounds are clearly dominated by strongly positive regions owing to the strongly electron-withdrawing groups. As electron-withdrawing substituents (such as –$NO_2$, –CN, –NC, –$ONO_2$, and –$NF_2$ groups) are introduced into the DDNP molecule, the balance parameter $\nu$ generally decreases; in other words, the charge imbalance becomes more anomalous, and the positive regions become stronger and more variable than the negative ones. On the other hand, electron-pushing substituents such as –$NH_2$ generally make the $\nu$ value increase, which illustrates that the position of the substituent on the benzene ring is also crucial to the charge imbalance. When the H atom on C3 is substituted with an electron-withdrawing group, the –$NO_2$ at the opposite C6 position counteracts the electron-withdrawing effect, so the charge balance is better than that for DDNP. However, when the H atom on C5 is substituted, all electron-withdrawing groups are situated on one side of the benzene ring, which causes a poorly balanced charge distribution.

Moreover, the charge redistributions caused by the substituents may change the electrostatic potentials and weaken the strength of the bond, including the weakest bond. Since the detonation of many explosives initiates

**Table 4** Computed parameters relating to the molecular surface electrostatic potentials for the title compounds

| Compound | $V_{s,\ min}$ | $V_{s,\ max}$ | $\overline{V}_S^+$ | $\overline{V}_S^-$ | $\overline{V}_S$ | $\sigma_+^2$ | $\sigma_-^2$ | $\sigma_{tot}^2$ | $\nu$ | $\Pi$ |
|---|---|---|---|---|---|---|---|---|---|---|
| DDNP | −36.9 | 59.3 | 24.3 | −17.0 | 8.2 | 151.6 | 73.8 | 225.5 | 0.220 | 20.2 |
| A1 | −27.6 | 58.2 | 23.3 | −11.5 | 8.2 | 158.2 | 46.9 | 205.1 | 0.176 | 17.9 |
| A2 | −24.7 | 65.4 | 28.0 | −13.6 | 7.3 | 231.5 | 37.8 | 269.3 | 0.121 | 21.3 |
| A3 | −19.6 | 65.9 | 22.5 | −8.9 | 7.9 | 231.2 | 24.9 | 256.1 | 0.088 | 16.7 |
| B1 | −40.7 | 79.1 | 25.1 | −19.6 | 7.5 | 256.3 | 95.4 | 351.7 | 0.198 | 21.7 |
| B2 | −42.9 | 55.4 | 21.3 | −18.6 | 6.8 | 141.0 | 112.1 | 253.1 | 0.247 | 18.8 |
| B3 | −46.0 | 74.9 | 21.0 | −20.7 | 6.0 | 215.4 | 131.1 | 346.6 | 0.235 | 19.6 |
| C1 | −28.4 | 50.8 | 24.5 | −13.1 | 8.4 | 149.7 | 42.7 | 192.4 | 0.173 | 19.1 |
| C2 | −31.3 | 64.6 | 28.0 | −14.9 | 7.6 | 208.3 | 52.2 | 260.5 | 0.160 | 21.8 |
| C3 | −26.2 | 56.3 | 26.0 | −10.9 | 9.3 | 219.1 | 34.0 | 253.1 | 0.116 | 19.3 |
| D1 | −30.1 | 50.9 | 23.3 | −13.2 | 8.8 | 142.7 | 51.9 | 194.6 | 0.196 | 18.3 |
| D2 | −27.4 | 63.3 | 26.0 | −13.4 | 7.5 | 222.1 | 47.9 | 270.0 | 0.146 | 20.2 |
| D3 | −23.0 | 54.2 | 22.1 | −10.8 | 7.9 | 195.4 | 33.1 | 228.5 | 0.124 | 17.1 |
| E1 | −35.7 | 51.9 | 19.9 | −15.5 | 8.0 | 145.4 | 73.1 | 218.6 | 0.223 | 17.0 |
| E2 | −28.6 | 61.6 | 24.0 | −12.0 | 6.6 | 217.8 | 54.4 | 272.1 | 0.160 | 18.5 |
| E3 | −27.8 | 53.8 | 18.7 | −10.9 | 6.9 | 170.5 | 50.1 | 220.6 | 0.175 | 15.1 |
| F1 | −28.7 | 42.6 | 20.4 | −12.8 | 8.1 | 124.7 | 48.2 | 172.8 | 0.201 | 16.4 |
| F2 | −25.0 | 63.3 | 26.8 | −12.5 | 6.9 | 217.7 | 40.3 | 258.0 | 0.132 | 20.0 |
| F3 | −20.5 | 45.9 | 17.8 | −9.2 | 6.7 | 158.2 | 29.9 | 188.1 | 0.134 | 14.2 |
| AB | −29.4 | 83.6 | 29.5 | −16.9 | 6.7 | 294.7 | 48.5 | 343.2 | 0.121 | 23.5 |
| BA | −36.3 | 54.6 | 21.1 | −13.0 | 7.4 | 128.1 | 77.3 | 205.4 | 0.235 | 17.0 |
| CD | −26.1 | 56.0 | 24.3 | −11.3 | 9.2 | 197.3 | 36.2 | 233.5 | 0.131 | 18.5 |
| DC | −22.6 | 54.6 | 23.5 | −10.2 | 8.7 | 200.5 | 29.8 | 230.3 | 0.113 | 17.7 |

[a] Units: $V_{s,\ min}$, $V_{s,\ max}$, $\overline{V}_S^+$, $\overline{V}_S^-$, $\overline{V}_S$, and $\Pi$ are in kcal mol$^{-1}$; $\sigma_+^2$, $\sigma_-^2$ and $\sigma_{tot}^2$ are in (kcal mol$^{-1}$)$^2$; $\nu$ is unitless

with the rupture of the weakest bond, and the sensitivity of the explosive is related to the strength of the trigger bond, there is certain relationship between the sensitivity and the electrostatic potential. According to Rice et al. [42], the most sensitive molecules among all of the types considered in their study had higher anomalous charge imbalances. Hence, the majority of the DDNP derivatives (excluding B2, B3, E1, and BA) may have impact sensitivities that are slightly higher than DDNP, which would make them favorable for use as solid propellants in micro rockets.

## Detonation performance

The detonation pressure ($P$) and detonation velocity ($D$) are the most important parameters when evaluating the performance of an energetic material. $P$ is proportional to the square of the density and $D$ to the density according to the K-J equations, so the density of an explosive is a critical parameter. The method mentioned in the "Computational method" section, based on quantum chemistry calculations, was employed here to predict the density. Previous studies have proven that this method is relatively reliable and simple compared with experiments [28–31]. The result obtained (1.757 g cm$^{-3}$) is in good agreement with the experimental value (1.727 g cm$^{-3}$); the deviation is only 1.74%. The densities of DDNP obtained from the improved procedure proposed by Politzer et al. [32] are both 1.80 g cm$^{-3}$, which deviates more from the experimental result (1.727 g cm$^{-3}$) than the density obtained using $M/V_m$ (1.757 g cm$^{-3}$). For comparison and validation, the crystal density of DDNP was also predicted via rigorous molecular packing calculations under the space groups of $P2_1$, $P2_1/c$, $P$-1, $P2_12_12_1$, $C2/c$, $Pbca$, $Pna2_1$, $Pbcn$, $Cc$, and $C2$ using the Compass, Universal, and Dreiding force fields.

Table 5 lists the volumes ($V$), densities ($\rho$), unit cell total energies ($E_T$), and lattice parameters of DDNP obtained under ten space groups with the Compass, Universal, and Dreiding force fields. The total energies of the most likely packing structures are quite different, even under the same force field; for example, the energies of the crystals obtained by the Compass force field are in the range of −148.92 to −150.91 kcal (mol asym cell)$^{-1}$, and that obtained with the $P2_12_12_1$ space group has the lowest energy. Generally, the one with the lowest energy is the most likely experimental structure. Thus, according to the results obtained with the Compass force field, DDNP tends to crystallize in the $P2_12_12_1$ space group, consistent with experimental results [38, 51]. The predicted lattice parameters are also close to the experimental values, but there is a large discrepancy between the predicted (1.81 g cm$^{-3}$) and the experimental (1.727 g cm$^{-3}$ [38] and 1.719 g cm$^{-3}$ [51]) densities. Under the Universal force field, the molecular

**Table 5** The volumes, densities, total energies, and lattice parameters of various molecular packings for DDNP

| Force field | Space group | $V$ (Å$^3$) | $\rho$ (g cm$^{-3}$) | $E_T$ (kcal (mol·asym·cell)$^{-1}$ | $a$ (Å) | $b$ (Å) | $c$ (Å) | $\alpha$ (°) | $\beta$ (°) | $\gamma$ (°) |
|---|---|---|---|---|---|---|---|---|---|---|
| Compass | $P2_1$ | 388.52 | 1.80 | −149.52 | 8.63 | 5.49 | 8.20 | 90.00 | 90.39 | 90.00 |
| | $P2_1$/c | 823.58 | 1.69 | −149.73 | 13.17 | 15.65 | 15.65 | 90.00 | 136.28 | 90.00 |
| | $P$-1 | 385.68 | 1.81 | −149.43 | 11.45 | 6.56 | 5.98 | 59.25 | 87.95 | 91.49 |
| | $P2_12_12_1$ | 770.60 | 1.81 | −150.91 | 6.06 | 8.12 | 15.66 | 90.00 | 90.00 | 90.00 |
| | $C$2/c | 1535.75 | 1.82 | −149.77 | 25.02 | 5.96 | 11.23 | 90.00 | 113.54 | 90.00 |
| | $P$bca | 1631.17 | 1.71 | −149.87 | 16.24 | 16.51 | 6.09 | 90.00 | 90.00 | 90.00 |
| | $P$na2$_1$ | 789.36 | 1.77 | −149.68 | 11.08 | 14.35 | 4.96 | 90.00 | 90.00 | 90.00 |
| | $P$bcn | 1585.68 | 1.76 | −149.50 | 17.22 | 9.45 | 9.74 | 90.00 | 90.00 | 90.00 |
| | $C$c | 818.05 | 1.71 | −148.93 | 8.30 | 16.47 | 6.17 | 90.00 | 104.21 | 90.00 |
| | $C$2 | 813.10 | 1.72 | −148.92 | 20.48 | 5.92 | 8.55 | 90.00 | 51.66 | 90.00 |
| Universal | $P2_1$ | 413.74 | 1.69 | −1.57 | 11.54 | 3.44 | 10.49 | 90.00 | 82.60 | 90.00 |
| | $P2_1$/c | 827.48 | 1.69 | −1.57 | 10.49 | 3.44 | 23.08 | 90.00 | 97.41 | 90.00 |
| | $P$-1 | 413.03 | 1.69 | −1.59 | 13.23 | 11.56 | 3.48 | 94.48 | 51.55 | 97.77 |
| | $P2_12_12_1$ | 829.47 | 1.68 | −1.19 | 3.54 | 23.95 | 9.80 | 90.00 | 90.00 | 90.00 |
| | $C$2/c | 1667.91 | 1.67 | −1.23 | 23.11 | 3.64 | 29.79 | 90.00 | 138.20 | 90.00 |
| | $P$bca | 1712.73 | 1.63 | −0.47 | 6.95 | 23.50 | 10.48 | 90.00 | 90.00 | 90.00 |
| | $P$na2$_1$ | 833.27 | 1.67 | −1.06 | 9.79 | 23.91 | 3.56 | 90.00 | 90.00 | 90.00 |
| | $P$bcn | 1700.09 | 1.64 | −0.61 | 14.50 | 6.90 | 16.98 | 90.00 | 90.00 | 90.00 |
| | $C$c | 865.24 | 1.61 | −0.41 | 10.05 | 13.94 | 6.87 | 90.00 | 64.10 | 90.00 |
| | $C$2 | 839.91 | 1.66 | −0.93 | 30.09 | 3.58 | 7.79 | 90.00 | 91.80 | 90.00 |
| Dreiding | $P2_1$ | 398.54 | 1.75 | 6.14 | 10.23 | 3.46 | 11.34 | 90.00 | 97.13 | 90.00 |
| | $P2_1$/c | 797.08 | 1.75 | 6.14 | 40.37 | 3.46 | 32.38 | 90.00 | 169.85 | 90.00 |
| | $P$-1 | 398.54 | 1.75 | 6.14 | 14.71 | 10.23 | 3.46 | 90.00 | 103.61 | 53.13 |
| | $P2_12_12_1$ | 803.56 | 1.74 | 6.63 | 3.48 | 21.94 | 10.54 | 90.00 | 90.00 | 90.00 |
| | $C$2/c | 1615.68 | 1.73 | 6.64 | 35.05 | 3.48 | 47.88 | 90.00 | 163.93 | 90.00 |
| | $P$bca | 1642.65 | 1.70 | 7.39 | 25.32 | 16.17 | 4.01 | 90.00 | 90.00 | 90.00 |
| | $P$na2$_1$ | 809.90 | 1.72 | 6.63 | 27.81 | 8.48 | 3.44 | 90.00 | 90.00 | 90.00 |
| | $P$bcn | 1632.87 | 1.71 | 7.01 | 14.09 | 6.92 | 16.74 | 90.00 | 90.00 | 90.00 |
| | $C$c | 830.01 | 1.68 | 7.56 | 14.17 | 8.97 | 6.70 | 90.00 | 76.89 | 90.00 |
| | $C$2 | 801.71 | 1.74 | 6.70 | 30.48 | 3.79 | 27.09 | 90.00 | 165.18 | 90.00 |

packing with the lowest energy has the $P$-1 space group, which disagrees with the experimental fact and the predicted density is lower than seen in the experiment results. Under the Dreiding force field, the structure with the lowest energy does not belong to the $P2_12_12_1$ space group either, though the predicted density is quite close to the experimental value. In conclusion, the densities calculated with the quantum chemistry method and the Dreiding force field are very close to and agree with the experimental results. In addition, the former method is much faster and simpler than force field calculations of molecular packings. These results indicate that the method used here to calculate the density is acceptable.

The calculated densities ($\rho$), molecular total energies ($E_0$), zero-point energies (ZPE), thermal corrections ($H_T$), heats of formation (HOF), and detonation properties of the title compounds are presented in Table 6 and Fig. 2 in order to illustrate the tendencies more intuitively.

It is evident that different substituent groups exert different influences on the HOF, $\rho$, $Q$, $D$, and $P$. The introduction of isocyan (−NC) and cyan (−CN) groups greatly increases the HOF—especially the former, which increases the HOF by about 300 kJ mol$^{-1}$ when one more –NC is introduced. The effect of a nitro group (−$NO_2$) on the HOF is essentially the same as the effect of difluoroamino (−$NF_2$), and the increase in the HOF is about 40 kJ mol$^{-1}$ when one –$NO_2$ or –$NF_2$ is added. However, the introduction of –$ONO_2$ causes a significant decrease in the HOF, and –$NH_2$ leads to a slight decrease in the HOF of a dozen kJ mol$^{-1}$, which can also be derived based on the fact that the HOFs of AB and BA are slightly lower than those of A1 and A2. The contributions of the substituents to HOF

**Table 6** The calculated total energies ($E_0$), zero-point energies (ZPE), thermal corrections ($H_T$), heats of formation (HOF), densities ($\rho$), detonation energies ($Q$), detonation velocities ($D$), and detonation pressures ($P$) of the title compounds

| Compound | $E_0$ (a.u.) | ZPE (a.u.) | $H_T$ (a.u.) | HOF (kJ mol$^{-1}$) | $Q$ (cal g$^{-1}$) | $\rho$ (g cm$^{-3}$) | $D$ (km s$^{-1}$) | $P$ (GPa) |
|---|---|---|---|---|---|---|---|---|
| DDNP | −824.69 | 0.09538 | 0.10857 | 134.22 | 1322.97 | 1.757 | 7.29 | 23.70 |
| A1 | −1029.16 | 0.09698 | 0.11302 | 172.88 | 1473.36 | 1.880 | 8.15 | 30.87 |
| A2 | −1029.16 | 0.09698 | 0.11299 | 177.88 | 1478.05 | 1.858 | 8.09 | 30.19 |
| A3 | −1233.62 | 0.09861 | 0.11744 | 221.87 | 1586.88 | 1.972 | 8.80 | 36.98 |
| B1 | −880.06 | 0.11227 | 0.12680 | 121.00 | 1244.62 | 1.781 | 7.38 | 24.49 |
| B2 | −880.06 | 0.11289 | 0.12701 | 119.18 | 1242.69 | 1.771 | 7.35 | 24.20 |
| B3 | −935.42 | 0.12961 | 0.14504 | 109.32 | 1177.60 | 1.771 | 7.36 | 24.29 |
| C1 | −916.91 | 0.09331 | 0.10847 | 324.76 | 1353.01 | 1.755 | 7.18 | 22.94 |
| C2 | −916.91 | 0.09308 | 0.10843 | 329.99 | 1358.32 | 1.762 | 7.20 | 23.19 |
| C3 | −1009.13 | 0.09095 | 0.10827 | 516.86 | 1378.70 | 1.767 | 7.12 | 22.67 |
| D1 | −916.88 | 0.09267 | 0.10803 | 436.98 | 1467.08 | 1.764 | 7.35 | 24.14 |
| D2 | −916.87 | 0.09250 | 0.10796 | 440.82 | 1470.98 | 1.762 | 7.35 | 24.13 |
| D3 | −1009.06 | 0.08982 | 0.10741 | 741.16 | 1584.77 | 1.749 | 7.32 | 23.82 |
| E1 | −1104.35 | 0.10066 | 0.11777 | −4.12 | 1403.83 | 1.934 | 8.32 | 32.67 |
| E2 | −1104.35 | 0.10074 | 0.11782 | 38.73 | 1441.60 | 1.881 | 8.21 | 31.31 |
| E3 | −1384.01 | 0.10590 | 0.12696 | −94.69 | 1489.35 | 2.003 | 8.91 | 38.23 |
| F1 | −1078.36 | 0.09367 | 0.10999 | 178.16 | 1539.89 | 1.930 | 8.46 | 33.75 |
| F2 | −1078.35 | 0.09373 | 0.11009 | 171.50 | 1533.80 | 1.967 | 8.57 | 34.99 |
| F3 | −1332.02 | 0.09182 | 0.11153 | 215.81 | 1681.09 | 2.059 | 9.32 | 42.47 |
| AB | −1084.53 | 0.11442 | 0.13136 | 164.74 | 1404.19 | 1.873 | 8.14 | 30.68 |
| BA | −1084.52 | 0.11393 | 0.13118 | 168.62 | 1407.63 | 1.879 | 8.16 | 30.94 |
| CD | −1009.09 | 0.09038 | 0.10783 | 628.35 | 1481.13 | 1.773 | 7.26 | 23.67 |
| DC | −1009.09 | 0.09043 | 0.10787 | 629.85 | 1482.51 | 1.752 | 7.21 | 23.12 |

follow the order: –NC>−CN>−$NO_2$≈−$NF_2$>−$NH_2$>−$ONO_2$, which is the same as that found for 2,2′,4,4′,6,6′-hexanitroazobenzene (HNAB) derivatives [28]. Furthermore, the HOF values for isomers such as A1 and A2, B1 and B2, etc., are somewhat different, indicating that HOF is slightly influenced by the position of the substituting group.

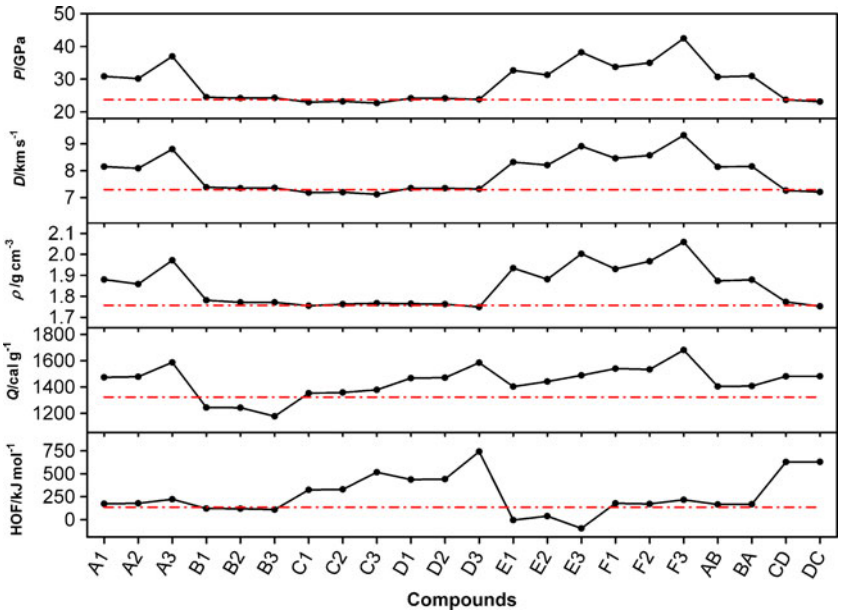



**Fig. 2** The calculated HOFs, densities, detonation energies, detonation velocities, and detonation pressures of the DDNP derivatives. The *dashed lines* represent the corresponding values for DDNP

**Table 7** $I_{r,sp}$, $T_c$, $\Delta H_{comb}$, and $C_{p,\ products}$ values for the DDNP derivatives

| Compound | $\Delta H_{comb}$ (kJ mol$^{-1}$) | $C_{p,\ products}$ (J mol$^{-1}$ K$^{-1}$) | $T_c$ (K) | $I_{r,\ sp}$ |
|---|---|---|---|---|
| DDNP | −1163.1 | 200.35 | 6103.3 | 1.00 |
| A1 | −1577.7 | 234.49 | 7026.5 | 1.09 |
| A2 | −1572.7 | 234.49 | 7005.2 | 1.09 |
| A3 | −1992.7 | 268.62 | 7716.3 | 1.15 |
| B1 | −1170.6 | 224.14 | 5520.7 | 0.99 |
| B2 | −1172.4 | 224.14 | 5528.9 | 0.99 |
| B3 | −1183.2 | 247.92 | 5070.8 | 0.97 |
| C1 | −1336.3 | 213.90 | 6545.6 | 1.00 |
| C2 | −1331.1 | 213.90 | 6521.2 | 1.00 |
| C3 | −1500.6 | 227.44 | 6896.1 | 1.00 |
| D1 | −1447.1 | 213.90 | 7063.8 | 1.04 |
| D2 | −1443.3 | 213.90 | 7045.8 | 1.04 |
| D3 | −1724.9 | 227.44 | 7882.3 | 1.07 |
| E1 | −1635.3 | 249.05 | 6864.5 | 1.08 |
| E2 | −1592.5 | 249.05 | 6692.4 | 1.07 |
| E3 | −2069.6 | 297.74 | 7249.2 | 1.13 |
| F1 | −1132.7 | 186.60 | 6368.7 | 1.07 |
| F2 | −1139.4 | 186.60 | 6404.4 | 1.07 |
| F3 | −1109.4 | 172.84 | 6717.0 | 1.13 |
| AB | −1591.0 | 258.27 | 6458.3 | 1.07 |
| BA | −1587.1 | 258.27 | 6443.3 | 1.07 |
| CD | −1613.6 | 227.44 | 7392.9 | 1.04 |
| DC | −1612.1 | 227.44 | 7386.3 | 1.04 |

Though $Q$ is deduced from HOF, the effects of the substituent groups on $Q$ are different from their effects on HOF, so the contribution of molecular composition to $Q$ cannot be underestimated. All substituent groups except for $-NH_2$ increase $Q$ as compared to DDNP. The order of the contributions of various substituent groups to $Q$ is: $-NF_2$ > $-NO_2$ > $-NC$ > $-ONO_2$ > $-CN$ > $-NH_2$, which is in accord with the order observed for the HNAB derivatives [28], except that the order of –NC and $-ONO_2$ is reversed. The contributions of $-NF_2$ and $-NO_2$ to $Q$ are the largest, so their presence in energetic materials is favorable on thermodynamic grounds.

It is evident that the plots of $\rho$, $D$, and $P$ in Fig. 2 are very similar in shape, which indicates that the substituent groups all have similar influences on the values of $\rho$, $D$, and $P$, and that $D$ and $P$ depend on $\rho$ more than $Q$. Because of this, the contributions of $-NF_2$ to $D$ and $P$ are the largest, though it has only a small effect on the HOF. $-ONO_2$ leads to a decrease in HOF, which is disadvantageous with respect to $D$ and $P$, but the large increase in $\rho$ caused by $-ONO_2$ is inferior only to that caused by $-NF_2$, which makes $D$ and $P$ of $-ONO_2$ derivatives second in the standings. $-NO_2$ increases HOF and $\rho$ moderately, so the derivatives of $-NO_2$ have the third largest $D$ and $P$ values. Though $-NH_2$ decreases the HOF, it increases $\rho$ to a certain extent, so the detonation properties of $-NH_2$ derivatives are enhanced to a small degree. On the other hand, the contribution of –NC to the HOF is the largest, and it only makes a minor contribution to $\rho$, so its contributions to $D$ and $P$ are very small. As for –CN derivatives, a large increase in the HOF cannot make up for the decrease in $\rho$, so their $D$ and $P$ values are smaller than those of DDNP.

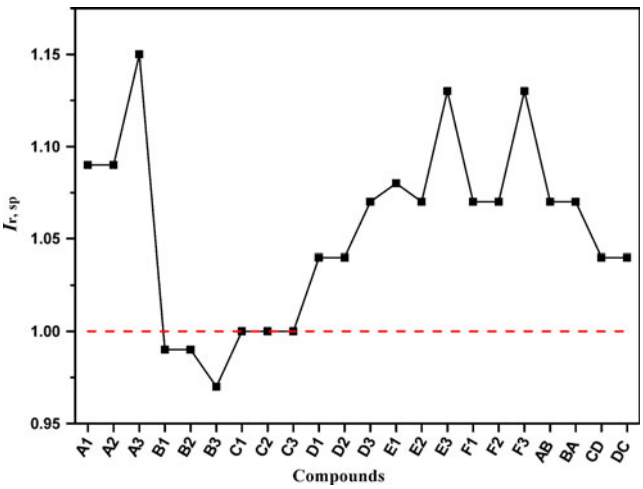



**Fig. 3** The calculated relative specific impulses ($I_{r,sp}$) of the title compounds. The *horizontal dashed line* represents the corresponding value for DDNP

It is worth noting that $\rho$, $D$, and $P$ of the derivatives of $-NF_2$, $-NO_2$, and $-ONO_2$ are higher than those of DDNP, and those of $-NH_2$, –NC, and –CN derivatives are similar in magnitude to those of DDNP, which indicates that these compounds may be good candidates for propellants.

Specific impulse

The specific impulse is a critical parameter for propellants. Essentially, the greater the specific impulse of a propellant, the smaller the amount needed to produce a desired momentum. The relative specific impulses ($I_{r,sp}$), the absolute temperatures in the combustion chamber ($T_c$), the enthalpies of combustion for the detonation reaction ($\Delta H_{comb}$), and the total heat capacities of the gaseous products ($C_{p,products}$) are listed in Table 7 for the title compounds. Figure 3 shows the effects of substituents on $I_{r,sp}$ more clearly. Only the $I_{r,sp}$ values of the $-NH_2$ derivatives are less than 1, which indicates that the $-NH_2$ group decreases the specific impulse. This is consistent with the effect of $-NH_2$ on the detonation energy $Q$. Introducing –CN has little influence on $I_{r,sp}$, so the $I_{r,sp}$ values of –CN derivatives are close to 1. The contributions of the various substituent groups to $I_{r,sp}$ can be ordered as follows: $-NO_2 > -NF_2 \approx -ONO_2 > -NC > -CN > -NH_2$, which is not precisely in accord with the orders seen for $D$ and $P$. The derivatives of $-NO_2$ have the largest $I_{r,sp}$ values; i.e., introducing $-NO_2$ greatly increases the specific impulse. Though the $-NF_2$ derivatives exhibited $I_{r,sp}$ values that were a little lower than those of the $-NO_2$ derivatives, their sensitivities were superior to those of the nitro derivatives, so incorporating $-NF_2$ may be a feasible route to improving the properties of propellants.

## Conclusions

Based on the theoretical studies of the structures and the properties of the DDNP derivatives, the following conclusions can be drawn:

(1) Most of the substituents improve $\rho$, $D$, and $P$, except for –CN and –NC, and their contributions can be ordered as follows: $-NF_2 > -ONO_2 > -NO_2 > -NH_2$. All of the substituents have similar influences on $\rho$, $D$, and $P$. The $\rho$, $D$, and $P$ values of the $-NF_2$, $-NO_2$, and $-ONO_2$ derivatives are higher than those of DDNP, and those of the derivatives of $-NH_2$, –NC, and –CN are close to those of DDNP.

(2) The sensitivities of DDNP derivatives were correlated with the Wiberg bond indices, bond dissociation enthalpies, and molecular surface electrostatic potentials. The C–$NO_2$ bond appears to be the trigger bond during the thermolysis initiation process for the title compounds, except for the $-ONO_2$ and $-NF_2$ derivatives. Electron-withdrawing substituents make the charge imbalance more anomalous, while electron-donating substituents make the balance parameter increase. The charge redistributions caused by the substituents may change the strength of the trigger bond. All of the DDNP derivatives except for B2, B3, E1, and BA have the impact sensitivities that are slightly higher than DDNP, which makes them favorable for use as solid propellants in micro rockets.

(3) $NH_2$ decreases $I_{r,sp}$, and –CN has little influence on $I_{r,sp}$. The order of the contributions of various substituents to the specific impulse is: $-NO_2 > -NF_2 \approx -ONO_2 > -NC > -CN > -NH_2$.

By all accounts, the detonation properties, sensitivities, and specific impulses of $-NO_2$ and $-NF_2$ derivatives are superior to those of DDNP, so they are potential candidates for micro rocket propellants. We hope that this work will provide some useful information for those who are looking for novel solid propellants for micro rockets, and will help them to save experimental time and resources.

**Acknowledgments** We thank the National Natural Science Foundation of China (grant no. 11076017) for their support of this work. Yan Liu gratefully thanks Professor Bulat for providing the related program.

## References

1. Rossi C, Briand D, Dumonteuil M, Camps T, Pham PQ, Rooij NF (2006) Sens Actuator A 126:241–252
2. Cass S (2001) IEEE Spectr 7:56–61
3. Tanaka S, Hosokawa R, Tokudome S, Hori K, Saito H, Watanabe M, Esashi M (2003) Trans Japan Soc Aero Space Sci 46:47–51
4. Takahashi K, Okada T, Ikuta T, Nagayama K, Yamada Y, Mitarai Y (2006) In: Proc Conf on Aerospace Propulsion, Beijing, China April 20-30, 2006, CD-ROM AJCPP2006-22174
5. Takahashi K, Okada T, Ikuta T, Nishiyama T, Nagayama K, Yamada Y, Mitarai Y (2006) Proc 25th Int Symp on Space Technology and Science, Kanazawa, Japan, 4–11 June 2006, pp 155–160
6. Espinosa H, Hrabe N, Hung A, Mehling J, Merkle A (2002) http://clifton.mech.northwestern.edu/~me381/project/02fall/Microrockets.pdf
7. Orieux S, Rossi C, Estève D (2002) Sens Actuator A 101:383–391
8. Teasdale D, Pister KSJ (2000) Masters report. University of California, Berkeley (see http://bsac.eecs.berkeley.edu)
9. Larangot B, Rossi C, Camps T, Berthold A, Pham PQ, Briand D, Rooij NF, Puig-Vidal M, Miribel P, Montané E, López E, Samitier J (2002) Solid propellant micro rockets—towards a new type of power MEMS. In: Proc Nanotech AIAA, Houston, TX, USA, 9–12 Sept 2002
10. Chaalane J, Rossi C, Estève D (2006) Nanotechnology 3:340–343
11. Sudweeks WB, Chen FF, McPherson MD (2007) Chemical explosives and rocket propellants. In: Kent JA (ed) Kent and

Riegel's handbook of industrial chemistry and biotechnology. Springer, Berlin, pp 1742–1793
12. Takahashi K, Ikuta T, Okada T, Nagayama K (2006) In: 6th Int Workshop on Micro and Nanotechnology for Power Generation and Energy Conversion Applications, Berkeley, CA, USA, 29 Nov–1 Dec 2006, pp 65–68
13. Takahashi K (2006) In: Proc CANEUS 2006 Conf Micro-Nano Technologies for Aerospace Applications, Toulouse, France, 27 Aug–1 Sept 2006, pp 1–5
14. Zhang K, Chou SK, Ang S (2007) J Micromech Microeng 17:322–332
15. Okada T, Yamada Y, Ebisuzaki H, Ikuta T, Takahashi K (2004) In: Proc PowerMEMS, Kyoto, Japan, 28–30 Nov 2004, pp 56–59
16. Yang ZW, Liua YC, Liu DC, Yan LW, Chen J (2010) J Hazard Mater 177:938–943
17. Griefs P (1858) Justus Liebigs Ann Chem 106:123–125
18. Clark LV (1933) Ind Eng Chem 25:663–669
19. Chaalane J, Rossi C, Estève D (2007) Sens Actuator A 138:161–166
20. Okada T, Yamada Y, Mitarai Y, Ikuta T, Nagayama K, Takahashi K (2005) In: Proc PowerMEMS, Tokyo, Japan, 28–30 Nov 2005, pp 161–164
21. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Zakrzewski VG, Montgomery JA, Stratmann RE, Burant JC, Dapprich S, Millam JM, Daniels AD, Kudin KN, Strain MC, Farkas O, Tomasi J, Barone V, Cossi M, Cammi R, Mennucci B, Pomelli C, Adamo C, Clifford S, Petersson GA Ochterski J, Ayala PY, Cui Q, Morokuma K, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Cioslowski J, Ortiz JV, Baboul AG, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Gomperts R, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Gonzalez C, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Andres JL, Gonzalez C, Head-Gordon M, Replogle ES, Pople JA (2004) Gaussian 03, revision C.02. Gaussian Inc., Wallingford
22. Reed AE, Curtiss LA, Weinhold F (1988) Chem Rev 88:899–926
23. Politzer P, Murray JS (1998) J Mol Struct THEOCHEM 425:107–114
24. Bulat FA, Toro-Labbé A, Brinck T, Murray JS, Politzer P (2010) J Mol Model 16:1679–1691
25. Brinck T, Murray JS, Politzer P (1992) Mol Phys 76:609–617
26. Murray JS, Brinck T, Lane P, Paulsen K, Politzer P (1994) J Mol Struct THEOCHEM 307:55–64
27. Gong XD (2007) Potden v.2.0, Nanjing University of Science and Technology, Nanjing
28. Liu Y, Gong XD, Wang LJ, Wang GX, Xiao HM (2011) J Phys Chem A 115:1754–1762
29. Xu XJ, Xiao HM, Ju XH, Gong XD, Zhu WH (2006) J Phys Chem A 110:5929–5933
30. Qiu L, Xiao HM, Gong XD, Ju XH, Zhu WH (2006) J Phys Chem A 110:3797–3807
31. Wang GX, Shi CH, Gong XD, Xiao HM (2009) J Phys Chem A 113:1318–1326
32. Politzer P, Martinez J, Murray JS, Concha MC, Toro-Labbé A (2009) Mol Phys 107:2095–2101
33. Accelrys, Inc. (2008) Materials Studio 4.4. Accelrys, Inc., San Diego
34. Kamlet MJ, Jacobs SJ (1968) J Chem Phys 48:23–35
35. Mayer R (1987) Explosives. VCH, Weinheim
36. Zheng W, Wong NB, Liang X, Long X, Tian A (2004) J Phys Chem A 108:840–847
37. Zheng W, Wong NB, Wang W, Zhou G, Tian A (2004) J Phys Chem A 108:97–106
38. Holl G, Klapotke TM, Polborn K, Rienacker C (2003) Propellants Explos Pyrotech 28:153–156
39. Harcourt RD (1982) Qualitative valence bond descriptions of electron-rich molecules (Lecture Notes in Chemistry, vol 30). Springer, Berlin
40. Sikder AK, Sikder N (2004) J Hazard Mater 112:1–15
41. Rice BM, Byrd EFC (2006) J Mater Res 21:2444–2452
42. Rice BM, Hare J (2002) J Phys Chem A 106:1770–1783
43. Li JS (2010) J Hazard Mater 174:728–733
44. Rice BM, Sahu S, Owens FJ (2002) J Mol Struct Theochem 583:69–72
45. Owens FJ, Jayasuriya K, Abrahmsen L, Politzer P (1985) Chem Phys Lett 116:434–438
46. Murray JS, Lane P, Politzer P, Bolduc PR (1990) Chem Phys Lett 168:135–139
47. Murray JS, Lane P, Politzer P (1995) Mol Phys 85:1–8
48. Politzer P, Murray JS (1995) Mol Phys 86:251–255
49. Politzer P, Murray JS (1996) J Mol Struct 376:419–424
50. Murray JS, Lane P, Politzer P (1998) Mol Phys 93:187–194
51. Lowe-Ma CK, Nissan RA, Wilson WS, Houk KN, Wang X (1988) J Chem Res 214:1740–1760

J Mol Model (2012) 18:1573–1582
DOI 10.1007/s00894-011-1190-5

ORIGINAL PAPER

# Docking-based 3D-QSAR analyses of pyrazole derivatives as HIV-1 non-nucleoside reverse transcriptase inhibitors

**Elena Cichero · Paola Fossa**

Received: 18 May 2011 / Accepted: 14 July 2011 / Published online: 30 July 2011


**Abstract** 1,3,4,5-tetrasubstituted-pyrazoles (TPs) have been recently identified as a new class of potent non-nucleoside HIV-1 reverse transcriptase (RT) inhibitors. A computational strategy based on molecular docking studies, followed by docking-based comparative molecular fields analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA), has been used to elucidate the atomic details of the RT/TP interactions and to identify the most important features impacting the TP antiretroviral activity. The final CoMSIA model resulted to be the more predictive, showing $r_{ncv}^2$=0.97, $r_{cv}^2$=0.723, SEE=0.248, F=240.291, and $r^2_{pred}$=0.77. The results allowed us to obtain useful information for the design of new compounds with improved potency toward WT HIV-1 and also against clinically relevant resistant mutants.

**Keywords** 3D-QSAR · CoMFA · CoMSIA · Docking · HIV-1 · Pyrazoles · Reverse transcriptase

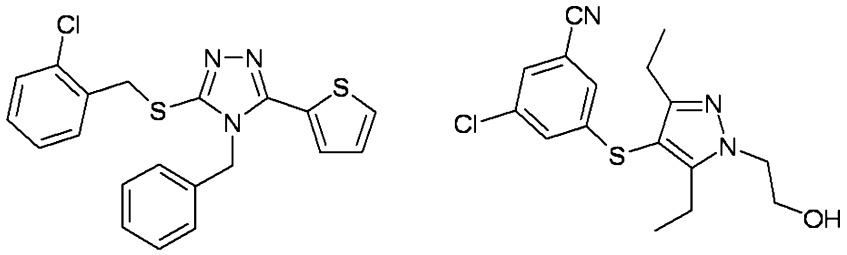
## Introduction

Reverse transcriptase (RT) is a key enzyme in the HIV replication cycle and is one of the main targets in the development of drugs for treating HIV-infection and AIDS [1–5]. RT catalyzes the conversion of viral RNA into double stranded DNA that is then integrated in the host genome. Non-nucleoside RT inhibitors (NNRTIs) bind to an allosteric hydrophobic pocket (NNRTI binding site: NNIBS), located at about 10 Å far from the polymerase active site and created upon inhibitor interaction, and lock the enzyme into an inactive form by affecting the geometry of the polymerase active site aspartyl residues [6]. In the past 15 years, more than 50 structurally diverse NNRTIs have been described [6–12]. The fact that cross-resistance extends to the whole NNRTI class calls for development of new agents capable of inhibiting clinically relevant NNRTI-resistant mutants [13, 14].

Up to now, several heterocyclic scaffolds have been studied and developed as NNRTI, including triazoles and pyrazoles [15–19]. In particular, in 2009 Kirschberg and co-workers published a series of triazole derivatives as NNRTI and also discussed the X-ray coordinates of RT in complex with 4-benzyl-3-[(2-chlorobenzyl)sulfanyl]-5-thien-2-yl-4*H*-1,2,4-triazole (Fig. 1, I) [16]. Unfortunately, the series of compounds described by the authors was characterized by low potency levels (micromolar log order) toward the WT. As previously reported by us, the low affinity value of that series of compounds could probably be elucidated on the basis of the (weak) interactions between the substituents located at the triazole position 3 and the enzyme [20]. According to our results, the absence of H-bond acceptor moieties at that triazole position resulted to be detrimental for the RT inhibitor activity. As concerns the pyrazole core, 1,3,4,5-tetrasubstituted-pyrazoles (TPs), enriched by H-bond acceptor moieties, have been recently described as a novel class of potent NNRTIs [17–19] (Fig. 1, compound **36** is depicted as the most potent of the series).

Thus, with the aim to elucidate the most important RT/TP interactions and to identify the features really impacting the TP antiretroviral activity, we performed a computational study on a dataset of 52 pyrazoles. Our aim was also to elaborate a quantitative structure-activity relationship (QSAR) model for enabling the TP antiretroviral activity

E. Cichero (✉) · P. Fossa
Dipartimento di Scienze Farmaceutiche,
Università degli Studi di Genova,
Viale Benedetto XV n.3,
16132 Genova, Italy
e-mail: cichero@unige.it




**I** ($pEC_{50}$ = 6.36) **36** ($pEC_{50}$ = 7.47)

**Fig. 1** Chemical structure of 4-benzyl-3-[(2-chlorobenzyl)sulfanyl]-5-thien-2-yl-4*H*-1,2,4-triazole (**I**) and of the pyrazole derivative **36** (showing the highest $pEC_{50}$ value in the TP series)

prediction prior to synthesis, and to obtain useful suggestions for the design of new TPs with improved potency also against clinically relevant resistant mutants.

On the basis of the crystal structure of RT in complex with **I** [16], a flexible docking simulation was performed on compounds **1**–**52**. The most probable docking poses, which resulted to be aligned in the NNIBS, were selected and submitted to the following docking-based 3D-QSAR studies, involving comparative molecular fields analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA). The results allowed us to obtain useful information for the design of new compounds with improved potency toward WT HIV-1 or potentially active against clinically relevant resistant mutants.

## Materials and methods

### Data set

A dataset of 52 pyrazoles, which were tested on the basis of the same assay, was selected from literature [17–19]. Compounds **1**–**52** (Table 1) were built, parameterized (Gasteiger-Hückel method) and energy minimized within MOE using MMFF94 force field [21].

### Docking-based ligand alignment

To locate the appropriate binding orientations and conformations of pyrazole derivatives within the NNIBS, a computational searching method was applied. Starting from a database including all 52 compounds, a docking procedure was performed. Thus, on the basis of the three-dimensional structure co-ordinates of the RT/**I** complex (PDB entry 2RKI) [16] (**I** in the bioactive conformation), each inhibitor was docked into the NNIBS using the flexible docking module implemented in Surflex.

Surflex-dock scores are expressed in -log10(Kd) units to represent binding affinities [22].

Since for all compounds the best-docked geometries were in agreement with the crystallographic data of the RT/**I** complex (and thus already aligned), they were directly submitted to CoMFA [23] and CoMSIA [24] studies by Sybyl-X 1.0 software [25].

### 3D- QSAR analyses

#### *Training set and test set*

All the compounds were grouped into a training set, for model generation, and a test set, for model validation, containing 40 and 12 compounds respectively. Both the training and the test set were divided manually according to a representative range of biological activities and structural variations. For QSAR analysis, $EC_{50}$ values have been transformed into $pEC_{50}$ values and then used as response variables. Compound $pEC_{50}$ range covered 4 log orders of magnitude.

#### *CoMFA and CoMSIA interaction energies*

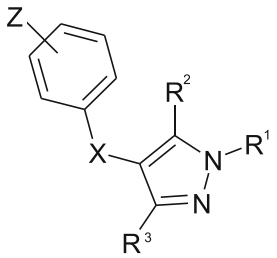
CoMFA method [23] is a widely used 3D-QSAR technique to relate the biological activity of a series of molecules to their steric and electrostatic fields, which are calculated placing the aligned molecules, one by one, into a 3D cubic lattice with a 2 Å grid spacing. The van der Waals potential and Coulombic terms, which represent steric and electrostatic fields, respectively, were calculated using the standard Tripos force field method. The column-filtering threshold value was set to 2.0 kcal $mol^{-1}$ to improve the signal-noise ratio. A methyl probe with a +1 charge was used to calculate the CoMFA steric and electrostatic fields. A 30 kcal $mol^{-1}$ energy cut-off was applied to avoid infinity of energy values inside the molecule. The CoMSIA method [24] calculates five descriptors, namely steric, electrostatic and hydrophobic parameters, and the H-bond donor and H-bond acceptor properties. The similarity index descriptors were calculated using the same lattice box employed for the CoMFA calculations and a $sp^3$ carbon as probe atom with a +1 charge, +1 hydrophobicity and +1 H-bond donor and +1 H-bond acceptor properties.

#### *Partial least square (PLS) analysis and models validation*

The partial least-squares (PLS) approach, an extension of the multiple regression analysis, was used to derive the 3D-

**Table 1** Molecular structure of pyrazole derivatives **1**–**52**

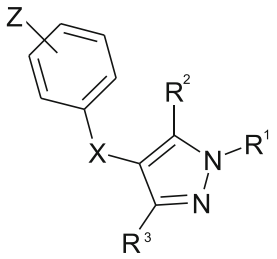
**a**




| Compound | $R_1$ | $R_2$ | $R_3$ | X | Z | $pEC_{50}$ |
|---|---|---|---|---|---|---|
| **1** | H | Et | Et | O | 3-CN | 6.05 |
| **2** | H | Et | Et | O | 3-CN, 5-Me | 7.13 |
| **3** | H | Et | Et | O | 3-CN, 5-F | 6.00 |
| **4** | H | Et | Et | O | 3,5-diCN | 6.62 |
| **5** | H | Et | Et | O | 3-CN, 5-Cl | 6.89 |
| **6** | $CH_2CH_2NH_2$ | Et | Et | O | 3-CN | 6.22 |
| **7** | $CH_2CH_2NH_2$ | Et | Et | O | 3-CN, 5-Me | 7.40 |
| **8** | $CH_2CH_2NH_2$ | Et | Et | O | 3-CN, 5-F | 6.59 |
| **9** | $CH_2CH_2NH_2$ | Et | Et | O | 3,5-diCN | 6.51 |
| **10** | $CH_2CH_2NH_2$ | Et | Et | O | 3-CN, 5-Cl | 7.37 |
| **11** | $CH_2CH_2OH$ | Me | Me | $CH_2$ | 3,5-diCl | 5.68 |
| **12** | $CH_2CH_2OH$ | Me | Me | O | 3,5-diCl | 5.96 |
| **13** | $CH_2CH_2OH$ | Me | Me | S | 3,5-diCl | 6.44 |
| **14** | $CH_2CH_2OH$ | Me | Me | $SO_2$ | 3,5-diCl | 4.82 |
| **15** | $CH_2CH_2OH$ | Me | OMe | $CH_2$ | 3,5-diCl | 5.20 |
| **16** | $CH_2CH_2OH$ | Me | OEt | $CH_2$ | 3,5-diCl | 4.27 |
| **17** | $CH_2CH_2OH$ | Me | $Pr^i$ | $CH_2$ | 3,5-diCl | 5.19 |
| **18** | $CH_2CH_2OH$ | Et | Et | $CH_2$ | 3,5-diCl | 6.18 |
| **19** | $CH_2CH_2OH$ | Et | Et | O | 2-F | 5.48 |
| **20** | $CH_2CH_2OH$ | Et | Et | O | 3-F | 5.92 |
| **21** | $CH_2CH_2OH$ | Et | Et | O | 3-CN | 6.46 |
| **22** | $CH_2CH_2OH$ | Et | Et | O | 2-Cl | 5.42 |
| **23** | $CH_2CH_2OH$ | Et | Et | O | 3-Cl | 6.25 |
| **24** | $CH_2CH_2OH$ | Et | Et | O | 3,5-diMe | 6.29 |
| **25** | $CH_2CH_2OH$ | Et | Et | O | 4-F,3-Me | 4.62 |
| **26** | $CH_2CH_2OH$ | Et | Et | O | 2,3-diF | 5.59 |

**Table 1** (continued)

**b**




| Compound | $R_1$ | $R_2$ | $R_3$ | X | Z | $pEC_{50}$ |
|---|---|---|---|---|---|---|
| **27** | $CH_2CH_2OH$ | Et | Et | O | 2,5-diF | 6.09 |
| **28** | $CH_2CH_2OH$ | Et | Et | O | 2,6-diF | 6.02 |
| **29** | $CH_2CH_2OH$ | Et | Et | O | 3,5-diF | 6.22 |
| **30** | $CH_2CH_2OH$ | Et | Et | O | 3-CN,5-Me | 7.38 |
| **31** | $CH_2CH_2OH$ | Et | Et | O | 4-F,3-CN | 4.46 |
| **32** | $CH_2CH_2OH$ | Et | Et | O | 5-F,3-CN | 7.09 |
| **33** | $CH_2CH_2OH$ | Et | Et | O | 3,5diCN | 6.92 |
| **34** | $CH_2CH_2OH$ | Et | Et | O | 4-CN-2,6di Me | 5.49 |
| **35** | $CH_2CH_2OH$ | Et | Et | O | 3-CN,4-Cl | 4.51 |
| **36** | $CH_2CH_2OH$ | Et | Et | O | 3-CN,5-Cl | 7.47 |
| **37** | $CH_2CH_2OH$ | Et | Et | O | 2,3diCl | 5.37 |
| **38** | $CH_2CH_2OH$ | Et | Et | O | 2,5 diCl | 5.92 |
| **39** | $CH_2CH_2OH$ | Et | Et | O | 2,6 diCl | 5.21 |
| **40** | $CH_2CH_2OH$ | Et | Et | O | 3,4 diCl | 4.51 |
| **41** | $CH_2CH_2OH$ | Et | Et | O | 3,5 diCl | 6.92 |
| **42** | $CH_2CH_2OH$ | Et | Et | CO | 3,5 diCl | 5.70 |
| **43** | $CH_2CH_2OH$ | Et | Et | S | 3,5 diCl | 7.19 |
| **44** | $CH_2CH_2OH$ | Et | Et | CHOMe | 3,5 diCl | 5.52 |
| **45** | $CH_2CH_2OH$ | Et | $NMe_2$ | $CH_2$ | 3,5 diCl | 5.01 |
| **46** | $CH_2CH_2OH$ | Et | $CF_3$ | $CH_2$ | 3,5 diCl | 6.27 |
| **47** | $CH_2CH_2OH$ | Et | NHCOOEt | $CH_2$ | 3,5 diCl | 4.11 |
| **48** | $CH_2CH_2OH$ | Et | $NHCOCH_2OMe$ | $CH_2$ | 3,5 diCl | 4.26 |
| **49** | $CH_2CH_2OH$ | $NMe_2$ | Me | $CH_2$ | 3,5 diCl | 4.44 |
| **50** | $CH_2CH_2OH$ | $CF_3$ | Me | $CH_2$ | 3,5 diCl | 6.04 |
| **51** | $CH_2CH_2OH$ | $CF_3$ | Et | $CH_2$ | 3,5 diCl | 6.20 |
| **52** | $CH_2CH_2OH$ | COOEt | Me | $CH_2$ | 3,5 diCl | 4.57 |

QSAR models. CoMFA and CoMSIA descriptors were used as independent variables and $pEC_{50}$ values were used as dependent variables.

The leave one out (LOO) cross-validation method was used to check the predictivity of the derived model and to identify the optimal number of components (ONC) leading to the highest cross-validated $r^2$ ($r^2_{cv}$). In the LOO methodology, one molecule is omitted from the dataset and a model is derived involving the rest of the compounds. Employing this model, the activity of the omitted molecule is then predicted.

The ONC obtained from cross-validation methodology was used in the subsequent regression model. Final CoMFA and CoMSIA models were generated using non-cross-validated PLS analysis. To further assess the statistical confidence and robustness of the derived models, a 100-cycle bootstrap analysis was performed. This is a procedure in which *n* random selections out of the original set of *n* objects are performed several times (100-times were required to obtain good statistical information). In each run, some objects may not be included in the PLS analysis, whereas some others might be included more then once. The mean correlation coefficient is represented as bootstrap $r^2$($r^2_{boot}$).

*Predictive correlation coefficient ($r^2_{pred}$)*

To further validate the CoMFA and CoMSIA derived model, the predictive ability for the test set of compounds (expressed as $r^2_{pred}$) was determined by using the following equation:

$$r^2_{pred} = (SD - PRESS)/SD. \quad (1)$$

SD is the sum of the squared deviations between the biological activities of the test set molecules and the mean activity of the training set compounds and PRESS is the sum of the squared deviation between the observed and the predicted activities of the test set compounds.

All calculations were carried out using a PC running the Windows XP operating system and an SGI O2 Silicon Graphics.

## Results and discussion

### Docking-based ligand alignment

In this work, for the molecular modeling analysis we used the X-ray co-ordinates of the crystal structure of **I**/RT complex. As shown in Fig. 2a, according to the experimental data, the triazole **I** displays H-bond contact between the N1 nitrogen atom and the K103 ε-amino group, and Van der Waals and π–π interactions with the following three hydrophobic pockets: (*i*) $P_1$, including the residues K104, V106, F227, L234, H235, P236 and Y318 (*ii*) P2, consisting of P95, P97, Y181, Y188 and W229 residues, (*iii*) and P3 delimited by L100, V179, I180, V189, G190. As far as TP derivatives are concerned, according to our calculations, all the inhibitors display at least an H-bond between the K103 ε-amino group and the N2 atom of the pyrazole ring, while several van der Waals and π–π interactions with the three hydrophobic pockets $P_1$, $P_2$ and $P_3$ are detected (Fig. 2b). In details, region $P_1$ is properly occupied by the TP R1 and R2 substituents, while $P_2$ is occupied by the TP phenyl ring. The R3 substituent is oriented toward the residues belonging to the P3 pocket.

Furthermore, compound **36** (the most active of the series, $pEC_{50}$=7.47) displays two additional H-bond contacts between: (i) the R1 hydroxyl group and the K103 carbonyl moiety, (ii) the phenyl cyano group and the Y188 sidechain hydroxyl. Accordingly, compound **36** displays higher $pEC_{50}$ value of the reference compound **I**.

In order to validate the docking procedure, we compared the most probable binding pose, selected for each compound, with the X-ray structure of **I** in complex with RT, obtaining ligand conformations perfectly superposed on the available crystallographic data. Thus, they resulted to be aligned together inside the NNIBS for the subsequent 3D-QSAR analysis.

### 3D-QSAR analyses

CoMFA and CoMSIA analyses were performed dividing compounds **1**–**52** into a training set (**2, 4–10, 12–17, 19–22, 24, 27–29, 32–39, 41, 43–45, 47–52**) for model generation and into a test set (**1, 3, 11, 18, 23, 25, 26, 30, 31, 40, 42, 46**) for model validation.

The final CoMFA model was generated employing non-cross-validated PLS analysis with the optimum number of components (ONC=6) to give a non-cross validated $r^2$ ($r^2_{ncv}$)=0.95, a test set $r^2$ ($r^2_{pred}$)=0.71, standard error of estimate (SEE)=0.276, steric contribution=0.440 and electrostatic contribution=0.560. The reliability of the model thus generated was supported by bootstrapping results. All statistical parameters supporting CoMFA model are reported in Table 2. The final CoMSIA model consisting of steric, electrostatic, hydrophobic, H-bond donor and H-bond acceptor fields with a $r^2_{ncv}$=0.97, SEE=0.248, steric contribution=0.086, electrostatic contribution=0.285, hydrophobic contribution=0.210, H-bond acceptor contribution=0.214 and H-bond donor contribution=0.204 was derived. All statistical parameters supporting CoMSIA model are reported in Table 2.

Distribution of experimental and predicted $pEC_{50}$ values for training set and for test set compounds according to CoMFA and CoMSIA analyses are reported in Fig. 3.

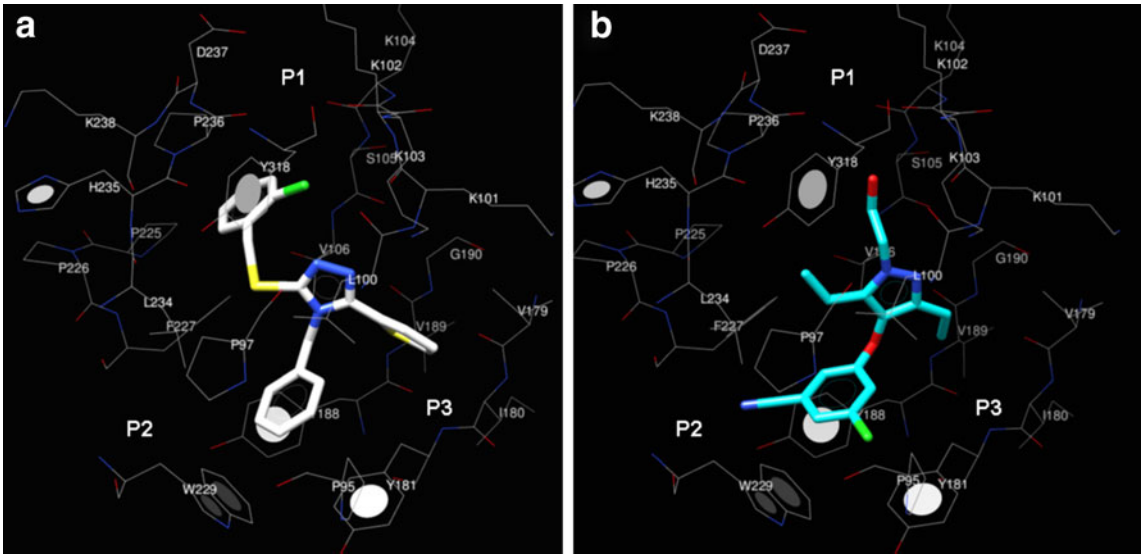



**Fig. 2** X-ray pose of compound **I** into the NNIBS (**a**) and the selected docking pose of **36** (showing the highest $pEC_{50}$ value in the TP series) into the NNIBS

*CoMFA steric and electrostatic regions*

As shown in Fig. 4a (for simplicity, only the structure of compound **36** is depicted as representative), the steric contour map predicts favorable interaction polyhedra (green) in proximity of the two *ortho* and both the two *meta* positions of the phenyl moiety (at the pyrazole position 4) and around the methyl located on the R2 and R3 ethyl substituents. On the contrary, unfavorable polyhedra (yellow) are positioned near the *para* position of the phenyl ring and in the area located between the positions 3 and 4 of the pyrazole ring.

Accordingly, compounds **31** (Z=4-F,3-CN; $pEC_{50}$=4.46) and **35** (Z=3-CN,4-Cl; $pEC_{50}$=4.51) show higher potency than **32** (Z=5-F,3-CN; $pEC_{50}$=7.09) and **36** (Z=3-CN,5-Cl; $pEC_{50}$=7.47). These results are also in agreement with the higher $pEC_{50}$ values of those compounds bearing a 3,5 di-substituted phenyl ring at the pyrazole position 4, than those inhibitors bearing a 3-substituted aromatic ring at the same pyrazole position [compare **2, 4, 5** (Z=3-CN,5-substituted; $pEC_{50}$=6.62-7.13), **7–10** (Z=3-CN,5-substituted; $pEC_{50}$=6.34-7.24), **30, 32, 33, 36** (Z=3-CN,5-substituted; $pEC_{50}$=6.92 – 7.47), **29**, **32** (Z=3-substituted,5-F; $pEC_{50}$=6.22 – 7.09) and **41** (Z=3,5-diCl; $pEC_{50}$=6.22 – 7.09) with compounds **1** (Z=3-CN; $pEC_{50}$=6.05), **6** (Z=3-CN; $pEC_{50}$=6.22), **21** (Z=3-CN; $pEC_{50}$=6.46), **20** (Z=3-F; $pEC_{50}$=5.92) and **23** (Z=3-Cl; $pEC_{50}$=6.25), respectively].

The reliability of the CoMFA steric map is also verified by the higher $pEC_{50}$ values of those pyrazoles

**Table 2** Summary of CoMFA and CoMSIA results

| | CoMFA model | CoMSIA model |
|---|---|---|
| No. compounds | 40 | 40 |
| Optimal number of components (ONC) | 5 | 5 |
| Leave one out $r^2$ ($r^2_{loo}$) | 0.603 | 0.622 |
| Cross-validated $r^2$ ($r^2_{cv}$) | 0.678 | 0.723 |
| Std. error of estimate (SEE) | 0.276 | 0.248 |
| Non cross-validated $r^2$ ($r^2_{ncv}$) | 0.95 | 0.97 |
| F value | 191.370 | 240.291 |
| Steric contribution | 0.440 | 0.086 |
| Electrostatic contribution | 0.560 | 0.285 |
| Hydrophobic contribution | - | 0.210 |
| H-bond acceptor contribution | - | 0.214 |
| H-bond donor contribution | - | 0.204 |
| Bootstrap $r^2$ ($r^2_{boot}$) | 0.97 | 0.98 |
| Standard error of estimate $r^2_{boot}$ (SEE $r^2_{boot}$) | 0.185 | 0.182 |
| Test set $r^2$ ($r^2_{pred}$) | 0.72 | 0.77 |

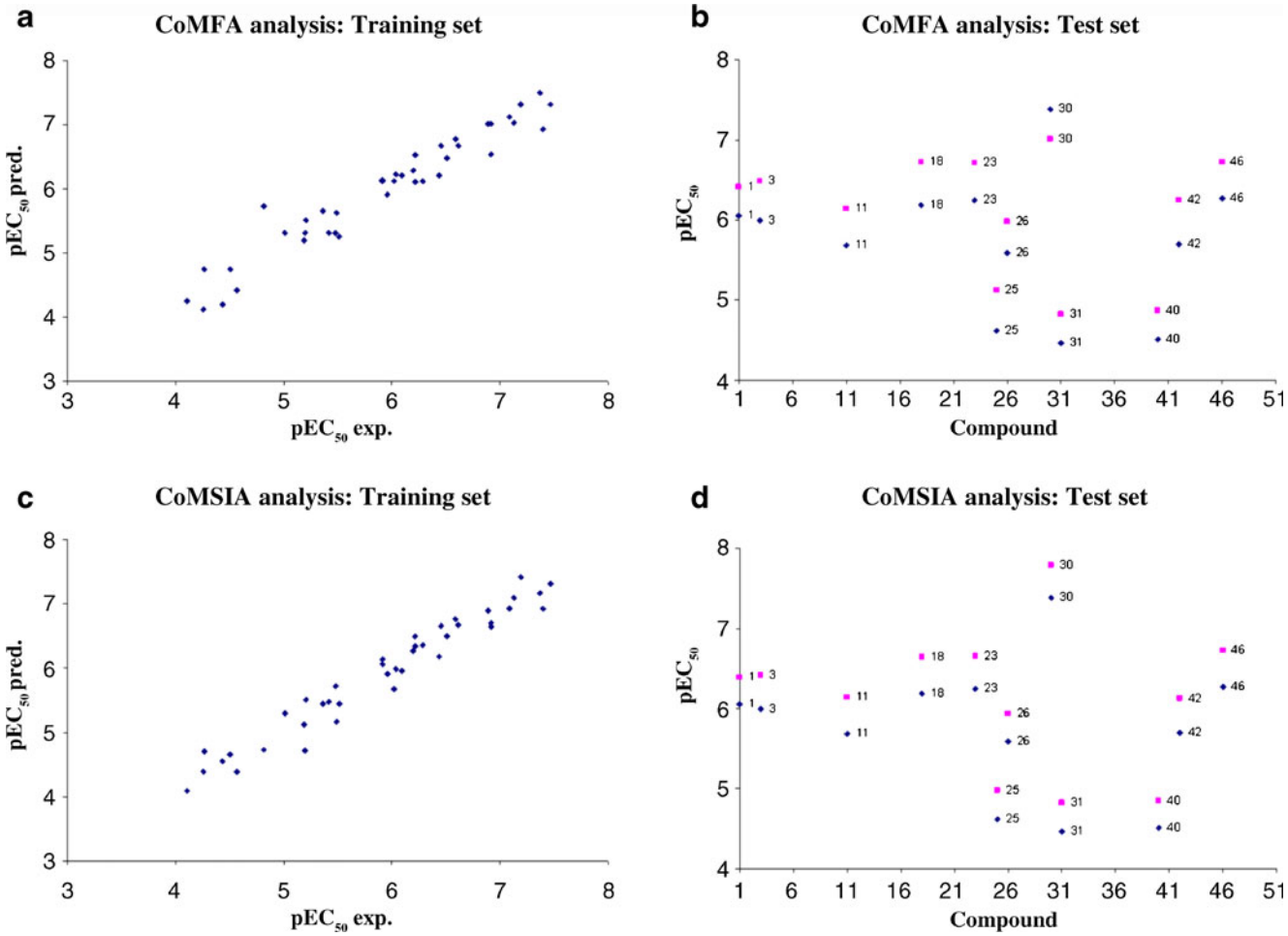



**Fig. 3** Distribution of experimental and predicted $pEC_{50}$ values for training set compounds according to CoMFA analysis (**a**), for test set compounds according to CoMFA analysis (**b**), for training set compounds according to CoMSIA analysis (**c**), and for test set compounds according to CoMSIA analysis (**d**)

3,5-diethyl-substituted than the corresponding compounds 3,5-dimethyl-substituted.

On the basis of the electrostatic fields contour map of the CoMFA analysis plotted in Fig. 4b (for simplicity, only the structure of compound **36** is depicted as representative), less positive moieties would be favored (red polyhedra) in proximity of one of the two *meta* positions of the phenyl ring, around the linker located between the phenyl ring and

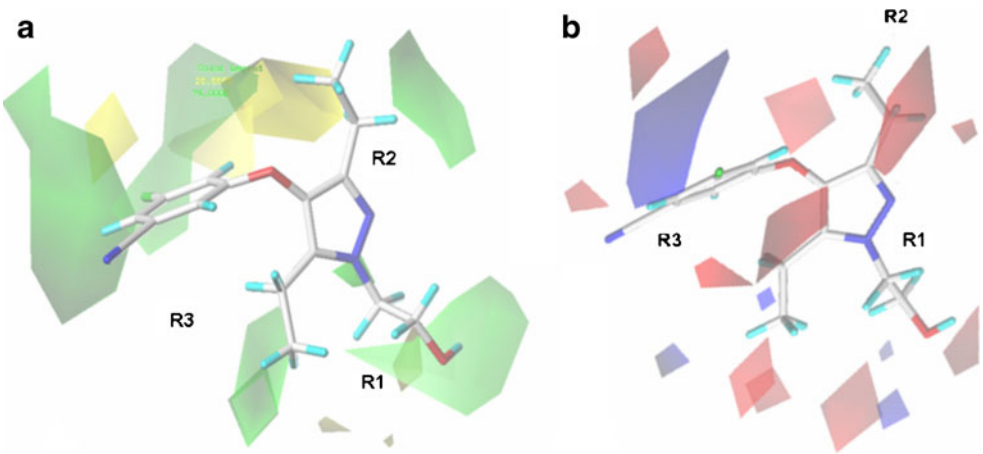



**Fig. 4** Contour maps of CoMFA steric regions (**a**) (green, favored; yellow, disfavored) and CoMFA electrostatic areas (**b**) are displayed around compound **36**. Blue regions are favorable for more positively charged groups; red regions are favorable for less positively charged groups. Ligands are depicted in stick mode and colored by atom type

the pyrazole core, and near the R1 hydroxyl moiety. On the other hand, more electropositive substituent are predicted to be beneficial (blue area) around the R2 and R3 substituents and near the two phenyl *ortho* positions.

These results are supported by the higher $pEC_{50}$ values of those pyrazoles bearing an alkyl group on the R2 or R3 substituents, than those bearing hetero atoms at the corresponding positions [compare **11** (R2=R3=CH3; $pEC_{50}$=5.68) with the 3-alkoxy substituted pyrazole derivatives **15**, **16** (R2=CH3; $pEC_{50}$=4.27 - 5.20), and with **49** (R2=N(CH3)$_2$; $pEC_{50}$=4.44) and **52** (R2=COOEt; $pEC_{50}$=4.57)]. Accordingly, compound **18** (R2=R3=CH3; $pEC_{50}$=6.18) displays higher $pEC_{50}$ than **45**, **47**, **48** ($pEC_{50}$=4.11 – 5.01), bearing amino, carbamidic and amide functional groups at the pyrazole position 3.

The CoMSIA steric and electrostatic regions are in agreement with the CoMFA steric and electrostatic areas.

### *CoMSIA hydrophobic, H-bond acceptor and H-bond donor regions*

As shown in Fig. 5 (for simplicity, only the structure of compound **36** is depicted as representative), the introduction of hydrophobic substituents in proximity of the R2 and R3 substituents and around one of the two meta positions of the phenyl ring is favored (yellow polyhedra). On the contrary, the presence of hydrophilic groups is recommended (white polyhedra) around the other *meta* and one of the two *ortho* positions, and near the linker located between the phenyl ring and the pyrazole core (namely, X). The reliability of the hydrophobic map calculation is verified by the higher $pEC_{50}$ values of compounds **18** (R1=R2=Et; X=$CH_2$; $pEC_{50}$=6.18), **41** (R1=R2=Et; X=O; $pEC_{50}$=6.92), and **43** (R1=R2=Et; X=S; $pEC_{50}$=7.19) in comparison with those of **11** (R1=R2=Me; X=CH2; $pEC_{50}$=5.68), **12** (R1=R2=Me; X=O; $pEC_{50}$=5.96) and **13** (R1=R2=Me; X=S; $pEC_{50}$=6.44), respectively. Accordingly, compound **50** (R1 CF3, R2=Me; $pEC_{50}$=6.04) shows higher potency in comparison with **11** (R1=R2=Me; $pEC_{50}$=5.68). These results are also supported by the high $pEC_{50}$ values of those ligand bearing one polar and one hydrophobic group linked to the meta positions of the phenyl ring.

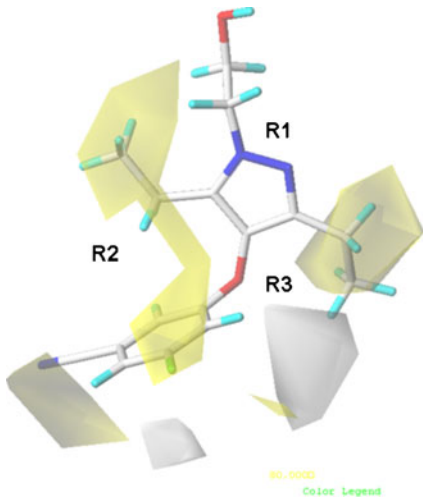



**Fig. 5** Contour maps of CoMSIA hydrophobic regions (yellow, favored; white, disfavored) are depicted around compounds **36**, shown in stick mode and colored by atom type

H-bond acceptor functions are predicted to be beneficial around the R1 hydroxyl group, the nitrogen atom at the pyrazole position 1 and near one of the two phenyl ring *meta* positions. On the contrary, the introduction of H-bond acceptor groups are disfavored in proximity of the methylene of the R2 and R3 substituents, and near the methylene linked to the R1 polar group (for compound **36**, the hydroxyl group). Accordingly, compounds **30** (Z=3-CN,5-Me, pEC50=7.38), **32** (Z=5-F,3-CN; $pEC_{50}$=7.09), **36** (Z=3-CN,5-Cl; $pEC_{50}$=7.47) show higher pEC50 values than those of compounds **24** (Z=3,5-diMe, pEC50=6.29), **29** (Z=3,5-diF; pEC50=6.22), and **41** (Z=3-CN,5-Cl; pEC50=6.92), respectively.

H-bond donors are predicted to be favorable and unfavorable near the R1 group and around the X linker, respectively.

### *A comparison between the CoMFA and CoMSIA analyses and the NNIBS/***36** *docking model*

In order to verify the reliability of the 3D-QSAR analyses, CoMFA and CoMSIA maps have been compared with the results of docking analysis. Since the CoMSIA study resulted to be more predictive, for simplicity only the CoMSIA hydrophobic and H-bond acceptor maps are reported in Figs. 6–7. Briefly, the CoMFA steric model proves to match with the NNIBS 3D topology, suggesting bulky substitution: (*i*) around the phenyl group, which is engaged in π–π interactions with the P2 aromatic residues (Y181, Y188 and W229), (*ii*) and in proximity of the R1 substituent, which resulted to be involved in hydrophobic contacts with the $P_1$ pocket (V106, F227, L234, H235, P236 and Y318). On the contrary, the R2 substituent is surrounded by a yellow area since a bulky group at this position could be in clash contacts with Y318.

Notably, W229, L234 and P236 are highly conserved aminoacids in the NNIBS and therefore they are recognized of strategic relevance for the design of new NNRTIs more resilient to the effects of RT mutations in this site [26]. With this purpose, the introduction of a ethylamide or propylamide group as R1 substituent and/or the substitution of the phenyl ring with a naphthyl one, could enhance the inhibitor potency. Furthermore, the presence of an amide function at the R1 substituent might also allow TPs to keep

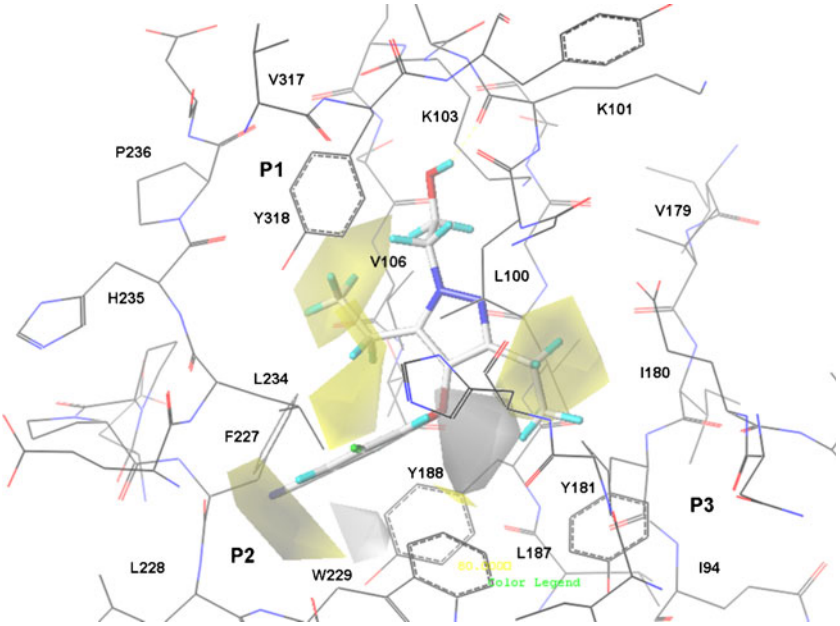



**Fig. 6** The CoMSIA hydrophobic map (yellow, favored; white, disfavored) is displayed around the compound **36** selected docking pose (in stick), located into the NNIBS

high level of activity against the clinically relevant resistant mutant K103N (the terminal amidic group of asparagine is a good H-bond donor).

The CoMFA electrostatic map points out the beneficial presence of small electropositive moieties in proximity of one of the two phenyl *meta* positions, at the R1 substituent and between the phenyl ring and the pyrazole core. Accordingly, the compound **36** R1 hydroxyl group and the cyano group are involved in H-bond contacts with K103 and Y318, respectively.

As shown in Fig. 6, the CoMSIA hydrophobic maps underline the importance of lipophilic group (yellow areas) at one of the two phenyl *meta* positions and at the TP R2 and R3 substituents, which appeared properly oriented toward the L228, L234 and to the L100 sidechain. Thus, the presence of *t*-butil substituents at the R2 and R3 substituents could be beneficial. The CoMSIA H-bond acceptor (Fig. 7) and H-bond donor maps confirm the importance of the formation of hydrogen bond interactions to stabilize the RT/TP complex. Accordingly, compound **36**

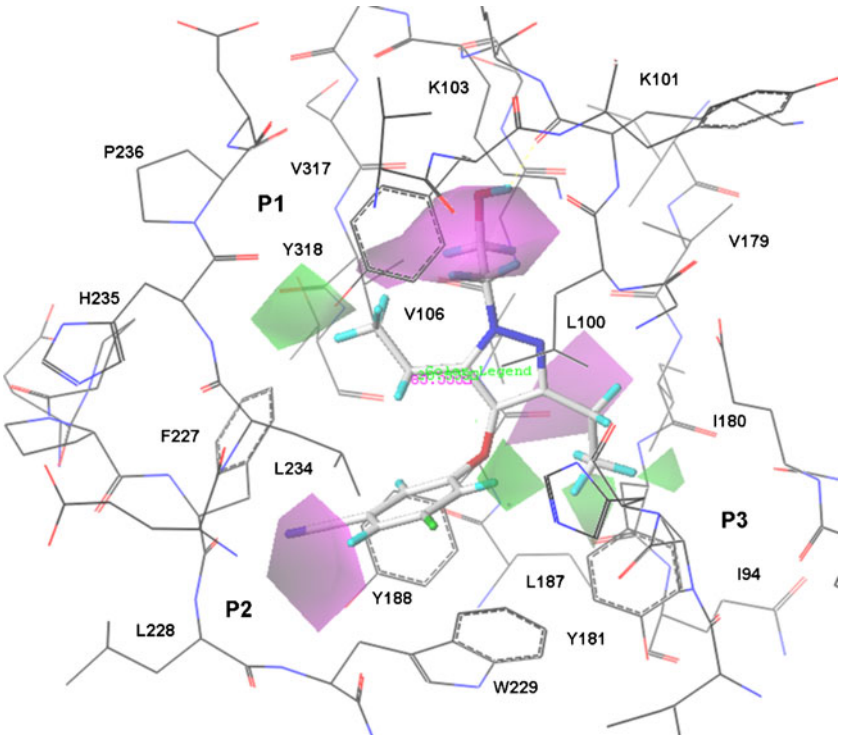



**Fig. 7** The CoMSIA H-bond acceptor map is displayed around the compound **36** selected docking pose (in stick), located in the NNIBS. H-bond acceptor groups: magenta, favored; green, disfavored

is the most active TP of the series ($pEC_{50}$ range=7.47), bearing the following three H-bond acceptor centers: (*i*) at one of the two *meta* position of the phenyl ring (H-bond with Y318), (*ii*) at the R1 substituent (H-bond with K103), (*iii*) and at the pyrazole ring (K103).

On the basis of these data, a carbonyl function, such as a methylcarbonyl group, could be introduced at one of the two phenyl *meta* positions, being engaged in H-bond with Y318.

## Conclusions

The computational studies here presented analyze the main interactions responsible for TP activity and give useful suggestions for the synthesis of new analogues with improved potency against WT HIV-1 and also against clinically relevant resistant mutants. The models elaborated will be exploited to design new TPs and predict their activity prior to synthesis.

**Acknowledgments** This work was financially supported by the University of Genova.

## References

1. Jonckheere H, Anne J, De Clercq E (2000) The HIV-1 reverse transcription (RT) process as target for RT inhibitors. Med Res Rev 20:129–154
2. De Clercq E (2001) New developments in anti-HIV chemotherapy. Farmaco 56:3–12
3. De Clercq E (2005) Emerging anti-HIV drugs. Expert Opin Emerg Drugs 10:241–273
4. Clercq E (2005) New approaches toward anti-HIV chemotherapy. Eur J Med Chem 48:1297–1313
5. Barbaro G, Scozzafava A, Mastrolorenzo A, Supuran CT (2005) Highly active antiretroviral therapy: current state of the art, new agents and their pharmacological interactions useful for improving therapeutic outcome. Curr Pharm Des 11:1805–1843
6. Balzarini J (2004) Current status of the non-nucleoside reverse transcriptase inhibitors of human immunodeficiency virus type 1. Curr Top Med Chem 4:921–944
7. De Clercq E (1998) The role of non-nucleoside reverse transcriptase inhibitors (NNRTIs) in the therapy of HIV-1 infection. Antiviral Res 38:153–179
8. Pedersen OS, Pedersen EB (1999) Non-nucleoside reverse transcriptase inhibitors, the NNRTI boom. Antiviral Chem Chemother 10:285–314
9. Pedersen OS, Pedersen EB (2000) The flourishing syntheses of non-nucleoside reverse transcriptase inhibitors. Synthesis 4:479–495
10. Campiani G, Ramunno A, Maga G, Nacci V, Fattorusso C, Catalanotti B, Morelli E, Novellino E (2002) Non-nucleoside HIV-1 reverse transcriptase (RT) inhibitors: past, present, and future perspectives. Curr Pharm Des 8:615–657
11. De Clercq E (2004) Non-nucleoside reverse transcriptase inhibitors (NNRTIs): past, present, and future. Chem Biodiversity 1:44–64
12. Pauwels R (2004) New non-nucleoside reverse transcriptase inhibitors (NNRTIs) in development for the treatment of HIV infections. Curr Opin Pharmacol 4:437–446
13. Leigh Brown AJ, Frost SD, Mathews WC, Dawson K, Hellmann NS, Daar ES, Richman DD, Little SJ (2003) Transmission fitness of drug-resistant human immunodeficiency virus and the prevalence of resistance in the antiretroviral-treated population. J Infect Dis 187:683–686
14. Richman DD, Morton SC, Wrin T, Hellmann N, Berry S, Shapiro MF, Bozzette SA (2004) The prevalence of antiretroviral drug resistance in the United States. AIDS 18:1393–1401
15. De La Rosa M, Kim HW, Gunic E, Jenket C, Boyle U, Koh YH, Korboukh I, Allan M, Zhang W, Chen H, Xu W, Nilar S, Yao N, Hamatake R, Lang SA, Hong Z, Zhang Z, Girardet JL (2006) Trisubstituted triazoles as potent non-nucleoside inhibitors of the HIV-1 reverse transcriptase. Bioorg Med Chem Lett 16:4444–4449
16. Kirschberg TA, Balakrishnan M, Huang W, Hluhanich R, Kutty N, Liclican AC, McColl DJ, Squires NH, Lansdon EB (2007) Triazole derivatives as non-nucleoside inhibitors of HIV-1 reverse transcriptase–structure-activity relationships and crystallographic analysis. Bioorg Med Chem Lett 18:1131–1134
17. Mowbray CE, Burt C, Corbau R, Perros M, Tran I, Stupple PA, Webster R, Wood A (2009) Pyrazole NNRTIs 1: design and initial optimisation of a novel template. Bioorg Med Chem Lett 19:5599–5602
18. Mowbray CE, Corbau R, Hawes M, Jones LH, Mills JE, Perros M, Selby MD, Stupple PA, Webster R, Wood A (2009) Pyrazole NNRTIs 3: optimisation of physicochemical properties. Bioorg Med Chem Lett 19:5603–5606
19. Mowbray CE, Burt C, Corbau R, Gayton S, Hawes M, Perros M, Tran I, Price DA, Quinton FJ, Selby MD, Stupple PA, Webster R, Wood A (2009) Pyrazole NNRTIs 4: selection of UK-453,061 (lersivirine) as a development candidate. Bioorg Med Chem Lett 20:5857–5860
20. Cichero E, Buffa L, Fossa P (2010) 3,4,5-Trisubstituted-1,2,4-4H-triazoles as WT and Y188L mutant HIV-1 non-nucleoside reverse transcriptase inhibitors: docking-based CoMFA and CoMSIA analyses. J Mol Model. doi:10.1007/s00894-010-0857-7
21. MOE: Chemical Computing Group Inc . Montreal. H3A 2R7 Canada. http://www.chemcomp.com
22. Jain AN (1996) Scoring noncovalent protein-ligand interactions: A continuous differentiable function tuned to compute binding affinities. J Comput Aided Mol Des 10:427–440
23. Cramer RD III, Patterson DE, Bunce JD (1989) Recent advances in comparative molecular field analysis (CoMFA). Prog Clin Biol Res 291:161–165
24. Klebe G, Abraham U, Mietzner T (1994) Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J Med Chem 37:4130–4146
25. Sybyl-X 1.0. Tripos Inc 1699 South Hanley Road. St Louis, MO, 63144, USA
26. Pauwels R (2004) New non-nucleoside reverse transcriptase inhibitors (NNRTIs) in development for the treatment of HIV infections. Curr Opin Pharmacol 5:437–446

J Mol Model (2012) 18:1583–1589
DOI 10.1007/s00894-011-1191-4

ORIGINAL PAPER

# Trimethylaurintricarboxylic acid inhibits human DNA methyltransferase 1: insights from enzymatic and molecular modeling studies

**Jakyung Yoo · José L. Medina-Franco**

Received: 2 June 2011 / Accepted: 18 July 2011 / Published online: 30 July 2011


**Abstract** DNA methyltransferase 1 (DNMT1) is an emerging target for the treatment of cancer, brain disorders, and other diseases. Currently, there are only a few DNMT1 inhibitors with potential application as therapeutic agents or research tools. 5,5-Methylenedisalicylic acid is a novel scaffold previously identified by virtual screening with detectable although weak inhibitory activity of DNMT1 in biochemical assays. Herein, we report enzyme inhibition of a structurally related compound, trimethylaurintricarboxylic acid (NSC97317) that showed a low micromolar inhibition of DNMT1 ($IC_{50}$=4.79 μM). Docking studies of the new inhibitor with the catalytic domain of DNMT1 suggest that NSC97317 can bind into the catalytic site. Interactions with amino acid residues that participate in the mechanism of DNA methylation contribute to the binding recognition. In addition, NSC97317 had a good match with a structure-based pharmacophore model recently developed for inhibitors of DNMT1. Trimethylaurintricarboxylic acid can be a valuable biochemical tool to study DNMT1 inhibition in cancer and other diseases related to DNA methylation.

**Keywords** Cancer · DNMT · Docking · Enzyme inhibition · Epigenetics · Methyltransferase

## Introduction

DNA methylation is an epigenetic change that results in the addition of a methyl group at the carbon-5 position of cytosine residues. DNA methyltransferase enzymes (DNMTs) mediate this process. To date, three types of DNMTs have been identified in the human genome, including the two de novo methyltransferases (DNMT3A and DNMT3B) and the maintenance methyltransferase (DNMT1), which is the most abundant among the three [1–3]. DNMT1 is responsible for duplicating the pattern of DNA methylation during replication, it is essential for proper mammalian development, and it has been proposed as the more interesting target for experimental cancer therapies [4]. DNA methylation represents a central mechanism for mediating epigenetic gene regulation. Thus, the development of DNMT inhibitors provides novel opportunities for cancer therapy [1, 5–7]. It is worth nothing that epigenetic alterations are not only associated with cancer but also with psychiatric and other diseases [8–10].

Human DNMT1 is a protein with 1616 amino acids whose structure can be divided into an N-terminal regulatory domain and a C-terminal catalytic domain [11–13]. The mechanism of DNA cytosine-C5 methylation is schematically depicted in Fig. 1 [4, 14, 15]. DNMT forms a complex with DNA and the cytosine which will be methylated flips out from the DNA. The thiol of the catalytic cysteine acts as a nucleophile that attacks the 6-position of the target cytosine to generate a covalent intermediate. The 5-position of the cytosine is activated and conducts a nucleophilic attack on the cofactor *S*-adenosyl-L-methionine (SAM) to form the 5-methyl covalent adduct and *S*-adenosyl-L-homocysteine (SAH). The attack on the 6-position is assisted by a transient protonation of the cytosine ring at the endocyclic nitrogen atom N3, which is stabilized by a glutamate residue. An arginine residue may participate in the stabilization of the intermediate making a hydrogen bonding interaction with the carbonyl oxygen of cytosine. The covalent complex

J. Yoo · J. L. Medina-Franco (✉)
Torrey Pines Institute for Molecular Studies,
11350 SW Village Parkway,
Port St. Lucie, FL 34987, USA
e-mail: jmedina@tpims.org

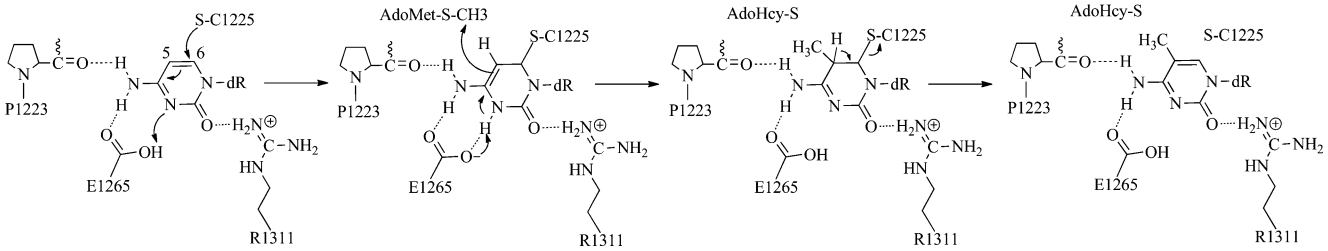



**Fig. 1** Proposed mechanism of DNA cytosine-C5 methylation. Amino acid residue numbers are based on the homology model

between the methylated base and the DNA is resolved by deprotonation at the 5-position to generate the methylated cytosine and the free enzyme.

To date, only the DNMT inhibitors 5-azacytidine and 5-aza-2′-deoxycytidine have been developed clinically. These two drugs are nucleoside analogues, which, after incorporation into DNA, cause covalent trapping and subsequent depletion of DNMTs [5, 15]. Aza nucleosides are FDA approved for the treatment of myelodysplastic syndrome, where they demonstrate significant, although usually transient improvement in patient survival and are currently being tested in many solid cancers [16–18]. 5-Aza-2′-deoxycytidine however, has relatively low specificity and is characterized by substantial cellular and clinical toxicity [5].

There are an increasing number of substances reported to inhibit DNMTs [19–21]. Representative DNMT inhibitors and other candidate demethylating agents are depicted in Fig. 2. Computational screening of chemical databases has contributed to the discovery of lead compounds with distinct scaffolds such as RG108 [22]. Similarly, molecular modeling and docking are playing a key role in the understanding of the mechanism of action of a number of DNMT inhibitors at the molecular level [23, 24].

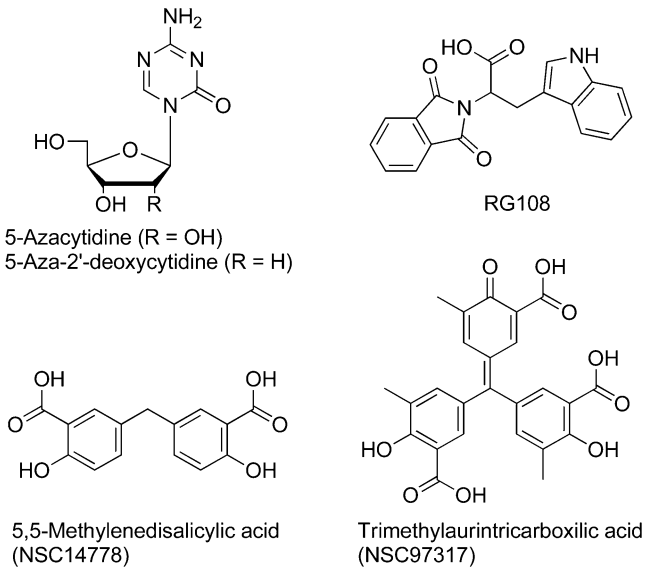



**Fig. 2** Chemical structures of representative inhibitors of DNA methyltransferases. NSC97317 is described in this work

We recently conducted a virtual screening of the National Cancer Institute (NCI) database identifying 5,5-methylenedisalicylic acid (NSC14778) (Fig. 2) as a DNMT1 inhibitor with an $IC_{50}$=92 μM [25]. Molecular docking of 5,5-methylenedisalicylic acid with an homology model of the catalytic domain of DNMT1 indicated the key role of the carboxylic acids in the binding recognition including hydrogen bonds with amino acid residues that are involved in the catalytic mechanism of methylation [25]. Recently, researchers have shown an increased interest in the biological activity of 5,5-methylenedisalicylic acid and related compounds. For example, 5,5-methylenedisalicylic acid also emerged as an experimentally validated hit of a virtual screening with the viral NS5 RNA methyltransferase, a promising drug target against flaviviruses which are the causative agents of severe diseases such as Dengue or Yellow fever [26]. Moreover, in an independent virtual screening, the aurintricarboxylic acid (which is a structural analogue of 5,5-methylenedisalicylic acid) was identified as a highly potent inhibitor of the methyltransferase activities on flaviviral methyltransferases [27]. Also, a recent study shows that aurintricarboxylic acid inhibit two AdoMet-dependant RNA methyltransferases of the severe acute respiratory syndrome (SARS) coronavirus [28].

We hypothesize that trimethylaurintricarboxylic acid (NSC97317; Fig. 2) is an inhibitor of DNMT1. To note, this compound is a structural analogue of aurintricarboxylic acid and 5,5-methylenedisalicylic acid. Moreover, NSC97317 is *freely available* at the NCI Drug Synthesis and Chemistry Branch [29]; this fact may increase the impact and applicability of the insights of this work. In order to test this hypothesis and identify a novel DNMT inhibitor, herein we report enzyme inhibition and molecular modeling studies that confirm this hypothesis.

## Methods

### Experimental

Trimethylaurintricarboxylic acid (NSC97317; Fig. 2) was obtained from the NCI Drug Synthesis and Chemistry

Branch [29]. The inhibition of the enzymatic activity of DNMT1 was tested using the HotSpot$^{SM}$ platform for methyltransferase assays available at Reaction Biology Corporation [30]. HotSpot$^{SM}$ is a low volume radioisotope-based assay which uses tritium-labeled AdoMet ($^{3}$H-SAM) as a methyl donor. NSC97317 diluted in DMSO was added by using acoustic technology (Echo550, Labcyte) into enzyme/substrate mixture in nano-liter range. The reaction was initiated by the addition of $^{3}$H-SAM, and incubated at 30°C. Total final methylations on the substrate (Poly(dI-dC)) were detected by a filter binding approach. Data analysis was performed using Graphed Prism software (La Jolla, CA) for curve fits. Reactions were carried out at 1 μM of *S*-adenosyl-L-methionine (SAM). *S*-adenosyl-L-homocysteine (SAH) is used as standard positive control in these assays. NSC97317 was tested in 10-dose $IC_{50}$ (effective concentration to inhibit DNMT1 activity by 50%) with three-fold serial dilution starting at 100 μM.

## Computational

### *Homology model*

We employed a homology model of the catalytic domain of human DNMT1 recently developed by our group; the homology model has been used to rationalize, at the molecular level, the binding mode of established DNMT1 inhibitors [31]. Despite the fact that a recent crystallographic structure for human DNMT1 has been recently published (*vide infra*), the structure corresponds to the enzyme bound to *unmethylated DNA* and this structure it is not suitable to model small-molecule inhibitors of DNMT1. This is because in the crystallographic structure the catalytic loop has an open conformation and the catalytic cysteine is far from the binding site (more than 9 Å) [32]. Therefore, the geometry of the catalytic site does not represent the catalytic mechanism of DNA methylation. Briefly, to build the homology model, the catalytic domain of the human DNMT1 was taken from the UniProt (UniProt ID: P26358) [33]. The DNMT1 sequence was aligned based on the sequence of DNA methyltransferases M.HhaI (PDB ID: 6MHT), M.HaeIII (PDB ID: 1DCT) and DNMT2 (PDB ID: 1G55) and built based on the template 3D structures using Prime (PRIME, version 2.2, Schrödinger, LLC, New York, NY, 2010). The co-factor was included in this model and the DNA double helix was constructed from the structure of M.HhaI. The variable small loops and gaps were filled by knowledge-based, homology or ab initio approach of ORCHESTRAR, and then missing long loop was modeled using Loop Search module implemented in Sybyl 8.0. The loops showing the highest homology and the lowest root mean square deviations were selected. The side chains and hydrogen atoms were added and the stability of the homology model was validated by checking the geometry using PROCHECK. The homology model coordinates were then energy minimized with Macromodel (MACROMODEL, version 9.8, Schrödinger, LLC, New York, NY, 2010) using MMFF94s force field in a water environment (until converging at a termination gradient of 0.05 kJ mol$^{-1}$-Å) and the H-bonds were fixed using the SHAKE algorithm during molecular dynamics.

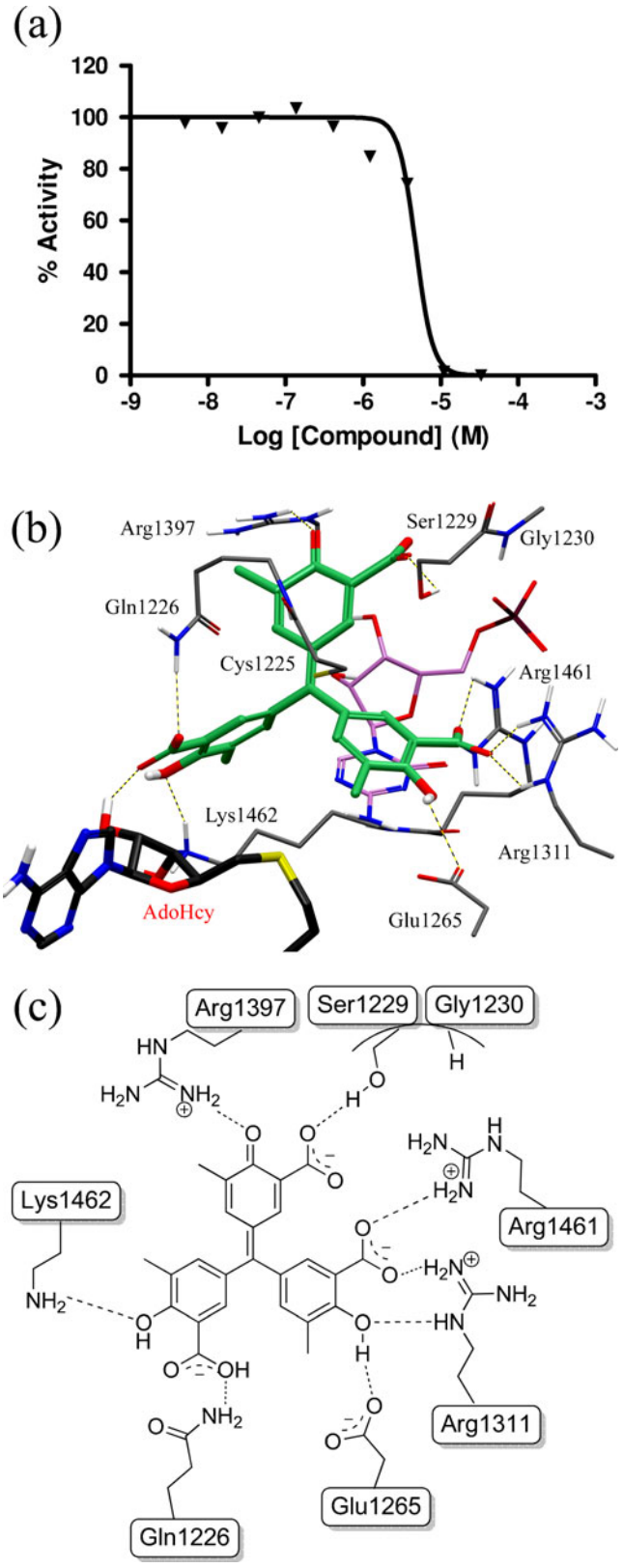
### *Molecular docking*

The starting conformation of NSC97317 was obtained by the conformational search in MacroModel and possible tautomers were explored using LigPrep (LigPrep, version 2.4, Schrödinger, LLC, New York, NY, 2010). The conformational analysis was carried out with Monte Carlo Multiple Minimum and Low-Mode conformational search method, employing the OPLS force field using GB/SA water solvation model. The lowest energy conformation of NSC97317 was docked into the catalytic site of the DNMT1 homology model using Glide extra precision (XP) (GLIDE, version 5.6, Schrödinger, LLC, New York, NY, 2010). We also performed flexible docking of other low-energy conformers of NSC97317 generated during the conformational analysis. Other known DNMT1 inhibitors were used as a reference (*vide infra*). To generate the grids for docking, the complexed DNA structure was removed from the homology model of human DNMT1, excluding the target cytosine. The shape and properties of the catalytic site were mapped onto the grids with default dimensions centered on the target cytosine. The best docking poses were selected and compared to the structure-based pharmacophore we recently reported for inhibitors of DNMT1 [31]. The same docking protocol was used recently to develop binding models for other DNMT inhibitors [31].

## Results and discussion

In order to experimentally test the hypothesis that NSC97317 is an inhibitor of DNMT1, the enzymatic inhibitory activity of this compound was measured using the HotSpot$^{SM}$ platform for methyltransferase assay described in Methods. NSC97317 showed a low micromolar inhibition of DNMT1 with $IC_{50}$ value of 4.79 μM. The corresponding dose-response plot is shown in Fig. 3a. To note, the enzymatic inhibitory activity of NSC97317 was better than the activity previously reported for 5,5-methylenedisalicylic acid ($IC_{50}$=92 μM) [25].

In order to rationalize the putative binding mode of NSC97317 with DNMT1 at the molecular level, we conducted molecular docking simulations. Despite the fact that a recent crystallographic structure for human DNMT1




◀ **Fig. 3** Enzymatic activity of NSC97317 and docking model with the catalytic domain of human DNMT1. (**a**) Inhibition curve obtained using the HotSpot$^{SM}$ platform for methyltransferase assays available at Reaction Biology Corporation (see text for details). (**b**) Comparison of the binding modes of NSC97317 (carbon atoms in green) with 5-azacytidine (carbon atoms in pink). Hydrogen bonds are depicted with dashes. Non-polar hydrogen atoms are omitted for clarity. (**c**) 2D representation showing the hydrogen bonding network with the binding pocket. Deprotonated form of NSC97317 is shown in the 3D and 2D binding models

has been recently published [34], the structure corresponds to the enzyme bound to *unmethylated DNA* and it is *not suitable to model small-molecule inhibitors of DNMT1* (*vide supra*). Therefore, we employed a homology model we recently developed and used to model several established DNMT1 inhibitors (cf. Methods). The geometry of the catalytic site, including the position of the target cytosine modeled in the complex, is in agreement with the catalytic mechanism of DNA methylation. Further details are provided elsewhere [31]. Of note, homology models of human DNMT1 have been used previously to identify novel inhibitors [22] and study the mechanism of action of small-molecule inhibitors [31].

Flexible docking studies of NSC97317 with DNMT1 were performed with Glide XP following the protocol described in Methods. The efficacy of the docking protocol was recently demonstrated in a recent published docking and structure-based pharmacophore study of several known DNMT1 inhibitors [31]. Figure 3 shows three- and two-dimensional (3D and 2D) representations of the optimized docked model of NSC97317 with DNMT1. Selected residues are shown. Of note, the deprotonated form of NSC97317 (expected in aqueous solution) is shown in the binding models. The best docking conformation of NSC97317 was generated by flexible docking. According to this binding model, NSC97317 makes hydrogen bond interactions with Ser1229 and Arg1397. The phosphate group of nucleoside analogues as well as the carboxylate anion of 5,5-methylenedisalicylic acid and RG108 also make hydrogen bonds with Ser1229 [31]. One of the salicylic acid moieties of NSC97317 forms a hydrogen bond network with key residues Glu1265, Arg1311, and Arg1461 (Fig. 3b). It is worth pointing out that Glu1265 and Arg1311 seem to play a crucial role in catalytic mechanism (Fig. 1). Furthermore, hydrogen bonds with Glu1265, Arg1311, and Arg1461 have also been predicted for other DNMT1 inhibitors and could be pharmacophoric. The second methylbenzoic acid moiety of NSC97317 makes hydrogen bonds with Gln1226, Lys1462, and 2′-OH of the co-factor (Fig. 3b). A similar hydrogen bond interaction with the co-factor has been reported for aurintricarboxylic acid, which is a very similar compound to NSC97317, with other methyltransferase. These hydrogen bond interactions are in agreement with the interactions found in the docking models of nucleoside analogues which are the most potent DNMT1 inhibitors. A comparison of the optimized binding mode of

NSC97317 with 5,5-methylenedisalicylic acid revealed that both compounds adopt a different docking position in the binding pocket (*vide infra*). Of note, the docking score of NSC97317 with DNMT1 (-9.45 kcal $mol^{-1}$) was better than the docking score of 5,5-methylenedisalicylic acid (-7.73 kcal $mol^{-1}$) and similar with that of nucleoside analogues (data not shown).

Figure 4 shows the comparison of the optimized binding mode of NSC97317 and 5,5′-methylenedisalicylic acid with a recently developed pharmacophore hypothesis for known DNMT inhibitors [31]. The pharmacophore model includes five-point features with one negative charge (N), one hydrogen bond acceptor (A), one aromatic ring (R) and two hydrogen bond donors (D). Considering a distance matching tolerance of 2.0 Å, a criteria commonly accepted in pharmacophore modeling [31, 35, 36], only one carboxylic acid group of 5,5′-methylenedisalicylic acid matches with the negative charged feature (Fig. 4b). In contrast, the binding position of NSC97317 matches (considering a distance matching tolerance of 2.0 Å) with the aromatic ring (R), hydrogen bond donor (D) (interaction with Glu1265) and hydrogen bond acceptor (A) (interaction

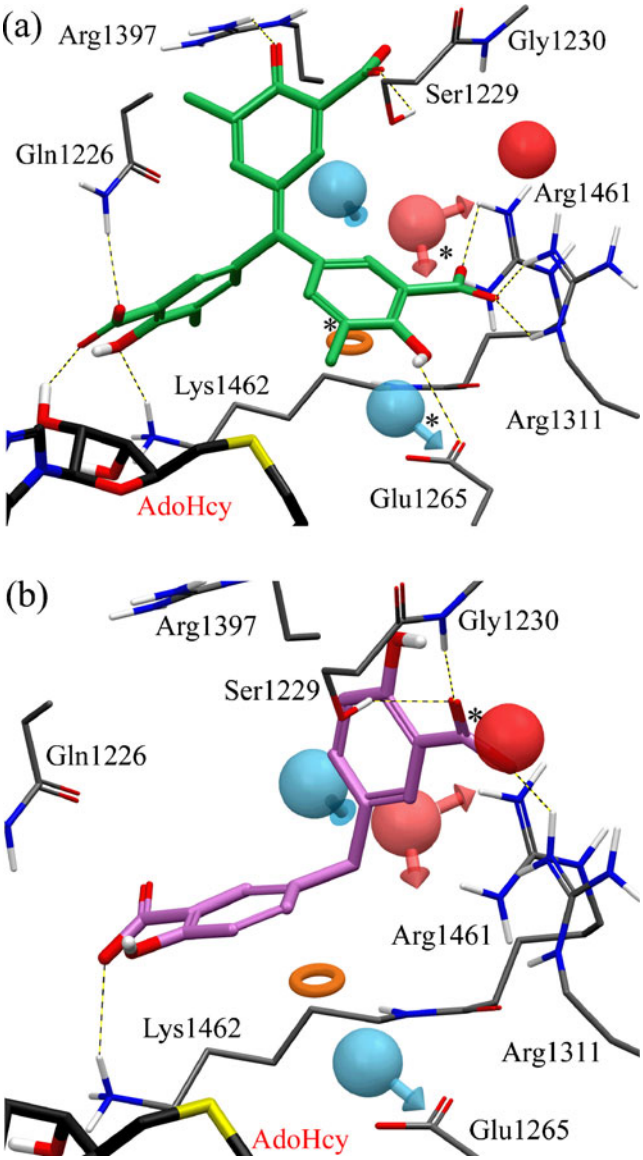



**Fig. 4** Comparison of the binding modes with pharmacophore hypothesis for (**a**) NSC97317 and (**b**) 5,5-methylenedisalicylic acid (deprotonated forms). Hydrogen bonds are depicted with dashes. Non-polar hydrogen atoms are omitted for clarity. Selected residues are displayed for reference. *Red sphere:* negative ionizable; *pink sphere with vectors:* hydrogen bond acceptor; *blue sphere with vector:* hydrogen bond donors; and *orange ring:* aromatic ring. Matching features, considering a distance matching tolerance of 2.0 Å, are marked with asterisks

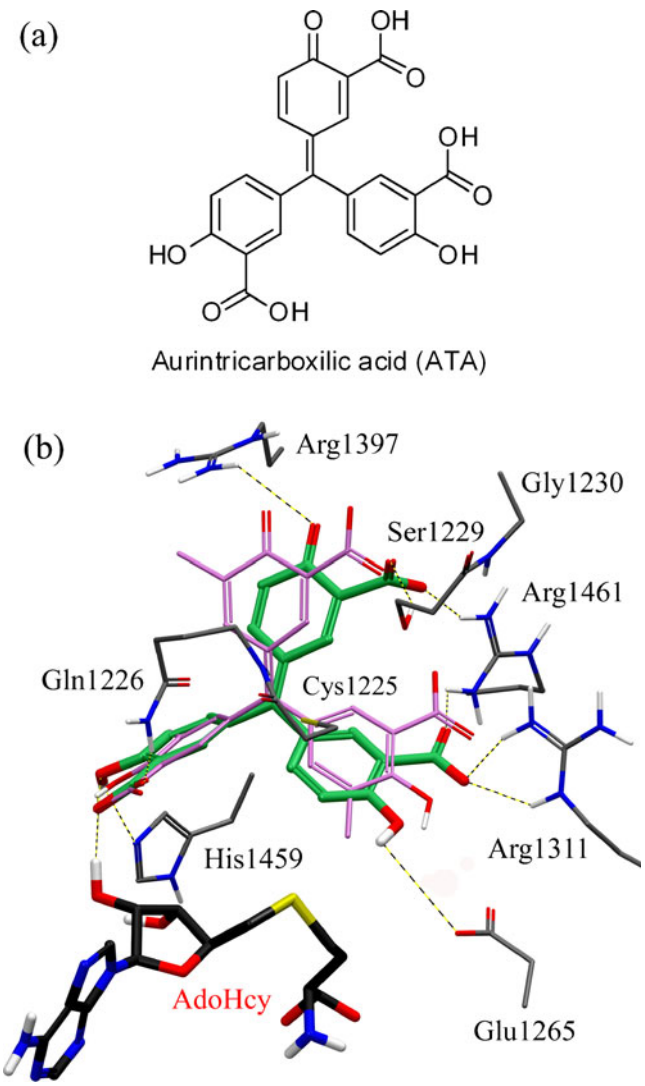



**Fig. 5** (**a**) Chemical structure of aurintricarboxilic acid (ATA). (**b**) Comparison of the binding model of ATA (carbon atoms in green) and NSC97317 (carbon atoms in pink) with the catalytic domain of human DNMT1. Deprotonated forms of ATA and NSC97317 are shown in the binding models. Selected residues in the binding pocket are shown. Hydrogen bonds are depicted with dashes. Non-polar hydrogen atoms are omitted for clarity

with Arg1311 and Arg1461) of the pharmacophore hypothesis (Fig. 4a). These results further supported the experimental observation that NSC97317 is a better DNMT1 inhibitor than 5,5′-methylenedisalicylic acid.

It seems likely that aurintricarboxylic acid will also inhibit DNMT1 [27]. Preliminary docking studies showed that aurintricarboxylic acid (Fig. 5a) has a very similar binding mode than NSC97317 making almost the same interactions with the catalytic site. Figure 5b shows the predicted binding mode of aurintricarboxylic acid with the catalytic domain of DNMT1. The binding model of the trimethyl analog is shown for comparison. Of note, the deprotonated forms of both ligands (expected in aqueous solution), are shown in the 3D binding model. In addition, the calculated binding score for aurintricarboxylic acid (-8.77 kcal $mol^{-1}$) is also better than the binding score for 5,5′-methylenedisalicylic acid (−7.62 kcal $mol^{-1}$). It remains to be determined the experimental enzymatic DNMT1 inhibition of aurintricarboxylic acid.

Collectively, these results support that NSC97317 impair the correct location of DNA in the catalytic site by steric hindrance of this site and through the interaction with residues involved in DNA-DNMT1 recognition. As such, NSC97317, *freely available* at the NCI/DTP, is a valuable biochemical tool to study DNMT1 inhibition in cancer and other diseases related to DNA methylation.

## Conclusions and perspectives

We report enzyme inhibition and molecular modeling studies of trimethylaurintricarboxylic acid (NSC97317) with the catalytic site of human DNMT1. NSC97317 is a low micromolar inhibitor of human DNMT1. Enzymatic and molecular modeling studies confirmed our hypothesis that this compound is a more potent inhibitor than 5,5-methylenedisalicylic acid, a structurally related compound. Molecular docking studies with the catalytic domain of DNMT1 indicate that NSC97317 makes hydrogen bond interactions with amino acid residues involved in the mechanism of DNA methylation. The polyanionic structure of NSC97317 and the electropositive binding region of the catalytic site of DNMT1, due to several arginine residues, also seem to play a key role in the binding recognition. The key interactions of the analogue of aurintricarboxylic acid with the binding pocket of DNMT1 can be exploited in the structure-based discovery and optimization of DNMT1 inhibitors. NSC97317 is a promising research tool to study enzymatic inhibition of DNMT1 in the research of cancer and other diseases related to DNA methylation. Follow up studies include the evaluation of NSC97317 as a demethylating agent in cell based assays. Additional work planned is to determine the potential biochemical activity of NSC97317, aurintricarboxylic acid, and other analogues toward different isoforms of DNMT including DNMT3A and DNMT3B.

**Acknowledgments** NSC97317 was kindly supplied by the Drug Synthesis and Chemistry Branch, Developmental Therapeutics Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute. This work was supported by the Menopause & Women's Health Research Center and the State of Florida, Executive Office of the Governor's Office of Tourism, Trade, and Economic Development. Authors are grateful to Kyle Kryak for proofreading the manuscript.

## References

1. Robertson KD (2001) DNA methylation, methyltransferases, and cancer. Oncogene 20:3139–3155
2. Yokochi T, Robertson KD (2002) Preferential methylation of unmethylated DNA by mammalian de novo DNA methyltransferase Dnmt3a. J Biol Chem 277:11735–11745
3. Goll MG, Bestor TH (2005) Eukaryotic cytosine methyltransferases. Annu Rev Biochem 74:481–514
4. Sippl W, Jung M (2009) DNA Methyltransferase inhibitors, vol. 42 Epigenetic Targets in Drug Discovery. Wiley-VCH, Weinheim
5. Stresemann C, Lyko F (2008) Modes of action of the DNA methyltransferase inhibitors azacytidine and decitabine. Int J Cancer 123:8–13
6. Kelly TK, De Carvalho DD, Jones PA (2010) Epigenetic modifications as therapeutic targets. Nat Biotechnol 28:1069–1078
7. Portela A, Esteller M (2010) Epigenetic modifications and human disease. Nat Biotechnol 28:1057–1068
8. Zawia NH, Lahiri DK, Cardozo-Pelaez F (2009) Epigenetics, oxidative stress, and Alzheimer disease. Free Radical Biol Med 46:1241–1249
9. Miller CA, Gavin CF, White JA, Parrish RR, Honasoge A, Yancey CR, Rivera IM, Rubio MD, Rumbaugh G, Sweatt JD (2010) Cortical DNA methylation maintains remote memory. Nat Neurosci 13:664–666
10. Narayan P, Dragunow M (2010) Pharmacology of epigenetics in brain disorders. Br J Pharmacol 159:285–303
11. Cheng XD, Blumenthal RM (2008) Mammalian DNA methyltransferases: A structural perspective. Structure 16:341–350
12. Lan J, Hua S, He XN, Zhang Y (2010) DNA methyltransferases and methyl-binding proteins of mammals. Acta Biochim Biophys Sin 42:243–252
13. Jurkowska RZ, Jurkowski TP, Jeltsch A (2011) Structure and function of mammalian DNA methyltransferases. ChemBioChem 12:206–222
14. Vilkaitis G, Merkiene E, Serva S, Weinhold E, Klimasauskas S (2001) The mechanism of DNA cytosine-5 methylation - Kinetic and mutational dissection of HhaI methyltransferase. J Biol Chem 276:20924–20934
15. Schermelleh L, Spada F, Easwaran HP, Zolghadr K, Margot JB, Cardoso MC, Leonhardt H (2005) Trapped in action: direct visualization of DNA methyltransferase activity in living cells. Nat Methods 2:751–756
16. Issa JPJ, Kantarjian HM, Kirkpatrick P (2005) Azacitidine. Nat Rev Drug Discovery 4:275–276
17. Schrump DS, Fischette MR, Nguyen DM, Zhao M, Li XM, Kunst TF, Hancox A, Hong JA, Chen GA, Pishchik V, Figg WD, Murgo

AJ, Steinberg SM (2006) Phase I study of decitabine-mediated gene expression in patients with cancers involving the lungs, esophagus, or pleura. Clin Cancer Res 12:5777–5785
18. Issa J-PJ, Kantarjian HM (2009) Targeting DNA methylation. Clinical Cancer Res 15:3938–3946
19. Lyko F, Brown R (2005) DNA methyltransferase inhibitors and the development of epigenetic cancer therapies. J Natl Cancer Inst 97:1498–1506
20. Yu N, Wang M (2008) Anticancer drug discovery targeting DNA hypermethylation. Curr Med Chem 15:1350–1375
21. Hauser AT, Jung M (2008) Targeting epigenetic mechanisms: Potential of natural products in cancer chemoprevention. Planta Med 74:1593–1601
22. Siedlecki P, Boy RG, Musch T, Brueckner B, Suhai S, Lyko F, Zielenkiewicz P (2006) Discovery of two novel, small-molecule inhibitors of DNA methylation. J Med Chem 49:678–683
23. Kuck D, Caulfield T, Lyko F, Medina-Franco JL (2010) Nanaomycin A selectively inhibits DNMT3B and reactivates silenced tumor suppressor genes in human cancer cells. Mol Cancer Ther 9:3015–3023
24. Medina-Franco JL, Caulfield T (2011) Advances in the computational development of DNA methyltransferase inhibitors. Drug Discovery Today 16:418–425
25. Kuck D, Singh N, Lyko F, Medina-Franco JL (2010) Novel and selective DNA methyltransferase inhibitors: Docking-based virtual screening and experimental evaluation. Bioorg Med Chem 18:822–829
26. Podvinec M, Lim SP, Schmidt T, Scarsi M, Wen D, Sonntag L-S, Sanschagrin P, Shenkin PS, Schwede T (2010) Novel inhibitors of dengue virus methyltransferase: Discovery by in vitro-driven virtual screening on a desktop computer grid. J Med Chem 53:1483–1495
27. Milani M, Mastrangelo E, Bollati M, Selisko B, Decroly E, Bouvet M, Canard B, Bolognesi M (2009) Flaviviral methyltransferase/RNA interaction: Structural basis for enzyme inhibition. Antiviral Res 83:28–34
28. Bouvet M, Debarnot C, Imbert I, Selisko B, Snijder EJ, Canard B, Decroly E (2010) In vitro reconstitution of SARS-coronavirus mRNA cap methylation. Plos Pathogens 6:13
29. Developmental Therapeutics Program. National Cancer Institute (NCI) http://dtp.cancer.gov
30. Reaction Biology Corporation. http://www.reactionbiology.com
31. Yoo J, Medina-Franco JL (2011) Homology modeling, docking and structure-based pharmacophore of inhibitors of DNA methyltransferase. J Comput Aided Mol Des 25:555–567
32. Yoo J, Medina-Franco JL (2011) Rundfeldt C (ed) Discovery and optimization of inhibitors of DNA methyltransferase as novel drugs for cancer therapy. Drug Development. InTech (in press)
33. The UniProt C The Universal Protein Resource (UniProt) (2010) Nucleic Acids Res 38(suppl 1):D142–D148
34. Song J, Rechkoblit O, Bestor TH, Patel DJ (2011) Structure of DNMT1-DNA complex reveals a role for autoinhibition in maintenance DNA methylation. Science 331:1036–1040
35. Loving K, Salam NK, Sherman W (2009) Energetic analysis of fragment docking and application to structure-based pharmacophore hypothesis generation. J Comput Aided Mol Des 23:541–554
36. Salam NK, Nuti R, Sherman W (2009) Novel method for generating structure-based pharmacophores using energetic analysis. J Chem Inf Model 49:2356–2368

J Mol Model (2012) 18:1591–1596
DOI 10.1007/s00894-011-1192-3

ORIGINAL PAPER

# Dynamics comparison of two myoglobins with a distinct heme active site

**Ying-Wu Lin · Yi-Mou Wu · Li-Fu Liao**

Received: 25 January 2011 / Accepted: 18 July 2011 / Published online: 30 July 2011


**Abstract** Sperm whale myoglobin (swMb) is a well-studied heme protein, both experimentally and theoretically. Comparatively, little attention has been paid to another member of Mb family, *Aplysia limacina* myoglobin (apMb). swMb and apMb have the same overall structure and perform the same biological function, i.e., $O_2$ carrier, while using a distinct heme active site. To provide insights into the structure-function relationship for these two Mbs, we herein made a comparison in terms of their dynamics properties using molecular dynamics simulation. We analyzed the overall structure and protein motions, as well as the intramolecular contacts, namely salt-bridges and hydrogen bonds, especially the interactions between the heme propionate groups and the polypeptide chain. The internal cavities in apMb were also compared to the well-known xenon and other cavities in swMb. Based on current simulations, we propose a unique "arginine gate" for apMb, which has a similar function to the histidine gate observed for swMb in previous studies. This study provides valuable information for understanding the homology of heme proteins, and also aids in rational design of structural and functional heme proteins by alternating the heme active site.

Y.-W. Lin (✉) · L.-F. Liao
School of Chemistry and Chemical Engineering,
University of South China,
Hengyang 421001, China
e-mail: linlinying@hotmail.com

Y.-W. Lin
State Key Laboratory of Coordination Chemistry,
Nanjing University,
Nanjing 210093, China

Y.-M. Wu
Institute of Pathogenic Biology, University of South China,
Hengyang 421001, China

**Keywords** Heme active site · Heme proteins · Homology · MD simulation · Myoglobin

## Introduction

Using the same heme group, heme proteins perform a large array of biological functions, including oxygen binding and transport, electron transport, catalysis, and signaling [1]. The various functions of heme proteins are owing to the diverse interactions between the polypeptide chain and the heme group. For example, an electron transporter such as cytochrome $b_5$ has an active site with a heme group coordinated by two axial histidines [2], and a catalytic cytochrome P450 has a five-coordinate heme group with an open site for substrate binding and oxidization [3]. As inspired by the distinct heme active sites, researchers can now achieve structural and functional conversions between different heme proteins by alternating their active sites [1–4].

A further interesting observation is that the same functional heme proteins may have a distinct heme active site. One such example is oxygen carriers, sperm whale myoglobin (swMb) [5] and *Aplysia limacina* myoglobin (apMb) [6]. Like most other myoglobins [7], swMb has a five-coordinate heme group with a distal histidine (H64) responsible for the stabilization of $O_2$ binding (Fig. 1a). Meanwhile, apMb lacks the conserved distal histidine. Instead, a valine residue (V63) is present in the distal pocket at the corresponding position and bound $O_2$ is stabilized by hydrogen bonding to the non-homologous aginine (R66) (Fig. 1b). During the last few decades, swMb has been well studied as a represent heme protein. In addition, it has been widely used as a protein model for rational design of artificial heme proteins with novel functions [1, 4, 8–11]. Comparatively, the structure-

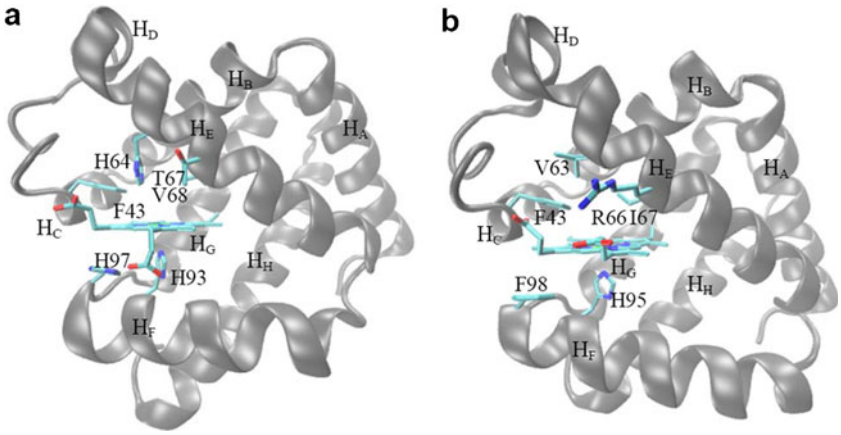



**Fig. 1** Heme environments in swMb (**a**) (PDB entry 1JP6) and apMb (**b**) (PDB entry 1MBA). The polypeptide chain with eight α-helices is labeled as $H_A$-$H_H$

activity relationship has not been fully understood for the unique apMb [6, 12–16].

As a complimentary approach, molecular dynamics (MD) simulations have become a powerful tool to provide atomic level and time-dependent information for protein structure and functions that otherwise is difficult to obtain from experiments [17–21]. To date, plenty of MD simulations have been performed for swMb to gain insights into its folding behaviors and functioning as an $O_2$ carrier [21–27]. While to the best of our knowledge, there is no report on MD simulation for apMb yet. However, it is interesting to make a dynamics comparison of apMb and swMb because they perform the same function using a distinct heme active site. We herein performed such a study, aimed at gaining deep insight into the structure-activity relationship for these two Mbs, and also for other heme proteins homogenous to one another.

## Methods

The initial structures of swMb and apMb for simulation were taken from the X-ray structures, PDB entries 1JP6 [5] and 1MBA [6], respectively. The psfgen plug-in of program VMD 1.8.7 (visual molecular dynamics) [28] was used to add hydrogen atoms and assign charges to the protein, which was set up according to pH 7. Note that according to previous experimental and theoretical results [5, 21–27], the ε-nitrogen of histidine was protonated for the heme distal His64 in swMb and other histidines in two Mbs, i.e., HSE in topology file, while His93 in swMb and His95 in apMb that coordinate directly to the heme iron were modeled with hydrogen at the δ-position (HSD). The protein was then solvated in a cubic box of TIP3 water, which extended 10 Å away from any given protein atom. A total of 17 (8 $Na^+$ and 9 $Cl^-$) and 15 (7 $Na^+$ and 8 $Cl^-$) counter ions were further added to swMb and apMb water system, respectively, to obtain the physiological ionic strength of 0.15 M by using the autoionize plug-in of VMD 1.8.7.

The resultant system was first minimized for 5000 ps with conjugate gradient method, and subsequently subjected to MD simulation with program NAMD 2.7 (nanoscale molecular dynamics) [29] at 300 K for 10 ns by employing the classical force filed CHARMM27 [30]. The Langevin dynamics was used for constant temperature control, with a value of the Langevin coupling coefficient setting to 5 ps. The simulations were performed with an integration time step of 2 fs. Atom-based cutoff of 12 Å with switching at 10 Å was used for non-bonded van der Waals interactions. The particle mesh Ewald (PME) method was applied to treat the electrostatic interactions. It is an efficient way of calculating long-range forces in the periodic system [29]. The PME grid size was set to be 64 Å in the direction along cell basis vector. The trajectory data was saved every 2500 steps (corresponding to 5 ps), generating 2,000 structures over 10 ns of simulation for analysis. Equilibrated swMb as well as the close and open forms of apMb were submitted to computed atlas of surface topography of proteins (CASTp) server (http://cast.engr.uic.edu) [31] to identify internal cavities with a default probe radius 1.4 Å. The pocket information was analyzed by program PyMOL 0.99rc6 [32]. Visualization and data analysis were done with VMD 1.8.7. Sequence alignment was performed using Clustal 2.0.12 program server (http://www.ebi.ac.uk/Tools/clustalw2/index.html) [33].

## Results and discussion

### Overall structures of swMb and apMb

As one member of Mb family, apMb has a typical Mb fold with eight α-helices ($H_A$-$H_H$), which is identical to swMb as shown in the X-ray structure (Fig. 1). However, amino acid sequence alignment shows that the sequence homology between apMb and swMb is only 21% (Fig. 2). These observations suggest that low homologous polypeptide chains may also fold into similar 3D structures. To better

```
swMb VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASED 60
apMb SLSAAEADLAGKSWAPVFANKNANGLDFLVALFEKFPDSANFFADFKG-KSVADIKASPK 59
      ** .* :*. : ** * *:  .:* *:*: **:..*:: : *  **  *: *::*** .

swMb LKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHP 120
apMb LRDVSSRIFTRLNEFVNNAANAGKMSAMLSQFAKEHVGFGVGSAQFENVRSMFPGFVASV 119
     *:. .  ::* *. :::: .:       *:*    :  : :   :* .   :.  .

swMb GDFGADAQGAMNKALELFRKDIAAKYKELGYQG 153
apMb AAPPAGADAAWIKLFGLIIDALKAAGA------ 146
     .   *.*:.* .* : *: . : *
```

**Fig. 2** Sequence alignment of swMb and apMb by program Clustal. "*", ":" and "." indicate high, middle, and low conserved residues, respectively

compare the overall structure, we performed structural alignment of these two Mbs after energy minimization and MD simulation instead of overlapping the initial X-ray structures directly. As exhibited in Fig. 3, the overall structure overlaps well with each other except for a short $H_D$ adjacent to a flexible loop region ($H_C$-$H_D$), and all eight α-helices have similar spatial orientations. In addition, the heme group has similar conformation in two proteins. Meanwhile, the heme propionate groups interact differently with the polypeptide chains (see following sections for details).

## Protein motions of swMb and apMb

Because no direct dynamics information can be gained from sequence or structural alignments, we monitored the $C_\alpha$ root mean square deviation (RMSD) for both Mbs from the starting structure during 10 ns MD simulations. As shown in Fig. 4a, the RMSD profiles are very similar to each other, with the deviation fluctuating around 1 Å after 1000 ps, which indicates that both Mbs have similar conformational stabilities of the overall structure. Furthermore, an average RMSD over time of each residue provides more detailed motions for both Mbs. As illustrated in Fig. 4b, it can be observed that eight α-helices in both proteins have small deviations between 0.5 and 1 Å, which agrees well with previous MD studies on swMb [22]. Note that there is no MD simulation of apMb available from literature for comparison yet.

In contrast, large deviations were observed for their loop regions, particularly for the loop of $H_G$-$H_H$, with RMSD around 3.5 Å for G121 and D127 in swMb and apMb, respectively. It is interesting to observe that more dramatic deviations occurred to swMb in loops of $H_A$-$H_B$ (residues 16–19) and $H_E$-$H_F$ (residues 75–81) compared to apMb. Sequence alignment shows that swMb has two charged residues (K16 and E18) in loop of $H_A$-$H_B$, where the corresponding residues are replaced by rigid P16 and

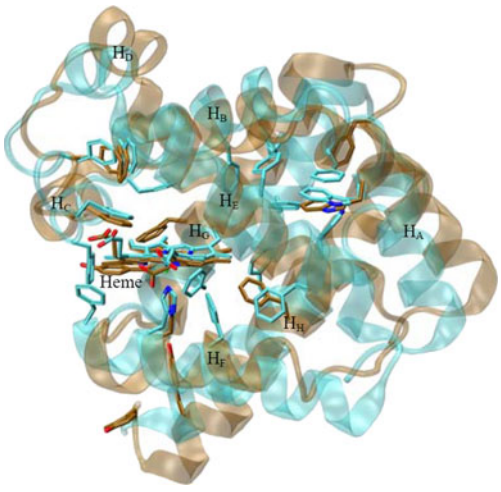



**Fig. 3** Structural alignment of swMb (in ochre) and apMb (in cyan) by program VMD after MD simulation from X-ray structure. Eight α-helices ($H_A$-$H_H$), the heme group and axial histidine, as well as the location of hydrophobic residues, Phe/Trp/Tyr, are highlighted

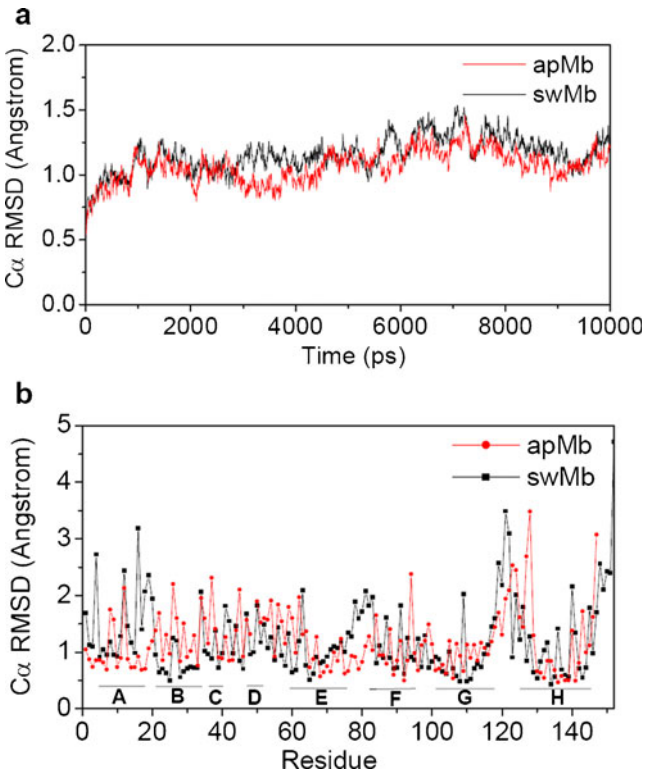



**Fig. 4** RMSD for swMb and apMb from the starting structure as a function of time (**a**) and the average RMSD over time of each residue (**b**). The eight α-helices ($H_A$-$H_H$) are indicated by short solid lines at the bottom

hydrophobic F18, respectively. At the same time, swMb has a unique tripeptide of lysine (K77-K78-K79) in loop of $H_E$-$H_F$, where K77 interacts with E18 in loop of $H_A$-$H_B$, K78 interacts with E85 in the N-terminus of $H_F$, and K79 interacts with E4 in the N-terminus of $H_A$, respectively. The tripeptide thus plays a key role in folding of $H_A$, $H_B$ and $H_F$ to a functional state of swMb. These unique interactions are not observed in apMb, since the corresponding peptide is N76-N77-A78 as shown by sequence alignment (Fig. 2). Both N76 and N77 form a part of the C-terminus of $H_E$ and thus show relative low RMSD during MD simulations.

## Intramolecular contacts of swMb and apMb

To further compare the dynamic behaviors of swMb and apMb, we analyzed their intramolecular contacts, namely salt-bridges and hydrogen bonds, with a cutoff set to be 3.2 Å during the MD simulation. With the help of VMD, 26 different salt-bridges were identified for swMb involving surface residues between Asp/Glu (O donors) and Arg/Lys (N donors). Meanwhile, only 11 were identified for apMb under the same conditions. For hydrogen bonds, analysis revealed that an average number of 81 were formed in swMb after 2000 ps in 10-ns simulation, which is slight larger than that in apMb with an average number of 72, as shown in Fig. 5a. These observations suggest that both salt-bridges and hydrogen bonds potentially play more roles in maintaining the structure and function of swMb in comparison to apMb. As a complement, more hydrophobic interactions were identified to exist in the scaffold of apMb. Sequence alignment shows that apMb has nine more Phe residues than that in swMb (15 vs. 6, Fig. 2). A detailed view of the Phe/Trp/Tyr locations in both proteins is exhibited in Fig. 3. These differences may also close link to protein properties such as internal cavities, as analyzed in the following section.

On the other hand, one might be interested in the interactions between the heme group and the polypeptide chain. An inspection of the heme group in swMb reveals that two propionate groups (6-P and 7-P) interact with R45 and H97, respectively, through hydrogen bonding interactions, as shown by the time dependent formation of hydrogen bond during MD simulation in Fig. 5b and a snapshot view in Fig. 6a. In apMb, however, the corresponding residues are replaced by D45 and F98, respectively (Fig. 2), and the same interactions do not exist in apMb. Instead, the heme 7-P group was found to be able to form a hydrogen bond with the side chain of R66 located in the middle of $H_E$ in apMb. Figure 5c shows the distance between R66(HH12) and heme 7-P(O2A) atoms as a function of simulation. It is interesting to find that the hydrogen bond was formed and broken frequently as the distance fluctuates between 1.7 Å and 6.4 Å in a relative

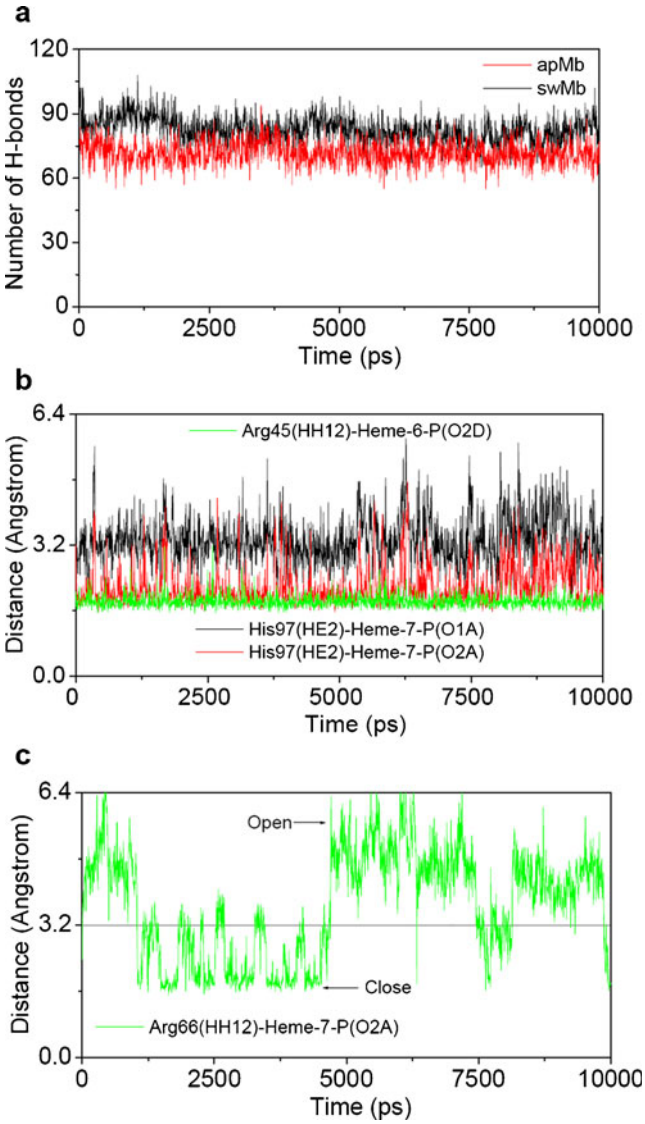



**Fig. 5** Time-dependent of hydrogen bonds number in swMb and apMb (**a**), hydrogen bonds distance of heme 6-P-R45 and heme 7-P-H97 in swMb (**b**), and heme 7-P-R66 in apMb (**c**) in MD simulation. The close and open conformations of R66 are indicated by arrows

short time (10-ns). As indicated by arrows in Fig. 5c, the hydrogen bonding form and the separated form are named as close form and open form, respectively. A detailed view of these two distinct forms is shown in Fig. 6b by spatial alignment. These observations indicate that dramatic conformational changes occurred to R66 and heme 7-P in apMb during MD simulation, which in turn suggest that R66 plays a vital role in manipulating the protein structure and function by conformational changes.

Although hydrogen bonding interaction was also observed for R45 and heme 6-P in swMb, no dramatic conformational change occurred to R45 or 6-P, as no break of the hydrogen bond was observed during the entire simulation process (Fig. 5b). As a matter of fact, the distal

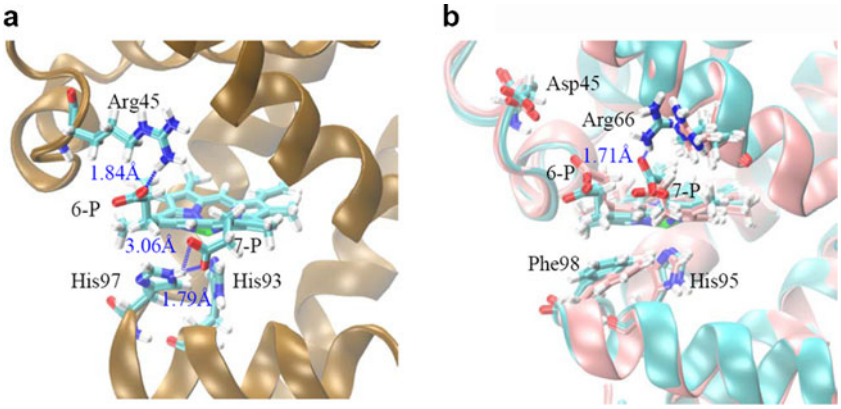



**Fig. 6** A detailed view of hydrogen bonds, indicated by dotted line and corresponding distance, of heme 6-P-R45 and heme 7-P-H97 in swMb (**a**) and heme 7-P-R66 in apMb (**b**). The close (cyan) and open (pink) conformations of R66 in apMb are overlapped in (**b**) for comparison

His64 in swMb plays a vital role in protein functioning as an $O_2$ carrier. In the absence of a distal histidine, the non-homologous R66 in apMb likely plays a corresponding role by forming an "arginine gate" instead of the so-called histidine gate in swMb, allowing access of small ligands to the heme iron. The present simulations support the interpretation presented in 1990 by Bolognesi et al. [12], and extended later by crystallography and kinetics studies [13].

It should be noted that in the present study the distal His64 in swMb was found to be in a close conformation during the entire 10-ns MD simulation process. In a previous study, Bossa et al. [22] showed that the histidine gate opens in deoxy swMb after 60-ns by extended MD simulations (80-ns). Note that the author used GROMOS96 force field and dummy hydrogen atoms, which is very different from present simulations. On the other hand, Smith et al. [34] showed that the presence of substrates such as straight-chain alkyl isocyanides can open the distal histidine gate in the crystal structure of swMb. In present simulations, although with a relative short time (10-ns) and no substrates presented, a closure-opening transition of conformation was observed for the proposed "arginine gate" in apMb. This finding may provide valuable insight into the experimental observation that apMb has a high $O_2$ dissociation rate constant ($K_{off}$=70 $s^{-1}$) compared to that for swMb ($K_{off}$=15 $s^{-1}$) [14, 35]. The ligand stabilization mechanism adopted by apMb through R66 was thought to be not as efficient as that based on H64 in swMb [13].

## Internal cavities of swMb and apMb

The presence and function of internal cavities in Mb have attracted much attention recently, in particular focused on swMb [22–27, 36–39]. To provide cavity information for apMb, we analyzed the cavities in both close and open forms of apMb and compared to that in swMb with energy minimization and MD simulation. In the crystal structure of swMb, Tilton et al. [36] observed four cavities occupied by xenon (Xe) atoms (Xe1-4), which was confirmed recently by Savino et al. [38] (PDB entry 2W6W) (Fig. 7a). As shown in Fig. 7b, these four cavities can be well-identified in present simulations of swMb. In addition, both a distal pocket (DP) and a side pocket (SP) can be identified, correlating with the theoretical CO migration pathways inside swMb [27].

For apMb, although it shows poor (21%) sequence conservation with respect to swMb, the Xe cavities as well as DP and SP can all be identified in both close and open forms, as shown in Fig. 7c and d, respectively. Meanwhile, with a distinct active site, apMb exhibits a larger DP (84 $Å^3$) compared to that in swMb (60 $Å^3$), even in a close form. With the opening of R66 gate, apMb exhibits a DP as

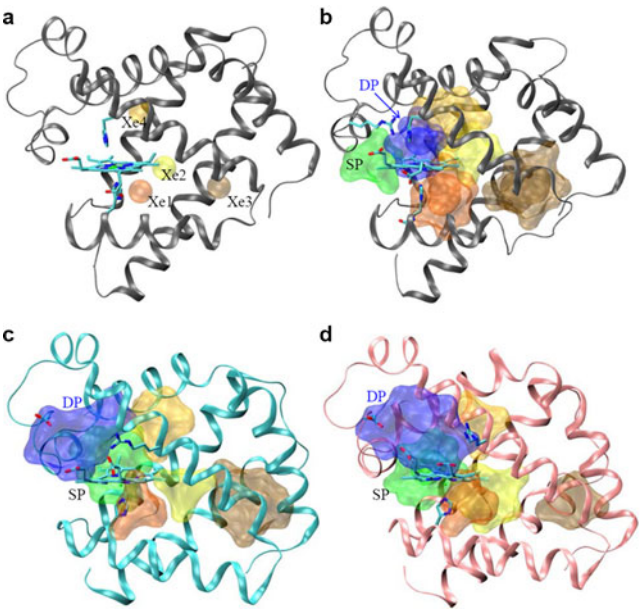



**Fig. 7** A view of cavities inside Mbs: The locations of four Xe cavities in the crystal structure of swMb (PDB entry 2W6W) are represented as colored spheres (**a**); The distal pocket (DP, blue) and side pocket (SP, green), as well as the cavities corresponding to the Xe cavities are represented as colored surface for swMb (**b**), and for apMb in the close (**c**) and open (**d**) forms

large as 154 $Å^3$. Moreover, the position of SP is found to move backward in apMb, likely due to the replacement of K42 with F42 and the absence of Y103 as in swMb. On the other hand, although four Xe cavities locate at similar positions in both Mbs, the residues lining the cavity are not conserved. More hydrophobic residues such as F28, F112 and F134 locate among the cavities in apMb, resulting in a lower connectivity of these cavities. These differences may be related to the ligand binding kinetics and migration pathways inside the protein [22–27, 37–39].

## Conclusions

In summary, the dynamics properties of two myoglobins, swMb and apMb, were investigated by MD simulations and compared to each other. Although with a low homology, these two proteins have similar folding behaviors as well as internal cavities in general. Meanwhile, swMb was found to have a dominant stabilization contributed by salt-bridges and hydrogen bonds, whereas apMb has more hydrophobic interactions. Additionally, with a distinct heme active site, different interactions were observed for the heme propionate groups and the polypeptide chains of swMb and apMb. More interestingly, dramatic conformational changes occurred to R66 in apMb in a relative short simulation time (within 10-ns), indicating that an "arginine gate" may play a vital role in performing the function of apMb as an $O_2$ carrier. apMb was also found to have a larger DP and four similar Xe cavities with a lower connectivity with respect to that in swMb. The current simulations provide valuable insights into the structure, function and homology relationship for heme proteins, especially for those that perform the same function with a similar protein scaffold and a distinct heme active site. The similarities and differences revealed in this study are also useful for rational design of structural and functional heme proteins such as advanced biocatalysts and biosensors.

**Acknowledgments** We thank Dr. Tianlei Ying at National Institutes of Health (NIH), USA, for helpful discussions. NAMD and VMD were developed by the Theoretical Biophysics Group in the Beckman Institute for Advanced Science and Technology at the University of Illinois at Urbana-Champaign, USA. This work is supported by Hunan Provincial Natural Science Foundation of China, No. 11JJ4017.

## References

1. Lu Y, Berry SM, Pfister TD (2003) Chem Rev 101:3047–3080
2. Sligar SG, Egeberg KD, Sage JT, Morikis D, Champion PM (1987) J Am Chem Soc 109:7896–7897
3. Sun L, Wang Z-H, Ni F-Y, Tan X-S, Huang Z-X (2010) Protein J 29:32–43
4. Lu Y, Yeung N, Sieracki N, Marshall NM (2009) Nature 460:855–862
5. Urayama P, Phillips GN Jr, Gruner SM (2002) Structure 10:51–60
6. Bolognesi M, Onesti S, Gatti G, Coda A, Ascenzi P, Brunori M (1989) J Mol Biol 205:529–544
7. Hirota S, Azuma K, Fukuba M, Kuroiwa S, Funasaki N (2005) Biochemistry 44:10322–10327
8. Abe S, Ueno T, Watanabe Y (2009) Top Organomet Chem 25:25–43
9. Yeung N, Lin Y-W, Gao Y-G, Zhao X, Russell BS, Lei L, Miner KD, Robinson H, Lu Y (2009) Nature 462:1079–1082
10. Lin Y-W, Yeung N, Gao YG, Miner KD, Tian S, Robinson H, Lu Y (2010) Proc Natl Acad Sci USA 107:8581–8586
11. Lin Y-W, Yeung N, Gao YG, Miner KD, Lei L, Robinson H, Lu Y (2010) J Am Chem Soc 132:9970–9972
12. Bolognesi M, Coda A, Frigerio F, Gatti G, Ascenzi P, Brunori M (1990) J Mol Biol 213:621–625
13. Conti E, Moser C, Rizzi M, Mattevi A, Lionetti C, Coda A, Ascenzi P, Brunori M, Bolognesi M (1993) J Mol Biol 233:498–508
14. Federici L, Savino C, Musto R, Travaglini-Allocatelli C, Cutruzzolà F, Brunori M (2000) Biochem Biophy Res Commun 269:58–63
15. Musto R, Bigotti MG, Travaglini-Allocatelli C, Brunori M, Cutruzzolà F (2004) Biochemistry 43:230–236
16. Svistunenko DA, Reeder BJ, Wankasi MM, Silaghi-Dumitrescu RL, Cooper CE, Rinaldo S, Cutruzzolà F, Wilson MT (2007) Dalton Trans 8:840–850
17. Karplus M, McCammon JA (2002) Nat Struct Biol 9:646–652
18. Daggett V (2006) Chem Rev 106:1898–1916
19. Sotomayor M, Schulten K (2007) Science 316:1144–1148
20. Kandt C, Monticelli L (2010) Methods Mol Biol 654:423–440
21. Lin Y-W (2011) Proteins 79:679–684
22. Bossa C, Amadei A, Daidone I, Anselmi M, Vallone B, Brunori M, Di Nola A (2005) Biophys J 89:465–474
23. Cohen J, Arkhipov A, Braun R, Schulten K (2006) Biophys J 91:1844–1857
24. Anselmi M, Di Nola A, Amadei A (2008) Biophys J 94:4277–4281
25. Ruscio JZ, Kumar D, Shukla M, Prisant MG, Murali TM, Onufriev AV (2008) Proc Natl Acad Sci USA 105:9204–9209
26. Ceccarelli M, Anedda R, Casu M, Ruggerone R (2008) Proteins 71:1231–1236
27. Maragliano L, Cottone G, Ciccotti G, Vanden-Eijnden E (2010) J Am Chem Soc 231:1010–1017
28. Humphrey W, Dalke A, Schulten K (1996) J Mol Graph 14:33–38
29. Kalé L, Skeel R, Bhandarkar M, Brunner R, Gursoy A, Krawetz N, Phillips J, Shinozaki A, Varadarajan K, Schulten K (1999) J Comput Phys 151:283–312
30. MacKerell AD Jr, Bashford D, Bellott M Jr, Dunbrack RL, Evanseck J, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher IWE, Roux B, Schlenkrich M, Smith J, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M (1998) J Phys Chem B 102:3586–3616
31. Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, Liang J (2006) J Nucleic Acids Res 34:W116–W118
32. DeLano WL, The PyMOL Molecular Graphics System (2002) DeLano Scientific, San Carlos, CA, USA. http://www.pymol.org
33. Thompson JD, Higgins DG, Gibson TJ (1994) Nucleic Acids Res 22:4673–4680
34. Smith RD, Blouin GC, Johnson KA, Phillips GN, Olson JS (2010) Biochemistry 49:4977–4986
35. Cutruzzolà F, Travaglini Allocatelli C, Brancaccio A, Brunori M (1996) Biochem J 314:83–90
36. Tilton RF Jr, Kuntz ID Jr, Petsko GA (1984) Biochemistry 23:2849–2857
37. Teeter MM (2004) Protein Sci 13:313–318
38. Savino C, Miele AE, Draghi F, Johnson KA, Sciara G, Brunori M, Vallone B (2009) Biopolymers 91:1097–1107
39. Tomita A, Kreutzer U, Adachi S, Koshihara S, Jue T (2010) J Exp Biol 213:2748–2754

J Mol Model (2012) 18:1597–1610
DOI 10.1007/s00894-011-1173-6

ORIGINAL PAPER

# A novel, customizable and optimizable parameter method using spherical harmonics for molecular shape similarity comparisons

**Chaoqian Cai · Jiayu Gong · Xiaofeng Liu · Hualiang Jiang · Daqi Gao · Honglin Li**

Received: 30 April 2011 / Accepted: 30 June 2011 / Published online: 30 July 2011


**Abstract** A novel molecular shape similarity comparison method, namely SHeMS, derived from spherical harmonic (SH) expansion, is presented in this study. Through weight optimization using genetic algorithms for a customized reference set, the optimal combination of weights for the translationally and rotationally invariant (TRI) SH shape descriptor, which can specifically and effectively distinguish overall and detailed shape features according to the molecular surface, is obtained for each molecule. This method features two key aspects: firstly, the SH expansion coefficients from different bands are weighted to calculate similarity, leading to a distinct contribution of overall and detailed features to the final score, and thus can be better tailored for each specific system under consideration. Secondly, the reference set for optimization can be totally configured by the user, which produces great flexibility, allowing system-specific and customized comparisons. The directory of useful decoys (DUD) database was adopted to validate and test our method, and principal component analysis (PCA) reveals that SH descriptors for shape comparison preserve sufficient information to separate actives from decoys. The results of virtual screening indicate that the proposed method based on optimal SH descriptor weight combinations represents a great improvement in performance over original SH (OSH) and ultra-fast shape recognition (USR) methods, and is comparable to many other popular methods. Through combining efficient shape similarity comparison with SH expansion method, and other aspects such as chemical and pharmacophore features, SHeMS can play a significant role in this field and can be applied practically to virtual screening by means of similarity comparison with 3D shapes of known active compounds or the binding pockets of target proteins.

**Keywords** Molecular shape · Similarity comparison · Spherical harmonic · Genetic optimization · Virtual screening

## Introduction

Virtual screening (VS) [1]—the rapid in silico assessment of large libraries of chemical structures in order to identify those most likely to bind to a drug target [2]—can be categorized into structure-based and ligand-based methods according to the availability of target protein structures. Although data on the three dimensional (3D) crystallographic structure of proteins increases every year due to the continuing development of X-ray and NMR technology, many structures still remain unknown, especially those of the transmembrane proteins that constitute a high proportion of important targets for drug discovery programs [3]. Structure-based methods [5], such as docking methods [6, 7], require accurate 3D information. Ligand-based methods [4] are frequently applied when precise information about the 3D structure of the

**Electronic supplementary material** The online version of this article (doi:10.1007/s00894-011-1173-6) contains supplementary material, which is available to authorized users.

C. Cai · J. Gong · D. Gao (✉)
School of Information Science and Engineering,
East China University of Science and Technology,
Shanghai 200237, China
e-mail: gaodaqi@ecust.edu.cn

X. Liu · H. Jiang · H. Li (✉)
State Key Laboratory of Bioreactor Engineering,
Shanghai Key Laboratory of Chemical Biology,
School of Pharmacy,
East China University of Science and Technology,
Shanghai 200237, China
e-mail: hlli@ecust.edu.cn

biological target of interest is lacking. Nevertheless, in numerous cases reported in the literature, the performance of ligand-based methods can be comparable to, or even exceed, that of docking methods [8–11]. However, current docking scoring approaches are still in need of further optimization in order to correctly predict binding affinities [12, 13].

Surface complementarity (generally a small molecule fitting into the cavity of a protein target) is pertinent to molecular recognition [14] and usually a prerequisite for protein–ligand binding. Molecular shape has long been a hot topic, and has been represented and compared by various approaches and plays a significant role in protein–ligand and protein–protein recognition and interaction. Currently, the approach most commonly used for molecular similarity comparison is molecular fingerprinting, using techniques such as Daylight [15], UNITY [16] and MACCS [17] fingerprints, which employ a binary bit string to represent the properties and topologies of a two-dimensional (2D) structure without considering 3D structure information. Recently, a field-based approach was applied to VS, using a set of field points to represent a molecule, with the VS process being achieved by time-consuming field alignment and scoring [18]. Likewise, some approaches [19–22] based on molecular electrostatic potential (MEP) fields, such as the shape and electrostatic potential (ShaEP) algorithm [23], are expressed as a rectangular or spherical grid around the molecule, involving tens of thousands of grid points for a typical representation, also come at the expense of high computation costs. From purely the shape perspective, replacing the traditional fused hard sphere model, the Gaussian function was devised to depict molecules, making the molecular volume overlap calculation highly efficient [24]. One such example is the rapid overlay of chemical structures (ROCS) algorithm [25], which uses Gaussian overlap to direct molecule alignment. The ultra-fast shape recognition (USR) algorithm [26–28] is another approach concentrating only on molecular shape comparisons. USR first identifies four locations for each molecule: the molecular centroid (ctd), the closest atom to ctd (cst), the farthest atom from ctd (fct) and the farthest atom from fct (ftf). For each of these locations, three moments are calculated to form a shape descriptor: the average atomic distance to this location, the square root of the atomic distances to this location, and the cube root of the skewness of the atomic distances to this location.

In mathematics, spherical harmonics (SH) are a special set of orthogonal spherical functions that can easily represent the shape of a closed curved surface (Fig. 1), e.g., a molecular surface. SH expansion theory has been applied successfully in the domain of drug discovery, including high throughput VS [29], protein–ligand recognition [30], binding pocket modeling [31], molecular fragment similarity [32], depiction of electron density [33], and so on. In nearly all extant SH based methods, an abstract molecular surface is generated and sampled, then SH expansion coefficients, which are transformed into various forms of descriptors, are acquired, and then used directly unchanged in a subsequent similarity comparison process. This scheme is sometimes unreasonable as will be explained later in the section on "Shape comparison and weighted similarity".

This study presents a novel SH-expansion-based molecular similarity method (SHeMS) that focuses on expansion coefficient optimization, and greatly enhances performance over the original spherical harmonics (OSH) method. The proposed method features two key aspects (Fig. 2). Firstly, unlike traditional SH methods using SH expansion coefficients directly to calculate molecular shape similarity, a weighted coefficient is used instead, thus giving each coefficient a distinct importance in contributing to the final similarity score. Choosing the appropriate weights value depends on the specific query conformer and a set of reference active and negative conformers, thus the resultant similarity score will be more suited for the actual system under consideration if the active and negative reference set is properly selected. Secondly, the weight choosing process can be completely customized. Provided with a user-customized reference set containing actives and negatives with respect to a certain query molecule, a genetic algorithm is

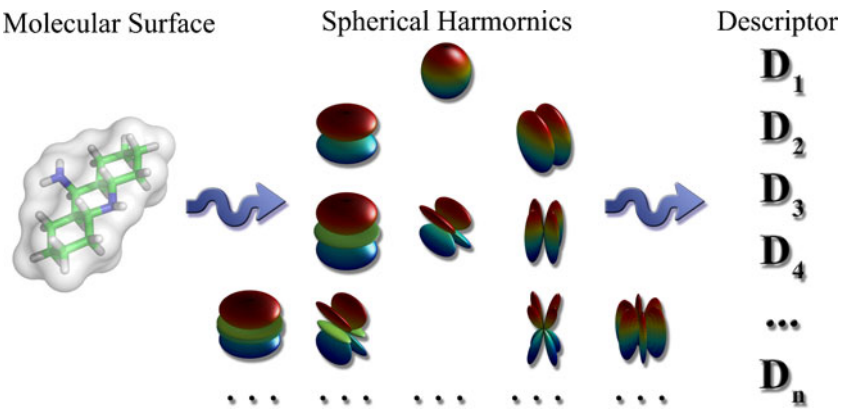



**Fig. 1** Schematic diagram of the spherical harmonic (SH) descriptor calculation. The molecular surface is projected to groups of SH producing a series of projection coefficients, which are used to calculate SH descriptors

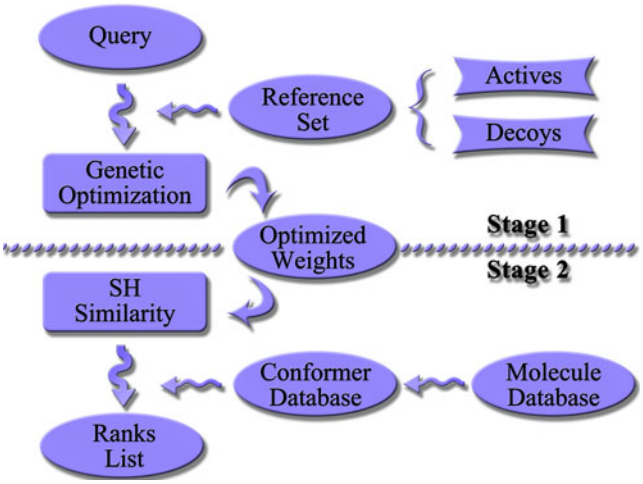



**Fig. 2** Flow chart of the spherical harmonic (SH)-expansion-based molecular similarity method (SHeMS). The SHeMS process can be divided roughly into two stages: Stage 1 is concerned with similarity weights optimization, which applies a genetic search to a user-defined reference set to obtain a group of optimal weights for the given query. Stage 2 is the virtual screening (VS) process, which uses SH descriptors combined with the optimal weights from stage 1 to evaluate shape similarity. The final output is a ranking list ordered by weighted SH similarity

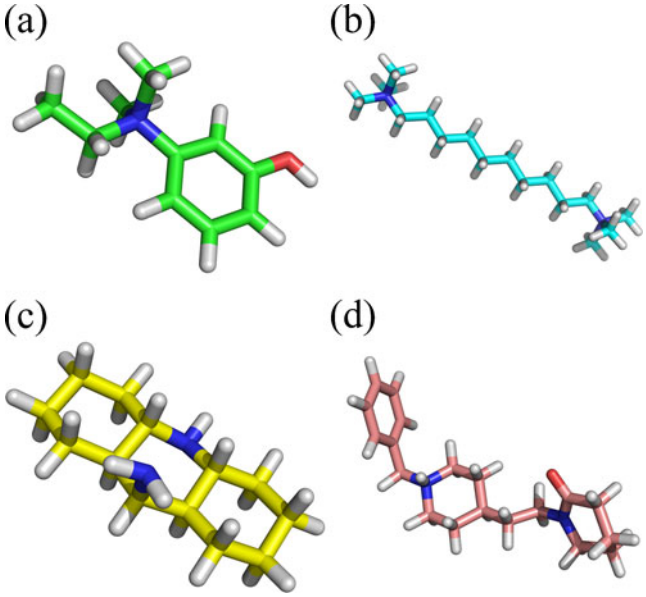



**Fig. 3** Sample actives from the acetylcholinesterase (ACHE) sub dataset. Four active molecules with representative structures from the ACHE sub dataset: **a** ZINC00001341; **b** ZINC01532339; **c** ZINC03795925; **d** ZINC03814250

applied to search for the best weights combination that well separates the reference actives and negatives, and the resulting weights are applied to the query when the real screening process takes place. This optimized weights combination can reasonably reflect the pattern shared with query molecule and given reference actives to help retrieve similar actives when screening. A retrospective VS experiment on a subset of the directory of useful decoys (DUD) data set [34] shows that the SHeMS method proposed in this study is capable of identifying a substantial number of active compounds in a database of drug-like molecules.

## Materials and methods

### Data preparation

To evaluate the ability of SHeMS to discriminate between actives and decoys, the DUD database (DUD LIB VS 1.0 [34–37]) was chosen as the source of molecules. As the raw DUD database is not suitable for shape-based screening due to the significant variation in the shape of actives [Fig. 3 shows four representative active molecules from the acetylcholinesterase (ACHE) dataset with significant discrepancy in shape], it is unrealistic to retrieve all these molecules based on some common pattern in shape. To make shape-based screening more reasonable, a cluster procedure, accomplished using the Tanimoto Coefficient of the functional-class fingerprint_4 (FCFP_4) molecular fingerprint as cluster criteria, was pre-applied to the actives set of each DUD sub dataset. FCFP_4 is a variant of the extended connectivity fingerprint (ECFP) [38], which belongs to the category of topological fingerprints implemented in Pipeline Pilot [39]. For each dataset, only one representative cluster (containing the largest number of actives) was picked as a new actives set from that dataset to participate in the evaluation, and all actives coming from other clusters were excluded. Note that some DUD datasets do not contain a sufficient number of actives (less than 20) to perform supervised weight searching and these datasets were not considered further. Finally, 15 sub datasets with adequate actives (see electronic supplementary material) from DUD were carefully selected for further testing, as shown in Table 1. A multi-conformer collection was then generated using the "3D Conformation" component of Pipeline Pilot for each selected sub dataset, with a maximum of 100 conformers per molecule. All data preparation was implemented in Pipeline Pilot [39].

### SH projection and shape descriptor

The SH theory, (for a detailed SH expansion, see Ref [40]), proves that any spherical function $f(\theta, \varphi)$ can be represented by its harmonics:

$$f(\theta,\varphi)=\sum_{l=0}^{\infty}\sum_{m=-l}^{l}c_{lm}y_{lm}(\theta,\varphi) \quad (1)$$

**Table 1** Selected directory of useful decoys (DUD) sub datasets from the DUD database [34]

| Data set | Actives total | Decoys total | Actives train | Decoys train |
|---|---|---|---|---|
| ACHE | 88 | 3,859 | 29–30 | 120 |
| AR | 32 | 2,848 | 10–12 | 120 |
| COX2 | 50 | 12,606 | 16–18 | 120 |
| DHFR | 53 | 8,350 | 17–19 | 120 |
| EGFR | 124 | 15,560 | 41–42 | 120 |
| GPB | 36 | 2,135 | 12 | 120 |
| INHA | 39 | 2,707 | 13 | 120 |
| P38 | 65 | 6,779 | 21–23 | 120 |
| PARP | 28 | 1,350 | 9–10 | 120 |
| PDGFRB | 37 | 5,603 | 12–13 | 120 |
| PNP | 24 | 1,036 | 8 | 120 |
| PR | 17 | 920 | 5–7 | 120 |
| SAHH | 23 | 1,346 | 7–9 | 120 |
| SRC | 40 | 5,679 | 13–14 | 120 |
| VEGFR2 | 27 | 2,712 | 9 | 120 |

Where, $y_{\mathrm{lm}}(\theta, \varphi)$ is a SH function located at band $l$ and component $m$, $\theta$ is inclination angle, $\varphi$ is azimuth angle, and $c_{\mathrm{lm}}$ is a coefficient constant. Note that Eq. 1 is the real form of spherical harmonics and will be referred to as spherical harmonics throughout this paper.

The corresponding expansion coefficient, $c_{\mathrm{lm}}$, is calculated by projecting a spherical function into a specified SH, as follows:

$$c_{lm} = \int_s f(s) y_{lm}(s) ds \tag{2}$$

Where, $f(s)$ is molecular surface shape function, $s$ is a distinct point on the surface. Due to the fact that $f(s)$ is a non-specific function, it is difficult to perform numerical integration directly; Monte Carlo integration can then be applied to the SH projection as follows:

$$c_{lm} \approx \frac{4\pi}{N} \sum_{i=1}^{N} f(x_i) y_{lm}(x_i) \tag{3}$$

where $N$ is the Monte Carlo samples count and $x_{\mathrm{i}}$ is the sample point. So the surface integral calculation is converted to summation over a series of sample points on the surface. The solvent-accessible surface (SAS) [41] can be obtained by a ray-atom intersection procedure [40], which is not the emphasis of this study and will not be elaborated here.

The SH shape descriptor in our study is generated from the above-mentioned expansion coefficient, $c_{\mathrm{lm}}$ . According to SH theory, the squared sum $d_{\mathrm{l}}$ of all expansion coefficients at a certain band remains unchanged under arbitrary rotation, and thus is a good candidate for generating rotation invariant shape descriptor, as follows:

$$d_l = \sum_{m=-l}^{l} c_{lm}^2 \tag{4}$$

The SH shape descriptor $D$ can then be written as a vector:

$$D = \lfloor d_0, d_1, d_2, \ldots, d_{L_{\max}} \rfloor \tag{5}$$

Where, $L_{\max}$ is the maximum band index of the SH projection. Each dimension of the descriptor corresponds to a certain band of SH. Generally speaking, the lower band coefficients depict the overall shape of the target molecular surface (e.g., the zero band coefficient gives the overall size of the surface) while, on the contrary, higher band coefficients are related to detailed features (they continue to impose refinement on the current course surface). Choosing an $L_{\max}$ value is a tradeoff between representation accuracy and computation cost: too small an $L_{\max}$ value will not represent the surface precisely; whereas too large an $L_{\max}$ value will improve accuracy a little but will also increase computation time dramatically. Based on a previous study [31], $L_{\max}$=6 is usually sufficient to represent any small molecule.

## Shape comparison and weighted similarity

Shape comparison is achieved by measuring the similarity between the corresponding SH shape descriptors. Given two shape descriptors $A = [a_0, a_1, a_2, \ldots, a_{L_{\max}}]$ and$B = [b_0, b_1, b_2, \ldots, b_{L_{\max}}]$, due to the rotation invariant property, their SH similarity score can be calculated directly without an alignment process as follows:

$$\begin{aligned} &S(A, B) = w_0 s(a_0, b_0) + w_1 s(a_1, b_1) + \ldots + w_{L_{\max}} s(a_{L_{\max}}, b_{L_{\max}}) \\ &w_0 + w_1 + \ldots + w_{L_{\max}} = 1 \\ &w_1, w_2, \ldots, w_{L_{\max}} \geq 0 \end{aligned} \tag{6}$$

where $w_{\mathrm{i}}$ is the corresponding weight constant measuring the contribution, which is made up of the current dimension (band) of the final similarity score, and $s(a_{\mathrm{i}}, b_{\mathrm{i}})$ is a bivariate function scaled in [0, 1] measuring the discrepancy (value 0 for most dissimilar, and value 1 for identical) of each dimension of shape descriptor, defined as follows:

$$s(a_i, b_i) = \frac{\min\{a_i, b_i\}}{\max\{a_i, b_i\}}. \tag{7}$$

It can clearly be inferred from the definition of the SH similarity score that all SH similarity scores are within the

range [0, 1]—0 for most dissimilar and 1 for identical—and have symmetrical properties, i.e., $S(A, B)=S(B, A)$. This differs from the traditional SH method, which treats all expansion coefficients equally. SHeMS gives each band different importance in terms of its contribution to the final similarity score, and this "importance" is determined by the weights term. Weighted similarity can be explained in practice: it is reasonable to deem that the overall features (lower bands) and detailed features (higher bands) should not be accounted equal weight when comparing two shapes, e.g., in a data set consisting of actives and negatives that vary significantly in size, it is the overall features that will be more helpful in discriminating them, so the lower bands coefficients will be more important in shape comparison and thus should get a higher weight. Conversely, when faced with actives and negatives of similar sizes, their overall features will almost have no effect in distinguishing them.

## Genetic optimization of weights

Given a pair of SH shape descriptors, their SH similarity score varies if their corresponding weights change. That is, for a given specific query molecule, the resulting ranking list will differ if the weight combinations vary. The purpose of weights optimization is to find the group of weights that best separates actives and negatives in screening a data set, which means that this group of weights gives a high SH similarity score for actives, but a low score for negatives, and thus the actives can be retrieved easily with fewer false actives.

In order to carry out genetic optimization, a reference set consisting of actives and negatives is needed (stage 1 in Fig. 2). This reference set can be totally customized regardless of structure or the number of reference molecules. In this paper, the reference set was selected randomly (using a 3-fold cross-validation scheme) from clustered DUD actives and corresponding decoys (Table 1). For the purpose of identifying patterns shared by the query conformer and reference actives, a pre-process was applied to all reference actives: first, the conformation space was enumerated using Pipeline Pilot [39] (with parameter *Maximum* set to 200). Second, for each active, we searched for the enumerated conformation space for one conformer that was most similar to the query conformer in shape, and selected a representation of that active in the reference set by means of USR [26, 27]. To reduce artificial bias, the molecules in this reference set were not used again later for further VS.

Assuming there are $N$ actives in the reference set, the ideal case is that the top $N$ hits of the ranking list with high shape similarity are active compounds. Based on this proposal, given a group of weights and the corresponding query, fitness can be calculated from the result ranking list over reference set (stage 2 in Fig. 2). Thus, the fitness score can be expressed as follows:

$$Fitness = \sum_i S_{i,\,active} + F_{scale} \sum_i \left(1 - S_{i,\,negative}\right) - Penalty \quad (8)$$

where $S_{\mathrm{i,actives}}$ is the SH similarity score for every correct active prediction, $S_{\mathrm{i,negative}}$ is the SH similarity score for every correctly classified negative, $F_{\mathrm{scale}}$ is a scale factor for adjusting the respective contribution of active and negative predictions to final fitness. To cope with the problem of imbalanced data sets [42], $F_{\mathrm{scale}}$ is set as (no. of reference actives) / 2×(no. of reference negatives). The purpose of the extra term *Penalty* is to provide a positive penalty term for weight violation to decrease the fitness of weights violating the constraints, $w_i < 0$ or $\sum_i w_i \neq 1$ for example. The penalty term is defined as follows:

$$\begin{aligned} Penalty &= P_1 + P_2 + P_3 \\ P_1 &= F_{penalty} \sum_i |w_i|,\ w_i < 0 \\ P_2 &= F_{penalty} \sum_i |w_i - B_{upper}|,\ w_i > B_{upper} \\ P_3 &= F_{penalty} \left| 1 - \sum_i w_i \right| \end{aligned} \quad (9)$$

where $P_1$ is the penalty for negative weights; $P_2$ is the penalty for weights exceeding a specified upper bound, $B_{\mathrm{upper}}$; $P_3$ is the penalty for normalizing violation; and $F_{\mathrm{penalty}}$ is a scale factor determining the overall penalty magnitude for fitness. In this paper, $B_{\mathrm{upper}}$ is set as 0.4, which means any weight larger than 0.4 will get some degree of penalty in fitness, and $F_{\mathrm{penalty}}$ is set as 40, which is roughly the maximum number of reference actives in each reference set.

The weight optimization is implemented in a standard real value based genetic algorithm (GA) framework described briefly as follows. Each chromosome is encoded as an array of $L_{\mathrm{max}}$ real values corresponding to each weight. In the whole evolution process, a standard fitness based Roulette wheel selection is applied to select fitted individuals. A standard single point crossover operator is applied to each selected chromosome to produce offspring. All offspring then pass a mutation operator to enter the next generation.

## Results evaluation

A principle component analysis (PCA) approach was adopted to validate the SH shape descriptor, which lays the foundation for further weights optimization and shape comparison. To present the screening result visually, a receiver operating characteristic (ROC) curve provides a useful qualitative guide [43, 44], plotted as the true positive

(TP) rate against the false positive (FP) rate. However, for quantitative assessment, the area under the curve (AUC) was calculated for each ROC curve [44], which provides a comparable overall performance measurement. The enrichment factor (EF) [45] at different points in the ranking list was calculated to assess preliminary recognition ability. Note that AUC is just an overall measurement and cannot reflect the early recognition ability, which was measured by means of EF and ROC plots in our study. The efficiency and convergence capability of weights optimization are demonstrated by evaluating the trend of the fitness changes during the genetic searching process.

## Results and discussion

### Descriptor analysis

For any category of molecular shape descriptor, the essence is a mapping from the original abstract shape space to an elaborately designed descriptor space, which is often lower-dimensional, and, due to the dimension reduction, information loss is inevitable. Thus, a qualified descriptor should always contain core information for shape comparison. To show that the SH shape descriptor in this study preserves enough shape information to separate the actives and negatives, PCA analysis was applied to the 7-dimensional descriptors generated. Figure 4 plots the spatial distribution of the actives and negatives from the ACHE dataset in SH descriptor space. The three axes in Fig. 4 correspond to the top three principle components with the most significance (explained variation, namely the eigenvalue of covariance matrix) calculated from PCA. It can be seen that the actives exhibit some kind of perceptible aggregation, while the negatives are dispersed in a rather vast space. Although there is some slight overlap between actives and decoys, it does not mean that the descriptors fail to preserve enough shape information to allow classification. It is important to note that, although some slight overlap exists, the distributions of actives and decoys have already presented some clear patterning: actives are clustered and distinguished from decoys. It is very common to see overlap in data analysis, and overlapped samples can still be classified correctly using certain non-linear approaches. The intention of Fig. 4 is not to prove that the actives and decoys represented by SH descriptors are linearly separable, but to demonstrate that the shape information preserved by the descriptors is enough (if not, all actives and decoys would be distributed randomly and no clear pattern would be observed) to perform VS.

The results of PCA of the ACHE dataset, i.e., the explained variation of each principal component (presented as bars) and the accumulated explained variation (presented as a curve), are shown in Fig. 5. It can be seen that the top three principle components explain more than 80% of the variation of samples in the descriptor space, which means that the 3D spatial distribution illustrated in Fig. 4 can precisely reflect the real spatial distribution in the descriptor space.

### Performance analysis

The overall performance of SHeMS on the selected DUD sub datasets is presented as a summary of average AUC values in Table 2, along with corresponding AUC values achieved by OSH and USR methods adopting identical actives, negatives and queries as references. The OSH method was implemented in the same way as the SHeMS

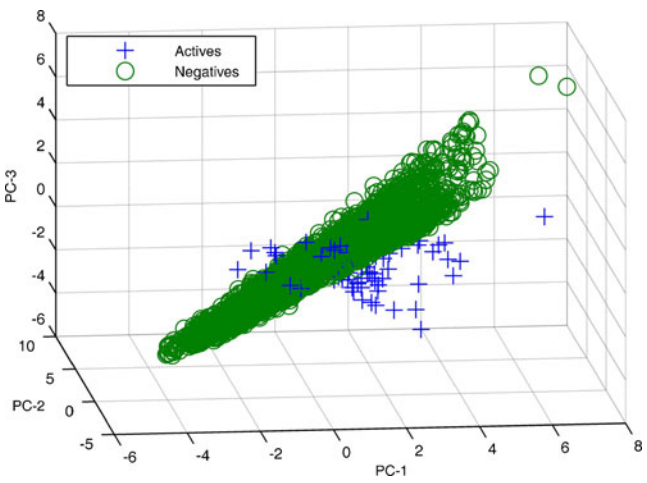



**Fig. 4** Scattergram of SH shape descriptor for actives and negatives in the ACHE dataset; + actives; ○ negatives

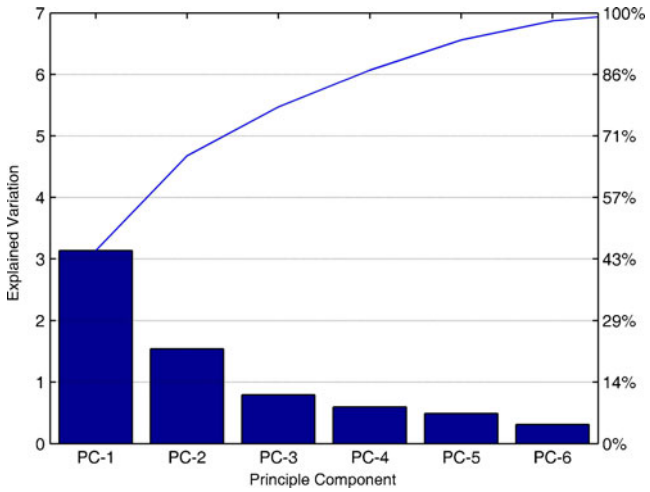



**Fig. 5** Explained variations for each principle component of the SH shape descriptor in the ACHE dataset. *Bars* Explained variations for each principle component, *curve* accumulated explained variation

**Table 2** Average area under the curve (AUC) values for rapid overlay of chemical structures (ROCS), shape and electrostatic potential (ShaEP), original spherical harmonics (OSH), SH-expansion-based molecular similarity (SHeMS) and ultra-fast shape recognition (USR) methods against the selected DUD sub datasets

| Dataset | AUC ROCS | AUC ShaEP | AUC OSH | AUC SHeMS | AUC USR | Enhance Rate[a] | ms/conf.[b] |
|---|---|---|---|---|---|---|---|
| ACHE | 0.701 | 0.712 | 0.590 | 0.633 | 0.657 | 7.31% | 96 |
| AR | 0.854 | 0.848 | 0.509 | 0.565 | 0.557 | 11.06% | 92 |
| COX2 | 0.977 | 0.981 | 0.704 | 0.805 | 0.754 | 14.34% | 95 |
| DHFR | 0.484 | 0.513 | 0.361 | 0.376 | 0.428 | 4.10% | 94 |
| EGFR | 0.951 | 0.964 | 0.626 | 0.661 | 0.686 | 5.63% | 93 |
| GPB | 0.719 | 0.709 | 0.537 | 0.565 | 0.568 | 5.23% | 90 |
| INHA | 0.530 | 0.526 | 0.492 | 0.516 | 0.565 | 4.91% | 97 |
| P38 | 0.953 | 0.957 | 0.683 | 0.709 | 0.484 | 3.76% | 96 |
| PARP | 0.688 | 0.679 | 0.523 | 0.571 | 0.509 | 9.28% | 90 |
| PDGFRB | 0.973 | 0.981 | 0.715 | 0.731 | 0.507 | 2.18% | 94 |
| PNP | 0.789 | 0.778 | 0.532 | 0.578 | 0.689 | 8.61% | 89 |
| PR | 0.919 | 0.923 | 0.689 | 0.710 | 0.603 | 3.02% | 94 |
| SAHH | 0.967 | 0.956 | 0.616 | 0.628 | 0.662 | 2.02% | 89 |
| SRC | 0.617 | 0.609 | 0.318 | 0.321 | 0.308 | 0.87% | 97 |
| VEGFR2 | 0.543 | 0.552 | 0.461 | 0.472 | 0.485 | 2.30% | 96 |

[a] Average AUC enhancement achieved by SHeMS method over OSH method

[b] Computation time (milliseconds per conformer) for SHeMS method

method except that it does not contain the weights optimization process, and a group of equal weights (equal to 1/7) are applied. The AUC values listed in Table 2 were calculated as mean values of all queries applied corresponding to a specific dataset. All AUC data in Table 2 were obtained from a 3-fold cross-validation. This metric was chosen on purpose, aiming to reduce the AUC variation caused by individual query templates and to reflect the true overall performance without bias being introduced by the query template.

Alignment based methods (ROCS and ShaEP) outperformed alignment-free methods (SHeMS, USR and OSH) in most of the datasets, which is not surprising since the alignment achieves high precision at the expense of computation time. Among all 15 selected DUD sub datasets, SHeMS achieves an average AUC of 0.589, and with a median of 0.578. The average AUC for the OSH method and USR method are 0.557 and 0.564, with a median of 0.537 and 0.565, respectively. Both the average and median AUC values for SHeMS are well above 0.5, which corresponds to a random selection process, indicating adaptability in retrieving actives of diverse shapes. On the other hand, the average and median AUC values for the OSH method are relatively low compared with that of SHeMS, suggesting a lack of pertinence for a specific query. This result illustrates well the efficiency of weights optimization and rationality of weighted shape similarity. For the SHeMS method, the overall performance is obviously target dependent: among all 15 dataset specific tests, 3 (DHFR, SRC and VEGFR2 datasets) generate an AUC value below 0.5, indicating a complete failure to retrieve actives (both other reference methods also failed in these three datasets). The OSH and USR methods failed in four datasets each. Compared with the OSH method, SHeMS gains some degree of performance improvement in all selected DUD sub datasets to a greater or lesser extent (there are two cases with an average AUC enhancement higher than 10%: AR and COX2 datasets). The AUC enhancement ratio of SHeMS with optimized method over the OSH method is presented in Table 2.

Various shape-based VS methods have been validated in many other studies. For example, Nicholls et al. [46] evaluated various shape-based methods using the DUD database and reported similar AUC values. Kahraman et al. [47] tested an SH-based method for protein–ligand shape matching on a private dataset, and reported a relatively high AUC value. This variation in AUC seems to have been caused by the difference in datasets as follows. Each decoy in the DUD data set was chosen to be similar in size (a crucial factor affecting performance) and other physical properties to the corresponding ligand, which makes DUD a challenging benchmark for VS. The dataset adopted by Kahraman et al. [47] contains no dedicated decoys, and the authors claim that the ligands in the dataset were of different sizes, which makes this dataset less challenging than DUD.

The computational time (4,900 surface points per conformer—a dominant factor affecting computation time) on each dataset is also listed in Table 2. For all 15 selected DUD sub datasets, the computation time was no more than 100 ms per conformer, varying slightly between datasets. Nearly all computation time is consumed by descriptor generation (descriptor comparison and ranking consume less than 0.01 ms per conformer), which means the VS process can be performed at extremely high speed if the descriptor (only seven float numbers) is calculated and stored in advance. The computation time for descriptor comparison and ranking, which depends largely on the underlying sort algorithm, is reported, not intending to demonstrate efficiency but rather to keep the assessment intact.

Early recognition is also a critical aspect of evaluating a ranking method [44, 45, 48, 49], especially the retrieval of actives in the top 20% of the rank-order list [45]. In order to demonstrate that SHeMS is capable of discriminating actives in the early stage of the ranking list, and that it is meaningful for a practical pharmaceutical program, we focused on average EF values corresponding to each selected DUD sub dataset at 0.1%, 2%, 5% and 10% of the ranking list in Table 3. The equivalent analysis was also applied to the ROCS, ShaEP, OSH and USR methods for comparison, with the same actives, negatives and queries. The EF values listed in Table 3 were also calculated as mean values of all queries applied corresponding to a specific dataset. All EF data in Table 3 were obtained from a 3-fold cross-validation.

The alignment-based methods (ROCS and ShaEP) achieve far higher EF values compared to alignment-free methods (SHeMS, USR and OSH), especially at the 0.1% level. The EF values at the 0.1% level vary drastically due to the fact that only a very small portion of the rank list is taken into consideration. The USR method performs very well at this level. At the 2% EF level, neither a significant enhancement nor a drop is observed. Most EF values at this level maintain or show only very slight perturbation after weights optimization. Considering the relatively small percentage level, this variation is rather trivial. The USR method outperforms SHeMS in 11 out of 15 of the datasets tested, but seems unable to maintain this advantage at the 5% EF level. At the 5% EF level, the EF values of SHeMS show a clear enhancement over the OSH method. Among all 15 selected datasets, 11 achieve a higher or equal EF value after optimized weights were applied. Reviewing the four cases (namely PARP, PDGFRB, PR and SAHH dataset) with a decreased EF value, the average amount of the EF decrease is only 0.12, which is fairly low compared with the average EF value. At this level, SHeMS outperforms the USR method on seven datasets. The EF value of SHeMS exceeds 6.0 on four datasets, while for the USR method there is only one case in which the EF value is just around 6.0.

**Table 3** Average enrichment factor (EF) values at different cutoffs (0.1%, 2%, 5% and 10%) for the ROCS, ShaEP, OSH, SHeMS and USR methods against the selected DUD sub datasets

| | EF @ 0.1% | | | | | EF @ 2% | | | | | EF @ 5% | | | | | EF @ 10% | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | ROCS | ShaEP | SHeMS | USR | OSH | ROCS | ShaEP | SHeMS | USR | OSH | ROCS | ShaEP | SHeMS | USR | OSH | ROCS | ShaEP | SHeMS | USR | OSH |
| ACHE | 40.33 | 35.56 | 4.02 | 12.64 | 5.17 | 12.65 | 13.70 | 3.24 | 4.21 | 2.45 | 6.56 | 6.61 | 2.54 | 3.06 | 2.28 | 3.88 | 3.96 | 2.16 | 2.45 | 1.86 |
| AR | 45.45 | 40.45 | 11.72 | 23.44 | 15.63 | 25.68 | 23.56 | 4.68 | 5.67 | 1.77 | 11.91 | 12.13 | 2.97 | 2.65 | 2.44 | 6.36 | 6.50 | 2.21 | 1.77 | 1.84 |
| COX2 | 255.51 | 250.11 | 16.00 | 76.80 | 12.80 | 46.05 | 45.78 | 12.78 | 10.29 | 3.45 | 18.90 | 17.55 | 7.79 | 5.38 | 4.34 | 9.58 | 9.32 | 4.92 | 3.45 | 3.17 |
| DHFR | 75.16 | 75.10 | 0.00 | 5.71 | 0.00 | 8.66 | 8.71 | 0.51 | 0.99 | 0.51 | 4.02 | 4.17 | 0.57 | 0.63 | 0.54 | 2.42 | 2.78 | 0.50 | 0.51 | 0.46 |
| EGFR | 156.04 | 140.43 | 0.35 | 53.55 | 0.71 | 37.69 | 35.23 | 8.24 | 10.76 | 4.16 | 16.51 | 15.69 | 4.90 | 6.40 | 4.50 | 8.73 | 8.35 | 3.28 | 4.16 | 3.08 |
| GPB | 65.97 | 60.78 | 3.09 | 12.35 | 12.35 | 15.80 | 16.47 | 4.40 | 7.52 | 2.99 | 8.13 | 8.01 | 3.19 | 4.75 | 2.75 | 4.79 | 4.54 | 2.26 | 2.99 | 2.13 |
| INHA | 56.21 | 55.21 | 23.68 | 10.53 | 21.05 | 6.21 | 6.33 | 3.55 | 2.22 | 1.33 | 3.02 | 2.82 | 2.15 | 1.72 | 2.17 | 2.01 | 2.28 | 1.51 | 1.33 | 1.60 |
| P38 | 125.54 | 119.65 | 42.86 | 44.76 | 68.57 | 34.20 | 32.92 | 12.57 | 4.93 | 1.67 | 15.87 | 16.23 | 6.91 | 2.65 | 6.93 | 8.64 | 8.71 | 4.34 | 1.67 | 4.13 |
| PARP | 52.63 | 53.29 | 30.61 | 45.92 | 35.71 | 16.08 | 16.57 | 4.59 | 5.68 | 1.77 | 7.37 | 7.11 | 2.99 | 2.77 | 2.73 | 4.21 | 4.20 | 2.38 | 1.77 | 2.15 |
| PDGFRB | 200.00 | 182.57 | 122.02 | 71.43 | 113.10 | 41.83 | 40.38 | 15.31 | 8.17 | 2.59 | 17.87 | 16.45 | 8.13 | 4.19 | 7.85 | 9.33 | 9.28 | 4.95 | 2.59 | 4.95 |
| PNP | 54.69 | 50.13 | 13.89 | 41.67 | 20.83 | 19.92 | 20.26 | 3.26 | 8.98 | 3.49 | 10.31 | 11.28 | 2.29 | 4.90 | 1.72 | 6.25 | 6.35 | 1.77 | 3.49 | 1.69 |
| PR | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 44.17 | 40.79 | 10.88 | 11.76 | 3.39 | 18.00 | 18.43 | 6.78 | 5.73 | 6.65 | 9.00 | 9.24 | 4.59 | 3.39 | 4.43 |
| SAHH | 62.50 | 43.12 | 22.73 | 37.88 | 30.30 | 27.68 | 30.36 | 6.64 | 8.58 | 3.34 | 15.36 | 1424 | 4.52 | 4.64 | 4.72 | 9.02 | 9.17 | 3.26 | 3.34 | 3.18 |
| SRC | 113.96 | 102.23 | 15.00 | 15.00 | 25.00 | 13.39 | 15.28 | 2.16 | 2.39 | 0.85 | 6.44 | 5.78 | 1.07 | 1.35 | 1.07 | 3.68 | 3.56 | 0.69 | 0.85 | 0.80 |
| VEGFR2 | 49.38 | 50.10 | 10.99 | 21.98 | 10.99 | 6.79 | 8.66 | 2.37 | 3.60 | 1.91 | 3.46 | 3.10 | 1.77 | 2.30 | 2.02 | 2.16 | 2.10 | 1.32 | 1.91 | 1.42 |

**Table 4** Average AUC values for training sets with different number of decoys

| Dataset | Decoys 60 | Decoy 120 | Decoys 240 | Decoys 480 |
|---|---|---|---|---|
| ACHE | 0.63 | 0.63 | 0.63 | 0.62 |
| AR | 0.60 | 0.57 | 0.60 | 0.58 |
| COX2 | 0.78 | 0.81 | 0.72 | 0.72 |
| DHFR | 0.36 | 0.38 | 0.36 | 0.35 |
| EGFR | 0.64 | 0.66 | 0.63 | 0.61 |
| GPB | 0.61 | 0.57 | 0.59 | 0.61 |
| INHA | 0.50 | 0.52 | 0.50 | 0.51 |
| P38 | 0.73 | 0.71 | 0.68 | 0.68 |
| PARP | 0.59 | 0.57 | 0.56 | 0.54 |
| PDGFRB | 0.64 | 0.73 | 0.64 | 0.66 |
| PNP | 0.61 | 0.58 | 0.61 | 0.60 |
| PR | 0.70 | 0.71 | 0.71 | 0.68 |
| SAHH | 0.63 | 0.63 | 0.62 | 0.59 |
| SRC | 0.33 | 0.32 | 0.33 | 0.34 |
| VEGFR2 | 0.49 | 0.47 | 0.49 | 0.48 |
| Avg. | 0.59 | 0.59 | 0.58 | 0.57 |

To validate the performance of SHeMS supplied with training sets of different sizes, various numbers of decoys are chosen to repeat weight optimization, and the average VS results for each dataset are listed in Table 4. Depending on the dataset size for each DUD target, these chosen decoys take up roughly 0.4% to 50% of the target-specific decoys, and the ratio of actives to decoys ranges from 1:60 to 2:3. As can be seen from Table 4, the AUC value for each DUD dataset shows only very slight variation among the different training sets, and the average AUC value over all datasets differs only 0.01–0.02, which demonstrates that the proposed method is insensitive to the size and imbalance rate of the training set.

To carry out further analysis of the detailed performance and demonstrate the optimization process, 15 representative query templates (one for each DUD sub dataset) were selected from the respective reference set to re-execute the screening process with SHeMS and four reference methods (ROCS, ShaEP, USR and OSH). The AUC values of the screening results are summarized in Table 5 along with the ZINC [50] code of the query molecule. Note that all validations in our study are purely shape-based, which means only the shape scores are taken into consideration for ROCS and ShaEP.

The ROC curve for the four methods for each selected query is plotted in Fig. 6. Generally, the ShaEP and ROCS methods (both methods based on Gaussian density alignment and combined with other aspects such as electrostatic potential, hydrogen donor / acceptor, hydrophobic groups, etc., but these aspects are not taken into consideration in this study) outperform the SHeMS, USR and OSH methods, which consider only shape similarity, in identifying bioactive compounds. In addition, the fact that SHeMS gains improvement over the OSH and USR methods is undoubtedly through parameter optimization. For SHeMS, the best performance is observed on the PDGFRB, P38 and COX2 dataset, showing both high early enrichment (the curve is rather steep at the initial stage with low false positive rate) and high AUC (the main body of the curve shifts towards the top left, enclosing a large area of the *x*-

**Table 5** AUC values for the ROCS, ShaEP, OSH, SHeMS and USR methods against the selected DUD sub datasets

| Dataset | Query | AUC ROCS | AUC ShaEP | AUC OSH | AUC SHeMS | AUC USR | Opt time (s)[a] |
|---|---|---|---|---|---|---|---|
| ACHE | ZINC03814252 | 0.767 | 0.789 | 0.647 | 0.725 | 0.734 | 7 |
| AR | ZINC03814383 | 0.875 | 0.901 | 0.614 | 0.713 | 0.581 | 6 |
| COX2 | ZINC00838696 | 0.973 | 0.987 | 0.798 | 0.892 | 0.700 | 5 |
| DHFR | ZINC03815003 | 0.678 | 0.929 | 0.459 | 0.451 | 0.500 | 6 |
| EGFR | ZINC00118581 | 0.977 | 0.988 | 0.689 | 0.791 | 0.896 | 7 |
| GPB | ZINC04134456 | 0.866 | 0.847 | 0.795 | 0.815 | 0.792 | 5 |
| INHA | ZINC03832029 | 0.550 | 0.600 | 0.625 | 0.657 | 0.633 | 5 |
| P38 | ZINC04617900 | 0.908 | 0.952 | 0.800 | 0.868 | 0.465 | 6 |
| PARP | ZINC03832188 | 0.774 | 0.798 | 0.548 | 0.716 | 0.611 | 6 |
| PDGFRB | ZINC00008076 | 0.998 | 0.999 | 0.920 | 0.958 | 0.710 | 5 |
| PNP | ZINC00391097 | 0.607 | 0.609 | 0.556 | 0.678 | 0.670 | 5 |
| PR | ZINC03832325 | 0.898 | 0.907 | 0.743 | 0.775 | 0.513 | 5 |
| SAHH | ZINC01483030 | 0.907 | 0.875 | 0.750 | 0.807 | 0.800 | 5 |
| SRC | ZINC04617935 | 0.476 | 0.552 | 0.356 | 0.351 | 0.235 | 6 |
| VEGFR2 | ZINC00008981 | 0.741 | 0.772 | 0.667 | 0.674 | 0.557 | 6 |

[a] Time (seconds per run) consumed by the weights optimization process

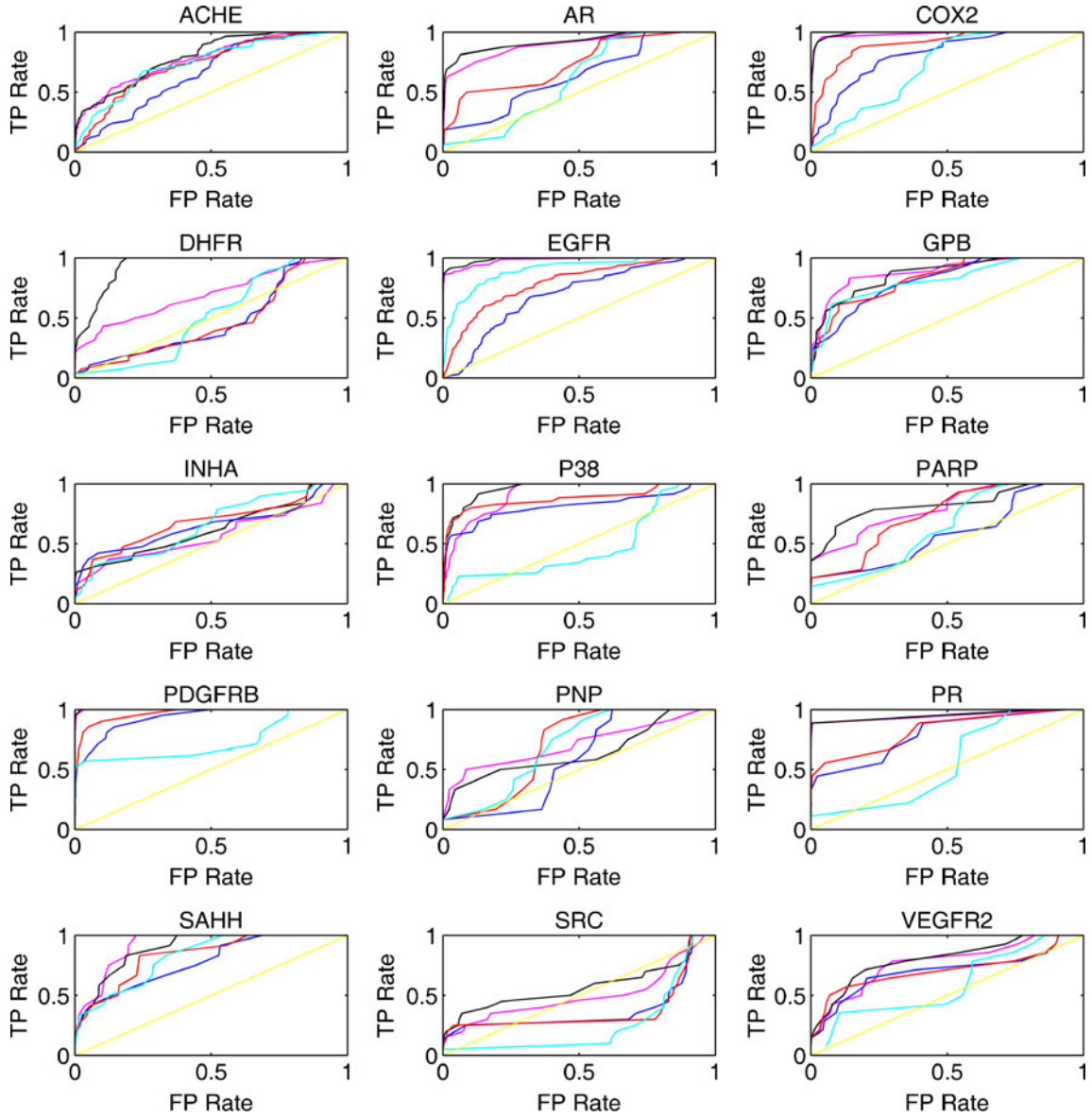



**Fig. 6** ROC plot for the selected DUD sub dataset. *Red* SHeMS, *blue* OSH method, *cyan* USR method, *black* ShaEP method, *magenta* ROCS method, *yellow* random selection

axis). With the PDGFRB dataset, the ShaEP and ROCS methods achieve a nearly ideal result, with the ROC curve appearing as almost a vertical straight line passing through the origin. SHeMS is next, giving an AUC value over 0.95, and with early enrichment far exceeding other two methods. On the P38 dataset, the SHeMS and ShaEP methods yield an almost identical ROC curve, with an FP rate ranging from 0 to 10% achieving a TP rate around 80%, which is clearly higher than the ROCS method, while the USR method performs no better than random pick up. Another similar case can be found on dataset VEGFR2, with SHeMS showing a better performance at an FP rate of 10%. On the GPB and INHA datasets, the performance of all methods differs slightly. Moreover, for specific kinase datasets, such as EGFR, SRC and VEGFR2, since the shapes of actives and decoys are highly similar, the above mentioned methods that merely compare the 3D shape of the molecule, cannot effectively identify those actives with low EF values (Table 5).

## Optimization analysis

To gain an overview of the genetic optimization process and to show the effectiveness of the optimization, the fitness trend of the whole evolution course for each selected dataset is presented in Fig. 7. The fitness is plotted against

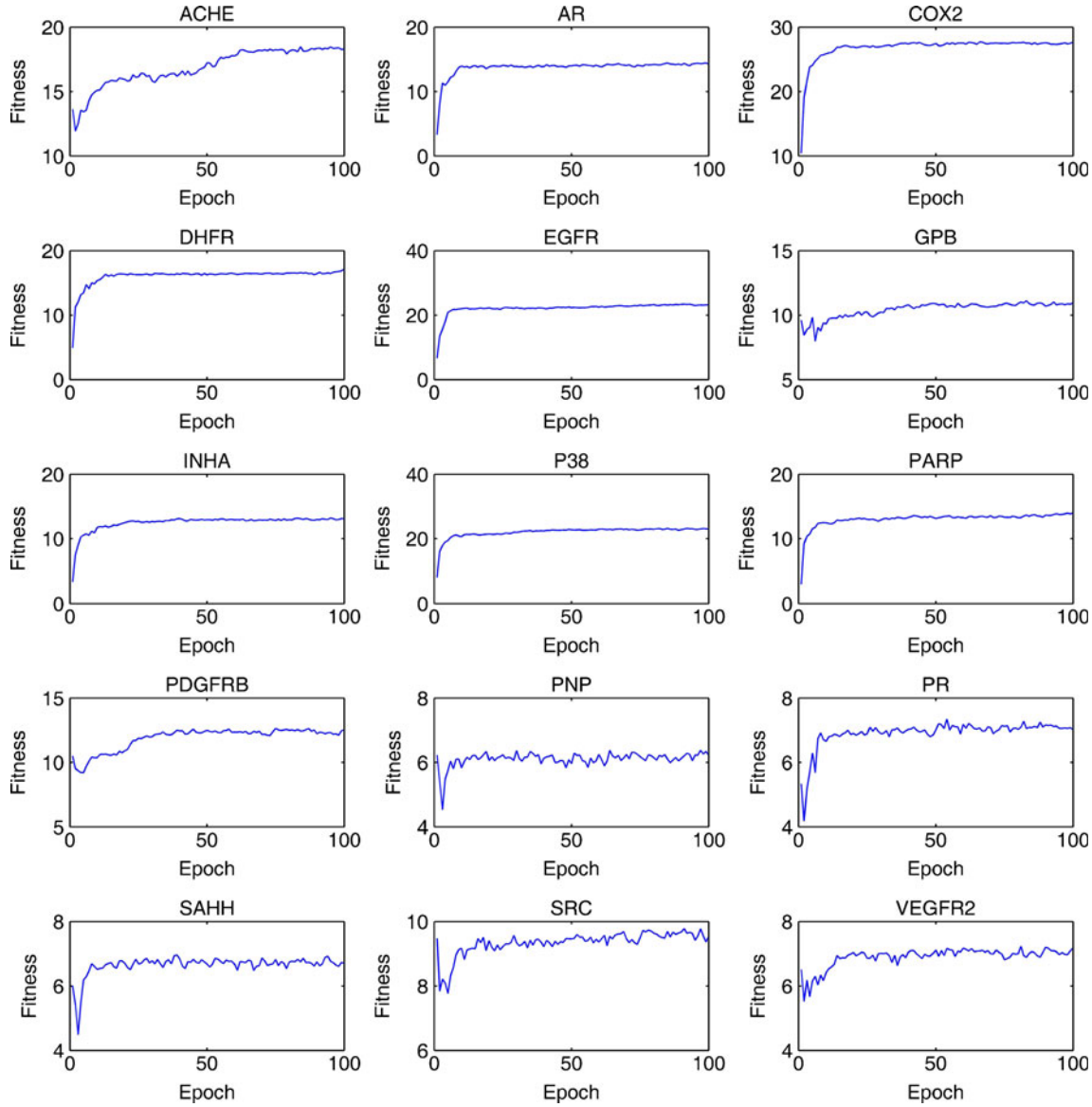



**Fig. 7** Optimized fitness plot for each selected DUD sub dataset

each epoch (generation) in the time series of the evolution, forming a converged curve. The fitness plotted here is the overall fitness (see "Materials and methods" section for precise definition) related to all individuals in the current generation, which can be converted to a mean fitness scaled in a fixed domain if divided by individual amounts. Owing to the nature of the genetic algorithm used, the fitness alteration of each epoch is never continuous, resulting in a large amount of minor peaks in the fitness curve instead of one smooth curve.

All the 15 fitness plots in Fig. 7 display an upward trend of fitness, with the epoch moving forward. Because fitness is related directly to the ability of distinguishing actives from negatives, the efficiency of weights optimization is definitely positive. Figure 7 depicts the changing fitness situation only within the first 100 generations, which is also the maximum generation limit in our test runs (this parameter is user configurable), but the fitness curve has already converged well. In fact, no obvious upward trend is observed after 50 generations (even after a mere 25 generations) indicating a convergence of optimization. This quick convergence phenomenon can be attributed to the relatively low dimensions of the searching space (seven dimensions in total, one for each weight), which means the computation time for optimization process can be further reduced substantially in real applications.

The time required for each optimization run (100 generations) is also listed in Table 5. The optimization time

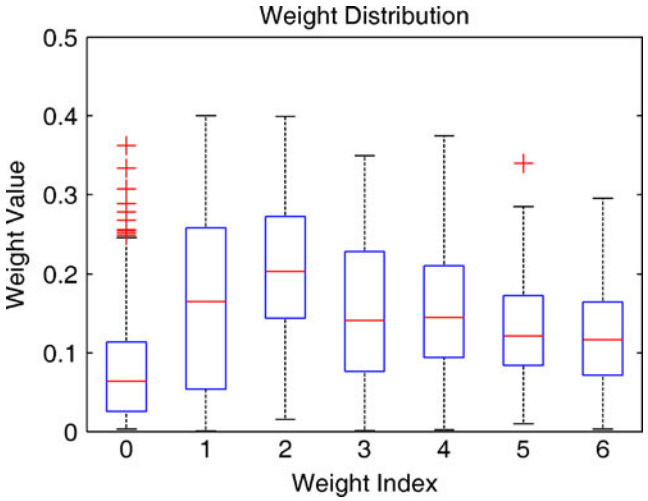



**Fig. 8** Distribution of optimal weights. Each SH weight (order 0 to order 6) is represented as a box indicating its overall distribution. Each *box* corresponds to one weight for a certain order; the lower and upper horizontal edge of the box mark the 25th and 75th percentiles, respectively. The *horizontal bar* located in each box marks the mean value for the weights. The whisker drawn as a *dotted line* for each box spans 1.5 times the height of the box. Any weight located outside the whisker is considered an outlier and marked with a *plus sign*

depends largely on the maximum number of evolution generations allowed and on the size of the reference set. For most cases in our experiments, the optimization process takes less than 7 s per query, which can be further reduced substantially (at least by half) due to the quick convergence mentioned above.

The distribution of optimal weights obtained from GA optimization for all 15 selected DUD datasets is plotted in Fig. 8. Each box in Fig. 8 corresponds to one weight for a certain order, and the lower and upper horizontal edge of the box mark the 25th and 75th percentiles, respectively, thus 50% weights are distributed within the box. The horizontal bar located in each box marks the mean value for the weights. The whisker drawn as a dotted line for each box spans 1.5 times the height of the box, which indicates a 99.3% coverage if the weights are distributed normally. Any weight located outside the whisker is considered an outlier and marked with a plus sign. It can be seen that the mean values of weights for lower (0) and higher (5 and 6) orders are relatively smaller than for the others, and the corresponding weights are distributed over a very narrow range. This observation indicates that the SH coefficients corresponding to these weights are less important for distinguishing actives from decoys, which can be explained as follows. The coefficient for order 0 is the ground level for all other SH expansions and depicts only the overall size of a shape. Because all decoys in the DUD datasets are selected to have a size similar to that of actives, then the coefficient for order 0 is similar for both actives and decoys, hence this coefficient is less helpful for classification. The coefficients for higher orders (order 5 and 6) depict more detailed features that are trivial between actives and decoys, and thus provides no further information for classification.

## Conclusions

In this study, a novel SH expansion based weighted molecular shape similarity metric, SHeMS, was devised, and a genetic optimization approach was carried out to find the best combination of weights according to a user customized reference molecule set. These two aspects comprise SHeMS, which is adaptable to specific systems. The discriminatory power of the SH shape descriptor was verified through a PCA approach forming the basis for weights optimization, and this shape descriptor, which possesses the characteristic of rotation invariant, preserves abundant information for shape comparison and reduces storage requirements. A retrospective test was implemented on selected DUD datasets, demonstrating a significant improvement over OSH and USR methods in either overall performance or early enrichment. In many cases, the performance of the proposed method was comparable to other popular shape methods. However, as it merely compares the 3D shape of the molecule, SHeMS cannot effectively identify all actives from a large number of decoys with similar shapes against the test set in this study. Through combining efficient shape similarity comparison with SH expansion methods and other aspects such as chemical and pharmacophore features, SHeMS can make a significant contribution and can be practically applied to virtual screening, by means of similarity comparison with the 3D shape of the known active compound or binding pocket of a protein. This is the focus of further research in our laboratory.

**Acknowledgments** This work was supported by the Special Fund for Major State Basic Research Project (grant 2009CB918501), the National Natural Science Foundation of China (grants 20803022), the Shanghai Committee of Science and Technology (grants 09dZ1975700 and 10431902600), the 863 Hi-Tech Program of China (grant 2007AA02Z304), and the Major National Scientific and Technological Project of China (grant 2009ZX09501-001). H.L. is also sponsored by Shanghai Rising-Star Program (grant 10QA1401800) and the Fundamental Research Funds for the Central Universities. The program and test sets of SHeMS are available from H.L. upon request.

**Author contributions** C.C. designed and validated the Cyndi method, and also contributed to analysis and data interpretation and co-drafted the manuscript with J.G. and X.L. J.G. contributed to the design of method. X.L. contributed to method validation. H.L. conceived the idea of the SHeMS and provided direction for its development and revised the subsequent drafts of this manuscript with D.G., and H.J. All authors read and approved the final manuscript.

## References


1. Rester U (2008) From virtuality to reality - Virtual screening in lead discovery and lead optimization: a medicinal chemistry perspective. Curr Opin Drug Discov Dev 11:559–568
2. Shoichet BK (2004) Virtual screening of chemical libraries. Nature 432:862–865
3. Kirchmair J, Markt P, Distinto S, Schuster D, Spitzer GM, Liedl KR, Langer T, Wolber G (2008) The Protein Data Bank (PDB), its related services and software tools as key components for in silico guided drug discovery. J Med Chem 51:7021–7040
4. Hristozov DP, Oprea TI, Gasteiger J (2007) Virtual screening applications: a study of ligand-based methods and different structure representations in four different scenarios. J Comput Aided Mol Des 21:617–640
5. Waszkowycz B (2008) Towards improving compound selection in structure-based virtual screening. Drug Discov Today 13:219–226
6. Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov 3:935–949
7. Lengauer T, Rarey M (1996) Computational methods for biomolecular docking. Curr Opin Struct Biol 6:402–406
8. Evers A, Hessler G, Matter H, Klabunde T (2005) Virtual screening of biogenic amine-binding G-protein coupled receptors: Comparative evaluation of protein- and ligand-based virtual screening protocols. J Med Chem 48:5448–5465
9. Hawkins PCD, Skillman AG, Nicholls A (2007) Comparison of shape-matching and docking as virtual screening tools. J Med Chem 50:74–82
10. McGaughey GB, Sheridan RP, Bayly CI, Culberson JC, Kreatsoulas C, Lindsley S, Maiorov V, Truchon JF, Cornell WD (2007) Comparison of topological, shape, and docking methods in virtual screening. J Chem Inf Model 47:1504–1519
11. Perez-Nueno VI, Ritchie DW, Rabal O, Pascual R, Borrell JI, Teixido J (2008) Comparison of ligand-based and receptor-based virtual screening of HIV entry inhibitors for the CXCR4 and CCR5 receptors using 3D ligand shape matching and ligand-receptor docking. J Chem Inf Model 48:509–533
12. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS (2006) A critical assessment of docking programs and scoring functions. J Med Chem 49:5912–5931
13. Tirado-Rives J, Jorgensen WL (2006) Contribution of conformer focusing to the uncertainty in predicting free energies for protein-ligand binding. J Med Chem 49:5880–5884
14. Gohlke H, Klebe G (2002) Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. Angew Chem Int Ed Engl 41:2645–2676
15. Daylight Chemical Information Systems, Inc. Daylight Chemical Information Systems Inc. http://www.daylight.com. Accessed 1 May 2010
16. Tripos, Inc. http://www.tripos.com. Accessed 1 May 2010
17. Durant JL, Leland BA, Henry DR, Nourse JG (2002) Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comput Sci 42:1273–1280
18. Cheeseright TJ, Mackey MD, Melville JL, Vinter JG (2008) FieldScreen: Virtual Screening Using Molecular Fields. Application to the DUD Data Set. J Chem Inf Model 48:2108–2117
19. Cheeseright T, Mackey M, Rose S, Vinter A (2006) Molecular field extrema as descriptors of biological activity: definition and validation. J Chem Inf Model 46:665–676
20. Marin RM, Aguirre NF, Daza EE (2008) Graph theoretical similarity approach to compare molecular electrostatic potentials. J Chem Inf Model 48:109–118
21. Ronkko T, Tervo AJ, Parkkinen J, Poso A (2006) BRUTUS: optimization of a grid-based similarity function for rigid-body molecular superposition. II. Description and characterization. J Comput Aided Mol Des 20:227–236
22. Thorner DA, Willett P, Wright PM, Taylor R (1997) Similarity searching in files of three-dimensional chemical structures: representation and searching of molecular electrostatic potentials using field-graphs. J Comput Aided Mol Des 11:163–174
23. Vainio MJ, Puranen JS, Johnson MS (2009) ShaEP: molecular overlay based on shape and electrostatic potential. J Chem Inf Model 49:492–502
24. Good AC, Hodgkin EE, Richards WG (1992) The utilisation of Gaussian functions for the rapid evaluation of molecular similarity. J Comput Inf Comput Sci 32:188–191
25. Grant JA, Gallardo MA, Pickup BT (1996) A fast method of molecular shape comparison: a simple application of Gaussian descriptor of molecular shape. J Comput Chem 17:1653–1666
26. Ballester PJ, Richards WG (2007) Ultrafast shape recognition to search compound databases for similar molecular shapes. J Comput Chem 28:1711–1723
27. Ballester PJ, Finn PW, Richards WG (2009) Ultrafast shape recognition: evaluating a new ligand-based virtual screening technology. J Mol Graph Model 27:836–845
28. Ballester PJ (2011) Ultrafast shape recognition: method and applications. Future Med Chem 3:65–78
29. Mavridis L, Hudson BD, Ritchie DW (2007) Toward high throughput 3D virtual screening using spherical harmonic surface representations. J Chem Inf Model 47:1787–1796
30. Cai WS, Shao XG, Maigret B (2002) Protein-ligand recognition using spherical harmonic molecular surfaces: towards a fast and efficient filter for large virtual throughput screening. J Mol Graph Model 20:313–328
31. Morris RJ, Najmanovich RJ, Kahraman A, Thornton JM (2005) Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. Bioinformatics 21:2347–2355
32. Jakobi AJ, Mauser H, Clark T (2008) ParaFrag–an approach for surface-based similarity comparison of molecular fragments. J Mol Model 14:547–558
33. DiMaio FP, Soni AB, Phillips GN, Shavlik JW (2009) Spherical-harmonic decomposition for molecular recognition in electron-density maps. Int J Data Min Bioinform 3:205–227
34. Huang N, Shoichet BK, Irwin JJ (2006) Benchmarking sets for molecular docking. J Med Chem 49:6789–6801
35. A directory of useful decoys. http://dud.docking.org/. Accessed 13 June 2010
36. Jahn A, Hinselmann G, Fechner N, Zell A (2009) Optimal assignment methods for ligand-based virtual screening. J Cheminform 1:14
37. Good AC, Oprea TI (2008) Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? J Comput Aided Mol Des 22:169–178
38. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50:742–754
39. Accelrys, Inc. http://accelrys.com/. Accessed 1 May 2010
40. Ritchie DW, Kemp GJL (1999) Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces. J Comput Chem 20:383–395
41. Connolly ML (1983) Solvent-accessible surfaces of proteins and nucleic acids. Science 221(4612):709–713
42. Laurikkala J (2001) Improving identification of difficult small classes by balancing class distribution. Lecture Notes Comput Sci 2101:63–66

43. Lee Y, Jeon K, Lee JT, Kim S, Kim VN (2002) MicroRNA maturation: stepwise processing and subcellular localization. EMBO J 21:4663–4670
44. Nicholls A (2008) What do we know and when do we know it? J Comput Aided Mol Des 22:239–255
45. Truchon JF, Bayly CI (2007) Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. J Chem Inf Model 47:488–508
46. Nicholls A, McGaughey GB, Sheridan RP, Good AC, Warren G, Mathieu M, Muchmore SW, Brown SP, Grant JA, Haigh JA, Nevins N, Jain AN, Kelley B (2010) Molecular shape and medicinal chemistry: a perspective. J Med Chem 53:3862–3886
47. Kahraman A, Morris RJ, Laskowski RA, Thornton JM (2007) Shape variation in protein binding pockets and their ligands. J Mol Biol 368:283–301
48. Kirchmair J, Markt P, Distinto S, Wolber G, Langer T (2008) Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection–what can we learn from earlier mistakes? J Comput Aided Mol Des 22:213–228
49. Jain AN, Nicholls A (2008) Recommendations for evaluation of computational methods. J Comput Aided Mol Des 22:133–139
50. Irwin JJ, Shoichet BK (2005) ZINC—a free database of commercially available compounds for virtual screening. J Chem Inf Model 45:177–182

J Mol Model (2012) 18:1611–1624
DOI 10.1007/s00894-011-1184-3

ORIGINAL PAPER

# Ligand-based 3-D pharmacophore generation and molecular docking of mTOR kinase inhibitors

**Karunakar Tanneeru · Lalitha Guruprasad**

Received: 3 May 2011 / Accepted: 11 July 2011 / Published online: 30 July 2011


**Abstract** The 3-D structure of the human mTOR kinase domain was modeled based on the crystal structure of PI3Kγ using comparative modeling methods, and the ATP-binding site of mTOR was characterized. The mTOR kinase 3-D model structure is similar to the structure of the PI3Kγ kinase domain, and exhibits great similarity to PI3Kγ at the active site of the kinase. Pharmacophore generation, the docking of mTOR inhibitors, and molecular dynamics (MD) simulations of mTOR–inhibitor docked complexes were carried out in this work. The best pharmacophore model generated from 27 ATP-competitive mTOR inhibitors comprised two hydrogen-bond acceptors, one aromatic ring, and one hydrophobic feature. These 27 inhibitors were docked into the ATP-binding site comprising the DFG motif, and the interactions in each protein–inhibitor complex were characterized. Mapping the pharmacophore model onto the docked inhibitors explained the specificity of the features of the pharmacophore and how they were arranged inside the active site of mTOR kinase. MD studies revealed important structural features, such as the large hydrophobic pocket "HP" and hydrophilic pocket "A1," and the solvent-exposed hydrophilic pocket "A2" at the active site of mTOR. Our results provide structural models of mTOR–inhibitor complexes and clues that should aid in the design of better mTOR kinase inhibitors.

**Keywords** mTOR kinase · Pharmacophore · Homology modeling · Molecular docking · Molecular dynamics

**Electronic supplementary material** The online version of this article (doi:10.1007/s00894-011-1184-3) contains supplementary material, which is available to authorized users.

K. Tanneeru · L. Guruprasad (✉)
School of Chemistry, University of Hyderabad,
Hyderabad 500046, India
e-mail: lgpsc@uohyd.ernet.in

## Introduction

Mammalian target of rapamycin (mTOR) is the central component of a complex signaling network initiated by nutrients, hormones and growth factors that regulate cell growth, proliferation, and animal size [1]. mTOR is a serine/threonine kinase, and is homologous to phosphatidylinositide 3-kinase (PI3K), with a sequence similarity of greater than 30%. Rapamycin is a macrocyclic lactone that binds to intracellular immunophilin FK506 binding protein 12 (FKBP12), and the resulting complex inhibits the protein kinase activity of mTOR. The functions of mTOR are dependent on its interaction with other adaptor proteins to form two signaling complexes with distinct functions [2]. The mTORC1 complex comprises of raptor (regulatory associated protein of mTOR), a 150 kDa binding partner of mTOR, and mammalian ortholog LST8/G-protein β-subunit-like protein [3]. In the mTORC1 complex, mTOR is activated by growth factors and nutrients to control protein translation and cell growth by regulating p70S6K and 4EBP1 through their respective signaling motifs, which are essential for mTOR-catalyzed phosphorylation [4, 5]. The mTORC2 complex comprises rictor (rapamycin-insensitive companion of mTOR), mLST8, and SIN1 [6]. In the mTORC2 complex, mTOR is activated by growth factors to control cell proliferation and cell survival by regulating protein kinase B through phosphorylation.

The PI3K-Akt-mTOR signaling pathway is one of the most important pathways in cell proliferation and angiogenesis, and has therefore been a rather attractive target for cancer treatment in recent years. mTOR is involved in the signaling pathways of several types of cancers, such as renal cell carcinoma, lung cancer, gastric cancer, colorectal cancer, neuroendocrine tumors, and chordoma [7, 8]. Rapamycin inhibits progression through the G1 phase of the cell cycle in

various cell types [3]. It is a clinically valuable drug that is used to block immune rejection of transplanted organs [9], as well as in clinical trials for the treatment of cancer [10, 11] and for the prevention of restenosis after angioplasty [12]. Currently, the dual mTOR/PI3K inhibitor NVP-BEZ235 from Novartis is in phase II clinical trials [13], and GDC-0941 has been reported by Genentech to be in clinical trials [14]. Blocking mTOR pathways interferes not only with cell growth and metabolism but also with angiogenesis, so rapamycin exhibits anti-cancer and anti-angiogenesis activities [15]. The inhibition of mTORC1 by rapamycin can also induce feedback activation of PI3K and ERK/MAPK [16, 17]. Hence, there is a need to develop new inhibitors of mTOR kinase activity. A series of ATP-competitive inhibitors based on morpholine-pyrimidines and triazines that are high-affinity and selective inhibitors of mTOR kinase have been discovered and optimized [18]. Another inhibitor of mTOR, Ku-0063794, has been shown to have high specificity for mTOR compared to other lipid kinases [19]. Nowak et al. and Zask et al. modeled the structure of mTOR based on the crystal structure of PI3$\gamma$ kinase in order to study the binding modes of inhibitors [20, 21]. Nowak et al. used high-throughput screening to identify mTOR inhibitors that, upon optimization, yielded selective mTOR inhibitors with high affinity compared to PI3$\alpha$ kinase. Further improvements involving replacing the morpholines with bridged morpholines in pyrazolopyrimidine inhibitors increased the selectivity towards mTOR by 26,000-fold. These studies reported the identification and activities of highly selective mTOR inhibitors.

In this work, we modeled the 3-D structure of the C-terminal region of mTOR, which contains the kinase domain, based on comparative modeling methods. ATP-competitive inhibitors of mTOR kinase were employed to generate a pharmacophore which could be used to identify the functional features responsible for binding to mTOR. These inhibitors were docked into the ATP-binding site of the kinase domain, and molecular dynamics (MD) simulations were carried out to gain a deep understanding of the protein–inhibitor interactions. We have also included previously known ATP-competitive inhibitors of mTOR kinase in the docking and MD studies. The inclusion of these independent molecules helped to validate the usefulness of the pharmacophore generated and to further our understanding of the binding modes of commercially important inhibitors, in order to gauge their interactions with the active site of the protein. To our knowledge, the generation of a pharmacophore for mTOR kinase, a study of the correlation between the arrangement of pharmacophore features and the orientations of docked inhibitor conformations at the ATP-binding site, and MD simulations of the protein–inhibitor complex have not been performed previously. Our results will aid the rational design of potential ATP-competitive mTOR inhibitors in future studies.

## Materials and methods

### Structure modeling and protein preparation

Comparative protein modeling is a reliable tool for building the 3-D structures of a protein based on the experimentally determined structure of another homologous protein. The human mTOR protein sequence (NCBI accession ID: P42345) was taken from the NCBI database (http://www.ncbi.nlm.nih.gov/). The C-terminal region of mTOR (residues 1330–2549) is homologous to PI3K$\gamma$ (NCBI accession ID: NP_002640.2). The kinase domain region of mTOR (2145–2446) was modeled based on the crystal structure of PI3$\gamma$ kinase bound to a pyrazolopyrimidine inhibitor (PDB_ID: 3IBE) as a template [21]. The template and mTOR sequences were aligned using CLUSTALW [22]. The protein's 3-D structure was modeled using Accelrys Discovery Studio 2.1 (DS 2.1; Accelrys Software Inc., San Diego, CA, USA), which implemented MODELLER [23]. The pyrazolopyrimidine inhibitor structure from PI3$\gamma$ kinase was also extracted and transferred into the modeled mTOR structure, in order to use it as a guide for the docking studies. We used the WHATIF online server (http://swift.cmbi.ru.nl/servers/html/index.html) to remove any atom clashes (bumps) in the protein model. To remove any further bad contacts in the structure, we refined the protein using steepest descent energy minimization and 100 ps position restraint dynamics using Gromacs 4.0.5 [24, 25].

### Pharmacophore generation

The ATP-competitive inhibitors of mTOR were taken from the literature [18–20]. These molecules were built, hydrogens were added, and the structures were minimized in the Minimization module using the steepest descent and conjugate gradient methods in Discovery Studio 2.1. We used the 27 inhibitors with high activities listed in Table 1 to build the pharmacophore model. The Hip-Hop pharmacophore model was built using the Common Feature Pharmacophore Generation module in Accelrys Discovery Studio 2.1. This pharmacophore model provides a geometrical representation of the features required for molecules to show biological activity and therefore interact favorably at a particular receptor site that possesses a geometry complementary to it. Hip-Hop identifies configurations or 3-D spatial arrangements of chemical features that are common to all molecules in the set. We generated ten pharmacophores using the Hip-Hop method. Under the Common Feature Pharmacophore Generation protocol, we used the features Hydrogen bond acceptor, Hydrogen bond donor, Ring aromatic, and Hydrophobic to generate the pharmacophore. In the pharmacophore generation protocol, we set the constraints such that there were a minimum of 4

**Table 1** List of 27 ATP-competitive inhibitors of mTOR kinase that were used to generate a pharmacophore and for molecular docking. The reported $pIC_{50}$ values from the literature and the GOLD docking scores are provided. The $pIC_{50}$ values were obtained as –log $IC_{50}$ (with the $IC_{50}$ values in nM)

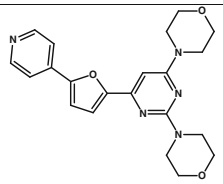
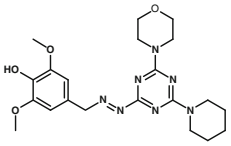
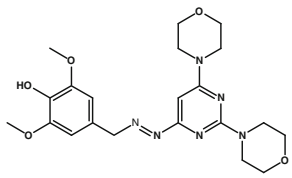
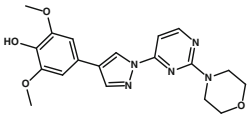
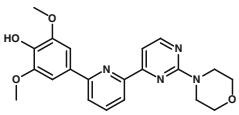
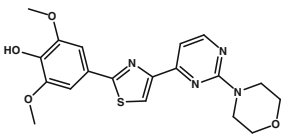
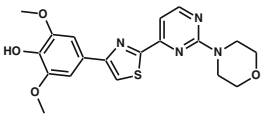
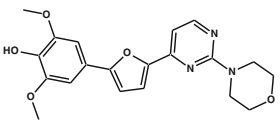
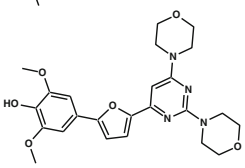
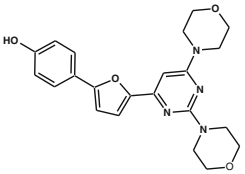
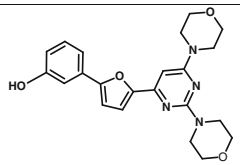
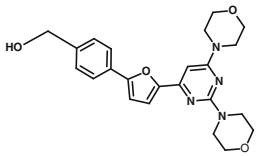
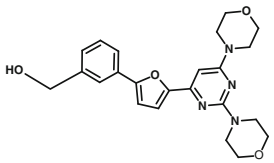
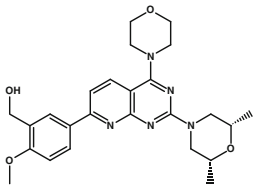
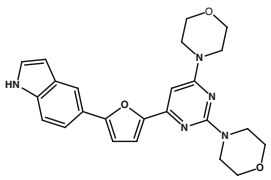
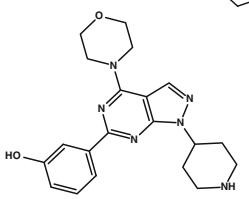
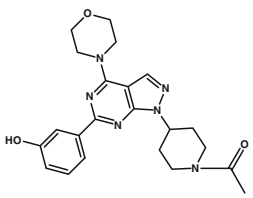
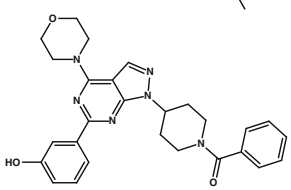
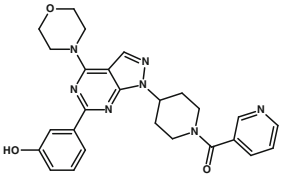
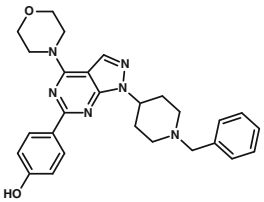
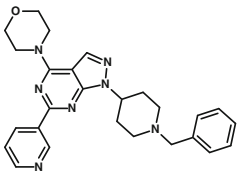
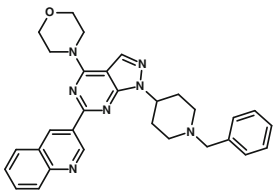
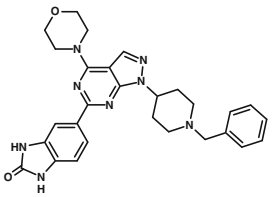
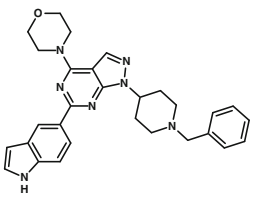
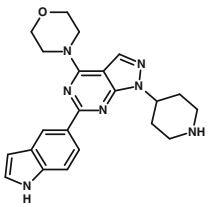
| S. No | Molecule | GOLD Score | Inhibition (pIC50) |
|---|---|---|---|
| 1 |  | 51.17 | 6.28 |
| 2 |  | 51.50 | 6.12 |
| 3 |  | 51.39 | 6.18 |
| 4 |  | 44.53 | 5 |
| 5 |  | 49.48 | 6.26 |
| 6 |  | 50.39 | 6.25 |
| 7 |  | 52.61 | 6.74 |
| 8 |  | 52.55 | 7.28 |
| 9 |  | 51.43 | 7.64 |
| 10 |  | 51.46 | 6.26 |

**Table 1** (continued)

| S. No | Molecule | GOLD Score | Inhibition (pIC50) |
|---|---|---|---|
| 11 |  | 46.74 | 5.82 |
| 12 |  | 46.08 | 5.26 |
| 13 |  | 51.06 | 6.29 |
| 14 |  | 48.55 | 7.52 |
| 15 |  | 42.63 | 5 |
| 16 |  | 53.90 | 7.44 |
| 17 |  | 53.82 | 7.31 |
| 18 |  | 57.93 | 8.4 |

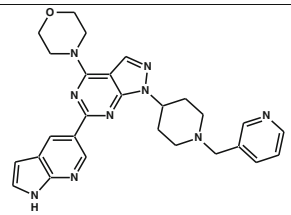
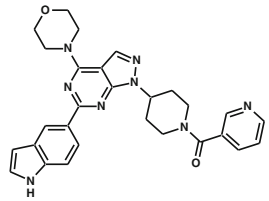
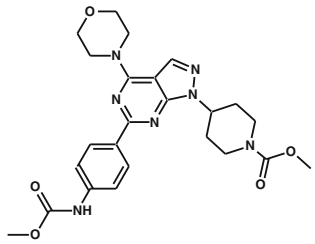
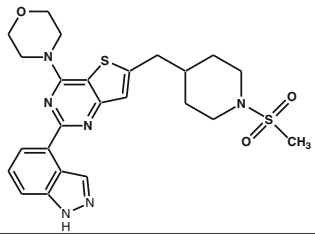
and a maximum of 10 features, and a minimum interphase distance of 2.97 Å between any two features in the pharmacophore. For the molecules that produced strong inhibition ($IC_{50}$ < 1 μM), we set the maximum omit features to zero and the principal value to 2. For the molecules with $IC_{50}$ > 1 μM, the principal value was set

**Table 1** (continued)

| S. No | Molecule | GOLD Score | Inhibition (pIC50) |
|---|---|---|---|
| **19** |  | **56.46** | **8.22** |
| **20** |  | **52.68** | **6.77** |
| **21** |  | **51.11** | **6.16** |
| **22** |  | **52.14** | **6.92** |
| **23** |  | **53.27** | **6.71** |
| **24** |  | **51.96** | **7.06** |
| **25** |  | **52.88** | **6.77** |

to 1. The protocol generated the maximum possible number of conformations for each molecule using the FAST method before generating a pharmacophore. A maximum limit of 255 conformations was set for each molecule, and an energy threshold of 20.0 kcal $mol^{-1}$ was employed.

**Table 1** (continued)

| S. No | Molecule | GOLD Score | Inhibition (pIC50) |
|---|---|---|---|
| **26** |  | **54.06** | **8.05** |
| **27** |  | **58.13** | **8.40** |
| **WYE - 354** |  | **57.84** | **8.40** |
| **GDC-0941** |  | **51.82** | **6.27** |

## Inhibitor docking

Molecular docking was used to determine the orientations of the inhibitors when binding to the protein, in order to predict their affinities and activities when bound to the protein. A cavity corresponding to 15 Å distance cut-off was defined around the ATP-binding site in the mTOR model structure. We performed molecular docking of all inhibitors using GOLD 5.0 (genetic optimization of ligand docking) software [26, 27]. Molecules GDC-0941 [14] and WYE-354 [28], which were known to be mTOR inhibitors, and are currently engaged in clinical trials, were also used for GOLD docking. The docking studies identified the inhibitor orientation in the model structure of the mTOR active site, and the results from GOLD docking were used for MD simulations. GOLD is a genetic algorithm (GA) for docking flexible ligands into protein-binding sites. For each of the ten independent GA runs, a maximum number of 100,000 GA operations were performed on a set of five groups with a population size of 100 individuals. Operator weights for crossover, mutation, and migration were set to 95, 95, and 10, respectively. Default distance annealing cut-off values of 2.5 Å (dH–X) for hydrogen bonds and 4.0 Å for van der Waals interactions were employed. During docking, the default algorithm speed was selected, and the ligand binding site was defined as being within 15 Å of the centroid of the pyrazolopyrimidine inhibitor at the active site. The number of poses for each inhibitor was set to 20, and early termination was allowed if the top five bound conformations of a ligand were within a RMSD of 1.5 Å. After docking, the individual binding poses of each ligand were re-ranked according to the GOLD score. The top-ranked conformation of each ligand was selected and analyzed using HERMES and Discovery Studio 2.1 to understand the mode of protein–inhibitor binding.

### Molecular dynamics

The molecular dynamics of the protein–inhibitor complex provides insights into the flexibility associated with changes in conformation, and therefore provides a molecular basis for inhibition. The best poses of molecule 18 and WYE-354 from Table 1 docked into the mTOR structure were used to perform MD simulations in Gromacs 4.0.5. We employed the GROMOS96 43a1 force field for the protein in all our simulations. The ligand pdb coordinates and topology files were generated using the Dundee PRODRG server (http://davapc1.bioch.dundee.ac.uk/prodrg/) using the GROMOS96.1 force field [29]. The coordinates of the molecules obtained from the GOLD docking were used to generate the topology from the Dundee PRODRG server. The PRODRG server has been successfully used to generate force fields for various ligands and inhibitors [30, 31]. During the topology generation, stereochemical changes and energy minimization of the molecule were not permitted. The protein was solvated in a cubic box with edges 1.5 nm in length using the explicit solvent–simple point charge model (SPC216 water molecules), which generated the water box for the mTOR structure. During MD simulations, we performed 2500 steps of steepest descent minimization and position-restrained dynamics to distribute water molecules throughout the system for 100 ps. Later, we performed MD simulations of the whole system for 15,000 ps, using 0.002 ps time steps. The long-range electrostatics were handled using the PME method, and the van der Waals interactions were computed within a cut-off distance of 1.4 nm. MD simulations were carried out at 300 K using the Nose–Hoover method [32, 33], and we used the Parrinello–Rahman method [34, 35] for pressure. The average structure obtained from the last 5000 ps of the MD simulation was refined by performing 1000 steps of steepest descent energy minimization. For active site analysis after the MD studies, the inhibitor in the energy-minimized average structure of the mTOR complex was superimposed on the initial conformation of the inhibitor mapped with the pharmacophore using Accelrys Discovery Studio 2.1.

## Results and discussion

### Homology modeling

Pair-wise sequence alignment between the kinase domains of mTOR and PI3γ kinase shown in Fig. 1a of the "Electronic supplementary material" (ESM) indicates an overall homology of 44% and a few insertion and deletion regions. The 3-D structures of the mTOR and PI3γ kinase domains are very similar, with a RMSD of 0.85 Å (Fig. 1b of the ESM), and their secondary structural elements are highly superimposable (Figs. 1a and 1b of the ESM). This is a measure of the structural conservation of these lipid family kinases. The mTOR kinase model structure was validated using PROCHECK [36], and it was observed that only two residues were in the generously allowed region of the Ramachandran plot [37] (see Fig. 1c of the ESM). This indicated that the mTOR model constructed possesses good geometric and stereochemical quality. The signature sequence motifs characteristic of kinases are present in both proteins. The hinge region GWVPHCDTL (2238–2246 in mTOR) connecting the N- and C-terminal domains is observed in both proteins. The gatekeeper residue Ile-2237 is observed close to the hinge region. The sequence motif ITSKQR (2163–2168) can be identified at the ATP binding site, and comprises the P loop. The catalytic loop YILGLGDRHPSN (2332–2343) observed in mTOR displays only minor sequence variations from PI3γ kinase. The highly conserved DGF motif is followed by the activation loop. The activation loops in mTOR kinase (2360–2379) and PI3γ kinase (966–987) show reduced sequence similarity. The activation loop that is missing from the PI3γ kinase crystal structure (PDB_ID: 3IBE) was modeled in the mTOR. The DFG loop exists in two conformations, DFG-in and DFG-out, in 3-D structures that determine whether the kinase is in the active or inactive state, respectively. The DFG (2357–2359) motif is characteristic of kinases, and we observed that the Asp-2357 side-chain is oriented into the active site while the side-chain of Phe-2358 is pointing outwards, revealing that the mTOR kinase is in the active conformation, similar to the conformation observed in PI3γ kinase (PDB_ID: 3IBE). The residues that line the ATP-binding pocket in the PI3γ kinase (Met-804, Ser-806, Lys-833, Asp-836, Tyr-867, Ile-881, Val-882, Thr-887, Asn-951, Met-953, Ile-963 and Asp-964) are highly conserved in mTOR kinase (Ile-2163, Ser-2165, Lys-2187, Glu-2190, Tyr-2225, Trp-2239, Val-2240, Thr-2245, Asn-2343, Met-2345, Ile-2356, and Asp-2357).

### Pharmacophore generation

The ten pharmacophores generated using the Hip-Hop method had four features, as indicated in Table 2. The pharmacophores ranked first and second according to the Hip-Hop score have three hydrogen-bond acceptors and one aromatic ring feature, and the pharmacophore models ranked third and fourth have identical scores and each has two hydrogen-bond acceptors, one aromatic ring, and one hydrophobic feature that occupies a 3-D geometry in space. The ranking is a measure of how efficiently the molecules map onto the proposed pharmacophore. For each molecule in the training set, the extent of overlap with the pharmacophore is indicated by a fit value. The pharmaco-

**Table 2** The rank scores and features present in ten pharmacophores generated using HIP-HOP

| S. no | Hypo features | Rank | Direct hit mask (DH) | Partial hit mask (PH) |
|---|---|---|---|---|
| 1 | RAAA | 286.47 | 11111111111111111111111111 | 00000000000000000000000000 |
| 2 | RAAA | 283.98 | 11111111111111111111111111 | 00000000000000000000000000 |
| 3 | RHPAA | 278.25 | 11111111111111111111111111 | 00000000000000000000000000 |
| 4 | RHPAA | 278.25 | 11111111111111111111111111 | 00000000000000000000000000 |
| 5 | RHPAA | 276.49 | 11111111111111111111111111 | 00000000000000000000000000 |
| 6 | RHPAA | 276.38 | 11111111111111111111111111 | 00000000000000000000000000 |
| 7 | RHPAA | 276.38 | 11111111111111111111111111 | 00000000000000000000000000 |
| 8 | RAAA | 275.61 | 11111111111111111111111111 | 00000000000000000000000000 |
| 9 | RAAA | 274.51 | 11111111111111111111111111 | 00000000000000000000000000 |
| 10 | RHPAA | 274.42 | 11111111111111111111111111 | 00000000000000000000000000 |

*R* aromatic ring, *A* hydrogen-bond acceptor, *H* hydrophobic. *Direct hit*: all the features are mapped (direct hit = 1 means "yes"). *Partial hit*: partial mapping of the hypothesis (partial hit = 0 means "no")

phores ranked first and second map to 11 and 3 molecules, respectively, in the training set with fit values of greater than 2.5, while the pharmacophores ranked third and fourth map to 19 and 13 molecules, respectively, in the training set with fit values of greater than 2.5. Also, it has been observed that kinases have a hydrophobic cavity in the hinge region, and we assumed that a hydrophobic pharmacophore feature is a requirement for the binding of ATP-competitive inhibitors. Therefore, based on the presence of a hydrophobic feature, and mapping with a large number of molecules with high fit values, we chose the third pharmacophore as the best pharmacophore model for the mTOR kinase inhibitors. Furthermore, when this pharmacophore was positioned into the active site of the mTOR model in the orientation of the docked molecule's conformation, we observed that its features correlate well with the nature of the active site residues. The hydrogen-bond acceptors were in the hydrophilic regions and the hydrophobic feature was positioned in the hydrophobic pocket. This observation further confirms that this is the best pharmacophore and that it is well correlated with the active site of mTOR kinase.

The best pharmacophore has two hydrogen-bond acceptors (A1 and A2), one aromatic ring (R), and one hydrophobic group (HP), as shown in Fig. 1. The distances between the four features are shown in Fig. 1. The two hydrogen-bond acceptors maintain a distance of ~10 Å from each other (A1...A2, ~10 Å), the distance between A2 and R is approximately 8 Å (A2…R, 8 Å), the distance between R and HP is ~4–5 Å (R…HP, 4–5 Å), A2 and HP are separated by 12 Å (A2…HP, 12 Å), and the distance between A1 and HP is 8–9 Å (A1...HP, 8–9 Å). We set the minimum distance between any two features during pharmacophore generation to be 2.97 Å, and found that the distance between the features exceeds the set limit.

## Molecular docking

We observed that almost all molecules had the same docking position in the ATP-binding site of the mTOR kinase. When inhibitors were mapped to the features of the pharmacophore and compared to the orientations of the docked inhibitors, it was clear that the pharmacophore features were satisfied by the nature of amino acid residues comprising the active site. Below, based on the GOLD docking results, we discuss the best docking poses of some of the inhibitors. We explain the interaction efficiency based on the orientation of the inhibitor and the hydrogen bonds between the inhibitor and its binding site in the mTOR kinase domain. In all cases, the best docking conformation was compatible with the features of the pharmacophore model for the active site of the mTOR. We found a reasonable correlation between the reported $IC_{50}$ values and the GOLD scores indicated in Table 1. The mTOR–inhibitor complexes were further used for MD simulations.

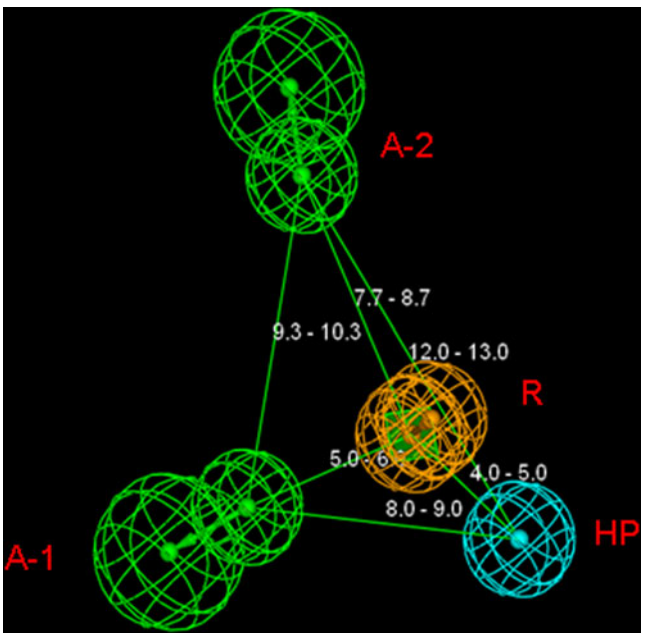



**Fig. 1** Pharmacophore generated from 27 ATP competitive inhibitors of mTOR using Hip-Hop. *A* hydrogen bond acceptor (*green*), *R* aromatic ring (*orange*), *HP* hydrophobic group (*blue*)

We discuss MD simulations of complexes of mTOR with molecule 18 and WYE-354, as shown in Table 1, which were chosen because of their good inhibitory activities.

*Molecule 3* The docking of molecule 3 into the mTOR active site is shown in Fig. 2a. The morpholine ring is docked into the pharmacophore's hydrophobic pocket (HP) formed by residues Met-2345, Trp-2239, Cys-2243, Leu-2354 and Val-2240. The main-chain NH of Val-2240 forms a hydrogen bond with O16 of the inhibitor (NH...O16, 2.9 Å). This hydrogen bond may stabilize the orientation of the inhibitor, and the morpholine ring is suitably positioned by the hydrophobic residues in the pocket. The side chains of Ile-2356 and Leu-2185 share space with the aromatic ring of the inhibitor on either side. The other morpholine group of the inhibitor is positioned in the pharmacophore's hydrophilic region A1. This region comprises the hydrophilic residues Tyr-2225, Lys-2187, Ile-2237, Gln-2167, Asp-2195 and Asp-2357. The DFG loop of the active form of mTOR kinase was observed near A1. The dimethoxyphenol group of the inhibitor is located in the hydrophilic pocket A2 that comprises the hydrophilic residues Ser-2165, Lys-2171, Ile-2163, Asn-2343 and Ser-2342. The side-chain hydroxyl group of Ser-2342 forms hydrogen bond with the methoxy oxygen of the inhibitor (OH…O28, 3.3 Å). The side-chain hydroxyl group of Ser-2165 forms a bifurcated hydrogen bond with the methoxy oxygen and the hydroxyl group of the inhibitor (OH…O29, 2.6 Å; OH…O28, 2.9 Å).

*Molecule 13* The docking of molecule 13 into the mTOR active site is shown in Fig. 2b. The morpholine ring is docked into the pharmacophore's hydrophobic pocket (HP) formed by residues Met-2345, Trp-2239, Cys-2243, Leu-2354 and Val-2240. The main-chain NH of Val-2240 forms a hydrogen bond with O12 of inhibitor (NH….O12, 2.6 Å). The other morpholine group of the inhibitor is positioned in the pharmacophore's hydrophilic region A1. This region comprises the hydrophilic residues Tyr-2225, Lys-2187, Ile-2237, Gln-2167, Asp-2195, and Asp-2357. The benzyl alcohol group of the inhibitor is located the hydrophilic pocket A2, formed by the residues Lys-2171, Gln-2161, Ile-2163, and Thr-2245. The side-chain NH of Lys-2171 is involved in a hydrogen bond with HO31 of the inhibitor (NH…O31, 3.3 Å).

*Molecule 18* The docking of molecule 18 into the mTOR active site is shown in Fig. 2c. The morpholine ring is docked into the pharmacophore's hydrophobic pocket (HP)

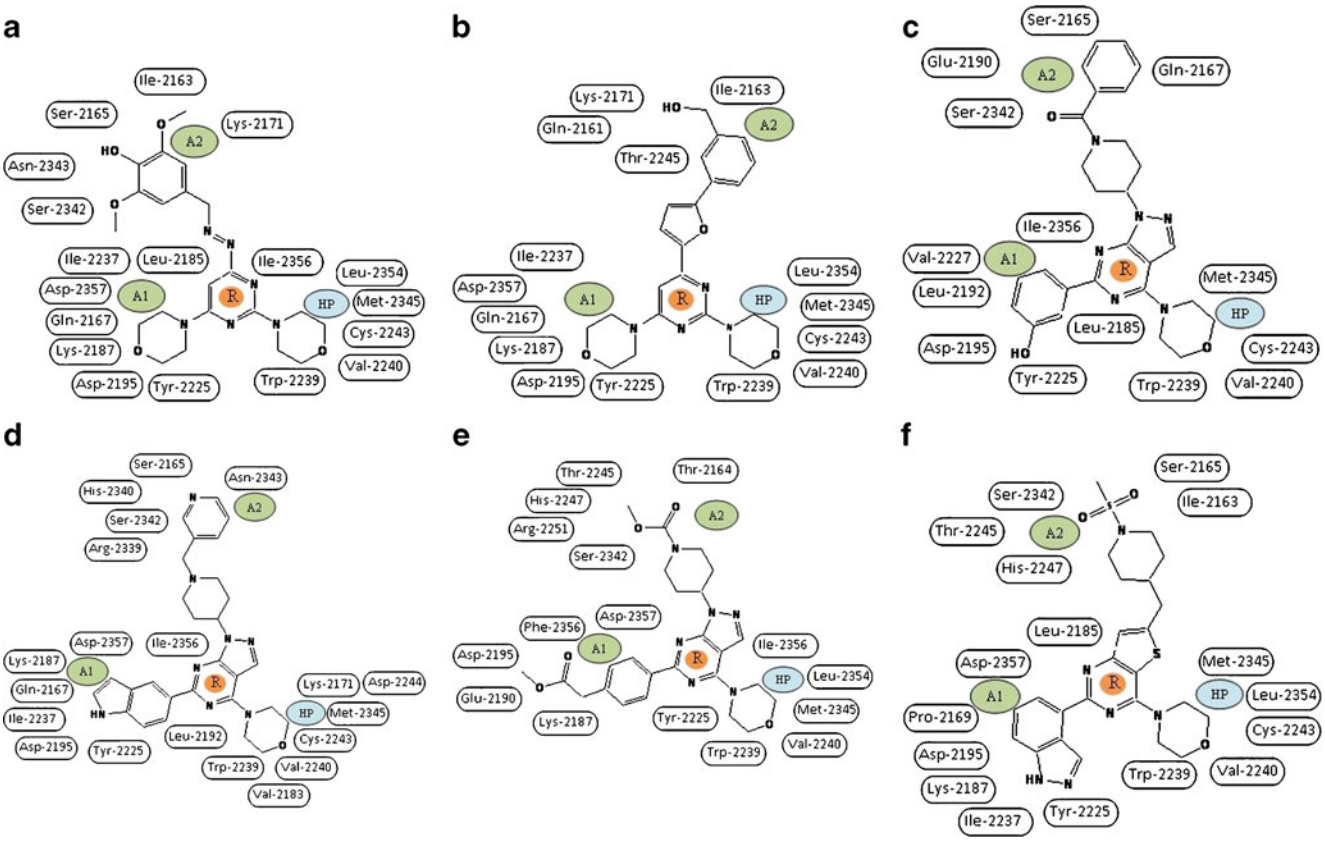



**Fig. 2a–f** GOLD docking orientations of inhibitors at the mTOR active site. **a** Molecule 3, **b** molecule 13, **c** molecule 18, **d** molecule 26, **e** WYE-354, and **f** GDC-0941

formed by residues Met-2345, Leu-2185, Trp-2239, Cys-2243, and Val-2240. The main-chain NH of Val-2240 forms a hydrogen bond with O7 of the inhibitor (NH….O7, 2.6 Å). The side-chain of Ile-2356 shares space with the aromatic ring of the inhibitor. The hydroxyl group of *m*-hydroxy phenyl is located in the pharmacophore's hydrophilic region A1. This region comprises the residues Asp-2195, Tyr-2225, Leu-2192, Val-2227, Ile2356, and Ile-2237. The side-chain carbonyl group of Asp-2195 is involved in a hydrogen bond with O23H of the inhibitor (CO…HO23, 2.1 Å). The benzoyl group carbonyl oxygen fits into the pharmacophore's hydrophilic pocket A2, formed by the hydrophilic residues Ser-2165, Ser-2342, Glu-2190, Lys-2166, and Gln-2167.

*Molecule 26* The best-docked conformation of molecule 26 is shown in Fig. 2d. The hydrophobic group morpholine of the inhibitor is docked into the pharmacophore's hydrophobic pocket (HP), surrounded by Met-2345, Val-2183, Trp-2239, Asp-2244, Lys-2171, Cys-2243, and Val-2240. The main-chain NH of Val-2240 forms a hydrogen bond with the morpholine oxygen (NH…O13, 3.3 Å). The residue Ile-2356 also shares the pocket of the aromatic ring group (R) of the pharmacophore, and interacts with the pyrazolopyrimidine ring. The indole group of the inhibitor is docked into the pharmacophore's hydrophilic pocket A1, created by the Asp-2357, Tyr-2225, Ile-2237, Gln-2167, Leu-2192, Lys-2187, and Asp-2195 residues in mTOR. The side-chain carbonyl oxygen of Asp-2195 forms a hydrogen bond with the NH of the indole (CO…HN37, 2.41 Å). The pyridine group fits into the pharmacophore's hydrophilic pocket A2, comprising the His-2340, Ser-2165, Arg-2339, Ser-2342, and Asn-2343 residues.

*WYE-354* The best-docked conformation of WYE-354 is shown in Fig. 2e. The morpholine ring of the inhibitor is docked into the pharmacophore's hydrophobic pocket (HP), which is formed by the residues Val-2240, Trp-2239, Tyr-2225, Met-2345, Leu-2354, and Ile-2356. Ile-2356 interacts with the ring aromatic group (R) of the pharmacophore. The main-chain NH of Val-2240 forms a hydrogen bond with O24 of the morpholine ring (NH…O24, 3.2 Å). The N-substituted methyl ester of aniline is positioned in the pharmacophore's hydrophilic pocket A1, formed by Lys-2187, Asp-2195, Asp-2357, Glu-2190, and Phe-2356. The main-chain NH of Phe-2356 forms a hydrogen bond with the carbonyl oxygen of the inhibitor (NH…O15, 3.0 Å). The piperidine-substituted ester is located in the pharmacophore's hydrophilic pocket A2, formed by His-2247, Arg-2251, Ser-2342, Thr-2245, and Thr-2164.

*GDC-0941* The best-docked conformation of GDC-0941 is shown in Fig. 2f. The hydrophobic morpholine group of the inhibitor is docked into the pharmacophore's hydrophobic

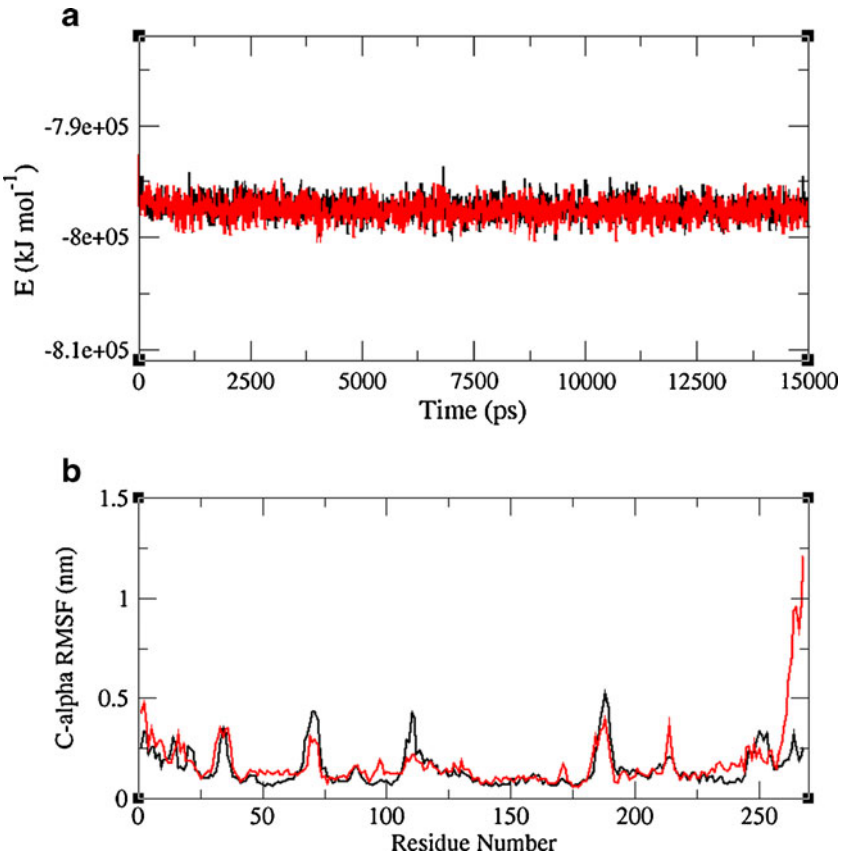



**Fig. 3a–b** The trajectory of the mTOR–inhibitor complex, plotted as a function of simulation time. **a** The potential energy, and **b** the RMSF values of the mTOR–inhibitor complexes. Molecule 18 is shown in *black*, and WYE-354 in *red*

pocket (HP), surrounded by Leu-2354, Met-2345, Trp-2239, Cys-2243, and Val-2240. The main-chain NH of Val-2240 forms a hydrogen bond with the morpholine oxygen (NH…O1, 2.5 Å). The benzopyrazole group of the inhibitor is docked into the pharmacophore's hydrophilic pocket A1, created by Asp-2357, Tyr-2225, Ile-2237, Pro-2169, Lys-2187, and Asp-2195. The side-chain carbonyl oxygen of Asp-2195 forms a hydrogen bond with the NH of the benzopyrazole (CO…HN30, 2.48 Å). The main-chain NH of Asp-2357 forms a hydrogen bond with N33 of the benzopyrazole of the inhibitor (NH…N33, 2.1 Å). The N-methanesulfonyl-substituted piperidine group fits into the pharmacophore's hydrophilic pocket A2, comprising Ser-2165, Ser-2342, His-2247, Thr-2245, and Ile-2163. The side-chain hydroxyl group of Ser-2342 is involved in a hydrogen bond with the methanesulfonyl oxygen of the inhibitor (S=O…O20, 2.1 Å). The side-chain hydroxyl group of Ser-2165 also forms a hydrogen bond with the methanesulfonyl oxygen of the inhibitor (S=O…O18, 2.9 Å).

## MD simulations

During the MD simulations, low RMSD values of ligands and minor variations in the active site residues of the mTOR kinase domain were observed. The energy-minimized average structure of the protein–inhibitor complex obtained from MD simulations was analyzed to estimate the movement of the inhibitor at the active site. The interactions were analyzed through comparison with the docking results. The results of the MD simulations of two protein–inhibitor (molecule 18 and WYE-354) complexes are described below.

The potential energy plots of the complexes during the MD simulations were obtained from the trajectory files for 15,000 ps, and are shown in Fig. 3a. The potential energy stabilized during the last 5000 ps of the MD (at approximately $-7.977\times 10^{-5}$ kJ mol$^{-1}$), indicating that the protein–inhibitor complex had stabilized. Variations in the protein were closely monitored in the trajectory files during the MD simulations, and the deviations of the Cα atoms are shown in Fig. 3b. These figures indicate that the protein deviated in the presence of each inhibitor during the MD simulations; more particularly, similar positions in the protein were affected during MD simulations with different inhibitors. From these plots, we can see that the residues in the inhibitor binding pocket were not in the highly variable regions (deviations were less than 0.15 nm), indicating that interactions between the inhibitor and most of the active site residues were well stabilized.

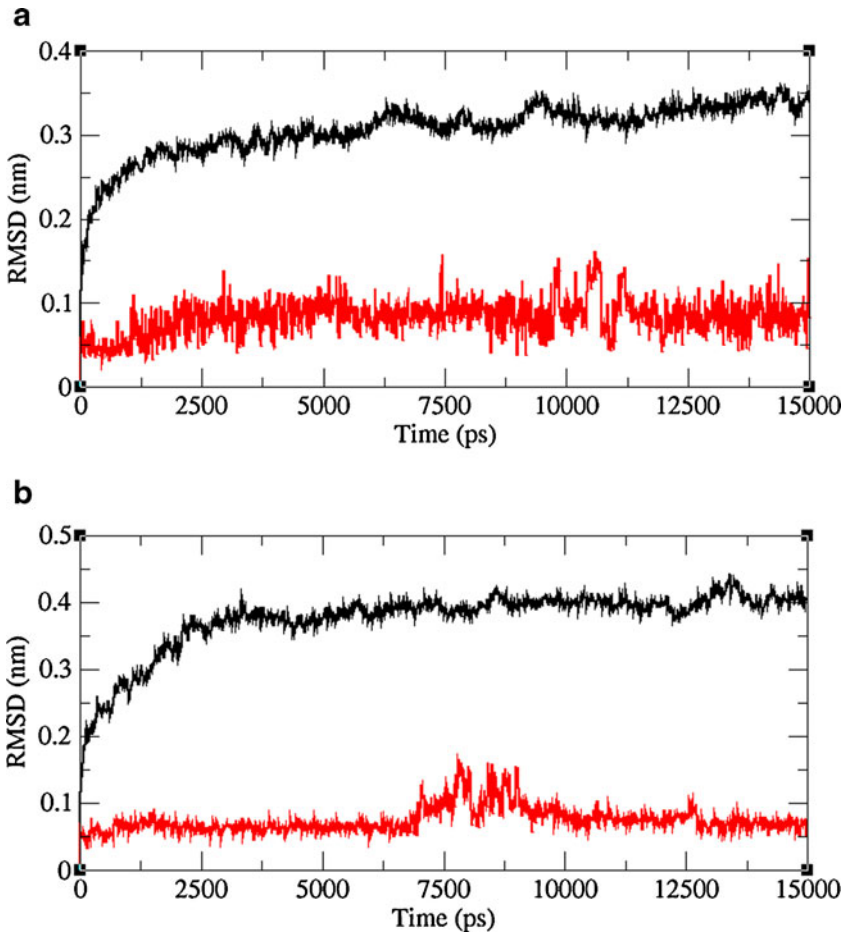



**Fig. 4a–b** The RMSD values of the inhibitors and the mTOR protein during molecular dynamics simulations. **a** Molecule 18, and **b** WYE-354. The inhibitor is presented in *red* and mTOR is presented in *black*

The RMSD plots mainly explain the details of the deviations of the inhibitor and protein during the MD simulations. The RMSD plots for molecule 18 and WYE-354 docked into the protein, as drawn based on the MD simulation trajectory files, are shown in Fig. 4a and b, respectively. These plots indicate that when docked into the protein, the two inhibitors deviated by less than 0.15 nm RMSD, and the RMSD of the protein also reached convergence at around 0.35–0.45 nm. During the 15,000 ps MD simulations, the inhibitors moved with the protein and showed several interactions with the active site. The pose and interactions of each inhibitor in the energy-minimized average structure of the protein–inhibitor complex explains its position in the active site.

*Molecule 18* All the hydrophobic residues covering the morpholine ring of the inhibitor retained their positions except Leu-2185. During the MD simulations, Val-2183 and Gly-2238 were newly introduced, thus increasing the size of the hydrophobic pocket. The hydrophilic pocket A1 surrounding the *m*-hydroxyphenyl group at the active site is displaced slightly compared to the docking orientation. The residues Tyr-2225 and Ile-2356 retained their positions, while the remaining residues moved away. Leu-2354 and His-2355 moved into the pocket. The main-chain carbonyl oxygen of Leu-2354 forms a hydrogen bond with the hydroxyl group of the inhibitor in the A1 pocket (C=O…O23, 3.0 Å). During dynamics simulations, Ser-2165 and Ser-2342 are retained in the A2 pocket. The residues Asn-2343 and Ile-2163 move into the pocket and are closer to the carbonyl group substitution.

*WYE-354* All residues except for Gly-2238 are retained in the hydrophobic pocket surrounding morpholine. In the hydrophilic pocket A1, the amino acid residues Asp-2195, Lys-2187 and Asp-2357 around the N-substituted ester group of the inhibitor retained their positions. The amino acid residues Tyr-2225, Cys-2327 and Ile-2356 are now found in the hydrophilic pocket, and they form hydrophobic interactions with the methyl group of the ester and the phenyl group. The side-chain NH of Lys-2187 forms a hydrogen bond with the carbonyl oxygen of the inhibitor (NH…O15, 3.0 Å). The side-chain hydroxyl OH of Tyr-2225 forms a hydrogen bond with N13 of the inhibitor (OH…N13, 3.15 Å). In the hydrophilic pocket A2, Thr-2164 retained its position in the active site, and Ser-2165, Val-2162 and Pro-2169 were newly introduced into the active site during the MD simulations.

The best-docked inhibitors, molecule 18 and WYE-354, were mapped onto the best pharmacophore model and superimposed onto the energy-minimized average structures of the inhibitors in complex with mTOR, as shown in Fig. 5a and b, respectively. These figures indicate that in the active site of the protein, the inhibitors show close fits to the features of the pharmacophore, signifying good correlation between the inhibitors, the pharmacophore model, and the active site of mTOR. This shows that there is good agreement between the ligand-based 3-D pharmacophore model of mTOR inhibitors and the receptor-based active site for inhibitor binding.

From the MD studies of the two mTOR–inhibitor complexes, we can see that the pharmacophore-mapped molecules fit into the active site well during the dynamics.

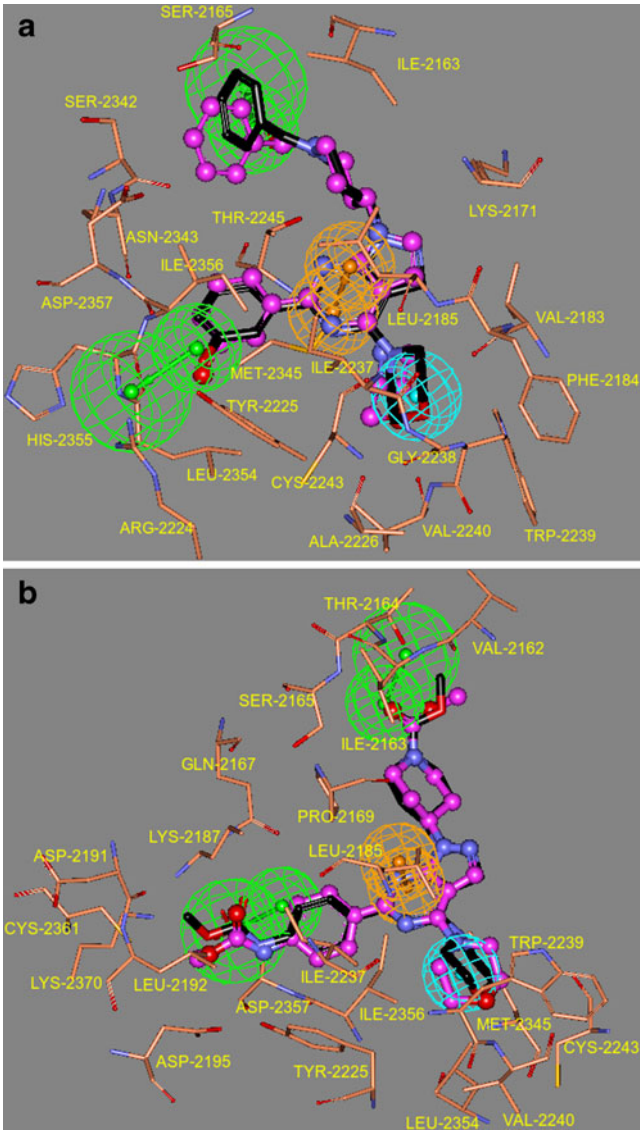



**Fig. 5a–b** Representation of the alignment of each inhibitor mapped onto the best pharmacophore, and the energy-minimized average structure of each mTOR–inhibitor complex. **a** Molecule 18, and **b** WYE-354. The inhibitor after molecular dynamics simulation is shown in *pink*. The inhibitor mapped to the best pharmacophore is colored as per element type. The pharmacophore features are depicted as follows: *green* hydrogen-bond acceptor, *orange* aromatic ring, *blue* hydrophobic group. The protein is colored according to element type

Although the active site of mTOR kinase after MD is very similar to the active site identified from the docking studies, minor differences in the protein's active site and the inhibitor's location and conformation were observed. The morpholine (molecule 18 and WYE-354) ring binds to the hydrophobic pocket, and we observed that the hydrophobicity and size of this pocket increased, comprising the amino acids Cys-2243, Val-2240, Trp-2239, and Met-2345. Zask et al. reported improved activity when the morpholine ring in the hydrophobic pocket is replaced with the bridged morpholine in pyrazolopyrimidine inhibitors, thus supporting our observations. The residue Ile-2356 is shared by the hydrophobic pocket and the aromatic ring region of the inhibitor. In the hydrophilic pocket A1, we observed that the residues Asp-2357 of the DFG motif and Asp-2195 form hydrogen-bond interactions with the inhibitor. We propose that the *m*-hydroxyphenyl group (molecule 18) at this subsite could be replaced with longer hydrophilic groups that facilitate enhanced interactions with mTOR, as observed in the case of WYE-354. In the hydrophilic pocket A2, hydrogen-bond interactions were observed between the inhibitor and the side-chain OH groups of Ser-2165 and Ser-2342, and the other residues also contribute to the stability of this pocket. Part of the inhibitor docked into this pocket is also partially exposed to solvent. Thus, we propose that hydrophilic substitutions on the inhibitor at this position would lead to greater activity.

Despite the great similarity between the active site regions of the PI3K$\gamma$ and mTOR structures, differences in their subsite residues were observed. Trp-2239, Gly-2238, and Cys-2243 are present in the HP pocket of mTOR, while that of PI3K$\gamma$ contains Ile-881, Glu-880, and Ala-885. These amino acid residue substitutions are responsible for the relatively dense and compact hydrophobic pocket in mTOR compared to PI3K$\gamma$. Thr-2164, Ser-2342, His-2247, and Arg-2251 are present in the A2 pocket of mTOR, while that of PI3K$\gamma$ comprises Ala-805, Asp-950, Ala-889 and Gln-893. These substitutions are responsible for the differences in the specific nonbonding interactions associated with these protein–inhibitor complexes. The results obtained from docking and dynamics simulations explain the best orientation of the inhibitor at the well-defined active site of mTOR, and can be applied to the design of better mTOR inhibitors.

## Conclusions

The 3-D model structure corresponding to the kinase domain of mTOR is similar to the crystal structure of PI3K$\gamma$. The best pharmacophore model generated from 27 ATP-competitive inhibitors comprises two hydrogen-bond acceptors, one hydrophobic feature and one aromatic ring feature. Docking these inhibitors (including the WYE-354 and GDC-0941 mTOR inhibitors that are currently in clinical trials) into the ATP-binding site of the mTOR kinase indicated that all of the inhibitors bind the protein in the same pocket and interact via nonbonding interactions. A reasonable correlation between the reported $IC_{50}$ values and the GOLD docking scores of the inhibitors was observed. MD simulations of two mTOR–inhibitor (molecule 18 and WYE-354) complexes in water confirmed that these complexes exhibit relatively stable interactions between the active site residues and inhibitors, and display only minor displacements in the active site. We observed that the pharmacophore features map well to the nature of the amino acid residues in the active site of mTOR kinase. The subtle differences in mTOR and PI3K$\gamma$ active site residues reported here may be responsible for mTOR inhibitor specificity. Our analysis will further aid the rational design of potent mTOR inhibitors.

**Acknowledgments** The authors thank the Department of Science and Technology (DST), India, for financial support (no. SR/S5/MBD-05/2007). The authors thank the Centre for Modelling Simulation and Design (CMSD), University of Hyderabad, Hyderabad, and the Centre for Development of Advanced Computing (CDAC), Pune, India, for providing access to their computational facilities. KT thanks the Council of Scientific and Industrial Research (CSIR), New Delhi, for research fellowship.

## References

1. Sarbassov DD, Ali SM, Sabatini DM (2005) Curr Opin Cell Biol 17:596–603
2. Jacinto E, Loewith R, Schmidt A, Lin S, Rüegg MA, Hall A, Hall MN (2004) Nat Cell Biol 6:1122–1128
3. Kim DH, Sarbassov DD, Ali SM, King JE, Latek RR, Erdjument-Bromage H, Tempst P, Sabatini DM (2002) Cell 110:163–175
4. Hara K, Maruki Y, Long X, Yoshino K, Oshiro N, Hidayat S, Tokunaga C, Avruch J, Yonezawa K (2002) Cell 110:177–189
5. Yonezawa K, Tokunaga C, Oshiro N, Yoshino K (2004) Biochem Biophys Res Commun 313:437–441
6. Sarbassov DD, Ali SM, Kim DH, Guertin DA, Latek RR, Erdjument-Bromage H, Tempst P, Sabatini DM (2004) Curr Biol 14:1296–1302
7. Guertin DA, Sabatini DM (2005) Trends Mol Med 11:353–361
8. Presneau N, Shalaby A, Idowu B, Gikas P, Cannon SR, Gout I, Diss T, Tirabosco R, Flanagan AM (2009) Br J Cancer 100:1406–1414
9. Saunders RN, Metcalfe MS, Nicholson ML (2001) Kidney Int 59:3–16
10. Dudkin L, Dilling MB, Cheshire PJ, Harwood FC, Hollingshead M, Arbuck SG, Travis R, Sausville EA, Houghton PJ (2001) Clin Cancer Res 7:1758–1764
11. Hidalgo M, Rowinsky EK (2000) Oncogene 19:6680–6686
12. Sousa JE, Costa MA, Abizaid AC, Rensing BJ, Abizaid AS, Tanajura LF, Kozuma K, Van Langenhove G, Sousa AG, Falotico R, Jaeger J, Popma JJ, Serruys PW (2001) Circulation 104:2007–2011
13. Maira SM, Stauffer F, Brueggen J, Furet P, Schnell C, Fritsch C, Brachmann S, Chène P, De Pover A, Schoemaker K, Fabbro D,

Gabriel D, Simonen M, Murphy L, Finan P, Sellers W, García-Echeverría C (2008) Mol Cancer Ther 7:1851–1863
14. Folkes AJ, Ahmadi K, Alderton WK, Alix S, Baker SJ, Box G, Chuckowree IS, Clarke PA, Depledge P, Eccles SA, Friedman LS, Hayes A, Hancox TC, Kugendradas A, Lensun L, Moore P, Olivero AG, Pang J, Patel S, Pergl-Wilson GH, Raynaud F, Robson A, Saghir N, Salphati L, Sohal S, Ultsch MH, Valenti M, Wallweber HJ, Wan NC, Wiesmann C, Workman P, Zhyvoloup A, Zvelebil MJ, Shuttleworth SJ (2008) J Med Chem 51:5522–5532
15. Chau CH, Wang W, Figg WD (2008) Cancer Biol Ther 7:1386–1387
16. Carracedo A, Ma L, Teruya-Feldstein J, Rojo F, Salmena L, Alimonti A, Egia A, Sasaki AT, Thomas G, Kozma SC, Papa A, Nardella C, Cantley LC, Baselga J, Pandolfi PP (2008) J Clin Invest 118:3065–3074
17. Yuan TL, Cantley LC (2008) Oncogene 27:5497–5510
18. Menear KA, Gomez S, Malagu K, Bailey C, Blackburn K, Cockcroft XL, Ewen S, Fundo A, Le Gall A, Hermann G, Sebastian L, Sunose M, Presnot T, Torode E, Hickson I, Martin NM, Smith GC, Pike KG (2009) Bioorg Med Chem Lett 19:5898–5901
19. García-Martínez JM, Moran J, Clarke RG, Gray A, Cosulich SC, Chresta CM, Alessi DR (2009) Biochem J 421:29–42
20. Nowak P, Cole DC, Brooijmans N, Bursavich MG, Curran KJ, Ellingboe JW, Gibbons JJ, Hollander I, Hu Y, Kaplan J, Malwitz DJ, Toral-Barza L, Verheijen JC, Zask A, Zhang WG, Yu K (2009) J Med Chem 52:7081–7089
21. Zask A, Verheijen JC, Curran K, Kaplan J, Richard DJ, Nowak P, Malwitz DJ, Brooijmans N, Bard J, Svenson K, Lucas J, Toral-Barza L, Zhang WG, Hollander I, Gibbons JJ, Abraham RT, Ayral-Kaloustian S, Mansour TS, Yu K (2009) J Med Chem 52:5013–5016
22. Thompson JD, Higgins DG, Gibson TJ (1994) Nucleic Acids Res 22:4673–4680
23. Sali A, Potterton L, Yuan F, Van Vlijmen H, Karplus M (1995) Proteins 23:318–326
24. Hess B, Kutzner C, Van der Spoel D, Lindahl E (2008) J Chem Theor Comput 4:435–447
25. Van der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ (2005) J Comput Chem 26:1701–1718
26. Jones G, Willett P, Glen RC (1995) J Mol Biol 245:43–53
27. Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) J Mol Biol 267:727–748
28. Yu K, Toral-Barza L, Shi C, Zhang WG, Lucas J, Shor B, Kim J, Verheijen J, Curran K, Malwitz DJ, Cole DC, Ellingboe J, Ayral-Kaloustian S, Mansour TS, Gibbons JJ, Abraham RT, Nowak P, Zask A (2009) Cancer Res 69:6232–6240
29. Schüttelkopf AW, van Aalten DM (2004) Acta Crystallogr D 60:1355–1363
30. Li J, Zhu X, Yang C, Shi R (2010) J Mol Model 16:789–798
31. Lemkul JA, Allen WJ, Bevan DR (2010) J Chem Inf Model 50:2221–2235
32. Nosé S (1984) Mol Phys 52:255–268
33. Hoover WG (1985) Phys Rev A 31:1695–1697
34. Parrinello M, Rahman A (1981) J Appl Phys 52:7182–7190
35. Nośe S, Klein ML (1983) Mol Phys 50:1055–1076
36. Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) J App Cryst 26:283–291
37. Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) J Mol Biol 7:95–99

J Mol Model (2012) 18:1625–1632
DOI 10.1007/s00894-011-1183-4

ORIGINAL PAPER

# A theoretical study on the hydrolysis mechanism of carbon disulfide

**Lixia Ling · Riguang Zhang · Peide Han · Baojun Wang**

Received: 30 November 2010 / Accepted: 7 July 2011 / Published online: 30 July 2011


**Abstract** The hydrolysis mechanism of $CS_2$ was studied using density functional theory. By analyzing the structures of the reactant, transition states, intermediates, and products, it can be concluded that the hydrolysis of $CS_2$ occurs via two mechanisms, one of which is a two-step mechanism ($CS_2$ first reacts with an $H_2O$, leading to the formation of the intermediate COS, then COS reacts with another $H_2O$, resulting in the formation of $H_2S$ and $CO_2$). The other is a one-step mechanism, where $CS_2$ reacts with two $H_2O$ molecules continuously, leading to the formation of $H_2S$ and $CO_2$. By analyzing the thermodynamics and the change in the kinetic function, it can be concluded that the rate-determining step involves H and OH in $H_2O$ attacking S and C in $CS_2$, respectively, causing the C=S double bond to change into a single bond. The two mechanisms are competitive. When performing the hydrolysis of $CS_2$ with a catalyst, the optimal temperature is below 252°C.

L. Ling
Research Institute of Special Chemicals,
Taiyuan University of Technology,
Taiyuan 030024 Shanxi, People's Republic of China

L. Ling · P. Han
College of Materials Science and Engineering,
Taiyuan University of Technology,
Taiyuan 030024 Shanxi, People's Republic of China

R. Zhang · B. Wang (✉)
Key Laboratory of Coal Science and Technology,
Ministry of Education and Shanxi Province,
Taiyuan University of Technology,
Taiyuan 030024 Shanxi, China
e-mail: wangbaojun@tyut.edu.cn

**Keywords** Carbon disulfide · Hydrolysis mechanism · DFT · Quantum chemistry calculation

## Introduction

$CS_2$ is a sulfur-containing organic compound that is often found in Claus tail gas, petroleum, natural gas, and various raw gases made from coal [1, 2]. Trace levels of $CS_2$ affects the environment and leads to economic problems within processes downstream, such as the syntheses of methanol, ammonia and urea. The detrimental effect that sulfur has upon many catalysts has been well documented [3]. Apart from affecting catalysts, the presence of sulfur can also lead to increased corrosion of the reactors used in refining processes [4]. The need to remove $CS_2$ has become increasingly important, and some technologies—including catalytic hydrolysis, oxidation conversion, and hydrogenation conversion [5–9]—have been developed to remove $CS_2$. Attention has recently focused on the hydrolysis of $CS_2$ due to the mild reaction conditions and relatively few side reactions involved, as well as the fact that it is inexpensive [10]. Zou et al. [11] studied the hydrolysis of $CS_2$, and found that the equilibrium constant for this hydrolysis $\log K_{CS2}$ is 12.0740 at 100 °C, but only 4.5314 at 600 °C. It is thus apparent that using a suitable temperature range is key to optimizing the conversion rate during $CS_2$ hydrolysis.

In addition, the kinetics of $CS_2$ hydrolysis is considered to be more complicated than the hydrolysis of another sulfur-containing organic compound, COS [6], so it is necessary to investigate the process of $CS_2$ hydrolysis in detail. Up to now, there have been some experimental

studies on the mechanism of $CS_2$ hydrolysis. Shangguang et al. [12] investigated the hydrolysis of $CS_2$, and noted both a two-step and a one-step mechanism, as shown in reactions (1) and (2) below, respectively:

$$\begin{aligned} CS_2 + H_2O &\rightarrow COS + H_2S \\ COS + H_2O &\rightarrow CO_2 + H_2S \end{aligned} \quad (1)$$

$$CS_2 + 2H_2O \rightarrow 2H_2S + CO_2. \quad (2)$$

While $CS_2$ hydrolysis involves these two mechanisms, they found that the hydrolysis of $CS_2$ to $H_2S$ and $CO_2$, not via the intermediate COS, is the major reaction according to mathematical analysis and experimental data fitting. However, other researchers have different opinions in this regard. $CS_2$ hydrolysis has been studied by analyzing the sulfur-containing compounds in tail gas by Wang et al. [13], and it was confirmed that $CS_2$ hydrolysis first gives rise to COS, then COS hydrolyzes via hydrogen thiocarbonate, leading to the formation of $H_2S$. In this process, the hydrolysis of $CS_2$ to COS is the rate-determining step, while COS to $H_2S$ is a fast reaction. This result was backed by Laperdrix et al. [14] and Sahibed-Dine et al. [15], who studied the process of $CS_2$ hydrolysis using IR. Tong et al. [6] studied the kinetics of $CS_2$ hydrolysis, and their results agreed with the two-step mechanism mentioned above. Thus, the predominant mechanism for $CS_2$ hydrolysis remains unknown. With the development of quantum chemistry, density functional theory (DFT) has proven capable of providing qualitative and quantitative insights into reaction mechanisms [16]. The addition of OH radicals to $CS_2$, and the saddle points for conversion of $CS_2$ to COS, which ultimately leads to the formation of $SO_2$, have been calculated at the modified G1 level of theory, and it was found that the addition of OH to a sulfur atom is more favorable than its addition to the carbon atom of $CS_2$ [17]. A theoretical study of the adduct H+$CS_2$ has been reported by Bohn et al. [18]. The major product in the reaction of H atoms with $CS_2$ has been identified as *trans*-HSCS from the matrix infrared spectrum and ab initio calculations of structure and the infrared spectrum. The reaction of N and $CS_2$ was studied using DFT at the B3LYP (full)/6-311+G** and CCSD (T)/6-311+G* levels of theory by Wang et al. [19]. The results showed that the reaction channel N+$CS_2$→CS+NS is easily achieved because of its low energy barrier, which is in agreement with experimental observations. Zheng et al. [20] also investigated the reaction mechanism of the reaction of O ($^3$P) with $CS_2$ using DFT at the B3LYP/6-311++G(d, p) level of theory. It was concluded that the products of the reaction of $CS_2$+O($^3$P) are CS and SO, in good agreement with experimental observations. Recently, the water-catalyzed hydrolysis reaction of $CS_2$ in aqueous solution ($\varepsilon$=78.4) has been investigated at the levels of HF and MP2 with the 6-311++G(d, p) basis set using the polarizable continuum model (PCM), and the activation Gibbs free energies in water solution for the rate-determining steps of 1–6 hydrolysis processes were obtained by Deng et al. [21]. The hydrolysis of COS was also studied by Deng et al. [22].

In the study described in this paper, the two mechanisms of $CS_2$ hydrolysis leading to the formation of $H_2S$ and $CO_2$ were studied using DFT. The main aims of our study were (i) to find the most favorable path for the hydrolysis of $CS_2$ by analyzing the microscopic structural parameters, the change in the thermodynamic function, and the change in the kinetic function, and (ii) to obtain the optimal temperature range for $CS_2$ hydrolysis.

## Computational methods

A DFT method was adopted and calculations were performed using the Dmol$^3$ program [23] running in the Materials Studio 4.0 package. The reactant, intermediates and products were optimized at the level of the generalized gradient approximation (GGA) using the Perdew–Wang 1991 (PW91) functional [24] and DND basis set. Unrestricted spin was chosen, the total SCF tolerance criteria, integration accuracy criteria and orbital cutoff quality criteria were set at medium, and multipolar expansion was set to octupole. TS search calculations were then carried out to find a possible transition state structure for each elementary reaction using the LST/QST method. All of the structures were calculated at the same level of theory, and vibration analysis was carried out for each structure to characterize it as either an equilibrium structure (no imaginary frequency) or a transition state (one imaginary frequency whose vibrational mode corresponds to the reaction coordinate). Meanwhile, the electronic energy ($E_{elec}$) and zero-point vibration energies (ZPVE) were calculated. TS confirmation calculations were carried out using the NEB method to confirm that every transition state leads to the desired reactant and product. All calculations were performed on a Dell Pentium D PC server system.

In order to evaluate the reliability of the selected calculation method and parameters, the bond lengths and bond angles of three sulfur-containing molecules were calculated, and the calculated and experimental results [25] are shown in Table 1. The calculated results are in agreement with the experimental structural parameters, the largest deviation in the bond angles is 1.323°, and the deviation in the bond lengths is smaller than 0.0254 Å. In addition, some results relating to the reaction mechanism were obtained using the above method and parameters in

**Table 1** Calculated and experimental bond lengths (Å) and bond angles (°) of COS, $CS_2$ and $H_2S$

| COS | | | | | $CS_2$ | | | | $H_2S$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| C—O | | C—S | | ∠ O—C—S | | C—S | | ∠ S—C—S | | H—S | | ∠ H—S—H | |
| $L_{cal}$ | $L_{exp}$ | $L_{cal}$ | $L_{exp}$ | $A_{cal}$ | $A_{exp}$ | $L_{cal}$ | $L_{exp}$ | $A_{cal}$ | $A_{exp}$ | $L_{cal}$ | $L_{exp}$ | $A_{cal}$ | $A_{exp}$ |
| 1.1730 | 1.1578 | 1.5720 | 1.5601 | 179.927 | 180.000 | 1.5660 | 1.5526 | 180.000 | 180.000 | 1.3610 | 1.3356 | 90.797 | 92.120 |

our previous studies [26–28]. At the same time, the bigger basis set, DNP, which is equivalent in accuracy to the 6-31G** Gaussian orbital basis set, was used to calculate the step IM1→IM2 (see Fig. 1). The results show that the activation energy is 168.04 kJ $mol^{-1}$ with the DND basis set, and 169.01 kJ $mol^{-1}$ with DNP basis set. The rate constants are $-45.19\ s^{-1}$ and $-47.14\ s^{-1}$, respectively. This indicates that the DND basis set can meet the requirements of the calculation.

## Results and discussion

### Reaction mechanism

The two-step mechanism and one-step mechanism for $CS_2$ hydrolysis (reactions (1) and (2), respectively) are shown in Fig. 1, the optimized geometries for the reactants, intermediates, transition states, products and their atomic numbers along the reaction pathways are shown schematically in Fig. 2, and the imaginary frequency of each transition state and the bonds corresponding to relative normal vibrations during the process of $CS_2$ hydrolysis are shown in Table 2.

In the two-step mechanism of $CS_2$ hydrolysis, first $CS_2$ reacts with an $H_2O$ molecule, leading to IM1. IM1 is the reactant-like intermediate in which two reactant molecules interact with each other by van der Waals forces of strength 7.49 kJ $mol^{-1}$, and their geometries remain almost the same as the corresponding free molecules. Then, H and OH in $H_2O$ attack S and C atoms in $CS_2$, respectively, and the attacked double bond in $CS_2$ changes into a single bond. The original linear $CS_2$ molecule is thus distorted, and a hydrogen dithiocarbonate species IM2 is formed via TS1. TS1 has one imaginary frequency of $-1290.68\ cm^{-1}$, corresponding to the stretch mode of O1–C5 and the migration of H3 from O1 to S4. Subsequently, H2 migrates from O1 to S4, leading to the formation of by-product IM3. The bond length of S4–C5 in IM3 is 3.667 Å, which shows that $H_2S$ separates from the original structure and COS is

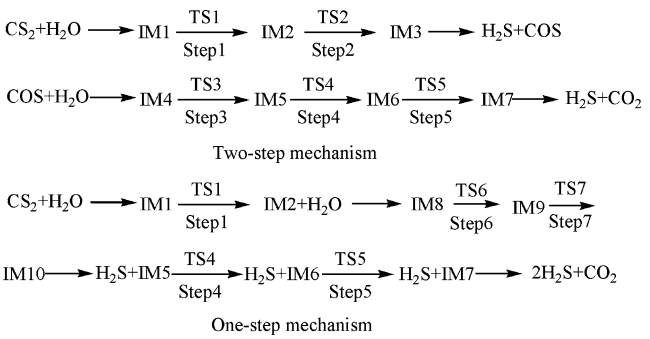



**Fig. 1** Mechanisms for the hydrolysis of $CS_2$

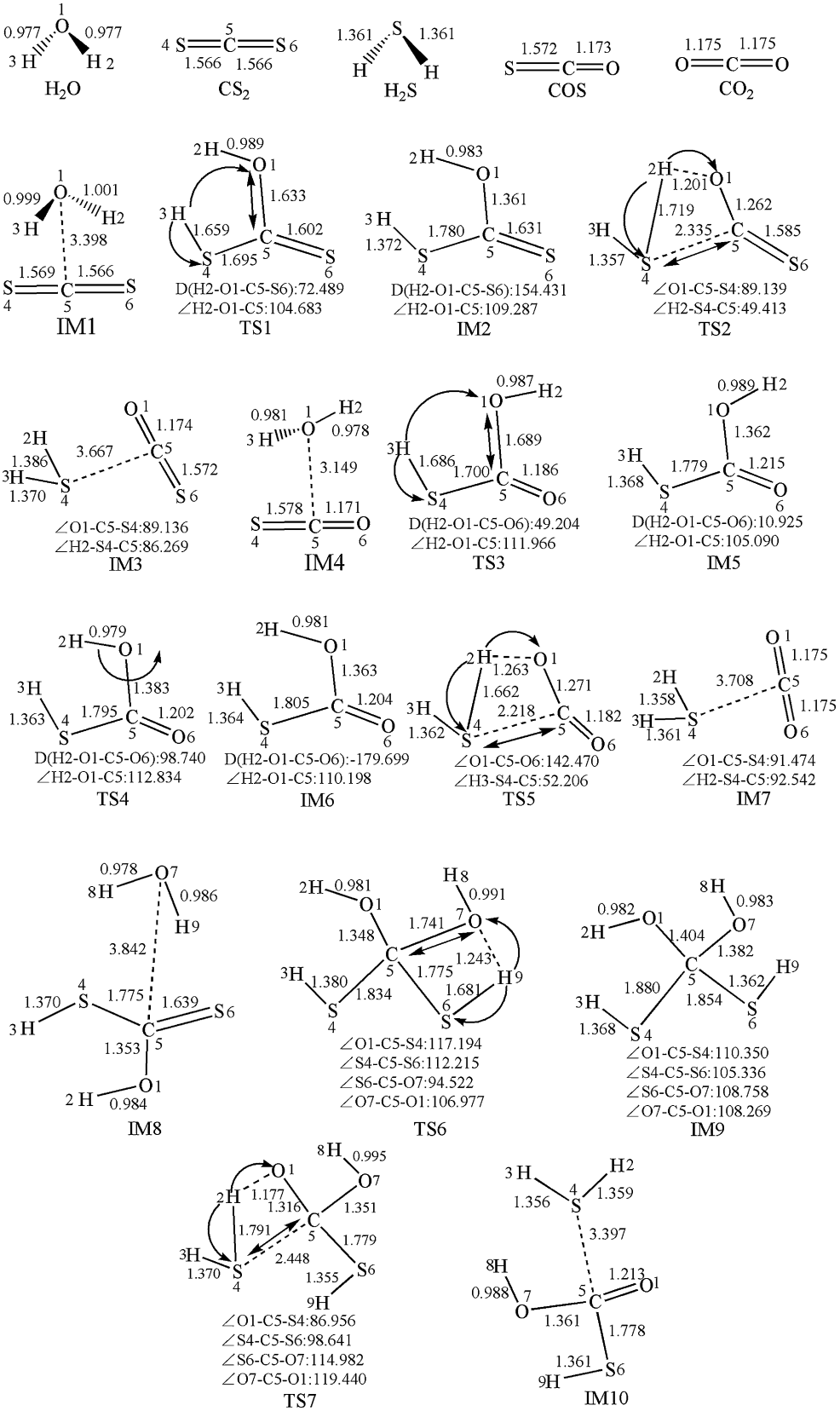



**Fig. 2** Optimized geometries for various species involved in the hydrolysis of $CS_2$

formed. The two product molecules, $H_2S$ and COS, interact with each other by van der Waals forces of strength 0.62 kJ $mol^{-1}$, and their geometries remain almost the same as those for the corresponding free molecules. The

**Table 2** Imaginary frequency of each transition state and the bonds corresponding to relative normal vibrations

| Transition state | Imaginary frequency ($cm^{-1}$) | Bonds corresponding to normal vibrations |
|---|---|---|
| TS1 | −1290.68 | H3–O1–S4; O1–C5 |
| TS2 | −1342.31 | H2–O1–S4; S4–C5 |
| TS3 | −1367.93 | H3–O1–S4; O1–C5 |
| TS4 | −616.76 | H2–O1–C5 |
| TS5 | −1407.03 | H2–O1–S4; S4–C5 |
| TS6 | −1268.34 | H9–O7–S6; C5–O7 |
| TS7 | −1258.29 | H2–O1–S4; C5–S4 |

corresponding transition state TS2 has a four-membered ring (H2–S4–C5–O1), which has one imaginary frequency of −1342.31 $cm^{-1}$ due to the stretch mode of S4–C5 and the migration of H2 from O1 to S4.

Next, COS is hydrolyzed by another $H_2O$ molecule, resulting in the reactant-like intermediate IM4, in which there is an interaction energy of 7.10 kJ $mol^{-1}$ between COS and $H_2O$. H and OH in $H_2O$ attack the S and C atoms in COS, respectively, leading to the formation of a hydrogen thiocarbonate species IM5. The bond lengths of O1–C5 and H3–S4 are 1.362 Å and 1.368 Å, respectively. In step 4, IM5 then isomerizes to IM6; it can be seen in Fig. 2 that H2 is deflected, causing the dihedral angle H2–O1–C5–O6 to change from 10.925° to −179.699°. Finally, H2 migrates from O1 to S4, leading to the formation of by-product IM7 via TS5, which has one imaginary frequency of −1407.03 $cm^{-1}$ corresponding to the stretch mode of S4–C5 and the migration of H2 from O1 to S4. In IM7, $H_2S$ and $CO_2$ are formed, and there is an interaction energy of 11.10 kJ $mol^{-1}$ between $H_2S$ and $H_2O$.

The one-step mechanism has the same steps as the two-step mechanism from reactants to IM2. IM2 then reacts with another $H_2O$ molecule, initially producing IM8, in which there is an interaction energy of 30.16 kJ $mol^{-1}$ between the two molecules. H and OH in $H_2O$ then attack the S6 and C5 atoms in IM2, respectively, leading to the formation of a tetrahedral intermediate, IM9, in which the angles O1–C5–S4, S4–C5–S6, S6–C5–O7 and O7–C5–O1 are 110.350°, 105.336°, 108.758° and 108.269°, respectively (similar to the angle H–C–H in $CH_4$). Subsequently, H2 in IM9 migrates from O1 to S4 via TS7, which has one imaginary frequency of −1258.29 $cm^{-1}$ corresponding to the stretch mode of S4–C5 and the migration of H2 from O1 to S4. Meanwhile, an $H_2S$ molecule separates from the original structure, a similar structure to IM5 is formed, and there is an interaction energy of 6.81 kJ $mol^{-1}$ between the two molecules. The similar structure to IM5 is then optimized, resulting in IM5 itself. From IM5 to the final products $CO_2$ and $H_2S$, the one-step mechanism exhibits the same steps as the two-step mechanism.

In these two mechanisms, H migration occurs in every step. There are two types of H migration: intermolecular H migration (for example, H in $H_2O$ migrates to S in $CS_2$, COS and IM8 in steps 1, 3 and 6, respectively), and intramolecular H migration (see steps 2, 4, 5 and 7). In steps 2, 5 and 7, the H in the OH group of the intermediate migrates to S, leading to the release of $H_2S$. It can be concluded that H migration plays an important role in $CS_2$ hydrolysis. However, it is difficult to determine which of the mechanisms is more favorable than the other from the microscopic structure. Therefore, analyses of the thermodynamic and kinetic functional changes during hydrolysis are needed.

## Thermodynamic analysis of $CS_2$ hydrolysis

The thermodynamic functional changes for the two mechanisms are listed in Table 3. In the two reaction pathways, the steps in which $H_2S$ is released via intramolecular H migration (steps 2, 5 and 7) have negative Gibbs free energy changes in Table 3, which shows that these steps are spontaneous processes under isothermal-isobaric conditions from a thermodynamic point of view. The steps involving the combination of two molecules—steps 1, 3 and 6—have positive Gibbs free energy changes, which shows that these steps are inhibited under isothermal-isobaric conditions. It can also be concluded that $CS_2$ hydrolysis is a feasible reaction from a thermodynamic point of view, as the total $\Delta_r S_m^\circ$ is positive and the total $\Delta_r G_m^\circ$ is negative, which means that it can occur at room temperature. However, despite the fact that hydrolysis of the $CS_2$ is thermodynamically feasible, the conversion rate within a certain time is low [29]. Kinetic data are therefore necessary to study the hydrolysis of $CS_2$ further.

## Kinetic analysis of $CS_2$ hydrolysis

A detailed kinetic analysis of the two mechanisms will be discussed in this section. According to transition state theory, the activation enthalpy $\Delta_r H_m^{\neq}$, activation entropy $\Delta_r S_m^{\neq}$ and activation energy $E_a$ can be obtained from Eqs. 3–5, respectively, while the rate constant $k$ can be expressed as in Eq. 6 [30, 31]:

$$\Delta_r H_m^{\neq} = H(TS) - H(R) \tag{3}$$

$$\Delta_r S_m^{\neq} = S(TS) - S(R) \tag{4}$$

$$E_a = \Delta_r H_m^{\neq} + nRT \tag{5}$$

**Table 3** The thermodynamic properties of $CS_2$ hydrolysis

| Steps | $\Delta_r H_m^\theta$ (kJ · mol$^{-1}$) | | $\Delta_r S_m^\theta$ (J · mol$^{-1}$ · K$^{-1}$) | | $\Delta_r G_m^\theta$ (kJ · mol$^{-1}$) | |
|---|---|---|---|---|---|---|
| | 298.15 K | 875 K | 298.15 K | 875 K | 298.15 K | 875 K |
| Two-step mechanism | | | | | | |
| Step 1 | 40.50 | 39.96 | −34.72 | −36.47 | 50.85 | 71.88 |
| Step 2 | −92.48 | −90.51 | 43.96 | 48.42 | −105.58 | −132.88 |
| Step 3 | −2.79 | −5.75 | −42.41 | −49.41 | 9.85 | 37.48 |
| Step 4 | 19.92 | 20.41 | 4.70 | 5.77 | 18.51 | 15.35 |
| Step 5 | −66.75 | −64.15 | 55.46 | 61.36 | −83.29 | −117.84 |
| One-step mechanism | | | | | | |
| Step 1 | 40.50 | 39.96 | −34.72 | −36.47 | 50.85 | 71.88 |
| Step 6 | 3.58 | 2.72 | −32.13 | −34.57 | 13.16 | 32.96 |
| Step 7 | −81.63 | −81.02 | 45.44 | 47.08 | −95.18 | −122.22 |
| Step 4 | 19.92 | 20.41 | 4.70 | 5.77 | 18.51 | 15.35 |
| Step 5 | −66.75 | −64.15 | 55.46 | 61.36 | −83.29 | −117.84 |
| Total | −110.10 | −110.97 | 20.10 | 17.45 | −116.10 | −126.26 |

$$k = \frac{k_b T}{h}\left(\frac{p^o}{RT}\right)^{1-n} \exp\left(\frac{-\Delta_r H_m^{\neq}}{RT}\right) \exp\left(\frac{\Delta_r S_m^{\neq}}{R}\right). \quad (6)$$

Here, $k$ is the rate constant, $T$ is the reaction temperature, $k_b$, $h$, $p^\circ$ and $R$ are Boltzmann's constant, Planck's constant, standard atmospheric pressure, and the fundamental gas constant, and $n$ is the number of reactants.

The activation parameters $\Delta_r H_m^{\neq}$, $\Delta_r S_m^{\neq}$, $E_a$, as well as the rate constant ln$k$ of every elementary reaction at the temperatures 298.15 K and 875 K, are shown in Table 4. According to the activation entropy ($\Delta_r S_m^{\neq}$) data in Table 4, it can be seen that these are entropy-decreasing pathways as the temperature increases. The activation enthalpy $\Delta_r H_m^{\neq}$ decreases slowly and monotonously with increasing temperature. We can also see that the activation energy of COS hydrolysis is lower than that of $CS_2$ hydrolysis, and the rate constant ln$k_1$ of step 1 is larger than the rate constant ln$k_3$ of step 3, which shows that $CS_2$ is more difficult than COS to hydrolyze, in agreement with the results of Huisman [7]. The activation energy of TS1 is the largest and the rate

**Table 4** The activation enthalpy, activation entropy, activation energy, and rate constant of each elementary reaction

| Elementary reaction | $\Delta_r H_m^{\neq}$ (kJ mol$^{-1}$) | | $\Delta_r S_m^{\neq}$ (J mol$^{-1}$ K$^{-1}$) | | $E_a$ (kJ mol$^{-1}$) | | ln$k$* | |
|---|---|---|---|---|---|---|---|---|
| | 298.15 K | 875 K | 298.15 K | 875 K | 298.15 K | 875 K | 298.15 K | 875 K |
| Two-step mechanism | | | | | | | | |
| Step 1 | 163.08 | 159.19 | −42.92 | −51.05 | 168.04 | 173.74 | −45.19 | −0.11 |
| Step 2 | 80.50 | 78.08 | 0.06 | −4.36 | 82.98 | 85.36 | −3.01 | 19.28 |
| Step 3 | 137.92 | 133.24 | −43.68 | −53.56 | 142.87 | 147.79 | −35.13 | 3.16 |
| Step 4 | 43.91 | 40.66 | −0.50 | −6.22 | 46.39 | 47.93 | 11.68 | 24.20 |
| Step 5 | 81.17 | 78.68 | −2.00 | −6.64 | 83.65 | 85.95 | −3.53 | 18.92 |
| One-step mechanism | | | | | | | | |
| Step 1 | 163.08 | 159.19 | −42.92 | −51.05 | 168.04 | 173.74 | −45.19 | −0.11 |
| Step 6 | 139.85 | 135.36 | −47.49 | −56.97 | 144.81 | 149.91 | −36.37 | 2.45 |
| Step 7 | 73.31 | 69.83 | −6.61 | −13.36 | 75.79 | 77.10 | −0.91 | 19.33 |
| Step 4 | 43.91 | 40.66 | −0.50 | −6.22 | 46.39 | 47.93 | 11.68 | 24.20 |
| Step 5 | 81.17 | 78.68 | −2.00 | −6.64 | 83.65 | 85.95 | −3.53 | 18.92 |

* Steps 1, 3 and 6 are second-order reactions: the units are mol$^{-1}$ dm$^3$ s$^{-1}$; steps 2, 4, 5 and 7 are first-order reactions: the units are s$^{-1}$

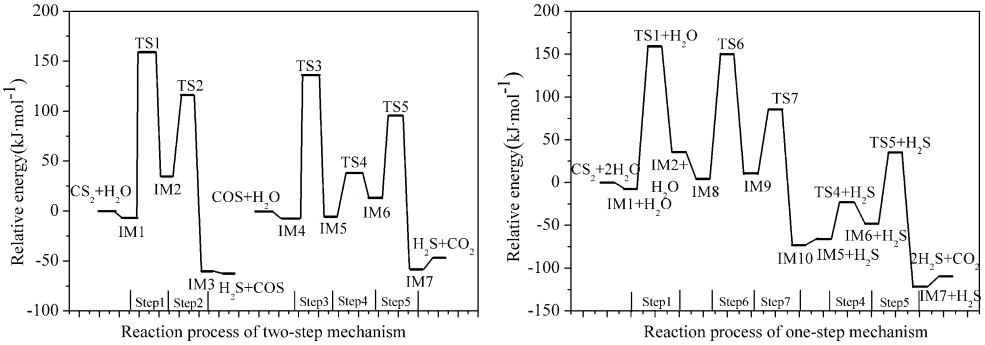



**Fig. 3** Sketch of the energies of the stationary points for $CS_2$ hydrolysis

constant $\ln k_1$ is the smallest in both mechanisms, which shows that step 1 (IM1→IM2) is the rate-determining step in both mechanisms. Although a higher barrier of 247.90 kJ $mol^{-1}$ to the proton transfer of $H_2O$ to $CS_2$ was obtained with the MP2 method by Deng et al. [21] compared to the value of 168.04 kJ $mol^{-1}$ obtained in our calculation, we reached the same conclusion: that the proton transfer process for the one-water hydrolysis mechanism of carbon disulfide is the rate-determining step. The energy barriers for the two mechanisms are similar, as shown in Fig. 3, which indicates that the two-step mechanism and the one-step mechanism are parallel and competitive. Increasing the reaction temperature favors $CS_2$ hydrolysis, as we can see by comparing to the rate constants at different temperatures in Table 4. However, the negative total $\Delta_r H°_m$ in Table 3 shows that $CS_2$ hydrolysis is an exothermic reaction, so increasing the temperature is disadvantageous for this reaction, in agreement with the results reported in [11]. So what is the optimal hydrolysis temperature of $CS_2$?

The optimal hydrolysis temperature of $CS_2$

In order to find the optimal hydrolysis temperature of $CS_2$, the equilibrium constants lg$K$ and the rate constants ln$k$ of the rate-determining step at different temperatures are shown in Fig. 4 (these values can be obtained from Eqs. 7 [32] and 6), respectively. It is clear that the equilibrium constant lg$K$ decreases monotonously, and the rate constant ln$k$ of the rate-determining step increases monotonously with increasing hydrolysis temperature. There is a crossing point at about 525 K, which shows that the optimal temperature is about 252 °C when $CS_2$ hydrolysis is performed without catalysts. The hydrolysis temperature should be below 252 °C when a catalyst is used, and the precise temperature range varies depending on the catalyst, equipment, etc. used. Wang et al. [13] found that the conversion of $CS_2$ reaches 91% at 130 °C when the catalyst is 5% Ce, 14% $K_2CO_3$ and 81% $Al_{20}O_3$, the space velocity is 21773 $h^{-1}$, the relative humidity is 1.6%, and the $CS_2$ concentration is 160 mgS/$m^3$. However, $CS_2$ conversion in the presence of $Al_2O_3$ with 6% $K_2CO_3$ can reach more than 98% at 160 °C and atmospheric pressure when the space velocity is 10000 $h^{-1}$ [11].

$$\Delta_r G^{\theta}_m(T) = -RT \ln K^{\theta} \tag{7}$$

## Conclusions

The process of $CS_2$ hydrolysis has been explored using density functional theory, and two detailed reaction mechanisms for $CS_2$ hydrolysis were presented. Ten intermediates and seven transition states occur along the reaction pathways. The results of a thermodynamic analysis show that $CS_2$ hydrolysis can occur at room temperature, but the reaction rate is very low. The rate-determining step of the two-step mechanism and one-step mechanism involves H and OH in $H_2O$ attacking S and C

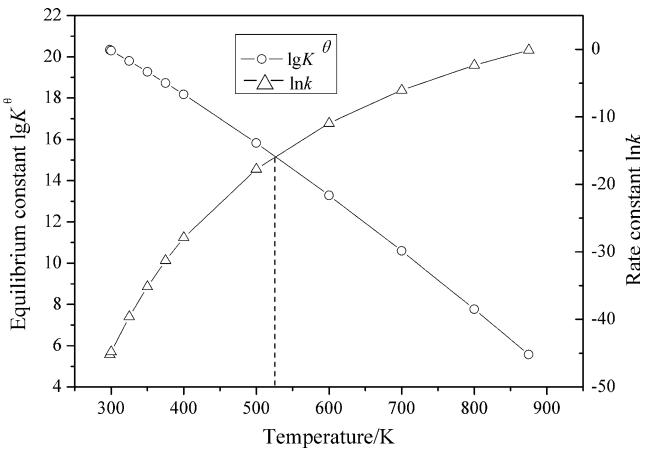



**Fig. 4** The equilibrium constant lg$K$ and the rate constant ln$k$ at different temperatures

in $CS_2$, respectively, so that the C=S double bond changes into a single bond. The two mechanisms are competitive. The optimal temperature is about 252 °C when $CS_2$ hydrolysis is performed without a catalyst, and the hydrolysis temperature should be below 252 °C when a catalyst is used.

**Acknowledgments** This work was supported by the National Natural Science Foundation of China (nos. 20776093, 20976115, 20906066), and the Foundation of Shanxi Province (no. 2009021015).

## References

1. Guo B, Chang LP, Xie KC (2006) Fuel Process Technol 87:873–881
2. Zou FL, Li CH, Guo HX (1997) J Mol Catal China 11:138–144
3. Yao XQ, Li YW (2009) J Mol Struct Theochem 899:32–41
4. Zhang RG, Ling LX, Wang BJ (2010) J Mol Model. doi:10.1007/s00894-010-0686-8
5. Shangguan J, Li CH, Guo HX (1998) J Nat Gas 7:24–30
6. Tong S, Dalla Lana IG, Chuang KT (1995) Can J Chem Eng 73:220–227
7. Huisman HM, van der Berg P, Mos R, van Dillen AJ, Geus JW (1994) Appl Catal A Gen 115:157–171
8. George ZM (1974) J Catal 32:261–271
9. Tong S, Dalla Lana IG, Chuang KT (1992) Can J Chem Eng 70:516–522
10. Wang L, Li FL, Wu DY, Wang SD, Yuan Q (2004) J Fuel Chem Technol 32:466–470
11. Zou FL, Li CH, Guo HX (1996) J Nat Gas 5:334–340
12. Shangguan J, Guo HX (1997) J Fuel Chem Technol 25:277–283
13. Wang L, Li FL, Wu DY, Wang SD, Yuan Q (2005) Nat Gas Chem Ind 30:1–5
14. Laperdrix E, Justin I, Costentin G, Saur O, Lavalley JC, Aboulayt A, Ray JL, Nédez C (1998) Appl Catal B Environ 17:167–173
15. Sahibed-Dine A, Aboulayt A, Bensitel M, Mohammed Saad AB, Daturi M, Lavalley JC (2000) J Mol Catal A Chem 162:125–134
16. Yang T, Wen XD, Huo CF, Li YW, Wang JG, Jiao HJ (2009) J Mol Catal A Chem 30:29–136
17. McKee ML (1993) Chem Phys Lett 201:41–46
18. Bohn RB, Brabson GD, Andrews L (1992) J Phys Chem 96:1582–1589
19. Wang YC, Lv LL, Dai GL, Wang DM, Geng ZY (2004) Chem Res 15:38–42
20. Zheng Y, Zhu YQ, Li LC (2005) J Sichuan Normal Univ (Nat Sci) 28:708–711
21. Deng C, Wu XP, Sun XM, Ren Y, Sheng YH (2009) J Comput Chem 30:285–294
22. Deng C, Li QG, Ren Y, Wong NB, Chu SY, Zhu HJ (2008) J Comput Chem 29:466–480
23. Delley B (2005) J Chem Phys 113:7756–7764
24. Perdew JP, Chevary JA, Vosko SH, Jackson KA, Pederson MP, Singh DJ, Fiolhais C (1992) Phys Rev B 46:6671–6687
25. Lide DR (2001) Handbook of chemistry and physics, 82nd edn. CRC, New York, pp 9-19–9-21
26. Ling LX, Zhang RG, Wang BJ, Xie KC (2009) J Mol Struct Theochem 905:8–12
27. Ling LX, Zhang RG, Wang BJ, Xie KC (2010) J Mol Struct Theochem 952:31–35
28. Ling LX, Zhang RG, Wang BJ, Xie KC (2009) Chin J Chem Eng 17:805–813
29. Rodes C, Riddel SA, West J, Williams BP, Hutchings GJ (2000) Catal Today 59:443–464
30. Hyre DE, Trong IL, Freitag S, Stenkamp RE, Stayton PS (2000) Protein Sci 9:878–885
31. Frisch C, Fersht AR, Schreiber G (2001) J Mol Biol 308:69–77
32. Fu XC, Shen WX, Yao TY (1990) Physical chemistry, 4th edn. Higher Education Press, Beijing, p 387

J Mol Model (2012) 18:1633–1644
DOI 10.1007/s00894-011-1178-1

ORIGINAL PAPER

# Proton transfer reaction and intermolecular interactions in associates of 2,5-dihydroxy-1,8-naphthyridine

**Borys Ośmiałowski**

Received: 29 March 2011 / Accepted: 4 July 2011 / Published online: 30 July 2011


**Abstract** Tautomerism in monomers/dimers and association of 2,5-dihydroxy-1,8-naphthyridine was studied at the DFT level recently recommended for studies of non-covalent interactions. Studied dimers are stabilized by double and triple hydrogen-bonding. In some associates the intermolecular proton transfer may take place. Transition state related to the double proton transfer reactions were calculated and discussed in terms of energetics, changes in atomic charges upon association, aromaticity (HOMA), properties of *hydrogen bond critical point* (QTAIM methodology) and geometry change during this reaction. It was found that double proton transfer is supported by third hydrogen bond or by weak secondary interaction. Some protons in transition states are shared between two basic atoms, while other are covalently bound only to one of them. The said process leads to replacement of secondary interactions of attractive character to repulsive and *vice versa*. Overall, results suggest that in subjected compound the triple hydrogen-bonded associate may be in equilibrium with double hydrogen-bonded dimer.

**Keywords** Dimerization · Hydrogen bond · Naphthyridine · Proton transfer · Self-association

**Electronic supplementary material** The online version of this article (doi:10.1007/s00894-011-1178-1) contains supplementary material, which is available to authorized users.

B. Ośmiałowski (✉)
Faculty of Chemical Technology and Engineering,
University of Technology and Life Sciences,
Seminaryjna 3,
PL-85-326 Bydgoszcz, Poland
e-mail: borys.osmialowski@utp.edu.pl

## Introduction

The association of molecules in solution and solid state depends on the balance of several types of intermolecular forces. One of the most frequently met, especially in crystal, [1–4] is the hydrogen bond (HB). Others such as π-π stacking and halogen bonding may play a role as a predictable force used in crystal engineering [5–19]. Despite that the hydrogen bonds are often used in non-covalent synthesis, organocatalysis and self-associated non-covalent polymers, these still need to be studied due to occurrence in biology [20–23]. Hydrogen bonds are often called "salt bridges" when they appear in peptides [24] and are responsible for their folding [25–27]. Although HBs are much weaker than the covalent bonds, if present in a great number they are able to control the shape of large molecules by the interaction of acidic and basic sites. Thus multiple HBing plays a very important role in self-organization [28–35]. It is obvious that the more hydrogen bonds are present in the associate, the more stable it is. In multiple HBed molecules, so called secondary interactions (SIs) play a role[36]. The SIs are even weaker than HBs but are also able to influence the aggregate stability [37–43]. The secondary interactions are those forces that act in the space between HBs. The discussion on the strength of interaction in these HBed patterns was a subject of several publications [9, 43–45]. Various interaction energies in these patterns are due to the presence of different types of secondary interactions (attractive ($SI_a$) secondary interaction or repulsive ($SI_r$) one). These are exemplified below (Fig. 1). Thus when one defines the HB donor as "D" and acceptor as "A", there are two possibilities of hydrogen bonding pattern in double-hydrogen-bonded complexes (AA/DD and AD/DA) and three in triple hydrogen bonded (AAA/DDD, AAD/DDA and ADA/DAD).

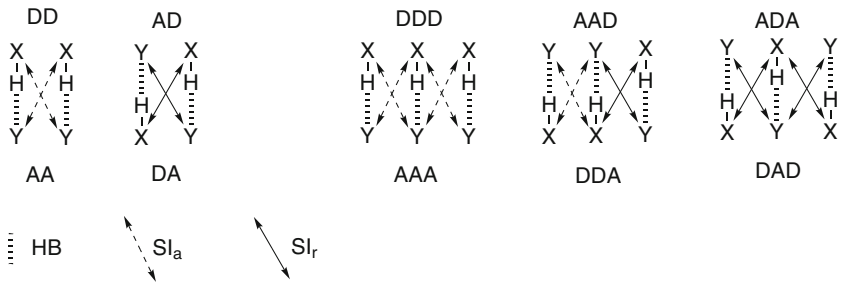



**Fig. 1** Double and triple-hydrogen-bonded complexes and secondary interactions within various HBed patterns

As was already said, the number of HBs and SIs might influence the intermolecular interactions. Also, the bulkiness of the substituent[29, 46] or the electronic repulsion[47–49] may be responsible for selective binding of one molecule versus another. Recently the effect of secondary interactions on overall stability of naphthyridine and pyridine associates was studied computationally [37, 38]. It was also shown that the SIs can influence the transition state (TS) in 2-hydroxypyridine/2-[1*H*]-pyridone equilibrium [37]. The studied here 2,5-dihydroxy-1,8-naphthyridine may exist as six tautomeric forms (Fig. 2). These tautomers may interact with another molecule by hydrogen bonding. Although the triple hydrogen-bonded complexes are more stable than double hydrogen-bonded, the latter should also be taken into account when studying a complicated equilibrium. All of the above-mentioned tautomeric forms of tautomers of 2,5-dihydroxy-1,8-naphthyridine consist of either a N(H)-C=O or N=C-OH fragment that is able to form homo- or heterocomplexes. This results in a full palette of probable intermolecular interactions. It is worth reminding that 2-[1*H*]-pyridone dimer is considered as the most stable among double-hydrogen-bonded dimers [42]. The molecule of 2-[1*H*]-pyridone also contains the N(H)-C=O fragment.

The aim of this work is to (a) describe the tautomeric preferences of 2,5-dihydroxy-1,8-naphthyridine, (b) study the intermolecular interactions of this molecule, (c) find out if the intermolecular proton transfer is possible.

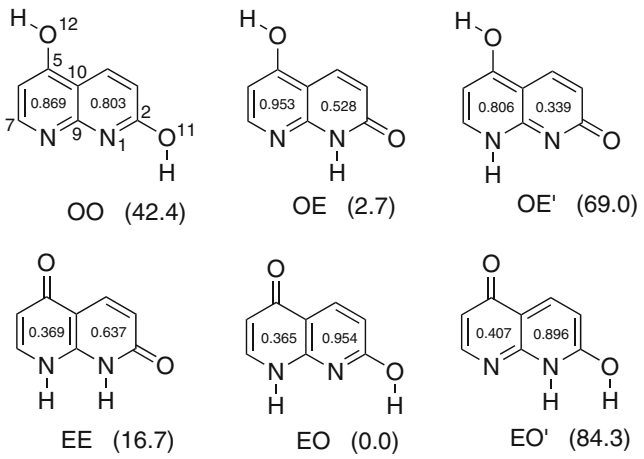



**Fig. 2** Possible tautomeric forms of 2,5-dihydroxy-1,8-naphthyridine, atom numbering, HOMA values (inside rings) and relative energies (in parentheses in kJ mol$^{-1}$)

## Methods

Computations were performed with the use of Gaussian 03 [50]. The M05 functional [51] was used as suggested previously for non-covalent interactions [52–54]. The M05 method was selected in order to sustain methodology with previous studies by the author for similar systems [37, 38]. All monomers and double- and triple-hydrogen-bonded dimers were optimized with the use of 6-31+G(2 d,2p) basis set [55–57] (diffuse functions added to heavy atoms and polarization functions on heavy atoms and hydrogens to describe the hydrogen bonding efficiently). The geometries obtained were used for the frequency calculations (only positive frequencies were obtained, *i.e.*, structures are in energy minima). The double proton transfer TSs for studied complexes were found with the use of synchronous transit-guided quasi-Newton algorithm [58]. The structures of substrate and product for optimization of transition state were taken from respective stable complexes optimized in a previous step. For all TSs, one imaginary frequency was obtained (the negative frequencies are given in SI file).

The energies of interaction between molecules were corrected to basis set superposition error (BSSE) using counterpoise method [59, 60] by single point calculations at the same level of theory with default settings. The geometry optimizations were carried out without BSSE correction to avoid time consuming computations. For all structures, the zero-point energy (ZPE) correction to the energy was applied. Charges were calculated with the use of natural bond orbital (NBO) [61, 62] while properties of the bond critical point (BCP) were calculated using AIM2000 [63] program. To analyze the aromaticity of heterocyclic rings, the geometry based index (HOMA) was used [64, 65].

## Results and discussion

Figure 2 shows the tautomeric forms of subjected compound, atom numbering, tautomer labels and HOMA

indexes for the heterocyclic rings (numbers inside each ring). The '**O**' refers to the hydroxyimine and '**E**' to the *oxo* form of the heterocyclic ring.

In general the aromatic character of the heterocyclic rings is lower in quinone-like forms than that of hydroxyimine. The O12 → N8 proton transfer causes higher changes in HOMA index than the O11 → N1 proton transfer, *i.e.*, the 1,4-quinone-like rings have significantly lower aromaticity than 1,2-quinone-like ones. This is caused by the higher effective conjugation between neighboring NH and CO moieties in 1,2-quinone structure. In case of **EO'** and **OE'**, the relatively high HOMA indices (ca. 0.8) for rings with protonated nitrogen atoms may be explained by the presence of two electronodonors (NH and OH) present in that ring capable for conjugation. In two remaining rings, the HOMA is much lower. Table 1 shows the selected distances, valence angles and charges at selected atoms within the monomeric species. Table 2 collects the data for double hydrogen-bonded associates. For dimer calculations, only these tautomers were taken whose relative energy for the most stable is lower than ca. 42 kJ $mol^{-1}$ (10 kcal $mol^{-1}$).

Geometry data in Table 1 shows that the distances that are crucial for intermolecular hydrogen bonding (N1–N8, N1–O11) are close to being constant independently on the tautomeric form. It is obvious that the C–O and C=O bond lenghts are different from one form to another. The C=O double bond is ca. 0.13 Å shorter than the C–O one in hydroxy form. The tautomeric state influences also the length of other bonds. These are C3–C4, C6–C7 and C9–C10. In the case of the first one, the changes are the smallest. The double bond localization is much better seen for C6–C7 and C9–C10 bonds. In general, the geometry parameters of the σ-skeleton are influenced by the tautomeric state and (de)localization of π electrons, but the N1–N8 and N1–O11 distances remain almost constant. In the light of intermolecular hydrogen bonding, it is important where the labile protons are located and what the order is of hydrogen bond donors and acceptors.

The relative energies ($E_{rel}$) of tautomeric forms show that the two most stable forms lose aromaticity in one heterocyclic ring. Moreover, it is worth stressing that the least stable form below the energy level cut-off is **OO** where both rings are aromatic. This is somehow similar to the isomeric 2,7-dihydroxy-1,8-naphthyridine [38]. Except tendency of cyclic, conjugated ring to be aromatic, the effect of neighborhood of electronodonor and acceptor may influence the relative tautomer stability. Thus, the conjugation in H–N–C=O moiety may influence the stability of the *oxo* form. Also, the benzoannulation affects the tautomeric equilibria [66–73]. In the studied structures, the aromatic character of the ring fused (let say A) to the one that loses aromaticity (B) is increased and that in turn compensates loss of the aromaticity in ring B and, to some extent, stabilizes the structure. Moreover, in the H–N–C=O moiety, the lone pair of nitrogen is perpendicular to the plane of the ring making these electrons easier to conjugate.

The atomic charges collected in Table 1 show some general behavior. As it was observed for the HOMA index before, the close proximity of OH or C=O group influences the charge at respective nitrogen atoms. Thus, the charge at N1 is higher than that on N8 (compare values in **OO** and **EE** forms). Except **OE'** form, the changes on NH nitrogen are higher than on –N = one. The same is realized for

**Table 1** The chosen geometry parameters for optimized monomers and atomic charges

| Form distance [Å] valence angle [°] | **OO** | **EE** | **OE** | **EO** | **OE'** | **EO'** |
|---|---|---|---|---|---|---|
| N1–N8 | 2.298 | 2.342 | 2.300 | 2.304 | 2.284 | 2.296 |
| N1–O11 | 2.273 | 2.269 | 2.273 | 2.276 | 2.281 | 2.316 |
| C2–O11 | 1.343 | 1.216 | 1.217 | 1.341 | 1.220 | 1.334 |
| C5–O12 | 1.348 | 1.225 | 1.345 | 1.226 | 1.342 | 1.230 |
| C3–C4 | 1.364 | 1.358 | 1.352 | 1.374 | 1.350 | 1.395 |
| C6–C7 | 1.402 | 1.348 | 1.389 | 1.354 | 1.364 | 1.367 |
| C9–C10 | 1.429 | 1.383 | 1.414 | 1.401 | 1.435 | 1.420 |
| C7N8C9 | 116.9 | 120.7 | 116.1 | 121.6 | 124.1 | 114.0 |
| C9N1C2 | 118.9 | 125.3 | 126.9 | 116.9 | 118.9 | 124.2 |
| charge at N1 | −0.530 | −0.641 | −0.629 | −0.579 | −0.596 | −0.590 |
| charge at N8 | −0.486 | −0.603 | −0.524 | −0.597 | −0.568 | −0.586 |
| charge at O11 | −0.708 | −0.621 | −0.617 | −0.702 | −0.607 | −0.684 |
| charge at O12 | −0.697 | −0.614 | −0.693 | −0.617 | −0.687 | −0.636 |
| charge at H1 | - | 0.448 | 0.453 | - | - | 0.454 |
| charge at H8 | - | 0.451 | - | 0.457 | 0.468 | - |
| charge at H11 | 0.531 | - | - | 0.527 | - | 0.527 |

**Table 2** Chosen geometry parameters for double hydrogen-bonded dimers and the energy of interaction ($E_{int}$ [kJ mol$^{-1}$])

| | **OO/OO$^1$** | **OO/OO$^2$** | **EE/EE** | **OE/OE$^1$** | **OE/OE$^2$** | **OE/OE$^3$** | **EO/EO$^1$** | **EO/EO$^{2,a}$** | **EO/EO$^{3,a}$** |
|---|---|---|---|---|---|---|---|---|---|
| OH···N | 1.804 | 1.919 | - | - | - | - | 1.803 | - | 1.838 |
| NH···N | - | - | 1.771 | - | 1.946 | 1.919 | - | 2.059 | 2.049 |
| NH···O | - | - | - | 1.861 | - | 1.917 | - | - | - |
| OH/OH $SI_r$ | 3.835 | - | - | - | - | - | 3.739 | - | 3.530 |
| NH/NH $SI_r$ | - | - | 3.485 | 3.891 | 3.882 | 3.897 | - | 3.598 | 3.966 |
| OH/NH $SI_r$ | - | - | - | - | - | - | 3.460 | 4.069 | 3.709 |
| O/O $SI_r$ | - | - | 3.709 | 3.434 | - | 3.779 | - | - | - |
| N/N $SI_r$ | 3.384 | 3.005/3.700[b] | - | - | 3.634 | 3.745 | 3.459 | 3.789 | 3.615 |
| O/N $SI_r$ | - | - | - | 3.780 | 3.711 | 3.551 | - | - | - |
| OH/N $SI_a$ | 3.349 | 3.752 | - | - | - | - | - | - | - |
| NH/O $SI_a$ | - | - | 3.369 | | | | | | |
| C2–C2' | 4.093 | 4.868 | 4.135 | 4.125 | 4.384 | 4.881 | 4.142 | [c] | [c] |
| C9–C9' | 5.970 | 4.929 | 6.031 | 6.530 | 6.474 | 4.988 | 6.037 | [c] | [c] |
| $E_{int}$ | −72.5 | −59.6 | −103.3 | −44.0 | −50.0 | −42.8 | −50.8 | −26.8 | −34.5 |

a – molecules in the dimer are not coplanar (C9N1N8'C9'=51.9° (**EO/EO$^2$**); C2N1N8'C9'=37.1° (**EO/EO$^3$**); see Supplementary material), b – the N1–N1′ and N1–N8′ distances, respectively, c – since molecules in the dimer are not coplanar the C2–C2′ and C9–C9′ are not indicative and cannot be discussed

oxygen atoms (charge at OH oxygen is higher than that on C=O). It is worth noting that the charge at the H atom capable for hydrogen bonding is always lower in OH groups than in NH ones. This shows that the OH groups are able to form stronger hydrogen bonds in dimers. Figure 3 shows the double hydrogen-bonded dimers and HOMA values for rings.

The changes of HOMA index values upon formation of double hydrogen-bonded dimers are very small. In general, the HOMA increases in 1,2-quinone-like rings of which atoms participate in hydrogen bonding (**EE/EE** vs. **EE**, **OE/OE$^{1-3}$** vs. **OE**). On the other hand, the HOMA decreases in OH carrying rings that participate in hydrogen bonding (**OO/OO$^{1-2}$** vs. **OO**).

The geometric data are collected in Table 2 and these are divided into sets. These sets contain: the (a) hydrogen bond distance, (b) heavy-atom-to-heavy-atom distance ($SI_r$ interaction) within the quasi-ring (see Fig. 3; in red) formed by HBs (except for **OO/OO$^2$** where said quasi-ring is much larger), and (c) set describing other heavy-atom-to-heavy-atom distances out of the mentioned earlier quasi-ring (except for **OO/OO$^2$**). The last row collects energy of interaction ($E_{int}$) between molecules within the complex.

The shortest HB in the series of double hydrogen-bonded structures is observed in **EE/EE** dimer. The same structure has the highest $E_{int}$ value. The next dimer in the sense of interaction energy is **OO/OO$^1$**. Two HBs, and two $SI_a$s stabilize these two double HBed dimers. The $SI_a$ in **EE/EE** has the character of weak hydrogen bond (see latter in text). No other double HBed dimers are stabilized by $SI_a$ of weak HBing character. To investigate the effect of SIs on

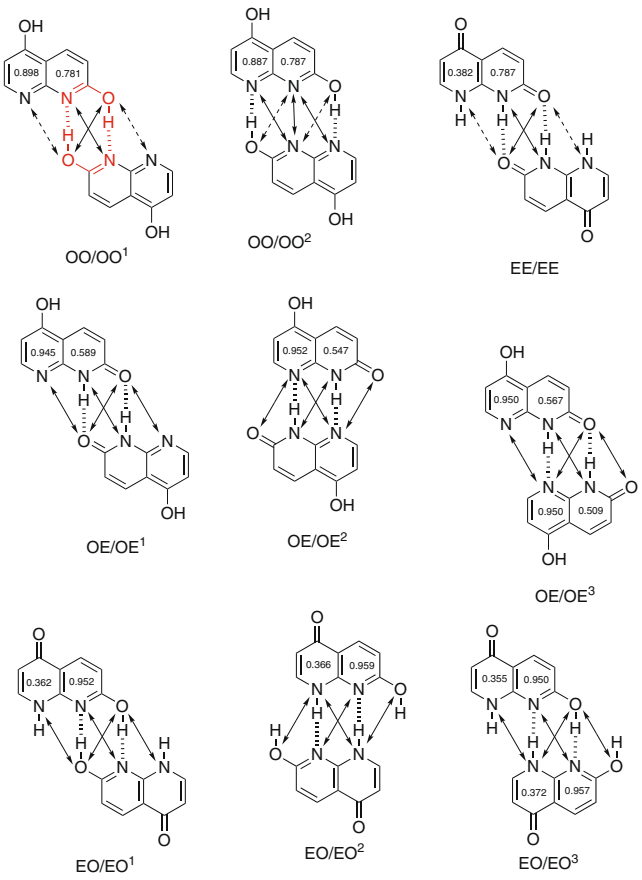



**Fig. 3** Double hydrogen-bonded dimers with $SI_r$ and $SI_a$ as solid and dashed arrows, respectively, and HOMA index values for heterocyclic rings

the structure and energy of interaction, the following dimers may be compared: **EE/EE** *vs.* **OE/OE$^1$** and **OO/OO$^1$** *vs.* **EO/EO$^1$**. It is easy to see that the change from $SI_a$ (in **EE/EE**) to $SI_r$ (in **OE/OE$^1$**) causes the shift of molecules, *i.e.*, the N1–N1′ distance increases while O11–O11′ decreases. From the comparison of **OO/OO$^1$** *vs.* **EO/EO$^1$** the same conclusion may be drawn but the geometrical effect is weaker. Also the changes in $E_{int}$ are higher in the **EE/EE** *vs.* **OE/OE$^1$** than in **OO/OO$^1$** *vs.* **EO/EO$^1$**. It is interesting to compare the **OE/OE$^2$** with **OE/OE$^3$**. Both structures are stabilized by two HBs and destabilized by four SIs but the difference is in the type of $SI_r$. Keeping in mind that $SI_r$s in **OE/OE$^2$** differs from **OE/OE$^3$** by replacement of one N/O $SI_r$ in the former by O/O $SI_r$ in the latter complex one can find the difference energy of interaction $\Delta E_{int}$ is equal to 8 kJ mol$^{-1}$.

The remaining dimers are twisted (in **EO/EO$^2$** and **EO/EO$^3$** two molecules do not occupy the common plane) or the direct lone-pair/lone-pair repulsion is present (**OO/OO$^2$**). Although, the strong repulsion is present in the last dimer, two $SI_a$s make this dimer relatively stable as compared to ones where lone-pair/lone-pair repulsion is absent.

The atomic charges in hydrogen bonded molecules are summarized in Table 3.

Charge at atoms involved in hydrogen bonding is influenced by the strength and type of this interaction. Thus, except for **OO/OO$^2$** and **EE/EE**, the charge at N1 is higher in double hydrogen-bonded associate than in respective monomeric species. For N8, the exceptions are **EE/EE** and **OE/OE$^1$**. The charge at O11 behaves irregularly from one type of associate to another, while charge at O12 is practically constant. Hydrogen bonding always lowers the electron charge at H atoms (except H8 and H11 in **EO/EO$^3$** and **EO/EO$^2$**, respectively).

To further study the properties of complexes, the density ($\rho$) and Laplacian of electron density ($\nabla^2\rho$) at hydrogen-bond critical point (H-BCP) according to the Quantum Theory of Atoms in Molecules (QTAIM) theory was calculated. These values are collected in Table 4.

The QTAIM-derived values show that the intermolecular hydrogen bonding is the major force that holds molecules together. There are, however, some additional interactions that influence the overall associate stabilization. These are weak CH···O interactions (in **OE/OE$^{2\text{-}3}$**) and the secondary attractive interaction ($SI_a$) in **EE/EE**. It is worth noting that $SI_a$ present in mentioned complex causes the molecules to get so close that the Laplacian of the electron density at NH···O BCP is higher than respective value in OH···N BCP in **OO/OO$^1$**. This is reflected also in hydrogen bonds lengths (Table 2). The CH···O weak interaction comes from the C–H bond polarization by the neighboring nitrogen atom. The properties of CH···O H-BCP are in line with criteria for weak interactions established by Koch and Popelier [74]. This type of interaction is met in complexes of pyridines with carboxylic acids [75]. It is also interesting that the QTAIM derived data for $SI_a$ in **EE/EE** (H8···O11') and CH···O (H7···O/N) interactions are very similar (Table 4).

Figure 4 and Table 5 collect the HOMA values, geometry of triple HBed associates and energy of interaction.

In triple hydrogen-bonded complexes, HOMA index behaves similarly as in double hydrogen-bonded ones. In general, the association increases the value of HOMA except for 2-hydroxypyridine ring.

In the triple HBed complexes, the effect of SIs is also easy to notice. The energy of interaction in **OO/EE** is much higher than in **OE/EO**. The presence of two $SI_a$s in **OO/EE** instead of two $SI_r$s in **OE/EO** brings molecules closer to each other. The lengths of HB bridges reflect this (all HBs in **OO/EE** are shorter that in **OE/EO**). It is necessary to keep in mind that, for example, the O/N $SI_r$ distance in **OO/EE** should be compared to OH/NH $SI_r$ in **OE/EO**. Thus, the presence of two $SI_a$s in **OO/EE** affects the remaining intermolecular distances in such a way as to compress the

**Table 3** NBO atomic charges in double hydrogen-bonded complexes

| | N1 | N8 | O11 | O12 | H1 | H8 | H11 |
|---|---|---|---|---|---|---|---|
| **OO/OO$^1$** | −0.580 | −0.501 | −0.704 | −0.697 | - | - | 0.547 |
| **OO/OO$^2$** | −0.517 | −0.539 | −0.724 | −0.696 | - | - | 0.553 |
| **EE/EE** | −0.638 | −0.601 | −0.715 | −0.614 | 0.472 | 0.474 | - |
| **OE/OE$^1$** | −0.631 | −0.508 | −0.664 | −0.695 | 0.492 | - | - |
| **OE/OE$^2$** | −0.643 | −0.557 | −0.633 | −0.692 | 0.476 | - | - |
| **OE/OE$^3$(t)** | −0.639 | −0.527 | −0.650 | −0.693 | 0.479 | - | - |
| **OE/OE$^3$(b)** | −0.640 | −0.560 | −0.609 | −0.694 | 0.488 | - | - |
| **EO/EO$^1$** | −0.617 | −0.598 | −0.733 | −0.614 | - | 0.459 | 0.532 |
| **EO/EO$^2$** | −0.607 | −0.614 | −0.700 | −0.614 | - | 0.465 | 0.524 |
| **EO/EO$^3$(t)** | −0.610 | −0.600 | −0.730 | −0.612 | - | 0.449 | 0.533 |
| **EO/EO$^3$(b)** | −0.610 | −0.613 | −0.696 | −0.614 | - | 0.467 | 0.530 |

"t" and "b" stand for *top* and *bottom* molecule, respectively (see Fig. 3)

**Table 4** The properties of H-BCP according to the QTAIM

| Complex type of interaction, $\rho$, $\nabla^2\rho$ | | |
|---|---|---|
| **OO/OO$^1$** | **OO/OO$^2$** | **EE/EE** |
| OH···N, 0.040, 0.092 | OH···N, 0.031, 0.071 | NH···O, 0.039, 0.112 |
| | | NH/O $SI_a$ (H8···O11'), 0.006 0.027 |
| **OE/OE$^1$** | **OE/OE$^2$** | **OE/OE$^3$** |
| NH···O, 0.030, 0.088 | NH···N, 0.030, 0.073 | NH···N, 0.032, 0.077 |
| | C–H7···O, 0.004, 0.015 | NH···O, 0.027, 0.077 |
| | | C–H7···N, 0.006, 0.016 |
| **EO/EO$^1$** | **EO/EO$^2$** | **EO/EO$^3$** |
| OH···N, 0.041, 0.096 | NH···N, 0.023, 0.060 | NH···N, 0.024, 0.061 |
| | | OH···N, 0.037, 0.089 |

N1–O11′ distance and to lengthen the N1'–O11 one. It is also worth mentioning that the N–H bond (N8′–H8′) is longer in the structure with higher $E_{int}$ value. The same is realized for O11–H11 *vs*. O11′–H11′ in **OO/EE** and **OE/EO**, respectively. The opposite behavior is observed for N1′–H1′ *vs*. N1–H1. Both double and triple hydrogen bonded structures presented here are quite stable and HB distances are small (1.65-1.77 Å). Thus, the intermolecular proton transfer may take place. The same behavior was studied before in 2-hydroxypyridine/2-[1*H*]-pyridone associates. It is worth mentioning that the triple proton transfer in the associates studied here is not possible. This reaction would lead to the ionic pair rather than to the neutral hydrogen bonded dimer. Schemes 1 and 2 show the TSs found for triple and double hydrogen-bonded structures, respectively, with QTAIM derived values and NBO atom charges.

Table 6 collects the chosen geometrical parameters for transition states and energy of activation ($E_a$).

Except the C9–C9′ distance in **TS2**, the C2–C2′ and C9–C9′ distances in TSs resulted from double hydrogen-bonded dimers (**TS2** and **TS3**) are shorter than respective values optimized for substrate and product of double proton transfer reaction. The C2–C2′ distances are shorter in transition state due to the shortening of hydrogen-bond bridges necessary to reach the TS. This is worth noting that the C9–C9′ distance in **TS2** is shorter by 0.54 Å than in **OE/OE$^1$**. On the other hand, the said distance is a little shorter in **OO/OO$^1$** than in **TS2**. This behavior tells that the OH/N $SI_a$ is the strongest in **OO/OO$^1$**, weakens in **TS2** and becomes O/N $SI_r$ in **OE/OE$^1$**. The C2–C2′ and C9–C9′ distances are very similar in **EO/EO$^1$** and **EE/EE**. It is even more interesting that the O/O $SI_r$ distance in **EE/EE** (Table 2) is similar to OH/OH $SI_r$ in **EO/EO$^1$**. The same is realized for NH/NH $SI_r$ (**EE/EE**) and N/N $SI_r$ (**EO/EO$^1$**). Thus, the different $E_{int}$ for the last two dimers does not affect the geometry of associate significantly. However, in

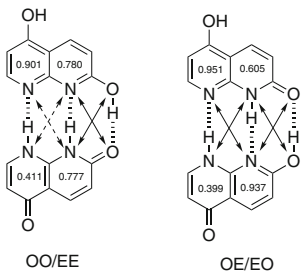


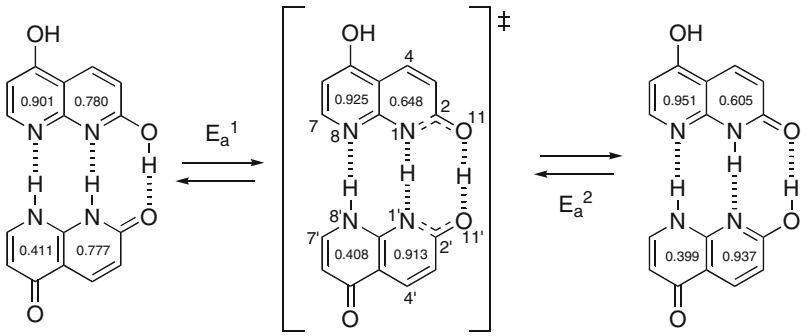

**Fig. 4** Triple hydrogen-bonded dimers, HOMA values (inside rings)

**Table 5** Chosen geometry parameters for triple hydrogen-bonded dimers and the energy of interaction ($E_{int}$ [kJ mol$^{-1}$])

| | **OO/EE** | **OE/EO** |
|---|---|---|
| OH···O | 1.653 | 1.764 |
| NH···N | 1.908, 1.976[a] | 1.960, 2.049[a] |
| NH/NH $SI_r$ | - | 3.788 |
| OH/NH $SI_r$ | 3.720 | 3.772 |
| N/N $SI_r$ | - | 3.822 |
| O/N $SI_r$ | 3.471 | 3.540 |
| NH/N $SI_a$ | 3.829, 3.687[b] | - |
| O11′–H11′ | - | 0.984 |
| N1–H1 | - | 1.029 |
| O11–H11 | 1.001 | - |
| N1′–H1′ | 1.027 | - |
| N8′–H8′ | 1.025 | 1.017 |
| N1–N1′ | 2.930 | 2.984 |
| N8–N8′ | 3.001 | 3.066 |
| O11–O11′ | 2.649 | 2.739 |
| C2–C2′ | 4.084 | 4.156 |
| C9–C9′ | 4.386 | 4.446 |
| $E_{int}$ | −111.6 | −66.6 |

a – N1···H1′ and N8···H8′, respectively, b – N1–N8′ and N8–N1′, respectively

**Scheme 1** Transition state found for triple hydrogen-bonded structure, HOMA values, $\rho$, $\nabla^2\rho$ calculated at H-BCP and NBO charges




**OO/EE** **TS1** **OE/EO**

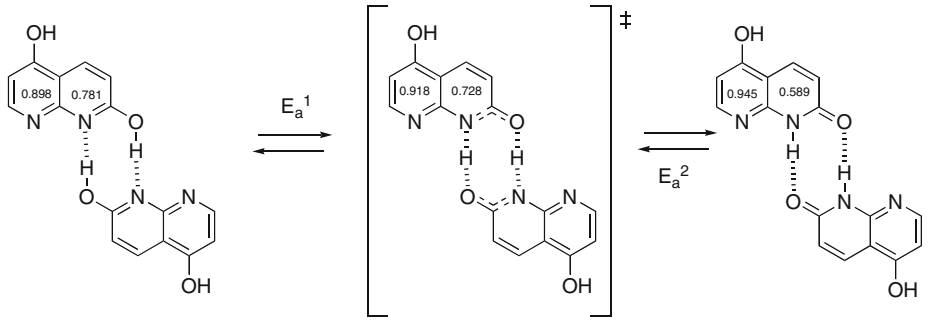
| | Property | | |
|---|---|---|---|
| | $\rho$, $\nabla^2\rho$ | $\rho$, $\nabla^2\rho$ | $\rho$, $\nabla^2\rho$ |
| H···N8 | 0.028, 0.068 | 0.042, 0.100 | 0.024, 0.061 |
| H–N8' | 0.335, -1.948 | 0.329, -1.904 | 0.343, -1.996 |
| H···N1/H–N1 | 0.033, 0.080 | 0.116, -0.066 | 0.331, -1.996 |
| H···N1'/H–N1' | 0.334, -1.940 | 0.182, -0.590 | 0.030, 0.070 |
| H···O11/H–O11 | 0.322, -1.987 | 0.092, 0.135 | 0.037, 0.111 |
| H···O11'/H–O11' | 0.048, 0.142 | 0.268, -1.354 | 0.344, -2.178 |
| | charge | charge | charge |
| N1 | -0.588 | -0.645 | -0.636 |
| N1' | -0.638 | -0.626 | -0.614 |
| N8 | -0.551 | -0.569 | -0.552 |
| N8' | -0.615 | -0.616 | -0.611 |
| O11 | -0.707 | -0.706 | -0.656 |
| O11' | -0.696 | -0.705 | -0.715 |

**TS3** molecules are much closer to one another. This, together with the low barrier between **EO/EO**[1] and **TS3**, makes the **EE/EE** form preferred. The similar low barrier was found between **OO/OO**[1] and **TS2** but the $E_{int}$ for **OE/OE**[1] is rather low when compared to other dimers. The above-mentioned observations lead to a conclusion that if only **EE** tautomeric form is present, the **EE/EE** double hydrogen-bonded dimer would be a major form or it would form triple hydrogen-bonded associate with **OO**.

The triple hydrogen-bonded structures are, in general, more stable that double hydrogen-bonded ones. Four tautomeric forms with relative low energy of constituents that may be present in the studied molecules can form two triple hydrogen-bonded associates. These may be in dynamic equilibrium exchanging two protons *via* **TS1**. As observed previously for **TS3**, both C2–C2′ and C9–C9′ are shortened in TS. Decrease of intermolecular distances in **TS1** is reflected also by other parameters collected in Table 6. The most crucial are: N1–N1′, N8–N8′, O11–O11′ and the hydrogen bond N8···H8′ (1.8 Å). The energetic and geometric data shows that the third hydrogen bond supports the double proton transfer in triple HBed structure.

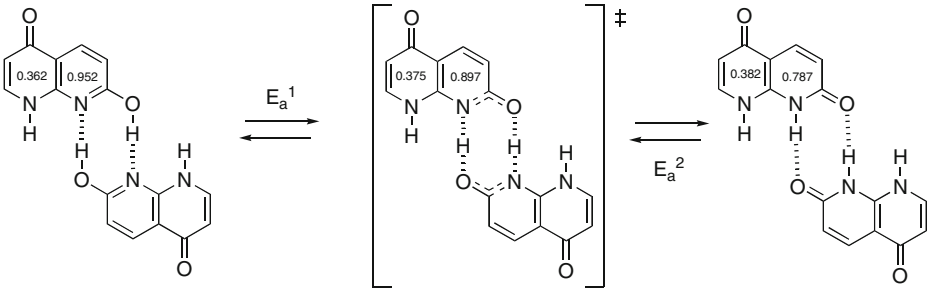
The character of double proton transfer may be studied by QTAIM. This methodology says that for shared type of interaction (covalent bond), the Laplacian of electron density at BCP is negative and for interactions of hydrogen-bonding type, it is positive. Thus, it is interesting to analyze how these values behave in optimized dimers and related transition states. The most interesting in this study is the **OO/EE**, **TS1**, **OE/EO** equilibrium. The values of electron density at H-BCP show that interaction between protons and basic center in the molecule of the counterpart are higher for H···O interaction than for H···N one. It is also important that said values are even higher in the transition state. The $\nabla^2\rho$ delivers information about the character of interaction in dimers and transition state. All hydrogen bonds present in the stable complexes are characterized correctly by positive values of Laplacian of electron density. In the transition state (**TS1**), the proton transferred between oxygen atoms is bonded to O11′ covalently (the O11′–H bond is slightly longer than in **OE/EO**, Tables 5 and 6) and forms a hydrogen bond with O11. On the other hand, the proton exchanged between nitrogen atoms is shared between both of them (both values of Laplacian at

**Scheme 2** Transition states in double hydrogen-bonded dimers




**OO/OO$^1$** **TS2** **OE/OE$^1$**

| | $\rho$, $\nabla^2\rho$ | $\rho$, $\nabla^2\rho$ | $\rho$, $\nabla^2\rho$ |
|---|---|---|---|
| H···N1/N1' or H–N1/N1' | 0.040, 0.092 | 0.148, -0.291 | 0.333, -1.970 |
| H···O11/O11' or H–O11/O11' | 0.329, -2.034 | 0.178, -0.503 | 0.030, 0.088 |
| | NBO charge | NBO charge | NBO charge |
| N1/N1' | -0.580 | -0.626 | -0.631 |
| N8/N8' | -0.501 | -0.504 | -0.508 |
| O11/O11' | -0.704 | -0.696 | -0.664 |




**EO/EO$^1$** **TS3** **EE/EE**

| | $\rho$, $\nabla^2\rho$ | $\rho$, $\nabla^2\rho$ | $\rho$, $\nabla^2\rho$ |
|---|---|---|---|
| H···N1/N1' or H–N1/N1' | 0.041, 0.096 | 0.182, -0.576 | 0.332, -1.929 |
| H···O11/O11' or H–O11/O11' | 0.336, -2.056 | 0.152, -0.235 | 0.039, 0.112 |
| $SI_a$ | - | - | 0.006, 0.027 |
| | NBO charge | NBO charge | NBO charge |
| N1/N1' | -0.617 | -0.644 | -0.638 |
| N8/N8' | -0.598 | -0.601 | -0.601 |
| O11/O11' | -0.733 | -0.735 | -0.715 |

**Table 6** The chosen geometry parameters for transition states and the energy of activation[a] ($E_a$ [kJ mol$^{-1}$])

| | TS1 | TS2 | TS3 |
|---|---|---|---|
| N1···H | 1.409 | 1.304 | 1.228 |
| N1′···H | 1.234 | 1.304 | 1.228 |
| O11···H | 1.412 | 1.192 | 1.257 |
| O11′···H | 1.052 | 1.192 | 1.257 |
| N8···H8′ | 1.803 | - | - |
| N1–N8′ | 3.554 | 5.118 | 5.037 |
| N1–O11′ | 3.485 | 2.495 | 2.483 |
| N1′–N8 | 3.615 | 5.118 | 5.037 |
| N1′–O11 | 3.366 | 2.495 | 2.483 |
| N1–N1′ | 2.642 | 3.337 | 3.266 |
| N8–N8′ | 2.834 | 7.197 | 7.127 |
| O11–O11′ | 2.461 | 3.427 | 3.459 |
| C2–C2′ | 3.835 | 3.772 | 3.822 |
| C9–C9′ | 4.156 | 5.989 | 5.873 |
| $E_a^1/E_a^2$ | 16.9, 27.1 | 1.0, 50.8 | 5.8, 24.3 |

a – two numbers for $E_a$ represent the energy of TS related to the energy of the substrate and the product (see Schemes 1 and 2). For graphic representation see Supplementary Material

H-BCP of H···N1 and H···N1′ are negative) but not equally. This observation resembles the strong, symmetrical hydrogen bond found in other complexes [76, 77]. The said proton is much more strongly bonded to the N1′ than to N1 resembling the **OO/EE**-like associate (compare the H to N distances in Tables 5 and 6). This behavior is understood in the sense of difference of atomic charge located at N1 and N1′ in **OO/EE** vs. **OE/EO** complexes and is in agreement with Leffler-Hammond postulate [78] which says that the TS is closer structurally to the structure with lower energy in exothermic reaction. This difference is much higher in **OO/EE** than in **OE/EO**. A similar situation is realized in explaining why the proton lying between oxygens is located much closer to O11' than O11 in the transition state. The above mentioned facts suggest that the double proton transfer is not simultaneous, but the O11–H···O11′ to O11···H–O11′ reaction initiates the whole process, while central proton (H1) is transferred. It is also important to say that the H···N8 hydrogen bond is stronger in the transition state than in the stable complexes. This hydrogen bond supports the proton transfer reaction.

The remaining two double proton transfer reactions are also interesting. In the **OO/OO$^1$**, **TS2**, **OE/OE$^1$** equilibrium, the transferred protons are shared between nitrogen and oxygen atoms (negative Laplacian). This is perfectly understood if one checks the geometry of the hydrogen bond bridge, *i.e.*, the H···O11/O11' distance is shorter than H···N1/N1′ one. The energy data (Table 2) also may explain why the transition state resembles **OO/OO$^1$** and why the proton is located closer to the oxygen than nitrogen. The energy of interaction is higher in **OO/OO$^1$** than in **OE/OE$^1$**, thus if the interaction is stronger, the energy of activation of the proton transfer reaction is easier to reach (Table 6). Thus, only a small O11–H/O11′–H bond elongation in associated molecules is responsible for proton transfer. The atomic charges (Scheme 2) in **TS2** also suggest that the mobile proton should be closer to the oxygen than to nitrogen.

The last transition state studied is **TS3**. Basically, it is the same OH···N to O···HN reaction as in **TS2**, but it differs from **TS2** by the presence of the proton at neighboring N8. This yields in additional stabilization of **EE/EE** complex by weak $SI_a$ interaction (Table 4). It is worth noting that the mobile proton exhibits shared type of interaction with nitrogen and oxygen as in **TS2**. The difference between **TS2** and **TS3** is that the values of Laplacian of electron density at H-BCP are reversed. Thus, the shared character of H-to-N interaction is higher than H-to-O, while in **TS2**, H-to-O interaction is stronger. Again, the atomic charges are in agreement with the position of mobile proton with respect to nitrogen and oxygen atoms.

Comparison between **TS1-3** shows that not only may the hydrogen bond support the proton transfer reaction, but this may also be supported by weak secondary interaction.

When the relative dimer energy ($E_{rel,dim}$) cut-off at the 20 kJ mol$^{-1}$ is set (**EE/EE**; $E_{rel,dim}$=0), structures that are below this energy level are **OO/EE** ($E_{rel,dim}$=6.0 kJ mol$^{-1}$), **OE/EO** ($E_{rel,dim}$=17.4 kJ mol$^{-1}$) and **EO/EO$^1$** ($E_{rel,dim}$= 19.0 kJ mol$^{-1}$), the last two being double hydrogen-bonded dimers. In the mentioned dimers, the double proton transfer may be observed with relatively low $E_a$. It is worth mentioning that in the above associates, the most abundant tautomeric form is **EE**. That shows that although **EE** form has a higher energy than the most stable **EO** (Table 1), it is still possible to observe it due to the intermolecular stabilization. Moreover, in all dimers below the 20 kJ mol$^{-1}$ barrier, at least two of heterocyclic rings lose the aromatic character and in the most stable dimer (**EE/EE**), all rings are not aromatic.

## Conclusions

The tautomeric preferences in heterocyclic systems may be crucial for intermolecular interactions and proton transfer reaction. The studied 2,5-dihydroxy-1,8-naphthyridine proves that. The prototropic equilibria, both intra- and intermolecular, influence the intermolecular interactions and the value of activation energy needed for intermolecular proton transfer. This proton transfer reaction may be supported by a third hydrogen bond or by weak secondary interactions. In the triple hydrogen bonded complex, the

only proton transfer that is possible is a double proton transfer. Data delivered by QTAIM methodology shows that such a proton transfer is not simultaneous, but the first step of this reaction takes place between atoms that form stronger hydrogen bond. Results show that, due to nearness of energy, a mixture of hydrogen-bonded dimers might be present in 2,5-dihydroxy-1,8-naphthyridine.

**Acknowledgments** Financial support from the Polish Ministry of Science and Higher Education (grant no. N N204 174138) is gratefully acknowledged. The authors are very much indebted to the ICM in Warsaw for providing computer time and programs.

## References

1. Desiraju GR (2011) Reflections on the hydrogen bond in crystal engineering. Cryst Growth Des 11:896–898. doi:10.1021/cg200100m
2. Desiraju GR (2002) Hydrogen bridges in crystal engineering: Interactions without borders. Acc Chem Res 35:565–573. doi:10.1021/ar010054t
3. Desiraju GR (1995) Supramolecular synthons in crystal engineering. A new organic synthesis. Angew Chem Int Ed 34:2311–2327
4. Moulton B, Zaworotko MJ (2001) From molecules to crystal engineering: Supramolecular isomerism and polymorphism in network solids. Chem Rev 101:1629–1658. doi:10.1021/cr9900432
5. Eckert J, Barthes M, Klooster WT, Albinati A, Aznar R, Koetzle TF (2000) No evidence for proton transfer along the N-H…O hydrogen bond in N-methylacetamide: Neutron single crystal structure at 250 and 276 k. J Phys Chem B 105:19–24. doi:10.1021/jp000267k
6. Katz JL, Post B (1960) The crystal structure and polymorphisc of n-methylacetamide. Acta Cryst 13:624–628
7. Adarsh NN, Sahoo P, Dastidar P (2010) Is a crystal engineering approach useful in designing metallogels? A case study. Cryst Growth Des 10:4976–4986. doi:10.1021/cg101078f
8. Paisner K, Zakharov LN, Doxsee KM A robust thiourea synthon for crystal engineering. Cryst Growth Des 10:3757–3762. doi:10.1021/cg100589n
9. Burrows AD (2004) Crystal engineering using multiple hydrogen bonds. In: Mingos DMP (ed) Supramolecular assembly via hydrogen bonds. Springer, Berlin / Heidelberg, pp 55–96
10. Smith G, Lynch D, Byriel K, Kennard C (1997) The utility of 4-aminobenzoic acid in promotion of hydrogen bonding in crystallization processes: The structures of the cocrystals with halo and nitro subsituted aromatic compounds, and the crystal structures of the adducts with 4-nitroaniline (1:1), 4-(4-nitrobenzyl)pyridine (1:2), and (4-nitrophenyl)acetic acid (1:1). J Chem Crystallogr 27:307–317. doi:10.1007/bf02575979
11. Auffinger P, Hays FA, Westhof E, Ho PS (2004) Halogen bonds in biological molecules. Proc Natl Acad Sci USA 101:16789–16794. doi:10.1073/pnas.0407607101
12. Aakeröy CB, Schultheiss NC, Rajbanshi A, Desper J, Moore C (2008) Supramolecular synthesis based on a combination of hydrogen and halogen bonds. Cryst Growth Des 9:432–441. doi:10.1021/cg8006712
13. Politzer P, Murray JS, Clark T (2010) Halogen bonding: An electrostatically-driven highly directional noncovalent interaction. Phys Chem Chem Phys 12:7748–7757
14. Prasang C, Nguyen HL, Horton PN, Whitwood AC, Bruce DW (2008) Trimeric liquid crystals assembled using both hydrogen and halogen bonding. Chem Comm:6164–6166
15. Rissanen K (2008) Halogen bonded supramolecular complexes and networks. Cryst Eng Comm 10:1107–1113
16. Ritter SK (2009) Halogen bonding begins to fly. Chem Eng News 87:39–42
17. Voth AR, Khuu P, Oishi K, Ho PS (2009) Halogen bonds as orthogonal molecular interactions to hydrogen bonds. Nature 1:74–79
18. Berl V, Krische MJ, Huc I, Lehn JM, Schmutz M (2000) Template-induced and molecular recognition directed hierarchical generation of supramolecular assemblies from molecular strands. Chem Eur J 6:1938–1946
19. van Beek DJM, Spiering AJH, Peters GWM, te Nijenhuis K, Sijbesma RP (2007) Unidirectional dimerization and stacking of ureidopyrimidinone end groups in polycaprolactone supramolecular polymers. Macromolecules 40:8464–8475. doi:10.1021/ma0712394
20. Richards FM, Kundrot CE (1988) Identification of structural motifs from protein coordinate data: Secondary structure and first-level supersecondary structure. Proteins Struct Funct Genet 3:71–84
21. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637
22. Watson JD, Crick FHC (1953) Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. Nature 171:737–738
23. Watson JD, Crick FHC (1953) Genetical implications of the structure of deoxyribonucleic acid. Nature 171:964–967
24. Jeffery GA, Saenger W (1994) Hydrogen bonding in proteins. In: Hydrogen bonding in biological structures. Springer-Verlag, Berlin, pp 351–394
25. de Groot BL, Daura X, Mark AE, Grubmüller H (2001) Essential dynamics of reversible peptide folding: Memory-free conformational dynamics governed by internal hydrogen bonds. J Mol Biol 309:299–313
26. Daura X, van Gunsteren WF, Rigo D, Jaun B, Seebach D (1997) Studying the stability of a helical β-heptapeptide by molecular dynamics simulations. Chem Eur J 3:1410–1417. doi:10.1002/chem.19970030907
27. Daura X, Jaun B, Seebach D, van Gunsteren WF, Mark AE (1998) Reversible peptide folding in solution by molecular dynamics simulation. J Mol Biol 280:925–932
28. Suárez M, Lehn J-M, Zimmerman SC, Skoulios A, Heinrich B (1998) Supramolecular liquid crystals. Self-assembly of a trimeric supramolecular disk and its self-organization into a columnar discotic mesophase. J Am Chem Soc 120:9526–9532
29. Ośmiałowski B, Kolehmainen E, Dobosz R, Gawinecki R, Kauppinen R, Valkonen A, Koivukorpi J, Rissanen K (2010) Self-organization of 2-acylaminopyridines in the solid state and in solution. J Phys Chem A 114:10421–10426
30. Lehn J-M (2002) Toward self-organization and complex matter. Science 295:2400–2403. doi:10.1126/science.1071063
31. Kuykendall DW, Anderson CA, Zimmerman SC (2009) Hydrogen-bonded DeUG–DAN heterocomplex: Structure and stability and a scalable synthesis of deug with reactive functionality. Org Lett 11:61–64. doi:10.1021/ol802344w
32. Hisamatsu Y, Fukumi Y, Shirai N, S-i I, Odashima K (2008) Five-membered heterocyclic ureas suitable for the donor-donor-acceptor hydrogen-bonding modules. Tetrahedron Lett 49:2005–2009
33. Hisamatsu Y, Shirai N, Ikeda S-I, Odashima K (2010) A new quadruple hydrogen-bonding module based on five-membered heterocyclic urea structure. Org Lett 12:1776–1779. doi:10.1021/ol100385b
34. Ong HC, Zimmerman SC (2006) Higher affinity quadruply hydrogen-bonded complexation with 7-deazaguanine urea. Org Lett 8:1589–1592. doi:10.1021/ol0601803

35. Mayer MF, Nakashima S, Zimmerman SC (2005) Synthesis of a soluble ureido-naphthyridine oligomer that self-associates via eight contiguous hydrogen bonds. Org Lett 7:3005–3008
36. Jorgensen WL, Pranata J (1990) Importance of secondary interactions in triply hydrogen bonded complexes: Guanine-cytosine vs uracil-2,6-diaminopyridine. J Am Chem Soc 112:2008–2010
37. Ośmiałowski B, Dobosz R The influence of secondary interactions on complex stability and double proton transfer reaction in 2-[1H]-pyridone/2-hydroxypyridine dimers. J Mol Model. doi:10.1007/s00894-010-0934-y
38. Ośmiałowski B (2009) Systematic investigation of 2,7-dihydroxy-1,8-naphthyridine dimerization - secondary interactions and tautomeric preferences calculations. J Mol Struct THEOCHEM 908:92–101
39. Lüning U, Kühl C, Uphoff A (2002) Four hydrogen bonds - DDAA, DADA, DAAD and ADDA hydrogen bond motifs. Eur J Org Chem 2002:4063–4070
40. Szyc Ł, Guo J, Yang M, Dreyer J, Tolstoy PM, Nibbering ETJ, Czarnik-Matusewicz B, Elsaesser T, Limbach H-H (2010) The hydrogen-bonded 2-pyridone dimer model system. 1. Combined NMR and FT-IR spectroscopy study. J Phys Chem A 114:7749–7760. doi:10.1021/jp103630w
41. Yang M, Szyc Ł, Dreyer J, Nibbering ETJ, Elsaesser T (2010) The hydrogen-bonded 2-pyridone dimer model system. 2. Femtosecond mid-infrared pump-probe study. J Phys Chem A 114:12195–12201. doi:10.1021/jp108096y
42. Müller A, Losada M, Leutwyler S (2004) Ab initio benchmark study of (2-pyridone)$_2$, a strongly bound doubly hydrogen-bonded dimer. J Phys Chem A 108:157–165
43. Alvares-Rua C, Garcia-Granda S, Goswami S, Mukherjee R, Dey S, Claramunt RM, Santa Maria MD, Rozas I, Jagerovic N, Alkorta I, Elguero J (2004) Multiple hydrogen bonds and tautomerism in naphthyridine derivatives. New J Chem 28:700–707
44. Newman SG, Taylor A, Boyd RJ (2008) Factors controlling extremely strong AAA-DDD triply hydrogen-bonded complexes. Chem Phys Lett 450:210–213
45. Blight BA, Camara-Campos A, Djurdjevic S, Kaller M, Leigh DA, McMillan FM, McNab H, Slawin AMZ (2009) AAA-DDD triple hydrogen bond complexes. J Am Chem Soc 131:14116–14122. doi:10.1021/ja906061v
46. Ośmiałowski B, Kolehmainen E, Gawinecki R, Dobosz R, Kauppinen R (2010) Complexation of 2,6-bis(acylamino)pyridines with dipyridin-2-ylamine and 4,4-dimethylpiperidine-2,6-dione. J Phys Chem A 114:12881–12887
47. Ośmiałowski B, Kolehmainen E, Gawinecki R, Kauppinen R, Koivukorpi J, Valkonen A (2010) Nmr and quantum chemical studies on association of 2,6-bis(acylamino)pyridines with selected imides and 2,2'-dipyridylamine. Struct Chem 21:1061–1067. doi:10.1007/s11224-010-9646-2
48. Chien C-H, Leung M-k SuJ-K, Li G-H, Liu Y-H, Wang Y (2004) Substituent effects on pyrid-2-yl ureas toward intramolecular hydrogen bonding and cytosine complexation. J Org Chem 69:1866–1871. doi:10.1021/jo0355808
49. Sartorius J, Schneider H-J (1996) A general scheme based on empirical increments for the prediction of hydrogen-bond associations of nucleobases and of synthetic host-guest complexes. Chem Eur J 2:1446–1452
50. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JA, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C, Pople JA (2004) Gaussian 03, Revision E.01. Gaussian Inc, Pittsburgh PA
51. Zhao Y, Schultz NE, Truhlar DG (2005) Exchange-correlation functional with broad accuracy for metallic and nonmetallic compounds, kinetics, and noncovalent interactions. J Chem Phys 123:161103–161104
52. Zhao Y, Truhlar DG (2006) Assessment of model chemistries for noncovalent interactions. J Chem Theor Comput 2:1009–1018
53. Zhao Y, Truhlar DG (2008) Exploring the limit of accuracy of the global hybrid meta density functional for main-group thermochemistry, kinetics, and noncovalent interactions. J Chem Theor Comput 4:1849–1868. doi:10.1021/ct800246v
54. Zhao Y, Truhlar DG (2008) The m06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: Two new functionals and systematic testing of four m06-class functionals and 12 other functionals. Theor Chem Acc 120:215–241
55. Ditchfield R, Hehre WJ, Pople JA (1971) Self-consistent molecular-orbital methods. Ix. An extended gaussian-type basis for molecular-orbital studies of organic molecules. J Chem Phys 54:724–728
56. Hehre WJ, Ditchfield R, Pople JA (1972) Self-consistent molecular orbital methods. XII. Further extensions of gaussian-type basis sets for use in molecular orbital studies of organic molecules. J Chem Phys 56:2257–2261
57. Francl MM, Pietro WJ, Hehre WJ, Binkley JS, Gordon MS, DeFrees DJ, Pople JA (1982) Self-consistent molecular orbital methods. XXIII. A polarization-type basis set for second-row elements. J Chem Phys 77:3654–3665
58. Peng C, Schlegel HB (1993) Combining synchronous transit and quasi-newton methods to find transition states. Isr J Chem 33:449–454
59. Boys SF, Bernardi F (1970) The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors. Mol Phys 19:553–566. doi:10.1080/00268977000101561
60. Simon S, Duran M, Dannenberg JJ (1996) How does basis set superposition error change the potential surfaces for hydrogen-bonded dimers? J Chem Phys 105:11024–11031
61. Foster JP, Weinhold F (1980) Natural hybrid orbitals. J Am Chem Soc 102:7211–7218. doi:10.1021/ja00544a007
62. Reed AE, Weinhold F (1983) Natural bond orbital analysis of near-hartree–fock water dimer. J Chem Phys 78:4066–4073
63. Biegler-König F, Schönbohm J, Bayles D (2001) AIM2000. J Comp Chem 22:545–559. doi:10.1002/1096-987x(20010415)22:5<545::aid-jcc1027>3.0.co;2-y
64. Kruszewski J, Krygowski TM (1972) Definition of aromaticity basing on the harmonic oscillator model. Tetrahedron Lett 13:3839–3842
65. Krygowski TM (1993) Crystallographic studies of inter- and intramolecular interactions reflected in aromatic character of π-electron systems. J Chem Inf Comput Sci 33:70–78
66. Kolehmainen E, Ośmiałowski B, Krygowski TM, Kauppinen R, Nissinen M, Gawinecki R (2000) J Chem Soc Perkin Trans 2:1259–1266
67. Kolehmainen E, Ośmiałowski B, Nissinen M, Kauppinen R, Gawinecki R (2000) J Chem Soc Perkin Trans 2:2185–2191

68. Gawinecki R, Kolehmainen E, Loghmani-Khouzani H, Ośmiałowski B, Lovasz T, Rosa P (2006) Effect of π-electron delocalization on tautomeric equilibria - benzoannulated 2-phenacylpyridines. Eur J Org Chem 2006:2817–2824
69. Ośmiałowski B, Krygowski TM, Dominikowska J, Palusiak M (2011) The effect of benzoannulation on the transition state and the proton transfer equilibrium in di(2-pyridyl)methane derivatives. New J Chem. doi:10.1039/C1NJ20108E
70. Gawinecki R, Kuczek A, Kolehmainen E, Ośmiałowski B, Krygowski TM, Kauppinen R (2007) Influence of bond fixation in benzo-annulated n-salicylideneanilines and their ortho-C(O)X derivatives (x=$CH_3$, $NH_2$, $OCH_3$) on tautomeric equilibria in solution. J Org Chem 72:5598–5607. doi:10.1021/jo070454f
71. Krygowski TM, Zachara JE, Ośmiałowski B, Gawinecki R (2006) Topology-driven physicochemical properties of π-electron systems. 1. Does the clar rule work in cyclic pi-electron systems with the intramolecular hydrogen or lithium bond? J Org Chem 71:7678–7682
72. Ośmiałowski B, Kolehmainen E, Nissinen M, Krygowski TM, Gawinecki R (2002) (1Z,3Z)-1,4-di(pyridin-2-yl)buta-1,3-diene-2,3-diol: The planar highly conjugated symmetrical enediol with multiple intramolecular hydrogen bonds. J Org Chem 67:3339–3345
73. Ośmiałowski B, Kolehmainen E, Gawinecki R (2003) Identity double-proton transfer in (3Z)-3-hydroxy-1,4-di(quinolin-2-yl)-but-3-en-2-one. Chem Eur J 9:2710–2716
74. Koch U, Popelier PLA (1995) Characterization of c-h-o hydrogen bonds on the basis of the charge density. J Phys Chem 99:9747–9754. doi:10.1021/j100024a016
75. Sugiyama T, Meng J, Matsuura T (2002) Intermolecular interactions in the formation of two-component molecular crystals composed of chloronitrobenzoic acids and 4-benzoylpyridine. J Mol Struct 611:53–64
76. Majerz I, Olovsson I (2010) Influence of proton transfer on the geometry of the donor and acceptor in $NHN^+$ hydrogen bonds. J Mol Struct 976:11–18
77. Marx D (2006) Proton transfer 200 years after von Grotthuss: Insights from ab initio simulations. Chem Phys Chem 7:1848–1870. doi:10.1002/cphc.200600128
78. Leffler JE (1953) Parameters for the description of transition states. Science 117:340–341

J Mol Model (2012) 18:1645–1659
DOI 10.1007/s00894-011-1161-x

ORIGINAL PAPER

# Theoretical studies on models of lysine-arginine cross-links derived from α-oxoaldehydes: a new mechanism for glucosepane formation

**Rasoul Nasiri · Mansour Zahedi · Hélène Jamet · Ali Akbar Moosavi-Movahedi**

Received: 28 April 2011 / Accepted: 21 June 2011 / Published online: 3 August 2011


**Abstract** Availability and high reactivity of α-oxoaldehydes have been approved by experimental techniques not only in vivo systems but also in foodstuffs. In this article we re-examine the mechanism of glucosepane formation by using computational model chemistry. Density functional theory has been applied to propose a new mechanism for glucosepane formation through reaction of α-oxoaldehydes with methyl amine (MA) and methyl guanidine (MGU) models of lysine and arginine residues respectively. This non enzymatic process can be described in three main steps: (1) Schiff base formation from methyl amine, methyl glyoxal (MGO) (2) addition of methyl guanidine and (3) addition of glyceraldehyde. We show that this process is thermodynamically possible and presents a rate-determining step with a reasonable free energy barrier equal to 37.8 kcal $mol^{-1}$ in water solvent. Comparisons were done with the mechanism formation of GODIC (glyoxal-derived imidazolium cross-link) and MODIC (methyl glyoxal-derived imidazolium cross-link), two other important cross-links in vivo.

**Keywords** α-oxoaldehydes · Cross-link · Density functional theory · Glucosepane · GODIC · MODIC · Physiological glucose levels

**Electronic supplementary material** The online version of this article (doi:10.1007/s00894-011-1161-x) contains supplementary material, which is available to authorized users.

R. Nasiri · M. Zahedi (✉)
Department of Chemistry, Faculty of Sciences,
Shahid Beheshti University,
G.C., Evin,
19839-63113, Tehran, Iran
e-mail: mansourzahedi@yahoo.com
e-mail: m-zahedi@sbu.ac.ir

R. Nasiri · H. Jamet
Département de Chimie Moléculaire, Chimie Théorique,
UMR-5250, Université Joseph Fourier,
BP 53, 38041, Grenoble Codex 9, France

A. A. Moosavi-Movahedi
Institute of Biochemistry and Biophysics, University of Tehran,
Tehran, Iran

## Introduction

The chemical modification and cross-linking of proteins by the Maillard reaction is the results of a cascade of reactions. Condensations, rearrangements, fragmentation and oxidation processes of reducing carbohydrates (glucose, ribose, fructose) or α-oxoaldehydes (methyl glyoxal (MGO), glyoxal (GO), 3-deoxyglucosone (3-DG) and glyceraldehyde (GLA)) occur with amino groups in lysine residues and guanidine groups in arginine residues [5, 7, 9, 25, 29, 33, 34, 46]. The mechanism of the glycation reaction was firstly tested by Maillard at the beginning of the twentieth century by incubation of amino acids with different sugars [37] and can be described at three stages. The early stage of reaction involves initiation of glycation through amino-carbonyl interaction leading to early glycation products, which contain an intact sugar moiety. In the intermediate stage, a variety of carbonyl compounds, which are generally more reactive than the original amino acids, are formed. Finally, successive rearrangements lead irreversibly to the formation of a variety of heterogeneous structures, known as the Advanced Glycation End-products (AGEs).

AGEs accumulate in body and have been implicated in the development of many pathophysiologies associated with ageing in general as well as with complications of diabetes, atherosclerosis and renal insufficiency [8, 56].

They are implicated in pathogenesis of various devastating disorders, such as coronary heart diseases [26], cancer growth [11] and metastasis, insulin resistance [56] and non alcoholic fatty liver disease [36]. They have also been shown to have a role in neuro degenerative disease such as Alzheimer's and Parkinson's diseases [22, 48].

According to the type of carbonyl compounds that can be a reducing agent in glycation reaction, six distinct classes of AGEs have been found in serum. They are glucose-derived AGEs (GLC-AGEs), MGO-derived AGEs (MGO-AGEs), GO-derived AGEs (GO-AGEs), 3-DG-derived AGEs (3-DG-AGEs), glyceraldehyde-derived AGEs (Glycer-AGEs) and glycolaldehyde-derived AGEs (Glycol-AGEs) [42, 49]. On the other hand, Ahmed et al. classified AGEs to the first four in aforementioned ones [1, 2]. Interestingly, clinical studies have shown that Glycer-AGEs are more important than GLC-AGEs in cardiovascular disease (CVD) and nonalcoholic steatohepatitis (NASH) in diabetes [31, 32].

An extense variety of AGEs has been discovered in vitro and in vivo systems; for example, glucosepane, DODIC, GOLD, MOLD, MODIC and GODIC, etc. These cross-links have also been identified in foodstuffs [4]. It is the case of the glucosepane, recognized as the major cross-link in vivo under non oxidative conditions [6, 46]. To explain its formation, Biemel et al. proposed a mechanism with lysine linked α-dicarbonyl compound as intermediate [7]. This mechanism involves formation of a dideoxyosone from glucose and lysine; dideoxyosone then undergoes cyclization into aldimine, which then reacts with the side chain of arginine to yield glucosepane (see Scheme 1).

Experimental studies have previously suggested that AGEs can arise not only from reducing sugars, but also from α-oxoaldehyde compounds obtained from the degradation of sugars [53–55, 58], lipids [3], Schiff base intermediate or Amadori product [38] in buffered solution at 37 °C. Decomposition of amadori products has been illustrated in Scheme 2. The α-oxoaldehyde compounds found in vivo may have a dietary origin [27, 35]. In the glycation of human serum albumin (HSA) under physicological conditions, separation of glucose from HSA [43][1] and its degradation to the α–oxoaldehydes (such as MGO, GO, 3-DG and GLA) [51], imply another possibility for glucosepane formation. It has been found that there are three pathways for the in vivo formation of glyceraldehydes: namely, glycolytic pathway, polyol pathway, and fructose metabolic pathway [42] and on the other hand, it plays a key role in Maillard cross linking of food proteins [17–19, 57].

According to these results which show availability of the α–oxoaldehydes, a new mechanism for the formation of glucosepane is re-examined by using computational model chemistry. This article is devoted to describe a novel mechanism for glucosepane in vivo and food systems. Computational model chemistry will be used to understand and establish proper reaction mechanism. It can give molecular level insights into a reaction through determination of intermediates and transition states structures and energies. The mechanism considers reactions between the α-oxoaldehydes MGO and GLA with lysine and arginine amino acids.

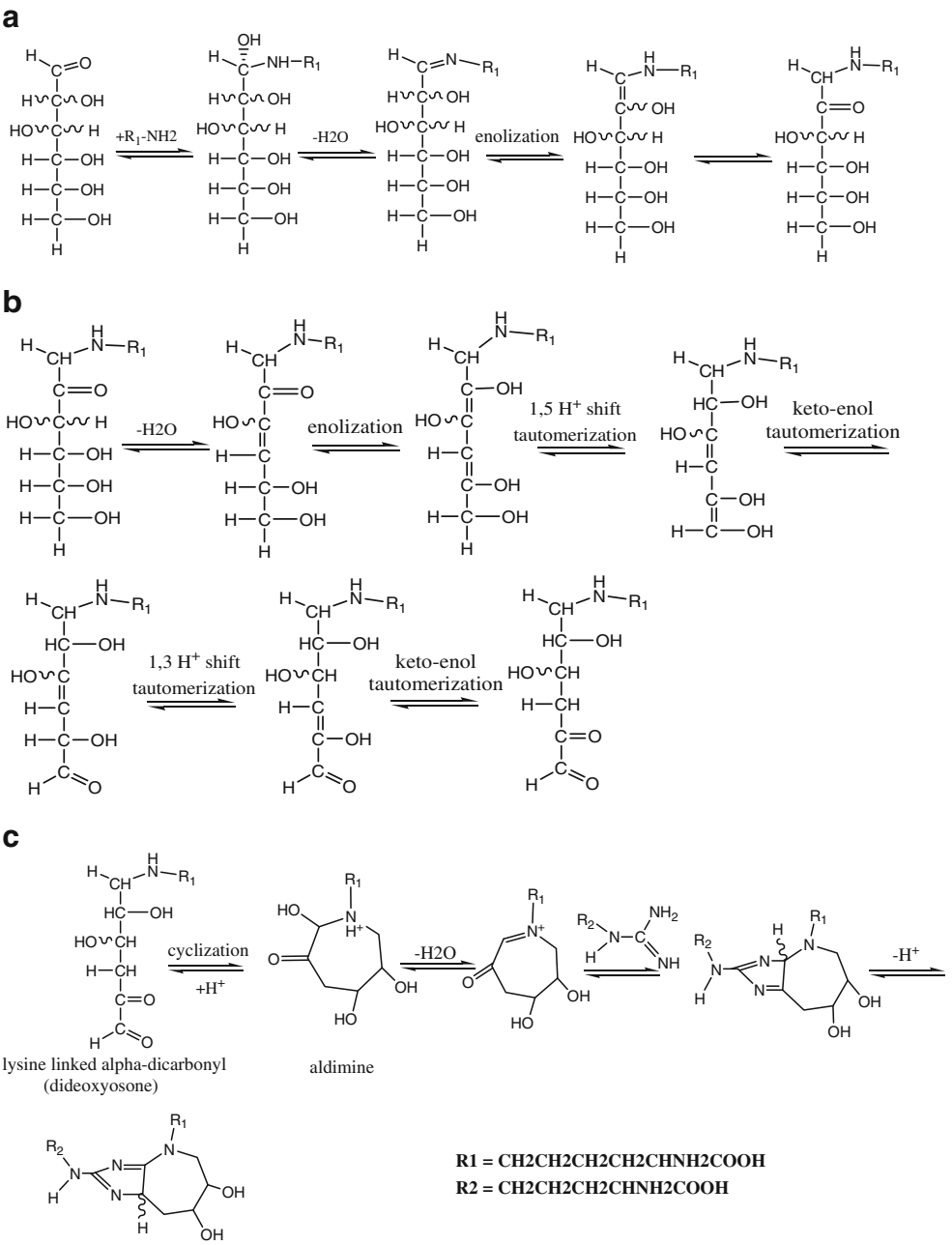
We have investigated mechanisms of formation of the cross-links with three main aims: first, to provide insight into the formation mechanism of glucosepane consistent with experimental results found from glycation of human serum albumin under physiological conditions (Scheme 3); second, to study and compare the reaction mechanisms of the glucosepane, GODIC, and MODIC; third, to analyze the changes in the thermodynamic and kinetic properties caused by the solvent.

## Computational details

The geometries of reactants, products, intermediates, and transition states involved in the proposed mechanism have been fully optimized with using the wB97XD functional [13]. This functional includes empirical dispersion and can better treat hydrogen bonding and van der Waals interactions than conventional DFT. The 6-31+G* basis set with both polarized and diffuse function is an effective choice for more accurate description of hydrogen transfer reactions and hydrogen-bonded species of this study.

For validation of method and basis set, all optimized structures in the first step of glucosepane formation process from wB97XD level were reoptimized with frozen core moller-plesset perturbation theory, MP2(FC)/6-31+G*, computational level. Moreover, CCSD(T)/6-31+G* single point calculations have been carried out for the same compounds to evaluate the reliability of barrier calculated at wB97XD and B3LYP/6-31+G* level. Extension of the basis set to 6-311++G (3df, 3p) in DFT calculations were also examined to compare to the same calculations with 6-31+G* basis set (see Table 1). This strategy showed that firstly, extension of basis set did not have significant effects on free energy diagram while obtained results from wB97XD/6-31+G* are more reliable than B3LYP/6-31+G

[1] In this stage of the reaction, there will be a competition between glucose and α-oxoaldehydes in the nucleophilic attack on amino groups in the proteins. Glucose, because of high stability in its cyclic form, will be less subject to the nucleophilic attacks compared to the α-oxoaldehydes and since α-oxoaldehydes are up to 20,000-fold more reactive than glucose [50]. Therefore, glucosepane can likely be derived through reaction of the α-oxoaldehydes with lysine and arginine amino acids




**Scheme 1** Scheme of proposed mechanism for glucosepane by Biemel, (**a**) Scheme of amadori formation step, (**b**) Scheme of lysine linked alpha-dicarbonyl formation step, (**c**) Scheme of addition of arginine step (R1 and R2 stand for CH2CH2CH2CH2CHNH2COOH and CH2CH2CH2CHNH2COOH, respectively)

ones since former considers non-bonded interactions along with long rang correlations better than latter one (compare relevant CCSD(T) values in Table 1).

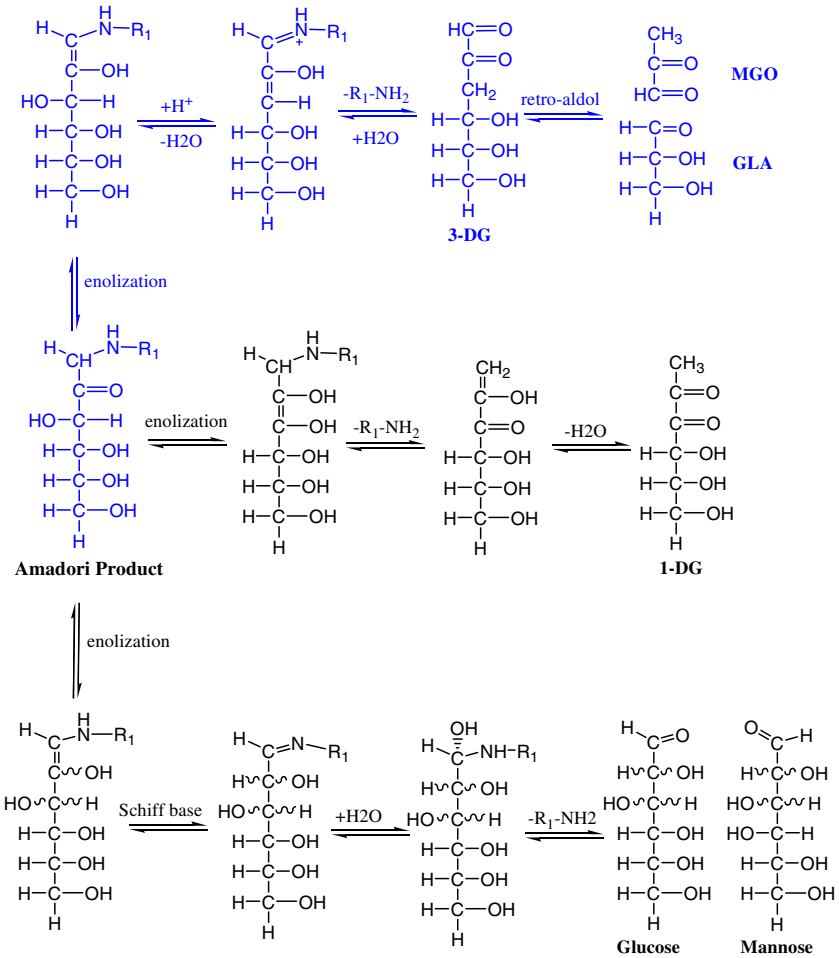
The intrinsic reaction calculations (IRC) [20, 21] in the gas phase confirmed that all TSs were in the right reaction pathways unambiguously.

To evaluate the effects of water as a high-permittivity solvent, on the cross-linking mechanisms we employed polarizable continuum model (PCM) calculations [52] based on solvent reaction field method (SCRF) to further optimize the geometries of the stationary points in water. Vibrational analyses were performed on all optimized structures with the same functional and basis set. The zero-point energy (ZPE) corrections were also included in the reported free energies with a scale factor of 0.89 at temperature 298.15 K and 1 atm pressure in water. All calculations were performed using the Gaussian 09 software [16] and the Gauss View program [15] by employing the software and hardware facilities of the theoretical group of the DCM, University of Grenoble I Codex, France.

## Results and discussion

Density functional theory (DFT) calculations were performed on methyl amine (MA) and methyl guanidine

**Scheme 2** The pathways of amadori products degradation




(MGU) which are models for the lysine and arginine amino acids with α-oxoaldehydes of MGO, GO and GLA to describe the mechanism of the formation of glucosepane, MODIC and GODIC. Water molecules are added in the mechanism, not to simulate the solvent but as reactive species. The proposed mechanism (Scheme 3) involves three main steps, each one of them involves several intermediates and transition state structures: (1) Schiff base formation (Scheme 3a), (2) addition of methyl guanidine (Scheme 3b) (only structures of Int3 to Int7) and (3) addition of glyceraldehyde (Scheme 3c). Because of similarity in new mechanism of Glucosepane with each one of MODIC and GODIC, we have put their mechanisms in the same schemes (Schemes 3a and b) but note that they are different-on the two last TS in Schemes 3b.

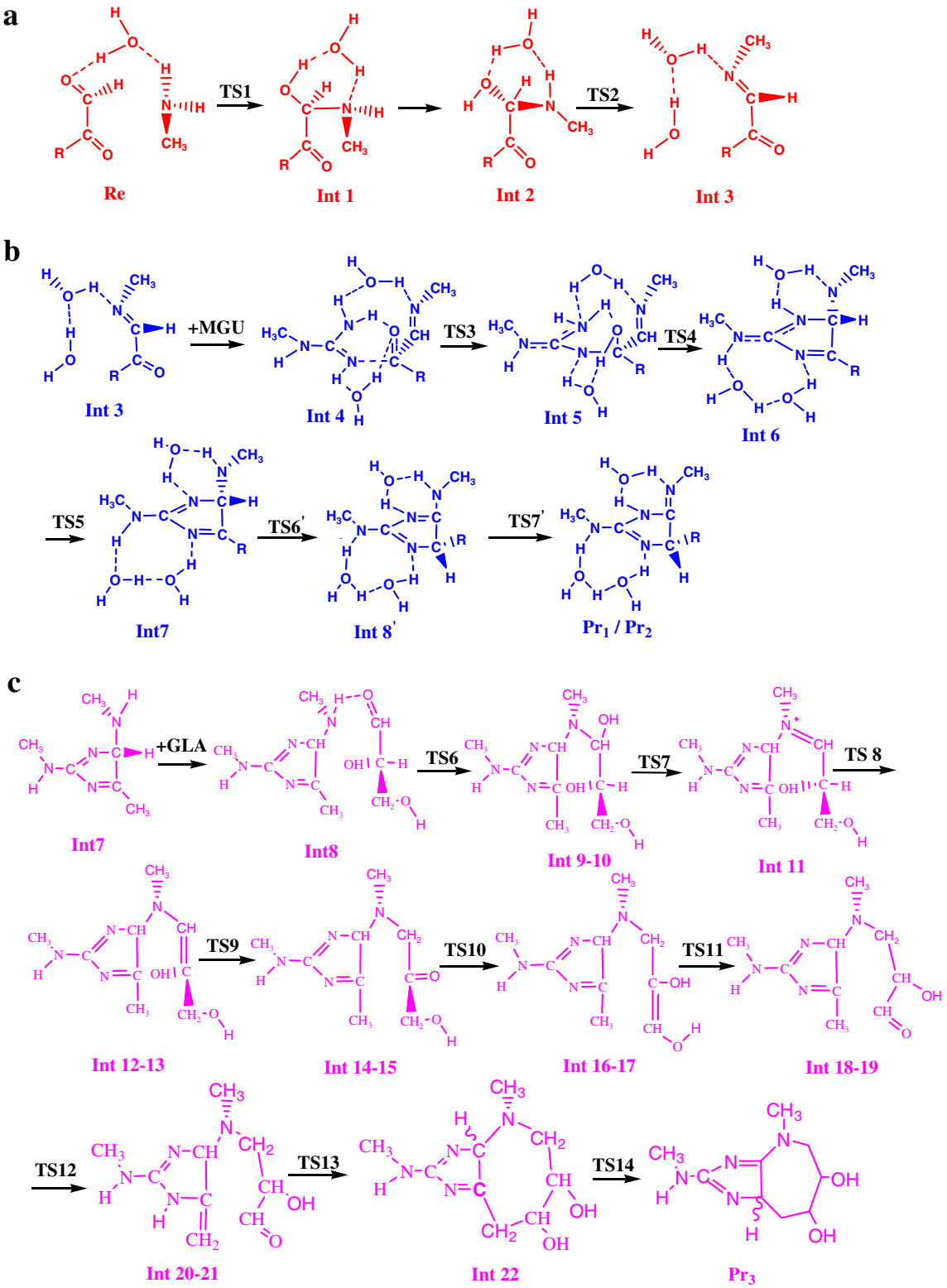
This section is divided as follows: First, the results and discussion of different steps of the glucosepane formation are presented (section A). Then formation mechanism of GODIC and MODIC is compared with glucosepane in section B. Finally, we discuss the solvent effects over the reaction mechanisms (section C).

## A. Formation mechanism of glucosepane

### Schiff base formation

A molecular complex consisting of MGO, MA and one water molecule was chosen as model compound to study the initial step of the cross linking process. The free energy profile in gas phase and solution for the process is depicted in Fig. 1. The electronic energy of the reactant complex is stabilized relative to each of the isolated species while its free energy is destabilized. The difference of values of the electronic and free energies are −11.9 and +9.5 kcal $mol^{-1}$ respectively. Hydrogen bond interactions between the molecules explain the stabilization of the electronic energy and the destabilization of the free energy reflects the negative entropic contribution associated to the transformation of three molecules to one specie. Because of instability of pre reactive complex, we do not include it in the discussion.

The process starts with carbinolamine formation by nucleophilic attack of the methyl amine molecule to the




**Scheme 3** Novel pathway for glucosepane formation proposed in this work, (**a**) Scheme of Schiff base formation step, (**b**) Scheme of addition of methylguanidine step (R for Pr1 and Pr2 is hydrogen and methyl respectively), (**c**) Scheme of addition glyceraldehyde step (All water molecules have been omitted for the sake of clarity)

**Table 1** Changes of electronic energy for the first step of proposed mechanism in the gas phase (in kcal $mol^{-1}$)

| | WB97XD[a] | WB97XD[b] | B3LYP[c] | MP2[d] | CCSD (T)[e] | CCSD (T)[f] |
|---|---|---|---|---|---|---|
| Re——TS1 | 5.4 | 4.2 | 4.5 | 5.7 | 7.2 | 7.3 |
| TS1——Int1 | 16.5 | 17.4 | 15.1 | 16.2 | 18.4 | 18.7 |
| Int1——Int2 | 4.8 | 4.4 | 4.8 | 4.7 | 3.5 | 4.3 |
| Int2——TS2 | 32.6 | 31.7 | 26.9 | 28.8 | 31.0 | 33.7 |
| TS2———Int3 | 24.7 | 26.4 | 23.9 | 25.5 | 26.3 | 28.7 |

Superscripts a, b, c, d, e and f indicate calculations at the WB97XD/6-31+G*, WB97XD/6-311++G (3df, 3p), B3LYP/6-31+G*, MP2 (FC)/6-31+G*, CCSD (T)/6-31+G*//WB97XD/6-31+G* and CCSD (T)/6-31+G*//B3LYP/6-31+G* levels, respectively

carbonyl carbon atom labeled as $C_4$ in Fig. 2a. This attack is performed via a proton transfer from $N_1$ to $O_{15}$ with a free energy barrier equals to 15.7 kcal $mol^{-1}$ (see Fig. 1). In this six-membered-ring TS called TS1 in Fig. 2a, the $C_4$-$N_1$ bond is almost formed (1.53 A) and the $C_4$-$O_7$ bond has a length close to that of a single bond (1.35 A). The Schiff base formation is accomplished with the dehydration of carbinolamine to yield a ketoimine. As Int1 cannot be dehydrated directly, it isomerizes to Int2 through two single bond rotations. The free energy barrier for this transformation is less than 4 kcal without participation of water molecule and can be done easily. Then Int2 evolves to the Schiff base as Hall and Smith reported this stage via reaction of MA and formaldehyde [24]. The reaction follows through a concerted proton transfer in which $H_{11}$ is transferred from $N_1$ to $O_{15}$ while $H_{10}$ from $O_{15}$ to $O_7$ TS2 in Fig. 2a illustrates these proton transfers. In this transition state, the double bond between $N_1$ and $C_4$ is formed. Finally, a ketoimine compound is created that has two electrophilic centers reacts with MGU. As can be seen from free energy profile (Fig. 1), the activation free energy barrier is high for the dehydration step and has a value equal to 33.5 kcal $mol^{-1}$ in the gas phase which is in good agreement with high-level *ab initio* results (33.2 kcal $mol^{-1}$) on similar reaction [24]. It is worth mentioning here that this barrier decreases significantly when more water molecules are mediated around the reactive center of reaction, but since water assisted effects are not the goal of this study we do not discuss it. In comparing to other barriers of the Schiff base formation (see Fig. 1) the last one is the rate-limiting step of this reaction in agreement with experimental kinetic studies [44, 45].

## Addition of methyl guanidine (MGU)

In the second step, methyl guanidine was chosen as molecular model of arginine residue and added to the Schiff base obtained from the first step. Because of existence five acidic protons on the guanidino group (see Scheme 4); there are five different neutral methyl guanidine tautomers. The guanidino function is usually assumed to be in the protonated form with an effective $pK_a$ of about 12.0 in aqueous solution at 298 K [40]. So the first calculations were performed with this form. However our calculations show the protonated form moving away from the Schiff base and we did not succeed to optimize an adduct from them. Hence, for this reason one of the neutral forms was taken into account and could explain why rate of the glycation reaction increases with the basicity [30]. Moreover, the $pK_a$ s values in biological systems are not the same as in the solution. A previous theoretical study shows that all the neutral forms of MGU are in equilibrium to each other [39]. Therefore, we considered only one of these tautomers[2]. Addition of MGU to the Schiff base leads to Int4 in which guanidino group is perpendicular to the ketoimine plan. Two water molecules assist this step, the water molecule was added in the Schiff base formation step, and the water molecule produced from the last dehydration. Since the NH function is more nucleophilic than the $NH_2$ one, first $N_2$ attack to $C_7$ of the Schiff base (see Fig. 2b for the notation). The first proton transfer in this step occurs through TS3 with a free energy barrier less than 8.4 kcal $mol^{-1}$. In this transition state, a twisted six-membered ring is formed in which the proton from NH is being transferred to the water molecule which in turn has already started to transfer its proton to $O_8$. The completion of this reaction leads to a carbinolamine (Int5) whose free energy is stabilized about 11.1 kcal $mol^{-1}$ relative to Int4. This stability is due to formation of bond $N_2$-$C_7$ completely in Int5 while distance of $N_2$-$C_7$ in Int4 is 1.54 Åand $N_2$ has also positive charge. A new step of dehydration occurs to give Int6. The calculated free energy barrier of this dehydration reaction which passes by TS4 (Fig. 1) is also high but 7.5 kcal $mol^{-1}$ less than

[2] Following the comment of a reviewer, we tested the influence of two different isomers of MGU (see tautomers C and E in Scheme 4) on the reaction profile in second step of cross-linking process. We found that activation barriers are to be in different values when tautomers are involved in the process (see Figure S1). It is worth noting that isomer F cannot be involved in the process actively since Lederer and Klaiber definitely excluded formation of MODIC or GODIC via endocyclic nitrogens of arginine residues (see 34 for more detail)

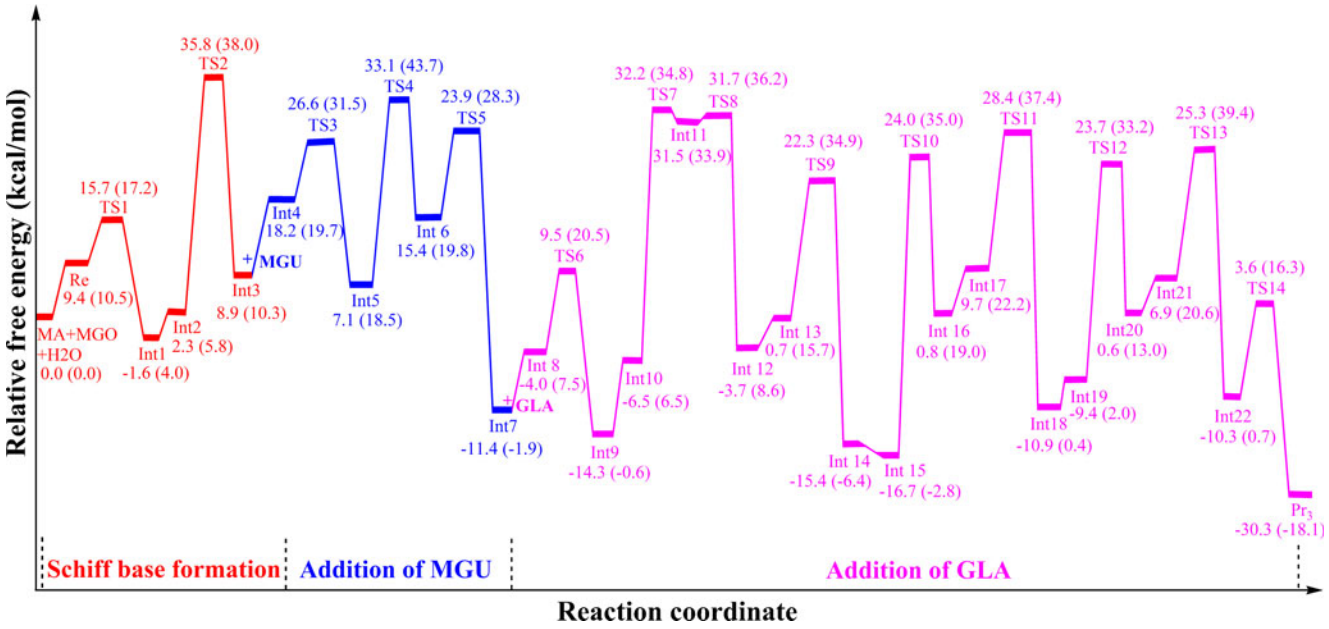



**Fig. 1** Free energy profile for the whole cross-linking process of **glucosepane** analog in gas phase and solvent (in parentheses) calculated at the wb97xd/6-31+G*level

similar one in the first step (TS2 in Fig. 1). The $NH_2$ group of the MGU acts as a proton transferring carrier, thereby lowering the barrier of this second dehydration reaction relative to the first dehydration reaction. After downhill from TS4, Int6 can produce a five-membered heterocyclic structure through TS5, which involves the concurrent transfer of two protons from $N_3$ to $N_1$. The free energy barrier for this reaction is calculated to be 8.5 kcal $mol^{-1}$ relative to Int6. From TS5 to Int7, distance of N3-C4 decreases from 1.51 to 1.46 Å and the free energy decreases to 35.3 kcal $mol^{-1}$ which could explain the irreversibility of the cross-linking process in vivo (see Fig. 1).

## Addition of glyceraldehyde (GLA)

Finally, we consider the formation of glucosepane by addition of glyceraldehyde to Int8 obtained from the previous step. Glyceraldehyde (GLA), the simplest reducing sugar, has been recognized as one of the most reactive species in the Maillard reaction in vivo and in vitro under physiological conditions [12]. Int8 is the reactant complex due to the addition of the GLA to Int7. $N_1$ in Int8 must be dehydrogenated before its addition to the carbonyl function. It is done with the help of the water molecule. Thanks to hydrogen bonds with $O_{19}$ and $H_{11}$, it acts as a bridge to facilitate the proton transfer through a concerted TS (TS6 in Fig. 2c). This proton transfer is also subject to a relatively small barrier (13.5 kcal $mol^{-1}$, Fig. 1).

In this step, because of the carbonyl mobility along the entire GLA backbone and the limited number of water molecules available, we had to change the position of water molecules relative to reactive center manually. This explains the apparition of new stationary points after intermediates 9, 12, 14, 16, 18, 20 (see Fig. 1). As the free energy difference between these stationary points is small, we can use this protocol to describe the mechanism. After formation of another carbinolamine via the reaction of GLA and the five-membered heterocyclic amine (Int7), dehydration occurs through two stepwise reactions. Firstly, dehydroxylation occurs that leads to the formation of the iminium (Int11) with a high barrier of 38.7 kcal/ mol (TS7). Secondly, a hydrogen abstraction is done with a very small free energy barrier. With incorporation of two water molecules, Int13 can be converted to a keto amine (Int14). This tautomerization reaction occurs through TS9 which involves three proton transfer reactions from $O_{16}$ to $C_{20}$ concurrently within the eight-membered ring (Fig. 2c). As it can be seen from Fig. 1, the barrier of this conversion is 21.6 kcal $mol^{-1}$ and the keto form (Int14) is stabilized about 16.1 kcal $mol^{-1}$ compared to the stabilization in the enol form (Int13). Conversion of this keto form to a new enol form (Int16) is performed through TS10 which has a high barrier of 40.8 kcal $mol^{-1}$ consistent with the fact that the $pK_a$ of $C_\alpha$ in GLA is about 10.3 [23]. Our DFT calculations show that the proton-transfer reaction in this tautomerisation is the rate-limiting step. This kind of tautomerization has also been reported by Biemel et al. [5, 7, 33] in the backbone of sugars (see Scheme 1 for comparison).

The tautomeric transformation of Int17 to Int18 is similar to the tautomerization of Int13 to Int14 and their free energy barriers are equal to 18.7 and 21.6 kcal $mol^{-1}$ respectively. An enamine (Int19) is

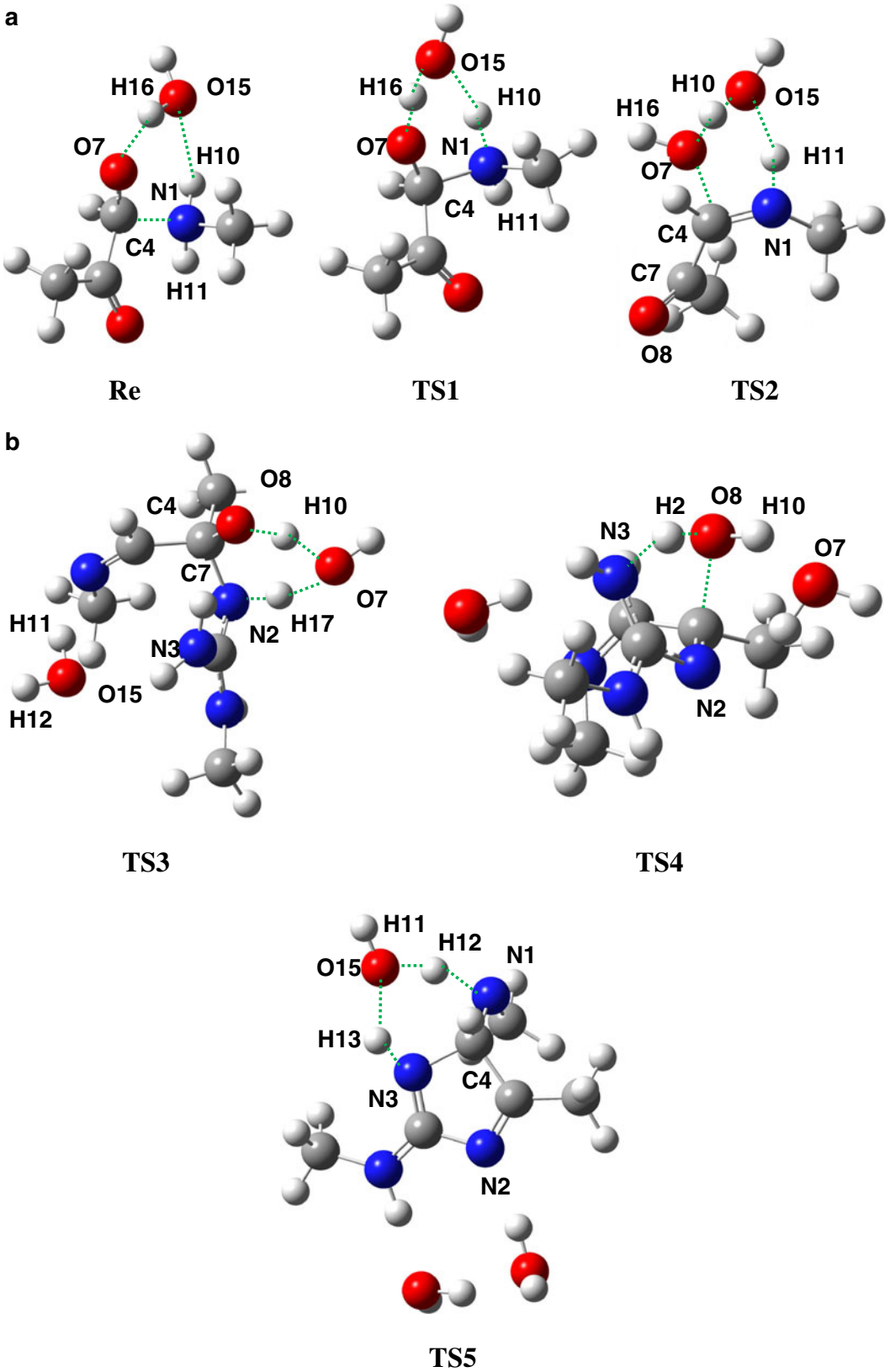



**Fig. 2** (**a**) Optimized (wb97xd/6-31+G*) geometries of the initial reactant complex and the transition states for Schiff base formation step. (**b**) Optimized (wb97xd/6-31+G*) geometries of the transition states for addition methyl guanidine (MGU) step, (**c**) Optimized (wb97xd/6-31+G*) geometries of the transition states and the product for addition of the glyceraldehyde step

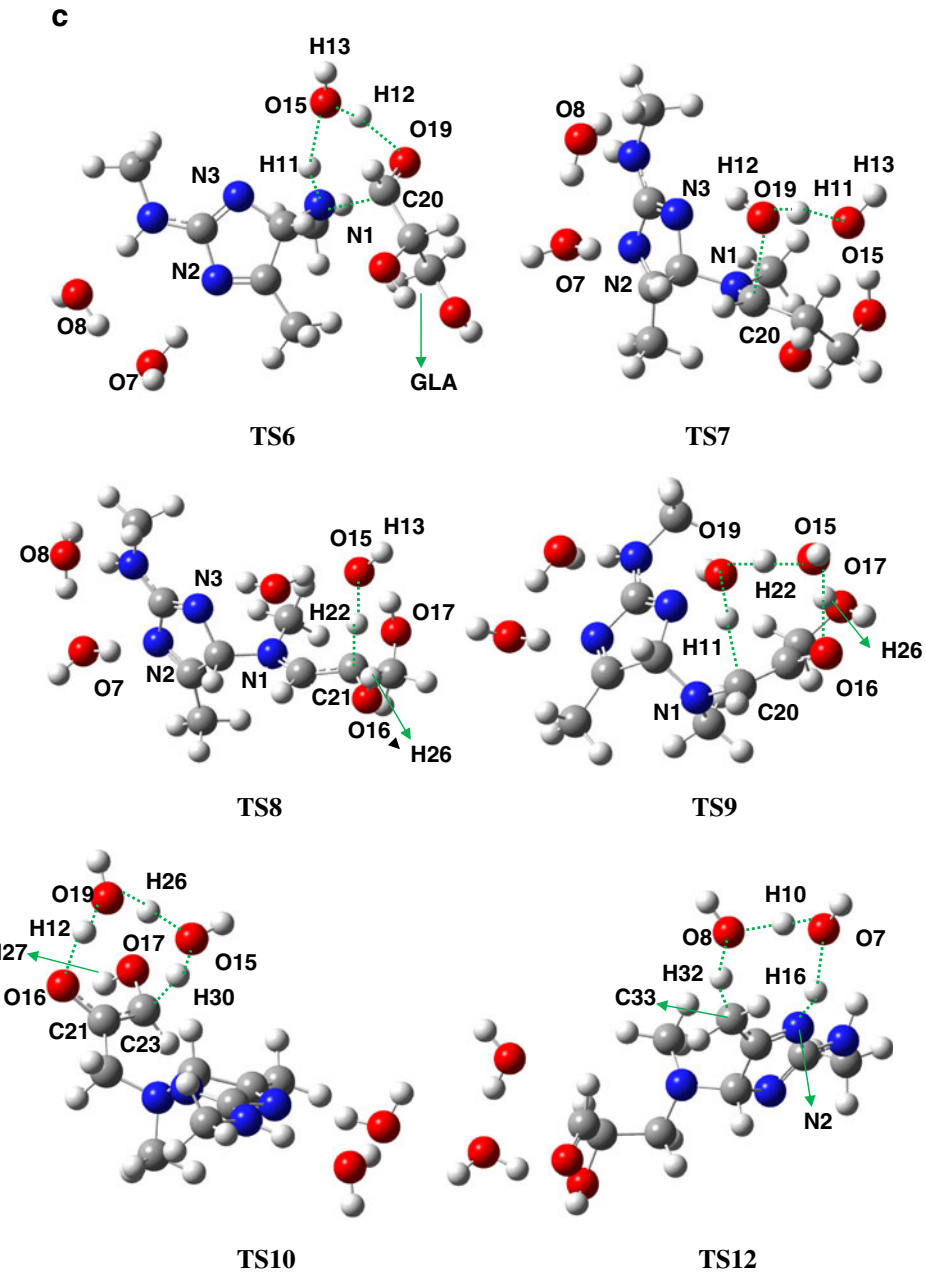



**Fig. 2** (continued)

formed after passing of TS12 which its formation barrier is 33.1 kcal $mol^{-1}$. In this step of reaction $H_{32}$, which is an acidic proton, is transferred to N2 (Fig. 2c). The last cyclization is done through addition of enamine to aldehyde in backbone of glyceraldehyde which is a kind of cross-aldol reaction.

This kind of phenomena has also been found in Morita–Baylis–Hillman reaction and mechanism of direct aldol reactions catalyzed by amino amide derivatives in other studies [14, 41]. In this step, $H_{16}$ is shuttled to $O_{17}$ through forming a less strained ten-membered-ring. Four water molecules assist this cyclization; two of these molecules serve as a bridge to help the proton transfer and the other ones make the cooperation effect to stabilize the structure via the formation of a hydrogen bond (Fig. 2c). As can be seen from Fig. 1, activation free energy for last cyclization (TS13) is about 18 kcal $mol^{-1}$. Distance between $C_{23}$-$C_{33}$ in TS13 reach to 1.7 and it shows that second cyclization through carbon-carbon bond formation mechanism can be done. Addition reactions of enamines through new C-C bond formation have been observed experimentally [10]. The reaction proceeds downhill from TS13 to form a seven-membered heterocyclic ring attached to a five-membered one (Int22) whose stability is 17.2 kcal $mol^{-1}$ more than Int21.

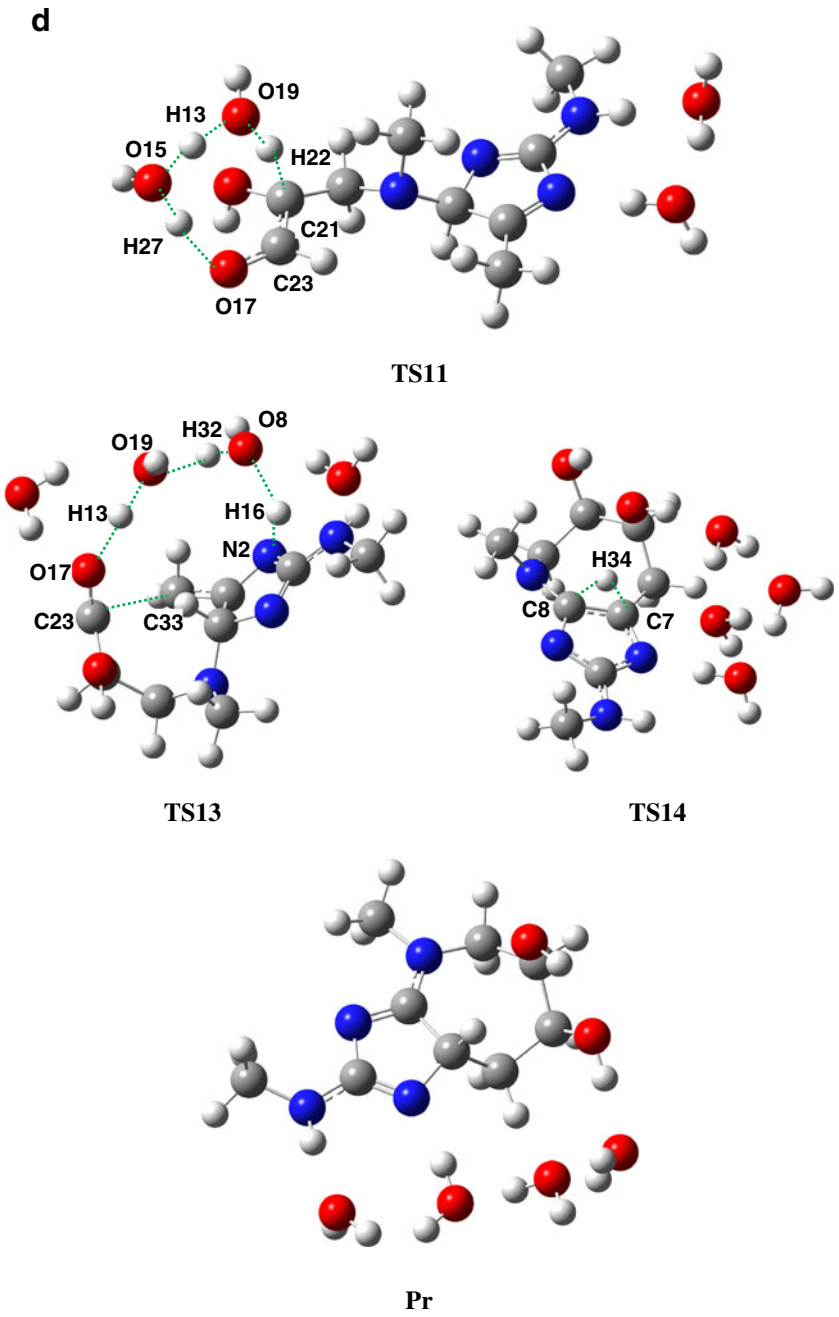
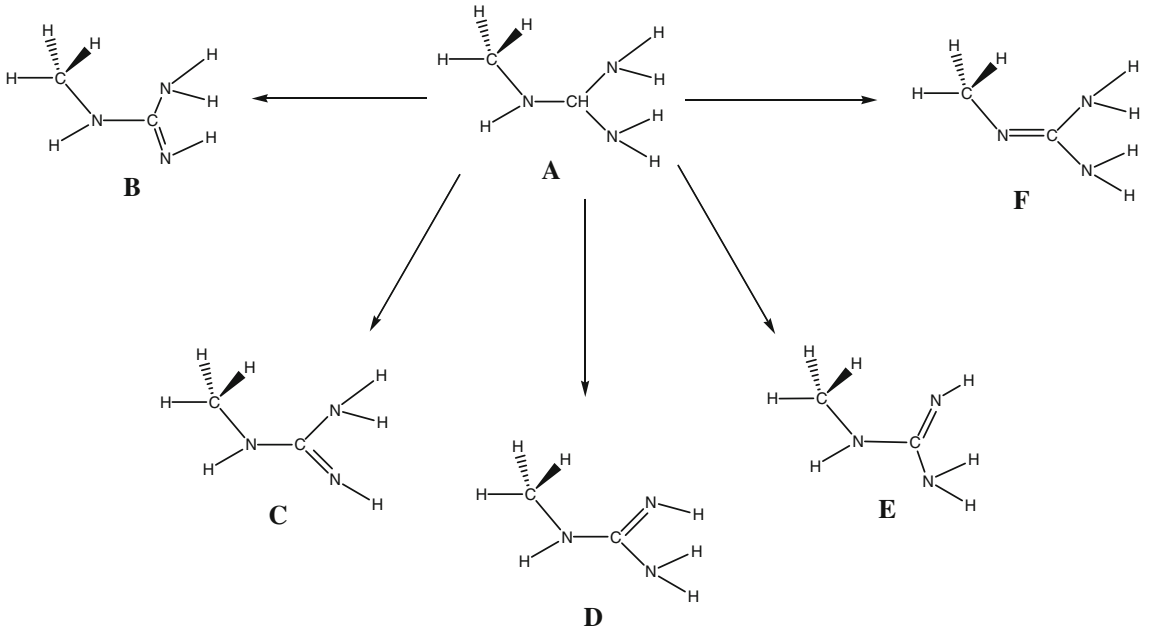



**Fig. 2** (continued)

The last proton transfer in this process is performed from $C_8$ to $C_7$ (see Fig. 2c TS14) and leads to the final product.

### B. Comparison of cross-links

Formation mechanism of GODIC and MODIC (Scheme 3) proceeds through two main steps namely: (1) Schiff base formation (Scheme 3a), (2) addition of methyl guanidine (Scheme 3b). All optimized structures for these cross-links are given in the Supporting information. Their free energy surfaces in gas and solvent phases are depicted in Figs. 3 and 4 respectively. When comparing these cross-links (see Scheme 3 along with Figs. 3 and 4), we find that in the first TS corresponding to the carbinolamine formation, relative free energies as compared to separated reactants are the same in both cross-links. Moreover, barriers of the first dehydration reaction (TS2) in the formation of GODIC and MODIC are equal to 32.3 and 33.5 kcal $mol^{-1}$ respectively. These barriers have highest




**Scheme 4** Structures of the protonated (A) and neutral forms of methyl guanidine (B-F)

values in free energy profile and are predicted to be the rate-limiting steps. There is a significant difference of 3.7 kcal $mol^{-1}$ between the free energy barriers of the second dehydration reaction (TS4) in GODIC and MODIC. This is due to the induction effect of methyl group on the $C_7$ in MODIC (see Fig. 2b for notation) which would lower positive charge from 0.337 in $C_7$ of GODIC to 0.001 in C7 of MODIC. Thus TS4 in GODIC is more advanced and the free energy barrier of this step is higher than that in MODIC formation.

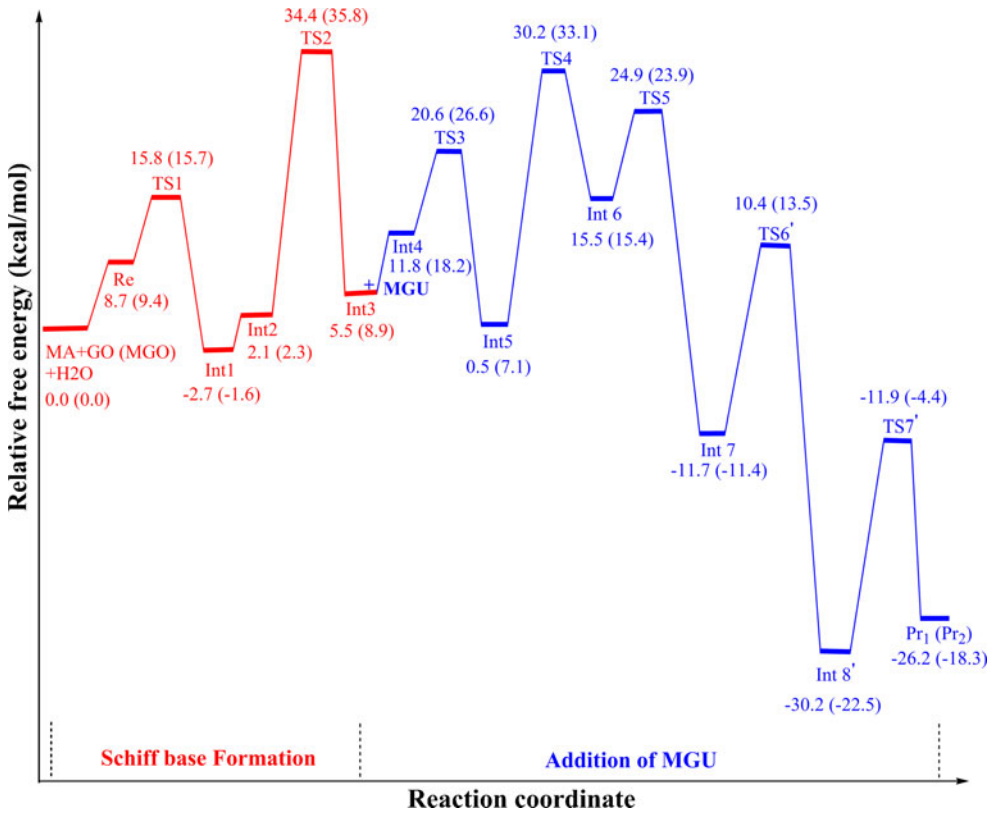



**Fig. 3** Free energy profile for the whole cross-linking process of **GODIC** and **MODIC** (in parentheses) analog in gas phase calculated at the wb97xd/6-31+G*level

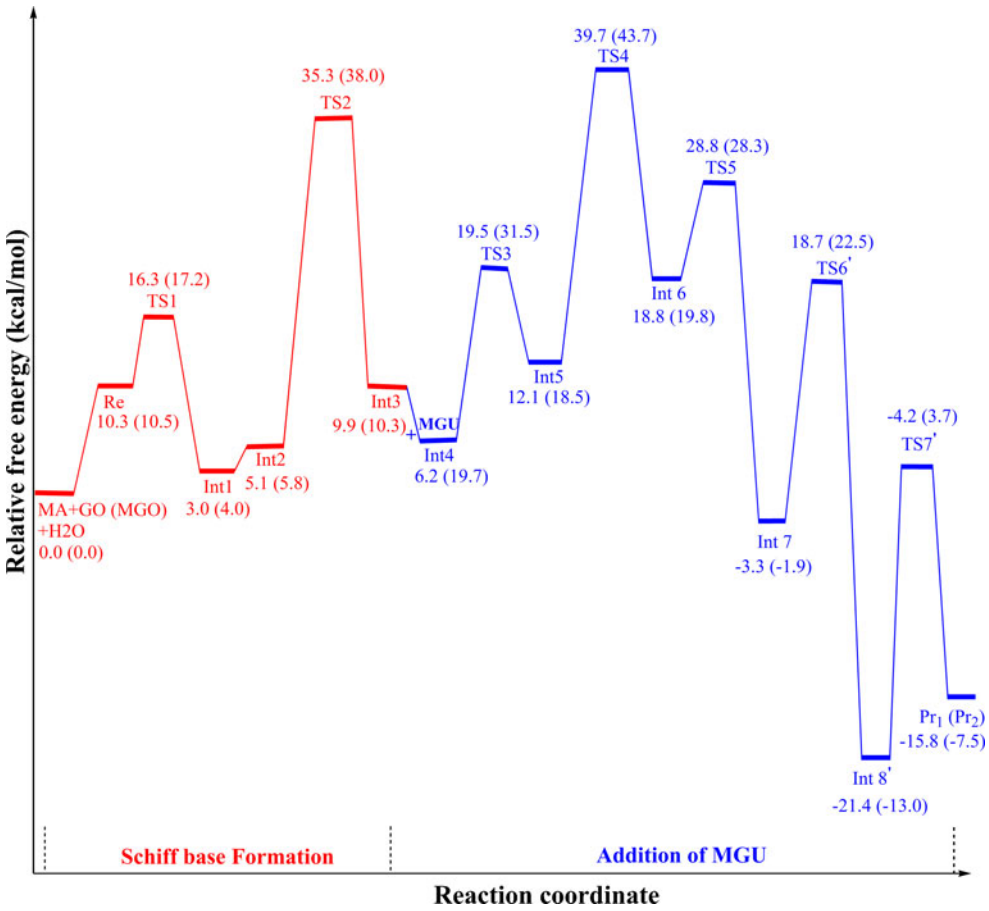



**Fig. 4** Free energy profile for the whole cross-linking process of **GODIC** and **MODIC** (in parentheses) analog in solvent phase calculated at the wb97xd/6-31+G*level

When we compare Figs. 1 and 3, it is clear that these cross-links share steps of Schiff base formation and addition of MGU with glucosepane. As it can be seen from the figures, from a thermodynamics point of view, the three processes are exothermic and the glucosepane formation is more favorable than each of GODIC and MODIC (by ca. 4.1 and 12.0 kcal $mol^{-1}$ respectively). however, the free energy barrier of the rate-limiting step in the glucosepane formation is higher than those in the GODIC and MODIC formation (40.8 kcal $mol^{-1}$ against 32.3 and 33.5 kcal $mol^{-1}$ respectively). So its formation is slower than others.

Table 2 summarizes levels of cross-links in vivo. In the different matrices of human, concentration of glucosepane is more important than each of MODIC and GODIC, which corroborates with the fact that its formation is highly exothermic. Moreover experimental results [6, 46] have confirmed that time incubation for its formation is higher than the two others cross-links. So our results are in good agreement with experimental evidence in vivo.

## C. Solvent effects

To estimate the solvent effect, the structures of all stationary points were reoptimized at the wb97xd/6-31+G* level with the PCM model. Cartesian coordinates optimized in the water are given in the Supporting information. The free energy profiles in solution are represented in Figs. 1 and 4 for glucosepane and GODIC (MODIC) respectively. The free energy barriers for each step of cross-linking processes in the gas and water phase for comparison are summarized in Table 3.

It is noteworthy that entropic contributions in aqueous solution can be somewhat overestimated since calculated in the ideal gas model[3] [47].

The presence of the polarizable dielectric of solvent increases the barriers when the dipole moment of the TS structures is smaller than that in their intermediates in solvent phase. For example, in glucosepane formation, relative free energy of TS1 with respect to separated

[3] Following the comment of a reviewer, we computed entropic contributions on free energy profile in the gas phase and solvent for glucosepan. Our calculations revealed that this contribution has maximum values at the beginning of any step; (i.e, in addition of MA, MGU and GLA on GO/MGO, Int4 and Int8, respectively) whereas in any other steps enthalpic contributions overcome on entropy penalty (see Table S1). Therefore, it can be concluded that entropic contribution in most cases can not affect the reaction profile (exception in beginning of each steps) while that enthalpic terms are more effective and determining on the proceeding of reaction

**Table 2** Levels of the cross-links in different matrices of human[a]

| | Glucosepane | GODIC | MODIC |
|---|---|---|---|
| Human skin collagen | 2000 (5000)[b] | 15 (<60) | 30 (<60) |
| Glomerular basement membrane (GBM) | 500 (900) | 80 | 60 |
| Human lens proteins | 185.1 | 21.1 | 65.5 |
| Human serum albumin (HSA) | 17.1 (29.2) | N.D[c] | 4.1 (3.9) |

[a] Data obtained from references 5 and 17

[b] All values are expressed as mmol/mg of protein and in parentheses are amount of cross-links in diabetic subjects

[c] N.D stands for "not detected"

reactants increases by 1.5 kcal $mol^{-1}$ in solution. It can be attributed to that the dipole moment of TS1 (5.1 Debye) is smaller than that of MA+MGO+$H_2O$ separated reactant (5.8 Debye). So the transition state is destabilized in solvent water by solvation. On the contrary, the dipole moment of TS7 (13.9 Debye) is larger than that of Int10 (6.2 Debye), and the solvation effect on TS7 is stronger than that on Int11. Therefore, the solvent water stabilizes the transition state TS11, leading to a decrease of free energy barrier by 10.4 kcal $mol^{-1}$. It can be conclude that by increasing of dipole moment of TSs relative to their intermediates (positive values of $\Delta\mu$ (sol) in Table 3), the free energy barriers in solution decrease and inverse too.

Thermodynamics of the whole cross-linking processes for glucosepane, GODIC and MODIC are disfavored by 12.2, 10.4 and 10.8 kcal $mol^{-1}$ when solvent is taken into account. However these reactions are still favorable (by ca. 18.1, 15.8 and 7.5 respectively for mentioned cross-links) (see Table 3 for comparison).

The rate-limiting steps stay the same for all three cross-links formation and the first dehydration reaction (TS2) for the formation of GODIC and MODIC and the last keto-enol conversion on the backbone of glyceraldehyde in the third

**Table 3** The reaction barriers and relative free energies respect to separated reactants expressed in kcal $mol^{-1}$ for each reaction step of GODIC (MODIC) and Glucosepane models in the gas and solvent phase at wb97xd/6-31+G* level of theory

| Mechanism of GODIC (MODIC) | | | | |
|---|---|---|---|---|
| | $\Delta G_{gas}^{\neq}$ | $\Delta G_{sol}^{\neq}$ | $\Delta G_{sol}^{\neq} - \Delta G_{gas}^{\neq}$ | $\Delta\mu_{SOL}$[a] |
| Re1 | 0.0 | 0.0 | 0.0 | 0.0 |
| Re1 — TS1 | 15.8 (15.7[b]) | 16.3 (17.2) | 0.5 (1.5) | −1.0 (−0.6) |
| INT2 — TS2 | 32.3 (33.5) | 30.2 (32.2) | −2.1 (−1.2) | 2.3 (3.4) |
| INT4 — TS3 | 8.8 (8.4) | 13.3 (11.8) | 4.5 (3.4) | −2.5 (−2.0) |
| INT5 — TS4 | 29.7 (26.0) | 27.6 (25.2) | −2.1 (−0.8) | 2.4 (2.1) |
| INT7 — TS5 | 0.4 (0.3) | −0.8 (0.9) | −1.2 (−1.1) | 0.5 (0.7) |
| INT8 — TS6 | 22.0 (25.0) | 22.0 (24.4) | 0.1 (−0.6) | −0.1 (0.1) |
| INT9 — TS7 | 18.3 (18.1) | 17.2 (16.7) | −1.1 (−1.4) | 1.0 (0.7) |
| Mechanism of glucosepane | | | | |
| Re2 | 0.0 | 0.0 | 0.0 | 0.0 |
| Re2 — TS1 | 15.7 | 17.2 | 1.5 | −0.6 |
| INT2 — TS2 | 33.5 | 32.2 | −1.2 | 3.4 |
| INT4 — TS3 | 8.4 | 11.8 | 3.4 | −2.0 |
| INT5 — TS4 | 26.0 | 25.2 | −0.8 | 2.1 |
| INT7 — TS5 | 0.3 | −0.9 | −1.1 | 0.7 |
| INT9 — TS6 | 13.5 | 13.0 | −0.5 | 1.5 |
| INT11——— TS7 | 38.8 | 28.3 | −10.4 | 7.8 |
| INT12——— TS8 | 0.2 | 2.3 | 2.1 | −1.3 |
| INT14——— TS9 | 21.5 | 19.2 | −2.4 | 13.0 |
| INT16——— TS10 | 40.8 | 37.8 | −3.0 | 0.2 |
| INT18——— TS11 | 18.7 | 15.2 | −3.5 | 0.2 |
| INT20——— TS12 | 33.1 | 31.1 | −1.9 | 0.7 |
| INT22——— TS13 | 18.4 | 18.8 | 0.4 | −2.4 |
| INT23——— TS14 | 13.8 | 15.5 | 1.7 | −3.3 |
| Thermodynamics of reactions | | | | |
| | $\Delta G_{gas}^{r}$ | $\Delta G_{sol}^{r}$ | $\Delta G_{sol}^{r} - \Delta G_{gas}^{r}$ | $\Delta\mu_{SOL}$[c] |
| $Re_3$ — $Pr_1$ ($Pr_2$) | −26.2 (−18.3) | −15.8 (−7.5) | 10.4 (10.8) | −7.8 (−6.4) |
| $Re_4$ — $Pr_3$ | −30.3 | −18.1 | 12.2 | −6.5 |

[a] $\Delta\mu_{SOL}=\mu^{TS} - \mu^{INT}$

[b] Value in parentheses are for MODIC model

[c] $\Delta\mu_{SOL}=\mu^{Pr} - \mu^{Re}$

Re1 is MA+GO (MGO)+H2O as separated

R2 is MA+MGO+H2O as separated

Re3 is R1+MGU as separated

Re4 is R2+MGU+GLA as separated

step of process for the Glucosepane are still rate-determining steps.

## Conclusions

In summary, our studies focused upon the cross linking reactions between α-oxoaldehydes and models of lysine and arginine residues in order to better understand cross-linking chemistry in vivo and food chemistry. Our results show that these reactions are thermodynamically and kinetically favorable in gas phase and solution with a high stabilization of the free energy of the glucosepane in respect to the separated reactants and values of free energy barriers are reasonable. Comparisons with the formation of two other important cross-links, GODIC and MODIC, show that the glucosepane formation is more exothermic than others while its formation kinetics is slower. These results could illuminate mechanisms of AGE formation in diabetes and aging as well as suggest potential therapeutic interventions in good agreement with experimental results in glycation of human serum albumin under physiological conditions.

**Acknowledgments** The financial support of Research Council of Shahid Beheshti University is gratefully acknowledged. Rasoul Nasiri would like to thank the Ministry of Science, Research, and Technology of Iran for financial support. Many thanks also go to Martin Field and Andrew Greene for their helpful discussions.

## References

1. Ahmed N, Argirov OK, Minhas HS, Cordeiro CAA, Thornalley PJ (2002) Assay of advanced glycation endproducts (AGEs): Surveying AGEs by chromatographic assay with derivatization by 6-aminoquinolyl-N-hydroxysuccinimidyl-carbamate and application to Nε-carboxymethyl-lysine- and nε-(1-carboxyethyl)lysine-modified albumin. Biochem J 364:1–14
2. Ahmed N, Thornalley PJ (2002) Chromatographic assay of glycation adducts in human serum albumin glycated in vitro by derivatization with 6-aminoquinolyl-N-hydroxysuccinimidyl-carbamate and intrinsic fluorescence. Biochem J 364:15–24
3. Beisswenger PJ, Szwergold BS, Yeo KT (2001) Glycated proteins in diabetes. Clin Lab Med 21:53–78
4. Biemel KM, Bühler HP, Reihl O, Lederer MO (2001) Identification and quantitative evaluation of the lysine-arginine crosslinks GODIC, MODIC, DODIC, and glucosepan in foods. Nahrung Food 45:210–214
5. Biemel KM, Conrad J, Lederer MO (2002) Unexpected carbonyl mobility in aminoketoses: The key to major maillard crosslinks. Angew Chem Int Ed 41:80–804
6. Biemel KM, Friedl DA, Lederer MO (2002) Identification and quantification of major maillard cross-links in human serum albumin and lens protein. Evidence for glucosepane as the dominant compound. J Biol Chem 277:24907–24915
7. Biemel KM, Reihl O, Conrad J, Lederer MO (2001) Formation pathways for lysine-arginine cross-links derived from hexoses and pentoses by maillard processes: Unraveling the structure of a pentosidine precursor. J Biol Chem 276:23405–23412
8. Bierhaus A, Schiekofer S, Schwaninger M, Andrassy M, Humpert PM, Chen J, Hong M, Luther HT, Klöting I, Morcos M, Hofmann M, Tritschler H, Weigle B, Kasper M, Smith M, Perry G, Schmidt AM, Stern DM, Häring HU, Schleicher E, Nawroth PP (2001) Diabetes-associated sustained activation of the transcription factor nuclear factor-κB. Diabetes 50:2792–2808
9. Bose T, Chakraborti AS (2008) Fructose-induced structural and functional modifications of hemoglobin: Implication for oxidative stress in diabetes mellitus. Biochim Biophys Acta 1780:800–808
10. Brannock KC, Bell A, Burpitt RD, Kelly CA (1964) Enamine chemistry. IV. Cycloaddition reactions of enamines derived from aldehydes and acyclic ketones. J Org Chem 29:801–812
11. Bucciarelli LG, Wendt T, Rong Lalla E, Hofmann MA, Goova MT, Taguchi A, Yan S, Yan SD, Stern DM, Schmidt AM (2002) RAGE is a multiligand receptor of the immunoglobulin superfamily: Implications for homeostasis and chronic disease. Cell Mol Life Sci 59:1117–1128
12. Candiano GM, Ghiggeri G, Delfino C, Quierolo C, Cuniberti E, Righetti PG (1985) Reaction of lysine with aldoses. Carbohydr Res 145:99–112
13. Chai JD, Head-Gordon M (2008) Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections. Phys Chem Chem Phys 10:6615–6620
14. Dong L, Qin S, Su Z, Yang H, Hu C (2010) Computational investigation on the mechanism and the stereoselectivity of Morita-Baylis-Hillman reaction and the effect of the bifunctional catalyst N-methylprolinol. Org Biomol Chem 8:3985–3991
15. Frisch AE, Dennington RD, Keith TA, Neilsen AB, Holder AJ (2003) GaussView, Rev 3.9. Gaussian Inc, Pittsburgh
16. Frisch MJ et al (2009) Gaussian 09, Revision A.01. Gaussian Inc, Wallingford
17. Gerrard JA, Brown PK, Fayle SE (2002) Maillard crosslinking of food proteins I: The reaction of glutaraldehyde, formaldehyde and glyceraldehyde with ribonuclease. Food Chem 79:343–349
18. Gerrard JA, Brown PK, Fayle SE (2003) Maillard crosslinking of food proteins II: The reactions of glutaraldehyde, formaldehyde and glyceraldehyde with wheat proteins in vitro and in situ. Food Chem 80:35–43
19. Gerrard JA, Brown PK, Fayle SE (2003) Maillard crosslinking of food proteins III: The effects of glutaraldehyde, formaldehyde and glyceraldehyde upon bread and croissants. Food Chem 80:45–50
20. Gonzalez C, Schlegel HB (1989) An improved algorithm for reaction path following. J Chem Phys 90:2154–2161
21. Gonzalez C, Schlegel HB (1990) Reaction path following in mass-weighted internal coordinates. J Phys Chem 94:5523–5527
22. Grillo MA, Colombatto S (2008) Advanced glycation end-products (AGEs): Involvement in aging and in neurodegenerative diseases. Amino Acids 35:29–36
23. Guthrie JP, Cossar J (1986) The chlorination of acetone: a complete kinetic analysis. Can J Chem 64:1250–1266
24. Hall NE, Smith BJ (1998) High-level ab initio molecular orbital calculations of imine formation. J Phys Chem A 102:4930–4938
25. Harrington CR, Colaco CA (1994) A glycation connection. Nature 370:247–248
26. Hartog JW, Voors AA, Bakker SJ, Smit AJ, van Veldhuisen DJ (2007) Advanced glycation end-products (AGEs) and heart failure: Pathophysiology and clinical implications. Eur J Heart Fail 9:1146–1155
27. Hayashi T, Shibamoto T (1985) Analysis of methyl glyoxal in foods and beverages. J Agric Food Chem 33:1090–1093

28. Henle T (2005) Protein-bound advanced glycation endproducts (AGEs) as bioactive amino acid derivatives in foods. Amino Acids 29:313–322
29. Hinton DJS, Ames JM (2006) Site specificity of glycation and carboxymethylation of bovine serum albumin by fructose. Amino Acids 30:425–433
30. Hodge JE (1953) Dehydrated foods: Chemistry of browning reactions in model systems. J Agric Food Chem 1:928–943
31. Kimura Y, Hyogo H, Yamagishi S-I, Takeuchi M, Ishitobi T, Nabeshima Y, Arihiro K, Chayama K (2010) Atorvastatin decreases serum levels of advanced glycation endproducts (AGEs) in nonalcoholic steatohepatitis (NASH) patients with dyslipidemia: Clinical usefulness of AGEs as a biomarker for the attenuation of NASH. J Gast 45:750–757
32. Kitahara Y, Takeuchi M, Miura K, Mine T, Matsui T, Yamagishi S (2008) Glyceraldehyde-derived advanced glycation end products (AGEs). A novel biomarker of postprandial hyperglycaemia in diabetic rats. Clin Exp Med 8:175–177
33. Lederer MO, Bühler HP (1999) Cross-linking of proteins by maillard processes characterization and detection of a lysine-arginine cross-link derived from D-glucose. Bioorg Med Chem 7:1081–1088
34. Lederer MO, Klaiber RG (1999) Cross-linking of proteins by Maillard processes: Characterization and detection of lysine-arginine cross-links derived from glyoxal and methylglyoxal. Bioorg Med Chem 7:2499–2507
35. Ledl F, Schleicher E (1990) New aspects of the Maillard reaction in foods and in the human body. Angew Chem Int Ed 29:565–594
36. Lim JS, Mietus-Snyder M, Valente A, Schwarz JM, Lustig RH (2010) The role of fructose in the pathogenesis of NAFLD and the metabolic syndrome. Nature 7:251–264
37. Maillard LC (1912) Action des acides amines sur le sucres: Formation des me'lanoidines per voie me'thodique. C R Acad Sci 154:66–68
38. Martins SI, Marcelis AT, van Boekel MA (2003) Kinetic modelling of Amadori N-(1-deoxy-D-fructos-1-yl)-glycine degradation pathways. Part I - Reaction mechanism. Carbohydr Res 338:1651
39. Norberg J, Foloppe N, Nilsson L (2005) Intrinsic relative stabilities of the neutral tautomers of arginine side-chain models. J Chem Theory Comput 1:986–993
40. Nozaki Y, Tanford C (1967) Examination of titration behavior. Methods Enzymol 11:715–734
41. Rao QQ, Luo SW, Gong LZ (2010) Water as dual functional cocatalyst: A theoretical study on the mechanism of direct aldol reaction on water catalyzed by a leucine derivative. Chin Sci Bull 55:1742–1752
42. Sato T, Iwaki M, Shimogaito N, Wu X, Yamagishi S-I, Takeuchi M (2006) TAGE (Toxic AGEs) theory in diabetic complications. Curr Mol Med 6:351–358
43. Sattarahmady N, Moosavi-Movahedi AA, Ahmad F, Hakimelahi GH, Habibi-Rezaei M, Saboury AA, Sheibani N (2007) Formation of the molten globule-like state during prolonged glycation of human serum albumin. Biochim Biophys Acta 1770:933–942
44. Sayer JM, Jenks WP (1977) Imine-forming elimination reactions. 2. Imbalance of charge distribution in the transition state for carbinolamine dehydration. J Am Chem Soc 99:464–474
45. Sayer JM, Pinsky B, Schonbrunn A, Washtien W (1974) Mechanism of carbinolamine formation. J Am Chem Soc 96:7998–8009
46. Sell DR, Biemel KM, Reihl O, Lederer MO, Strauch CM, Monnier VM (2005) Glucosepane is a major protein cross-link of the senescent human extracellular matrix: Relationship with diabetes. J Biol Chem 280:12310–12315
47. Strajbl M, Sham YY, Vill-a J, Chu Z-T, Warshel AJ (2000) Calculations of Activation Entropies of Chemical Reactions in Solution. J Phys Chem B 104:4578–4584
48. Takeuchi M, Kikuchi S, Sazaki N, Suzuki T, Watai T, Iwaki M, Bucala R, Yamagishi S (2004) Involvement of advanced glycation end-products (AGEs) in Alzheimer's disease. Curr Alzheimer Res 1:39–46
49. Takeuchi M, Yamagishi S (2008) Possible involvement of advanced glycation end-products (AGEs) in the pathogenesis of Alzheimer's disease. Curr Pharm Des 14:973–978
50. Thornalley PJ (2005) Dicarbonyl intermediates in the Maillard reaction. Ann NY Acad Sci 1043:111–117
51. Thornalley PJ, Langborg A, Minhas HS (1999) Formation of glyoxal, methylglyoxal and 3-deoxyglucosone in the glycation of proteins by glucose. Biochem J 344:109–116
52. Tomasi J, Persico M (1994) Molecular interactions in solution: An overview of methods based on continuous distributions of the solvent. Chem Rev 94:2027–2094
53. Wells-Knecht KJ, Zyzak DV, Litchfield JE, Thorpe SR, Baynes JW (1995) Mechanism of autoxidative glycosylation: Identification of glyoxal and arabinose as intermediates in the autoxidative modification of proteins by glucose. Biochem 34:3702–3709
54. Wolff SP, Crabbe MJC, Thornalley PJ (1984) The autoxidation of glyceraldehyde and other simple monosaccharides. Experientia 40:244–246
55. Wolff SP, Dean RT (1987) Glucose autoxidation and protein modification. The potential role of "autoxidative glycosylation" in diabetes. Biochem J 245:243–250
56. Yamagishi S-I (2011) Role of advanced glycation end products (AGEs) and receptor for AGEs (RAGE) in vascular damage in diabetes. Exp Gerontology 46:217–224
57. Yaylayan VA (1997) Classification of the Maillard reaction: A conceptual approach. Trend Food Sci Tech 8:13–18
58. Zyzak DV, Richardson JM, Thorpe SR, Baynes JW (1995) Formation of reactive intermediates from Amadori compounds under physiological conditions. Arch Biochem Biophys 316:547–554

J Mol Model (2012) 18:1661–1666
DOI 10.1007/s00894-011-1193-2

ORIGINAL PAPER

# BON-BONs: cyclic molecules with a boron-oxygen-nitrogen backbone. Computational studies of their thermodynamic properties

**Aloysus K. Lawong · David W. Ball**

Received: 13 June 2011 / Accepted: 19 July 2011 / Published online: 3 August 2011


**Abstract** Although they were first reported in 1963, molecules with a boron-oxygen-nitrogen dimeric backbone do not seem to have been investigated seriously in terms of thermodynamic properties. Here we report on the calculated structures and properties, including thermodynamics, of several so-called "BON-BON" molecules. With the popularity of nitrogen-containing substituents on new high-energy materials, nitro-substituted BON-BONs were a focus of our investigation. A total of 42 BON-BON molecules were evaluated, and thermochemical analysis shows a decrease in the specific enthalpy of combustion or decomposition with increasing $NO_2$ content, consistent with other systems.

**Keywords** B3LYP calculations · BON-BON molecules · High energy materials

## Introduction

The search for new high energy (HE) materials is ongoing [1]. Good HE materials share certain physical characteristics, including high density, thermodynamics favorability of decomposition or combustion, and the production of high-pressure products upon reaction. Models for these characteristics, and others, have allowed researchers to probe the probabilities that new, as yet unsynthesized materials may be good potential HE targets.

A. K. Lawong · D. W. Ball (✉)
Department of Chemistry, Cleveland State University,
2121 Euclid Avenue,
Cleveland, OH 44115, USA
e-mail: d.ball@csuohio.edu

Recently we have been interested [2–4] in boron-containing compounds as potential HE materials because the formation of $B_2O_3$ as a decomposition/combustion product is very energetically favored; at 18.3 kJ $g^{-1}$ [5], boron has one of the highest specific combustion energies of any element. For similar reasons, we have also investigated a series of boron/nitrogen compounds [6–8], taking advantage of the stability of $N_2$ as a product of decomposition or combustion.

In 1963, Kuhn and Inatome [9] published a report on an air-stable boron-oxygen-nitrogen molecule that they determined was composed of a $BH_2$-O-$NH_2$ dimer in which the nitrogen atom lone electron pair of one molecule was making a coordinate covalent bond with the empty *p* orbital of the boron atom of the other molecule, forming a six-membered ring; they later presented experimental evidence, in the form of measured dipole moments, that the ring existed as a chair conformer [10]. Kuhn et al. referred to these molecules as "BON-BON" species; unlike the original references, here we use all capital letters to emphasize the symbols of the elements in the six-membered ring. Kuhn et al.'s derivatives had several n-butyl groups bonded to either the boron atom or the nitrogen atom (or both) in the ring.

As far as we have been able to determine, thermodynamic properties of the BON-BON compounds have not been measured. We submit that the thermodynamic properties of these compounds may be intriguing, especially in regard to their enthalpies of combustion or decomposition. These molecules would be at least partially self-oxidizing, with O atoms in the ring, while at the same time forming stable $N_2$ as a combustion or decomposition product. As such, we have conducted a systematic study of a series of molecules with a BON-BON central ring. Because nitro

groups ($-NO_2$) are common in HE materials, [11] we have not only determined properties of the parent BON-BON molecule, but of BON-BON molecules containing up to eight $NO_2$ groups.

## Computational details

All calculations were performed using the Gaussian 09 computational chemistry program [12] on an IBM cluster 1350 supercomputer at the Ohio Supercomputer Center in Columbus, Ohio. The method used was the density functional theoretical method as defined using Becke's three-parameter exchange functional with the correlation functional of Lee, Yang, and Parr (abbreviated B3LYP in the Gaussian program) [13, 14] along with the standard Gaussian basis set labeled 6-31G(d,p) [15]. Minimum-energy geometries of the BON-BON molecules were determined using default settings, and vibrational frequency calculations were performed to verify that a minimum-energy geometry was found. Once the proper structure of the BON-BON molecule was established, the enthalpy of formation of was determined by calculating the enthalpy change for the molecule formed from its gaseous elements and then corrected for the enthalpy of formation of gas-phase boron. For example, the reaction for the parent molecule was:

$$4H_2(g) + O_2(g) + N_2(g) + 2B(g) \rightarrow \text{cyclo} - BH_2ONH_2BH_2ONH_2(g),$$

whose energy change was determined from the calculations and then corrected for the formation of two moles of B(g):

$$2[B(s) \rightarrow B(g)] \qquad \Delta H = 2\left[565.0 \text{ kJ mol}^{-1}\right].$$

The enthalpy of formation for B(g) was found at the NIST Chemistry Webbook website [5]. Once corrected for the formation of B(g), the energy represents the enthalpy of formation of the BON-BON molecule. After this, enthalpies of combustion and/or decomposition can be determined using standard balanced reactions, assuming that the products are $B_2O_3(s)$, $H_2O(\ell)$, and $N_2(g)$, whose thermodynamic properties were also found at the NIST Chemistry Webbook [5]. When necessary, $O_2(g)$ is added as a reactant for complete oxidation of B and H in the molecules.

## Results and discussion

In the course of our systematic study of BON-BON and its nitro derivatives, we explored all substitutional isomers and found a total of 42 isomers ranging from the parent molecule, cyclo-$BH_2ONH_2BH_2ONH_2$, to octanitro-BON-BON, cyclo-$B(NO_2)_2ON(NO_2)_2B(NO_2)_2ON(NO_2)_2$. Further, we only explored positional isomers, not axial-vs.-equatorial isomers that may be possible for multiple $NO_2$ groups (see below). Because of the sheer number of isomers in this study, in this report we will present specific data on only a few selected isomers and give complete results for all isomers only for thermodynamic data. Output files for all isomers are available upon request if additional analysis is desired.

Figure 1 shows the minimum-energy structures of the parent molecule, cyclo-$BH_2ONH_2BH_2ONH_2$; the symmetrically tetranitro-substituted cyclo-$B(NO_2)HON(NO_2)HB(NO_2)HON(NO_2)H$; and the octanitro-substituted cyclo-$B(NO_2)_2ON(NO_2)_2B(NO_2)_2ON(NO_2)_2$. We note that the BON-BON backbone shows the N atom making a

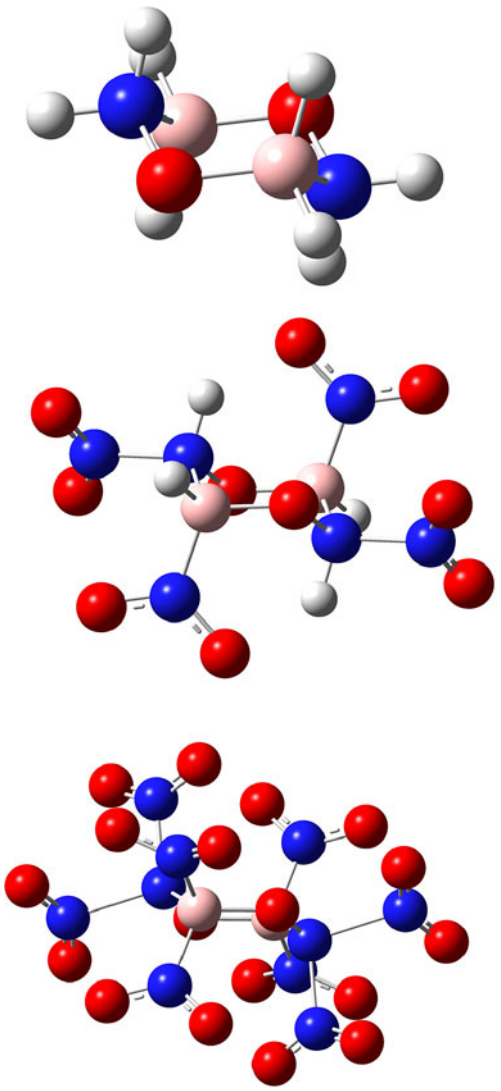


**Fig. 1** Optimized structures of the parent BON-BON molecule, the symmetrically-substituted tetranitro-BON-BON, and octanitro-BON-BON. In the bottom molecule, the BON-BON central ring is almost obscured by the $NO_2$ groups

coordinate covalent bond to the B atom of the other BON moiety, and the six-membered ring is puckered, resembling the chair conformation of cyclohexane. It is this conformation of the BON-BON backbone that leads to the possibility of axial or equatorial positions for substitution, which is the stereoisomerism type we are ignoring here.

Table 1 lists representative bonding parameters of these three molecules; bonding parameters of the other substitutions and their isomers are intermediate between these. There are several trends worth discussing. First, upon successive nitration of the BON-BON ring, the six-membered ring shrinks markedly (except for the B-O bond increasing 0.001 Å going from the tetranitro-BON-BON to the octanitro-BON-BON). The six-membered ring opens up a bit, as seen by the B-O-N angle going from 109.6º to 120.6º and the N-B-O angle increasing from 105.7º to 110.9º. The O-N-B bond angle, however, stays relatively constant from one extreme to the other, 111.9º to 112.8º. This opening-up, however, is compensated for by a decrease in the ring pucker, measured by the boron atom in the ring being out of the central plane by 63.0º for the parent BON-BON but only 45.7º for the octanitro-BON-BON.

Perhaps the most noteworthy bond parameter change relates to the bond to the $NO_2$ group. The B-N(nitro) bond distance stays relatively constant at around 1.56 – 1.56 Å. In N-nitro-BON-BON (not shown here), the N-N(nitro) bond optimized at 1.60 Å. In the tetranitro-BON-BON molecule (Fig. 1 and Table 1), the N-N(nitro) bond remains at about 1.60 Å. However, as the nitro content of the ring increases above four nitro groups, the N-N(nitro) bond increases, in some cases substantially. In the octanitro-BON-BON, one of the N-N(nitro) bonds increases markedly to almost 1.90 Å, an unusually long bond distance for any type of covalent bond. The CRC Handbook [16] list mean N-N bond distances of 1.454 Å for pyramidal-pyramidal N-N molecules (sample size $n$=44), 1.420 Å for planar-pyramidal N-N molecules ($n$=68), and 1.401 Å for planar-planar N-N molecules ($n$=40). In octanitro-BON-BON, we have four pyramidal-planar N-N bonds, two being relatively short (1.69 Å) and two being relatively long (1.89 Å). Even the "relatively short" bond distance of 1.69 Å is long for this type of N-N bond. Compare this with the N-N bond distances for some more simple compounds: 1.43 Å for nitroamine, $NH_2NO_2$ [17; experimental value] but 1.52 – 1.69 Å for 1,1-dinitrohydrazine [18; computational]. Thus, while there is some evidence for long N-N bonds in nitro-N compounds, an N-N bond length of 1.89 Å is extremely long for a covalent bond. However extreme, the trend seems unmistakable, as examination of the structures of the other derivatives demonstrates a slowly increasing N-N(nitro) bond distance, starting at 1.66 Å for N-nitro-BON-BON (that is, the lone $NO_2$ group is bonded to a nitrogen atom in the ring) and increasing as the number of nitro groups on the N atoms increase and as the nitro content of adjacent B and N atoms in the ring increases. The first effect is likely due to the electron-withdrawing nature of the $NO_2$ group, decreasing the internuclear electron density between the two nitrogen atoms. The second circumstance can be explained the bulky $NO_2$ group, as steric effects become important with nitro groups are bonded to adjacent B and N atoms in the ring.

**Table 1** Selected bonding parameters of the parent BON-BON, tetranitro-BON-BON, and octanitro-BON-BON (see Fig. 1; r in Å, angles in degrees)

| BON-BON derivative | Parent | Tetranitro- | Octanitro- |
|---|---|---|---|
| r(B-N) [ring] | 1.624 | 1.614 | 1.577 |
| r(N-O) [ring] | 1.429 | 1.399 | 1.351 |
| r(B-O) | 1.476 | 1.460 | 1.461 |
| r(B-H) | 1.205, 1.214 | 1.186 | — |
| r(N-H) | 1.020, 1.028 | 1.044 | — |
| r(B-N) [nitro] | — | 1.590 | 1.562, 1.595 |
| r(N-O) [nitro] | — | 1.189 - 1.249 | 1.174 - 1.235 |
| r(N-N) | — | 1.605 | 1.696 - 1.899 |
| α(B-O-N) | 109.6 | 112.8 | 120.6 |
| α(O-N-B) | 111.9 | 110.8 | 112.8 |
| α(N-B-O) | 105.7 | 105.5 | 110.9 |
| α(O-B-N) [nitro] | — | 112.1 | 108.6 |
| α(O-N-N) [nitro] | — | 110.9, 114.9 | 107.2 - 113.5 |
| α(O-N-O) [nitro] | — | 123.1, 134.2 | 124.3 - 140.6 |
| Ring angle | 63 | 61.8 | 45.7 |

The net effect suggests that the N-N(nitro) bond is the likely site for initiation of detonation or decomposition, a point already made by Murray et al. [19]. To support this, we estimated the N-N(nitro) bond energy by calculating the single-point energies of an $NO_2$ fragment and the rest of the octanitro-BON-BON molecule, with both radicals having the same geometries as they have in the optimized octanitro-BON-BON molecule (i.e. vertical bond energies as opposed to adiabatic bond energies, the energy change as determined when the fragments are allowed to geometrically re-optimize). Ignoring energy contributions due to thermal energies and differing multiplicities, we calculated a $\Delta E$ of 0.04685 h, or 123.0 kJ $mol^{-1}$. This is comparable to but less than the bond energy in elemental fluorine (E[F-F]= 158 kJ $mol^{-1}$) and the N-$NO_2$ bond in methyl nitramine (195 kJ $mol^{-1}$) and methyl dinitramine (185 kJ $mol^{-1}$; [19]), suggesting a stable but easily broken bond.

Table 2 lists the calculated enthalpies of formation, determined as described above, and the resulting enthalpies of combustion or decomposition for the various substitutional isomers of each level of nitration. In this table, the isomers are listed by the atom(s) that have a hydrogen atom substituted by a $NO_2$ group. Any BON-BON molecule can

**Table 2** Oxygen balances, enthalpies of formation and combustion/decomposition, and specific enthalpies of reaction for BON-BON and its nitrated derivatives

| Compound | OB% | $\Delta H_f$, kJ mol$^{-1}$ | $\Delta H_{comb/dec}$, kJ mol$^{-1}$ | $\Delta H_{comb/dec}$, kJ g$^{-1}$ |
|---|---|---|---|---|
| $BH_2ONH_2$-$BH_2ONH_2$ | −89.3 | −313.1 | −1490.2 | −16.6 |
| Nitro-BON-BON | −29.7 | | | |
| B- | | −442.9 | −1239.4 | −9.21 |
| N- | | −225.8 | −1456.6 | −10.8 |
| Dinitro-BON-BON | 0 | | | |
| B,B- | | −518.3 | −1043.2 | −5.81 |
| B,B′- | | −517.8 | −1043.6 | −5.81 |
| B,N- | | −345.3 | −1216.2 | −6.77 |
| B,N′- | | −336.8 | −1224.7 | −6.82 |
| N,N- | | −135.3 | −1426.1 | −7.94 |
| N,N′- | | −126.9 | −1434.6 | −7.99 |
| Trinitro-BON-BON | 17.8 | | | |
| B,B,B′- | | −607.1 | −833.5 | −3.71 |
| B,B,N- | | −412.5 | −1028 | −4.58 |
| B,B,N′- | | −400.6 | −1039.9 | −4.63 |
| B,N,B′- | | −230.7 | −1209.8 | −5.39 |
| B,N,N- | | −227.8 | −1212.7 | −5.4 |
| B,N,N′- | | −243.1 | −1197.4 | −5.33 |
| B,N′,N′- | | −227.8 | −1212.7 | −5.4 |
| N,N,N′- | | −33.5 | −1407.1 | −6.26 |
| Tetranitro-BON-BON | 29.7 | | | |
| B,B,B′,B′- | | −687.9 | −631.7 | −2.34 |
| B,B,B′,N′- | | −326.2 | −993.4 | −3.68 |
| B,B,N,N- | | −297.8 | −1021.8 | −3.79 |
| B,N,B′,B′- | | −475.9 | −843.7 | −3.13 |
| B,N,B′,N′- | | −341.5 | −978.2 | −3.63 |
| B,N,N′,N′- | | −495.3 | −824.3 | −3.06 |
| N,N,B′,N′- | | −130.8 | −1188.8 | −4.41 |
| N,N,N′,N′- | | 53.2 | −1372.8 | −5.09 |
| Pentanitro-BON-BON | 38.1 | | | |
| B,B,N,N,B′- | | −386.6 | −812.1 | −2.58 |
| B,B,N,B′,N′- | | −358.9 | −839.8 | −2.67 |
| B,B,N,B′,B′- | | −555.9 | −642.8 | −2.04 |
| B,B,N,N,N′- | | −172.4 | −1026.3 | −3.26 |
| B,B,N,N′,N′- | | −167.7 | −1031 | −3.28 |
| B,B,B′,N′,N′- | | −358.1 | −840.6 | −2.67 |
| B,N,N,B′,N′- | | −233.2 | −965.5 | −3.07 |
| B,N,N,N′,N′- | | 9.2 | −1207.9 | −3.84 |
| Hexanitro-BON-BON | 44.5 | | | |
| B,B,N,N,B′,B′- | | −418.9 | −658.9 | −1.83 |
| B,B,N,N,B′,N′- | | −222.6 | −855.2 | −2.38 |
| B,B,N,B′B′N′- | | −417.5 | −660.3 | −1.84 |
| B,B,N,B′,N′,N′- | | −226.1 | −851.7 | −2.37 |
| B,B,N,N,N′,N′- | | −56.3 | −1021.4 | −2.84 |
| B,N,N,B′,N′,N′- | | −49.7 | −1028.1 | −2.86 |
| Heptanitro-BON-BON | 49.4 | | | |
| B,B,N,N,B′,B′,N- | | −283.1 | −673.8 | −1.67 |
| B,B,N,N,B′,N′,N′- | | −112.1 | −844.8 | −2.09 |
| Octanitro-BON-BON | 53.4 | −96.5 | −739.5 | −1.64 |

be thought of as being composed of two BON units. Nitration in the initial BON unit is indicated with unprimed labels; if an isomer has $NO_2$ groups on both BON units, the sites in the other BON half of the ring are indicated with primes (′) on the label. The table also lists the oxygen balances (OB%) of each level of nitration. For these compounds, the oxygen balance is given by the expression [20]:

$$\mathrm{OB\%} = -\frac{3200\left(\frac{3}{4}b + \frac{1}{4}h + 0n - \frac{1}{2}o\right)}{\mathrm{MW}}, \quad (1)$$

where $b$, $h$, $n$, and $o$ are the numbers of boron, hydrogen, nitrogen, and oxygen atoms in the molecular formulas, respectively, and MW is the molar mass of the molecule. An OB% that is negative indicates that a molecular formula does not have sufficient oxidizer (for these molecules, oxygen) present to completely oxidize all other atoms present, while an OB% of zero or greater indicates that a molecular formula does contain sufficient oxidizer to completely oxidize all other atoms fully. Thus, substances with negative OB% values need additional oxidizer (assumed here to be molecular oxygen) and the enthalpy changes of reaction with said oxidizer are appropriately labeled enthalpies of combustion ($\Delta H_{comb}$). Substances with zero or positive OB% values have sufficient oxidizer to oxidize without the need for another reactant, so the enthalpy changes of reaction are more appropriately labeled enthalpies of decomposition ($\Delta H_{dec}$). As the OB% values in Table 2 demonstrate, only the parent BON-BON molecule and nitro-BON-BON need additional oxygen to react, so the reaction enthalpies for those compounds are properly termed $\Delta H_{comb}$. For all other BON-BON derivatives, OB% values are zero or positive, so the reaction enthalpies are properly termed $\Delta H_{dec}$.

The parent BON-BON, cyclo-$BH_2ONH_2BH_2ONH_2$, is calculated to have an enthalpy of formation of −313 kJ $mol^{-1}$ and a combustion enthalpy of −1490 kJ $mol^{-1}$. Because energy per unit mass is a useful concept for HE materials, this converts to −16.6 kJ $g^{-1}$ for the specific enthalpy of combustion for cyclo-$BH_2ONH_2BH_2ONH_2$. This is fairly high for an HE material; compare it to −6.2 kJ $g^{-1}$ for nitroglycerine and about −5.0 kJ $g^{-1}$ for both RDX and HMX. [11] However, each mole of cyclo-$BH_2ONH_2BH_2ONH_2$ requires 2.5 mol of $O_2$ to oxidize fully, which at ~80 g is almost as massive as the 1 mol of BON-BON. Still, the high value of its specific enthalpy of combustion makes cyclo-$BH_2ONH_2BH_2ONH_2$ worth additional study.

There are two possible isomers of nitro-BON-BON, one with the nitro group attached to a B atom and one with the nitro group attached to an N atom. Table 2 shows the B-bonded nitro-BON-BON has an enthalpy of formation of −443 kJ $mol^{-1}$, while the N-bonded nitro-BON-BON has an enthalpy of formation of −226 kJ $mol^{-1}$. This simplest nitro derivative clearly demonstrates the trend that is also found in derivatives with higher levels of nitration: bonding an $NO_2$ group to a B atom yields a molecule that is about 200 kJ $mol^{-1}$ more stable than bonding an $NO_2$ group to an N atom, all other things being the same.

Despite having strongly negative enthalpies of formation, the two nitro-BON-BON derivatives have large enthalpies of combustion and specific enthalpies of combustion. At −9 to -10 kJ per gram, nitro-BON-BON has a specific energy density larger than most current HE materials.

The isomers of more highly nitrated BON-BON mirror the trend in enthalpies of formation when going from a B-substituted BON-BON to an N-substituted BON-BON. In all cases, the isomers that have the most B-$NO_2$ groupings in the molecule are the most stable, while the ones having the most N-$NO_2$ groupings are the least stable. This trend reaches its extreme for tetranitro-BON-BON, for which the B,B,B′,B′-bonded isomer is over 700 kJ $mol^{-1}$ more stable than the N,N,N′,N′-bonded isomer. (Indeed, this last isomer is one of the few nitro-BON-BON molecules that has a positive enthalpy of formation.) This difference also has an impact on the specific enthalpy of decomposition, as the B,B,B′,B′-bonded isomer has less than half the energy per unit mass as the N,N,N′,N′-bonded isomer.

As the nitro content of the BON-BON molecule increases, the enthalpies of formation generally increase, although the exact value of $\Delta H_f$ depends strongly on how many B versus how many N sites are bonded to an $NO_2$. Because of this, the enthalpies of combustion or decomposition generally increase as well. It may be tempting to report average enthalpies of combustion or decomposition for each level of nitration, but given the variability in $NO_2$ sites in these molecules, we suggest that such an average would be misleading. One other obvious trend is that the specific enthalpy of combustion or decomposition, in units of energy per unit mass, decreases significantly as the level of nitration increases. While nitro-BON-BON has a specific enthalpy of combustion of around −10 kJ $g^{-1}$, this value drops to about one sixth of this value for octanitro-BON-BON. The molar enthalpy of decomposition does not decrease its magnitude as much, although there is a noticeable trend of such as the level of nitration increases. However, the relatively large mass of a $NO_2$ group (46 u) means that the energy change per unit gram drops faster than energy change per unit mole. This is similar to trends we have noted for other compounds having low to high $NO_2$ content [18, 21–23].

In summary, our results show that while the parent BON-BON compound or low-nitrated BON-BON compounds have energy densities that rival current HE materials, BON-BON compounds of higher $NO_2$ content have significantly lower specific enthalpies of decomposition. Of course, a good HE material must have other desirable characteristics, but it seems unlikely that a compound giving off low energy per unit mass would be a viable target for additional studies.

**Acknowledgments** We would like to thank the Ohio Supercomputer Center in Columbus, OH, for a grant of time to perform this work. AKL thanks the McNair Scholars Program at Cleveland State University for support.

## References

1. Badgujar DM, Talawar MB, Asthana SN, Mahulikar PP (2008) J Hazard Mater 151:289–305
2. Richard RM, Ball DW (2007) J Mol Struct THEOCHEM 814:91–98
3. Hillegass J Jr, Ball DW (2007) J Undergrad Chem Res 6:178–185
4. Richard RM, Ball DW (2008) J Mol Struct THEOCHEM 851:284–293
5. NIST Chemistry Webbook. http:// webbooknistgov. Accessed 20 April 2011
6. Richard RM, Ball DW (2008) J Mol Model 14:871–878
7. Richard RM, Ball DW (2007) J Mol Struct THEOCHEM 823:6–15
8. Tricamo AJ, Knaus KJ, Ball DW (2007) J Mol Struct THEOCHEM 807:67–72
9. Kuhn LP, Inatome N (1963) J Am Chem Soc 85:1206–1207
10. Thomson HB, Kuhn LP, Inatome N (1964) J Phys Chem 68:421–422
11. Akhavan J (2004) The Chemistry of Explosives 2nd edition. Royal Society of Chemistry, London
12. Frisch MJ, Trucks GW, Schlegel HB et al (2009) Gaussian 09, Revision A01. Gaussian Inc, Wallingford, CT
13. Becke AD (1993) J Chem Phys 98:5648–5652
14. Lee C, Yang W, Parr RG (1988) Phys Rev 37:785–789
15. Hariharan PC, Pople JA (1973) Theor Chem Acc 28:213–222
16. Lide DR (ed) (2008) CRC Handbook of Chemistry and Physics, 89th edn. CRC Press, Boca Raton, FL
17. Tyler JK (1963) J Mol Spectrosc 11:39–46
18. Ball DW (2006) J Mol Struct THEOCHEM 773:1–7
19. Murray JS, Concha MC, Politzer P (2009) Mol Phys 107:89–97
20. Persson PA, Holmberg R, Lee J (1993) Rock Blasting and Explosives Engineering. CRC Press, Boca Raton, FL
21. Richard RM, Ball DW (2008) J Mol Struct THEOCHEM 858:85–87
22. Lawong AK, Ball DW (2009) J Mol Struct THEOCHEM 916:33–36
23. Wagner RR, Ball DW (2011) J Mol Model. 2011 Feb 12 (Epub ahead of print)

J Mol Model (2012) 18:1667–1676
DOI 10.1007/s00894-011-1188-z

ORIGINAL PAPER

# Structural and electronic properties of Z isomers of (4α→6′′,2α→O→1′′)-phenylflavans substituted with R=H, OH and $OCH_3$ calculated in aqueous solution with PCM solvation model

**Rosana Maria Lobayan · Erika N. Bentz · Alicia H. Jubert · Alicia B. Pomilio**

Received: 26 May 2011 / Accepted: 12 July 2011 / Published online: 3 August 2011


**Abstract** In the search for new antioxidants, flavan structures called our attention, as substructures of many important natural compounds, including catechins (flavan-3-ols), simple and dimeric proanthocyanidins, and condensed tannins. In this work the conformational space of the Z-isomers of (4α→6′′, 2α→O→1′′)-phenylflavans substituted with R=H, OH and $OCH_3$ was scanned in aqueous solution, simulating the solvent by the polarizable continuum model (PCM). Geometry optimizations were performed at B3LYP/6-31 G** level. Electronic distributions were analyzed at a better calculation level, thus improving the basis set (6-311++G**). A topological study based on Bader′s theory (atoms in molecules) and natural bond orbital (NBO) framework was performed. Furthermore, molecular electrostatic potential maps (MEPs) were obtained and thoroughly analyzed. The stereochemistry was discussed, and the effect of the solvent was addressed. Moreover, intrinsic properties were identified, focusing on factors that may be related to their antioxidant properties. Hyperconjugative and inductive effects were described. The coordinated NBO/AIM analysis allowed us to rationalize the changes of MEPs in a polar solvent. To investigate the molecular and structural properties of these compounds in biological media, the polarizabilities and dipolar moments were predicted which were further used to enlighten stability and reactivity properties. All conformers were taken into account. Relevant stereoelectronic aspects were described for understanding the stabilization and antioxidant function of these structures.

**Keywords** (4α→6′′,2α→O→1′′)-Phenylflavans · Antioxidants · Aqueous solvent effect · Atoms in molecules · Molecular dipole moment · Natural bond orbital analysis

Rosana M. Lobayan and Erika N. Bentz contributed equally to this work

**Electronic supplementary material** The online version of this article (doi:10.1007/s00894-011-1188-z) contains supplementary material, which is available to authorized users.

R. M. Lobayan (✉) · E. N. Bentz
Facultad de Ingeniería, Universidad de la Cuenca del Plata,
Lavalle 50,
3400 Corrientes, Argentina
e-mail: rlobayan@ucp.edu.ar

A. H. Jubert
CEQUINOR Facultad de Ciencias Exactas y Facultad de Ingeniería, Universidad Nacional de La Plata,
CC 962,
1900 La Plata, Argentina

A. B. Pomilio
PRALIB (UBA, CONICET), Facultad de Farmacia y Bioquímica,
Universidad de Buenos Aires,
Junín 956,
C1113AAD Buenos Aires, Argentina

## Introduction

The role of flavonoids as antioxidants is the subject of intense theoretical and experimental research. The biological and pharmaceutical activities of these compounds have been related to their antioxidant activity as free radical scavengers. Free radicals can damage biomolecules such as lipids, amino acids, proteins, carbohydrates and nucleic acids by the process of oxidative damage resulting in several diseases and aging [1]. They were also associated with deterioration of plant tissues, such as fruits and vegetables. The biological activity of flavonoids is governed by electronic interactions of the biomolecules within the cell [2]. Their antioxidant activity

depends on the capacity to donate protons and electrons [3] and an adequate solubility [4]. Moreover, the ease of deprotonation of the flavonoid determines the intermediates that are formed, and in turn, the stability of the intermediates determines the antioxidant capacity [4, 5].

In the human body flavonoids show a lot of biological properties as antioxidants, antiallergenic, antibacterial, antifungal, antiviral and anticarcinogenic agents. These features provide them useful pharmacological properties to treat diseases ranging from allergies, bacterial and viral infections, to those of greater risk such as heart diseases, cancer and HIV [6].

The study of the electronic and molecular properties of flavonoids is of great importance since it enables a better understanding of the mechanism of their antioxidant activity, which is so far controversial due to the lack of a thorough knowledge of both the molecular structure of these compounds and the structure-property relationship. Therefore, theoretical investigations of the physical and chemical properties, and structural features of flavonoids are very important to disclose the relationship between the structure, properties and performance, and to assist in the design and synthesis of new derivatives with improved properties.

4-Phenylflavans with an additional ethereal bridge comprise an interesting structure, which is included in simple and dimeric A-type proanthocyanidins that have been previously synthesized with various substituents in the aromatic rings [7, 8].

To our knowledge there are no crystallographic reported data of the A-type proanthocyanidins, and there has been no previous study on A-type proanthocyanidins using high-level ab initio calculations, except for a few theoretical studies on their electronic, structural, and energetic properties, which we have earlier reported, where the conformational space of unsubstituted and substituted 4-phenylflavans was scanned in gas phase [9, 10]. In addition, intramolecular bonding interactions were thoroughly described and related to their stability and reactivity. The stereochemistry of the molecules and the analysis of the effect of substitution on the behavior of the electron distribution were also studied in our laboratories [10].

Theoretical results on isolated molecules are very useful for determining the intrinsic properties of the system free of any interactions but can only be compared with a few experimental cases, since most experiments are performed in solution. Furthermore, the analysis of the molecular behavior in the presence of water is of great interest owing to the stabilizing function of proanthocyanidins in solution, their ability to form complexes with other biomolecules, and to interfere with biological systems. For these reasons, this paper focuses on the analysis of solvent effects of water on the intrinsic electronic properties and molecular geometry, simulated by a polarizable continuum model.

The ultimate aim of this work is to complete the first step in studying the electronic structure, and the quantification of stereoelectronic and substituent effects to contribute to the study of new antioxidants. Furthermore, this knowledge will be a basis for studying in another step, other substitution patterns and the ability to scavenge free radicals.

## Methods

The geometries from simulation by the MD module of the HyperChem [11] package were optimized to an energy gradient less than 0.004 kJ $mol^{-1}$ $Å^{-1}$ at a semiempirical level (AM1). The lowest-energy conformers were further optimized in gas phase [9, 10] by the density functional theory (DFT), as implemented by the Gaussian 03 package [12]. Geometry optimizations were performed using the Becke three-parameter hybrid functional combined with Lee-Yang-Parr correlation functional [13], giving rise to the well-known B3LYP method. The 6-31 G** basis set was used for all atoms. The optimized geometric structures were confirmed as minimum on the potential energy surface through an analysis of harmonic vibrational frequencies.

The lowest-energy conformers obtained through gas-phase calculations were re-optimized taking into account the solvent effect by the conductor polarizable continuum model (PCM) [14] at the same level of theory. The PCM approach considers that the molecule under study is embedded in a polarizable dielectric representing the solvent, without explicit solvent molecules included. The polarizable dielectric medium was described by the dielectric constant of the solvent ($\varepsilon$=78.39 for water). The values of surface area and volume of the cavity were in the range of 333.61-334.85 $Å^2$ and 387.34-390.57 $Å^3$, respectively, for R=OH, and in the range of 333.87-396.92 $Å^2$ and 390.19-473.65 $Å^3$, respectively, for R=$OCH_3$. This model has proven to be a reliable tool for the description of electrostatic solute-solvent interactions. At the same level of calculation the zero-point energy (ZPE) was obtained, which was used to correct all energy terms.

Maps of electrostatic potential (MEP) were obtained in the van der Waals molecule surfaces using the Gaussian 03 software, and further visualized by MOLEKEL 4.0 [15].

Topological analysis and evaluation of local properties were performed by the PROAIM software [16] using the wave functions calculated at the B3LYP level and the improved 6-311++G** basis set implemented in the G03 program. NBO analysis was carried out at the same level [17]. The B3LYP functional has shown a good correlation with experimental data concerning studies of the activity of phenolic antioxidants [18-20]

## Results and discussion

The main substructure of the non-substituted (4α→6′′, 2α→*O*→1′′)-phenylflavan was a [3.1.3]bicycle (Fig. 1), which consisted of two 6-membered rings, C and E, each involved in a benzo-γ-pyran, fused by a $CH_2$-3-containing bridge with bridgehead C-2 and C-4. This bridge is a stereocenter that gives rise to *E* (*anti*)/*Z* (*syn*) isomers. Steric hindrance and high ring tension were typical of the *E*-isomer [9]. Consequently, only the conformers of Z-isomer were analyzed thoroughly.

The C-3—C-2—C-1′—C-6′ dihedral angle in relation to ring B had a mean value of 88° for the structures that made up the socalled Z1 group, and about 0° for the other Z2 group. Regardless of the substitution, Z1 conformers were on average 1.91 kcal $mol^{-1}$ more stable than the Z2. Substitution in the D ring led to eight conformers for R=OH (termed $Z1_{CT}$, $Z1_{CC}$, $Z1_{TC}$, $Z1_{TT}$, $Z2_{CT}$, $Z2_{CC}$, $Z2_{TC}$ y $Z2_{TT}$), and three conformers for R=$OCH_3$ (termed $Z1_{CT}$, $Z1_{CC}$ y $Z2_{CC}$) (see Supplementary material; Figs. S1 and S2, respectively). The distinction between the "T" and "C" subscripts for conformers refers to *anti* (*trans*) and *syn* (*cis*) configurations of H—O-3′′ and H—O-5′′ in relation to the C-3"—C-4" and C-5"—C-4" bonds, respectively (Fig. 1). The average values of the H—O-3′′—C-3′′—C-4′′ and H—O-5′′—C-5′′—C-4′′ dihedral angles were about 180° (*anti*) or near 0° (*syn*). The characteristics of the conformers obtained in aqueous phase were similar to those obtained in gas phase [10]. The Z1 rotamers were the most stable in aqueous phase as in gas phase, showing the following stability order, CT>CC>TC> TT (Table 1).

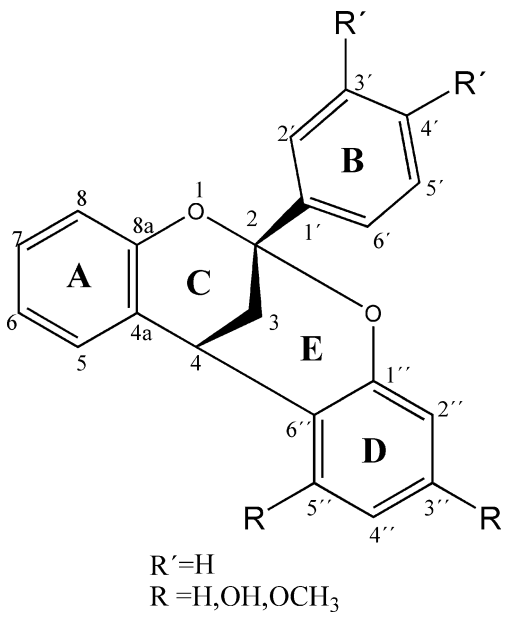



**Fig. 1** Structure of Z-isomers of (4α→6′′,2α→O→1′′)-phenylflavans substituted with R=H, OH and $OCH_3$. Atom numbering is indicated

The relative conformational energies between the two most stable conformers were reduced in solution. This finding agrees with the behavior found in anthocyanidins by Estévez and Mosquera [21].

The difference of energy values between the structures in aqueous solvent and gas phase are shown in Table 1. For example, -22 kcal $mol^{-1}$ on average for R=OH and −11 kcal $mol^{-1}$ on average for R=$OCH_3$. As expected, all structures were stabilized in solution. The stabilization was higher for R=OH than for R=$OCH_3$. The Z2 rotamers are always more stabilized than Z1.

The effect of solvent on the structure stabilization of each conformer was not uniform. Relative conformational populations at 298 K calculated with the Boltzmann distribution showed considerable variation in solution, especially in the structures of the most stable conformers. In gas phase, the relative population of $Z1_{CC}$ for R=OH was ca. 19% with 0.79 kcal $mol^{-1}$ difference with respect to $Z1_{CT}$ conformer ($\Delta E_{vacuum}$). Instead, the corresponding population in aqueous phase was ca. 42% with 0.06 kcal $mol^{-1}$ difference with respect to $Z1_{CT}$ ($\Delta E_{solution}$). The relative population of the $Z2_{CC}$ conformer in aqueous phase also increased, being ca. 1% in gas phase and ca. 4% in solution.

A similar behavior was found for the R=$OCH_3$ substitution. The relative population of $Z1_{CC}$ ranged from ca. 18% in gas phase ($\Delta E_{vacuum}$=0.91 kcal $mol^{-1}$) to ca. 24% in solution ($\Delta E_{solution}$=0.66 kcal $mol^{-1}$). The relative population of Z2 conformer accounted for ca. 1% in vacuum, and ca. 2% in solution (Table 1).

Regarding the geometries optimized in solution, the optimized structural parameters showed slight changes compared to those in gas phase. Bond lengthenings were less than 0.03 Å for C—C, C—H and O—H. Bond angle variations were less than 1.5°. Changes in the dihedral angles C3—C2—C1′—C6′, C3—C2—C1′—C2′, C3—C2—C1′—C2′ and O—C2—C1′—C2′ were lower than 2º. For $Z1_{TC}$ and $Z1_{TT}$ conformers, the dihedral angle H—O5′′—C5′′—C4′′ varied in 23º; for $Z2_{TC}$ and $Z2_{TT}$ conformers the dihedral angle varied between 47º and 50º. The dihedral angle H—O3′′—C5′′—C4′′ changed 2º for TT conformers (see Supplementary material; Table S1).

### Topology of the electronic charge density function and NBO analysis

Relevant information of the molecular structure can be extracted from an analysis of the topology of the molecular charge density, ρ(r), whose main features are summarized by the curvatures of ρ(r) at critical points (CPs) [22].

**Table 1** Energies calculated at B3LYP/6-31 G** level of theory for the (4α→6′′, 2α→*O*→1′′)-phenylflavans substituted with R=H, OH and $OCH_3$[a]

| | | Vacuum | | Solution | | |
|---|---|---|---|---|---|---|
| | Conformer | Relative population (%) | $\Delta E^b$ (kcal $mol^{-1}$) | Relative population (%) | $\Delta E^b$ (kcal $mol^{-1}$) | $\Delta^c$ |
| R=OH | | $E_0$=−696831.64 kcal $mol^{-1}$ | | $E_0$=−696853.90 kcal $mol^{-1}$ | | |
| | $Z1_{CT}$ | 72.60 | 0.00 | 46.99 | 0.00 | −22.26 |
| | $Z1_{CC}$ | 19.03 | 0.79 | 42.35 | 0.06 | −22.99 |
| | $Z1_{TC}$ | 2.23 | 2.06 | 1.12 | 2.21 | −22.11 |
| | $Z1_{TT}$ | 2.11 | 2.09 | 0.46 | 2.73 | −21.62 |
| | $Z2_{CT}$ | 3.07 | 1.87 | 4.64 | 1.37 | −22.76 |
| | $Z2_{CC}$ | 0.79 | 2.67 | 4.34 | 1.41 | −23.52 |
| | $Z2_{TC}$ | 0.08 | 4.02 | 0.05 | 4.05 | −22.23 |
| | $Z2_{TT}$ | 0.08 | 4.02 | 0.05 | 4.04 | −22.24 |
| $R=OCH_3$ | | $E_0$=−746124.30 kcal $mol^{-1}$ | | $E_0$=−746135.48 kcal $mol^{-1}$ | | |
| | $Z1_{CT}$ | 81.74 | 0.00 | 73.40 | 0.00 | −11.18 |
| | $Z1_{CC}$ | 17.55 | 0.91 | 24.13 | 0.66 | −11.43 |
| | $Z2_{CC}$ | 0.71 | 2.81 | 2.47 | 2.01 | −11.98 |
| R=H | | $E_0$=−602433.03 kcal $mol^{-1}$ | | $E_0$=−602433.17 kcal $mol^{-1}$ | | |
| | Z1 | 95.93 | 0.00 | 95.94 | 0.00 | −0.14 |
| | Z2 | 4.07 | 1.87 | 4.06 | 1.87 | −0.14 |

[a] Energies corrected for zero point energy (ZPE)

[b] ΔE represents the difference of energy between the most stable conformer ($Z1_{CT}$) and each of the remaining conformers

[c] Δ represents the difference of energy values between the structures in aqueous solvent and gas phase

Through the analysis of (3,-1) CPs (bond critical points or BCPs), polar covalent interactions were characterized for C—H bonds, and all C—C bonds of rings A, B, and D in all conformers under study. The C-2—O1 bond of ring C also showed the same typical behavior of a covalent bond. The C-8a—O1 bond of ring C, and the equivalent for ring E, C-1′′—O1 and C—O bonds accounting for $R=OCH_3$ and R=OH, did not behave like C-2—O1, according to values for an intermediate or shared polar covalent bond (see Supplementary material; Tables S2 and S3).

Relevant changes were found in (3,-3) CPs (nuclear critical points or NCPs) in solution, corresponding exclusively to hydrogen atoms. A decrease of 0.7-1.2% of the electron density at NCP of benzenic hydrogen was found in solution. At NCPs of hydrogen of the hydroxy group change was even more relevant; a decrease of 7.3% was observed with respect to the values in gas phase. These changes were remarkable in relation to the features of the maps of electrostatic potential (MEPs) as shown below.

As in gas phase [10], the average value of bond ellipticity (ε) in each ring increased in the order of B<A<D. In terms of the orbital model of electronic structure, the ellipticity is a measure of the extend of π-character up to the limit of a "double bond" for which ε reaches a maximum [22]. Therefore our results can be rationalized through an enhanced ability of ring D to act on electrophilic aromatic substitution reactions due to the increased availability of π-electrons also in aqueous phase. In addition, this feature was enhanced with R=OH substitution in solution (see Supplementary material; Tables S2 and S3), which was also found in vacuum [10]. In general, the average ε values decreased in solution (changes less than 1.7%). Ring B showed the greatest changes.

In the NBO analysis, the electronic wave functions are interpreted in terms of occupied Lewis-type and unoccupied non-Lewis localized orbitals. Delocalizations of electron density between occupied Lewis-type (bonds or lone pairs) NBO orbitals and formally unoccupied (anti-bonds or Rydberg) non-Lewis NBO orbitals account for stabilizing donor–acceptor interactions.

For each donor NBO (i) and acceptor NBO (j), the stabilization energy ($E^{(2)}$) associated with i/j delocalization is explicitly estimated by the following equation:

$$E^{(2)} = -n_i \frac{F_{ij}^2}{\varepsilon_j - \varepsilon_i}, \quad (1)$$

where $\eta_i$ is the ith donor orbital occupancy; $\varepsilon_i$, $\varepsilon_j$, are diagonal elements (orbital energies); and $F_{i,j}$, off-diagonal elements, respectively, associated with the Fock matrix in NBO.

A detailed NBO analysis allowed us to describe the resonance of the lone pairs of O and O1 oxygens with features similar to those found in gas phase [10]. In vacuum

as in solution we characterized two lone pairs for each oxygen atom. One of them, *1n*, is **sp**-type and the other one, *2n*, is **p**-type. Each one played a different role. Also, the substitution in solution led to C-2—C-3 bond more polarized than C-3—C-4, and that polarization increased as R=H<R=$OCH_3$<R=OH. We also found that the conformational changes were correlated with the relative stability of the structure through the same hyperconjugative mechanisms described in vacuum. Instead of these similarities we found interesting changes in solution.

Tables 2 and 3 show the percentage values of the main changes for second-order stabilization energies (*x*) of relevant $\sigma \rightarrow \sigma^*$ and $1,2n_{O,\ O1} \rightarrow \sigma^*$, $\pi^*$ charge transfers ($\%\Delta_{\sigma\rightarrow\sigma^*}$ and $\%\Delta_{1,2nO,\ O1\rightarrow\sigma^*,\ \pi^*}$, respectively) calculated from

$$\%\Delta x = \frac{x(solution) - x(vacuum)}{x(vacuum)} \times 100 \qquad (2)$$

in average for all conformers. The *x (vacuum)* values are those of the corresponding second-order energies in gas phase.

i) In solution as in vacuum the $\sigma_{C\text{-}3—H} \rightarrow \sigma^*_{C—O}$ transfers were higher than those of $\sigma_{C\text{-}3—H} \rightarrow \sigma^*_{C—C}$. Moreover, in solution both kinds of electron delocalizations were enhanced; e.g., transfers to C—O bonding orbital increased 3.4-8.0%, and electron transfers to C—C orbital bonding increased 2.2-5.8%, the effect being improved for R=OH (Table 2).

**Table 2** Percentage values of the main changes for second-order stabilization energies, $E^{(2)}$, calculated at B3LYP/6-311++G** level of theory for relevant transferences[a,b]

| Donor | Acceptor | Z1 | Z2 |
|---|---|---|---|
| $\sigma_{C\text{-}3—H}$ | $\sigma^*_{C\text{-}4—C\text{-}6''}$ | 5.8 | 2.6 |
| $\sigma_{C\text{-}3—H}$ | $\sigma^*_{C\text{-}2—O}$ | 6.3 | 8.0 |
| $\sigma_{C\text{-}3—H}$ | $\sigma^*_{C\text{-}4—C\text{-}4a}$ | 2.2 | 4.8 |
| $\sigma_{C\text{-}3—H}$ | $\sigma^*_{C\text{-}2—O1}$ | 4.9 | 3.4 |
| $\sigma_{O\text{-}3''—H}$ | $\sigma^*_{C—C}$[c] | 4.1 | 4.1 |
| $\sigma_{O\text{-}5''—H}$ | $\sigma^*_{C—C}$[c] | 5.0 | 1.4 |
| $\sigma_{C\text{-}2—C\text{-}1'}$ | $\sigma^*_{C\text{-}1'—C\text{-}6'}$ | 0.0 | −2.7 |
| $\sigma_{C\text{-}2—C\text{-}1'}$ | $\sigma^*_{C\text{-}1'—C\text{-}2'}$ | 2.2 | 1.4 |
| $\sigma_{C\text{-}2—C\text{-}1'}$ | $\sigma^*_{C\text{-}5'—C\text{-}6'}$ | −0.9 | −1.1 |
| $\sigma_{C\text{-}2—C\text{-}1'}$ | $\sigma^*_{C\text{-}2'—C\text{-}3'}$ | −0.7 | −0.3 |
| $\sigma_{C\text{-}5'—C\text{-}6'}$ | $\sigma^*_{C\text{-}4'—H}$ | −2.6 | −2.7 |
| $\sigma_{C\text{-}2'—C\text{-}3'}$ | $\sigma^*_{C\text{-}4'—H}$ | −2.7 | −2.9 |
| $\sigma_{C\text{-}2—C\text{-}3}$ | $\sigma^*_{C\text{-}1—C\text{-}6'}$ | 7.1 | — |

[a] All values are expressed in kcal $mol^{-1}$

[b] Values in average for all conformers of the (4α→6′′,2α→*O*→1′′)-phenylflavans substituted with R=OH and $OCH_3$

[c] C—C bond correspond to bonds of D ring

ii) The second-order stabilization energies of $\sigma_{O\text{-}3''—H} \rightarrow \sigma^*$ and $\sigma_{O\text{-}5''—H} \rightarrow \sigma^*$ transfers were increased up to 5.0% in solution (Table 2).

iii) Charge transfers from O-3′′ and O-5′′ lone pairs to the D-ring antibonding orbitals increased 2.3-8.6% for R=OH in solution. The same trend was observed for R=$OCH_3$, but the effect was attenuated (Table 3). Our results showed an increase in the delocalization effects for R=OH in solution, and higher solvent effects for these compounds. Interestingly, the improved donor role of O-3′′ and O-5′′ lone pairs for R=OH can be explained by their hybridization. In fact, we found that in solution the *p*-character of *1n* bonding orbitals of O-3′′ and O-5′′ significantly increased for compounds with R=OH. We propose that the enhanced delocalization in connection with O-3′′ and O-5′′ is based on the solvent effect on the hybridization of lone pairs. The increase of *p*-character of *1n* lone pairs allows a better overlap between them and the C—C antibonding orbitals of ring D (Table 4).

The effects of charge delocalization *i* - *iii* allowed us to rationalize the enhanced structure stabilization of compounds with R=OH in solution (Table 1).

iv) The lone pairs of the oxygens O and O1 showed a different behavior. In fact, the O and O1 lone pair transfers decreased 16.1% in solution (Table 3). Second-order stabilization energies for $1n_{O1} \rightarrow \sigma^*_{C\text{-}8a—C\text{-}8}$ transfers, and their symmetry equivalents ($1n_O \rightarrow \sigma^*_{C\text{-}1''—C\text{-}2''}$) decreased 5.2-12.1% for R=OH and 1.7-5.4% for R=$OCH_3$. Our results show a lower donor role of the O and O1 lone pairs in solution, an effect that is enhanced in compounds with R=OH.

v) In connection with the decrease of hyperconjugative interactions of the oxygen lone pair $1n_{O,\ O1} \rightarrow \sigma^*_{C\text{-}2—C\text{-}1'}$ in solution (see Table 3), we found that the $\sigma_{C\text{-}2—C\text{-}1'} \rightarrow \sigma^*_{C\text{-}1'—C\text{-}6'}$, $\sigma^*_{C\text{-}5'—C\text{-}6'}$, $\sigma^*_{C\text{-}2'—C\text{-}3'}$ transfers also decreased (Table 2). These findings explained the decrease in the average B-ring ellipticity shown above, and allowed us to explain the decrease in the π-character of ring B in terms of a reduction in the effects of charge delocalization related to O and O1 lone pairs in solution (anomeric effect).

## Molecular electrostatic potential maps

The molecular electrostatic potential maps have been used extensively to predict the behavior and reactivity of a wide variety of chemical systems [23-25]. The V(r) potentials, created in the space around a molecule by its nuclei and electrons, are a useful tool for the study of molecular reactivity. Unlike other current parameters used as reactivity indices, V(r) is an actual physical property that can be

**Table 3** Percentage values of the main changes for second-order stabilization energies, $E^{(2)}$, calculated at B3LYP/6-311++G** level of theory for donation transferences of oxygen lone pairs for Z1 conformers[a]

| | | R=OH | | | | R=$OCH_3$ | |
|---|---|---|---|---|---|---|---|
| Donor | Acceptor | $Z1_{CT}$ | $Z1_{CC}$ | $Z1_{TC}$ | $Z1_{TT}$ | $Z1_{CT}$ | $Z1_{CC}$ |
| 1nO1 | $\sigma^*_{C\text{-}8a—C\text{-}8}$ | −7.1 | −7.1 | −12.1 | −10.3 | −5.4 | −1.8 |
| 1nO1 | $\sigma^*_{C\text{-}4a—C\text{-}8a}$ | −0.8 | −1.1 | −3.0 | −2.6 | −1.1 | −0.1 |
| 2nO1 | $\pi^*_{C\text{-}4a—C\text{-}8a}$ | −2.6 | −3.2 | −5.1 | −4.4 | −3.3 | −2.4 |
| 1nO | $\sigma^*_{C\text{-}1''— C\text{-}6''}$ | — | −4.1 | −0.7 | 3.0 | −1.1 | — |
| 1nO | $\sigma^*_{C\text{-}1''—C\text{-}2''}$ | — | — | −5.2 | −6.8 | −5.4 | −1.7 |
| 2nO | $\pi^*_{C\text{-}1''—C\text{-}2''}$ | −6.2 | −5.4 | −2.6 | −3.4 | −3.2 | −1.7 |
| 1nO | $\sigma^*_{C\text{-}2—C\text{-}1'}$ | −2,4 | −3,6 | −10,3 | −9,3 | −3,6 | −2,4 |
| 2nO | $\sigma^*_{C\text{-}2—C\text{-}1'}$ | 3,1 | 1,0 | 16,2 | 17,7 | −10,3 | 1,0 |
| 1nO1 | $\sigma^*_{C\text{-}2—C\text{-}1'}$ | −9.6 | −8.5 | −2.5 | −2.5 | −7.4 | −2.5 |
| 2nO1 | $\sigma^*_{C\text{-}2—C\text{-}1'}$ | 2.5 | 3.5 | −15.8 | −16.1 | 13.2 | 0.8 |
| 1nO1 | $\sigma^*_{C\text{-}2—C\text{-}3}$ | −3.0 | −2.7 | −4.8 | −5.2 | −0.2 | −1.5 |
| 2nO1 | $\sigma^*_{C\text{-}2—C\text{-}3}$ | −7.5 | −7.4 | 9.2 | 9.3 | −13.4 | −1.3 |
| 2nO1 | $\sigma^*_{C\text{-}2—O}$ | −4.1 | −3.3 | 0.8 | 0.5 | −2.9 | 0.1 |
| 1n O-3′′ | $\sigma^*_{C\text{-}2''—C\text{-}3''}$ | 4.6 | 3.8 | 5.3 | 5.1 | 0.7 | 8.7 |
| 2n O-3′′ | $\pi^*_{C\text{-}3''— C\text{-}4''}$ | 2.8 | 2.3 | 2.8 | 3.5 | −0.8 | −1.1 |
| 1n O-5′′ | $\sigma^*_{C\text{-}4''—C\text{-}5''}$ | 7.1 | 5.3 | 8.1 | 7.8 | 2.2 | 1.2 |
| 2n O-5′′ | $\pi^*_{C\text{-}5''— C\text{-}6''}$ | 7.5 | 8.6 | −10.1 | −8.6 | 4.4 | 2.4 |
| 1n O-3′′ | $\sigma^*_{C\text{-}3a''—H}$ | — | — | — | — | −11.2 | −7.6 |
| 2n O-3′′ | $\sigma^*_{C\text{-}3a''—H}$ | — | — | — | — | −5.4 | −3.2 |
| 2n O-3′′ | $\sigma^*_{C\text{-}3a''—H}$ | — | — | — | — | −6.4 | −4.2 |
| 1n O-5′′ | $\sigma^*_{C\text{-}5''—C\text{-}6''}$ | — | — | — | — | 2.9 | — |
| 1n O-5′′ | $\sigma^*_{C\text{-}5a''—H}$ | — | — | — | — | −3.0 | −1.8 |
| 1n O-5′′ | $\sigma^*_{C\text{-}5a''—H}$ | — | — | — | — | −13.8 | −13.3 |
| 1n O-5′′ | $\sigma^*_{C\text{-}5a''—H}$ | — | — | — | — | — | 3.7 |
| 2n O-5′′ | $\sigma^*_{C\text{-}5a''—H}$ | — | — | — | — | −5.0 | −3.3 |
| 2n O-5′′ | $\sigma^*_{C\text{-}5a''—H}$ | — | — | — | — | −5.7 | −3.2 |

[a] All values are expressed in kcal $mol^{-1}$

determined experimentally and by computational methods. The calculation of physicochemical properties on the molecular surface, and its visualization by a color code allows a different view of molecular behavior.

In this paper we used the procedure proposed by Politzer et al. [23] to predict targeted sites for electrophilic attack in the regions of negative V(r) values. Moreover, the study was enriched by a thorough analysis of electrostatic potential maps, in light of the complementarities of the different theoretical tools used herein, for rationalizing other aspects of their distribution [9, 10].

MEPs in gas phase and in aqueous phase are shown in Fig. 2. In all species there were two very close main sites of attack, associated with the oxygen atoms of rings C and E. The highest negative V(r) values showed that Z1 were more reactive conformers for electrophilic attack. The highest values (positive) showed a 46% increase in solution, and the lowest (negative) showed a 31 %

**Table 4** Percentage values of the change of π-character of 1n lone pairs of oxygen atoms for the (4α→6′′,2α→*O*→1′′)-phenylflavans substituted with R=OH and $OCH_3$ in aqueous solution

| | R=OH | | | | | | | | R=$OCH_3$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $Z1_{CT}$ | $Z1_{CC}$ | $Z1_{TC}$ | $Z1_{TT}$ | $Z2_{CT}$ | $Z2_{CC}$ | $Z2_{TC}$ | $Z2_{TT}$ | $Z1_{CT}$ | $Z1_{CC}$ | $Z2_{CC}$ |
| $1n_O$ | 0.3 | 0.2 | 0.5 | 0.6 | 0.3 | 0.2 | 1.1 | 1.1 | 0.1 | 0.7 | 0.2 |
| $1n_{O1}$ | 0.2 | 0.3 | 0.0 | −0.1 | 0.1 | 0.2 | −0.7 | −0.8 | 0.5 | 0.9 | 0.3 |
| $1n_{O3''}$ | 2.5 | 2.6 | 2.5 | 2.6 | 2.5 | 2.7 | 2.6 | 2.6 | 0.5 | 1.1 | 0.5 |
| $1n_{O5''}$ | 2.1 | 2.2 | 4.7 | 4.7 | 2.1 | 2.2 | 2.5 | 2.5 | 0.2 | 0.2 | 0.2 |

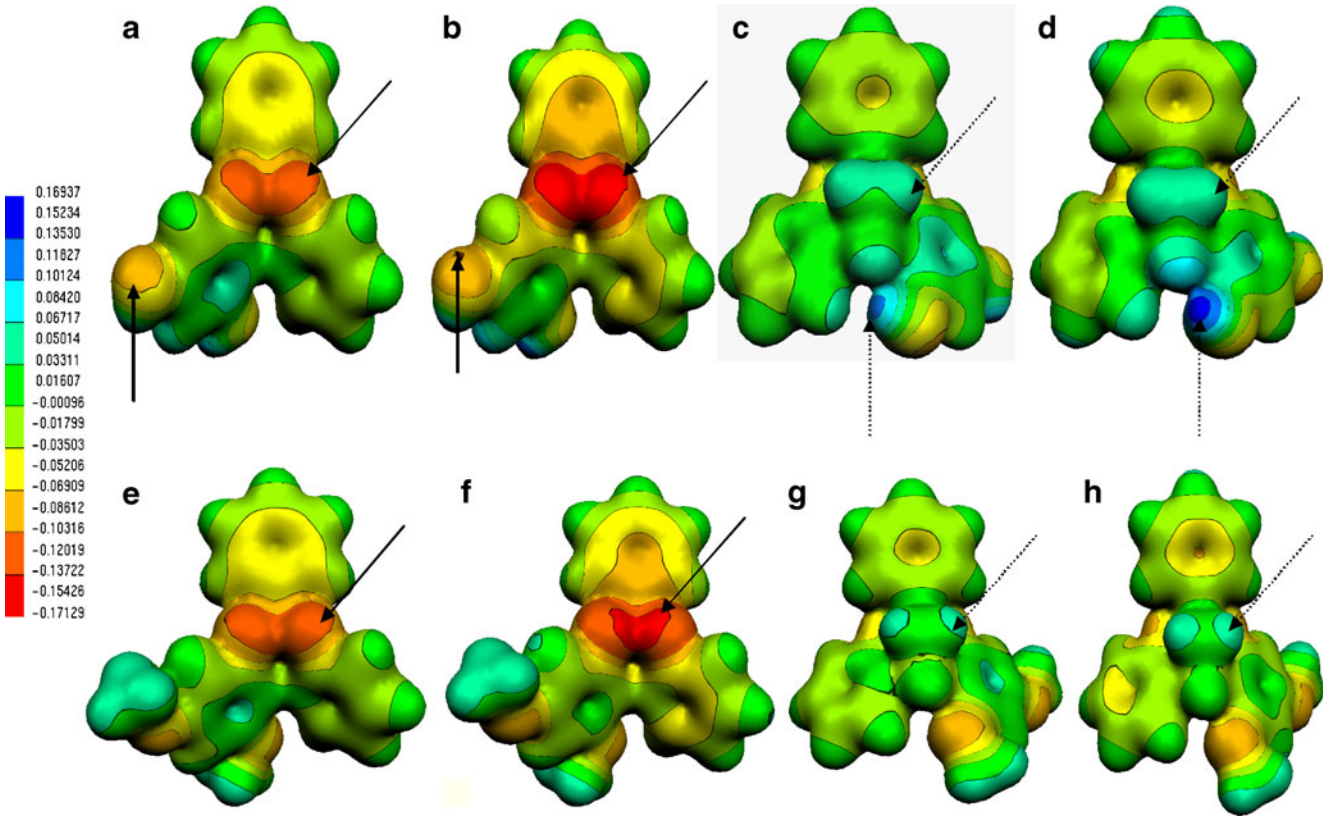



**Fig. 2** (**a**)-(**d**): Maps of molecular electrostatic potential (MEP) for $Z1_{CC}$ conformer of (4α→6′′,2α→*O*→1′′)-phenylflavans substituted with R=OH in gas phase (**a**) and water solution (**b**), and for $Z1_{TC}$ conformer in gas phase (**c**) and water solution (**d**). (**e**)-(**f**): MEPs for $Z1_{CT}$ conformer of (4α→6′′,2α→*O*→1′′)-phenylflavans substituted with R=$OCH_3$ in gas phase (**e**), (**g**), and water solution (**f**), (**h**). Solid arrows indicate the change of negative V(r) region over oxygen atoms. Dashed arrows indicate the change of positive V(r) region over hydrogen atoms in both media. The solid oblique arrows indicate the decrease in the lowest values (negative) on the O-O1 region, enhanced in compounds with R=OH than in those with R=$OCH_3$

decrease, thus demonstrating an enhanced reactivity in solution.

We also found other interesting features of MEPs, as follows:

- The most positive region on the hydrogens of $CH_2$-3 increased in solution, which can be explained by the increased donor role of the C-3—H bonding orbitals in solution, as discussed in item *i*.
- The most positive region on the hydrogens of $CH_2$-3 was higher for R=OH than for R=$OCH_3$ in vacuum as in solution, which agreed with the enhanced delocalizations found for R=OH in relation with C-3$H_2$ bonding orbitals, as described in item *i*.
- The increase in the highest values (positive) located on the hydrogens of the hydroxyls can be rationalized by charge delocalizations discussed in *ii*. In fact, these results allowed us to propose that the higher donor role of O-3′′—H and O-5′′—H bonding orbitals in solution was reflected in the increased V(r) values in this region. Interestingly, this finding correlated with the decrease of electron density at NCPs of those hydrogen atoms.
- The decrease in the lowest values (negative) on the O-O1 region can be rationalized by charge delocalizations discussed in *iv*. In fact, these results allowed us to propose that the minor donor role of the O and O1 lone pairs in solution was reflected in the increased reactivity of this region.
- The decrease in the lowest values (negative) on the O-O1 region was enhanced in compounds with R=OH. This feature can be explained by charge delocalization as discussed above, because they are lower in compounds with R=OH than in those with R=$OCH_3$ (see *iv*).

## Molecular polarizability and dipole moment

The values of dipole moment and the isotropic molecular polarizability are of interest as an indication of the solubility and chemical reactivity of the molecules under study. The isotropic polarizability is a measure of electronic distortion in a molecule, caused by an external electric field, and a good indicator of how the total charge

distribution of the flavonoid molecules is affected by the presence of different substituents at specific positions [26].

Calculation of the polarizability <α> (see Table 5) was performed from $\langle\alpha\rangle = \frac{1}{3}\left(\alpha_{xx} + \alpha_{yy} + \alpha_{zz}\right)$, where the tensor components were obtained from the second derivatives of the energy with respect to the Cartesian components of the applied electric field ε, ($\alpha = [\partial^2 E/\partial\varepsilon^2]$).

To the best of our knowledge, investigations concerning the molecular polarizabilities of proanthocyanidins have not yet been reported, and therefore, our results are the first prediction of them. The average polarizabilities of (4α→6′′,2α→O→1′′)-phenylflavans substituted with R=OH and R=$OCH_3$ were≈223 a.u. and≈249 a.u., respectively, indicating their nature of being soluble in polar solvents, and their ability to polarize other atoms or molecules.

Moreover, we found, as expected, that the Z2 structures, which are more stabilized in solution than the Z1 structures, showed the largest value of <α> in vacuum, while Z1 had minor <α>, which correlated with the lower stabilization in solution.

The permanent electric dipole moment of an isolated molecule (μ) is also an important predictor of its behavior in physical, chemical, and biological processes [27]. The dipole moment represents a generalized measure of the charge density in a molecule and is a reactivity index, which is very important to define biological properties, especially those related to the interaction with the active sites of an enzyme [2]. It can be determined by a variety of techniques. Because the magnitudes and directions of dipole moments are sensitive to molecular size and shape they can also serve as a useful tool in conformational analysis [28].

The values of the modulus of the permanent dipole moments (μ values) of Z-isomers of (4α→6′′,2α→O→1′′)-phenylflavans substituted with R=H, OH and $OCH_3$ are listed in Table 6. The observed variation allowed us to predict that the values of dipole moments can be used to distinguish between different conformers. In fact, in vacuum as in solution, μ decreased as CC>TC>TT>CT.

Besides the changes of the μ modulus, for R=OH conformers we found also an important variation of the μ vector projection on the X-Y plane, as shown in Fig. 3. This behavior can be rationalized by the analysis of the contribution of dipole moments associated to the hydroxyls following a scheme similar to that used in the vector addition model to predict the dipole moments of a compound, based on the known dipole moments of the "component parts" [28]. In fact, our results mainly showed that in vacuum as in solution, μ had two separable contributions, one of them associated to the changes in O—C-2—O1 moiety, which were uniform for all conformers (also in unsubstituted structures), and another related to the conformational changes of the hydroxy groups.

The μ variation found when all conformers are considered warns about the use of theoretical μ values considering only the most stable conformer. To take into account properly the conformational landscape, we propose a Maxwell-Boltzmann statistical average at 298.15 K of each μ Cartesian component according to:

$$\left\langle \mu_j \right\rangle = \sum_i \mu_{j,i} \exp\left(-\frac{E_i}{RT}\right) / \sum_i \exp\left(\frac{E_i}{RT}\right)$$

with *j*=*x, y, z*, where $E_i$ is the relative energy of the *i*th conformer, and $\mu_{j,i}$ is the component *j*th of the *i*th conformer. Then, obtaining the Maxwell-Boltzmann statistical average of the magnitude of total dipole moment:

$$\langle\mu\rangle = \sqrt{\langle\mu_x\rangle^2 + \left\langle\mu_y\right\rangle^2 + \langle\mu_z\rangle^2}.$$

The average Maxwell-Boltzmann values for Z1 conformers are shown in Table 6. We found an increase of 25.8% with respect to the value of the most stable conformer for R=OH, and a decrease of 43.4% with respect to the value of the most stable conformer for R=$OCH_3$. The same trend, but enhanced, was found in solution (54.8% for R=OH, and 47.1% for R=$OCH_3$).

Moreover, our calculations of molecular dipole moment and polarizabilities implied that the Z-isomers of

**Table 5** Polarizability data for the (4α→6′′,2α→*O*→1′′)-phenylflavans substituted with R=H, OH and $OCH_3$ calculated at B3LYP/6-31 G** level of theory in gas phase

| | R=OH | | | | | | | | R=$OCH_3$ | | | R=H | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $Z1_{CT}$ | $Z1_{CC}$ | $Z1_{TC}$ | $Z1_{TT}$ | $Z2_{CT}$ | $Z2_{CC}$ | $Z2_{TC}$ | $Z2_{TT}$ | $Z1_{CT}$ | $Z1_{CC}$ | $Z2_{CC}$ | Z1 | Z2 |
| $\alpha_{xx}$ | 275.64 | 279.29 | 275.45 | 271.82 | 277.60 | 281.05 | 274.94 | 271.25 | 298.65 | 311.71 | 308.03 | 260.02 | 264.69 |
| $\alpha_{yy}$ | 252.45 | 251.62 | 255.23 | 256.09 | 215.49 | 214.42 | 226.69 | 228.21 | 284.00 | 278.28 | 248.08 | 232.97 | 191.57 |
| $\alpha_{zz}$ | 137.90 | 136.20 | 136.72 | 138.47 | 176.41 | 175.10 | 170.23 | 171.49 | 162.20 | 157.63 | 194.97 | 124.66 | 165.01 |
| <α> | 222.00 | 222.37 | 222.46 | 222.13 | 223.17 | 223.52 | 223.95 | 223.65 | 248.28 | 249.20 | 250.36 | 205.89 | 207.09 |
| Average | 222.24 | | | | 223.57 | | | | 248.74 | | 250.36 | | |

[a] All values are expressed in a.u.

**Table 6** Values of the modulus of the permanent dipole moment for the (4α→6′′,2α→*O*→1′′)-phenylflavans substituted with R=H, OH and $OCH_3$ calculated at B3LYP/6-31 G** level of theory in aqueous solvent and gas phase[a]

| | | Vacuum | | Solution | | Δ[c] |
|---|---|---|---|---|---|---|
| R=OH | $Z1_{CT}$ | 1.45 | 1.83[b] | 2.32 | 3.59 [b] | 0.86 |
| | $Z1_{CC}$ | 3.56 | | 5.53 | | 1.97 |
| | $Z1_{TC}$ | 2.57 | | 5.08 | | 2.51 |
| | $Z1_{TT}$ | 2.57 | | 3.52 | | 0.95 |
| | $Z2_{CT}$ | 1.31 | | 2.18 | | 0.87 |
| | $Z2_{CC}$ | 3.28 | | 5.31 | | 2.03 |
| | $Z2_{TC}$ | 2.34 | | 4.36 | | 2.01 |
| | $Z2_{TT}$ | 2.63 | | 3.61 | | 0.98 |
| R=$OCH_3$ | $Z1_{CT}$ | 1.88 | 1.06 [b] | 1.89 | 1.00 [b] | 0.01 |
| | $Z1_{CC}$ | 3.88 | | 4.99 | | 1.11 |
| | $Z2_{CC}$ | 3.61 | | 4.77 | | 1.16 |
| R=H | Z1 | 1.20 | | 2.73 | | 1.53 |
| | Z2 | 1.89 | | 2.59 | | 0.70 |

[a] All values are expressed in Debye

[b] Values of the Maxwell-Boltzmann statistical averages for Z1 conformers

[c] Δ represents the difference of the modulus of the permanent dipole moment for the structure in aqueous solvent and gas phase

(4α→6′′,2α→O→1′′)-phenylflavans substituted with R=H, OH and $OCH_3$ are polarized systems, and consequently, soluble in polar solvents, and with ability to polarize other atoms.

As expected, increased dipole moment values were found in solution, however, the solvent effect was not uniform. In fact, we found higher changes in CC and TC structures, and lower changes when the configuration of the H—O-3′′bond was *anti* (*trans*) to the C-3”—C-4” bond.

The highest value was found in CC structures in agreement with the higher stabilization in aqueous solution. This finding explained the marked increase in relative population (ca. 19% in vacuum and ca. 42% in solution for $Z1_{CC}$; ca. 1% in vacuum and ca. 4% in solution for $Z2_{CC}$; (Table 1).

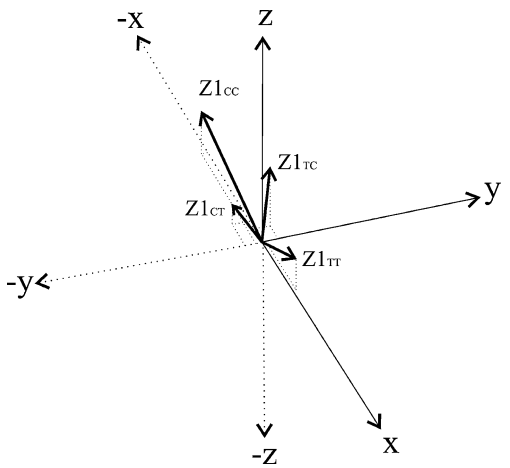



**Fig. 3** Variation of the electric dipole moment vector (μ) projection on the X-Y plane

## Conclusions

The stereochemistry of (4α→6′′,2α→*O*→1′′)-phenylflavan was studied in aqueous solution, emphasizing the description of the factors that determine it, and the changes that occur with R=$OCH_3$, R=H, and R=OH, R'=H as substituents. As in gas phase, two lowest-energy conformers were characterized for R=H, eight for R=OH, and three for R=$OCH_3$.

In relation to the structural solvation effects we reached the following conclusions:

- As expected, all structures were stabilized in solution, however, the stabilization in solution was higher for R=OH than for R=$OCH_3$. The effects of charge delocalization that we described, allowed us to rationalize it (improved structure stabilization of the compounds with R=OH in solution).
- The Z2 were always more stabilized than the Z1 rotamers. We rationalize this through the higher Z2 polarizability.
- The solvent effect on structure stabilization was not uniform for all conformers. The relative population of CC conformers increased significantly in aqueous solution in accordance with the highest values of dipole moment in CC structures.

Intramolecular interactions were studied and characterized by the theory of atoms in molecules, characterizations being similar to those in vacuum. Changes on electron density at nuclear critical points (NCPs) in solution were related to charge delocalization effects and changes at the electrostatic potential maps. We described the intrinsic properties of (4α→6′′,2α→*O*→1′′)-phenylflavan and also showed some delocalization improved (decreased) in solution as well as the topological and structural consequences.

Our study was enriched by a deep analysis of electrostatic potential maps and their changes in solution; with the description of the underlying charge delocalization effects.

Moreover, our findings allowed us to propose an enhanced stabilization in polar solvents of the cation radical and the radical generated from the scavenging process for the substitution with R=OH due to the higher resonance and further charge delocalizations found in (4α→6′′,2α→O→1′′)-phenylflavans substituted with R=OH.

To the best of our knowledge, our results are the first predictions of molecular polarizabilities and dipole moments of proanthocyanidins and phenylflavans, which led to propose the Z-isomers of (4α→6′′,2α→O→1′′)-phenylflavans substituted with R=H, OH and $OCH_3$ as

molecules soluble in polar solvents with the ability to polarize other atoms and molecules.

We remark on the use of dipole moments to distinguish between different conformers. In this regard, we calculated the dipole moment variation taking into account all conformers, using Maxwell-Boltzmann statistical average of each Cartesian μ component. Results warn about using μ values only calculated considering the most stable conformer.

In this paper we studied in depth the stereoelectronic features of the Z isomers of (4α→6′′, 2α→O→1′′)-phenylflavans substituted with R=H, OH and $OCH_3$ in a simulated aqueous solution. The study of substitution with R′=R=H, OH and $OCH_3$, and the ability to scavenge free radicals are in progress.

**Acknowledgments** Thanks are due to Agencia de Promoción Científica y Tecnológica Argentina (MINCYT), CONICET, Universidad Nacional de La Plata and Universidad de Buenos Aires (Argentina) for financial support. A.B.P. is a Senior Research Member of the National Research Council of Argentina (CONICET). A.H.J. is Member of the Scientific Research Career (CIC, Province of Buenos Aires). E.N.B. acknowledges a fellowship (IP-PRH N0 54) from Agencia de Promoción Científica y Tecnológica (Argentina) and Universidad de la Cuenca del Plata (Corrientes, Argentina). R.M.L. acknowledges Universidad de la Cuenca del Plata for facilities provided during the course of this work.

## References

1. Cotelle N (2001) Role of flavonoids in oxidative stress. Curr Top Med Chem 1:569–590
2. Olivero-Verbel J, Pacheco-Londoño L (2002) Structure−activity relationships for the anti-HIV activity of flavonoids. J Chem Inf Comput Sci 42:1241–1246
3. Martins HFP, Leal JP, Fernández MT, Lopes VHC, Cordeiro MNDS (2004) Toward the prediction of the activity of antioxidants: experimental and theoretical study of the gas-phase acidities of flavonoids. J Am Soc Mass Spectrom 15:848–861
4. Zhang HY, Sun YM, Wang XL (2003) Substituent effects on O—H bond dissociation enthalpies and ionization potentials of catechols: a DFT study and its implications in the rational design of phenolic antioxidants and elucidation of structure–activity relationships for flavonoid antioxidants. Chem A Eur J 9:502–508
5. Leopoldini M, Marino T, Russo N, Toscano M (2004) Antioxidant Properties of phenolic compounds: H-atom versus electron transfer mechanism. J Phys Chem A 108:4916–4922
6. Pinent M, Bladé C, Salvadó MJ, Blay M, Pujadas G, Fernandez-Larrea J, Arola L, Ardevol A (2006) Procyanidin effects on adipocyte-related pathologies. Crit Rev Food Sci Nutr 46:543–550
7. Pomilio A, Müller O, Schilling G, Weinges K (1977) Zur Kenntnis der Proanthocyanidine, XXII. Über die Konstitution der Kondensationsprodukte von Phenolen mit Flavyliumsalzen. Justus Liebigs Ann Chem: 597–601
8. Pomilio A, Ellmann B, Künstler K, Schilling G, Weinges K (1977) Naturstoffe aus Arzneipflanzen, XXI. 13 C-NMR-spektroskopische Untersuchungen an Flavanoiden. Justus Liebigs Ann Chem: 588–596
9. Lobayan RM, Jubert AH, Vitale MG, Pomilio AB (2009) Conformational and electronic (AIM/NBO) study of unsubstituted A-type dimeric Proanthocyanidin. J Mol Model 15:537–550
10. Bentz EN, Jubert AH, Pomilio AB, Lobayan RM (2010) Theoretical study of Z isomers of A-type dimeric proanthocyanidins substituted with R=H, OH and OCH3: stability and reactivity properties. J Mol Model 16:1895–1909
11. HyperChem Release 7.5, Hypercube Inc, Gainsville, FL
12. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JA, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C, Pople JA (2003) Gaussian 03, Revision B.02. Gaussia, Inc, Pittsburgh, PA
13. Becke AD (1993) Density-functional thermochemistry. III. The role of exact exchange. J Chem Phys 98:5648–5642
14. Miertus S, Scrocco E, Tomasi J (1981) Electrostatic interaction of a solute with a continuum. A direct utilizaion of ab initio molecular potentials for the prevision of solvent effects. J Chem Phys 55:117–129
15. Flúkiger P, Lúthi HP, Portmann S, Weber J (2000) MOLEKEL 4.0. Swiss Center for Scientific Computing, Manno, Switzerland
16. Biegler-Koning FW, Bader RFW, Tang TH (1982) Calculation of the average properties of atoms in molecules.II. J Comput Chem 3:317–328
17. Glendening ED, Reed AE, Carpenter JE, Weinhold F NBO 3.1. Program as implemented in the Gaussian 98 package
18. Wright JS, Johnson ER, DiLabio GA (2001) Predicting the Activity of phenolic antioxidants: theoretical method, analysis of substituent effects, and application to major families of antioxidants. J Am Chem Soc 123:1173–1183
19. Rasulev BF, Abdullaev ND, Syrov VN, Leszczynski J (2005) A quantitative structure-activity relationship (QSAR) study of the antioxidant activity of flavonoids. QSAR Comb Sci 24:1056–1065
20. Venkateswarlu D, Leszczynski J (1998) Tautomeric equilibria in 8-oxopurines: Implications for mutagenicity. J Comput Aided Mol Des 12:373–382
21. Estévez L, Mosquera R (2007) A Density Functional Theory Study on Pelargonidin. J Phys Chem A 111:11100–11109
22. Bader RFW (1990) A quantum theory of molecular structure and its applications. Chem Rev 91:893–928
23. Politzer P, Truhlar DG (eds) (1981) Chemical applications of atomic and molecular electrostatic potentials. Plenum, NY
24. Politzer P, Murray JS (1991) Theoretical biochemistry and molecular biophysics: a comprehensive survey, vol 2. In: Beveridge DL, Lavery R (eds) Protein. Adenine, Schenectady, pp 165–191
25. Roy DK, Balanarayan P, Gadre SR (2009) Signatures of molecular recognition from the topography of electrostatic potential. J Chem Sci 121:815–821
26. Weber KC, Honório KM, Bruni AT, da Silva ABF (2006) The use of classification methods for modeling the antioxidant activity of flavonoid compounds. J Mol Model 12:915–920
27. Bublitz GU, Boxer SG (1997) STARK SPECTROSCOPY: Applications in Chemistry, Biology, and Materials Science. Annu Rev Phys Chem 48:213–242
28. Nguyen TV, Pratt DW (2006) Permanent electric dipole moments of four tryptamine conformers in the gas phase: a new diagnostic of structure and dynamics. J Chem Phys 124:1216–1219

J Mol Model (2012) 18:1677
DOI 10.1007/s00894-012-1373-8

ERRATUM

# Erratum to: Computational prediction of selectivities in nonreversible and reversible hydroformylation reactions catalyzed by unmodified rhodium-carbonyls

**Giuliano Alagona · Raffaello Lazzaroni · Caterina Ghio**

Published online: 11 February 2012


**Erratum to: J Mol Model (2011) 17:2275–2284**
**DOI 10.1007/s00894-010-0864-8**

The original version of this article unfortunately contained a mistake. The last two terms of Eqs. (1) and (2) are not correct. Equations (1) and (2) should read, respectively:

$$\mathrm{B} : \mathrm{L} = k_B : k_L = \sum k_B[C] : \sum k_L[C] = \sum e^{-\Delta G_B^{\neq}/RT} : \sum e^{-\Delta G_L^{\neq}/RT} \approx \sum e^{-\Delta E_B^{\neq}/RT} : \sum e^{-\Delta E_L^{\neq}/RT} \quad (1)$$

and

$$b : b' = k_b : k_{b'} = \sum e^{-\Delta G_b^{\neq}/RT} : \sum e^{-\Delta G_{b'}^{\neq}/RT} \approx \sum e^{-\Delta E_b^{\neq}/RT} : \sum e^{-\Delta E_{b'}^{\neq}/RT} \quad (2)$$

The online version of the original article can be found at http://dx.doi.org/10.1007/s00894-010-0864-8.

G. Alagona (✉) · C. Ghio (✉)
Molecular Modeling Lab,
CNR-IPCF (Institute for Physico-Chemical Processes),
Via Moruzzi 1,
56124 Pisa, Italy
e-mail: G.Alagona@ipcf.cnr.it
e-mail: C.Ghio@ipcf.cnr.it

R. Lazzaroni
DCCI (Department of Chemistry and Industrial Chemistry),
University of Pisa,
Via Risorgimento 35,
56126 Pisa, Italy