Theor Chem Acc (2004) 111:270–279
DOI 10.1007/s00214-003-0522-7

Theoretical
Chemistry Accounts

Regular article

QM/MM connection atoms for the multistate treatment of organic and biological molecules

A. Toniolo[1], **C. Ciminelli**[1], **G. Granucci**[1], **T. Laino**[2], **M. Persico**[1]

[1] Dipartimento di Chimica e Chimica Industriale, Università di Pisa, v.Risorgimento 35, 56126, Pisa, Italy
[2] NEST-INFM Scuola Normale Superiore di Pisa, p.zza dei Cavalieri 7, 56125, Pisa, Italy

Received: 18 February 2003 / Accepted: 30 April 2003 / Published online: 31 December 2003

Abstract. We present an extension of our semiempirical floating occupation MO-CI approach for the determination of ground and excited state potential energy surfaces of interest in photochemistry. The QM/MM variant of the method, which allows for electrostatic and van der Waals interactions between the QM and MM subsystems, is supplemented with a treatment of covalent interactions based on Antes and Thiel's connection atom approach. We concentrate on the correct treatment of electrostatic interactions concerning the connection atom, on the specific requirements for the representation of excited states, and on the transferability of the optimal parameters. We show the viability of the method with four examples of connection atoms: S in a thioether bridge, acylic C, aliphatic C, and N in a peptide. The results obtained with the QM/MM treatment compare well with all-QM results of the same level.

Keywords: Hybrid QM/MM methods – Connection atom – Semiempirical methods – Photochemistry – Photobiology

Introduction

Ab initio Quantum Mechanical (QM) methods have been shown over the years to be effective in providing insight about the structure, energetics and dynamical processes of molecular systems. They have been successful in representing chemical events that cannot be modeled by simple force fields, like bond breaking and formation, and electronic transitions. In general, they are able to represent changes in the electronic and nuclear structure along a given reaction coordinate. Despite the fact that there has been remarkable improvement in ab initio methods and algorithms, applications are still limited by their computational cost. Aside from any consideration of the accuracy of a particular method, they still appear to be limited to systems containing tens or a few hundreds of atoms. However, if one is interested in studying excited states, the use of correlated electronic wave functions is mandatory for reliable results, so the computational demand is consequently increased. The applicability of ab initio techniques is further limited in the field of dynamic simulations, because of the very large number of calculations needed to build a potential energy surface (PES) or run direct dynamics [1, 2].

Semiempirical methods [3–5] have long been appreciated for their speed and overall reliability for ground state molecules. The NDO model in its standard implementations (MNDO [6, 7], PM3 [8], AM1 [9]) has been parameterized primarily for geometries and energies of organic molecules using an SCF wave function. However, when used in conjunction with correlated wave functions, and applied for instance to excited states, the regular parameters do not guarantee reliable results. Reoptimization of semiempirical parameters, usually for a specific molecule, is a necessary step in order to obtain a fast method able to give acceptable results for large systems. In particular, our group has developed a new technique which extends the applicability of Configuration Interaction in the semiempirical context. Instead of using canonical SCF monoelectronic functions, we compute Floating Occupation Molecular Orbitals (FOMO) [2, 10]. Their occupation numbers vary according to the molecular geometry: as a result, homolytic dissociation is correctly treated and orbital degeneracy is automatically satisfied if necessary. The virtual orbitals are partially optimized and the quality of the CI development is enhanced. In this framework, standard semiempirical parameters are even less suitable and a reoptimization is mandatory for good results. This strategy has already been applied to run "on the fly" dynamics of several processes, including the

Contribution to the Jacopo Tomasi Honorary Issue

Correspondence to: M. Persico
e-mail: mau@dcci.unipi.it

photodissociation of ClOOCl [11] and the decay of the first two excited states of benzene [12].

On the other hand, methods based on classical force fields (MM) have been parameterized in order to give quantitative agreement with structures and energetics of ground state molecules, as determined experimentally or through ab initio calculations. The molecular energy is decomposed into two-, three- and four-body interaction terms that can be computed extremely quickly. Of course, any classical force field is intrinsically inadequate to represent events where the quantum nature of the electronic motion is manifest, such as excited states, electronic transitions and bond rearrangements. However, if the portion of the molecular system where the electrons must be explicitly treated with a Quantum Mechanical approach is localized, then a mixed QM/MM method can be resorted to. Indeed, QM/MM strategies join the reliability and general applicability of quantum mechanics to the effectiveness of molecular mechanics, so that complex systems can be investigated. Since the pioneering work done in the early 70's (see for instance [13, 14]) and especially after the seminal paper of Warshel and Levitt (1976) [15], many variants of the method have been developed and successfully applied (for reviews see references [16, 17]). In a preceding paper [18], we described a QM/MM extension of our semiempirical FOMO-SCF-CI method. In our approach, as in some others too, the QM and MM subsystems interact through electrostatic terms which are included in the QM Hamiltonian, so as to influence directly the computed wavefunctions. This feature is essential for a correct treatment of surface crossing situations. Additional Lennard-Jones terms provide a representation of the dispersion-repulsion interactions between the two subsystems. This method can be applied to QM molecules interacting in non-covalent ways with MM solvents, solid surfaces, biological matrices, and so on. In the present paper, we want to focus on the QM/MM boundary, in cases where the two regions are linked together through a covalent bond. Different approaches have been proposed. The first idea was to saturate the valence of the QM part by adding a monoelectronic atom (link atom [19–22], usually a hydrogen), invisible to the MM part of the system. Several variants of the link atom method have been proposed, with the aim to overcome the ambiguities in the treatment of the interactions involving the link atom [23–26].

Another approach involves the use of localized orbitals centered on a boundary atom which has both QM and MM nature [27]. The boundary atom carries atomic basis functions but it also concurs in defining the MM potential through the frontier. In the Local SCF variant developed by Rivail and coworkers [28–30] the boundary atom is bound to three other QM and one MM atoms. It forms four hybrid sp^3 orbitals. Three of them, pointing toward the QM atoms, are included in the SCF and contribute to the definition of the molecular orbitals, while the fourth one is oriented along the QM-MM bond and is kept frozen throughout the SCF optimization. Direct comparison on benchmark compounds, with care taken in partitioning and treatment of electrostatic, link atom and Local SCF methods, have shown comparable results: neither gave systematically better results [31]. A similar approach has been developed by Gao who used Generalized Hybrid Orbitals [32]. Here, the boundary atom is connected to three MM atoms and one QM atom. Only the hybrid orbital pointing at the QM atom is included in the SCF procedure, while the other three are frozen.

A third strategy uses an adjusted connection atom, as demonstrated by Antes and Thiel in a semiempirical scheme [33]. Here, the boundary atom is a hydrogenoid atom with one *s* atomic orbital. Since it replaces a carbon atom, the basis function is of 2 *s* type. Its electron interacts quantum-mechanically with the rest of the QM portion of molecule while the interaction with the MM atoms is ensured by mechanical and electrostatic embedding (model B in [33]). The wave function is determined under the influence of an electric field produced by all the MM charges excluding those centered on atoms directly bound to the connection atom. This simple ansatz is quite adequate for alkyl residues, such as those treated by Antes and Thiel.

In this paper, we present a variant of the connection atom approach, the main difference with respect to Antes and Thiel residing in the treatment of the electrostatic interaction between the boundary atom and the MM environment and in the parameterization, which is adapted to our FOMO-SCF-CI treatment. We show how to obtain the semiempirical parameters in benchmark molecules and then we apply them to two molecular systems. The first one is a dipeptide containing a triptophane and a glutammic acid, which can be considered a general piece of a polypeptide with low lying excited states and the ability to establish strong intramolecular hydrogen bonds. It may constitute an example of how to split a proteic polymer into QM and MM parts. The second one is azobenzenophane, as an example of a big organic ring.

In the following section we describe our method, and in the section after that we discuss how the semiempirical parameters have been reoptimized for four types of connection atoms. Then, in the next two sections, we report applications of our method to two large molecules. Conclusions are offered in the final section.

Theory

The total Hamiltonian in any QM/MM partitioning strategy can be written as a sum of three terms:

$$H = H_{\mathrm{QM}} + H_{\mathrm{QM/MM}} + H_{\mathrm{MM}} \tag{1}$$

In our approach, as in Antes and Thiel's [33], the QM portion includes the connection atom (CA) and its electron. H_{QM} is the semiempirical Hamiltonian for all the QM atoms. H_{MM} contains all of the terms provided by the adopted force field and involving only MM atoms; since it does not depend on the electronic coordinates, it is a state-independent additive term.

$H_{QM/MM}$ contains all of the coupling terms between the QM and MM portions. At the present level of theory, these are of three kinds:

$$H_{QM/MM} = H^{Coul}_{QM/MM} + H^{vdW}_{QM/MM} + H^{CA}_{QM/MM} \quad (2)$$

The electrostatic term contains the Coulomb interactions of QM cores and electrons with MM atomic charges:

$$H^{Coul}_{QM/MM} = \sum_{\alpha,m} \frac{q_\alpha q_m}{R_{\alpha m}} - \sum_{i,m} \frac{q_m}{R_{im}} \quad (3)$$

Here and in the following i, α and m number electrons, QM nuclei and MM atoms, respectively; q_α is the core charge of the QM atom α and q_m is the atomic charge of the MM atom m. The second term in $H^{Coul}_{QM/MM}$ is added to H_{QM} to perform the calculation of the electronic wavefunctions and energies. According to the rules of NDO, only two-center one-electron integrals need to be evaluated; they are of the type $h_{\mu\nu} = \langle\mu|-q_m/R_{im}|\nu\rangle$, where μ and ν are basis functions centered on the same QM atom. The approximate form of the $h_{\mu\nu}$ integrals we have actually used is given in Appendix A.

The van der Waals interactions between QM and MM atoms are:

$$H^{vdW}_{QM/MM} = 4\sum_{\alpha,m} \varepsilon_{\alpha m}\left[\left(\frac{\sigma_{\alpha m}}{R_{\alpha m}}\right)^{12} - \left(\frac{\sigma_{\alpha m}}{R_{\alpha m}}\right)^{6}\right] \quad (4)$$

with $\varepsilon_{\alpha m} = \sqrt{\varepsilon_\alpha \varepsilon_m}$ and $\sigma_{\alpha m} = \sqrt{\sigma_\alpha \sigma_m}$. The summation runs over all α, m pairs. As an exception, as in most MM force fields, the 1–2 and 1–3 interactions (atoms connected by one or two covalent bonds) are discarded: this situation occurs across the QM/MM boundary and involves the CA. $H^{vdW}_{QM/MM}$, as well as the first term in $H^{Coul}_{QM/MM}$, is simply added to the total energy and is not state-specific.

The $H^{CA}_{QM/MM}$ term groups MM-like contributions which involve the CA, at least one MM atom, and possibly one or two QM atoms. It is mainly through this term that the PES's depend on distances, angles and dihedrals involving covalent bonds across the QM/MM boundary.

Contrary to all other QM atoms, the CA has a fractional core charge:

$$q_{CA} = Q + 1 - \sum_m q_m \quad (5)$$

In this way, we can attribute to each MM atom the atomic charge q_m as prescribed by the adopted force field, still preserving the total charge Q of the QM + MM system. Usually $Q = 0$; nevertheless, each of the QM and MM subsystems may be charged. In practice, the q_{CA} charge is mainly determined by the MM atomic charges of the groups directly bound to the CA: for instance, if the CA is a carbon atom bound to oxygen and/or nitrogen q_{CA} exceeds 1 and induces the expected polarization in the next QM bond. The CA also interacts electrostatically with all of the MM atoms, through both the q_{CA} charge and the electronic density, as expressed in Eq. 3. In most MM force fields (for instance, in AMBER) the electrostatic 1–2 and 1–3 interactions are not computed and the other MM interactions (bond stretching, bending, and so on) are parameterized accordingly. Since we prefer not to alter the MM parameters concerning the CA, we should treat the electrostatics with consistency; in other words we should include all of the interactions of the CA with MM atoms, except for its first and second neighbors. These interactions are essentially proportional to $q_{CA} + q_{el}$, where $q_{el} \simeq -1$ is the CA electronic charge. Therefore, when computing the 1–2 and 1–3 electrostatic terms in $H^{CA}_{QM/MM}$ (and only in these cases), we use a core charge $q'_{CA} = 1$, which cancels out q_{el} approximately and annihilates the total interaction.

Optimization of semiempirical parameters

The semiempirical parameters of the connection atoms (in a similar way to those of the other QM atoms) have to be optimized: an atom containing one electron only has to reproduce the geometry and the electronic structure of the entire environment, like in a full QM calculation. The general approach to this problem is to list a certain number of properties (at least geometrical data and energies) of one or more molecules that contain the same chemical environment for the connection atom, and reproduce the same properties at QM/MM level. The QM/MM calculations have been carried out with a development version of the MOPAC [34] semiempirical code merged with the AMBER [35] force field. In the following, all of the QM atoms in the QM/MM calculations, apart the CA's, are represented by regular or previously reoptimized AM1 parameters. The optimization consists of the minimization of the target function:

$$F = \sqrt{\frac{\sum_i \omega_i \left(d_{s,i} - d_{0,i}\right)^2}{\sum_i \omega_i}} \quad (6)$$

over all the molecules and data. Here d_s are the semiempirical results, d_0 is the reference data (obtained from experiment and/or good quality calculations) and w_i are weights. In this work, we consider a single molecule as the benchmark and we show that our approach generates parameters that are transferable to other, more complex, systems. This is reported in the next sections.

As a first case we optimized the sulfur atom in the simplest thioether, $(CH_3)_2S$, because of the interest in sulfur bridges in organic and biological chemistry. In the QM/MM calculation a methyl group was included in the QM part (semiempirical CAS-CI with four electrons in four orbitals, using standard AM1 parameters and floating occupation molecular orbitals with a Gaussian width equal to 0.2 hartree), the other methyl being treated at MM level with the GAFF-AMBER force field, explicitly developed for organic molecules, with atomic charges computed ab initio, as described in ref. [36]. The central sulfur atom was, of course, the connection atom, characterized by a single 3 *s* atomic basis function (see Fig. 1) and the core charge was equal to 0.8976. The data included in the optimization were the S-C(QM) distance, the Mulliken charges of all of the QM atoms, and the energy profile as a function of the S-C(QM) bond distance, computed with four rigid displacements of ± 0.1 and ± 0.05 Å from the equilibrium geometry. Note that the Mulliken charge of the connection atom incorporates its core charge q_c. The corresponding reference data were obtained with an MP2 ab initio calculation with the cc-pVTZ Dunning's basis set. The optimization was carried out with a simulated annealing procedure [37]. Table 1 reports reference ab initio and optimized semiempirical QM/MM

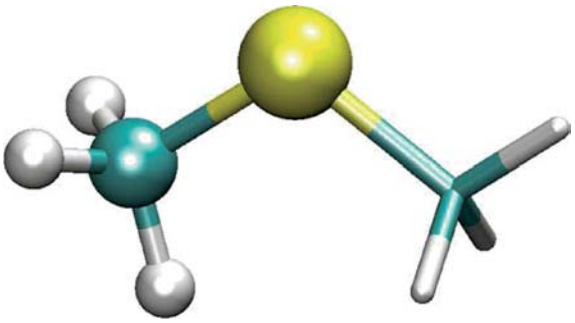

Fig. 1. The model compound $(CH_3)_2S$ for the parameterization of S as connection atom. Balls represent QM atoms, sticks MM atoms

results which agree very well. We took care to ensure that the lowest excited state lies at least 5.5 eV above the ground state at the equilibrium geometry, to ensure that the connection atom does not produce fictitious low-lying excited states involving the adjacent σ and σ^* orbitals.

The second benchmark system was a substituted glycine. It was saturated with a NH-CH_3 group on the acylic side and with a CO-CH_3 on the aminic one (see Fig. 2). This molecule was used to obtain three different sets of parameters for connection atoms: the amidic nitrogen (N), the acylic carbon (C) and the tetrahedral carbon (Ct). In the first case, the connection atom (N) is bound to a QM carbon with sp^3 hybridization, to an MM hydrogen, and to the MM CO-CH_3 group (see Fig. 2, upper panel). In the second case the CA (acylic carbon) was surrounded by an sp^3 QM carbon, the MM oxygen, and the MM NH-CH_3 group (see Fig. 2, middle panel). The aliphatic carbon Ct is bound to the acylic QM carbon, two MM hydrogens and an MM nitrogen (see Fig. 2, lower panel). These three connection atoms were provided with a 2 *s* atomic basis function. The QM reference calculation was a semiempirical AM1 CI with eight electrons in eight orbitals, limited to triple excitations only, with floating occupation molecular orbitals (width = 0.2 hartree). In the QM/MM calculations the CI active space was limited to four electrons in four orbitals. The MM force field was the standard AMBER for proteins, which also supplies the values of the atomic charges. Accordingly, the core charges of the connection atoms were 0.7003 for N, 1.7133 for C and 0.9764 for Ct.

As in the case of the sulfur atom, the target data were the distance between the CA and the adjacent QM sp^3 carbon, the Mulliken charges of the QM subsystem, and the energy profile of the CA-QM bond. In addition we considered the bond length with the isolated MM atom (O or H) and the vertical excitation energy to S_1. Actually, the system contains two peptidic groups and therefore has two low-lying $n \rightarrow \pi^*$ states. However, all of the QM/MM calculations include only one peptidic group in the QM subsystem and consequently only one $n \rightarrow \pi^*$ state will be found. The connection atoms Ct, N and C occupy the α, β and γ positions with respect to the QM carbonyl group, respectively. In all of the cases, we had to make sure that the σ and σ^* orbitals of the CA were not involved in the peptidic $n \rightarrow \pi^*$ excited state. Actually, this requirement is most easily fulfilled when the CA is far from the chromofore. In this scenario, the worse case is represented by the Ct atom which is directly bound to the carbonyl.

Table 1. Results for the optimization of the semiempirical parameters for dimethyl sulphide: target abinitio and reparameterized QM/MM results. Weights for the data are also reported

	Ab initio	QM/MM	Weight
S–C(QM) distance (Å)	1.804	1.777	50
Mulliken charges			
S	−0.1126	−0.1166	10
C	−0.2559	−0.2692	6
H1	0.0998	0.0930	4
H2	0.0998	0.0930	4
H3	0.1127	0.0972	4
Energies (eV) at distorted S–C(QM) distances			
−0.1 Å	0.1286	0.1098	2
−0.05 Å	0.0296	0.0272	2
0.05 Å	0.0254	0.0246	2
0.1 Å	0.0940	0.0961	2

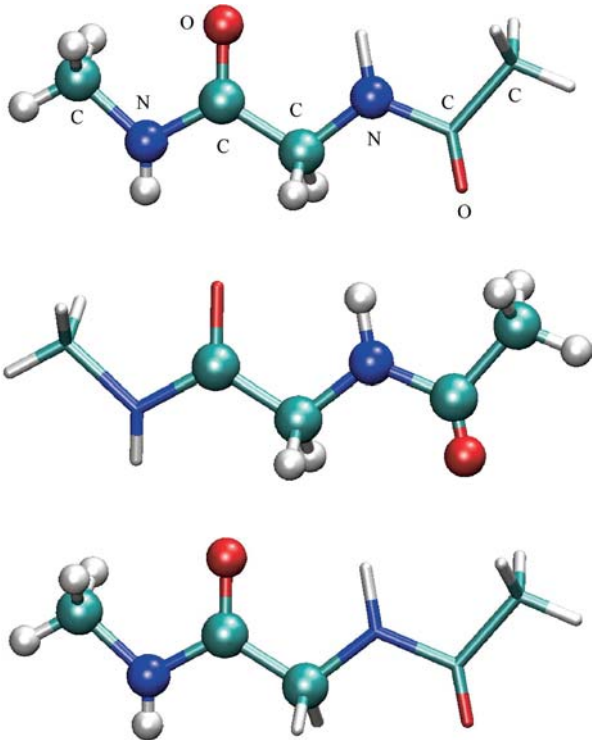

Fig. 2. The substituted glycine used to parameterize the peptidic N atom (upper panel) and C (middle panel) and the aliphatic C atom (Ct, lower panel). Balls and sticks represent QM and MM atoms respectively

The all-QM results confirmed that there are two almost degenerate $n \rightarrow \pi^*$ excited states characterized by localized excitation in the two peptidic groups. Consequently, the three optimizations have to provide approximately the same vertical transition energy for S_1.

Tables 2, 3 and 4 report the results for the N, C and Ct connection atoms, respectively.

As a general remark, we notice that the bond distances between the connection atoms and the closest QM and MM atoms are in very good agreement with the corresponding reference data. Despite the fact that the connection atom has only one electron and one basis function, and moreover its electrostatic MM embedding is very different with respect to a real environment constituted by nuclei and electrons, the electronic charge distribution in the QM subsystem is not substantially modified by the breaking of a covalent bond. The excitation energies are slightly underestimated: the largest error (– 0.12 eV) is obtained when Ct is the connection atom. Moreover, in this case, we find a non-negligible σ and σ^* contamination of the $n \rightarrow \pi^*$ wavefunction.

Table 5 reports the reoptimized semiempirical parameters.

Table 2. Results for the optimization of the semiempirical parameters for the amidic N in the substituted glycine (see Figure 2 upper panel): target AM1 and reparameterized QM/MM results. Weights for the data are also reported

	AM1	QM/MM	Weight
N–C(QM)distance (Å)	1.437	1.445	150
N–H(MM)distance (Å)	0.995	1.001	1
Mullik encharges			
N(Conn.At.)	−0.3903	−0.4464	1
sp^3C	−0.0214	−0.0232	1
H1	0.1010	0.1245	1
H2	0.1010	0.1245	1
Peptidic C	0.3661	0.3640	1
Peptidic O	−0.4428	−0.5012	1
Peptidic N	−0.3651	−0.3518	1
Peptidic H	0.2265	0.2362	1
Energies (eV) at distorted N–C(QM)distances			
−0.1 Å	0.2786	0.2628	4
−0.05 Å	0.0648	0.0642	3
0.05 Å	0.0571	0.0569	3
0.1 Å	0.2116	0.2160	8
Excited state (eV)	4.14	4.09	5

Table 3. Results for the optimization of the semiempirical parameters for the acylic C in the substituted glycine: target AM1 and reparameterized QM/MM results. Weights for the data are also reported

	AM1	QM/MM	Weight
C–C(QM) distance (Å)	1.532	1.534	150
C–O(MM) distance (Å)	1.246	1.236	2
Mulliken charges			
C (Conn.At.)	0.3661	0.6334	2.5
sp^3 C	−0.0214	−0.0205	2
H1	0.1010	0.1063	1.5
H2	0.1010	0.1063	1.5
Peptidic N	−0.3903	−0.3802	1.5
Peptidic H	0.2469	0.2612	1
Peptidic C	0.3674	0.3612	1
Peptidic O	−0.4482	−0.4499	0.5
Terminal sp^3 C	−0.2065	−0.2084	0.5
Energies (eV) at distorted C–C(QM)distances			
−0.1 Å	0.2490	0.2592	2
−0.05 Å	0.0570	0.0592	2
0.05 Å	0.0480	0.0482	2
0.1 Å	0.1730	0.1711	2
Excited state (eV)	4.14	4.08	5

Azobenzenophane

The 2,19-dithia[3.3](4,4′)-trans-diphenyldiazeno < 2 > phane consists of a ring which contains two azobenzene units linked by two -CH_2-S-CH_2- bridges. Our interest in this molecule comes from a broader study carried out in our laboratory on the photochemistry of azobenzene [38, 39, 40] and of supramolecular systems based on azobenzene functionality. This compound has been studied by Rau and Lüddecke to demonstrate that the mechanism of *trans* → *cis* photoisomerization of azobenzene in S_1 involves an inversion motion [41]. In fact, apparently the torsional motion along the C-N = N-C dihedral angle in this compound is hindered by the tension of the ring. Here we just focus on geometrical and electronic features to assess the reliability of the connection atoms, which are in this case two sulfur atoms.

Experimental data on the molecular geometry and electronic spectrum are available [41]. It is not easy to produce a reliable all-QM calculation to serve as reference for such a big system. On the other hand, we have already simulated the excited state dynamics of azobenzene using semiclassical [39] and quantum methods [40]. Therefore, we have already optimized the semiempirical parameters for the nitrogen and carbon atoms to get an accurate representation of the ground state and first three excited states of azobenzene [39]. We can directly compare the shifts in the transition energies from azobenzene to azobenzenophane using the corresponding experi-

Table 4. Results for the optimization of the semiempirical parameters for the tetrahedral C (Ct) in the substituted glycine: target AM1 and reparameterized QM/MM results. Weights for the data are also reported

	AM1	QM/MM	Weight
C–C(QM) distance(Å)	1.532	1.459	100
Mulliken charges			
C (Conn.At.)	−0.0214	−0.0214	2.5
Peptidic C	0.3661	0.4689	2
Peptidic O	−0.4428	−0.4702	1.5
Peptidic N	−0.3651	−0.3902	1
Peptidic H	0.2265	0.2316	1
Terminal sp^3 C	−0.0874	−0.1011	0.5
Energies (eV) at distorted C–C(QM)distances			
−0.1 Å	0.2490	0.1963	2
−0.05 Å	0.0570	0.0486	2
0.05 Å	0.0480	0.0430	2
0.1 Å	0.1730	0.1672	2
Excited state (eV)	4.14	4.02	15

Table 5. Optimized semiempirical parameters for sulfur and peptidic nitrogen (N) and carbon (C) and tetrahedral carbon (Ct). n indicates the principal quantum number of the atomic basis function used; for the other parameters see [43]

Parameter	S	N	C	Ct
U_{ss}	−11.4636158302	−11.5510913592	−25.9143353860	−13.4514522134
β_s	−13.6986805840	−12.9908116907	−23.5586541962	−14.2657034001
Z_s	1.1545008268	1.1242153967	1.0939735163	1.2234953769
α	1.3020718805	1.0764930858	1.3946964243	1.4695046726
G_{ss}	11.2252957230	16.3758167108	12.3359253484	11.8856958539
K_1	−0.2959768819	−0.5122426232	−0.2822886402	−0.3156048028
L_1	4.5892926321	10.6702903382	9.4980765320	5.8012199929
M_1	1.8983396405	1.3362713488	1.3796150939	1.5742505900
K_2	0.0418281873	0.0610837625	0.0629927643	0.0577247239
L_2	5.0154537672	2.6175313365	4.1885921105	5.4825078424
M_2	1.2828125314	−0.5105622887	2.1210480914	2.1468318567
K_3	−0.3198628377	−0.5835698069	−0.1439263952	−0.2334521642
L_3	5.1845917106	21.0534243879	6.8272068727	5.3825357844
M_3	1.8273677516	2.3642526844	2.3469666640	2.3951922703
K_4	0.0372125488	−0.0587377597	−0.0282225702	−0.1270557422
L_4	4.9292335733	4.9330597165	3.617491946	4.1139242811
M_4	2.8804361971	1.7776280934	3.7326939634	3.0841720178
n	3	2	2	2

mental data. The azobenzene was computed using a multi reference single CI starting from a complete active space with four electrons in four orbitals, adopting floating occupation MO's with width equal to 0.1 hartree. In the present QM/MM treatment of azobenzenophane, the QM subsystem includes one of the two azobenzene units, two standard AM1 CH_2 groups, and two sulfur atoms optimized as connection atoms, as described in the previous section. The MM subsystem consists of the other azobenzene unit plus two CH_2 groups. Notice that the QM treatment of both azobenzene units at a CI level comparable with the one adopted here, which is necessary to represent the S_1-S_3 excited states, would pose serious problems of efficiency and consistency. Moreover, the inclusion of the S lone pairs in the QM calculation would further complicate the definition of the MO active space. The force field used for the MM part of the ring was GAFF of AMBER 7 [36] with ESP charges obtained from an SCF/6–31G* calculation on a para-substituted azobenzene with two CH_3-S-CH_2- groups. The resulting core charge of sulfur was 0.9088. Figure 3 shows the molecule and how the atoms have been partitioned between QM (balls) and MM regions.

The optimized geometry at QM/MM level looks very similar to the one obtained at pure MM level. They are superimposed in Figure 4. This similarity could also serve as a check for the GAFF force field since the QM/MM optimized geometry compares favorably with experimental results [41]: the distances between the two azo groups are believed to be 3.86 Å and the QM/MM calculation yields 3.73 Å (an all-MM calculation with AMBER gives 3.65 Å). The two benzene rings have an average distance equal to 3.57 Å while the QM/MM computed quantity is 3.45 Å (AMBER: 3.40 Å). Consequently, the two azobenzene units are not perfectly planar, and the molecular geometry is vaguely elliptical. The S-C(QM) and S-C(MM) bond distances are 1.788 Å and 1.829 Å, respectively (AMBER: 1.827 Å): this is in line with the results obtained for the benchmark molecule, $(CH_3)_2S$.

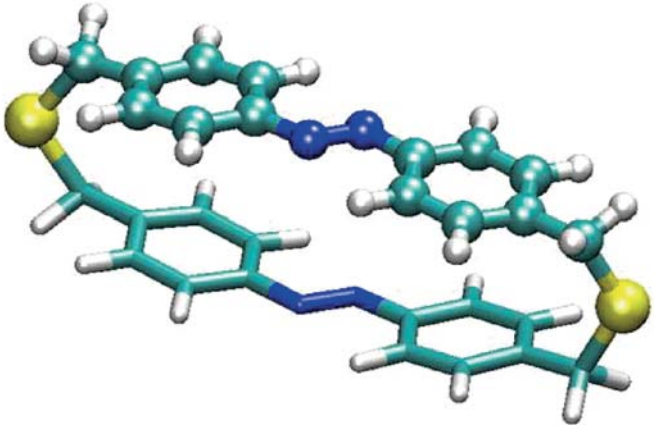

Fig. 3. Azobenzenophane used as a test case for the sulfur connection atom. Balls and sticks represent QM and MM atoms respectively

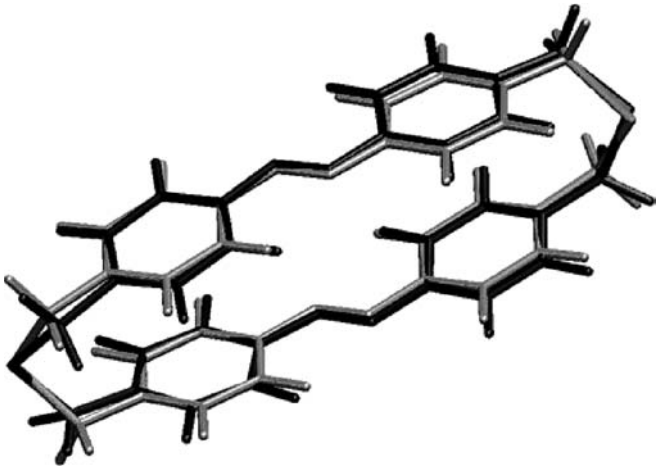

Fig. 4. Optimized geometries of the azobenzenophane: grey, all-MM calculation; black, QM/MM

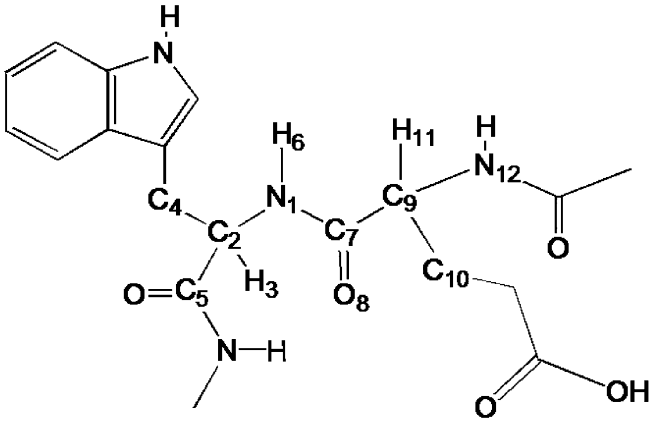

Fig. 5. Triptophane-glutammic acid dipeptide used as a test case for the peptidic N, C and Ct connection atoms

Moving from azobenzene to the azobenzenophane molecule, the first two excited states show a different behavior [41, 42]: the absorption maximum for S_1 is blue-shifted from 447 to 436 nm (from 2.77 to 2.84 eV) while S_2 shows a red-shift, from 316 to 340 nm (3.92 to 3.65 eV). The same trend is obtained with our calculations, although the shifts are somewhat overestimated: S_1 changes from 2.93 eV in the azobenzene molecule to 3.24 in the benzaphane compound (vertical excitation). For S_2 we obtain 4.27 and 3.58 eV for azobenzene and azobenzenophane respectively.

TRP-GLH dipeptide

We applied the peptidic C and N and tetrahedral Ct reparameterized connection atoms to a dipeptide containing glutammic acid and triptophane, the structure of which is drawn in Figure 5. This compound was chosen to test possible intramolecular hydrogen bonds between a QM and an MM part of the system. Triptophane has low-lying excited states and we wanted to verify that the QM/MM treatment is able to preserve the electronic structure of the ground and of the excited states. Three tests were carried out on this compound: the QM subsystem was defined in turn as one or the other of the two aminoacids, depending on the chosen connection atom. Figure 6 shows how the atoms are partitioned in the three calculations. The atoms represented by balls belong to the QM subsystem. One can see that the connection atoms have been chosen to be the peptidic C-7 and N-1 atoms and tetrahedral C-9. The corresponding QM parts of the compound contained glutammic acid for C and triptophane for N and Ct. Of course, in any application of the method, the appropriate QM/MM partition depends on the focus on the reactive and/or excited state processes of interest. In the following we shall focus on the properties of the QM-MM frontier regions. In fact, the further we go from the connection atom inside the MM or the QM regions, the easier it is to reproduce the expected QM or MM behavior. Before going into details, we stress that the results depend on the particular choice of the force field and the QM Hamiltonian chosen, in the sense that even pure QM and pure MM calculations give slightly different results.

We optimized a local minimum of the ground state but we did not need to find the absolute minimum, because there is no rule to ensure that all-QM and all-MM calculations (let alone the three QM/MM ones) would provide the absolute minimum at the same conformation. We just checked that the five local minima we found (QM, MM and three QM/MM) belonged to the same conformation. The three QM/MM calculations were performed using a CI wave function containing single and double excitations (CISD) in a window of orbitals: when the triptophane was included in the QM part, the window contained ten electrons in eight orbitals; for the glutamic acid we adopted an active space of eight electrons in seven orbitals. Consistently with these choices, the QM reference calculation was carried out with an 18 electron/15 orbital CISD. In all calculations the Hamiltonian was a standard AM1 and the Gaussian width for the floating occupation of molecular orbitals was 0.2 hartree. The MM force field was again AMBER. The CA core charges were 0.7281, 1.5670 and 1.0439 for N, C and Ct respectively.

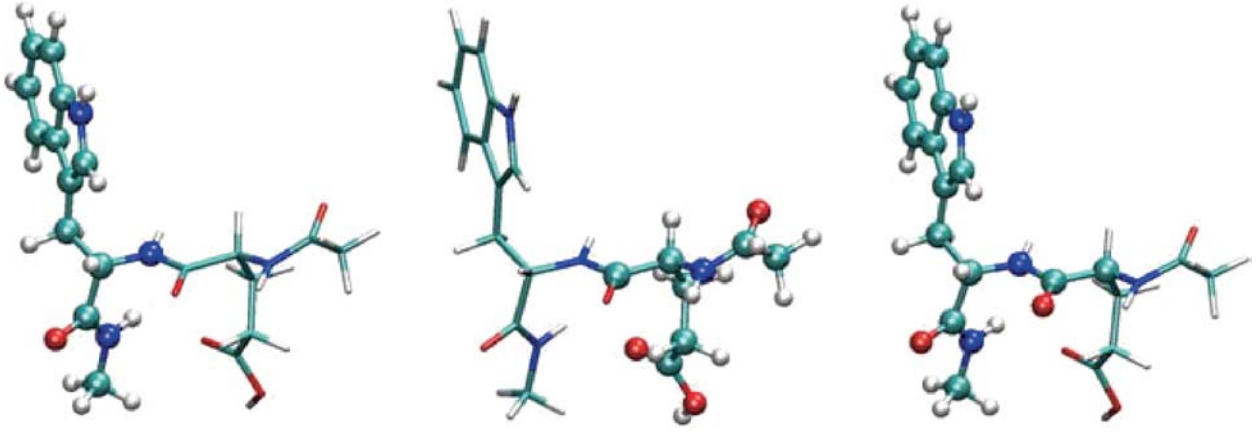

Fig. 6. Partition of the triptophane-glutammic acid dipeptide in QM (balls) and MM (sticks) subsystems, used to test the peptidic N (left panel), C (middle panel) and Ct (right panel) connection atoms

Table 6. Comparison between geometric and electronic parameters obtained at different computational levels: purely semiempirical AM1; QM/MM calculations with C, N and Ct as connection atoms (see Fig. 5 for the labels and Fig. 6 for the definition of the QM and MM subsystems); the entire system computed using the AMBER force field. Mulliken charges are reported for atoms in the QM region (bold) and MM atomic charges for the other ones. Bond distances (Å) are also reported (see Figure 5 for labeling). Note that the distances between two QM atoms are in bold and those involving the connection atom are in italics. The last five rows contain vertical transition energies: $\pi \rightarrow \pi^*$ states are located on the indole group of triptophane while $n \rightarrow \pi^*$ ones are localized on the peptidic groups

	AM1	N	C	Ct	AMBER
Charge C5	**0.285**	**0.362**	0.597	**0.315**	0.597
Charge C2	**0.015**	**0.034**	0.027	**0.012**	0.027
Charge C4	−0.096	−0.098	−0.005	−0.109	−0.005
Charge H3	**0.168**	**0.172**	0.078	**0.124**	0.078
Charge N1	−0.361	−0.434	−0.416	−0.404	−0.416
Charge H6	**0.220**	0.272	0.272	**0.231**	0.272
Charge C7	**0.331**	0.597	**0.386**	**0.320**	0.597
Charge O8	−0.376	−0.568	−0.568	−0.377	−0.568
Charge C9	**0.008**	0.014	**0.148**	**0.113**	0.014
Charge H11	**0.141**	0.078	**0.122**	0.078	0.078
Charge C10	−0.159	−0.007	−0.179	−0.007	−0.007
Charge N12	−0.382	−0.416	−0.397	−0.416	−0.416
Dist. C5–C2	**1.553**	**1.522**	1.544	**1.555**	1.542
Dist. C4–C2	**1.539**	**1.524**	1.537	**1.535**	1.537
Dist. C2–N1	**1.441**	***1.421***	1.469	**1.441**	1.471
Dist. N1–H6	**0.989**	*0.991*	1.018	**0.993**	1.001
Dist. N1–C7	**1.375**	*1.326*	*1.356*	**1.388**	1.335
Dist. C7–O8	**1.242**	1.232	*1.245*	**1.252**	1.228
Dist. C7–C9	**1.550**	1.540	***1.535***	***1.497***	1.541
Dist. H11–C9	**1.135**	1.091	**1.145**	*1.095*	1.091
Dist. C9–C10	**1.538**	1.543	**1.551**	*1.545*	1.542
Dist. N12–C9	**1.442**	1.467	**1.464**	*1.454*	1.467
$E(S_1)$	3.96 $\pi \rightarrow \pi^*$	3.88 $\pi \rightarrow \pi^*$	4.23 $n \rightarrow \pi^*$	3.87 $\pi \rightarrow \pi^*$	–
$E(S_2)$	4.22 $\pi \rightarrow \pi^*$	4.17 $\pi \rightarrow \pi^*$	–	4.17 $\pi \rightarrow \pi^*$	–
$E(S_3)$	4.66 $n \rightarrow \pi^*$	4.74 $n \rightarrow \pi^*$	–	5.34 $n \rightarrow \pi^*$	–
$E(S_4)$	4.81 $n \rightarrow \pi^*$	–	–	–	–
$E(S_5)$	4.83 $n \rightarrow \pi^*$	–	–	–	–

The results are summarized in Table 6 (see Figure 5 for the labels). Here we compare Mulliken charges (only for the QM atoms), bond distances, and the vertical excitation energies for the QM, the three QM/MM calculations and also the corresponding one performed with the force field only. The Mulliken charges determined by QM/MM methods are very similar to the corresponding ones obtained at all-QM level. This tells us that the QM-MM cut does not significantly alter the electronic distribution in the QM region of the system. In addition, the five lowest singlet excited states of the dipeptide are also obtained in one or another of the three QM/MM calculations: the two lowest states (3.96 and 4.22 eV) are $\pi \rightarrow \pi^*$ excitations of the indole group while the three next at 4.66–4.83 eV are $n \rightarrow \pi^*$ excitations of the peptidic groups. The same order is obtained when the triptophane only is computed at QM level (N and Ct connection atoms), and the energies of the two lowest states ($\pi \rightarrow \pi^*$) agree with the all-QM calculation. When the CA is N, we also get one $n \rightarrow \pi^*$ state, approximately at the right energy. When it is Ct the $n \rightarrow \pi^*$ energy is largely overestimated, showing that the contamination with σ and σ^* could spoil the transferability of the connection atom parameters, as far as the excited states are concerned. The calculation of the QM glutammic acid (C as CA) provides one peptidic $n \rightarrow \pi^*$ state, and its energy is underestimated by ~0.4 eV. Apparently the presence of the indole group influences the position of the $n \rightarrow \pi^*$ states. This is confirmed by an all-QM calculation for the same dipeptide without the indole group (alanine-glutammic acid dipeptide): here the lowest state is a peptidic $n \rightarrow \pi^*$ at 4.29 eV (the isolated indole has two $\pi \rightarrow \pi^*$ states at 3.90 and 4.23 eV).

The bond lengths obtained in the QM/MM treatments are very close to all-QM or all-MM results, respectively, when both atoms belong to the QM or to the MM region of the system. The bond lengths involving the CA (in italic in Table 6) should be compared with all-QM or all-MM results when the partner is a QM or an MM atom, respectively. The largest error (– 0.05 Å) occurs in the C-7–C-9 bond length when C-9 is the Ct connection atom. The second largest (– 0.03 Å) occurs between two QM atoms (C-5–C-2), with N as CA. All other errors do not exceed 0.02 Å.

Conclusions

We have conjugated our semiempirical QM/MM multistate method [2, 10, 18] with Antes and Thiel's connection atom approach [33]. As a result, we have described a viable QM/MM strategy, which can be applied when there are covalent bonds linking the QM and MM subsystems, and is able to treat excited states and bond breaking or formation in the QM portion.

The transferability of the connection atom parameters allows us to optimize them for a model compound and to use them in much larger molecules, for which structural and energetic data are not available and/or the optimization would be a burden. We have successfully applied this procedure for two examples, a large cyclic molecule composed of two azobenzene units joined by -CH_2-S-CH_2- bridges, and a triptophane-glu-

tammic acid dipeptide. In this way we have tested the method for four different connection atoms: S in the thioether group, and the acylic C, amidic N and aliphatic C in a peptide.

The results show that the most important goal (to reproduce the electronic structure of the QM subsystem, as it can be computed in an all-QM treatment) is achieved. The treatment of the excited states of a peptide is particularly exacting. In fact, each aminoacid has at least one chromophore and one should not place the connection atom too close to a chromophore belonging to the QM subsystem: this is shown by the example of the aliphatic carbon positioned in α with respect to a peptidic group. Anyway, this choice is quite acceptable for studying ground state phenomena.

The calculation of energies and energy gradients by the QM/MM procedure illustrated in this paper is very fast, even with a large MM environment. On the other hand, a specific parameterization of the semiempirical CI method allows one to obtain accurate ground and excited state PES for practically any molecule. These favorable features are currently being exploited in simulations of photochemical events in supramolecular systems of various kinds, including biological chromophores embedded in proteins.

Appendix A

Here we give the expressions we have used in the calculation of the monoelectronic integrals:

$$h_{\mu v} = \langle \mu | \frac{q_m}{R_{im}} | v \rangle \tag{7}$$

where μ and v are atomic orbitals centered on the same atom α, and R_{im} is the distance between the electron i and the point charge q_m.

Following the standard MNDO rules, the above integral is approximated by a bielectronic integral:

$$h_{\mu v} = q_m(\mu v | s_m s_m) \tag{8}$$

where s_m is an s-type atomic function centered on the MM atom m. The bielectronic integral (Eq. 8) is in turn evaluated using the standard MNDO rules; the only other approximation we made was to neglect the penetration effect of the s_m orbital into the charge distribution μv.

Considering only s and p type atomic functions we have

$$\langle s | \frac{q_m}{R_{im}} | s \rangle = \frac{q_m}{R} \tag{9}$$

$$\langle s | \frac{q_m}{R_{im}} | p_\sigma \rangle = \frac{q_m}{2} \left(\frac{1}{|R - \delta_d|} - \frac{1}{|R + \delta_d|} \right) \frac{\sigma}{R} \qquad (\sigma = x,y,z) \tag{10}$$

$$\langle p_\sigma | \frac{q_m}{R_{im}} | p_{\sigma'} \rangle = \frac{\beta_1 - \beta_2}{R^2} \sigma\sigma' + \beta_2 \delta_{\sigma,\sigma'} \qquad (\sigma, \sigma' = x,y,z) \tag{11}$$

where $R = \sqrt{x^2 + y^2 + z^2}$ is the distance between the atoms α and m and

$$\beta_1 = q_m \left(\frac{1}{|R - 2\delta_q|} + \frac{1}{|R + 2\delta_q|} - \frac{1}{R} \right) \tag{12}$$

$$\beta_2 = q_m \left(\frac{2}{\sqrt{R^2 + 4\delta_q^2}} - \frac{1}{R} \right) \tag{13}$$

where δ_d and δ_q are semiempirical derived parameters characterizing the *sp* and *pp* charge distributions, given in terms of the Slater exponent and the principal quantum numbers of the orbitals involved [43].

Acknowledgements. This paper is dedicated to our friend and teacher Jacopo Tomasi, a pioneer of the elaboration of electrostatics and QM/MM concepts and methods. Our research work owes a lot to Jacopo: inspiration, warnings, criticism and, sometimes, his most welcome appreciation.

References

1. Ben-Nun M, Quenneville J, Martinez TJ (2000) J Phys Chem A 104:5161
2. Granucci G, Persico M, Toniolo A (2001) J Chem Phys 114:10608
3. Stewart JJP (1990) In: Lipkowitz EKB, Boyd DB (eds) Reviews in computational chemistry. VCH, New York, pp 45–81
4. Zerner MC (1991) In: Lipkowitz EKB, Boyd DB (eds) Reviews in computational chemistry. VCH, New York, pp 313–65
5. Thiel W (1996) Adv Chem Phys 93:703
6. Dewar MJS, Thiel W (1977) J Am Chem Soc 99:4899
7. Dewar MJS, Thiel W (1977) J Am Chem Soc 99:4907
8. Stewart JJP (1989) J Comp Chem 10:209
9. Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP (1985) J Am Chem Soc 107:3902
10. Granucci G, Toniolo A (2000) Chem Phys Lett 325:79
11. Toniolo A, Granucci G, Inglese S, Persico M (2001) Phys Chem Chem Phys 3:4266
12. Toniolo A, Ben-Nun M, Martinez TJ (submitted for publication)
13. Bonaccorsi R, Petrongolo C, Scrocco E, Tomasi J (1971) Theor Chim Acta 20:331
14. Bonaccorsi R, Pullman A, Scrocco E, Tomasi J (1972) Theor Chim Acta 24:51
15. Warshel A, Levitt M (1976) J Mol Biol. 103:227
16. Gao J (1995) In: Lipkowitz EKB, Boyd DB (eds) Reviews in computational chemistry. VCH, New York, pp 119–85
17. Monard G, Merz KM (1999) Acc Chem Res 32:904
18. Persico M, Granucci G, Inglese S, Laino T, Toniolo A (2003) J Mol Struct (THEOCHEM) 621:119
19. Singh UC, Kollman PA (1986) J Comput Chem 7:718
20. Field M, Bash PA, Karplus M (1990) J Comput Chem 11:700
21. Hypercube, HYPERCHEM, Waterloo, Canada
22. Hall RJ, Hindle SA, Burton NA, Hillier IH (2000) J Comp Chem 21:1433
23. Eurenius KP, Chatfield DC, Brooks BR, Hodoscek M (1996) Int J Quantum Chem 60:1189

24. Harrison MJ, Burton NA, Hillier IH (1997) J Am Chem Soc. 119:12285
25. Antes I, Thiel W (1998) In: Gao J, Thompson MA (eds) Combined quantum mechanical and molecular mechanical methods (ACS Symposium). ACS, Washington DC, p 50
26. Das D, Eurenius KP, Billings EM, Sherwood P, Chatfield DC, Hodoscek M, Brooks BR (2002) J Chem Phys 117:10534
27. Phillip DM, Friesner RA (1999) J Comp Chem 20:1468
28. Thery V, Rinaldi D, Rivail L-J, Maigret B, Ferenczy GG (1994) J Comput Chem 15:269
29. Monard G, Loos M, Thery V, Baka K, Rivail L-J (1996) Int J Quantum Chem 58:153
30. Assfeld X, Rivail L-J (1996) Chem Phys Lett. 263:100
31. Reuter N, Dejaegere A, Maigret B, Karplus M (2000) J Phys Chem A 104:1720
32. Gao J, Amara P, Alhambra C, Field MJ (1998) J Phys Chem A 102:4714
33. Antes I, Thiel W (1999) J Phys Chem A 103:9290
34. Stewart JJP, MOPAC2000. Fujitsu Limited, Tokyo, Japan, 1999
35. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA (1995) J Am Chem Soc 117:5179
36. Case DA, Pearlman DA, Caldwell JW, Cheatham III TE, Wang J, Ross WS, Simmerling CL, Darden TA, Merz KM, Stanton RV, Cheng AL, Vincent JJ, Crowley M, Tsui V, Gohlke H, Radmer RJ, Duan Y, Pitera J, Massova I, Seibel GL, Singh UC, Weiner PK, Kollman PA (2002) AMBER 7. University of California, San Francisco, CA
37. Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) Numerical recipes. Cambridge University Press, Cambridge, UK
38. Cattaneo P, Persico M (1999) Phys Chem Chem Phys 1:4739
39. Ciminelli C, Granucci G, Persico M: Chem Eur J, accepted
40. Toniolo A, Martinez TJ, Persico M: in preparation
41. Rau H, Lüddecke E (1982) J Am Chem Soc 104:1616
42. Rau H (1990) In: Durr H, Bouas-Laurent H (eds) Photochromism: molecules and systems. Elsevier, Amsterdam, p 165
43. Stewart JJP (1999) MOPAC2000 User Manual. Fujitsu Limited, Tokyo