

Evaluating the enthalpic contribution to ligand binding using QM calculations: effect of methodology on geometries and interaction energies†

Duangkamol Gleeson,^a Ben Tehan,^{*b} M. Paul Gleeson^{*c} and Jumras Limtrakul^d

Received 2nd April 2012, Accepted 27th June 2012

DOI: 10.1039/c2ob25657f

As a result of research on ligand efficiency in the pharmaceutical industry, there is greater focus on optimizing the strength of polar interactions within receptors, so that the contribution of overall size and lipophilicity to binding can be decreased. A number of quantum mechanical (QM) methods involving simple probes are available to assess the H-bonding potential of different heterocycles or functional groups. However, in most receptors, multiple features are present, and these have distinct directionality, meaning very minimalist models may not be so ideal to describe the interactions. We describe how the use of gas phase QM models of kinase protein–ligand complex, which can more closely mimic the polar features of the active site region, can prove useful in assessing alterations to a core template, or different substituents. We investigate some practical issues surrounding the use of QM cluster models in structure based design (SBD). These include the choice of the method; semi-empirical, density functional theory or *ab-initio*, the choice of the basis set, whether to include implicit or explicit solvation, whether BSSE should be included, *etc.* We find a combination of the M06-2X method and the 6-31G* basis set is sufficiently rapid, and accurate, for the computation of structural and energetic parameters for this system.

1. Introduction

A variety of studies have helped to highlight the important contribution that individual interactions can have on the overall protein binding energy of a ligand. These include detailed studies on the characteristic interactions made by a variety of different functional groups^{1,2} with amino acid residues, the characteristics of π – π stacking,^{3–6} cation– π ⁷ and anion– π ⁸ interactions, halogen bond interactions,^{9,10} as well as the unique conformational preferences of different functional groups.^{11,12} Indeed, recent analyses of isothermal calorimetry data by Keseru *et al.*,^{13–15} who advocate the assessment of both the enthalpic and entropic contributions to the binding affinity, have noted how the focus on entropic gains in potency are not as productive

(*i.e.*, increasing lipophilicity and driving potency through the hydrophobic effect). Optimization efforts that focus on improving the enthalpic contribution to protein binding, by directly improving the polar interactions between the ligand and receptor, are preferable. In fact, the authors note that the undesirable focus on entropic potency gains is one of the key reasons for the increase in lipophilicity and molecular weight of drugs and drug candidates over time.¹⁵ In addition, it helps to explain the observation that historical drugs generally have lower potencies, lipophilicity and molecular weight compared to compounds in current, or recent development.^{16,17}

In light of the recent focus on ligand efficient molecules,^{18–23} there now appears to be a greater emphasis on improving the efficiency of the lead template/series, rather than achieving potency gains due to addition of lipophilicity. The latter is typically achieved by filling lipophilic pockets, displacing labile water, or incorporating extensive non-polar linkers to target more distant polar interactions, often resulting in questionable overall gain. This is because the resultant increase in overall lipophilicity and/or molecular weight to increase potency can have a significant detrimental effect on a wide variety of Adsorption Distribution Metabolism Excretion and Toxicity parameters (ADMET).^{24–26}

Understanding the interactions between functionality on a ligand with that in a protein active site is critical to improving potency in an efficient manner. A more ideal approach is to optimize the available enthalpic interactions present in a template, with the use of additional approaches to increase potency

^aDepartment of Chemistry, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand

^bHeptares Therapeutics, BioPark, Broadwater Road, Welwyn Garden City, Herts, AL7 3AX, United Kingdom.

E-mail: ben.tehan@heptares.com; Tel: +44-1707-358646

^cDepartment of Chemistry, Faculty of Science, Kasetsart University, 50 Phaholyothin Rd, Chatuchak, Bangkok 10900, Thailand.

E-mail: paul.gleeson@ku.ac.th; Fax: +66-2-5793955; Tel: +66-2-562-5555 extn 2210

^dDepartment of Chemistry, Faculty of Science, and Center for Advanced Studies in Nanotechnology and its Applications in Chemical, Food and Agricultural Industries, Kasetsart University, Bangkok 10900, Thailand

†Electronic supplementary information (ESI) available: All structural parameters and energies and energy correlation matrix plots and tables. See DOI: 10.1039/c2ob25657f

afterwards as needed.^{13–15} This is not a trivial task, but could be achieved by leveraging calorimetry binding data and structure based design (SBD) techniques. The latter technique is extensively used in drug discovery programs with structural data of the target, to rationally design increases in potency or selectivity into a lead series. The use of experimental structures derived from X-ray or NMR, can be used in isolation or in conjunction with computational chemistry.²⁷ The latter method presents program teams with a means to rationally design and test new molecules that can better leverage the interactions and steric features present in the protein.

Theoretical models of protein–ligand complexes can be generated in a number of different ways. Rapid, molecular mechanical (MM) methods can be used to sample whole protein models quickly (or over long timescales),²⁸ linear scaling semi-empirical methods can be used to simulate the whole protein system quantum mechanically (QM),^{29–32} or QM/MM methods^{33–40} can be used to simulate the active site using QM and the remainder using MM. Alternatively, smaller, approximate models can be used at higher levels of QM theory to evaluate particular regions of interest more rapidly.^{41–44}

Each of the methods discussed above offers distinct advantages in particular circumstances. For example, MM methods are very quick to evaluate, meaning extensive sampling is possible. However, non-standard templates, metals or certain interactions are not ideally described.^{9,32,45} Semi-empirical QM methods are relatively rapid, allowing large clusters to be considered or whole proteins in linear scaling form, but are not considered the most accurate as a result of the approximate method used.^{46,47} QM/MM allows the use of accurate QM methods to treat the important core regions, and take into account longer range effects using MM, however interactions across the boundary region can lead to issues.^{40,48} QM clusters allow the use of very accurate levels of theory to study the key interactions between a protein and ligand, however the effect of the surrounding protein is therefore completely neglected. QM cluster calculations are nevertheless employed for many tasks including the prediction of interaction strengths between model ligands and probes,^{42–44,49–51} to more complex tasks such as reaction mechanism elucidation^{52,53} and X-ray structure refinement.^{47,54}

In previous reports the authors have investigated the use of QM/MM methods to study protein kinase-inhibitor complexes, showing the distinct benefits of this method over traditional docking in ligand pose scoring.⁵⁵ A follow up to this study highlighted the potential use of this method in aiding refinement of the active site region where non-standard ligands are present.⁵⁶ Subsequent investigations were carried out on smaller, but more rapidly computable QM cluster models, consisting of the ligand and active site residues that make the key interactions.⁵⁷ While this approach neglects the effect of the protein and solvent, it allows a researcher to assess how optimal the interactions between the moieties present are, and whether they can be improved.

As illustrated in Fig. 1, if the ligand conformation or interactions in the optimized active site model differ significantly from the experimental protein–ligand structure: this suggests that either the conformation/interactions present are not optimal to make the best possible interactions, due to unfavourable sterics for example. Alternatively, the structure might change dramatically because the initial ligand parameters used in the refinement

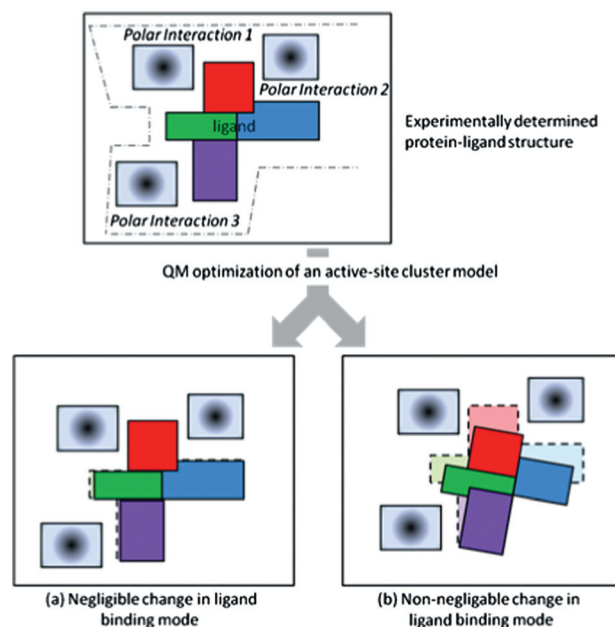


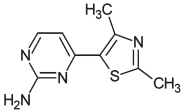
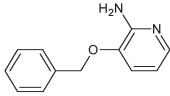
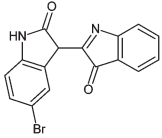
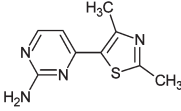

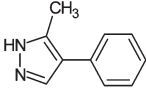
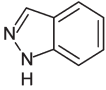
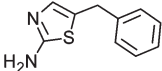
Fig. 1 An illustration of how QM active site models could be employed to aid in the optimization of the enthalpic contribution to overall binding energy. In case (a), a negligible change in the structure occurs on optimization suggesting the interactions present are optimal, since when the protein is removed they do not change dramatically. In case (b), the ligand conformation changes dramatically on optimization suggesting (1) the polar ligand interactions are not optimal and should be improved by alteration of the template substituents or (2) the refined ligand coordinates obtained from the experimental structure were sub-optimal.

step were sub-optimal. In contrast, a negligible change in structure without the surrounding protein present suggests that the enthalpic interactions between the ligand and the key residues are optimal. Thus, understanding the strength of interactions, and the preferred conformations adopted by a molecule in a receptor are important pre-requisites to allow the rational, efficient optimization of a lead series to be performed. A number of methods are available to predict the strength of interaction of individual functional groups which can prove extremely insightful in the design and modification of lead series.^{41,43,44,49–51}

In this work we consider the use of small QM models of receptors, consisting of the key polar active site interactions, rather than generic probes. The region selected is not as extensive as used in the approach of Gueto-Tettay, who used residues within 5 Å of the active site, which, due to the significant size, necessitates the use of the semi-empirical PM6 method.⁴¹ Here we investigate smaller, yet more interaction relevant active site models. We are not, *per se*, interested in predicting the much more challenging absolute binding free energy,²⁸ rather, the goal is to determine whether such methods could be used to assess the relative interaction strength of inhibitors with key polar elements of a receptor, with the view to using them to rapidly assess alternative modifications of lead series to improve the contribution to enthalpic binding.

For these initial studies, we have employed the cluster based approach using a variety of conditions to understand the impact

Table 1 Kinase-inhibitor structure used in this study. Reported are the PDB ID, inhibitor structure, resolution, kinase target, target pIC₅₀ and a description of the H-bonds mediated with the hinge. Outer (O), central (C) and inner (I) HBs correspond to those defined in Fig. 1. CH refers to a short interaction distance between a carbonyl group of the hinge and a CH hydrogen atom of the inhibitor

PDB ID	Inhibitor	Resolution	Target	Activity ^a	H bond pattern
1PXJ ⁶⁸		2.3	CDK2	IC ₅₀ = 6.5 uM ^{68,69}	O(CH), C, I
1W7H ⁷⁰		2.2	P38	IC ₅₀ = 1300 uM ⁷¹	O(CH), C, I
2BHE ⁷²		1.9	CDK2	IC ₅₀ = 2 uM ⁷²	O, C, I
2C5O ⁶⁹		2.1	CDK2	K _i = 6.5 uM ^{68,69}	O, C, I(CH)
2UVX ⁷³		2.0	PKA-B	IC ₅₀ > 100 uM	O(CH), C, I
2UW3 ⁷⁴		2.2	PKA-B	IC ₅₀ = 80 uM	C, I
2VTA ⁷⁵		2.0	CDK2	IC ₅₀ = 185 uM ⁷⁵	O, C, I(CH)
3DND ⁷⁶		2.3	CDK2	IC ₅₀ = 16 uM ⁷⁶	O(CH), C, I

^a SD in activity <1 log unit which means the binding energies of these molecules differ no more than 1.4 kcal mol⁻¹ on average.

of the choice of methods in such assessments. We have attempted to quantify the effect of using different methodologies on a set of cluster models generated from a set of 8 PDB structures we have previously reported on (Table 1). We consider a number of different factors in this study, including; (a) the choice of model system (*i.e.*, a QM active site model containing the key residues), (b) the choice of QM method (*i.e.*, semi-empirical, density functional theory or *ab-initio*), (c) the size of the basis set, (d) should solvation be included, (e) should basis set superposition error be considered when assessing binding energies.

2. Computational procedures

Crystal structures of the 8 protein-kinases listed were downloaded from the RCSB protein databank (www.rcsb.org) (Table 1). These structures were chosen such that the ligands only made polar interactions with the three amino acid residues that constitute the “hinge” region (*i.e.* no water mediated interactions were present). For a detailed description of the structural features of the protein kinase target class, see ref. 77.

The 8 truncated protein models consisted of the backbones of the 3 hinge amino acids involved in binding the adenine portion of ATP. The amino acid sidechains were replaced by hydrogen atoms. The QM representation used in this study is exemplified in Fig. 2 and has been employed by both us and others to elucidate aspects of non-bonded interactions in kinase-inhibitor complexes.⁵⁷ The C_α atoms of the truncated amino acids were frozen during geometry optimization.

Geometry optimization of QM models was performed using Gaussian 03⁵⁸ at the following levels of theory: MP2/6-31+G**, M06-2X/6-31G*, HF/6-31G*, HF/3-21G, AM1. These span the time-consuming, to very rapid methods. M06-2X is an increasingly popular, newer, DFT method that has performed better in recent benchmarking studies than the more common B3LYP method.^{59–61} For the purpose of comparison, a purely MM based approach was also investigated, consisting of the CHARMM force field as implemented in Discovery Studio 2.5 with empirically derived Momany-Rone atomic charges.⁶² The effect of including an implicit solvent model of water was also investigated for the M06-2X/6-31G* and HF/3-21G models using a polarizable continuum model (PCM).

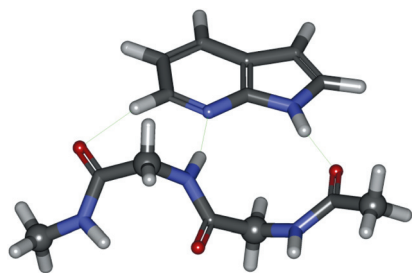


Fig. 2 An illustration of the QM model used in this study. The polar interactions of the proteins are denoted using the backbone atoms of 3 amino acids that constitute the hinge region. The amino acids sidechains were removed and replaced by hydrogen atoms. The AA chain was terminated one SP3 carbon atom after the nearest amide heteroatom. All atoms in the calculation were flexible except for the C_α atoms (denoted with a ball representation).

Computed interaction energies were obtained by subtracting the energy of the optimized, isolated ligand, and protein acid model, from the energy of the complex. In addition, the effect of correcting the energies for basis set superposition error (BSSE)⁶³ was considered (counterpoise correction) for the M06-2X/6-31G* and HF/3-21G models.

3. Results and discussion

We have evaluated a number of different methodologies that can be used to generate small QM models of the polar active site interactions found within typical protein ligand complexes. This was done in order to understand how ideal the polar interactions are in the absence of the extended protein environment, and discuss their suitability in terms of their RMSDs, in addition to analyzing H-bond distances and how they compare to their corresponding X-ray structures. It should be noted that minor changes in the distances and angles of a particular interaction can result in subtle differences in the positioning of a ligand within an active site pocket, and which in turn may significantly affect the choice of substituents, or where additional growth is considered.

A benefit of using such easy to construct, albeit approximate models, is that we can rapidly evaluate how ideal the interactions are between the polar active site features and the ligand. As discussed above, deviations in the interactions or binding conformation on removing the extended protein might suggest that the ligand binding mode is either sub-optimal due to high conformational strain or sub-optimal polar interactions due to steric constraints imposed by the extended protein, or potentially due to sub-optimal refinement.^{64,65}

As discussed in our previous study,⁵⁷ where we earlier reported the results at the MP2/6-31+G(d,p) level for this dataset alone, we found both of the scenarios above had occurred. Briefly, models for 3DND, and 1W7H, in particular, showed dramatic differences between the QM optimized models and X-ray results. Rotation of the non-polar benzyl and benzyloxy groups in the two complexes, respectively, led to lower energy, preferred conformations (*i.e.*, significant strain energy present). This also led to a dramatic change in the polar interactions in the case of the former, however this is also likely to be

affected by another factor. In 3DND, the ligand lies in a position rather distant from the hinge, but on QM optimization, the H-bonds (and the atypical C–H...O=C interaction that is frequently seen in kinases) decrease dramatically.

Other effects were independent of the protein and more likely due to issues regarding the ligand fitting to relatively poor density. In 2C5O for example, the pyrimidin-2-amine and thiazole group are planar with respect to each other. However, on optimization the groups adopt a more plausible angle of ~37°. Indeed, the same ligand was also found in the structure 1PXJ where it displayed an angle of 39°, apparently confirming that the refinement process led to the former result.

Furthermore, unusually short and long H-bond distances were observed in 2BHE, 3DND, 1PXJ and 2C5O. In particular 2BHE displays a very short H-bond to the central H-bond acceptor that on optimization increases to approximately 1.8 Å. In addition, 2C5O displays a very short C–H...O=C bond (~2.0 Å), which increases to the more realistic value (~2.4 Å) from an analysis of known kinase X-ray structures sourced from the PDB databank. These results also suggest that less attention was spent assessing the chemical accuracy of the interactions in question, compared to the empirical fitting to the density, which in these cases is not ideal.^{64–66}

It should be noted that the resolutions of the X-ray structures used here are typical of those used in SBD studies (~1.9–2.3 Å). However, those studied are not necessarily at an ideal standard to compare theoretical results to. This is because the structures are (a) not completely representative of a protein–ligand complex in solution, at 37° and (2) that the atomic coordinates that have been derived are not error free.^{36,64–66} Indeed, these structures typically lack any information regarding hydrogen atom positions, and sometimes contain poorly positioned ligands, especially in cases where inhibitors are non-standard,⁶⁶ or only weakly potent.⁶⁴ We therefore also make reference to higher resolution experimental structural data taken from comparable interactions⁵⁷ found in the Cambridge Structural Database (CSD) (www.ccdc.cam.ac.uk/products/csd). This contrasts to comparative studies by others who have used the original electron density as a reference.⁶⁵

3.1 Effect of methodology of QM active site structures

For molecular systems of the size employed here, the MP2/6-31+G(d,p) calculations are resource intensive, requiring days per complex to optimize on Intel Core i7 workstations. Thus, even these calculations might be prohibitive when used in SBD exercises, or in support refinement studies. In typical SBD applications, multiple template modifications or alternate substituents may require evaluation, and the use of less computationally expensive methods may therefore be warranted. For example, if 10 alternatives to the heterocyclic template were considered, and another 10 modifications in terms of the points substitution, or substituent types, 100 different calculations would be required. A solution would be to employ DFT, semi-empirical method, or MM based methods in such studies. We have therefore investigated the use of a number of different methodologies for use in probing active site models, ranging from the very slow (MP2/6-31+G(d,p)), to the very fast (AM1 and CHARMM calculations).

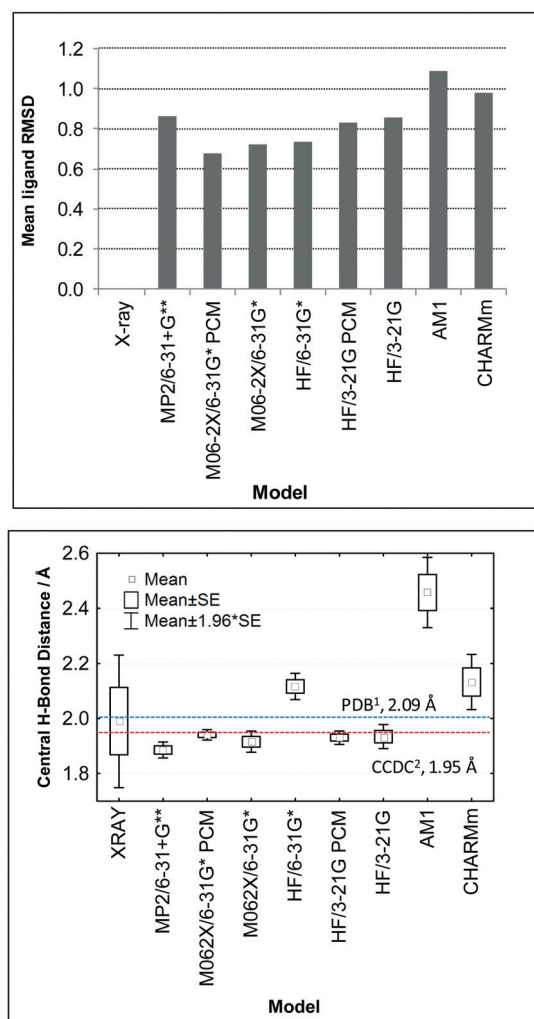


Fig. 3 (a) Plot of the mean RMSD of the optimized gas phase models to the original X-ray structures and (b) a box plot summary of the central hinge H-bond interactions. ¹Taken from analysis of PDB kinase complexes. ²Taken from analysis of CSD small molecule interactions.

A summary of the structural results obtained in this study is presented in Fig. 3 and 4, and the ESI Table S1.† We present the results in terms of the H-bond interaction distances from the different models and also the RMSD to the original X-ray coordinates (Fig. 3). We can also compare these distances to benchmark values obtained from a search of high resolution kinase X-ray structures containing heterocyclic inhibitors, and comparable interactions in high resolution, small molecule crystal structures. We observed that there is a general trend toward lower mean RMSDs with the increasing accuracy of the computation method used (Fig. 3a). However it is clear that the MP2 based results do not show the best agreement with the original X-ray results. Although the X-ray structures are not ideal standards, a general trend to lower RMSDs is still expected to be a reasonable measure of computational success, at least up to a point. A further measure that one can use to assess the overall quality of the method is the predicted interaction distances.

As a result of the ~ 2.0 Å resolution of the X-ray structures, and the lack of hydrogen information (atoms were added using

the AMBER forcefield⁵⁷), the X-ray structure derived distances are not ideal. We also make reference to distances derived from comparable high resolution small molecule crystal structures.⁷⁸ We can compare these values to (a) the mean value from the 8 X-ray structures used here, (b) the mean over comparable, high resolution structures reported in the PDB (c) and the mean distance between a heterocyclic nitrogen and an amide based on those reported in the small molecule Cambridge Structural Database (CSD).⁵⁷ In Fig. 3b it can be seen that the average distance between the ligand hetero-atom and the hinge H-bond donor, generally improves with increasing level of theory, albeit with the MP2 based method again being an outlier. AM1, CHARMM and HF/6-31G* show mean values higher than the mean of the original X-ray complexes, or benchmark values taken from the CSD and PDB sources. In contrast, the MP2/6-31+G**, M06-2X variants and HF/3-21G variants show lower means than the mean of the original X-ray complexes, or benchmark values from the PDB. Apart from the MP2/6-31+G** set, the mean values are very close to the values obtained from the CSD reference set suggesting the models here have lost a degree of their kinase character. These results also show that the neglect of the protein environment generally leads to a greater association between the protein model and the ligand. For the MP2 based result the effect is even more pronounced suggesting that the increased accuracy of the method is not beneficial since the structures deviate more significantly from those in the protein environment. Note, this does not mean that isolated QM active site models have no value in SBD. Indeed, if this was the case then data from the CSD would probably not prove useful in design efforts.⁷⁸ The value of a simplified QM model is that it represents the best case interaction between the moieties concerned, without external electrostatic or VDW constraints. Alteration of the real ligand in the protein environment, so that it can adopt the preferred low energy conformation observed in the gas phase, may help to maximise the intermolecular interaction.

Looking more broadly at the structural results, we can see that the trends identified using the computationally demanding MP2/6-31+G**,⁵⁷ are reproduced using the M06-2X methods and both HF 3-21G based models. AM1 and CHARMM models have generally larger, but also much more variable, distances between the ligand and protein hinge model compared to the other methods, and experimental benchmark values.⁵⁷ This is perhaps unsurprising in that rigorous charge derivation for MM methods is reported to be needed, or additional terms added. In addition, AM1 semi-empirical methods are being superseded by the newer PM6 variants, as well as PM6 with additional customization.^{46,54} HF/6-31G* models seem to systematically underestimate the association compared to the M06-2X and MP2 based models. The inclusion of water solvent was also investigated using an implicit PCM solvent model. The M06-2X/6-31G* and HF/3-21G models were reoptimized using the PCM model. The results in Fig. 3 show that the mean RMSD and central H-bond interaction at the hinge are slightly lower compared to the related gas phase optimization. Given the considerable computational overhead, such treatment may not therefore be warranted, at least in terms of an assessment of the structural features.

The results reported here indicate that the structures obtained can vary noticeably depending on the method used. Validation of the method for the system under investigation should be

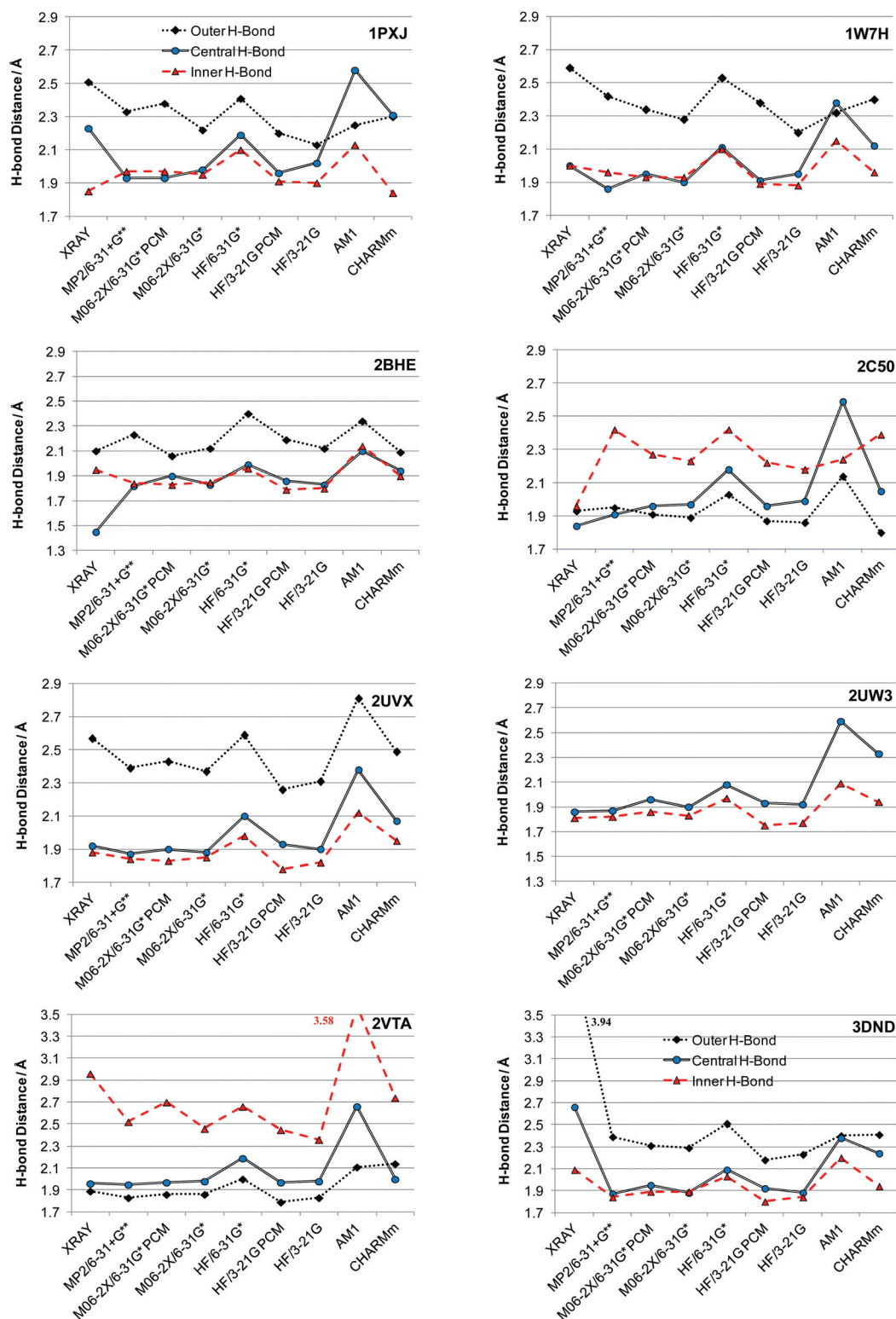


Fig. 4 A comparison of the H-bond distances obtained from 8 theoretical models and the original X-ray coordinates.

undertaken to ensure that the method can reliably account for the interactions present within the system in question. These results show that relatively rapid methods could be used to assess interactions of non-charged heterocycles. In particular, the well

validated M06-2X method with a modestly sized basis set, gave optimized structures with both the lowest RMSDs, and interaction distances closest to benchmark values, at relatively modest computational expense.

3.2 Effect of methodology of QM interaction energies

An additional application of QM cluster models is in the computation of enthalpies of binding. The goal of such a method is not to compute a realistic binding free energy, rather it is to try and assess the strength of polar interactions between a molecule and a probe (or active site representation in our case). For example, a number of methods are available to assess the H-bonding potential of different heterocycles or functional groups.^{41,43,44,49–51} The basis of such methods is that substituents or frameworks that have the optimal potential to interact with the polar features of an active site should lead to greater binding. However, in most receptors multiple features are present which have distinct directionality, meaning simple models may not be so ideal to compute interaction energies. The use of a model more closely mimicking the polar features of the active site might prove advantageous in assessing alterations to a core template, or different templates completely.

It is not expected that simplistic enthalpies will correlate strongly with the experimental free energy related parameters such as the K_i or IC_{50} , especially for such a diverse set of templates, across a range of protein kinases, as sampled in this study. Indeed, in this study it should be noted that the dataset chosen here consists of molecules with moderate to low potency for their particular kinase (Table 1). The observed standard deviation of 0.6 log units corresponds to just a 0.91 kcal mol⁻¹ difference in energy according to the Arrhenius equation, which is below the accuracy of many theoretical methods. Thus, even in the best case scenario, a correlation between the predicted interaction energy and the activity would not be expected (especially since the contribution of hydrophobic effects also need to be considered in any evaluation). Nevertheless, in this study we are interested in examining the magnitude of the differences in interaction energies for the different methods assessed, as each method treats H-bond interactions, bond lengths and angles *etc.* to different degrees of accuracy. These differences will have a dramatic effect on the rank ordering, which is especially pertinent if used in a design setting. For example, diffuse functions are suggested in cases where negative charges are present as delocalized can occur within the higher orbitals. The presence of halogen bonds necessitates additional parameters for PM6, and will be poorly described using MM methods for example.⁹

It is important to note that the MP2/6-31+G** energies are the most rigorous that have been obtained here. However, the optimized geometries deviate slightly more from the X-ray coordinates than those from M06-2X for example. Nevertheless, they are expected to be the most suitable here in terms of describing the interactions and conformational energies in the systems under investigation. Thus, we compare the interaction energies of all methods to these benchmark values (Table S2†).

The correlation between the energies obtained at the MP2/6-31+G**, M06-2X/6-31G*, HF/6-31G*, HF/3-21G, AM1 and CHARMM are reported in the ESI (Fig. S1†). The MP2/6-31+G** energies correlate well with those at M06-2X/6-31G* ($r^2 = 0.74$) and HF/6-31G* ($r^2 = 0.83$). Methods such as HF/3-21G and AM1, relying on smaller basis sets, do not correlate as well, with r^2 's of 0.54 and 0.36, respectively. The CHARMM

based energies show no correlation with the MP2 based results, or any other QM measure.

Also investigated was the effect of BSSE, a common artifact in QM calculations that can lead to inaccurate interaction energies. BSSE arises due to orbitals in the combined complex, which have negligible overlap, and can in fact lead to a lowering of the overall energy in the combined complex compared to the isolated components. This effect can be removed in the QM calculation of each individual component by including the ghost orbitals of the other component. The results from BSSE calculations at the M06-2X/6-31G* and HF/3-21G models are reported in the ESI (Fig. S2†). The correlation between the BSSE corrected energy and the uncorrected value for HF/3-21G displayed an r^2 of 0.74, while the value for the calculation at M06-2X/6-31G* level was 0.97. The effect of a common solvent model (PCM) was also investigated for both the M06-2X/6-31G* and HF/3-21G models, and these results are also reported in Fig. S2.† The M06-2X results including a PCM solvent model of water correlates only moderately well with the gas phase energies ($r^2 = 0.57$) while those at the HF/3-21G level display an r^2 of just 0.27. These results also highlight the dramatic effect the inclusion of solvent can have on the rank ordering for a given method.

The overall correlation between the different energies can be appreciated more clearly using principal components analysis (PCA). PCA is a method for identifying small numbers of correlated, orthogonal components for a dataset containing many descriptors. The QM energies (descriptors), and the kinase QM models (observations), that show a high degree of inter correlation will be located in the same region of component space on the combined scores/loadings bi-plot. In this case, a two component model can describe over 80% of the total variation in the dataset of 10 descriptors and 8 observations (Fig. 5). The

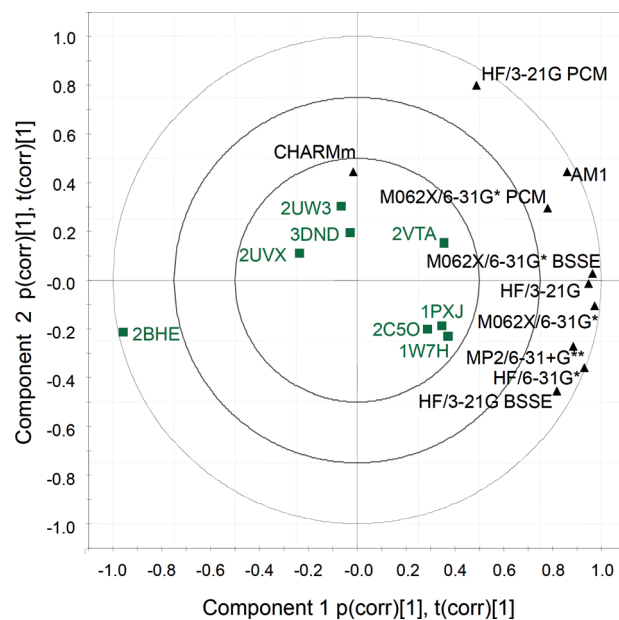


Fig. 5 PCA loadings bi-plot highlighting the inter-correlation between the different computed energies. The 2 component model describes 82% of the total variation (66 and 16% respectively for components 1 and 2) in the 10 energies computed for the 8 different model proteins.

combined loading/score plots show that all of the QM energies display a significant degree of correlation on component 1, as can be seen by their large positive loading. The CHARMM based result correlates poorly with the QM results, since it is located close to the origin on component 1. On component 2, the HF/3-21G PCM, AM1, M06/6-31G* PCM and the CHARMM model deviate more significantly from the other computed energies as can be discerned from their more positive loadings.

These results appear to suggest that the use of a moderately sized basis set, such as 6-31G* is preferred, especially as the effect of BSSE is minimal. The fact that the M06-2X method gives energies close to those of MP2/6-31+G**, and also good geometries (unlike HF/6-31G* for example which also correlates well), suggests it may be a preferred method to compute interactions energies. However, given the impact of the implicit solvent correction on the energies, it may also be beneficial to evaluate this term given its possible impact on rank ordering.

4. Conclusions

A number of methods are available to assess the H-bonding potential of different heterocycles or functional groups.^{41,43,44,49–51} The advantage of such methods is that substituents or frameworks that have the potential to more effectively interact, will presumably lead to greater binding efficiency with a receptor with an opposing feature (assuming it does not interact with water to a greater extent). However, in most receptors multiple features are present and these have distinct directionality meaning simple models may not be so ideal to compute interaction energies.

The use of more representative cluster models, more closely mimicking the polar features of a specific active site, might prove advantageous in assessing substituent alterations to a template, or different templates altogether. Understanding the strength of the polar interactions formed between a ligand and the active site is important if we wish to improve the formers enthalpic binding efficiency.^{13–15} Such an understanding would be beneficial in our attempts to increase the ligand efficiency of molecules in development and concomitantly improve their ADMET characteristics.^{16,67}

In this study we have assessed the effect of using a number of different theoretical methods to optimize QM active site models of protein kinase–ligand complexes. We employed MP2/6-31+G**, M06-2X/6-31G*, HF/6-31G*, HF/3-21G, AM1 and CHARMM methods, and considered the effect of BSSE and the inclusion of an implicit solvation model. We are interested in the effect these different choices have on the structures and energetics obtained for the systems in question. The results reported here on small, active site models, indicate that the structures obtained can vary noticeably depending on the method used. Validation of the method for the system under investigation should be undertaken to ensure that the method can reliably account for the interactions present within the system in question.

These results show that relatively rapid methods could be used to assess interactions of non-charged heterocycles, using the well validated M06-2X method with a modestly sized basis set,

giving optimized structures with both the lowest RMSDs, and interaction distances closest to benchmark values, at relatively modest computational expense. Analysis of the computed energies shows that a significant degree of correlation exists between the methods. The effect of BSSE on the rank ordering of the ligands in this study is negligible with a moderately sized basis set such as 6-31G*. The effect of PCM was shown to be more significant and may warrant consideration. The observation that the M06-2X method gives energies close to those of MP2/6-31+G**, and also reasonable optimized geometries, suggests it is the preferred method here for computing interactions energies.

The information derived from such models could be used to guide the ranking and selection of substituents or heterocyclic templates to improve their ligand efficiency by maximizing polar interactions. Alternately small QM models (or more descriptive QM/MM models⁴⁷) could be employed to benchmark ligand conformation and active site interactions which could be used to guide the refinement of X-ray structures, particularly of low to moderate resolution. We believe that while such calculations certainly have limitations, they have a place in SBD applications, alongside methods such as experimental X-ray structure, CSD structural analyses, QM/MM calculations of full protein–ligand complexes, with each offering a different insight into the interactions found within biological complexes.

Acknowledgements

This work was supported by the Thailand Research Fund grant RSA5480016 (MPG) and RMU5180032 (DG) and the Faculty of Science, Kasetsart University (KU).

Notes and references

- 1 Y. N. Imai, Y. Inoue and Y. Yamamoto, *J. Med. Chem.*, 2007, **50**, 1189–1196.
- 2 T. Zhou, D. Z. Huang and A. Caffisch, *Curr. Top. Med. Chem.*, 2010, **10**, 33–45.
- 3 S. Grimme, *Angew. Chem., Int. Ed.*, 2008, **47**, 3430–3434.
- 4 L. Estevez, N. Otero and R. A. Mosquera, *J. Phys. Chem. A*, 2009, **113**, 11051–11058.
- 5 E. A. Meyer, R. K. Castellano and F. Diederich, *Angew. Chem., Int. Ed.*, 2003, **42**, 1210–1250.
- 6 L. M. Salonen, M. Ellermann and F. Diederich, *Angew. Chem., Int. Ed.*, 2011, **50**, 4808–4842.
- 7 R. Wu and T. B. McMahon, *J. Am. Chem. Soc.*, 2008, **130**, 12554–12555.
- 8 Y. Zhao, H. T. Ng and E. Hanson, *J. Chem. Theory Comput.*, 2009, **5**, 2726–2733.
- 9 P. Zhou, J. W. Zou, F. F. Tian and Z. C. Shang, *J. Chem. Inf. Model.*, 2009, **49**, 2344–2355.
- 10 Y. Lu, T. Shi, Y. Wang, H. Yang, X. Yan, X. Luo, H. Jiang and W. Zhu, *J. Med. Chem.*, 2009, **52**, 2854–2862.
- 11 K. A. Brameld, B. Kuhn, D. C. Reuter and M. Stahl, *J. Chem. Inf. Model.*, 2008, **48**, 1–24.
- 12 M.-H. Hao, O. Haq and I. Muegge, *J. Chem. Inf. Model.*, 2007, **47**, 2242–2252.
- 13 G. G. Ferenczy and G. M. Keseru, *J. Chem. Inf. Model.*, 2010, **50**, 1536–1541.
- 14 G. M. Keseru and G. G. Ferenczy, *Drug Discovery Today*, 2010, **15**, 919–932.
- 15 G. M. Keseru and G. M. Makara, *Nat. Rev. Drug Discovery*, 2009, **8**, 203–212.
- 16 M. P. Gleeson, A. Hersey, D. Montanari and J. Overington, *Nat. Rev. Drug Discovery*, 2011, **10**, 197–208.
- 17 M. M. Hann, *Med. Chem. Commun.*, 2011, **2**, 349–355.

- 18 C. Abad-Zapatero and J. T. Metz, *Drug Discovery Today*, 2005, **10**, 464–469.
- 19 S. D. Bembek, B. A. Tounge and C. H. Reynolds, *Drug Discovery Today*, 2009, **14**, 278–283.
- 20 A. L. Hopkins, C. R. Groom and A. Alex, *Drug Discovery Today*, 2004, **9**, 430–431.
- 21 E. Perola, *J. Med. Chem.*, 2010, **53**, 2986–2997.
- 22 C. H. Reynolds, S. D. Bembek and B. A. Tounge, *Bioorg. Med. Chem. Lett.*, 2007, **17**, 4258–4261.
- 23 C. H. Reynolds and M. K. Holloway, *ACS Med. Chem. Lett.*, 2011, **2**, 433–437.
- 24 C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Adv. Drug Delivery Rev.*, 2001, **46**, 3–26.
- 25 M. P. Gleeson, *J. Med. Chem.*, 2008, **51**, 817–834.
- 26 P. Gleeson, G. Bravi, S. Modi and D. Lowe, *Bioorg. Med. Chem.*, 2009, **17**, 5906–5919.
- 27 A. R. Leach, *Molecular Modelling. Principles and Applications*, Longman limited, Harlow, 1996.
- 28 A. de Ruiter and C. Oostenbrink, *Curr. Opin. Chem. Biol.*, 2011, **15**, 547–552.
- 29 S. L. Dixon and K. M. J. Merz, *J. Chem. Phys.*, 1996, **104**, 6643–6649.
- 30 K. Raha, A. J. van der Vaart, K. E. Riley, M. B. Peters, L. M. Westerhoff, H. Kim and K. M. Merz, *J. Am. Chem. Soc.*, 2005, **127**, 6583–6594.
- 31 K. M. Merz and M. B. Peters, *J. Chem. Theory Comput.*, 2006, **2**, 383–399.
- 32 P. Dobeš, J. Řezáč, J. i. Fanfrlík, M. Otyepka and P. Hobza, *J. Phys. Chem. B*, 2011, **115**, 8581–8589.
- 33 U. C. Singh and P. A. Kollman, *J. Comput. Chem.*, 1986, **7**, 718–730.
- 34 A. Warshel, *Annu. Rev. Biophys. Biomol. Struct.*, 2003, **32**, 425–443.
- 35 T. Vreven, K. S. Byun, I. Komáromi, S. Dapprich, J. A. Montgomery, K. Morokuma and M. J. Frisch, *J. Chem. Theory Comput.*, 2006, **2**, 815–826.
- 36 X. Li, S. A. Hayik and K. M. Merz, *J. Inorg. Biochem.*, 2010, **104**, 512–522.
- 37 L. C. Menikarachchi and J. A. Gascon, *Curr. Top. Med. Chem.*, 2010, **10**, 46–54.
- 38 H. M. Senn and W. Thiel, *Angew. Chem., Int. Ed.*, 2009, **48**, 1198–1229.
- 39 H. Lin and D. G. Truhlar, *Theor. Chem. Acc.*, 2007, **117**, 185–199.
- 40 S. Shaik, S. Cohen, Y. Wang, H. Chen, D. Kumar and W. Thiel, *Chem. Rev.*, 2010, **110**, 949–1017.
- 41 C. Gueto-Tettay, J. Drosos and R. Vivas-Reyes, *J. Comput.-Aided Mol. Des.*, 2011, **25**, 583–597.
- 42 M. B. Peters, K. Raha and K. M. Merz, *Curr. Opin. Drug Discovery*, 2006, **9**, 370–379.
- 43 R. Villar, M. J. Gil, J. I. Garcia and V. Martinez-Merino, *J. Comput. Chem.*, 2005, **26**, 1347–1358.
- 44 J. Schwöbel, R.-U. Ebert, R. Kühne and G. Schüürmann, *J. Comput. Chem.*, 2009, **30**, 1454–1464.
- 45 R. S. Paton and J. M. Goodman, *J. Chem. Inf. Model.*, 2009, **49**, 944–955.
- 46 M. Korth, *J. Chem. Theory Comput.*, 2010, **6**, 3808–3816.
- 47 U. Ryde, *Dalton Trans.*, 2007, 607–625.
- 48 H. Hu and W. Yang, *Annu. Rev. Phys. Chem.*, 2008, **59**, 573–601.
- 49 J. Schwöbel, R.-U. Ebert, R. Kühne and G. Schüürmann, *J. Phys. Chem. A*, 2009, **113**, 10104–10112.
- 50 M. Nocker, S. Handschuh, C. Tautermann and K. R. Liedl, *J. Chem. Inf. Model.*, 2009, **49**, 2067–2076.
- 51 J. A. H. Schwöbel, R.-U. Ebert, R. Kühne and G. Schüürmann, *J. Phys. Org. Chem.*, 2011, **24**, 1072–1080.
- 52 P. E. M. Siegbahn and F. Himo, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2011, **1**, 323–336.
- 53 P. Siegbahn and F. Himo, *JBIC, J. Biol. Inorg. Chem.*, 2009, **14**, 643–651.
- 54 U. Ryde and K. Nilsson, *J. Mol. Struct. (THEOCHEM)*, 2003, **632**, 259–275.
- 55 M. P. Gleeson and D. Gleeson, *J. Chem. Inf. Model.*, 2009, **49**, 1437–1448.
- 56 M. P. Gleeson and D. Gleeson, *J. Chem. Inf. Model.*, 2009, **49**, 670–677.
- 57 M. P. Gleeson, S. Hannongbua and D. Gleeson, *J. Mol. Graphics Modell.*, 2010, **29**, 507–517.
- 58 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. M. Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski and D. J. Fox, Gaussian Inc., Wallingford CT, 2009.
- 59 Y. Zhao and D. G. Truhlar, *Theor. Chem. Acc.*, 2008, **120**, 215–241.
- 60 Y. Zhao and D. G. Truhlar, *J. Phys. Chem. C*, 2008, **112**, 6860–6868.
- 61 Y. Zhao and D. G. Truhlar, *Chem. Phys. Lett.*, 2011, **502**, 1–13.
- 62 Discovery Studio 2.5, Accelrys Inc, San Diego, CA, United States <http://accelrys.com/products/discovery-studio/>.
- 63 F. B. van Duijneveldt, J. G. C. M. van Duijneveldt-van de Rijdt and J. H. van Lenthe, *Chem. Rev.*, 1994, **94**, 1873–1885.
- 64 A. M. Davis, S. J. Teague and G. J. Kleywegt, *Angew. Chem., Int. Ed.*, 2003, **42**, 2718–2736.
- 65 D. Yusuf, A. M. Davis, G. J. Kleywegt and S. Schmitt, *J. Chem. Inf. Model.*, 2008, **48**, 1411–1422.
- 66 A. Malde and A. Mark, *J. Comput.-Aided Mol. Des.*, 2011, **25**, 1–12.
- 67 P. D. Leeson and B. Springthorpe, *Nat. Rev. Drug Discovery*, 2007, **6**, 881–890.
- 68 S. Wang, C. Meades, G. Wood, A. Osnowski, S. Anderson, R. Yuill, M. Thomas, M. Mezna, W. Jackson, C. Midgley, G. Griffiths, I. Fleming, S. Green, I. McNae, S.-Y. Wu, C. McInnes, D. Zheleva, M. D. Walkinshaw and P. M. Fischer, *J. Med. Chem.*, 2004, **47**, 1662–1675.
- 69 G. Kontopidis, C. McInnes, S. R. Pandalaneni, I. McNae, D. Gibson, M. Mezna, M. Thomas, G. Wood, S. Wang, M. D. Walkinshaw and P. M. Fischer, *Chem. Biol.*, 2006, **13**, 201–211.
- 70 M. J. Hartshorn, C. W. Murray, A. Cleasby, M. Frederickson, I. J. Tickle and H. Jhoti, *J. Med. Chem.*, 2004, **48**, 403–413.
- 71 A. L. Gill, M. Frederickson, A. Cleasby, S. J. Woodhead, M. G. Carr, A. J. Woodhead, M. T. Walker, M. S. Congreve, L. A. Devine, D. Tisi, M. O'Reilly, L. C. A. Seavers, D. J. Davis, J. Curry, R. Anthony, A. Padova, C. W. Murray, R. A. E. Carr and H. Jhoti, *J. Med. Chem.*, 2004, **48**, 414–426.
- 72 R. Jautelat, T. Brumby, M. Schäfer, H. Briem, G. Eisenbrand, S. Schwahn, M. Krüger, U. Lücking, O. Prien and G. Siemeister, *Chem-BioChem*, 2005, **6**, 531–540.
- 73 A. Donald, T. McHardy, M. G. Rowlands, L.-J. K. Hunter, T. G. Davies, V. Berdini, R. G. Boyle, G. W. Aherne, M. D. Garrett and I. Collins, *J. Med. Chem.*, 2007, **50**, 2289–2292.
- 74 G. Saxty, S. J. Woodhead, V. Berdini, T. G. Davies, M. L. Verdonk, P. G. Wyatt, R. G. Boyle, D. Barford, R. Downham, M. D. Garrett and R. A. Carr, *J. Med. Chem.*, 2007, **50**, 2293–2296.
- 75 P. G. Wyatt, A. J. Woodhead, V. Berdini, J. A. Boulstridge, M. G. Carr, D. M. Cross, D. J. Davis, L. A. Devine, T. R. Early, R. E. Feltell, E. J. Lewis, R. L. McMenamin, E. F. Navarro, M. A. O'Brien, M. O'Reilly, M. Reule, G. Saxty, L. C. A. Seavers, D.-M. Smith, M. S. Squires, G. Trewartha, M. T. Walker and A. J. A. Woolford, *J. Med. Chem.*, 2008, **51**, 4986–4999.
- 76 J. Orts, J. Tuma, M. Reese, S. K. Grimm, P. Monecke, S. Bartoschek, A. Schiffer, K. U. Wendt, C. Griesinger and T. Carlomagno, *Angew. Chem., Int. Ed.*, 2008, **47**, 7736–7740.
- 77 J. J. L. Liao, *J. Med. Chem.*, 2007, **50**, 409–424.
- 78 F. H. Allen and R. Taylor, *Chem. Soc. Rev.*, 2004, **33**, 463–475.