MedChemComm

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: D. Cole, I. Cabeza de Vaca and W. L. Jorgensen, *Med. Chem. Commun.*, 2019, DOI: 10.1039/C9MD00017H.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/medchemcomm

View Article Online
DOI: 10.1039/C9MD00017H

Journal Name

COMMUNICATION

Computation of Protein-Ligand Binding Free Energies using Quantum Mechanical Bespoke Force Fields

Daniel J. Cole,*[a] Israel Cabeza de Vaca[b] and William L. Jorgensen[b]

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

A quantum mechanical bespoke molecular mechanics force field is derived for the L99A mutant of T4 lysozyme and used to compute absolute binding free energies of six benzene analogs to the protein. Promising agreement between theory and experiment highlights the potential for future use of system-specific force fields in computer-aided drug design.

Alchemical free energy methods are a widely used computational tool for predicting binding affinities of biomolecular complexes.[1-3] The ability to accurately predict protein-ligand binding free energy would have a profound impact on our ability to guide molecular design and prioritize compounds in early stage drug discovery.[4-7] One of the many computational techniques developed for such a purpose is free energy perturbation (FEP) theory. FEP, in combination with molecular dynamics (MD) or Monte Carlo (MC) sampling of biomolecular conformation, provides a rigorous means to compute protein-ligand binding free energy. Its accuracy is limited, in practice, only by the accuracy of the underlying molecular mechanics (MM) force field used to compute atomistic interactions and dynamics, and by finite sampling that can limit the exploration of conformational space.

Biomolecular MM force fields, such as OPLS, AMBER, CHARMM and GROMOS, have been under development for ca. 40 years. In these force fields, covalent bonding is described using harmonic bond stretching and angle bending components, in conjunction with anharmonic torsion rotation terms, while the non-bonded component is a sum of electrostatic interactions between atom-centered point charges and Lennard-Jones interactions. Strategies vary, but in general the parameters used to describe these interatomic potentials are fit to reproduce the quantum mechanical (QM) and experimental properties of small organic molecules.[8-11] This combination of a physically motivated functional form, alongside very careful parameterization, means that MM force fields generally describe biomolecular dynamics and interactions well.[12,13]

However, there is also potential room for improvement. Implicit in the described parameterization process is the assumption of transferability. That is, parameters that are fit to reproduce QM or experimental properties of small molecules are assumed to also be optimal for describing macromolecular dynamics. While methods are under development for rigorously correcting computed MM free energies by processing snapshots using hybrid QM/MM,[14-16] or even large-scale QM,[17] these approaches require many evaluations of the QM energy and it is as yet unclear how accurate they are when applied to protein-ligand binding.

An alternative to the use of transferable MM force fields for sampling molecular conformational space is the derivation of a system-specific, or bespoke, force field directly for the molecule under study, thus avoiding the parameter transferability assumption. Several different approaches to system-specific force field derivation have been developed.[18-20] All of these have in common the use of QM for parameter derivation, since experimental measurements are unlikely to be available for any but the simplest organic molecules. However, the applicability of these methods is limited to small system sizes (e.g. <100 atoms), either by the expense of the QM calculation or the sheer number of parameters required to describe biomolecular dynamics.

To address these issues, the QUantum mechanical BEspoke (QUBE) force field for protein modelling has recently been developed, as described in detail elsewhere.[21,22] In contrast to standard biomolecular force fields, non-bonded parameters are derived specifically for the system under study. This is made possible by advances in linear-scaling QM software, which enable the computation of the ground state electron density of systems comprising many thousands of atoms.[23] In QUBE, the

[a] School of Natural and Environmental Sciences, Newcastle University, Newcastle upon Tyne NE1 7RU, U.K.
[b] Department of Chemistry, Yale University, New Haven, Connecticut 06520-8107, U.S.A.

† Footnotes relating to the title and/or authors should appear here.
Electronic Supplementary Information (ESI) available: Computational methods, force field parameters and structure files. See DOI: 10.1039/x0xx00000x

View Article Online
DOI: 10.1039/C9MD00017H

total electron density is partitioned into atom-centered basins[24,25] and used to derive atoms-in-molecule charge and Lennard-Jones parameters.[26] In contrast, local covalent interactions (bond, angle and torsion components) are treated in a similar manner to transferable force fields. Bond and angle parameters are derived from the QM Hessian matrices of small, model dipeptides using the so-called modified Seminario method.[27] To complete the first version of the QUBE protein force field, torsion parameters have been extensively re-fit for compatibility with the rest of the QUBE parameter set, using QM dihedral scans.[22] Bonded parameters are read from a library and assumed to be transferable. The QUBE method has been validated and shown to be similar in accuracy to widely-used transferable force fields by comparing the condensed phase properties of a database of small, organic molecules with experiment, and by comparing computed protein dynamics with experimental NMR observables.[21,22] However, QUBE has not been used before now to predict protein-ligand binding affinity and it remains to be seen how accurate it is for potential use in computer-aided drug design applications.

One problem with the use of protein-ligand binding free energies for force field validation is that it is difficult to separate possible force field inaccuracy from the aforementioned limitations in conformational sampling, which can lead to incorrect predicted binding poses if not treated carefully. To get around this problem, in this paper, we investigate the binding of several small, rigid benzene derivatives to an engineered binding site of the T4-lysozyme L99A protein (hereinafter referred to as lysozyme) using the QUBE force field. This is a relatively simple binding site, and there is a wide range of experimental structural and calorimetric data available.[28,29] For these reasons, lysozyme has been recommended as a potential community benchmark test system,[3] and it is commonly used to validate new computational methodologies.[30-32] In particular, standard MM force fields yield reasonably accurate free energies,[33-35] though even for this relatively simple system, larger ligands can present challenging sampling requirements.[36]

The lysozyme binding site has recently been used to develop and validate new computational algorithms for exploring protein-ligand conformational space with the MCPRO software,[37] with emphasis on improved backbone and side chain Monte Carlo moves.[38] Alongside the recently implemented replica exchange with solute tempering (REST) method,[39-44] which is employed here to enhance sampling in localized regions of the system, we now have the ability to employ extensive sampling of protein-ligand binding modes and hence focus our efforts on validation and improvement of the QUBE force field.

The QUBE force field is designed specifically for the system under study through post-processing of the QM electron density, and due to its implementation in linear-scaling QM software it may be applied to a range of proteins comprising many thousands of atoms. Here, the entire lysozyme protein (164 residues, 2628 atoms) was parameterized using the QUBE force field (see the ESI for system preparation and methods). The QM electron density was obtained using the ONETEP linear-scaling density functional theory software,[45] and atoms-in-molecule analysis[24,25] was performed to extract atomic charges and Lennard-Jones parameters from the partitioned atomic electron densities.[26] Figure 1(a) compares the derived QUBE charges with widely-used OPLS force field parameters. In line with previous studies,[26] we observe a strong correlation between QUBE and OPLS charges. This is despite the very different parameterization strategies employed. While QUBE charges are derived directly from the QM electron density, OPLS charges have been very carefully fit to reproduce the experimental condensed phase properties of small organic molecules.[8] Also noteworthy is the spread of charges along the y-axis, which demonstrates the ability of QUBE charges to adapt to their chemical and electrostatic environment, which is expected to improve the point charge description of the electrostatic potential at the protein surface.

Figure 1(b) displays the correlation between summed QUBE and OPLS C_6 parameters per residue, which provides a measure of the interatomic van der Waals interaction, a fundamental driving force behind protein folding and dynamics (see the ESI for a full description). Again, strong correlation is observed between the two parameterization strategies. However, the QUBE van der Waals coefficients are approximately 1.5x lower than in the OPLS force field. This is in very good agreement with previous comparisons between QM-derived C_6 parameters and standard force fields, and does suggest that force field parameters have been elevated to effectively account for higher-order QM interactions absent in the simple force field functional form.[46] Nevertheless, experimental condensed phase properties of organic molecules are reproduced to a high degree of accuracy using the current QUBE model,[21] and so it is informative to ascertain its accuracy in describing protein-ligand binding.

Protein dynamics using the QUBE force field are extensively validated elsewhere,[22] however protein-ligand binding has not been assessed before. The software tool QUBEKit[21] has been employed to parameterize six small, rigid benzene analogs, and extensive MC sampling of protein-ligand conformational space has been performed using the MCPRO software and the REST enhanced sampling algorithm (see the ESI). Figure 2(a) shows an overlay of the simulated lysozyme-benzene complex with the experimental crystal structure (PDB: 181L). The agreement between the two structures is promising. Notably, the V111 side chain is known to undergo reorientation upon binding larger benzene derivatives (in particular, *o*-xylene and *p*-xylene from the set studied here).[29] Figure 2(b) indicates that this reorientation has been captured in the case of *o*-xylene. Since all input structures used here were based on the benzene-bound form, this is encouraging both for the QUBE force field description of the complex and the enhanced MC sampling employed (see the ESI). To study this reorientation in more detail, Figure 3 shows the distribution of side chain dihedral angles sampled during the MC simulations of lysozyme bound to *o*-xylene and *p*-xylene, and in its apo form. The latter structures are extracted from the final window of the free energy calculations, with the ligand fully annihilated. Bound to *o*-xylene, the V111 side chain explores three different rotamers with a preference for 300°. Bound to *p*-xylene, the V111 side

chain prefers the 180° rotamer, though some exploration of the 300° rotamer is observed. Experimentally, in both cases, the 300° rotamer is expected, though populations of alternative conformations may not be visible in the X-ray electron density. In all cases, in the unbound form, V111 has a strong preference for the 180° rotamer, in good agreement with experiment.[29] For the other four ligands studied here (data not shown), the 180° rotamer is preferred, and only indene shows any significant sampling of the 300° rotamer. Thus, simulated protein-ligand conformations are in broad agreement with experimental observations. However, a much more reliable indicator of force field accuracy is the use of free energy calculations to compute protein-ligand binding affinity.

Absolute binding free energies of benzene, and five further derivatives, to lysozyme were computed using the QUBE force field, and compared with experiment (Table 1). Overall, the mean unsigned error (MUE) of 0.85 kcal/mol is extremely competitive with established, widely-used transferable force fields. For comparison, the MUEs obtained for the same dataset in recent literature studies are 1.26 kcal/mol using the OPLS force field,[38] and 2.09 kcal/mol using the AMBER/AM1-BCC force field and the confine-release protocol (Table S1).[33] In the current study, we did not observe any significant outliers from experiment. Indole, indene and benzofuran each have UE > 1 kcal/mol. Inaccuracy in the description of indole binding may be explained by the neglect of off-site charges to model anisotropic electron density. We have shown previously that such effects are important in the accurate description of interactions of the sulfur atom in the neighboring binding site residue M102.[26] The difference between the computed binding free energies of indene and benzofuran are more difficult to explain, and may reflect inaccuracy in this first-generation protein force field. A previous study examining small molecule properties has shown that computed hydration free energies of small non-polar molecules using QUBE tend to be too positive compared with experiment.[21] Similar results are found here (Table S2), which may explain some of the deviations from experiment, though some cancellation of errors is expected in the free energy cycles used to compute protein-ligand binding. Indeed, the binding free energies of the non-polar *o*-xylene and *p*-xylene are in particularly close agreement with experiment, which indicates not only that the QUBE force field is able to describe changes in van der Waals interactions upon moving between water and the protein binding site (Figure 1(b)), but also that the use of REST enhanced sampling is able to capture the V111 reorientation upon binding these larger analogs (Figure 3).

Conclusions

In summary, the quantum mechanical bespoke force field, QUBE, has been used to compute the absolute binding free energies of six benzene derivatives to the engineered L99A mutant of T4 lysozyme. Encouraging agreement is obtained between theory and experiment with a MUE of just 0.85 kcal/mol. Known challenges regarding the reorientation of V111 in the protein binding site have been addressed through recent improvements to MC configurational sampling algorithms and the use of a local enhanced sampling (REST) method. It is important to note that, since the majority of the force field parameters used in the current study are derived from quantum mechanics rather than fit to experiment, there is significant potential for systematic improvement of the results. For example, new force fields employing more accurate underlying quantum mechanical methods or alternative atoms-in-molecule analysis schemes[47] may be employed to compute the partitioned atomic electron density. More fundamentally, the functional form of the force field may be updated by, for example, including terms beyond the C_6r^{-6} interaction (Figure 1(b)) to describe van der Waals interactions,[48,49] or altering the short-range repulsion term,[50] or including more advanced torsion potentials to account for coupling between internal coordinates.[51,52] Beyond improvement of the QUBE force field, we are also planning further validation of protein-ligand binding free energies beyond the relatively simple lysozyme model and extension of the applicability of the method to, for example, the description of metal centers in protein binding sites.[53] Overall, however, the current study represents an encouraging first step toward the use of system-specific QM-derived force fields in molecular design and prospective drug discovery efforts.

Conflicts of interest

There are no conflicts to declare.

Notes and references

‡ The authors are grateful to Alice Allen (ORCID: 0000-0002-8727-8333) and Joshua Horton (ORCID: 0000-0001-8694-7200) for helpful discussions, and to the EPSRC (EP/R010153/1) and the National Institutes of Health (GM32136) for funding. This research made use of the Rocket High Performance Computing Service at Newcastle University.

1 J. D. Chodera, D. L. Mobley, M. R. Shirts, R. W. Dixon, K. Branson and V. S. Pande, *Current Opinion in Structural Biology*, 2011, **21**, 150.
2 W. L. Jorgensen, *Acc. Chem. Res.*, 2009, **42**, 724.
3 D. L. Mobley and M. K. Gilson, *Annual Review of Biophysics*, 2017, **46**, 531.
4 L. Wang, Y. Wu, Y. Deng, B. Kim, L. Pierce, G. Krilov, D. Lupyan, S. Robinson, M. K. Dahlgren, J. Greenwood, D. L. Romero, C. Masse, J. L. Knight, T. Steinbrecher, T. Beuming, W. Damm, E. Harder, W. Sherman, M. Brewer, R. Wester, M. Murcko, L. Frye, R. Farid, T. Lin, D. L. Mobley, W. L. Jorgensen, B. J. Berne, R. A. Friesner and R. Abel, *J. Am. Chem. Soc.*, 2015, **137**, 2695.
5 C. D. Christ and T. Fox, *J. Chem. Inf. Model.*, 2014, **54**, 108.
6 W.-G. Lee, R. Gallardo-Macias, K. M. Frey, K. A. Spasov, M. Bollini, K. S. Anderson and W. L. Jorgensen, *J. Am. Chem. Soc.*, 2013, **135**, 16705.
7 B. J. Williams-Noonan, E. Yuriev and D. K. Chalmers, *J. Med. Chem.*, 2018, **61**, 638.
8 W. L. Jorgensen, D. S. Maxwell and J. Tirado-Rives, *J. Am. Chem. Soc.*, 1996, **118**, 1225.
9 D. Shivakumar, E. Harder, W. Damm, R. A. Friesner and W. Sherman, *J. Chem. Theory Comput.*, 2012, **8**, 2553.
10 L. S. Dodda, J. Z. Vilseck, K. J. Cutrona, W. L. Jorgensen, *J. Chem. Theory Comput.*, 2015, **11**, 4273.
11 J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case, *J. Comput. Chem.*, 2004, **25**, 1157.

12 K. A. Beauchamp, Y.-S. Lin, R. Das and V. S. Pande, *J. Chem. Theory Comput.*, 2012, **8**, 1409.
13 M. J. Robertson, J. Tirado-Rives, W. L. Jorgensen, *J. Chem. Theory Comput.*, 2015, **11**, 3499.
14 F. R. Beierlein, J. Michel and J. W. Essex, *J. Phys. Chem. B*, 2011, **115**, 4911.
15 M. A. Olsson and U. Ryde, *J. Chem. Theory Comput.*, 2017, **13**, 2245.
16 M. Wang, Y. Mei and U. Ryde, *J. Chem. Theory Comput.*, 2018, **14**, 6613.
17 C. Sampson, T. Fox, C. S. Tautermann, C. Woods and C.-K. Skylaris, *J. Phys. Chem. B*, 2015, **119**, 7030.
18 S. Grimme, *J. Chem. Theory Comput.*, 2014, **10**, 4497.
19 A. J. Misquitta and A. J. Stone, *J. Chem. Theory Comput.*, 2016, **12**, 4184.
20 I. Cacelli, A. Cimoli, P. R. Livotto and G. Prampolini, *J. Comp. Chem.*, 2012, **33**, 1055.
21 J. T. Horton, A. E. A. Allen, L. S. Dodda and D. J. Cole, *J. Chem. Inf. Model.*, 2019, https://doi.org/10.1021/acs.jcim.8b00767
22 A. E. A. Allen, M. J. Robertson, M. C. Payne and D. J. Cole, 2019, https://doi.org/10.26434/chemrxiv.7565222.v1
23 D. J. Cole and N. D. M. Hine, *J. Phys.: Condens. Matter*, 2016, **28**, 393001.
24 T. A. Manz and D. S. Sholl, *J. Chem. Theory Comput.*, 2012, **8**, 2844.
25 L. P. Lee, N. G. Limas, D. J. Cole, M. C. Payne, C.-K. Skylaris and T. A. Manz, *J. Chem. Theory Comput.*, 2014, **10**, 5377.
26 D. J. Cole, J. Z. Vilseck, J. Tirado-Rives, M. C. Payne and W. L. Jorgensen, *J. Chem. Theory Comput.*, 2016, **12**, 2312.
27 A. E. A. Allen, M. C. Payne and D. J. Cole, *J. Chem. Theory Comput.*, 2018, **14**, 274.
28 A. Morton, W. A. Baase and B. W. Matthews, *Biochemistry*, 1995, **34**, 8564.
29 A. Morton and B. W. Matthews, *Biochemistry*, 1995, **34**, 8576.
30 E. Gallicchio, M. Lapelosa and R. M. Levy, *J. Chem. Theory Comput.*, 2010, **6**, 2961.
31 K. Wang, J. D. Chodera, Y. Yang and M. R. Shirts, *J Comput Aided Mol Des*, 2013, **27**, 989.
32 W. Jiang and B. Roux, *J. Chem. Theory Comput.*, 2010, **6**, 2559.
33 D. L. Mobley, A. P. Graves, J. D. Chodera, A. C. McReynolds, B. K. Shoichet and K. A. Dill, *J. Mol. Biol.*, 2007, **371**, 1118.
34 I. V. Khavrutskii and A. Wallqvist, *J. Chem. Theory Comput.*, 2011, **7**, 3001.
35 Y. Deng and B. Roux, *J. Chem. Theory Comput.*, 2006, **2**, 1255.
36 N. M. Lim, L. Wang, R. Abel and D. L. Mobley, *J. Chem. Theory Comput.*, 2016, **12**, 4620.
37 W. L. Jorgensen and J. Tirado-Rives, *J. Comput. Chem.*, 2005, **26**, 1689.
38 I. Cabeza de Vaca, Y. Qian, J. Z. Vilseck, J. Tirado-Rives and W. L. Jorgensen, *J. Chem. Theory Comput.*, 2018, **14**, 3279.
39 P. Liu, B. Kim, R. A. Friesner and B. J. Berne, *PNAS*, 2005, **102**, 13749.
40 L. Wang, R. A. Friesner and B. J. Berne, *J. Phys. Chem. B*, 2011, **115**, 9431.
41 L. Wang, B. J. Berne and R. A. Friesner, *PNAS*, 2012, **109**, 1937.
42 D. J. Cole, J. Tirado-Rives and W. L. Jorgensen, *J. Chem. Theory Comput.*, 2014, **10**, 565.
43 D. J. Cole, J. Tirado-Rives and W. L. Jorgensen, *Biochimica et Biophysica Acta (BBA) - General Subjects*, 2015, **1850**, 966.
44 D. J. Cole, M. Janecek, J. E. Stokes, M. Rossmann, J. C. Faver, G. J. McKenzie, A. R. Venkitaraman, M. Hyvönen, D. R. Spring, D. J. Huggins and W. L. Jorgensen, *Chem. Commun.*, 2017, **53**, 9372.
45 C.-K. Skylaris, P. D. Haynes, A. A. Mostofi and M. C. Payne, *J. Chem. Phys.*, 2005, **122**, 084119.
46 M. Mohebifar, E. R. Johnson and C. N. Rowley, *J. Chem. Theory Comput.*, 2017, **13**, 6146–6157.
47 T. A. Manz and N. Gabaldon Limas, *RSC Advances*, 2016, **6**, 47771.
48 A. J. Stone, *Science*, 2008, **321**, 787.
49 A. J. Misquitta and A. J. Stone, *J. Chem. Theory Comput.*, 2016, **12**, 4184.
50 M. J. Van Vleet, A. J. Misquitta, A. J. Stone and J. R. Schmidt, *J. Chem. Theory Comput.*, 2016, **12**, 3851.
51 J. Cerezo, G. Prampolini and I. Cacelli, *Theor. Chem. Acc.*, 2018, **137**, 80.
52 W. Kang, F. Jiang, Y.-D. Wu, *J. Chem. Theory Comput.*, 2018, **14**, 4474.
53 L. Hu and U. Ryde, *J. Chem. Theory Comput.*, 2011, **7**, 2452.

View Article Online DOI: 10.1039/C9MD00017H

Table 1. Comparisons between computed (QUBE) absolute free energies of binding (kcal/mol) and experiment.[28] Where computations have been run in triplicate, the standard error in the mean free energy is quoted in parentheses.

	ΔG^{exp}	ΔG^{QUBE}	UE
benzene	-5.19	-5.97 (0.2)	0.78
***p*-xylene**	-4.67	-4.41	0.26
***o*-xylene**	-4.60	-4.98	0.38
benzofuran	-5.46	-6.95 (0.2)	1.49
indole	-4.89	-3.84 (0.4)	1.05
indene	-5.13	-4.01 (0.3)	1.12
MUE			**0.85**

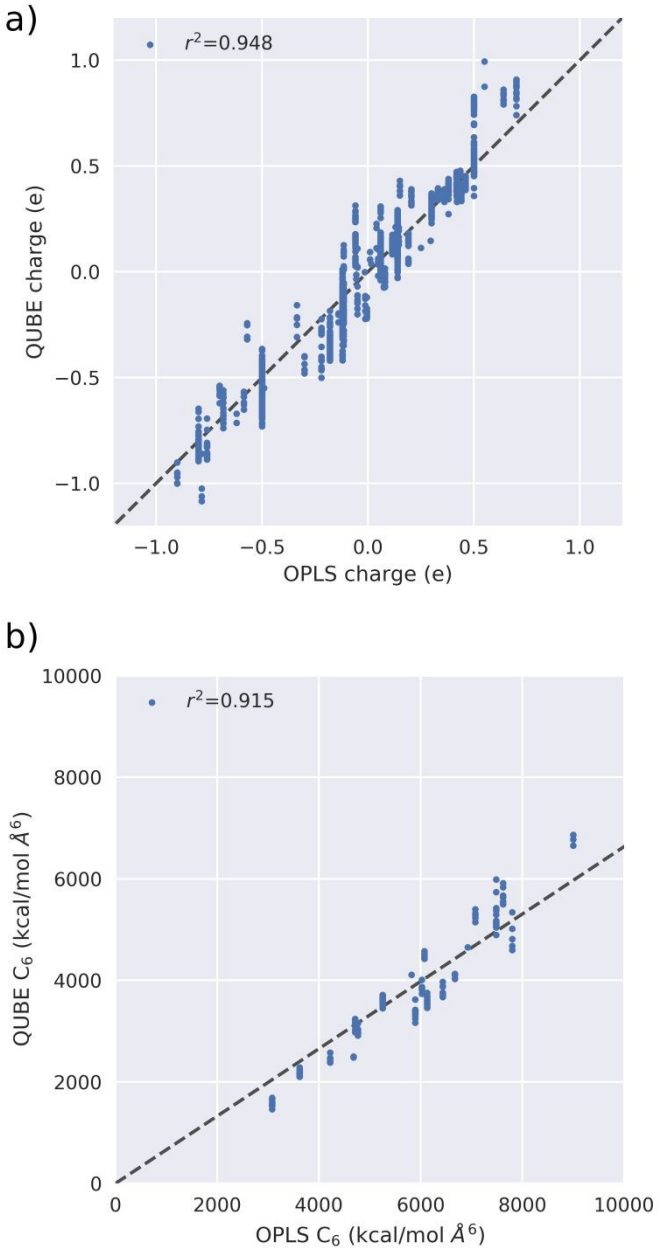


Figure 1. a) Correlation between derived QUBE and OPLS atomic charges for the lysozyme protein. The dashed line is given by the equation y = x. b) Correlation between derived QUBE and OPLS C6 parameters per residue. The dashed line is given by the equation y = 0.66x.

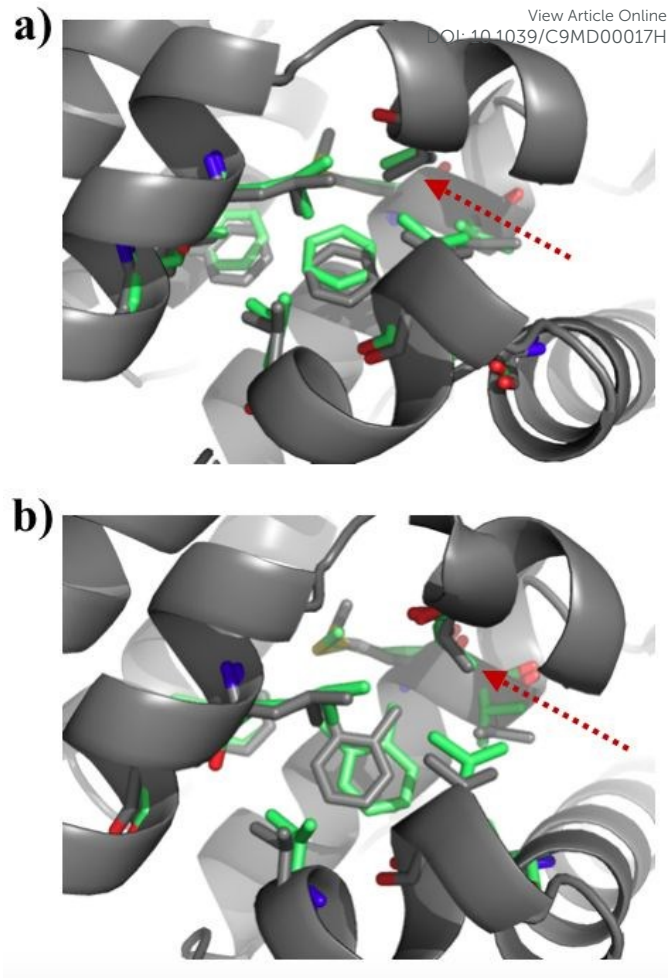


Figure 2. a) Overlay of crystal structures (grey) and binding site residues from final snapshots of MC simulations (green) of (a) lysozyme-benzene (after 30 million Monte Carlo steps) and (b) lysozyme-*o*-xylene (after 80 million Monte Carlo steps) complexes. The residue V111 is indicated by an arrow.

View Article Online
DOI: 10.1039/C9MD00017H

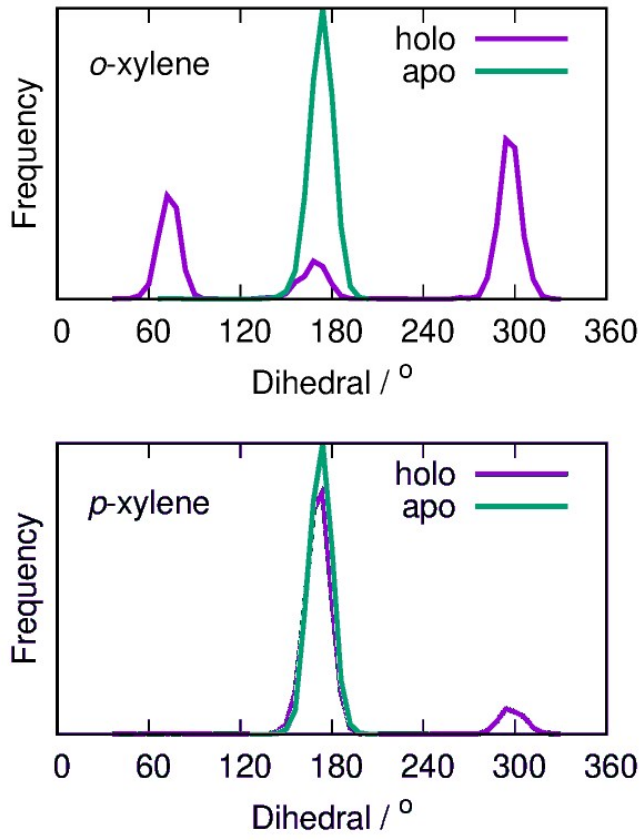
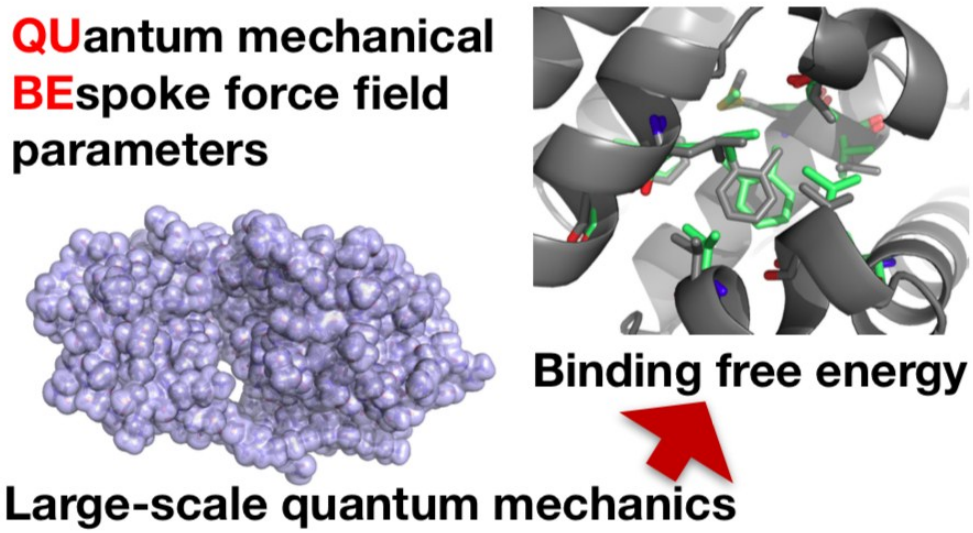


Figure 3. Sampled V111 side chain orientations in lysozyme from simulations of lysozyme-*o*-xylene and lysozyme-*p*-xylene. Experimentally, the apo protein side chain has a dihedral angle of 180° and the holo has a dihedral angle close to 300° for both *o*-xylene and *p*-xylene.

View Article Online
DOI: 10.1039/C9MD00017H



320x175mm (72 x 72 DPI)