Published in final edited form as:
*J Phys Chem B*. 2009 April 9; 113(14): 4533–4537. doi:10.1021/jp806838b.

# Predictions of hydration free energies from all-atom molecular dynamics simulations

**David L. Mobley**[†,*], **Christopher I. Bayly**[‡], **Matthew D. Cooper**[‡], and **Ken A. Dill**[§]

[†] *Department of Chemistry, University of New Orleans, New Orleans, LA 70148*

[‡] *Merck-Frosst Canada Ltd., 16711 TransCanada Highway, Kirkland, Quebec, Canada H9H 3L1*

[§] *Department of Pharmaceutical Chemistry, University of California at San Francisco, San Francisco, CA 94158*

## Abstract

Here, we computed the aqueous solvation (hydration) free energies of 52 small drug-like molecules using an all-atom force field in explicit water. This differs from previous studies in that: (1) this was a blind test (in an event called SAMPL sponsored by OpenEye Software), and (2) the test compounds were considerably more challenging than have been used in the past in typical solvation tests of all-atom models. Overall, we found good correlations with experimental values which were subsequently made available, but the variances are large compared to in previous tests. We tested several different charge models, and found that several standard charge models performed relatively well. We found that hypervalent sulfur and phosphorous compounds are not well handled using current force field parameters, and suggest several other possible systematic errors. Overall, blind tests like these appear to provide significant opportunities for improving force fields and solvent models.

## 1 Introduction

Hydration free energies provide an important metric of the accuracy of physics-based methods used in molecular simulations. Since these can now be calculated very precisely, they can be compared with experiment to test force fields and identify systematic errors[1,2,3,4]. They also can provide insight into the underlying solvation effects such as hydrophobicity[5], surface effects[6], and solvent asymmetries[7]. For these and other reasons, there have been a wide range of recent computational studies of small molecule hydration free energies from explicit solvent simulations[1,8,4,9,10,11,2,12,13,14].

A major advantage of physical methods is their potential ability to predict properties of compounds that have not been previously studied. Ideally this ability could be used in drug discovery and other applications. With this in mind, it is important to test methods not only in retrospective tests, but also prospectively, as they would be used in real applications. Here we report the results of a blind test for computing hydration free energies with explicit solvent molecular dynamics simulations. This test was done as part of OpenEye's Statistical Assessment of Modeling of Proteins and Ligands (SAMPL) challenge. Hydration free energies were computed with no knowledge of experimental values, then submitted to the moderators of the SAMPL project, who then provided the experimental values.

[*] Author to whom correspondence is to be addressed, at E-mail: dmobley@gmail.com.

Supporting Information Available: A full table of all the computed hydration free energies and components, for all charge models, along with the experimental values; computed BEDROC values and uncertainties by functional group and charge model; AMBER parameter and coordinate files for the molecules in the set.

## 2 Methods

Starting mol2 files were provided by the organizers of OpenEye's SAMPL event; names and 2D structures are provided in the work of Guthrie[15]. We then prepared five partial charge sets for use with AMBER small molecule parameters – a negative control, two positive controls, and two sets for testing. The negative control was Merck Molecular Mechanics Force Field (MMFF) charges, which we expected to perform poorly[16]. Positive controls were RESP HF/6-31G* and AM1-BCC partial charges. We tested PM3-BCC v0.2 and PM3-BCC v0.3 partial charges, which are under development by C. I. Bayly and collaborators as potential successors to AM1-BCC. MMFF charges were computed using routine Merck & Co. internal software. RESP charges were computed as described previously[16], except that a B3LYP (cc-pVTZ) minimization was done on an extended conformation holding all non-H-containing dihedrals constant; the restraint weight was 0.001 in both stages; and for all topologically equivalent atoms, charges were averaged as the last step. And, for time reasons, geometry optimization was not entirely completed for molecule 23, though the forces were in the last significant figure before the convergence threshold. AM1-BCC charges were computed as described previously[17,16].

The approach for the free energy calculations here was very similar to that in several previous studies of hydration free energies[4,2,11]. We used explicit solvent molecular dynamics simulations with the TIP3P water model[18] and Amber GAFF[19,20] small molecule parameters. Simulations were conducted using the April 2, 2007 CVS version of the GROMACS 3.3.1 software package[21] (which incorporated several bugfixes past the 3.3.1 release itself). The hydration free energy calculations involved several components as described previously[4], with each simulation conducted independently from the same starting structure. First, solute electrostatics are turned off in water linearly with the variable $\lambda$ (where $\lambda = 0, 0.25, 0.5, 0.75, 1.0$ in turn). Second, solute-water Lennard-Jones interactions are turned off in water using soft core potentials[22] with the parameters suggested by Shirts[8] ($\alpha = 0.5$, with a soft core exponent of 1), as previously[4]. For this step, $\lambda$ values were 0.0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1.0. Finally, solute electrostatics were turned back on in vacuum (with $\lambda = 0, 0.25, 0.5, 0.75, 1.0$). The free energy of each of these component steps was computed using the Bennett acceptance ratio (BAR)[23] and then the total hydration free energy was computed as $\Delta G_{hyd} = \Delta G_{chg,vac} - \Delta G_{chg} - \Delta G_{LJ}$, where $\Delta G_{chg}$ denotes the free energy of turning off the electrostatics in water, $\Delta G_{chg,vac}$ denotes the same quantity for vacuum, and $\Delta G_{LJ}$ denotes the free energy of turning off the solute-water Lennard-Jones interactions in water.

Protocols were generally as described previously[4]. Briefly, at each lambda value, the (same) starting structure was minimized using steepest descents minimization. The resulting structures were then run through an equilibration procedure consisting of 10 ps of constant volume equilibration, followed by 100 ps of constant pressure equilibration. Production simulations were 5 ns at each $\lambda$. We did make some minor modifications to our previous protocols. We replaced the L-BFGS minimizations with up to 5,000 steps of steepest descents minimization for each molecule at each $\lambda$ value (because the GROMACS L-BFGS minimizer would often terminate too early, resulting in forces that remained too large; we achieved better minimization with the steepest descents minimizer). For simulations in water, we we used a neighbor list cutoff of 1.0 angstroms and an electrostatic cutoff of 1.0 angstroms (this change was because the 3.3.1 version of GROMACS requires these two cutoffs be equal when using lattice-sum electrostatics). Small molecules were solvated using GROMACS utilities in a dodecahedral simulation box with at least 1.2 nm from the solute to the nearest simulation box edge; in some cases previous simulations used slightly smaller box sizes. For each charge set, a seperate set of electrostatic annihilation calculations was performed, rather than computing the free energy of changing the charges from a reference set as in the previous work (this change was to avoid introducing the potential for additional error by adding an extra step to the calculation). Additionally, following constant pressure equilibration, at each lambda value, we performed an affine transformation on the atomic coordinates to scale the volume to the average box volume from the constant pressure equilibration. The box was then fixed at this size during the subsequent constant volume simulations, and an additional 100 ps of data was discarded to equilibration before collecting data for analysis. This change was made because occasionally box volumes at the end of constant pressure equilibration could be far from the mean, and fixing the box volume to this value for constant volume production could lead to artifactual densities. Adding the affine transformation ensures the box volume, and hence density, is correct for the constant volume production. Data and error analysis was as described previously[4,2,11]; computed uncertainties reported in the Supporting Information represent the estimated standard error in the mean. Nonpolar components, which do not depend on the charge model, were only calculated once.

With the data we generate, we want to be able to identify whether there are particular functional groups that tend to cause systematic errors, or whether errors are not particularly linked to functional groups. We begin with the realization that, if a functional group is not associated with systematic errors, it should be roughly as likely to occur in compounds that have large errors relative to experiment as in compounds with small errors relative to experiment. For example, a previous study found that whether a compound is aromatic or not has no bearing on whether it is well- or poorly-predicted[11]. So, to identify systematic errors, what we seek to find is chemical groups that are statistically over-represented in the compounds with the largest errors. There are many potential ways to perform such a search, and here we choose just one such way that appears to work well for us. We first sort the molecules by the absolute value of the error relative to experiment, from largest to smallest errors. We then use the package Checkmol[24] to group compounds by functional group. We then want a statistical metric to assess which are over-represented at the largest errors. We choose the BEDROC metric[25] for this task, as in one previous study[11]. Basically, BEDROC computes a Boltzmann-weighted area under the cumulative probability distribution function for finding compounds (with a particular chemical group) at a particular error, then rescales the resulting numbers to fall between 0 and 1. The weighting simply makes the early (high-error) part of the curve dominate. Here, we compute BEDROC values (with $\alpha = 1$) for different functional groups. Those functional groups which have particularly high BEDROC values (relative to random) are typically associated with large errors (and thus may have parameter or other problems), as noted previously[11]. Thus, the BEDROC values we report here are simply a numerical metric that tells us whether or not a particular functional group is especially likely to be associated with large errors relative to experiment.

Experimental results are taken from the tables of Guthrie[15], and experimental error bars shown in the plot are taken from the uncertainty estimates described in that work. Some potential sources of error are discussed there as well.

## 3 Results and Discussion

A variety of previous explicit solvent hydration free energy studies had RMS errors relative to experiment in the 0.8–1.6 kcal/mol range[4,2,11], which might have implied similar accuracies here. However, the composition of this test set is very different. Prior test sets contained mainly monofunctional molecules with relatively standard or common functional groups. They were, in some ways, unlike typical drugs because they lacked the polyfunctionality common in many drugs. On the other hand, the SAMPL set is much different, and in some respects more drug-like (many of the compounds are pesticides). Most of the SAMPL molecules are larger (16.3 heavy atoms on average, compared to 7.1 in a previous extensive test[11]), highly polyfunctional and very polar (see the discussion and structures in [15]). Also, a number of these functional groups have been rarely, if ever, studied with fixed-charge force fields. Finally, this test set takes us fairly far afield from the usual functional groups (such as amino acid sidechain analogs, nucleic acids, and common cosolvents) which have been studied in developing the GAFF parameter set[19,20]. These factors make this SAMPL test much more challenging. To illustrate the difference, representative molecules are shown in Figure 1.

Here, we tried several different charge models with the same GAFF bonded and nonbonded parameters. We expected the RESP HF/6-31G*[26] and AM1-BCC[17,16] charge models would do fairly well, and as a negative control we used MMFF charges, which we expected to perform relatively poorly, as they were developed for a different force field. We also tested two other charge models to see how they compared against these others. Statistics are shown in Table 1. We use RMS error and $R^2$, the correlation coefficient, as our metrics for quality, and also report mean error to show whether there is a systematic offset in computed values. We find that, as expected, MMFF performs worst. Both PM3-BCC charge models are intermediate in terms of RMS error, and comparable to RESP and AM1-BCC in terms of $R^2$, and RESP and AM1-BCC have the lowest RMS errors. Here, RESP has the lowest RMS error − 3.5 ± 0.2 kcal/mol – and an $R^2$ value of 0.76 ± 0.08. Except for MMFF, RMS errors fall between 3.5 and 4.1 kcal/mol, and $R^2$ values are decent, running from 0.76±0.08 to 0.83±0.09. Computed hydration free energies versus experiment are shown in Figures 2 and 3, and a full table of computed values and components is provided in the Supporting Information.

Overall, RMS errors here are markedly higher than in previous studies[4,2,11], probably reflecting the difficulty of this highly polar and polyfunctional test set, as well as its deviations from the regions of chemical space the force field has been tested in. Previous work on more typical functional groups showed that computed results for more polar compounds with more negative hydration free energies had larger errors[11]; this set has a higher proportion of highly polar compounds, which may have played a significant role in the lower accuracy here. Here, we group the compounds by functional group, sort the list by the magnitude of the error relative to experiment, and use the BEDROC metric to look for functional groups that are disproportionately associated with large errors. High BEDROC values mean a particular functional group occurs mostly in compounds with large errors, while low BEDROC values mean it occurs mostly in compounds with small errors, and intermediate values mean the functional group is distributed roughly randomly. Thus high BEDROC values may be an indication of force field errors for a particular functional group.

BEDROC values for the functional groups we examined are shown in Figure 4 by charge model. A random distribution (for these numbers of compounds) gives a BEDROC value of 0.49–0.50. Some functional groups show particularly significant deviations from random. BEDROC values for ureas, compounds with hypervalent sulfur, sulfonamides, and compounds with hypervalent phosphorous are all significantly worse than random with AM1-BCC. In contrast, nitrates are particularly well predicted. These trends are consistent across all the charge models, except that nitro-containing compounds also perform poorly in MMFF and PM3-BCC v0.2, and hypervalent phosphorous compounds are reasonably well predicted with PM3-BCC and RESP.

It seems clear that something is significantly wrong with the calculations or experiments for the compounds with hypervalent sulfur. Figure 2(a) shows the compounds with hypervalent sulfur for the AM1-BCC charge set, with those containing hypervalent sulfur and phosphorous highlighted with a different color and symbol. All of the computed values for these compounds are off from the experimental values in the same direction by several kcal/mol (mean error −8.10 +/− 0.40 kcal/mol with AM1-BCC). Hypervalent phosphorous compounds are also particularly poorly predicted with AM1-BCC and several other charge models (Figure 2(a)). Together, these two groups account for the worst outliers – if hypervalent sulfur and phosphorous compounds are excluded, the AM1-BCC RMS error is 1.6 ± 0.2 kcal/mol (down from 3.8 ± −0.2) and the $R^2$ increases to 0.9 ± 0.1 (from 0.8 ± 0.1), more in line with the accuracies seen in previous studies[2,11]. Of course, excluding outliers always makes results better, and is only possible retrospectively.

Should we have known that hypervalent sulfur and phosphorous compounds might be a problem? Our previous retrospective study[11] only had five hypervalent sulfur compounds, and we find a BEDROC value of 0.6 ± 0.2, within uncertainty of random, so the data is inconclusive. The situation was even worse for hypervalent phosphorous compounds, for which there were only two representatives.

We believe this analysis suggests a systematic problem with force field parameters for hypervalent sulfur and possibly hypervalent phosphorous compounds that was statistically insignificant in earlier work, essentially due to the small number of such compounds in the earlier test set.

What might be the problem with these parameters? While AMBER[27] and GAFF[28] use a variety of atom types for sulfur and phosphorous, the Lennard-Jones parameters for all sulfur atom types are identical. Similarly, the Lennard-Jones parameters for all phosphorous atom types are identical. This seems surprising, as the chemical environment seems likely to affect the strength of dispersion interactions between these atoms and their surroundings. Another study recently found that the Lennard-Jones parameters for triple bonded carbons had been taken from those for aromatic carbons in AMBER and GAFF and that this underestimated the attractive interactions between, for example, alkynes and water, leading to systematic errors[4]. Something similar could be going on here. Apparently the original AMBER sulfur parameters were taken from OPLS[27] (though the AMBER force field files indicate that free energy perturbation calculations also played a role), but now OPLS has moved to using different Lennard-Jones parameters for different sulfur atom types, while AMBER has maintained the single set of parameters. The (single set of) phosphorous parameters for AMBER were originally developed by Weiner *et al.*[29] and GAFF simply took this set and applied it to all the phosphorous atom types[28]. This practice of taking Lennard-Jones parameters derived for an element in one particular environment and applying it to the same element in a substantially different chemical environment, differing even in terms of the number of bonds, could be the cause of some of these systematic errors.

RMS errors in the range of those reported here (3.5 kcal/mol and up) are large for practical applications. For example, a 3.5 kcal/mol error in a binding free energy calculation would be larger than the range in binding affinities in many lead series! In some sense, the compounds tested here are ”drug-like”, so this is at first a discouraging result. But as noted, RMS errors are much better (and more in line with previous studies) if the hypervalent sulfur and phosphorous compounds are excluded, so the poor accuracy seen here may be simply pointing the way towards the need for refinements of the force field for these particular compounds.

We are not aware of any force fields or methods that would be expected to give better results for this set. Other participants in the SAMPL challenge seemed to achieve at best comparable results[15]. The same held true in another recent prospective test[2], where the best Poisson-Boltzmann based approach gave accuracies no better than molecular dynamics free energy calculations. Even a later retrospective study using a quantum mechanical continuum solvation model gave accuracies that were roughly comparable (RMS errors between 1.08 and 1.88 kcal/mol[30], versus RMS errors between 1.33 and 2.0 kcal/mol with explicit solvent in the previous study[2]). So, from a methods point of view, there is no reason to expect that other methods should perform any better on this set. However, if the dominant source of error here is indeed a parameter problem for a small subset of the compounds, it might suggest that a method more grounded in quantum mechanics might do substantially better here.

## 4 Conclusions

Opportunities for prospective or blind tests of computational free energy methods have been relatively rare. This represents the second such test for calculations of hydration free energies[2]. These tests are helpful, as they provide a way to avoid any possibility of being influenced by knowledge of the "right answer" and thus to genuinely test the method with no adjustments to parameters.

Overall, this prospective test has provided an opportunity to test explicit solvent simulations in a region of chemical space in which they have been rarely applied. Stepping into the "wilderness" in this way appears to present some risks, as errors were larger here than in previous studies considering simpler, often mono-functional, compounds that were more similar to typical protein or nucleic acid components. It was encouraging that correlations with experimental values remained fairly strong ($R^2$ of 0.75 and higher, except for our negative control charge model), though errors were relatively high. Some functional groups were particularly poorly predicted, suggesting that further force field development for these functional groups may improve accuracies. We believe that regular studies of this nature will provide substantial benefits for the development of solvation models and force fields, and will aid in identifying systematic errors with force fields and making improvements.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

We are grateful to J. Peter Guthrie (University of Western Ontario) for curating and providing the experimental data, and we appreciate OpenEye Software for organizing SAMPL; particular thanks go to Geoff Skillman and Anthony Nicholls (OpenEye). We appreciate the support of NIH grant GM 63592 to KAD.

## References

1. Shirts MR, Pitera JW, Swope WC, Pande VS. J Chem Phys 2003;119:5740–5761.
2. Nicholls A, Mobley DL, Guthrie JP, Chodera JD, Pande VS. J Med Chem 2008;51:769–778. [PubMed: 18215013]
3. Mobley DL, Chodera JD, Dill KA. J Phys Chem B 2008;112:938–946. [PubMed: 18171044]
4. Mobley DL, Dumont È, Chodera JD, Dill KA. J Phys Chem B 2007;111:2242–2254. [PubMed: 17291029]
5. Ashbaugh HS, Kaler EW, Paulaitis ME. J Am Chem Soc 1999;121:9243–9244.
6. Chorny I, Dill KA, Jacobson MP. J Phys Chem B 2005;109:24056–24060. [PubMed: 16375397]
7. Mobley DL, Barber AE II, Fennell CJ, Dill KA. J Phys Chem B 2008;112:2405–2414. [PubMed: 18251538]
8. Shirts MR, Pande VS. J Chem Phys 2005;122:134508. [PubMed: 15847482]
9. Hess B, van der Vegt NFA. J Phys Chem B 2006;110:17616–17626. [PubMed: 16942107]
10. Deng Y, Roux B. J Chem Phys 2004;108:16567–16576.
11. Mobley DL, Bayly CI, Cooper MD, Shirts MR, Dill KA. 2008submitted
12. Villa A, Mark AE. J Comp Chem 2002;23:548–553. [PubMed: 11948581]
13. Xu Z, Luo HH, Tieleman DP. J Comp Chem 2007;28:689–697. [PubMed: 17195160]
14. Maccallum JL, Tieleman DP. J Comp Chem 2003;24:1930 – 5. [PubMed: 14515375]
15. Guthrie JP. J Phys Chem B. 2009accepted

16. Jakalian A, Jack DB, Bayly CI. J Comput Chem 2002;23:1623–1641. [PubMed: 12395429]

17. Jakalian A, Bush BL, Jack DB, Bayly CI. J Comput Chem 2000;21:132–146.

18. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. J Chem Phys 1983;79:926–935.

19. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. J Comput Chem 2004;25:1157–1174. [PubMed: 15116359]

20. Wang J, Wang W, Kollman PA, Case DA. J of Mol Graphics Modell 2006;26:247260.

21. van der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC. J Comput Chem 2005;26:1701–1718. [PubMed: 16211538]

22. Beutler TC, Mark AE, van Schaik RC, Gerber PR, van Gunsteren WF. Chem Phys Lett 1994;222:529–539.

23. Bennett CH. J Comp Phys 1976;22:245–268.

24. Haider, N. Checkmol. http://merian.pch.univie.ac.at/nhaider/cheminf/cmmm.html

25. Truchon JF, Bayly C. J Chem Inf Model 2007;47:488–508. [PubMed: 17288412]

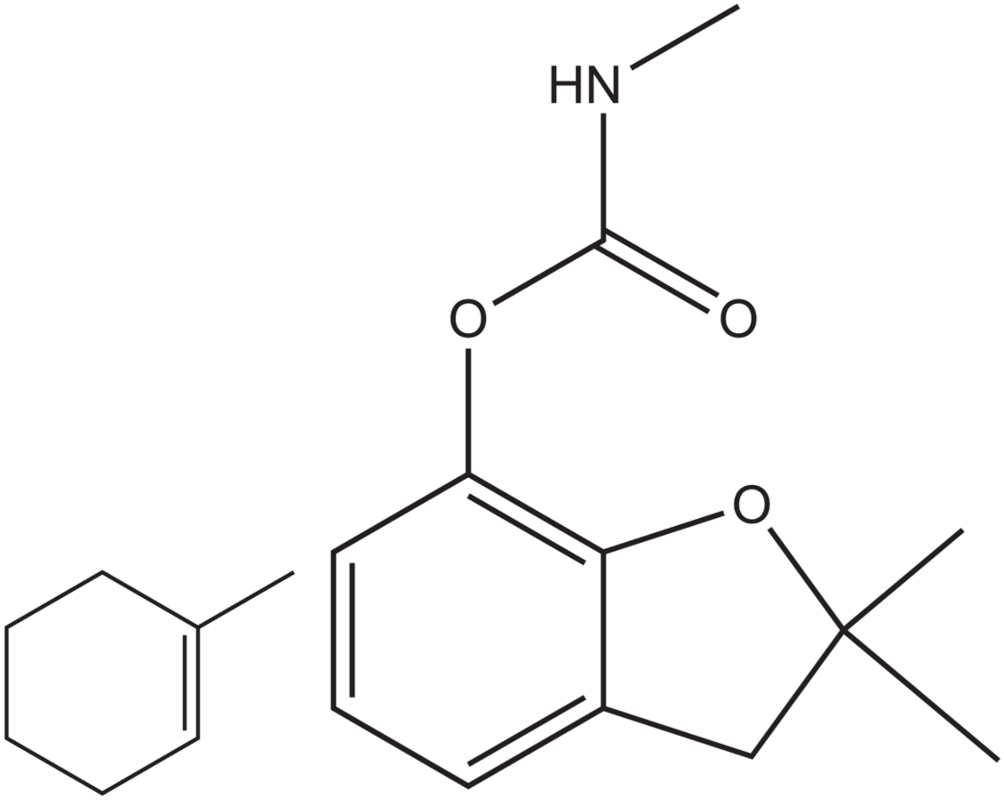
26. Bayly CI, Cieplak P, Cornell WD, Kollman PA. J Phys Chem 1993;97:10269–10280.

27. Cornell W, Cieplak P, Bayly CI, Gould IRKMM Jr, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. J Am Chem Soc 1995;117:5179–5197.

28. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. J Comput Chem 2004;25:1157–1174. [PubMed: 15116359]

29. Weiner SJ, Kollman PA, Nguyen DT, Case DA. J Comp Chem 1986;7:230–252.

30. Chamberlin AC, Cramer CJ, Truhlar DG. J Phys Chem B 2008;112:8651–8655. [PubMed: 18582013]

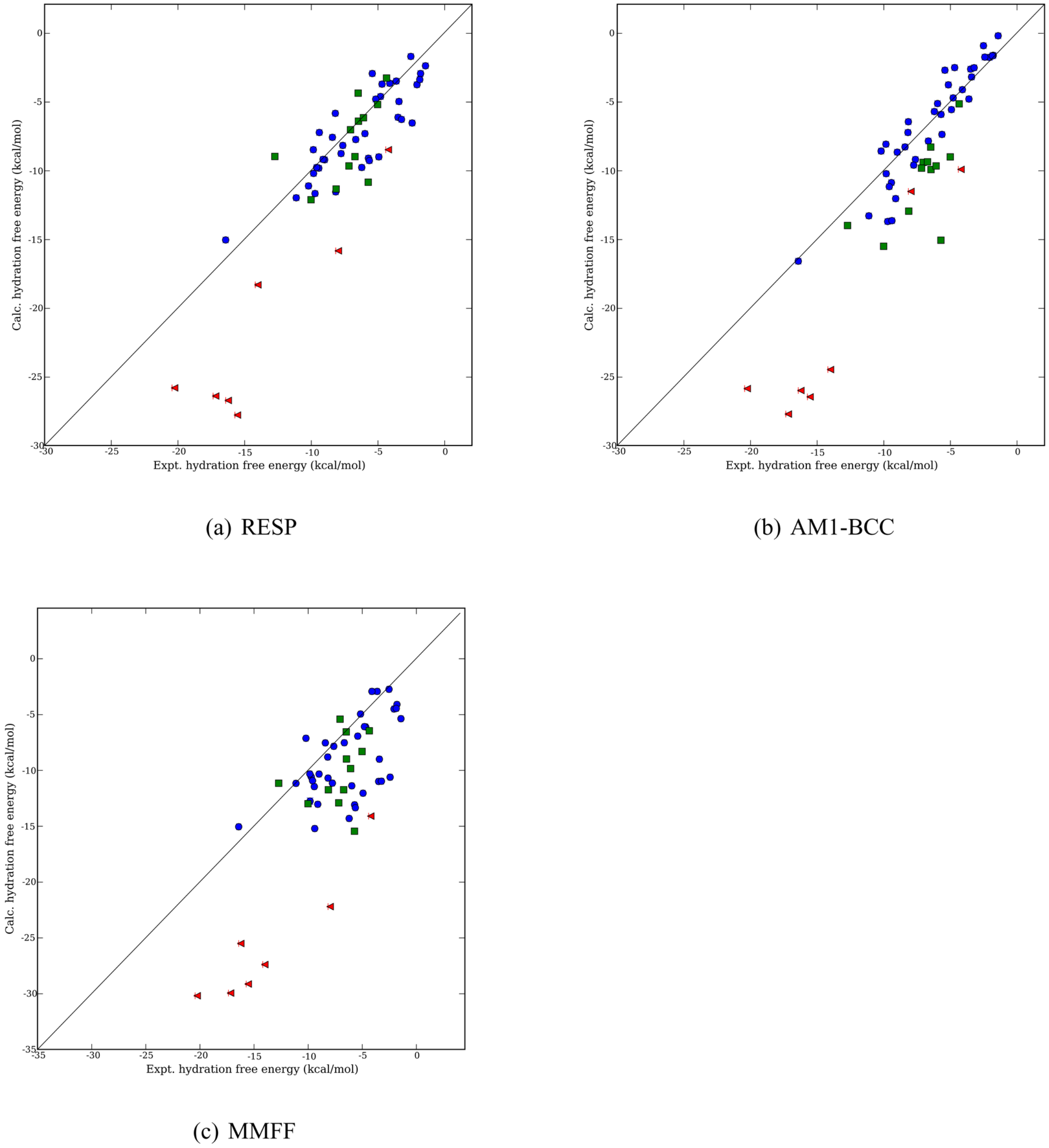


(a) 2-methylcyclohexene

(b) 2,2-dimethyl-2,3-dihyrdobenzofuran-7-yl methylcarbamate

**Figure 1. Representative molecules from the test sets**

Shown are reference molecules with the typical number of heavy atoms in the previous (a) and this (b) test sets. 1-methylcyclohexene is shown in (a) and has 7 heavy atoms; 2,2-dimethyl-2,3-dihydrobenzofuran-7-yl methylcarbamate (also known as carbofuran) is shown in (b) and has 16 heavy atoms.



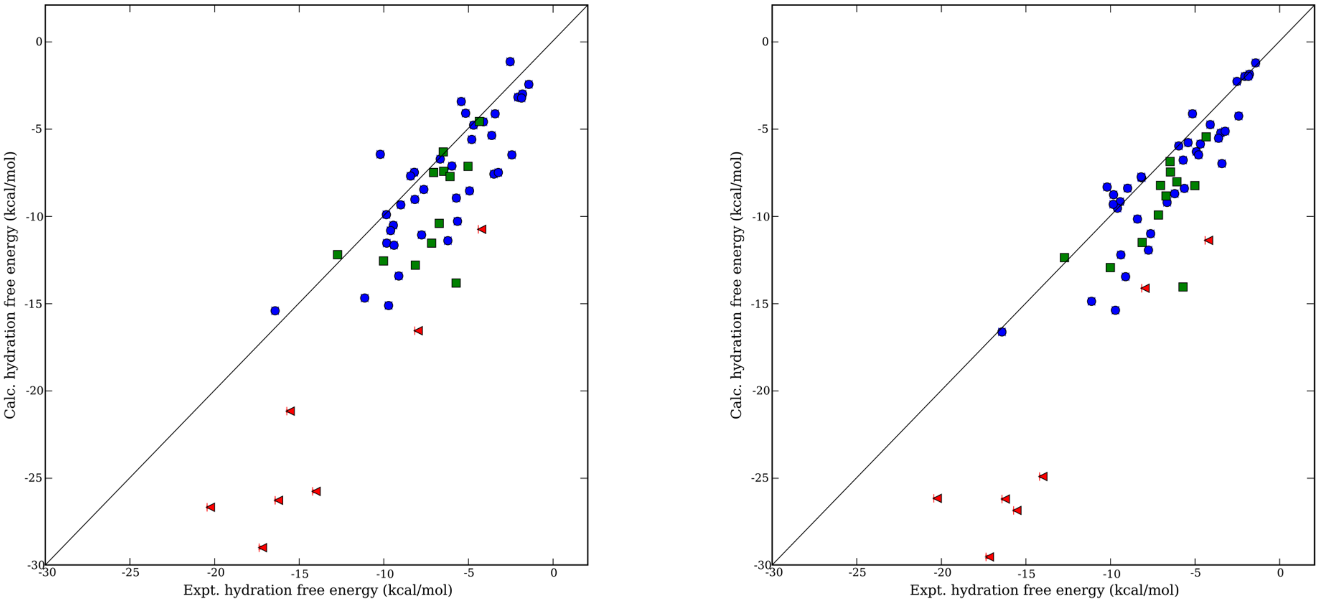
(a) RESP

(b) AM1-BCC

(c) MMFF

**Figure 2. Computed versus experimental hydration free energies for the charge models studied**

Shown are computed hydration free energies versus experiment, for partial charges from (a) RESP HF/6-31G*; (b) AM1-BCC; and (c) MMFF. Red triangles denote compounds containing hypervalent sulfur; green squares denote those containing hypervalent phosphorous, and blue circles denote the remainder of the compounds.



(a) PM3-BCC v0.2

(b) PM3-BCC v0.3

**Figure 3. Computed versus experimental hydration free energies for the PM3-BCC charge models**

Shown are computed hydration free energies versus experiment, for partial charges from (a) PM3-BCC v0.2 and (b) PM3-BCC v0.3. Red triangles denote compounds containing hypervalent sulfur; green squares denote those containing hypervalent phosphorous, and blue circles denote the remainder of the compounds.

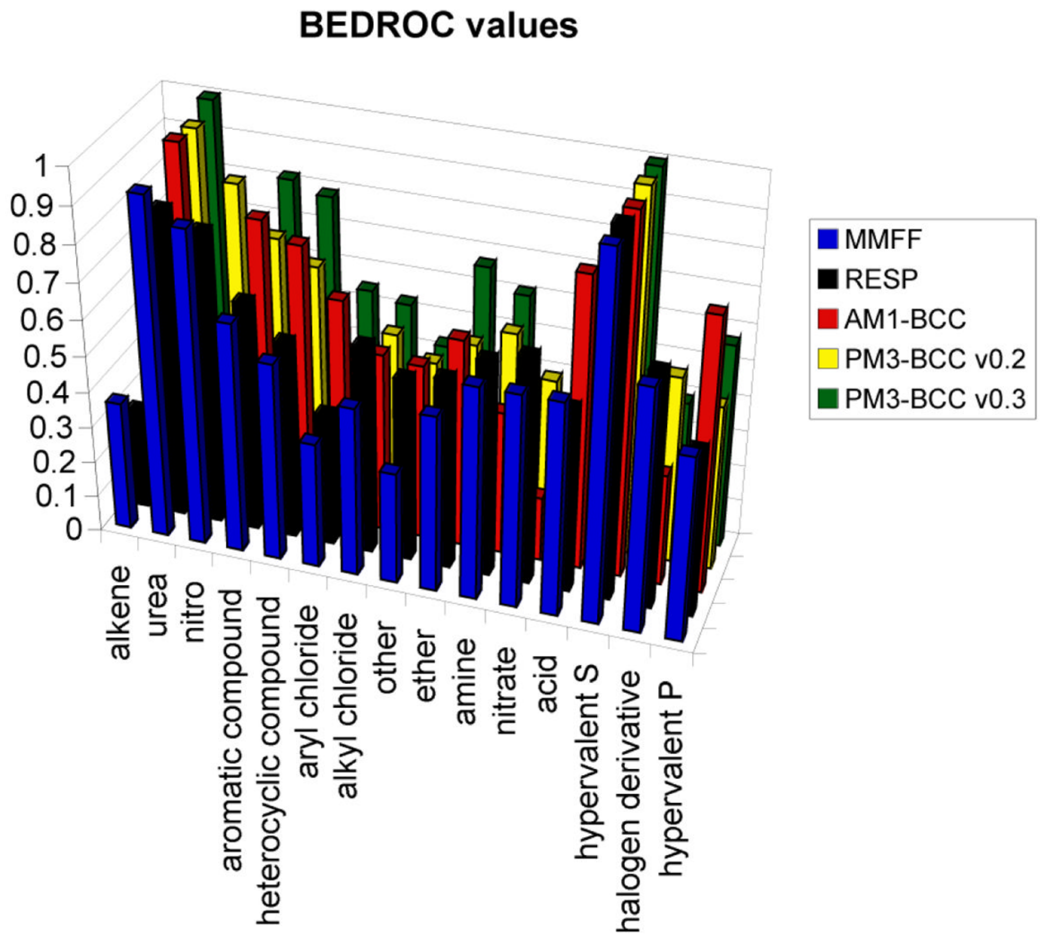


**Figure 4. BEDROC values by functional group and charge model**

Shown are computed BEDROC values ($\alpha = 1.0$) for different functional groups represented in the test set, for all of the charge models examined. A table of values, including uncertainties, is shown in the Supporting Information.

**Table 1**

Statistics for the charge models tested in this study. Shown are RMS error, correlation coefficient ($R^2$), and mean error.

| Charge model | RMS error (kcal/mol) | $R^2$ (kcal/mol) | Mean error (kcal/mol) |
|---|---|---|---|
| RESP | 3.51 ± 0.20 | 0.76 ± 0.08 | −1.68 +/− 0.42 |
| AM1-BCC | 3.82 ± 0.21 | 0.83 ± 0.09 | −1.88 ± 0.45 |
| MMFF | 5.75 ± 0.20 | 0.60 ± 0.08 | −3.92 ± 0.57 |
| PM3BCC v0.2 | 4.13 ± 0.22 | 0.76 ± 0.09 | −2.57 ± 0.44 |
| PM3BCC v0.3 | 4.05 ± 0.21 | 0.80 ± 0.09 | −2.47 +/− 0.43 |