Supporting Information
Approaches for calculating solvation free energies and enthalpies demonstrated with an update of the FreeSolv database

Guilherme Duarte Ramos Matos,[1] Daisy Y. Kyu,[2] Hannes H. Loeffler,[3]
John D. Chodera,[4] Michael R. Shirts,[5] and David L. Mobley[6]

[1] *Department of Chemistry, University of California, Irvine*
[2] *Department of Pharmaceutical Sciences, University of California, Irvine*
[3] *Scientific Computing Department, STFC Daresbury, Warrington, WA4 4AD, United Kingdom*
[4] *Computational and Systems Biology Program, Sloan Kettering Institute*
[5] *Department of Chemical and Biological Engineering, University of Colorado Boulder, Boulder*
[6] *Departments of Pharmaceutical Sciences and Chemistry, University of California, Irvine**

I. FREESOLV HAS HYDRATION FREE ENERGIES FOR NEUTRAL COMPOUNDS

FreeSolv focuses on hydration free energies of neutral compounds. While many studies have computed hydration free energies for charged species, measuring hydration free energies for charged species in isolation is impossible, so extracting these can require extrathermodynamic assumptions or introduce other complexities. Thus, we agree with previous work suggesting that the main focus should be on hydration free energies of neutral compounds[1] (see particularly footnote 61), as also discussed elsewhere[2].

It is worth noting, however, that the database does contain a variety of carboxylic acids. In solution, these are typically charged at neutral pH. However, hydration free energies are typically reported for the neutral form of the molecule[2] so those are the values used here.

II. ADDITIONAL PRACTICAL CONSIDERATIONS FOR CALCULATION OF SOLVATION FREE ENERGIES

One of the appeals of hydration free energies is that they could be relatively free of the protonation state and tautomer issues which can challenge predictions of protein-ligand binding; however, this seems unlikely to be true in general (though it may be true for many of the relatively small, fragment-like compounds in FreeSolv). Particularly, small molecules can certainly have multiple relevant tautomers in solution, tautomers which change on transfer between environments (such as gas to water transfer), or tautomers which are uncertain yet important for solvation and transfer properties. While these issues may not play a major role in solvation of the present compounds, they certainly can become a factor elsewhere, as was amply illustrated in the recent SAMPL5 challenge, which focused on calculation of cyclohexane-water distribution coefficients. Many participants estimated these from solvation free energies in both solutes, and protonation and tautomer issues played an important role[3].

It is also worth noting one important issue that can affect interpretation of literature solvation free energies – these can use different standard states. Values reported in FreeSolv are for transfer free energies from gas (at a 1 M standard state) to solution (at a 1 M standard state). It is also possible to report and/or calculate values for transfer from an alternate 1 atm standard state in gas to a 1M standard state in solution[1], resulting in values which differ by an additive constant relating to the difference in gas phase standard state. Thus, care must be taken when pulling values from the literature in order to ensure a consistent standard state is used.

III. REBUILDING THE FREESOLV DATABASE

All input files deposited in FreeSolv were re-generated using the `rebuild_freesolv.py` script deposited on our GitHub repository at `github.com/mobleylab/FreeSolv`. To rebuild the input files, one can simply run this script, which requires the Chodera lab's 'openmoltools' package and the Mobley Lab's 'SolvationToolkit', both of which are conda installable from the omnia channel, and are also available on GitHub at `github.com/choderalab/openmoltools` and `github.com/mobleylab/solvationtoolkit` respectively. In this particular iteration of rebuilding FreeSolv and re-running the calculations, we used openmoltools version 0.6.7.

IV. ADDITIONAL PLOTS

A. Statistics

Figure S1(a) statistics:
- Kendall $\tau = 0.76 \pm 0.02$
- Pearson $R = 0.943 \pm 0.005$

Figure S1(b) statistics:
- Kendall $\tau = 0.40 \pm 0.02$
- Pearson $R = 0.60 \pm 0.03$

* dmobley@mobleylab.org

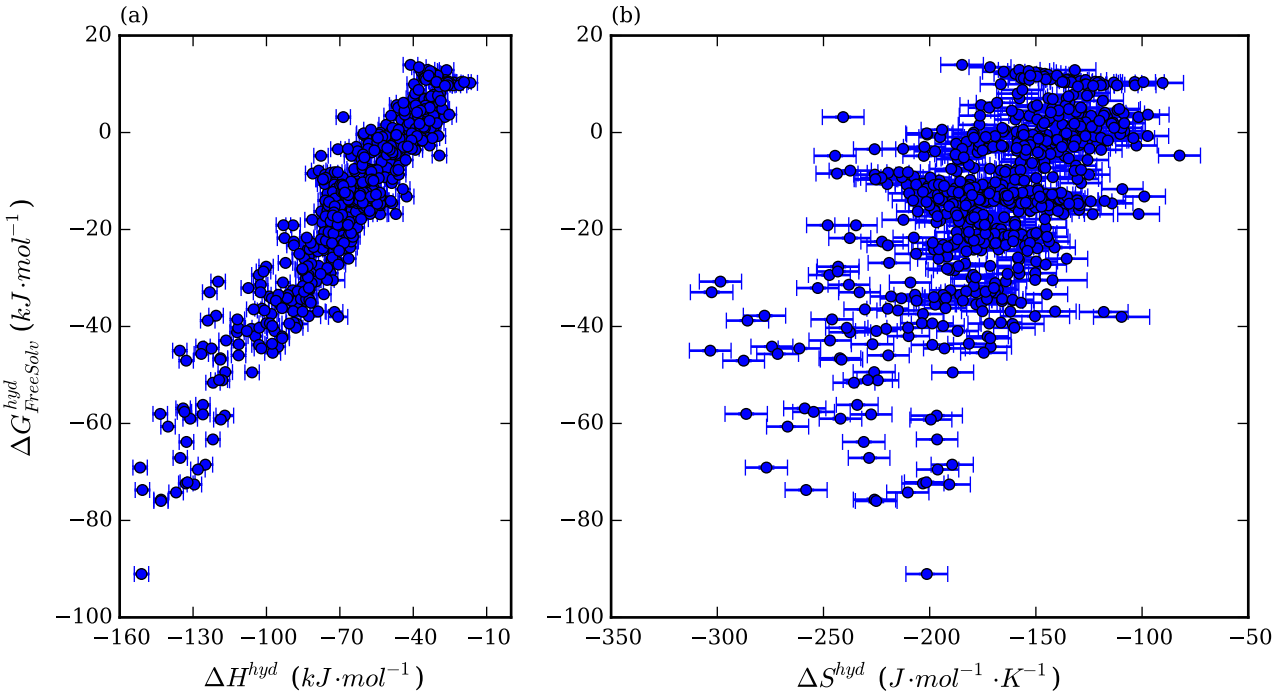

FIG. S1. Correlation plots between (a) calculated enthalpies and hydration free energies in FreeSolv, and (b) calculated entropies and hydration free energies in FreeSolv. Error bars are given as standard errors in the mean.

Figure S2(a) statistics:

- Average error $= -5 \pm 2\, kJ \cdot mol^{-1}$
- RMS $= 9 \pm 2\, kJ \cdot mol^{-1}$
- Average unsigned error $= 7 \pm 2\, kJ \cdot mol^{-1} \cdot K^{-1}$
- Kendall $\tau = 0.7 \pm 0.2$
- Pearson $R = 0.88 \pm 0.08$

Figure S2(b) statistics:

- Average error $= 2.3 \pm 0.8\, kJ \cdot mol^{-1}$
- RMS $= 3.2 \pm 0.9\, kJ \cdot mol^{-1}$
- Average unsigned error $= 2.3 \pm 0.8\, kJ \cdot mol^{-1} \cdot K^{-1}$
- Kendall $\tau = 0.85 \pm 0.2$
- Pearson $R = 0.96 \pm 0.05$

V. SIMULATION DETAILS

The following are GROMACS 4.6.7 simulation input parameters, as are the MDP files with full details which are deposited in the Supporting Information and on GitHub.

General information

- Friction coefficient $= \text{mass}_{particle}/\tau_t$, $\tau_t = 2.0\, ps$.
- Parrinello-Rahman barostat: $\tau_p = 10\, ps$ and compressibility $= 4.5 \cdot 10^{-5}\, bar^{-1}$.

Electrostatics

- PME cut-off: 1.2 nm.
- PME order: 6
- Fourier spacing = 0.10 nm
- additional details can be found in the MDP files deposited with this paper and on GitHub at `github.com/mobleylab/freesolv`.

vdW interactions

- Cut-off: $1.0\, nm$
- Switch at $0.9\, nm$
- DispCorr = AllEnerPres

Free Energy calculation control parameters

- vdW lambda schedule: 0.0, 0.0, 0.0, 0.0, 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
- FEP lambda schedule (all non-specified lambdas use this schedule): 0.0, 0.25, 0.5, 0.75, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0
- soft-core $\alpha = 0.5$
- soft-core power (m in Equation 8) $m = 1$
- additional details can be found in the MDP files deposited with this paper and on GitHub at `github.com/mobleylab/freesolv`.

All input files were generated (as noted above) via the `rebuild_freesolv.py` script deposited in the FreeSolv GitHub repository. This relies on `openmoltools`; we used version 0.6.7. As noted in the main body of the text, AM1-BCC charges were assigned with OpenEye's quacpac python module; we used `openmoltools` to drive this process. Specific source code used for charging is available at

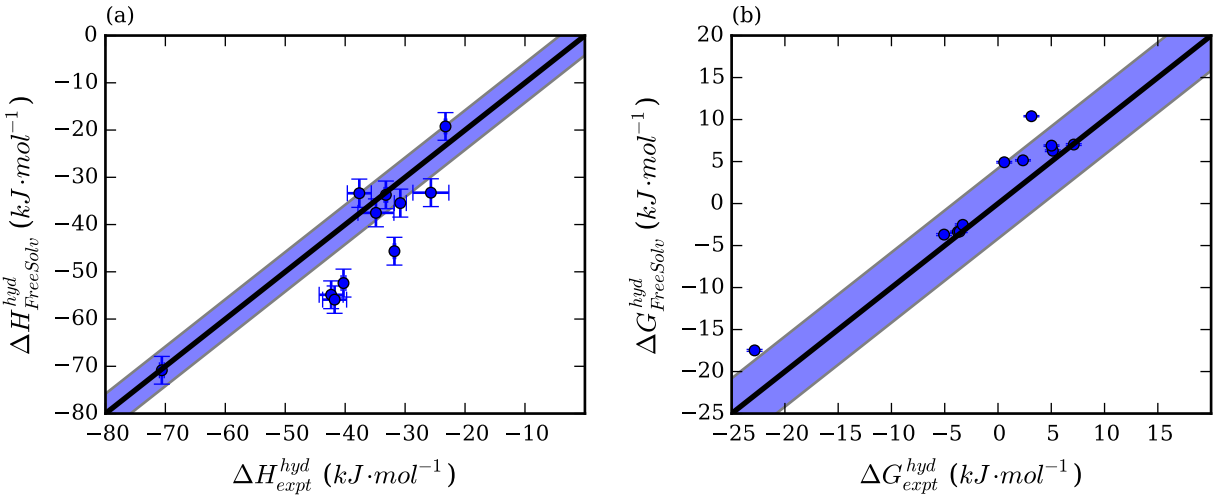

FIG. S2. Correlation plots between (a) the 11 calculated enthalpies in FreeSolv, and their corresponding experimental values from ORCHYD, and (b) calculated hydration free energies for these same 11 compounds, and their corresponding experimental values. The shaded area indicates values within 4 kJ/mol of the $x = y$ line.

TABLE S1. The 11 FreeSolv compounds with known experimental enthalpies from ORCHYD. All values are in kJ·mol^{-1}. Sources of experimental data are given in the FreeSolv database itself.

FreeSolv key	**CID**	**SMILES**	$\Delta G^{hyd}_{FreeSolv}$	ΔG^{hyd}_{expt}	$\Delta H^{hyd}_{FreeSolv}$	ΔH^{hyd}_{expt}
mobley_2689721	8078	C1CCCCC1	6.3 ± 0.1	5 ± 3	-34 ± 3	-33.2 ± 0.3
mobley_2784376	6351	C1CC1	10.40 ± 0.07	3 ± 3	-19 ± 3	-23.3 ± 0.2
mobley_3053621	241	c1ccccc1	-3.4 ± 0.1	-3.8 ± 0.8	-46 ± 3	-31.8 ± 0.2
mobley_3183805	7247	Cc1ccc(c(c1)C)C	-3.3 ± 0.1	-4 ± 3	-55 ± 3	-42 ± 2
mobley_3211679	8079	C1CCC=CC1	4.9 ± 0.1	0.6 ± 0.4	-38 ± 3	-35 ± 3
mobley_3452749	10686	Cc1cccc(c1C)C	-3.7± 0.1	-5 ± 3	-56 ± 3	-42 ± 2
mobley_7010316	7966	C1CCC(CC1)O	-17.5 ± 0.1	-23 ± 3	-71 ± 3	-70.6 ± 0.4
mobley_8006582	9253	C1CCCC1	6.90 ± 0.09	5 ± 3	-35 ± 3	-31 ± 1
mobley_8127829	7500	CCc1ccccc1	-2.5 ± 0.1	-3 ± 3	-52 ± 3	-40.3 ± 0.4
mobley_8885088	8882	C1CC=CC1	5.15 ± 0.08	2 ± 3	-33 ± 3	-26 ± 3
mobley_9100956	7962	CC1CCCCC1	7.0 ± 0.1	7 ± 3	-33 ± 3	-37 ± 2

`https://github.com/choderalab/openmoltools/blob/v0.6.7/openmoltools/openeye.py#L13` and generates molecular conformations prior to charging, as was recommended at `http://docs.eyesopen.com/toolkits/cookbook/python/modeling/am1-bcc.html`. We have found this procedure considerably more robust than the Antechamer AM1-BCC procedure used in earlier versions of the database, in part because it removes the dependence of charges on the input conformation.

For solvated systems, all solutes were placed in cubic boxes with at least 1.5 nm from the solute to the nearest box edge, and then solvated with TIP3P water using the gromacs tool `genbox`, so the number of water molecules used varied depending on the solute (but can be obtained from the topology and coordinate files deposited in the database).

VI. ABSOLUTE DIFFERENCES BETWEEN OLD AND NEW FREESOLV ΔG^{hyd} VALUES

Table S2 shows the largest differences between the calculated values previously deposited in FreeSolv and those shown here. For *most* compounds in the set, differences are relatively modest, but for this particular group some of the changes are quite significant. Some of these compounds are carboxylic acids in their neutral form, which can suffer from slow sampling of the orientation of the hydroxyl proton[4] so that may be one possible explanation for some of the discrepancies.

It is possible that other discrepancies could result from parameter differences, though we have not been able to identify any clear origins of differences. Lennard-Jones parameters seem to be identical between the (potentially different) GAFF versions used in these setups, though potentially there could be differences in bonded param-

TABLE S2. Fifteen biggest differences between old and new ΔG^{hyd} values, in kJ·mol^{-1}.

FreeSolv key	**name**	**old** ΔG^{hyd}	**new** ΔG^{hyd}	$\Delta\Delta G^{hyd}$
mobley_2099370	ketoprofen	-49.82	-72.19	22.37
mobley_1527293	flurbiprofen	-36.43	-58.42	21.99
mobley_820789	butyric acid	-22.86	-39.50	16.64
mobley_2078467	ibuprofen	-28.93	-45.46	16.53
mobley_2850833	2-hydroxybenzaldehyde	-20.47	-36.88	16.41
mobley_4792268	pentanoic acid	-22.48	-37.90	15.42
mobley_2929847	3-methylbutanoic acid	-23.07	-37.03	13.96
mobley_1735893	hexanoic acid	-21.27	-32.98	11.71
mobley_7758918	propionic acid	-26.84	-38.05	11.21
mobley_8207196	simazine	-36.13	-45.69	9.56
mobley_2913224	acetylsalicylic acid	-47.10	-39.35	7.75
mobley_8916409	malathion	-54.39	-46.88	7.52
mobley_1821184	3-methyl-1H-indole	-27.42	-34.17	6.74
mobley_7690440	methyldisulfanylmethane	6.20	-0.39	6.59
mobley_1792062	1,2-dibromoethane	0.80	-5.34	6.13

eters (because of differences in how input files were generated between when the database was originally constructed and now, GROMACS topologies use different function types for these parameters so equivalent parameters will not appear identical). However, a more likely origin of discrepancies is partial charges, as charging procedures for the studies originally used in constructing FreeSolv in some cases used Antechamber's AM1-BCC charging procedure on a database conformation of the molecule, rather than our current, more modern charging procedure which uses reasonable molecular conformations before assigning AM1-BCC charges with the OpenEye toolkits. However, we have not yet verified whether these issues can definitively be linked to the charging procedure.

Another possibility is simply protocol differences and differences in software versions. For example, some of our early work used constant volume simulations for our free energy calculations (after equilibration at constant pressure) which we later found could, in some cases, introduce additional noise to calculated hydration free energies due to artifactual densities at some λ values[5].

DISCLOSURE STATEMENT

DLM is a member of the Scientific Advisory Board for OpenEye Scientific Software. JDC is a member of the Scientific Advisory Board for Schrödinger, LLC.

[1] Mobley, D. L.; Dill, K.; Chodera, J. D. Treating Entropy and Conformational Changes in Implicit Solvent Simulations of Small Molecules. *J Phys Chem B* **2008**, *112*, 938–946.

[2] Mobley, D. L.; Guthrie, J. P. FreeSolv: A Database of Experimental and Calculated Hydration free Energies, with Input Files. *J. Comput.-Aided Mol. Des.* **2014**, *28*, 711–720.

[3] Bannan, C. C.; Burley, K. H.; Chiu, M.; Shirts, M. R.; Gilson, M. K.; Mobley, D. L. Blind Prediction of Cyclohexane-water Distribution Coefficients from the SAMPL5 Challenge. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 927–944.

[4] Klimovich, P. V.; Mobley, D. L. Predicting Hydration Free Energies Using All-Atom Molecular Dynamics Simulations and Multiple Starting Conformations. *J Comput Aided Mol Des* **2010**, *24*, 307–316.

[5] Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Dill, K. A. Predictions of Hydration Free Energies from All-Atom Molecular Dynamics Simulations. *J Phys Chem B* **2009**, *113*, 4533–4537.