Correspondence: dmobley@uci.edu (David Mobley)

Abstract:

This work provides a curated database of experimental and calculated hydration free energies for small molecules in water, along with experimental values and input files. Experimental values are taken from prior literature and will continue to be curated, with updated experimental references and data added as it becomes available. Calculated values are based on the GAFF small molecule force field in TIP3P water with AM1-BCC charges, as in the provided parameter files. Values were calculated using the GROMACS simulation package, with full details given in references cited within the database itself. This database builds on previous work from the Mobley lab and others, and extends the prior database. With deposition in eScholarship, the database is now versioned, allowing citation of specific versions of the database, and easier updates.

Background:

This page provides an update of David Mobley's hydration free energy database. The current goal is to provide curated calculated and experimental values for every molecule which the Mobley group has studied at any point, and to allow these to be updated in a versioned manner as issues are found, better experimental data is tracked down or obtained, and so on.

The prior database gives calculated and experimental values for a 504 molecule set which has been called the "504 molecule set", or "the Mobley set" or similar variants. The explicit solvent study on this set was published in (1) and the implicit solvent version in (2), and the full database is in the supporting information. The "504 molecule set" built on earlier sets, notably that from Rizzo (3) and earlier hydration studies by David Mobley and collaborators.

The current set and format is motivated by several factors:

- There were several problems with specific molecules and/or experimental values in the 504 molecule set which needed correcting
- We have studied many additional molecules since then and these need adding to the set
- We need a way to continue sharing and expanding our set, providing *both* experimental data with references and calculated values (with parameters) as these are used as inputs to test other methods
- We want to be able to update the set in a versioned manner without having to write a new paper for every update, which necessitates migrating away from journal supporting information.

What we provide:

The database consists of a .tar.gz file containing:

- database.txt: A semicolon delimited text file containing compound IDs, SMILES, IUPAC

names or similar, experimental values and uncertainties, calculated values, DOIs for references, and notes. Format described in the header
- database.pickle: Python pickle file containing the same database, with some extra fields as well (notably, 'groups', which provides functional groups for the compounds as assigned by checkmol)
- groups.txt: Functional groups for compounds as assigned by checkmol. Semicolon delimited. First field is compound ID, second field is compound name, and subsequent fields are functional groups.
- iupac_to_cid.pickle, smiles_to_cid.pickle: Python pickle files containing conversion of IUPAC name to compound id and SMILES string to compound id, stored in dictionaries
- Structure files:
    - mol2files_sybyl: Mol2 files with partial charges as written by OEChem in Sybyl format/Sybyl atom types
    - mol2files_gaff: Mol2 files with partial charges as used for our hydration free energy calculations (AMBER GAFF atom types)
    - sdffiles: sdf files with partial charges as written by OEChem
    - topgro: GROMACS format topology and coordinate files as used for our AM1-BCC GAFF hydration free energy calculations. Technical note: There may be some variation as to whether water molecules are or are not included in these files; these are intended to be used for the small molecule parameters only.

The future:

We plan to write a paper reporting the work done to create this database. This paper will provide a static version of the database, and reference a cite-able website for future updates to the database. Once this website is online, this note will be redirected to point to the website, which will contain similar information to what is here.

In the meantime, if you write a paper referencing this database we ask that you contact us for appropriate citation information.

Change log/version history:

This dataset started by taking all of the compounds we have studied previously with hydration free energies (references 1, 2, 4-9) including those from SAMPL4 and compiling them all into one big set, removing any redundancies and providing data, references, etc. for all of them. Details of changes for specific versions are found below.

Version 0.1:

- We corrected the following problems from the 504 molecule set (1-2):
    - Removal of 504/triacetyl glycerol, which was not the intended molecule (and the intended molecule, glycerol triacetate, is present in v0.1 anyway as it comes in via reference (5)
    - Correction of the experimental value for hexafluoropropene, which had (via (3)) incorrectly been the value for hexafluoro-propan-2-ol
    - Removed several duplicates within the set:
        - 2-methylbut-2-ene under two names

        - 3-methylbut-1-ene
        - benzonitrile vs cyanobenzene
    - Removed a "duplicate" butanal which had an incorrect experimental value
- We also corrected issues from other sets:
    - The molecule labeled penta-2-one in the set of (4) was penta-3-one; the corresponding experimental value was corrected from -3.52 kcal/mol to -3.41 kcal/mol.
    - The molecule "lindane" was removed from the set of reference (6) because the 3D structure has the incorrect stereoisomer and thus the calculations were wrong; this issue seems to have originated with the Guthrie 2009 experimental paper providing the source data.
    - We removed 'prometryn' (set of reference 6) because chemical structure (3D/2D) does not match the name -- an ethyl where there should be a dimethyl. Again this seems to have originated from Guthrie 2009 experimental paper with the source data.
    - We removed 'ethylene glycol diacetate' from the set of reference (5) because the 3D structure does not match the 2D structure as indicated in the paper
    - Sulfonyl urea compounds with questionable vapor pressure were removed from the set of SAMPL1 (6) after consultation with J. Peter Guthrie, who had concerns about the quality of this data.
- Based on a cross-comparison with data from J. Peter Guthrie's dataset (in preparation), we updated several experimental values. Details of how these were changed and why are provided in the 'notes' field within the database itself. The compounds affected were:
    - 4-propylphenol
    - 4-bromophenol
    - 3-hydroxybenzaldehyde
    - 2-methoxyethanol
    - dimethyl sulfoxide (methanesulfinylmethane)
- Notes were added in a few other cases, especially for formaldehyde, and a number of IUPAC names were standardized

Currently this set contains 642 molecules. Full details will be provided in a paper reporting this database.
Please also note that some discrepancies between experimental values here and values in J. Peter Guthrie's database are still being investigated, so we expect that a new version will be released relatively shortly which will update some subset of the experimental values (less than 60, but more than zero).

References:

(1) Mobley, D. L., Bayly, C. I., Cooper, M. D., Shirts, M. R., & Dill, K. A. (2009). Small Molecule Hydration Free Energies in Explicit Solvent: An Extensive Test of Fixed-Charge Atomistic Simulations. *Journal of Chemical Theory and Computation*, *5*(2), 350–358.
(2) Mobley, D. L., Dill, K., & Chodera, J. D. (2008). Treating entropy and conformational changes in implicit solvent simulations of small molecules. *The Journal of Physical Chemistry B*, *112*(3), 938.
(3) Rizzo, R. C., Aynechi, T., Case, D. A., & Kuntz, I. D. (2006). Estimation of Absolute Free Energies of Hydration Using Continuum Methods: Accuracy of Partial Charge Models and Optimization of Nonpolar Contributions. *Journal of Chemical Theory and Computation*, *2*(1), 128–139. doi:10.1021/ct050097l
(4) Mobley, D. L., Dumont, É., Chodera, J. D., & Dill, K. (2007). Comparison of charge models for fixed-

charge force fields: Small-molecule hydration free energies in explicit solvent. *The Journal of Physical Chemistry B*, *111*(9), 2242–2254.
(5) Nicholls, A., Mobley, D. L., Guthrie, J. P., Chodera, J. D., Bayly, C. I., Cooper, M. D., & Pande, V. S. (2008). Predicting small-molecule solvation free energies: an informal blind test for computational chemistry. *Journal of Medicinal Chemistry*, *51*(4), 769–779. doi:10.1021/jm070549+
(6) Mobley, D. L., Bayly, C. I., Cooper, M. D., & Dill, K. A. (2009). Predictions of hydration free energies from all-atom molecular dynamics simulations. *The Journal of Physical Chemistry B*, *113*(14), 4533–4537. doi:10.1021/jp806838b
(7) Klimovich, P., & Mobley, D. L. (2010). Predicting hydration free energies using all-atom molecular dynamics simulations and multiple starting conformations. *Journal of Computer-Aided Molecular Design*, *24*(4), 307–316.
(8) Mobley, D. L., Liu, S., Cerutti, D. S., Swope, W. C., & Rice, J. E. (2012). Alchemical prediction of hydration free energies for SAMPL. *Journal of Computer-Aided Molecular Design*, *26*(5), 551–562. doi:10.1007/s10822-011-9528-8
(9) Mobley et al., details TBA, SAMPL4 hydration free energies, 2014.