This repository Search Pull requests Issues Gist

MobleyLab / **FreeSolv**

Unwatch 17 Star 11 Fork 8

Code Issues 8 Pull requests 0 Projects 0 Wiki Pulse Graphs Settings

Branch: master FreeSolv / **README.md** Find file Copy path

davidlmobley Update README.md with paper citation info. 3db971c a day ago

5 contributors

195 lines (155 sloc) 24.6 KB Raw Blame History

FreeSolv: Experimental and Calculated Small Molecule Hydration Free Energies

This repository provides an issue tracker and revision control for the FreeSolv database, initially described in JCAMD (10): http://dx.doi.org/10.1007/s10822-014-9747-x. If you find any issues, please raise an issue in the issue tracker or file a pull request!

Releases are automatically assigned unique DOIs via Zenodo. Latest release: DOI 10.5281/zenodo.495235

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 including 'groups', which provides functional groups for the compounds as assigned by checkmol), PubChem compound IDs, calculated enthalpies of hydration, some experimental enthalpies of hydration (from ORCHYD), and components of the enthalpy of hydration and hydration free energy (as described in our forthcoming paper, to be linked here soon).
- `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.tar.gz` : `mol2` files with partial charges as written by OEChem in Sybyl format/Sybyl atom types
  - `mol2files_gaff.tar.gz` : `mol2` files with partial charges as used for our hydration free energy calculations (AMBER GAFF atom types)
  - `sdffiles.tar.gz` : `sdf` files with partial charges as written by OEChem
  - `gromacs_original.tar.gz` : 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.

(See the Manifest below for a more complete list of all available files.)

The future:

The database is maintained on the cite-able eScholarship repository of the University of California. It is currently available on that site at www.escholarship.org/uc/item/6sd403pz. Updated versions will be maintained there, mirroring point releases provided via this GitHub site.

Please cite:

> Mobley, David L. (2013). Experimental and Calculated Small Molecule Hydration Free Energies. UC Irvine: Department of Pharmaceutical Sciences, UCI. Retrieved from: http://www.escholarship.org/uc/item/6sd403pz

Manifest

- `gromacs_analysis` : Contains plots resulting from GROMACS analysis of some of the data in FreeSolv.
- `gromacs_energies` : Contains XVG files associated with the most recent (2017) update of FreeSolv calculated values; these files are large and are only available in the archived version of the database and not on GitHub.
- `gromacs_mdpfiles` : Contains GROMACS run (.mdp) files used for the calculations connected with the most recent (2017) update of the calculated hydration free energies and enthalpies reported here.
- `mol2files_gaff.tar.gz` : contains mol2 files for all compounds with AM1-BCC charges and GAFF atom types
- `mol2files_sybyl.tar.gz` : contains mol2 files for all compounds with AM1-BCC charges and SYBYL atom types

- `primary-data` : Primary data from which the contents of this database can be re-generated; obtained from full database via `scripts/extract-primary-data.py`
- `scripts` : Scripts pertaining to the material deposited here
- `sdffiles.tar.gz` : SDF-format files for all of the molecules deposited here (as in `mol2files_gaff` and `mol2files_sybyl`)
- `amber.tar.gz` : AMBER format parameter, coordinate, and frcmod files corresponding to the systems we ultimately simulated in GROMACS.
- `gromacs_original.tar.gz` : GROMACS format topology and coordinate files for the calculations associated with the computed values in FreeSolv, for calculations in gas phase. These were generated from AMBER files via acpype, prior to our more recent migration to ParmEd.
- `gromacs_solvated.tar.gz` : GROMACS format topology and coordinate files for the calculations associated with the computed values in FreeSolv, for calculations in solution, again generated from AMBER files via acpype.
- `lammps.tar.gz` : LAMMPS format topology and coordinate files for the calculations associated with the computed values in FreeSolv, automatically converted using InterMol from the AMBER files
- `charmm.tar.gz` : CHARMM format topology and coordinate files for the calculations associated with the computed values in FreeSolv, automatically converted using ParmEd (via InterMol) from the AMBER files
- `gromacs.tar.gz` : GROMACS format topology and coordinate files for the calculations associated with the computed values in FreeSolv, automatically converted using ParmEd (via InterMol) from the AMBER files
- `desmond.tar.gz` : DESMOND format topology and coordinate files for the calculations associated with the computed values in FreeSolv, automatically converted using InterMol from the AMBER files
- `simulation_comparison_input/` : directory containing input files used for the validation of the input conversion files by comparing energy files, description of automated conversion process, and the energy comparisons. See `simulation_comparison_input/README.md` for more details.
- `README.md` : This file
- `database.pickle` : Python pickle file of the FreeSolv database
- `database.json` : JSON format version of the FreeSolv database also stored in `database.pickle`
- `database.txt` : Text format version of some of the fields from the database
- `groups.txt` : Functional groups assigned to the different compounds in the database
- `iupac_to_cid.pickle` and `.json` : Python pickle file and JSON file containing a dictionary for converting IUPAC names to FreeSolv compound IDs
- `smiles_to_cid.pickle` and `.json` : Python pickle and JSON file containing a dictionary for converting SMILES strings to FreeSolv compound IDs
- `notebooks/OrionDB.ipynb` : iPython notebook providing an example of concatenating molecules and associating generic data.

Rebuilding FreeSolv

The input files deposited here can be rebuilt (from SMILES strings) using the script `scripts/rebuild_freesolv.py` , which requires the Chodera lab's `openmoltools` package and the Mobley Lab's `SolvationToolkit` , both of which are `conda` installable from the `omnia` channel.

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.

On 12/20/2013 this database was moved to the eScholarship site of the University of California, at http://www.escholarship.org/uc/item/6sd403pz.

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 pentan-2-one in the set of (4) was pentan-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. [See v0.2 notes -- this revision was actually a mistake, and in fact this was the correct compound, though the tools we were using did not properly parse the alternative name, "ethylene glycol diacetate".
  - 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).

Version 0.2:

- Corrected the experimental references from one of our earlier papers (10.1021/jp0667442) which incorrectly reported the data as having come from the Rizzo set, but it instead came from Abraham et al. 1990 (10.1039/P29900000291). Updated experimental uncertainty estimates for this set to match the Abraham et al. "suggestion" of 0.2 kcal/mol
- Corrected the experimental value for 1,3-butadiene (and the experimental reference), as pointed out by Christopher Bayly (OpenEye Software). Specifically, the Hine and Mookerjee paper (JOC (1975) 40:292) finds two experimental values for 1,3-butadiene: -log(cg)=1.39 and -log(cw)=1.87. From these, he derives a value of -0.41 for the former minus the latter, which leads to a transfer free energy of 0.56 kcal/mol. The correct difference is -0.48 not -0.41, which leads to a transfer free energy of 0.65 kcal/mol. This applies to compound mobley_511661, IUPAC 'butadiene'. The prior value was listed as 0.6 kcal/mol in this set (0.56 kcal/mol rounded). The citation was updated as well to point to this original experimental data.

- Updating 2,6-dichlorosyringaldehyde (mobley_6195751) and 3,5-dichloro-2,6-methoxyphenol (mobley_6688723) with improved values from J. Peter Guthrie's SAMPL4 writeup which were NOT used in the SAMPL4 challenge, as he didn't make final changes until many people's manuscripts were submitted. These took the values for 2,6-dichlorosyringaldehyde from -8.24+/-0.76 to -8.68+/-0.76, and 3,5-dichloro-2,6-methoxyphenol from -6.24+/-0.38 kcal/mol to -6.44+/-0.38 kcal/mol
- Updated (2E)-hex-2-enal, mobley_2792521, with detailed experimental references and a slight update to the hydration free energy (-3.60 kcal/mol, vs previous value of -3.68 kcal/mol) based on a weighted average of the available experimental data.
- Re-added "ethylene glycol diacetate" (which was removed under v0.1) from reference (5) as this was in fact the correct compound, and had been removed because of issues relating to handling of the name. This has been assigned the more standard name, "2-acetoxyethyl acetate".
- Updated uncertainty estimates for experimental values in the set of reference (5) to 0.2 rather than 0.6 kcal/mol, to match the estimate given in reference (5).
- Experimental references were updated/corrected, typically by drilling down (for example, in v0.1, the experimental citation for the 504 molecule set was listed as reference (2); now, references point to reference (3), the reference for the Bordner set, and to original source data, depending on the compound). Much more could be done here, but as substantial manual intervention is needed it is unlikely to happen soon.
- In v0.1, IUPAC names for various compounds were supposed to have been modified to make them easier to parse (essentially, standardization of various nonstandard names) and this was reflected in the notes field for these compounds. However, the IUPAC names themselves were never updated. These have now been corrected.
- In preparation for adding PubChem compound IDs, we detected several IUPAC name/SMILES string pairs which did not lead to a compound on PubChem. Alternate IUPAC names were assigned as follows:
  - mobley_2636578, formerly 1,3-bis-(nitrooxy)propane, renamed as 3-nitrooxypropyl nitrate
  - mobley_819018, formerly trans-3,7-Dimethylocta-2,6-dien-1-ol, renamed as (2E)-3,7-dimethylocta-2,6-dien-1-ol
- PubChemIDs for all compounds were added automatically using PubChemPy by looking up compounds via IUPAC name, with a fallback to SMILES string. In several cases cases (mobley_6843802, [(1R)-1,2,2-trifluoroethoxy]benzene; mobley_7869158, [(2S)-butan-2-yl] nitrate; and mobley_9741965, 1,3-bis-(nitrooxy)butane) the PubChem ID was assigned manually because of issues with PubChem's name for the compound and/or issues relating to PubChem not specifying stereochemistry for a chiral center.

Version 0.21:

- The structure files for 2-acetoxyethyl acetate, mobley_4689084, SMILES CC(=O)OCCOC(=O)C, contained multiple conformations of the molecule. This was corrected. Additionally, the .sdf file for this molecule had been written in mol2 format.

Version 0.3 (Feb. 4, 2014):

- Due to bug(s) in Checkmol and issues with its handling of the `.mol2` file format, functional groups assigned to some molecules were incorrect (for example, around eight molecules were incorrectly labeled as cations, with no other groups correctly assigned). After correspondence with the authors, we switched to running checkmol on the associated `.sdf` files, which are better supported by the program, eliminating these problems. All functional groups were re-computed and re-stored.

Version 0.31 (Sept. 25, 2014):

- Repaired partial charges in some .mol2/.sdf files: Due to a human error in retrieving old files, the .mol2 and .sdf files for the compounds from the Dumont set (calculated value reference key 10.1021/jp0667442) contained partial charges which were inconsistent with those used for the calculated values. In six cases, the partial charges in the distributed files were zero, whereas in the remainder of cases they were only slightly different due to use of an apparently different charge calculation procedure. The six compounds with zero charges were mobley_186894, mobley_2005792, mobley_3738859, mobley_5157661, mobley_5449201, and mobley_9055303, while the full list of affected compounds was IDs mobley_1323538, mobley_5449201, mobley_3053621, mobley_3738859, mobley_8427539, mobley_1873346, mobley_5157661, mobley_9979854, mobley_2005792, mobley_9055303, mobley_1923244, mobley_3727287, mobley_20524, mobley_2068538, mobley_1875719, mobley_186894, mobley_2049967, mobley_511661, mobley_2972906, mobley_4035953, mobley_525934, mobley_1728386, mobley_2178600.

- Some .mol2 files had residue names listed as <0>, which can cause problems for some codes. All .mol2 files were standardized to use the residue name "MOL".
- Corrected expt_reference field for 423 molecules to correctly point to the Rizzo et al. work (10.1021/ct050097l) rather than the Mobley et al. 504 molecule study. Corrected expt_reference field for mobley_8809274 as it had been in error in the Rizzo work (personal correspondence, RC Rizzo)
- Minor details:
  - Updated database.txt to have correct release date and version, and to list units of free energies in the headers.
  - Added citation detail for FreeSolv to the References section below.
  - Corrected IUPAC name of 'biphenyle' to 'biphenyl'; the notes already said this had been done, but the name had not been updated.

Version 0.32 (Sept. 29, 2015):

- Corrected SMILES strings (and other files) for nitro-containing compounds mobley_3802803 and mobley_9741965. Due to some type of earlier error, the GAFF and SYBYL .mol2 files for these contained incorrect bonding in the nitro group(s), which resulted in generation of incorrect SMILES when generating FreeSolv. These SMILES strings have now been corrected, as has the bond type in the .mol2 files. Partial charges in the topology files and .mol2 files were retained as use for the calculations reported here, and will be updated in a subsequent release when the calculations are repeated. Checkmol groups for these compounds were also updated. Thanks to Christopher Bayly for noticing these issues. (9/29/15)
- Added (temporarily?) unique, short nicknames to all compounds in database.txt and database.pickle; these consist of IUPAC names when short, or common/other names which are unique and lead to useful hits when used as search terms. (Approx. 10/21/14)
- Removed mobley_4689084, which duplicates mobley_352111 (same experimental value and source data, but the calculated value of the former is older, and topology/coordinate file were less well curated). (10/24/14)

Version 0.320:

Same as the above but initiates Zenodo DOIs. DOI http://dx.doi.org/10.5281/zenodo/159499

Version 0.5 (Jan. 26, 2017) (10.5281/zenodo.264280):

- Re-generates all input files (`.mol2` , `.sdf` , GROMACS and AMBER format files, etc.) from primary data (SMILES strings)
- Deposits scripts used for re-generating the database in the `scripts` directory
- Re-calculates all calculated values (in conjunction with forthcoming paper)
- Adds calculated enthalpies of hydration and components of enthalpy
- Adds charge and non-polar components of hydration free energy
- Adds a few experimental enthalpies of hydration obtained from the ORCHYD dataset
- Adds `README.md` files in some of the sub-directories better indicating their contents
- Corrects `tripos_mol2` back to `mol2files_sybyl` for consistency with `mol2files_gaff` (as in a prior version, but we had lost this change)
- Provides JSON versions of database files

Version 0.51 (April 5, 2017) (10.5281/zenodo.495235):

- Introduced automatically-generated input files for CHARMM, DESMOND, and LAMMPS, and alternate GROMACS files generated via ParmEd rather than acpype
- Reorganizes naming convention of simulation structure files

- Provides energy comparison of all automatically generated files in `simulation_comparison_input`
- Addition of `notebooks` directory

The changes made in the Version 0.5 and 0.51 updates are described in our recent FreeSolv update/mini-review paper in the Journal of Chemical and Engineering Data.

Changes not yet in a formal release:

Contributors

(Please let us know if your name should be on this list but isn't)

- David L. Mobley (UC Irvine)
- J. Peter Guthrie (University of Western Ontario)
- The many people who contributed to the SAMPL challenges over the years and our early studies on hydration free energies, prior to construction of this database.
- Guilherme Duarte Ramos Matos (UC Irvine)
- Daisy Y. Kyu (UC Irvine)
- John D. Chodera (MSKCC)
- Michael R. Shirts (Colorado)
- Hannes H. Loeffler (STFC Daresbury)
- Nathan M. Lim (UC Irvine)

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, D. L., Wymer, K. L., Lim, N. M., Guthrie, J. P. (2014) "Blind prediction of solvation free energies from the SAMPL4 challenge", Journal of Computer-Aided Molecular Design, 28:135-150 (2014).
- (10) Mobley, D. L., and Guthrie, J. P., "FreeSolv: A database of experimental and calculated hydration free energies, with input files", Journal of Computer-Aided Molecular Design, 28(7):711-720 (2014)
- (11) Duarte Ramos Matos, G. et al., "Approaches for calculating solvation free energies and enthalpies demonstrated with an update of the FreeSolv database", bioRxiv 10.1101/104281