Published in final edited form as:
J Comput Aided Mol Des. 2014 July ; 28(7): 711–720. doi:10.1007/s10822-014-9747-x.

FreeSolv: A database of experimental and calculated hydration free energies, with input files

David L. Mobley and
Department of Pharmaceutical Sciences and Department of Chemistry, 147 Bison Modular, University of California, Irvine, Irvine, CA 92697, Tel.: 949-824-6383, Fax: 949-824-2949. Department of Chemistry, University of New Orleans, 2000 Lakeshore Drive, New Orleans, LA 70148

J. Peter Guthrie
Department of Chemistry, University of Western Ontario, London, ON, Canada

David L. Mobley: dmobley@mobleylab.org

Abstract

This work provides a curated database of experimental and calculated hydration free energies for small neutral molecules in water, along with molecular structures, input files, references, and annotations. We call this the Free Solvation Database, or FreeSolv. Experimental values were taken from prior literature and will continue to be curated, with updated experimental references and data added as they become available. Calculated values are based on alchemical free energy calculations using molecular dynamics simulations. These used the GAFF small molecule force field in TIP3P water with AM1-BCC charges. Values were calculated with the GROMACS simulation package, with full details given in references cited within the database itself. This database builds in part on a previous, 504-molecule database containing similar information. However, additional curation of both experimental data and calculated values has been done here, and the total number of molecules is now up to 643. Additional information is now included in the database, such as SMILES strings, PubChem compound IDs, accurate reference DOIs, and others. One version of the database is provided in the Supporting Information of this article, but as ongoing updates are envisioned, the database is now versioned and hosted online. In addition to providing the database, this work describes its construction process. The database is available free-of-charge via http://www.escholarship.org/uc/item/6sd403pz.

Keywords

hydration free energy; transfer free energy; alchemical; molecular dynamics; free energy calculations

6 Supporting Information
In the Supporting Information, we provide version 0.3 of the FreeSolv database, released Feb. 3, 2014, and a PDF file detailing changes leading up to this database.

1 Introduction

Hydration free energies have been of substantial interest to the molecular simulations and computer-aided drug discovery communities for many years. These free energies describe the transfer of small molecules between gas to water, or their relative populations in gas and water at equilibrium. This interest stems from both practical and scientific reasons. Water is of considerable interest as a solvent, and these free energies can be used to probe aspects of solvation we do not yet understand [6, 30, 9, 2, 10]. Furthermore, since biomolecular binding interactions involve at least partial transfer of a molecular ligand from solution into a binding site, our ability to accurately model solvation and desolvation is thought to provide insight into the level of accuracy we could expect under ideal circumstances in a binding free energy calculation. That is, we should not expect to have substantially *higher* accuracy in binding calculations than we can when computing hydration free energies. At a more practical level, these calculations are interesting in part simply because they can be calculated extremely precisely from molecular simulations for many small molecules [48, 32], enabling quantitative comparison to experiment. This comparison can provide insight into where and how to improve our underlying solvation models and force fields[32, 23, 24, 49, 51, 50, 35, 37, 11, 21].

For these reasons, the Mobley lab has spent a good deal of effort on hydration free energy calculations. Our approach to calculating these typically involves alchemical free energy calculations based on classical molecular dynamics (MD) simulations[5, 29, 7, 47], usually with a fixed-charge force field in explicit solvent. While other methods such as implicit-solvent calculations[49, 40, 33, 44] and MD simulations based on polarizable force fields[41, 42] or QM-MM approaches[60] are also of considerable interest, this has not been a major emphasis of our work.

Because of our interest in all-atom MD simulations, we previously compiled a database of roughly 504 neutral small molecules with experimental hydration free energies, and we computed hydration free energies of all of these compounds in both implicit solvent[33] and explicit solvent[32] using the GAFF small molecule force field[59, 58], AM1-BCC partial charges[17, 18], and the AMBER (implicit solvent case)[3] and GROMACS (explicit solvent case)[54] simulation packages. This dataset, typically called the “504 molecule set” or the “Mobley set”, has seen substantial use as a benchmark and test set in a reasonably wide variety of applications. We attribute this use partly to the substantial size of the set, but also partly because it includes both experimental and calculated values for all of the compounds, as well as input files. So, for example, it has been used to test and/or train implicit solvent models to reproduce explicit solvent results with the same parameters, as well as for direct comparisons of new or existing force fields against experiment[11, 13, 24, 8, 55, 28, 38, 1, 57, 27, 25].

While this previous set, which we here call “the 2008 set”, has been useful, it has several deficiencies. First, there are several errors in the set itself, in terms of duplicate compounds, incorrect values, and so on. While these issues are being corrected via an erratum, it seems likely that further updates will be needed in the future (especially if new experiments begin being done), and there is no obvious mechanism for keeping the database updated when its

main repository is the Supporting Information of a particular paper. Second, the format is less than ideal (in that much of the key information is embedded in PDF files within the Supporting Information), making it difficult to deal with in an automated manner. While we have provided this information in alternate formats such as plain text to individual researchers, this is hardly an ideal solution. Third, we now have additional experimental and calculated values[1] and we would like to extend the set to include these. Fourth, an ideal database would also include additional information to improve ease-of-use, such additional compound identifiers like SMILES strings or identifiers from other databases such as PubChem, and better handling of experimental sources. Finally, an ideal database should be extensible in a straightforward manner.

To improve on the current situation, we have moved our database online to a permanent, cite-able URL (http://www.escholarship.org/uc/item/6sd403pz) and simultaneously updated, expanded, and curated the set, also adding additional, smaller sets we have studied previously and since. This paper reports on the update and curation process. The final product includes a variety of changes described below, to deal with limitations of the previous database. Additionally, the database is now versioned. While one specific version of the database is deposited in the Supporting Information associated with this paper, the full database now has a permanent, cite-able repository online which will allow further updates. Here, we describe our curation and construction process for this database, which we call the "Free Solvation Database" or FreeSolv.

2 Database construction

2.1 Starting points

The starting point in constructing the FreeSolv database was to pull together all of the lead author's previous work calculating hydration free energies in explicit solvent. This included calculated values, experimental values, and structures and input files[2] from several previous studies[34, 39, 33, 32, 31, 22, 35, 36]. To simplify the following discussion, we will refer to the set represented in each study by one of the author's names[3], except for the large 2008 set [33, 32] as noted above. Specifically, we drew on the Dumont set [34], the Nicholls set [39], the 2008 set [33, 32], the Mobley set [31], the Klimovich set [22], the Liu set [35], and the Wymer set [36].

For all of these sets except one, we had retained not only calculated and experimental hydration free energies and original coordinate files (.mol2 format) containing geometries and partial charges, but also input files in the form of GROMACS topology and coordinate files. However, for the Nicholls set, we no longer had topology and coordinate files, so these were re-generated using Antechamber and ACPYPE[52].

[1]Obtained using essentially the same protocols
[2]with one exception described below
[3]The author selected is usually one of those involved in running the calculations represented; for most of these sets, J. Peter Guthrie was key in determining the composition of the set.

After pulling together all these files, we found we had source files for 736 compounds. However, no cross-checking had been done at this point to ensure uniqueness of compounds. Uniqueness will be addressed below.

It is worth highlighting that this database contains only *neutral* solutes[4]. This is driven by two main considerations. In part, a variety of technical issues make alchemical free energy calculations for charged solutes extremely challenging[20, 19, 45] and we have only recently begun to understand the necessary corrections. Secondarily, experimental measurements of ionic hydration free energies are typically not possible, and typically must be obtained from decomposing solvation of ionic pairs into solvation of the individual compounds. This step can involve assumptions which are controversial. Hence, here, our focus has been on hydration free energies of neutral compounds. It is worth noting, however, that the Rizzo lab database[44] (http://ringo.ams.sunysb.edu/index.php/Rizzo_Lab_Downloads) contains in excess of 50 ions, including monoatomic and polyatomic ions, so the interested reader is referred there.

2.2 Error correction

We were already aware of several errors which we corrected in construction of the FreeSolv set. These will also be addressed in errata to the relevant individual studies. Specifically:

- A human error had resulted in an incorrect structure and name (triacetyl glycerol) of the molecule which was intended to be triacetin/glycerol triacetate, in the 2008 set[33]. This compound had originated from the Nicholls set[39], where it was correct. The incorrect structure/name is now removed but the correct molecule from the Nicholls set is retained.
- The experimental value for hexafluoropropene was corrected from −3.76 kcal/mol to 2.31 kcal/mol; it had incorrectly been assigned the value for hexafluoropropan-2-ol due to human error interpreting abbreviations in reference [44], as per personal communication[43].
- Several duplicates within the 2008 set[33] were removed, including 2-methylbut-2-ene under slight variants of the same name, 3-methylbut-1-ene in similar circumstances, and benzonitrile which is equivalent to cyanobenzene.
- From the 2008 set[33], we removed a duplicate butanal entry which had an incorrect experimental value
- The molecule labeled pentan-2-one in the Dumont set[34] was actually pentan-3-one, so the name and experimental value were updated to reflect the correct compound
- The molecules labeled “lindane” and “prometryn” from the Mobley set were removed because of incorrect stereochemistry in the former case, and a swap between a dimethyl and an ethyl in the latter case. This issue appears to have originated in conversion of .xyz format files to 3D structures when the organizers

[4]It does contain a variety of carboxylic acids which would be expected to be charged in solution at neutral pH, but hydration free energies of these are typically reported for the neutral form of the molecule

were preparing for the Statistical Assessment of Modeling of Proteins and Ligands challenge[14], and will likely require errata to several papers utilizing the relevant set[14]. This was caught during the curation process discussed below.

2.3 Initial construction process

While ideally each compound might be identified by its IUPAC name or SMILES string, different schemes for constructing these can lead to different names or strings. Every compound in the set needs a unique identifier, however, so our first step in updating the set was to assign each compound a compound identifier, consisting of the prefix “mobley_” followed by a unique random integer between 0 and 1 billion. These compound IDs serve as the basic identifiers of compounds in the set, and also serve as file names for structures and molecule files. These IDs were assigned automatically via Python script.

Once compound identifiers were assigned, we pulled experimental and calculated values, as well as their uncertainties (when applicable – experimental uncertainties were not always available) and names (some followed IUPAC conventions; others did not) from the sets studied previously via custom Python scripts, with one script handling each prior database separately (since data formats differed). The resulting data was stored into a Python dictionary, keyed by compound ID, along with separate digital object identifiers (DOIs) for the sources of the experimental and calculated values. Our Python scripts also organized the supporting files (3D structures and parameter files), ensuring we had .mol2 files with both SYBYL and GAFF atom naming conventions for each molecule, and organizing the appropriate GROMACS topology and coordinate files. As noted above, in the case of the Nicholls set[39], the relevant script also re-generated topology files. A note of this was added to the ’notes’ field in the database for each of the affected compounds.

2.4 Curation process

Following initial construction of the database, we used a Python script drawing on OpenEye software’s Python toolkits[53] to curate the database.

Before doing anything else, this script removed the entry corresponding to 4-nitroaniline from the 2008 set[33], since the Mobley set[31] had this as well with an experimental value which had been more carefully curated[14].

After this, we used OpenEye tools to attempt to parse all of the compound names. Any names which did not parse correctly at this stage were flagged for attention, and these were typically dealt with in one of two ways. First, some of the failures were because stereochemistry information was unspecified by the compound name, but specified in our existing 3D structures. In these cases (1,2-dichloroethylene, nerol) we re-generated IUPAC names from the 3D structure using OpenEye tools. Second, the remaining cases were dealt with manually. There seemed to be several major sources of problems. There were a handful of typos (5-flurouracil rather than 5-fluorouracil, for example), and a variety of other cases where a common name had been used for the compound which was not recognized by the OpenEye toolkits (carbaryl, trifluralin, pirimor, etc.). The Mobley set[31, 14] was the origin of many of these. These were typically resolved by finding alternate names. Our default procedure was to generate the compound from its common name in MarvinSketch[4], and

then compute an IUPAC name within MarvinSketch and check if the OpenEye toolkit could parse it back into the correct structure. When this procedure failed, we resorted to searching Wikipedia or PubChem for alternate compound names and checking that we obtained one which the OpenEye toolkits could parse back into the correct structure. In any case where the IUPAC name was edited as described here, a note to this effect was added in the 'notes' field of the database. All compound names were stored to the 'iupac' field in the database, though not all of these are technically IUPAC names. Additionally, alternate IUPAC names were assigned manually in two additional cases when PubChem lookup (discussed in Section 2.5, below) by the name failed. Specifically, mobley_2636578, 1,3-bis-(nitrooxy)propane, was renamed as 3-nitrooxypropyl nitrate, and mobley_819018, trans-3,7-dimethylocta-2,6-dien-1-ol, was renamed as (2E)-3,7-dimethylocta-2,6-dien-1-ol.

Following this check of compound names, we then generated canonical isomeric SMILES strings for each compound from the 3D structure and stored this to the database. We also then generated an analogous SMILES string for each compound from its stored name. In any case where SMILES generation from the name failed, a new name was generated from the 3D structure and stored, with the 'notes' field updated accordingly. In cases where SMILES were generated from both the name and the 3D structure (the vast majority of cases), we cross-checked these and ensured that they matched. This was the step where we caught the errors relating to lindane and prometryn noted above. Aside from that, no errors were found at this step.

Since for the vast majority of compounds, we now had two isomeric SMILES strings – one generated from the name, and one from the 3D structure – this provided an ideal opportunity check for redundancy in the set. Many compounds at this point appeared multiple times. For example, almost all of the compounds from the Dumont set[34] also appeared in the 2008 set[33, 32]. Some of the compounds from the 2008 set appeared in later sets as well. Thus, our next step was to remove duplicate compounds. This was made slightly more difficult by the fact that in some cases, the experimental data had a different origin (typically because an alternate name for the compound had led us to overlook the duplication initially), and thus the experimental values were potentially different. We dealt with this by identifying compounds which were identical (i.e. their canonical isomeric SMILES strings or chemical names were equivalent and cross-checking their experimental values. In any case where the difference in experimental values was larger than the tabulated experimental uncertainty, the case was flagged for further investigation. This was not true for any of the compounds in the set except 4-nitroaniline, which occurred in both the 2008 and Mobley sets[31, 14]. After investigation, it was concluded that the later value is probably superior and this was retained. The remaining duplicates, where differences were not statistically significant (approximately 72), were removed from the set automatically.

In separate work, J. Peter Guthrie is compiling an extensive, carefully curated database of experimental hydration free energies. We cross-compared experimental values in our set to a pre-release version of the Guthrie database, and flagged discrepancies above 1 kcal/mol. (The number of discrepancies below 1 kcal/mol numbered over 100, and falls within the scope of Guthrie's database curation work rather than the scope of this paper). In these cases we obtained details of the data from Guthrie and in some cases updated experimental values

and references. When we did so, this is shown in the ’notes’ field of the database. This was true for 4-propylphenol, 4-bromophenol, 3-hydroxybenzaldehyde, 2-methoxyethanol, (2E)-hex-2-enal, and dimethyl sulfoxide/methylsulfinylmethane.

Additionally, after consultation with Guthrie, we removed a series of sulfonylurea compounds from the Mobley set[31, 14], because of concerns about the quality of the underlying vapor pressure measurements, especially Figures 2–5 of reference [46]. Specifically, we removed the compounds called sulfometuron-methyl, metsulfuronmethyl, chlorimuronethyl, thifensulfuron, and bensulfuron. Unfortunately this means that we now only have two sulfones in our set, and in general have far too few sulfur-containing compounds, as we discuss below.

We also updated the experimental details for 1,3-butadiene. Specifically, we updated the reference to point to the original experimental data of Hine and Mookerjee[16], and updated our previous hydration free energy of 0.6 kcal/mol to 0.65 kcal/mol. As pointed out by Christopher I. Bayly in personal correspondence, the raw data there for activity coefficients in gas and water ($-\log c_g = 1.39$ and $-\log c_w = 1.87$) leads to a difference of −0.48 rather than the stated value of −0.41, which is apparently a typo. The former leads to a hydration free energy of 0.65 kcal/mol, the correct value, while the latter would yield 0.56 kcal/mol.

As a final step, we also generated SDF format files for all of the molecules in the set using the OpenEye toolkits. These supplement the .mol2 files we already had available.

Any further curation done will be documented in the database documentation distributed with each database version.

2.5 Annotation

In the past, we have found it useful to focus analysis on just a fraction of the database, such as by examining systematic errors organized by functional group[32]. To aid further such analysis, we used Checkmol[15] to assign functional groups to all of the compounds in the set. The resulting functional group identifiers were stored to the database in the ‘groups’ field.

We also decided to link compounds in our set to alternate databases to simplify future work relating to compound identification, so we chose PubChem compound identifiers as an alternate way of referencing compounds. We assigned PubChem compound IDs to all of the compounds in our set using PubChemPy[56] automatically. Our script first attempted lookup by the assigned compound name (usually IUPAC name) and in cases where this did not result in a match in PubChem, it fell back to lookup via SMILES string. In several cases, typically due to unspecified stereochemistry in PubChem, we had to assign a PubChem ID manually. This was the case for mobley_6843802 ([(1R)-1,2,2-trifluoroethoxy]benzene); mobley_7869158, [(2S)-butan-2-yl] nitrate; and mobley_9741965, 1,3-bis-(nitrooxy)butane. PubChem IDs are thus stored in the database for all compounds in the set.

2.6 Database format

Currently, the database is stored within Python as a dictionary, keyed by compound ID, with each compound having keys for the various entries (SMILES string, experimental value and uncertainty, calculated value and uncertainty, (IUPAC) name, functional groups, PubChem ID, and notes). This database is then stored as a Python pickle file, and in a semicolon delimited text file. In the latter format, functional groups are stored to a separate file, groups.txt, to ensure the number of fields in the database text file is manageable. The semicolon delimited format was chosen because other common delimiters (spaces, commas) often occur in compound names making them unsuitable as delimiters.

3 Database contents

Currently, the database contains 643 neutral compounds which can mostly be considered fragment-like from a drug discovery perspective. The range in molecular weight from methane (16.04 Daltons, compound mobley_9055303) to 1,2,3,4,5-pentachloro-6-(2,3,4,5,6-pentachlorophenyl)benzene (that is, decachlorobiphenyl, at 498.66 Daltons, compound mobley_5456566) (Figure 1). The compounds also span a range of polarities. While experimental dipole moments are not part of our data set, we can compute dipole moments based on the AM1-BCC partial charges assigned to molecules, and we find that dipole moments range from 0.0 (methane and many others) to 7.14 for 4-nitroaniline (mobley_6082662). Experimental hydration free energies cover a range of approximately 29 kcal/mol, from 3.43 kcal/mol for octafluorocyclobutane (mobley_1723043) to −25.47 kcal/mol for (2R,3R,4S,5S,6R)-6-(hydroxymethyl)tetrahydropyran-2,3,4,5-tetrol[5] (mobley_9534740). Calculated hydration free energies range from 3.43 kcal/mol for decane (mobley_2197088) to −21.71 kcal/mol for cyanuric acid (mobley_6239320). The distribution of these properties is shown in Figure 2.

While calculated and experimental hydration free energies for the compounds in this set have been compared before, this analysis is spread across several studies and aggregate statistics are not available. Figure 3 compares calculated and experimental values for the set. Here, we find an overall average error of 0.47 ± 0.06 kcal/mol, an RMS error of 1.51 ± 0.07 kcal/mol, an average unsigned error of 1.14 ± 0.04 kcal/mol, a Kendall τ of 0.80 ± 0.01, and a Pearson R of 0.94 ± 0.01.

As noted previously[32, 25], having such a large set of data makes it possible to look for systematic errors in the force field description of particular functional groups. This can also be seen in Figure 4, where we look at the average unsigned error by functional group (as assigned by Checkmol)[6]. Previously, we have used information from similar tests to isolate systematic errors for alkynes[32] and alcohols[12] and taken some steps towards addressing

[5] tetrahydropyran numbering is used here

[6] Various groups used extremely long names and were abbreviated, while some other groups which were underrepresented were filtered out. We provide statistics only for groups occurring in more than 5 compounds, and we renamed “tertiary aliphatic amine (trialkylamine)” to “trialkylamine”, “halogen derivative” to “halogenated”, “tertiary aliphatic/aromatic amine (alkylarylamine)” to “alkylarylamine (3rd)”, “primary aliphatic amine (alkylamine)” to “alkyl amine”, “phenol or hydroxyhetarene” to “phenolic”, “secondary aliphatic/aromatic amine (alkylarylamine)” to “alkylarylamine (2nd)”, “secondary aliphatic amine (dialkylamine)” to “dialkylamine”, “orthocarboxylic acid derivative” to “ca-ortho”, and “carboxylic acid ester” to “ca-ester”

these issues. However, further work in this direction is needed, as it seems fairly clear that some functional groups tend to have particularly large errors.

One reason hydration free energies are of such interest is that they provide a test of potential relevance to binding affinity calculations for drug discovery. But is this set relevant to drug discovery? The typical size of molecules in the set is substantially smaller than typical small-molecule drugs. As noted, many of these molecules are more like “fragments” than drugs. But this may not be a problem as long as we cover all the common chemical functionalities found in drug molecules. For example, if we know that each hydroxyl group typically leads to a systematic error of just over 1 kcal/mol in fragment-like molecules[12], there is no reason to assume the error should be more or less in larger, drug-like molecules. But if there are some functional groups which frequently occur in drug-like molecules but are missing from the present set, then we have very little insight into what level of performance to expect on compounds containing these functional groups.

To compare functional group representation in typical drugs with that in our set, we downloaded the set of small molecule drugs from DrugBank 3.0[26]. This contains over 1500 approved small-molecule drugs and a larger number of experimental drugs, with some 6583 molecules in total. We then compared the functional group distribution seen in these molecules with that represented in our set (Figure 5)[7] On the whole, results are mixed. The present set does cover a reasonably broad range of functional groups, and even has *more* of some functional groups than in typical drugs (chlorinated compounds are a good example of this). But some functional groups are underrepresented by far or do not appear at all, such as aminals/hemiaminals, boronic acid and boronic acid esters, enamines, enols, enol ethers, hemithioaminals, and many sulfur-containing compounds, especially sulfonamides, sulfonic acids, sulfuric acid monoesters, and thiocarboxylic acid esters. If we want to truly understand how our methods can do at predicting thermodynamic properties for molecules containing these functional groups, we will need more data. These classes of compounds are also particularly concerning in that they are further away from the region of chemical space we have studied the most – specifically, current biomolecular force fields have typically started with proteins and sometimes nucleic acids and branched out from there. As we move further from that region of chemical space, we know less about how well we can expect our force fields to work. And thus we particularly need more data for these types of compounds.

4 Conclusions

Here, we provide FreeSolv, an updated database of calculated and experimental hydration free energies for a large set of 643 neutral molecules which are mostly fragment-like. This

[7]As was the case when we examined the average error in our set by functional group, we simplified and shortened a variety of group names, as well as merging some groups and passing over others which contained too few or too many compounds. Specifically, every “carboxylic acid” was abbreviated “ca”, so “carboxylic acid amidine” became “ca-amidine”, etc. Other names were simplified to aid alphabetizing, such as “primary aliphatic amine (alkylamine)” being replaced by “amine, alkyl”, and similar changes for other alcohols and amines. “carbamic acid ester (urethane)” became “urethane”, and “halogen derivative” became “halogenated”. We otherwise retained only groups which occurred in at least 30 compounds in DrugBank, and passed over groups labeled “aromatic”, “heterocyclic”, “anion”, “cation”, and “alkene” because they tended to hit too many compounds or (in the case of “anion” and “cation”) were assigned in error. Other groups were merged to save space, either because they involved sub-categories (i.e. “carboxylic acid imide, N-unsubstituted” and “carboxylic acid imide, N-substituted” just became “carboxylic acid imide”) or to reduce the number of categories (“acetal” and “hemiacetal” became “acetal or hemiacetal”).

database is freely available at http://www.escholarship.org/uc/item/6sd403pz and updates will be posted there when available.

While this database builds on our previously published work, it corrects a number of errors and redundancies and is more carefully curated. It is also designed to allow easy automated use via programs and scripts, and contains a variety of supporting files including molecular structures, topology and coordinate files, parameter files, and so on. We also provide SMILES strings and PubChem compound IDs for all the compounds in the set to allow easier cross-linking to other sources of chemical information.

We hope that the availability of the FreeSolv dataset will drive future force field development, development and testing of new methods, and potentially even new experimental work to fill in gaps in the available data. For example, we have highlighted functional groups which are common in drugs, and which are underrepresented or not present in this set.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Robert C. Rizzo (Stony Brook University) for help tracking down an issue with hexafluoropropene, and many others who have been involved in work on the experimental and calculated values represented in this database, including Élise Dumont, John D. Chodera, Ken A. Dill, Alan E. Barber, II, Anthony Nicholls, Christopher I. Bayly, Matthew D. Cooper, Vijay S. Pande, Michael R. Shirts, Pavel V. Klimovich, Shuai Liu, David S. Cerutti, William C. Swope, Julia E. Rice, Christopher J. Fennell, Nathan M. Lim, and Karisa L. Wymer. We also appreciate work done by Karisa Wymer and Jessica Fuselier towards initial curation of the set. DLM appreciates financial support from the National Institutes of Health (1R15GM096257-01A1), and computing support from the UCI GreenPlanet cluster, supported in part by NSF Grant CHE-0840513.

References

1. Aguilar B, Onufriev AV. Efficient Computation of the Total Solvation Energy of Small Molecules via the R6 Generalized Born Model. J Chem Theory Comput. 2012; 8(7):2404–2411.
2. Baron, R.; Setny, P.; McCammon, JA. Hydrophobic Association and Volume-Confined Water Molecules. In: Gohlke, H., editor. Protein-Ligand Interactions. Wiley-VCH; 2012.
3. Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, Merz KM Jr, Onufriev A, Simmerling C, Wang B, Woods RJ. The Amber biomolecular simulation programs. J Comput Chem. 2005; 26(16): 1668–1688. [PubMed: 16200636]
4. ChemAxon: MarvinSketch (2013)
5. Chodera JD, Mobley DL, Shirts MR, Dixon RW, Branson K, Pande VS. Alchemical free energy methods for drug discovery: progress and challenges. Curr Opin Struct Biol. 2011; 21(2):150–160. [PubMed: 21349700]
6. Chorny I, Dill K, Jacobson MP. Surfaces affect ion pairing. J Phys Chem B. 2005; 109(50):24,056–24,060. [PubMed: 16850976]
7. Christ CD, Mark AE, van Gunsteren WF. Basic ingredients of free energy calculations: A review. J Comput Chem. 2010; 31(8):1569–1582. [PubMed: 20033914]
8. Corbeil CR, Sulea T, Purisima EO. Rapid Prediction of Solvation Free Energy. 2. The First-Shell Hydration (FiSH) Continuum Model. J Chem Theory Comput. 2010; 6(5):1622–1637.
9. Fennell CJ, Bizjak A, Vlachy V, Dill KA. Ion Pairing in Molecular Simulations of Aqueous Alkali Halide Solutions. J Phys Chem B. 2009; 113:6782–6791. [PubMed: 19206510]

10. Fennell CJ, Kehoe CW, Dill KA. Oil/Water Transfer Is Partly Driven by Molecular Shape, Not Just Size. J Am Chem Soc. 2010; 132(1):234–240. [PubMed: 19961159]
11. Fennell CJ, Kehoe CW, Dill KA. Modeling aqueous solvation with semi-explicit assembly. Proceedings of the National Academy of Sciences. 2011; 108(8):3234–3239.
12. Fennell CJ, Wymer KL, Mobley DL. A Fixed-Charge Model for Alcohol Polarization in the Condensed Phase, and Its Role in Small Molecule Hydration. J Phys Chem B. 2014
13. Gallicchio E, Paris K, Levy RM. The AGBNP2 Implicit Solvation Model. J Chem Theory Comput. 2009; 5:2544–2564. [PubMed: 20419084]
14. Guthrie JP. A Blind Challenge for Computational Solvation Free Energies: Introduction and Overview. J Phys Chem B. 2009; 113(14):4501–4507. [PubMed: 19338360]
15. Haider, N. Checkmol. merian.pch.univie.ac.at
16. Hine J, Mookerjee PK. Structural effects on rates and equilibriums. XIX. Intrinsic hydrophilic character of organic compounds. Correlations in terms of structural contributions. J Org Chem. 1975; 40(3):292–298.
17. Jakalian A, Bush B, Jack D, Bayly CI. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. J Comput Chem. 2000; 21(2):132–146.
18. Jakalian A, Jack D, Bayly CI. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. J Comput Chem. 2002; 23(16):1623–1641. [PubMed: 12395429]
19. Kastenholz M, Hünenberger P. Computation of methodology-independent ionic solvation free energies from molecular simulations. I. The electrostatic potential in molecular liquids. The Journal of Chemical Physics. 2006; 124:124, 106.
20. Kastenholz M, Hünenberger P. Computation of methodology-independent ionic solvation free energies from molecular simulations. II. The hydration free energy of the sodium cation. The Journal of Chemical Physics. 2006; 124:224, 501.
21. Kehoe CW, Fennell CJ, Dill KA. Testing the semi-explicit assembly solvation model in the SAMPL3 community blind test. J Comput Aided Mol Des. 2012; 26(5):563–568. [PubMed: 22205387]
22. Klimovich P, Mobley DL. Predicting hydration free energies using all-atom molecular dynamics simulations and multiple starting conformations. J Comput Aided Mol Des. 2010; 24(4):307–316. [PubMed: 20372973]
23. Knight JL, Brooks CL III. Validating CHARMM Parameters and Exploring Charge Distribution Rules in Structure-Based Drug Design. J Chem Theory Comput. 2009; 5:1680–1691. [PubMed: 20046995]
24. Knight JL, Brooks CL III. Surveying implicit solvent models for estimating small molecule absolute hydration free energies. J Comput Chem. 2011; 32(13):2909–2923. [PubMed: 21735452]
25. Knight JL, Yesselman JD, Brooks CL III. Assessing the quality of absolute hydration free energies among CHARMM-compatible ligand parameterization schemes. J Comput Chem. 2013; 34(11): 893–903. [PubMed: 23292859]
26. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djoumbou Y, Eisner R, Guo AC, Wishart DS. Drug-Bank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic acids research. 2011; 39(Database issue):D1035–41. [PubMed: 21059682]
27. Li L, Fennell CJ, Dill KA. Field-SEA: A Model for Computing the Solvation Free Energies of Nonpolar, Polar, and Charged Solutes in Water. J Phys Chem B. 2013:131213113930002.
28. Liu Y, Fu J, Wu J. High-Throughput Prediction of the Hydration Free Energies of Small Molecules from a Classical Density Functional Theory. The Journal of Physical Chemistry Letters. 2013; 4(21):3687–3691.
29. Michel J, Essex JW. Prediction of protein–ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations. J Comput Aided Mol Des. 2010; 24:649–658.
30. Mobley DL, Barber AE II, Fennell CJ, Dill KA. Charge Asymmetries in Hydration of Polar Solutes. J Phys Chem B. 2008; 112:2404–2414.
31. Mobley DL, Bayly CI, Cooper MD, Dill KA, Dill KA. Predictions of Hydration Free Energies from All-Atom Molecular Dynamics Simulations. J Phys Chem B. 2009; 113:4533–4537. [PubMed: 19271713]

32. Mobley DL, Bayly CI, Cooper MD, Shirts MR, Dill KA. Small Molecule Hydration Free Energies in Explicit Solvent: An Extensive Test of Fixed-Charge Atomistic Simulations. J Chem Theory Comput. 2009; 5(2):350–358. [PubMed: 20150953]
33. Mobley DL, Dill K, Chodera JD. Treating entropy and conformational changes in implicit solvent simulations of small molecules. J Phys Chem B. 2008; 112(3):938. [PubMed: 18171044]
34. Mobley DL, Dumont É, Chodera JD, Dill K. Comparison of charge models for fixed-charge force fields: Small-molecule hydration free energies in explicit solvent. J Phys Chem B. 2007; 111(9): 2242–2254. [PubMed: 17291029]
35. Mobley DL, Liu S, Cerutti DS, Swope WC, Rice JE. Alchemical prediction of hydration free energies for SAMPL. J Comput Aided Mol Des. 2012; 26(5):551–562. [PubMed: 22198475]
36. Mobley DL, Wymer KL, Lim NM. Blind prediction of solvation free energies from the SAMPL4 challenge. J Comput Aided Mol Des. 2014
37. Mukhopadhyay A, Fenley AT, Tolokh IS, Onufriev AV. Charge hydration asymmetry: the basic principle and how to use it to test and improve water models. J Phys Chem B. 2012; 116(32): 9776–9783. [PubMed: 22762271]
38. Nerenberg PS, Jo B, So C, Tripathy A, Head-Gordon T. Optimizing Solute–Water van der Waals Interactions To Reproduce Solvation Free Energies. J Phys Chem B. 2012:120406101304000.
39. Nicholls A, Mobley DL, Guthrie JP, Chodera JD, Bayly CI, Cooper MD, Pande VS. Predicting small-molecule solvation free energies: an informal blind test for computational chemistry. J Med Chem. 2008; 51 (4):769–779. [PubMed: 18215013]
40. Nicholls A, Wlodek S, Grant JA. The SAMP1 Solvation Challenge: Further Lessons Regarding the Pitfalls of Parametrization. J Phys Chem B. 2009; 113:4521–4532. [PubMed: 19281198]
41. Ponder JW, Wu C, Ren P, Pande VS, Chodera JD, Schneiders MJ, Haque I, Mobley DL, Lambrecht DS, DiStasio RA Jr, Head-Gordon M, Clark GNI, Johnson ME, Head-Gordon T. Current status of the AMOEBA polarizable force field. J Phys Chem. 2010
42. Ren P, Chun J, Thomas DG, Schnieders MJ, Marucho M, Zhang J, Baker NA. Biomolecular electrostatics and solvation: a computational perspective. Quarterly Reviews of Biophysics. 2012; 45(04):427–491. [PubMed: 23217364]
43. Rizzo, RC. Hexafluoropropene correction (2013). Personal communication. Jun. 2013
44. Rizzo RC, Aynechi T, Case DA, Kuntz ID. Estimation of Absolute Free Energies of Hydration Using Continuum Methods: Accuracy of Partial Charge Models and Optimization of Nonpolar Contributions. J Chem Theory Comput. 2006; 2(1):128–139.
45. Rocklin GJ, Mobley DL, Dill KA, Hünenberger PH. Calculating the binding free energies of charged species based on explicit-solvent simulations employing lattice-sum methods: An accurate correction scheme for electrostatic finite-size effects. The Journal of Chemical Physics. 2013; 139(18):184, 103.
46. Schmuckler ME, Barefoot AC, Kleier DA, Cobranchi DP. Vapor pressures of sulfonylurea herbicides. Pest Manag Sci. 2003; 56(6):521–532.
47. Shirts, MR.; Mobley, DL. Biomolecular Simulations. Methods in Molecular Biology. 2013. An Introduction to Best Practices in Free Energy Calculations.
48. Shirts MR, Pitera JW, Swope WC, Pande VS. Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins. The Journal of Chemical Physics. 2003; 119(11):5740–5761.
49. Shivakumar D, Deng Y, Roux B. Computations of absolute solvation free energies of small molecules using explicit and implicit solvent model. J Chem Theory Comput. 2009; 5(4):919–930.
50. Shivakumar D, Harder E, Damm W, Friesner RA, Sherman W. Improving the Prediction of Absolute Solvation Free Energies Using the Next Generation OPLS Force Field. J Chem Theory Comput. 2012; 8(8):2553–2558.
51. Shivakumar D, Williams J, Wu Y, Damm W, Shelley J, Sherman W. Prediction of Absolute Solvation Free Energies using Molecular Dynamics Free Energy Perturbation and the OPLS Force Field. J Chem Theory Comput. 2010; 6(5):1509–1519.
52. Sousa da Silva AW, Vranken WF. ACPYPE - AnteChamber PYthon Parser interfacE. BMC Res Notes. 2012; 5(1):367. [PubMed: 22824207]
53. Software, O.S. OpenEye Python Toolkits (2013)

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

55. Sulea T, Corbeil CR, Purisima EO. Rapid prediction of solvation free energy. 1. An extensive test of linear interaction energy (LIE). J Chem Theory Comput. 2010; 6(5):1608–1621.

56. Swain, M. PubChemPy. 2013. URL https://pypi.python.org/pypi/PubChemPy/1.0

57. Truchon JF, Pettitt BM, Labute P. A Cavity Corrected 3D-RISM Functional for Accurate Solvation Free Energies. J Chem Theory Comput. 2014:140114120800002.

58. Wang J, Wang W, Kollman P, Case DA. Automatic atom type and bond type perception in molecular mechanical calculations. Journal of Molecular Graphics and Modelling. 2006; 25:247–260. [PubMed: 16458552]

59. Wang J, Wolf R, Caldwell J, Kollman P, Case D. Development and testing of a general amber force field. J Comput Chem. 2004; 25(9):1157–1174. [PubMed: 15116359]

60. Zheng L, Yang W. Practically Efficient and Robust Free Energy Calculations: Double-Integration Orthogonal Space Tempering. J Chem Theory Comput. 2012; 8:810–823.

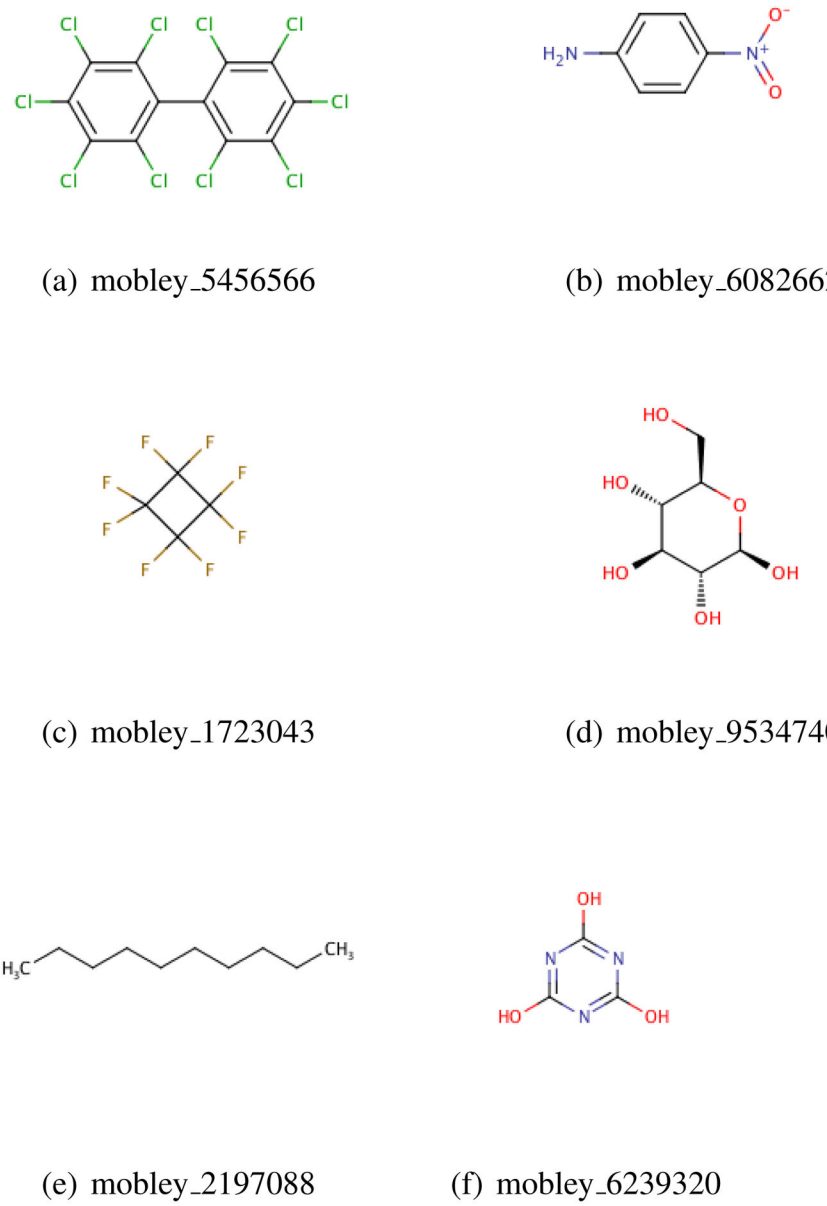
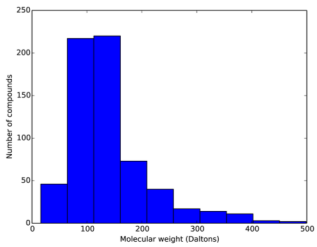
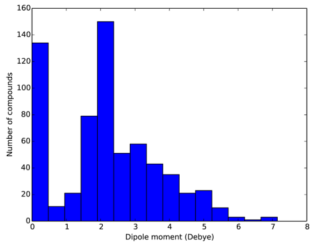


Fig. 1.

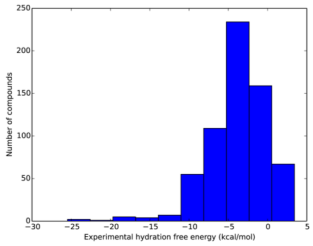
Shown are compounds representing some of the extrema in the set. 1,2,3,4,5-pentachloro-6-(2,3,4,5,6-pentachlorophenyl)benzene (mobley_4546566) has the largest molecular weight, while methane has the smallest. Methane, among others, has the smallest dipole moment, while 4-nitroaniline (mobley_6082662) has the largest. Experimental hydration free energies range from 3.43 kcal/mol for octafluorocyclobutane (mobley_1723043) to −25.47 kcal/mol for (2R,3R,4S,5S,6R)-6-(hydroxymethyl)tetrahydropyran-2,3,4,5-tetrol (mobley_9534740), while calculated values range from 3.43 kcal/mol for decane (mobley_2197088) to −21.71 kcal/mol for cyanuric acid (mobley_6239320).



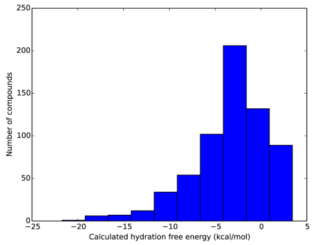
(a) Molecular weight distribution



(b) Dipole moment distribution



(c) Experimental hydration free energy



(d) Calculated hydration free energy

Fig. 2.

Distributions of molecular weight, dipole moment, and hydration free energies for the set described here.

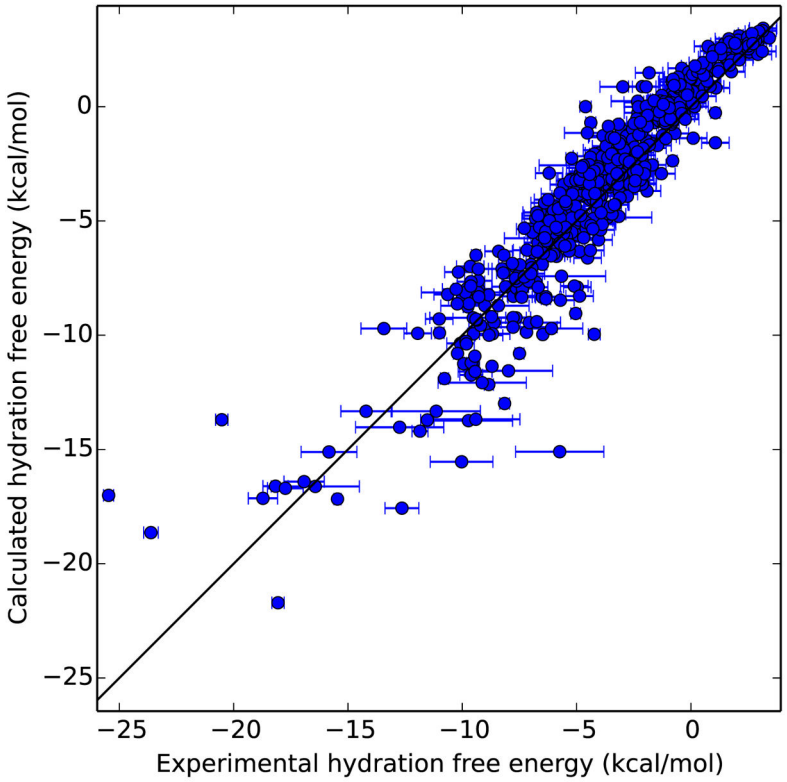


Fig. 3.

Calculated versus experimental hydration free energies for the compounds in the set. Error bars are present for both calculated and experimental values, but statistical uncertainties in the calculated values are extremely small, which typically makes it difficult to see the error bars.

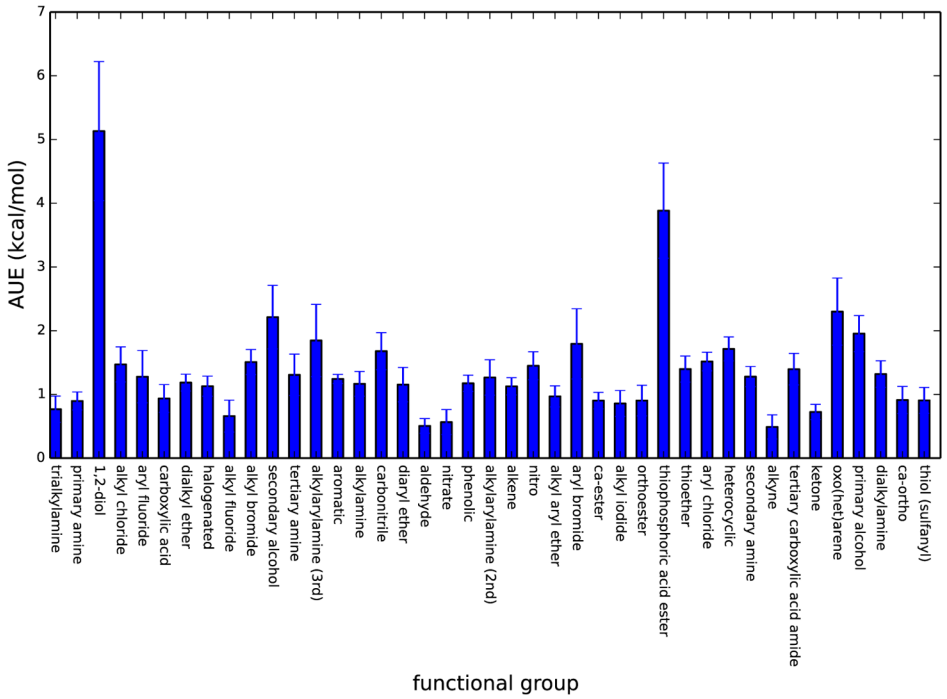


Fig. 4.

Average unsigned error by functional group. Shown is the average unsigned error for compounds in the set by functional group (as assigned by Checkmol) for functional groups represented in at least 5 compounds in the set. Alcohols tend to be particularly problematic, as we are addressing elsewhere[12], but a variety of other functional groups appear particularly challenging as well. Error bars were computed via 10000 iterations of a bootstrapping procedure described elsewhere[36], where we construct new data sets with replacement while resampling the experimental data with Gaussian noise and look at the standard deviation over trials

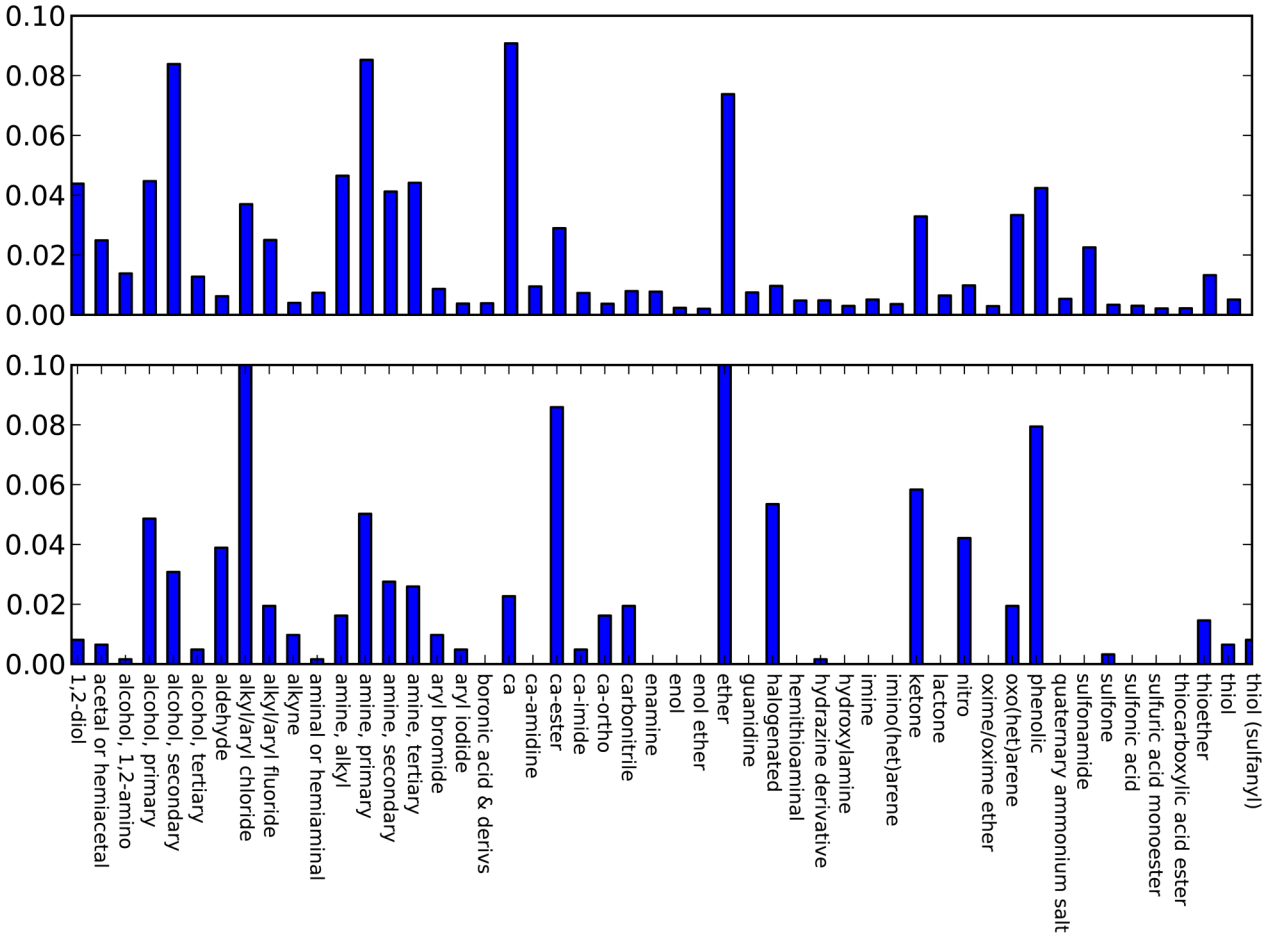


Fig. 5.

Distribution of functional groups in DrugBank versus our dataset. At top is the distribution of functional groups (assigned by checkmol) in DrugBank, and at bottom, the distribution of functional groups in our small-molecule hydration set. Functional groups with fewer than 30 occurrences in DrugBank are excluded for space reasons, and a variety of other functional groups have been merged or skipped as described in the text, again for space reasons. The abbreviation “ca” is short for carboxylic acid.