Noname manuscript No.
(will be inserted by the editor)

Blind prediction of cyclohexane-water distribution coefficients from the SAMPL5 challenge

I don't think we need a subtitle...

Caitlin C Bannan · Kalistyn H Burley · David L Mobley

Received: date / Accepted: date

Abstract Keywords distribution coefficient · blind challenge · free energy

1 Introduction

SAMPL is a blind challenge, has included solvation free energy in the past

What is a distribution coefficient? Relate to free energy, distiguish from logP

More detail about why logD

Possibly brief statement about samples set, site Bas' experimental paper. We provide a detailed look at our submissions to the SAMPL5 challenge and an analysis of submitted results

2 Challenge Logistics

SAMPL5 began on when the specifications for the challenge became available on the D3R website (ww...). The challenge deadline was and experimental results were provided to participants not long after. As in past SAMPL challenges the same group could submit multiple sets of predictions and opt to remain anonymous. A total of 76 prediction sets from 18 participants were submitted and assigned a 2 digit ID number 01 to 76 that will be used throughout this paper. After the challenge was complete, the D3R resource hosted a meeting at University of California, San Diego March 9-11, 2016.

The logD part of SAMPL5 consisted of 53 molecules divided into batches 0, 1, and 2 containing 13, 20, and 20 molecules respectively. Participants could

C. C. Bannan, K. H. Burley, and D. L. Mobley
University of California, Irvine
Tel.: +123-45-678910
Fax: +123-45-678910
E-mail: dmobley@mobleylab.org

submit just batch 0, batches 0 and 1, or batches 0, 1, and 2. Analysis in this paper will focus on the complete set of molecules, but the separate analysis for batch 0 and batches 0 and 1 is available in the supporting information. Molecules were assigned an identifier in the form SAMPL5_XXX, the complete table can be found below and in the supporting information. Included in the challenge information was the SMILES string for each molecule as well as mol2 and sdf files. Also provided were GROMACS, AMBER, and files prepared for each molecule in a solvated box of water or cyclohexane. All information provided to challenge participants is included in supporting information.

Participants were asked to report a cyclohexane/water distribution coefficient for each molecule. As discussed above, distribution coefficients are the ratio of concentrations for all forms of the solute in cyclohexane and the aqueous layer. During the experimental measurements, the water layer was an aqueous phosphate buffer at 7.4 pH. We also required participants to provide two estimates for uncertainty, a statistical uncertainty for their computational method and a model uncertainty that estimates agreement with experiment. The statistical uncertainty should be the variation expected from repeated computational calculations. The model uncertainty, on the other hand, is an estimate of how well the calculated value will agree with experiment. For example, in a recent study we computed cyclohexane/water partition coefficients using alchemical solvation free energy calculations in GROMACS where the statistical uncertainties were around but the root mean squared error was around 1.4 log units. An important part of creating predictive models is the ability to know when it will fail. Analysis of model uncertainties then, is an important part of evaluating any model.

3 Error metrics and ...

Similar to past SAMPL challenges, we considered a large number of error metrics in analyzing all predictions submitted to SAMPL5. For each prediction set we calculate the root-mean-squared error (RMSE), average unsigned error (AUE), average signed error (ASE), Pearson's R (R), Kendall's tau (tau). Uncertainty in each metric was calculated as the standard deviation in 1000 bootstrap trials. This bootstrapping technique included variation in the experimental values based on their reported uncertainties.

As discussed above, an important evaluation of a predictive tool is the ability to estimate how well the computational method will agree with experiment. As in SAMPL4, a QQ Plot was created for each prediction set. The fraction of predictions in an uncertainty range is plotted against the expected fraction of predictions within that range, assuming a gaussian distribution around the experimental value with the model uncertainty. Then the slope of data in the QQ plot was stored for each prediction, we will refer to this as the "error slope." An error slope of greater than one indicates that the calculated values are with uncertainty of experiment more often than expected, or in other

words the model uncertainty was over estimated. Oppositely, an error slope less than one indicates the model uncertainty was underestimated.

Where possible, error analyses were repeated for each molecule where the data set is a complete list of all predicted values for the $\log D$ for that compound. By evaluating each molecule, we can highlight molecules that many groups struggled to accurately predict and possibly highlight trends on where most methods need to improve.

4 Submissions from the Mobley Group

We also participated in the challenge, submitting one complete set of predictions before experimental results were provided to the Mobley group. In addition, KHB, a graduate student in the Mobley group performed the calculations submitted to the SAMPL5 challenge. CCB and DLM performed a series of other calculations after the challenge, which were not included in the prediction sets. We considered a null hypothesis where all molecules are assumed to distribute equally between cyclohexane and water. Many fast structural based tools for octanol/water partition coefficients exist, which we compared with little correction for cyclohexane. We also included a number of post challenge corrections for protonation and tautomeric states which were not included in the original prediction set.

4.1 Calculating partition coefficients from solvation free energies

The Mobley group submitted prediction set 39, a calculated partition coefficient between cyclohexane and water. Partition coefficients are the ratio of concentrations in a single tautomeric state of a solute distributed between two solvents. Before the challenge, each molecule was taken directly as the provided SMILES string with no further tautomer enumeration. As demonstrated in the literature, they are directly proportional to the difference between the solvation free energy for the solute into each solvent. We use previously established and automated protocols to calculate the solvation free energy of each molecule into water and cyclohexane. Then the calculated partition coefficient was reported as an estimate for $\log D$.

To calculate solvation free energies, we used automated tools created by the Mobley lab. Molecular dynamics simulations were performed in GROMACS with the General AMBER Force Field (GAFF) with AM1-BCC charges. Topology and coordinate files for the solvated boxes with 1 solute molecule and 500 cyclohexane or 1000 water molecules were built using the Solvation Toolkit. The Solvation Toolkit takes advantage of many open source Python modules. It convert SMILES strings or IUPAC names of any mixture of compounds to parameterized molecules and builds topology and coordinate files for a variety of simulation packages. All molecular dynamics parameters are identical

to previous studies. The molecule is taken from the solvated box to a non-interacting gas phase in 20 lambda values. Solvation free energies are calculated with Alchemical Analysis tool using the multi-state Bennet acceptance ratio to extract free energy difference between the beginning and end state. The partition coefficient was calculated as the difference in the cyclohexane solvation free energy and the hydration free energy. Statistical uncertainty was reported as the propagated uncertainty from the solvation free energy calculations. Model uncertainty was estimated to be the same for all molecules and reported as the root-mean-squared error from a recent study on calculating cyclohexane/water partition coefficient, specifically 1.4 log units.

As a part of this study, we also wanted to verify that a change in the simulation box size does not affect the calculated solvation free energy in cyclohexane. Hydration free energies were previously shown to be independent of box sizes from 2 to 9 nanometers, within calculated uncertainties. Using we calculated the dipole moment for each SAMPL5 molecule. Then the solvation free energy calculations discussed above were repeated with 150, 200, 300, 400, 500, ... cyclohexane molecules in the box.

4.2 Consideration of tautomers after SAMPL

As a follow-up study to our initial SAMPL5 prediction submission we wanted to explore how correcting for changes in protonation or tautomeric state would have affected the partition coefficient predictions we originally submitted. A common way to correct between experimentally measured distribution coefficients and partition coefficients is with pKa values for the solute. This is a simple correction using the Henderson-Hasselbalch equation:

$$\log D \quad (1)$$

to relate the concentration of neutral species to the charged species at a given pH. Therefore a distribution coefficient can be calculated from a partition coefficient as ... for a basic solute and ... for an acidic solute. To follow this trend, we decided to calculate pKa values for each compound and

corrections based on pKa

corrections based on tautomer enumeration

4.3 Comparing to fast, structural based partition coefficient calculators

Many structural based tools exist for octanol/water partition coefficients; they are very fast and generally accurate. However, these tools are all trained on empirical data, meaning they are limited by the training data. We chose the OpenEye tool XlogP as an example of such a tool. Two post prediction sets

were prepared with the XlogP tool. First, the predicted octanol/water partition coefficient was considered an estimate for $\log D$. In the second set, we calculated a correction for the bias between the calculated XlogP values and a set of experimental cyclohexane/water partition coefficients from a previous study.

5 Results and Discussion

To compare each prediction set to experiment compare plots, QQ, error metric calculated for each prediction set

To visualize how prediction sets compared to each other, explain histograms

Some submissions did not include batch 1 or 2, how many of each and approximately where they were on the histograms...

Error Slope, only one group (2 submissions) did a good job with this

5.1 Null Hypothesis

One way of evaluating predictive models is to compare them to a null hypothesis, or default result of some kind. In the case of distribution coefficients, we chose a null hypothesis where we assume all solute molecules distribute equally between cyclohexane and water, corresponding to a $\log D = 0$. We performed all error analyses discussed above on this pretend data set as a point of comparison.

5.2 Prediction sets that performed most strongly

We want to consider how close to experiment RMSE/AUE and how well correlated with experiment tau/R

16 did best across both metric, COSMO-RS, brief statement about procedure

14 and 36 also did very well, making "top 10" by at least 3 of those 4 metrics, 14 is one of Frank Pickard's and 36 is Chris Fennell

but really null did best across the board so we have a lot of work to do... best RMSE over 2.0 log units and average around 3.5

5.3 Compounds that were difficult to accurately predict

Full error analysis repeated for individual molecules, there aren't very many "simple" molecules, almost all have hetro atoms and rotatable bonds...

$\tilde{5}$-10 worst, I'm looking into if there are of tuatomer or functional group similarities...

$\tilde{5}$-10 best, still looking for trends, we know 083, 074. 015 also did poorly.

5.4 Classes of methods...

Broad range of methods, split into classes: MD all atom, MD hybrid, quantum?, Anything similar to COSMO? Pie chart by number maybe?

Clear trends on which are doing well?

5.5 Mobley group prediction results

How did logP do, not bad in general, include our plots

Slight bias for cyclohexane over water, probably because tautomer enumeration will increase concentrations in water?

pKa corrections tautomer corrections

Biases, and major shifts for some compounds - tautomers in cyclohexane not being accounted for

5.6 Reanalysis of difficult tautomers

It was clear from epik and discussions with other SAMPL5 participants (cite Klamt?) that 050 and 083 had dominant tautomers other than provided SMILE

We reran these tautomers to calculate solvation free energies, used state penalty

Generally hard to tell if its tautomer enumeration that isn't good or the solvation free energies

6 Conclustion

Overall, range of methods and performance

Compare to dGhydration in past SAMPL challenges? using average errors, possibly what methods/FF are top ranked?

Tautomer and/or pKa predictions are going to be an important part of improving these

We, as a communitte, need to improve error estimation, both how we do and how we evaluate it...

logP/logD seem to be good options for future blind challenges

Acknowledgements John Chodera and Bas MSKCC Andreas Klamt Chris Fennell Samuel Genheden D3R team people who set up the automated submission system

6.1 Available in supporting info

things provided to participants all scripts used for error analysis all participant files? Can we include the anonymous one? triple check no names/e-mails/institutions/etc in the final submitted data all plots not in the paper

all input/output files for schrodinger calculations all input files and results files for 'logP' calculations, tautomer redos, box size simulations (PME too?) example MDP and run scripts