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

Caitlin C Bannan · **Kalistyn H Burley** ·
Michael Chiu · **Michael K Gilson** ·
David L Mobley

Abstract describe DC part of SAMPL5 challenge

We analyze submissions and provide reference calculations

Results, range of RMSE or AUE? Better or worse than expected from past sample challenges?

Conclusions: tautomeric enumeration,

Keywords distribution coefficient · blind challenge · free energy

1 Introduction

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

Briefly describe value of past hydration challenges and what happened over the years (i.e. very valuable tests; more methods began to agree well with experiment as challenge went on)

Briefly explain problem with continuing hydration challenges (no new measurements, so no ability to select particular types of functionality for follow up or check measured values)

C. C. Bannan
Department of Chemistry, University of California, Irvine

K. H. Burley
Department of Pharmaceutical Sciences, University of California, Irvine

M. Chiu

M. K. Gilson
University of California, San Diego

D. L. Mobley
Department of Chemistry and Department of Pharmaceutical Sciences, University of California, Irvine
147 Bison Modular, Irvine, CA 92697 Tel.: +949-824-6383
Fax: +949-824-2949
E-mail: dmobley@mobleylab.org

Explain motivation for this challenge and how data was obtained

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

More detail about why logD

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

Drug Design Data Resource (D3R)

2 Challenge Logistics

SAMPL5 began on September 15, 2015 when the specifications for the challenge became available on the D3R website (http://drugdesigndata.org), these are also provided in the supporting information. The challenge deadline was February 1, 2016 and experimental results were provided to participants not long after. As in past SAMPL challenges, each group could submit multiple sets of predictions. There was also the option to remain anonymous. A total of 76 prediction sets from 18 participants or participating groups were submitted and assigned a 2 digit ID number 01 to 76 that will be used throughout this paper. Predictions were analyzed and overview statistics, as well as individual analysis of each submission by various error metrics (as detailed below) were returned to each participant. The challenge culminated with discussions of participants experiences and results at the 1st Annual D3R Workshop at the University of California, San Diego March 9-11, 2016.

For the prediction of distribution coefficients in SAMPL5, a total of 53 molecules were considered. Molecules were assigned an identifier in the form SAMPL5_XXX; the complete table can be found below and in the supporting information. The 53 molecules were divided into batches 0, 1, and 2 containing 13, 20, and 20 molecules respectively. We wanted each batch to have a similar dynamic range and for the molecules to increase in size, so on average the smallest molecules are in batch 0 and the largest in batch 2. To control for dynamic range, molecules were grouped by calculated octanol/water partition coefficient and then by molecular weight. The smallest molecules from each partition coefficient group were added to batch 0, then batch 1, the rest of the molecules compromise batch 2.

Participants could submit just batch 0, batches 0 and 1, or batches 0, 1, and 2. The idea was that all participants should attempt predictions on the full set if at all possible, but grouping into batches would allow people with particularly demanding methods (such as polarizable force fields or methods requiring intensive quantum mechanics) to focus on smaller compounds and still be evaluated. Eight submissions from two participants submitted results for only batch 0, an additional five submissions from two participants provided only batch 0 and batch 1. Here we focus on the results for the complete set of molecules (batches 0, 1, and 2). Separate analysis for the other submission options (batch 0 or batches 0 and 1) is available in the supporting information.

Included in the challenge information was the SMILES string for each molecule as well as mol2 and SDfiles. 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 buffer at pH 7.4. 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 0.05 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

Similar to past SAMPL challenges, we considered a large number of error metrics in analyzing all predictions submitted to SAMPL5. Each error metric was calculated for all submissions, by batch and distributed to challenge participants before the workshop. Here we will focus primarily on six error metrics: the root-mean-squared error (RMSE), average unsigned error (AUE), average signed error (ASE), Pearson's R (R), Kendall's tau (tau), and the 'error slope' explained in depth below. We also calculated maximum absolute error and percent of predictions with the correct sign which are not included in the analysis here, but were provided to challenge participants and available in the supporting information. Uncertainty in each metric was calculated as the standard deviation in 1000 bootstrap trials. As described previously, 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 and when a computational method will agree with experiment. As in SAMPL4, a quantile-quantile plot (QQ Plot) was created for each prediction set. QQ Plots show how the distribution of predictions with their model uncertainty compare to expectation, assuming a gaussian distribution. For example, consider the number of predictions within one standard deviation of the expected value, the value on the x-axis represents the normal distribution, or 0.68, the value on the y-axis will represent the fraction of predictions that are within one model uncertainty of the experimental value. A regression analysis helps summarize these results, assuming a y-intercept

of zero. The 'error slope' is the slope of the line comparing the fraction of predictions within range of experiment to the expected fraction in a normal distribution. An error slope of greater than one indicates that the calculated values are within uncertainty of experiment more often than expected, or in other words the model uncertainty was over estimated. Oppositely, an error slope less than one suggests the model uncertainty was underestimated.

The last goal of prediction analysis was to identify any individual molecules where predicting distribution coefficients was difficult. To accomplish this a data set was created for each molecule consisting of all predictions submitted for that molecule, then all the error metrics discussed above were calculated for each molecule. Here we will primarily focus on just average unsigned error for molecules, but all other error metrics were provided to participants and available in supporting information.

4 Reference calculations from the Mobley group

We calculated distribution coefficients through a few different methods as a reference. KHB, a graduate student in the Mobley group, performed a set of blind calculations estimating the $\log D$ as a partition coefficient between cyclohexane and water calculated from solvation free energies. In addition, CCB and DLM performed calculations after the challenge which were not included with 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 no and little correction for cyclohexane. We also included post challenge analysis of protonation and tautomeric states as a correction from calculated partition coefficients to distribution coefficients

4.1 Calculating partition coefficients from solvation free energies

Partition coefficients are the ratio of concentrations of a solute in a single tautomeric state 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, partition coefficients 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 GRO-MACS [1–7] with the General AMBER Force Field (GAFF) [8] with AM1-BCC charges[9,10]. Topology and coordinate files for the solvated boxes with 1 solute molecule and 500 cyclohexane or 1000 water molecules were built

Fig. 1 This is a figure for the box size study

using the Solvation Toolkit. These files were then converted to AMBER, DESMOND, and LAMMPS formats and provided to SAMPL5 participants. The Solvation Toolkit takes advantage of many open source Python modules and is available at https://github.com/MobleyLab/SolvationToolkit. It converts 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 [11,12]. 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 [13] using the multi-state Bennett acceptance ratio to extract free energy difference between the beginning and end state. The partition coefficient was calculated as the difference between the cyclohexane solvation free energy and the hydration free energy. The statistical uncertainty was reported as the propagated uncertainty from the solvation free energy calculations. The 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. These were assigned submission ID 39 and included in the error analysis performed on all submissions.

Simulation box size does not affect the calculated solvation free energy. Hydration free energies were previously shown to be independent of box sizes from 2 to 9 nanometers, within calculated uncertainties [14]. Polar solutes are more likely to significantly affect long range interactions so we calculated the dipole moment of each SAMPL5 molecule using the position and charges on atoms in the mol2 files. SAMPL5_024 had the largest dipole moment so it was used as the solute for the box size investigation. The solvation free energy calculations were set up as described above with 150, 200, 300, 400, 500, ... cyclohexane molecules in the box. We also performed these calculations in a reaction field with the dielectric coefficient for cyclohexane, 2.... We found that for ... to ... nm box sizes there is no significant change in the calculated solvation free energy for SAMPL5_024 in cyclohexane with ... or ... The input, results, molecular dynamics parameter files, and tables of solvation free energies are available in the supporting information.

4.2 Consideration of tautomers after SAMPL5

To help understand the results from our partition coefficient calculations could have been improved, we considered corrections for changes in the solutes protonation or tautomeric states. Distribution coefficients different from partition coefficients in that they include all forms of the solute in both solvents. 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:

$$pH = pK_a + \log \frac{X}{HX} \qquad (1)$$

to relate the concentration of neutral species to the charged species at a given pH. This correction will depend on if the neutral solute is acidic or basic. The equation used to calculate a distribution coefficient ($\log D$) from a partition coefficient ($\log P$) for a basic solute is below.

$$\log D = \log P - log(1 + 10^{pK_a - pH}) \qquad (2)$$

Alternatively for an acid solute we would instead use:

$$\log D = \log P - log(1 + 10^{pH - pK_a}) \qquad (3)$$

We use Schrödinger's Epik tool to calculate pKa values for each molecule according to experimental conditions. We then estimated a $\log D$ using the equations above. Using pKa values only accounts for one change in protonation, whereas a correct distribution coefficient should include all relevant tautomers and protonation states of the molecule in both solvents. To correct for all other tautomer states we used Schrödinger's LigPrep to enumerate tautomers for each molecule in the aqueous solution. The results of the enumeration includes an energy penalty for the predicted population of each tautomer at the given conditions. LigPrep can only perform the tautomer enumeration with water or DMSO as a solvent, so we were unable to predict tautomers in cyclohexane. Therefor both of these corrections account for the protonation or tautomer states only in the aqueous layer and assume the tautomer remains fixed in cyclohexane as the one used in the initial simulation.

4.3 Estimating distribution coefficients with a fast, structural based partition coefficient calculator

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 OEXlogP [15,16] as an example of such a tool. Two post prediction sets were prepared with the OEXlogP 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 [17]. For the rest of this paper we will refer to the octanol/water partition coefficient set as $XlogP_{oct}$ and the bias corrected set as $XlogP_{corr}$.

5 Results and Discussion

Past SAMPL challenges have involved hydration free energy calculations, but this is the first to include any partitioning between multiple solvents. Distribution coefficients can be related to transfer free energy between solvents, which allows us to estimate an expected uncertainties from errors in past hydration free energy calculations. In SAMPL4 [18], the average root-mean-squared error (RMSE) for the top half of submissions was about 1.5 kcal/mol which would correspond to 1.54 log units. In contrast, here there are very few submissions with an RMSE less than 2.5 log units.

A total of six error metrics are used to evaluate all sets of predictions: root-mean-squared error (RMSE), average unsigned error (AUE), average signed error (ASE), Pearson's R (R), Kendall's tau (tau), and the slope from the QQ plot (error slope) (Table 5). For each group, we also created a plot comparing their predictions to experimental results. A few example plots are provided (Fig. 2) these represent a typical submission, in that these groups were in the middle of the pack by most error metrics. Comparison and QQ plots for every submission are available in the supporting information as well as error metric tables broken down by batch.

To help visualize all of the error metrics, the data was compiled into a histogram where results are sorted by what would be ideal for that metric (closest to 1 for error slope for example). These metrics are split into measurements of deviation from experiment (Fig. 3) and correlation with experiment (Fig. 4) distinctions which helped in identifying high performing groups. Most submissions included data for all three batches so these histograms are limited to those submissions. A total of eight submissions from two participants that only included data from batch 0, then an additional 5 submissions from 2 participants with only batches 0 and 1. These submissions are indicated in Table 5 for clarity.

In considering the results for the error slope analysis, participants generally tend to do poorly estimating model uncertainty. The top three submissions are the only within uncertainty of 1, submissions 53 and 60 from Andrew Paluch and submission 43 from Gerhard Koenig Only one submission (40) significantly overestimated their model uncertainty. The rest are below 1, indicating a significant underestimation of the model uncertainty.

5.1 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

ID	Ave. err.	RMS	AUE	tau	R	Err. slope
01^1	2.3 ± 0.8	5.1 ± 0.5	4.3 ± 0.5	0.13 ± 0.12	0.20 ± 0.17	0.44 ± 0.09
02	-0.5 ± 0.3	2.3 ± 0.3	1.7 ± 0.2	0.48 ± 0.07	0.63 ± 0.07	0.69 ± 0.07
03^1	-7.6 ± 3.5	21.3 ± 2.6	15.9 ± 2.5	0.52 ± 0.10	0.59 ± 0.12	-0.00 ± 0.00
04^0	1.6 ± 0.5	2.5 ± 0.6	1.9 ± 0.4	0.77 ± 0.12	0.87 ± 0.05	0.77 ± 0.12
05	-8.2 ± 0.4	8.7 ± 0.5	8.2 ± 0.4	0.29 ± 0.09	0.39 ± 0.11	0.21 ± 0.03
06	1.8 ± 0.5	4.0 ± 0.3	3.4 ± 0.3	0.46 ± 0.09	0.61 ± 0.10	0.58 ± 0.07
07	0.5 ± 0.4	3.3 ± 0.4	2.5 ± 0.3	0.34 ± 0.08	0.51 ± 0.11	0.33 ± 0.07
08	-1.7 ± 0.4	3.5 ± 0.5	2.5 ± 0.3	0.58 ± 0.06	0.70 ± 0.06	0.60 ± 0.08
09	6.5 ± 0.6	7.8 ± 0.6	6.5 ± 0.6	-0.29 ± 0.08	-0.40 ± 0.10	0.35 ± 0.07
10	0.3 ± 0.4	3.1 ± 0.3	2.6 ± 0.3	0.51 ± 0.07	0.69 ± 0.07	0.79 ± 0.07
11	-4.4 ± 1.8	13.3 ± 2.6	6.9 ± 1.6	0.45 ± 0.09	0.53 ± 0.09	0.39 ± 0.08
12	-5.5 ± 2.5	19.4 ± 1.8	15.0 ± 1.6	0.37 ± 0.09	0.39 ± 0.12	-0.00 ± 0.00
13^0	-11.1 ± 5.0	21.0 ± 4.9	12.2 ± 4.8	0.56 ± 0.16	0.43 ± 0.22	0.59 ± 0.17
14	-0.7 ± 0.3	2.7 ± 0.4	2.0 ± 0.3	0.57 ± 0.06	0.72 ± 0.06	0.66 ± 0.08
15	-1.4 ± 0.4	3.3 ± 0.5	2.3 ± 0.3	0.57 ± 0.07	0.70 ± 0.06	0.61 ± 0.07
16	0.5 ± 0.3	2.1 ± 0.2	1.7 ± 0.2	0.73 ± 0.04	0.84 ± 0.04	0.46 ± 0.08
17	-4.2 ± 0.4	5.0 ± 0.4	4.2 ± 0.4	0.36 ± 0.08	0.51 ± 0.10	0.50 ± 0.07
18	-0.8 ± 0.4	2.7 ± 0.4	2.0 ± 0.3	0.47 ± 0.07	0.60 ± 0.08	0.62 ± 0.08
19	1.5 ± 0.3	2.7 ± 0.2	2.3 ± 0.2	0.54 ± 0.07	0.75 ± 0.07	0.83 ± 0.06
20	-2.3 ± 0.4	3.6 ± 0.5	2.7 ± 0.3	0.55 ± 0.07	0.70 ± 0.06	0.48 ± 0.08
21	-1.2 ± 0.5	3.4 ± 0.7	2.4 ± 0.3	0.44 ± 0.08	0.45 ± 0.16	0.58 ± 0.08
22	1.6 ± 0.5	3.9 ± 0.3	3.1 ± 0.3	0.29 ± 0.09	0.48 ± 0.11	0.68 ± 0.08
23	1.9 ± 0.5	4.0 ± 0.4	3.0 ± 0.4	0.42 ± 0.07	0.58 ± 0.08	0.78 ± 0.08
24^0	2.3 ± 0.7	3.3 ± 0.8	2.5 ± 0.6	0.77 ± 0.13	0.88 ± 0.05	0.67 ± 0.15
25	0.0 ± 0.5	3.6 ± 0.3	2.9 ± 0.3	0.53 ± 0.07	0.70 ± 0.07	0.71 ± 0.07
26	2.3 ± 0.7	5.6 ± 0.4	4.6 ± 0.4	0.25 ± 0.08	0.37 ± 0.11	0.46 ± 0.07
27	-0.2 ± 0.4	2.6 ± 0.4	1.8 ± 0.2	0.49 ± 0.07	0.61 ± 0.08	0.66 ± 0.08
28	-2.3 ± 0.4	3.6 ± 0.4	2.7 ± 0.3	0.54 ± 0.07	0.69 ± 0.06	0.47 ± 0.07
29	-6.7 ± 0.4	7.2 ± 0.4	6.7 ± 0.4	0.33 ± 0.08	0.45 ± 0.11	0.28 ± 0.04
30	2.5 ± 0.5	4.3 ± 0.3	3.7 ± 0.3	0.39 ± 0.10	0.52 ± 0.11	0.53 ± 0.07
31	-1.0 ± 0.3	2.7 ± 0.3	2.0 ± 0.3	0.56 ± 0.07	0.72 ± 0.06	0.63 ± 0.08
32	2.5 ± 0.4	3.5 ± 0.3	3.1 ± 0.2	0.47 ± 0.06	0.64 ± 0.07	0.25 ± 0.06
33	-0.1 ± 0.5	3.4 ± 0.3	2.8 ± 0.3	0.53 ± 0.08	0.71 ± 0.08	0.73 ± 0.07
34	-1.3 ± 0.4	3.0 ± 0.4	2.2 ± 0.3	0.56 ± 0.06	0.69 ± 0.07	0.61 ± 0.08
35	0.5 ± 0.4	2.9 ± 0.3	2.2 ± 0.2	0.36 ± 0.08	0.54 ± 0.10	0.35 ± 0.07
36	1.1 ± 0.3	2.6 ± 0.2	2.1 ± 0.2	0.57 ± 0.06	0.75 ± 0.06	0.50 ± 0.07
37^0	-7.1 ± 5.1	19.6 ± 4.3	13.9 ± 3.9	0.59 ± 0.16	0.41 ± 0.22	-0.00 ± 0.00
38	0.8 ± 0.4	3.3 ± 0.3	2.7 ± 0.3	0.41 ± 0.08	0.58 ± 0.08	0.78 ± 0.07
39	1.6 ± 0.3	2.6 ± 0.2	2.1 ± 0.2	0.49 ± 0.08	0.65 ± 0.10	0.63 ± 0.08
40	0.4 ± 0.4	2.6 ± 0.3	1.9 ± 0.2	0.48 ± 0.08	0.61 ± 0.08	1.16 ± 0.05
41	0.3 ± 0.4	3.2 ± 0.3	2.7 ± 0.3	0.53 ± 0.07	0.69 ± 0.07	0.77 ± 0.07
42	4.6 ± 0.4	5.3 ± 0.4	4.6 ± 0.4	0.50 ± 0.08	0.61 ± 0.12	0.15 ± 0.05
43	-0.7 ± 0.4	3.0 ± 0.3	2.3 ± 0.3	0.51 ± 0.08	0.67 ± 0.09	0.94 ± 0.07
44	-0.6 ± 0.3	2.4 ± 0.3	1.8 ± 0.2	0.47 ± 0.07	0.63 ± 0.07	0.70 ± 0.07
45	0.9 ± 0.5	3.6 ± 0.3	2.9 ± 0.3	0.38 ± 0.08	0.58 ± 0.10	0.71 ± 0.07
46	-8.3 ± 0.5	9.1 ± 0.6	8.3 ± 0.5	0.23 ± 0.08	0.31 ± 0.10	0.14 ± 0.03
47	-1.3 ± 0.4	3.3 ± 0.5	2.2 ± 0.3	0.58 ± 0.07	0.71 ± 0.06	0.62 ± 0.08
48	1.5 ± 0.4	3.0 ± 0.3	2.3 ± 0.3	0.38 ± 0.07	0.55 ± 0.08	0.42 ± 0.07
49	-1.1 ± 0.4	3.3 ± 0.4	2.6 ± 0.3	0.42 ± 0.07	0.58 ± 0.07	0.78 ± 0.07
50^1	-7.1 ± 2.5	16.6 ± 3.0	9.2 ± 2.3	0.60 ± 0.08	0.66 ± 0.08	0.38 ± 0.09
51	1.7 ± 0.7	5.2 ± 0.4	4.3 ± 0.4	0.31 ± 0.08	0.46 ± 0.10	0.46 ± 0.08
52^0	-3.5 ± 1.2	5.4 ± 0.7	4.8 ± 0.7	0.56 ± 0.14	0.59 ± 0.13	0.23 ± 0.10
53	0.5 ± 0.4	2.8 ± 0.3	2.2 ± 0.2	0.44 ± 0.09	0.58 ± 0.10	1.00 ± 0.06
54	-1.0 ± 0.3	2.7 ± 0.3	1.9 ± 0.2	0.56 ± 0.07	0.70 ± 0.06	0.65 ± 0.08
55^1	-11.6 ± 3.4	22.3 ± 3.1	13.7 ± 3.1	0.59 ± 0.09	0.61 ± 0.12	0.38 ± 0.09
56	-1.1 ± 0.4	3.3 ± 0.5	2.2 ± 0.3	0.57 ± 0.06	0.71 ± 0.06	0.67 ± 0.08
57	-10.2 ± 2.4	20.2 ± 2.2	12.6 ± 2.2	0.43 ± 0.09	0.42 ± 0.12	0.38 ± 0.07
58	-2.9 ± 0.5	4.8 ± 0.5	3.8 ± 0.4	0.30 ± 0.10	0.44 ± 0.12	0.55 ± 0.08
59^0	-4.2 ± 1.0	5.6 ± 0.6	5.2 ± 0.6	0.54 ± 0.15	0.55 ± 0.14	0.13 ± 0.08
60	0.2 ± 0.4	2.5 ± 0.4	1.9 ± 0.2	0.49 ± 0.08	0.60 ± 0.08	1.02 ± 0.06
61	-1.2 ± 0.4	3.4 ± 0.6	2.4 ± 0.3	0.44 ± 0.08	0.45 ± 0.16	0.53 ± 0.07
62	0.7 ± 0.5	3.5 ± 0.4	2.7 ± 0.3	0.27 ± 0.09	0.38 ± 0.12	0.73 ± 0.07
63	-4.5 ± 1.8	13.3 ± 2.6	6.9 ± 1.6	0.45 ± 0.09	0.52 ± 0.09	0.41 ± 0.08
64	1.3 ± 0.7	5.2 ± 0.4	4.4 ± 0.4	0.35 ± 0.08	0.51 ± 0.11	0.43 ± 0.07
65	-2.2 ± 0.5	4.4 ± 0.5	3.5 ± 0.4	0.24 ± 0.10	0.35 ± 0.12	0.61 ± 0.08
66	1.4 ± 0.7	5.4 ± 0.4	4.6 ± 0.4	0.34 ± 0.08	0.51 ± 0.10	0.41 ± 0.07
67^0	-5.0 ± 3.0	11.9 ± 4.3	6.2 ± 2.8	0.59 ± 0.18	0.58 ± 0.13	0.56 ± 0.17
68	2.5 ± 0.4	3.6 ± 0.3	3.1 ± 0.2	0.47 ± 0.07	0.64 ± 0.07	0.25 ± 0.06
69^0	-5.1 ± 3.0	11.9 ± 4.5	6.2 ± 2.9	0.59 ± 0.16	0.57 ± 0.12	0.59 ± 0.17
70^1	-7.0 ± 2.7	16.5 ± 3.2	9.2 ± 2.4	0.60 ± 0.09	0.67 ± 0.08	0.36 ± 0.09
71	-10.7 ± 0.4	11.2 ± 0.5	10.7 ± 0.4	0.22 ± 0.08	0.29 ± 0.11	0.16 ± 0.03
72	-2.6 ± 0.4	4.2 ± 0.6	3.0 ± 0.4	0.56 ± 0.06	0.70 ± 0.06	0.45 ± 0.07
73	0.3 ± 0.3	2.4 ± 0.3	1.8 ± 0.2	0.48 ± 0.08	0.64 ± 0.08	0.50 ± 0.08
74	-2.7 ± 0.4	4.2 ± 0.5	3.0 ± 0.4	0.56 ± 0.07	0.70 ± 0.06	0.44 ± 0.08
75	4.1 ± 0.4	5.1 ± 0.3	4.4 ± 0.3	0.23 ± 0.09	0.34 ± 0.11	0.29 ± 0.06
76	1.7 ± 0.7	5.3 ± 0.4	4.3 ± 0.4	0.32 ± 0.08	0.47 ± 0.11	0.47 ± 0.08

Table 1 Error metrics were calculate for each set of predictions, including root-mean-squared error (RMSE), average unsigned error (AUE), average signed error (ASE), Kendall's
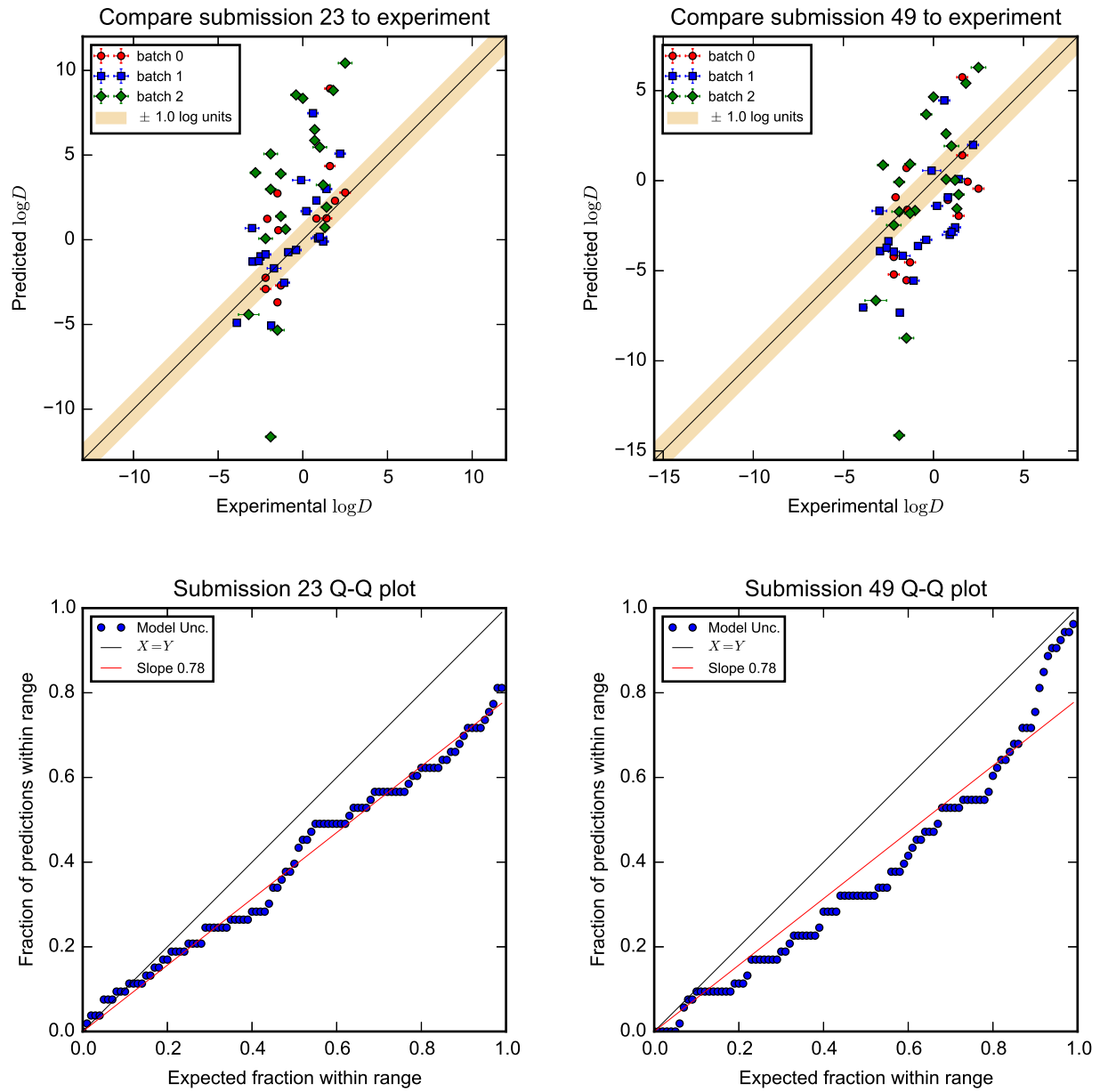
Fig. 2 These are examples of plots created for each set of predictions. They were chosen to try to represent the average submissions, those that were in the middle by most error metrics. a and b) comparison plots showing how predicted distribution coefficients compared to experiment for both groups. c,d) QQ Plots showing how their actual predictions were distributed compared to expectations given the model uncertainty.
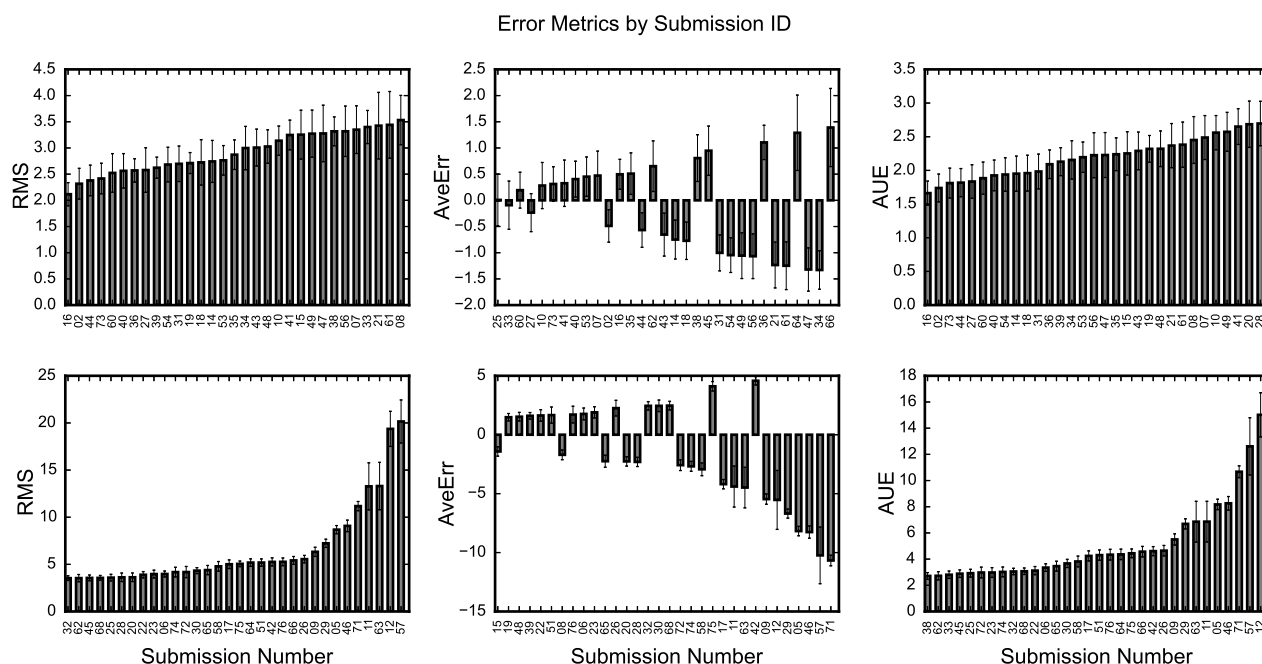
Fig. 3

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

5.2 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.

5.3 Compounds that were difficult to accurately predict

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

about 5-10 worst, I'm looking into if there are trends in number of tautomer or functional group similarities...

about 5-10 best, still looking for trends, we know 083, 074. 015 also did poorly amino acid? I don't know why that would cause more problems in general.
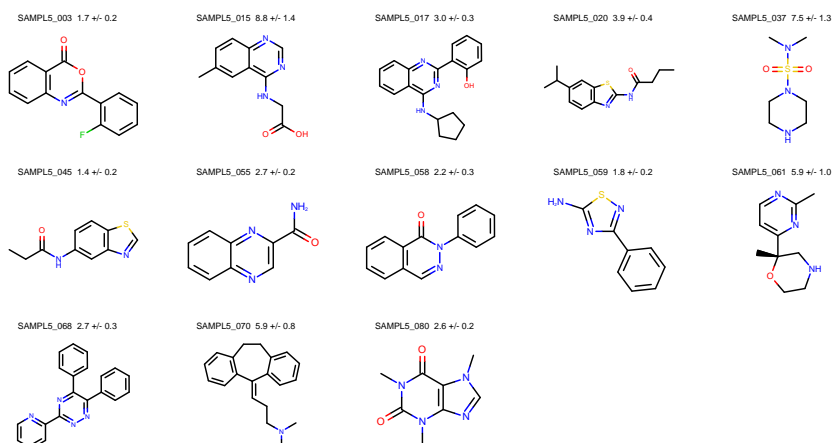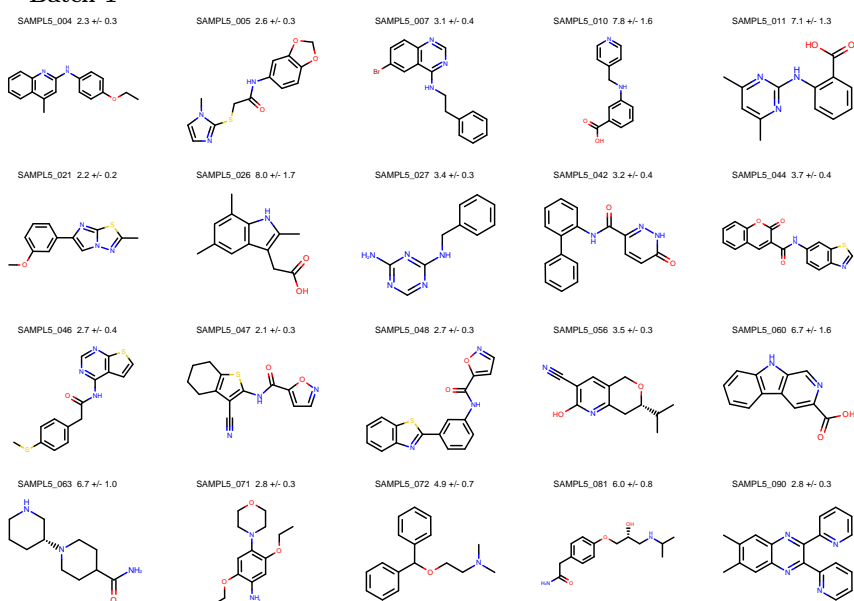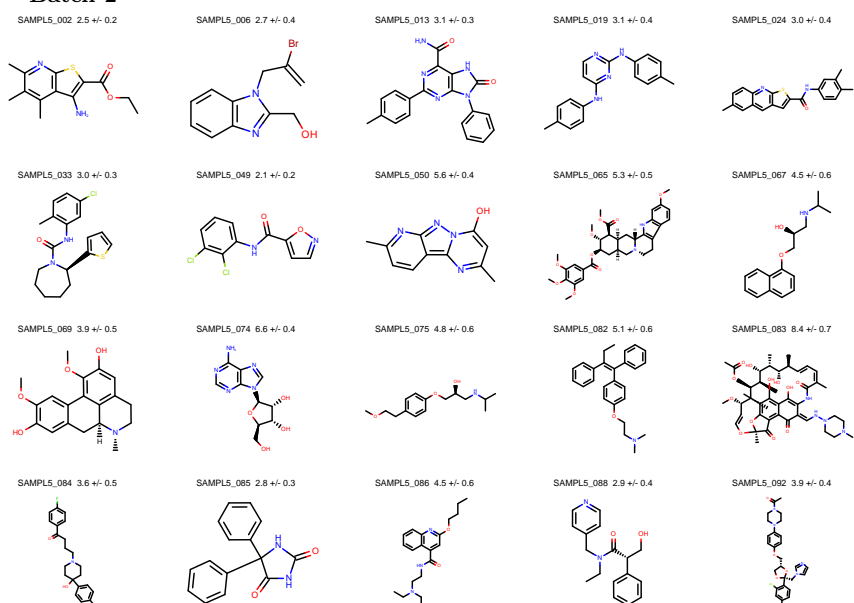
Batch 0

SAMPL5_003 1.7 +/- 0.2

SAMPL5_015 8.8 +/- 1.4

SAMPL5_017 3.0 +/- 0.3

SAMPL5_020 3.9 +/- 0.4

SAMPL5_037 7.5 +/- 1.3

SAMPL5_045 1.4 +/- 0.2

SAMPL5_055 2.7 +/- 0.2

SAMPL5_058 2.2 +/- 0.3

SAMPL5_059 1.8 +/- 0.2

SAMPL5_061 5.9 +/- 1.0

SAMPL5_068 2.7 +/- 0.3

SAMPL5_070 5.9 +/- 0.8

SAMPL5_080 2.6 +/- 0.2

Batch 1

SAMPL5_004 2.3 +/- 0.3

SAMPL5_005 2.6 +/- 0.3

SAMPL5_007 3.1 +/- 0.4

SAMPL5_010 7.8 +/- 1.6

SAMPL5_011 7.1 +/- 1.3

SAMPL5_021 2.2 +/- 0.2

SAMPL5_026 8.0 +/- 1.7

SAMPL5_027 3.4 +/- 0.3

SAMPL5_042 3.2 +/- 0.4

SAMPL5_044 3.7 +/- 0.4

SAMPL5_046 2.7 +/- 0.4

SAMPL5_047 2.1 +/- 0.3

SAMPL5_048 2.7 +/- 0.3

SAMPL5_056 3.5 +/- 0.3

SAMPL5_060 6.7 +/- 1.6

SAMPL5_063 6.7 +/- 1.0

SAMPL5_071 2.8 +/- 0.3

SAMPL5_072 4.9 +/- 0.7

SAMPL5_081 6.0 +/- 0.8

SAMPL5_090 2.8 +/- 0.3

Batch 2

SAMPL5_002 2.5 +/- 0.2

SAMPL5_006 2.7 +/- 0.4

SAMPL5_013 3.1 +/- 0.3

SAMPL5_019 3.1 +/- 0.4

SAMPL5_024 3.0 +/- 0.4

SAMPL5_033 3.0 +/- 0.3

SAMPL5_049 2.1 +/- 0.2

SAMPL5_050 5.6 +/- 0.4

SAMPL5_065 5.3 +/- 0.5

SAMPL5_067 4.5 +/- 0.6

SAMPL5_069 3.9 +/- 0.5

SAMPL5_074 6.6 +/- 0.4

SAMPL5_075 4.8 +/- 0.6

SAMPL5_082 5.1 +/- 0.6

SAMPL5_083 8.4 +/- 0.7

SAMPL5_084 3.6 +/- 0.5

SAMPL5_085 2.8 +/- 0.3

SAMPL5_086 4.5 +/- 0.6

SAMPL5_088 2.9 +/- 0.4

SAMPL5_092 3.9 +/- 0.4

Table 2 A complete list of compounds used in the SAMPL5, sorted by batch. The average unsigned error, reported in log units, was calculated with all predictions for that compound.
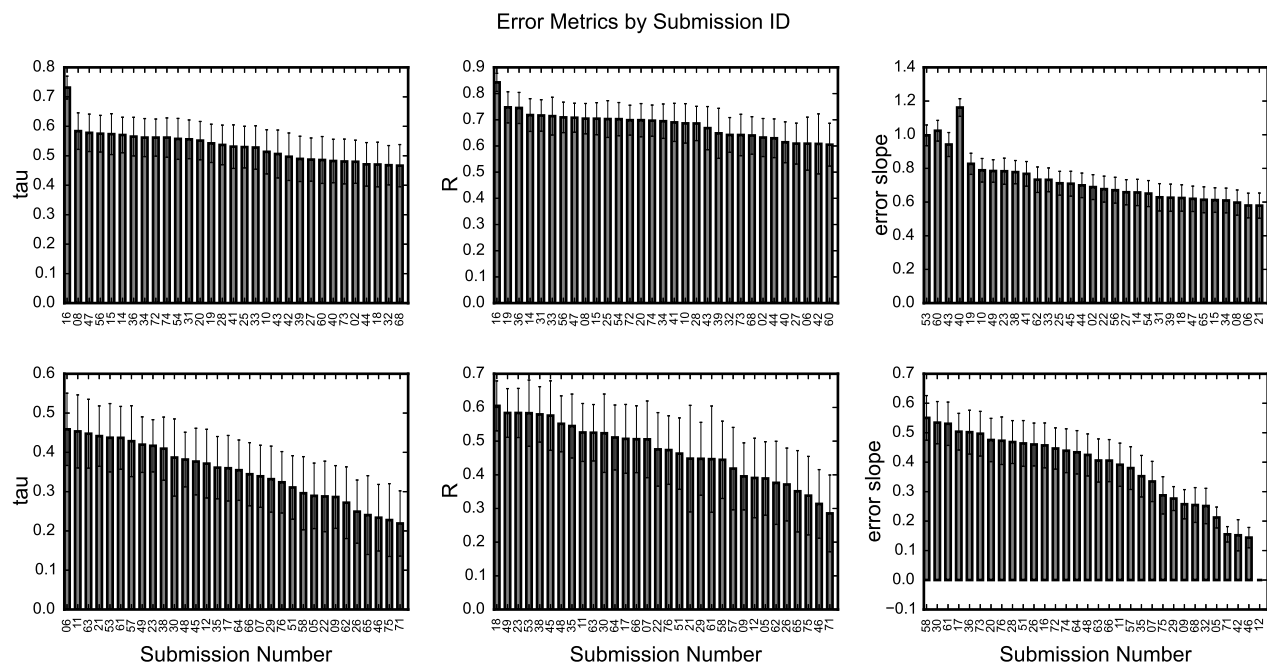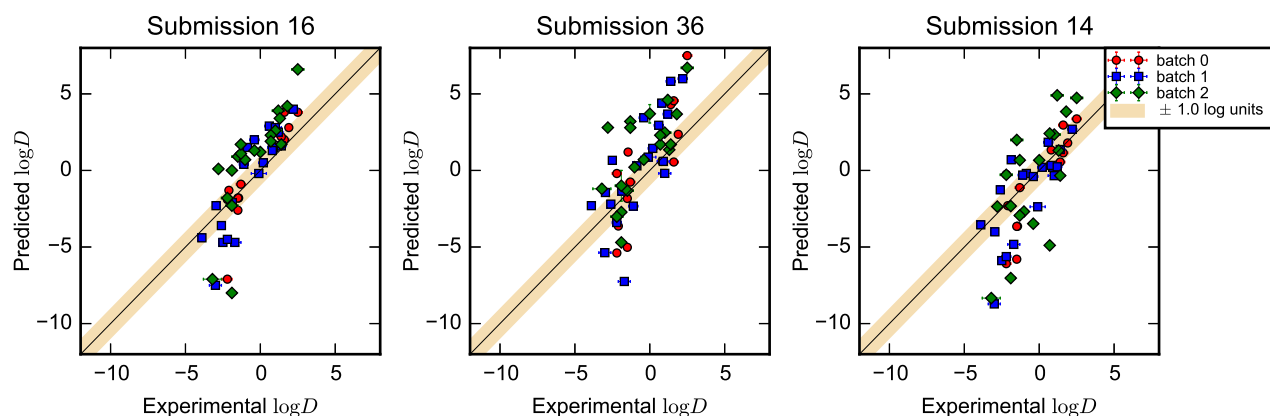
Error Metrics by Submission ID

Fig. 4

Fig. 5

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?

Fig. 6 Plots showing our predictions compared to experiment. a) submission 39 to SAMPL5, with no tautomer correction. b) distribution coefficient corrected from calculated partition coefficient based on pKas. c) distribution coefficient correct from calculated partition coefficient with state penalties

5.5 Mobley group prediction results

We submitted a set of blind predictions (39) to the challenge. Solvation free energies were calculated using GROMACS with GAFF and AM1-BCC charges. The initial set of predictions were partition coefficients, determined from the difference in solvation free energies without correcting for variation in tautomers. 39 was within the top 15 submissions for all error metrics.

After the challenge we explored how correcting for protonation states would have affected the our initial predictions. The first set of corrections involved calculating the pKa for each molecule using Schrödinger's Epik tool. Next, $\log D$ was calculated using the pKa and partition coefficient determined in submission 39 using eqn. For compounds with more than one pKa, the one which caused the largest change in $\log D$ was used, this would represent the most acidic proton leaving or most basic functional group becoming protonated. This correction showed a slight improvement by most error metrics (Table including a decrease in the average error (value) indicating less bias toward concentration.

For the next set of corrections, we used Schrödinger's Ligprep tool to calculate a state penalty, which gives the relative population of tautomers in water at a given pH. The state penalty was used to correct the concentration in the aqueous layer, according to eqn. State penalties improved predictions from the original partition coefficient coefficients and showed a slight improvement over the pKa corrections (Fig. 6). Both of these correction methods only adjust the concentration in the aqueous layer, however there may be tautomer affects that also affect the concentration in cyclohexane as well. There are a few molecules where the state penalty correction caused a significant bias for the concentration in the aqueous layer (SAMPL5_050). One explanation for these extreme examples is that the solute might have other neutral tautomers that would affect the concentration in cyclohexane, which we did not correct for.

5.6 Reanalysis of difficult tautomers

From our tautomer enumeration and discussions with other SAMPL5 participants it became clear that the provided SMILES string may not be the most popular tautomeric form of the molecule. If we could perfectly calculate solvation free energies and tautomer populations in both solvents, the starting tautomer should not effect the final calculated distribution coefficient. Our initial solvation free energy calculations used provided SMILES strings without any consideration of other tautomers. To explore how this may have affected our $\log D$ calculation, we decided to repeat a few solvation free energy cal-

	SAMPL5_050		SAMPL5_083	
	tautomer 1	tautomer 2	tautomer 1	tautomer 2
$\Delta G_{hydration}$	1 ± 1	1 ± 1	1 ± 1	1 ± 1
$\Delta G_{cyclohexane}$	1 ± 1	1 ± 1	1 ± 1	1 ± 1
$\log P_{cyc/wat}$	1 ± 1	1 ± 1	1 ± 1	1 ± 1
Correction	1 ± 1	1 ± 1	1 ± 1	1 ± 1
$\log D_{cyc/water}$	1 ± 1	1 ± 1	1 ± 1	1 ± 1
experimental $\log D$	1 ± 1		1 ± 1	

Table 3 Simulations with different tautomers...

culations with different tautomers. We repeated calculations with different tautomers of SAMPL5_050 and SAMPL5_083 that could be present in both the water and cyclohexane solutions. To explore how the tautomer used to calculate the solvation free energy might effect the estimate of a distribution coefficient.

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

5.7 Considering how solvent interactions could possibly affect results

6 Conclusions

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 community, 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

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 Schrödinger calculations all input files and results files for 'logP' calculations, tautomer redos, box size simulations example MDP and run scripts

References

1. H.J.C. Berendsen, D. Van Der Spoel, R. van Drunen, Comput. Phys. Commun. **91**(1-3), 43 (1995)
2. B. Hess, C. Kutzner, D. van der Spoel, E. Lindahl, J. Chem. Theory Comput. **4**(3), 435 (2008)
3. E. Lindahl, B. Hess, D. van der Spoel, J. Mol. Model. **7**(8), 306 (2001)
4. D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A.E. Mark, H.J.C. Berendsen, J. Comput. Chem. **26**(16), 1701 (2005)
5. S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M.R. Shirts, J.C. Smith, P.M. Kasson, D. van der Spoel, B. Hess, E. Lindahl, Bioinformatics (Oxford, England) **29**(7), 845 (2013)
6. S. Páll, M.J. Abraham, C. Kutzner, B. Hess, E. Lindahl, in *Solving Software Challenges for Exascale*, vol. 8759 (Springer International Publishing, Stockholm, Sweden, 2014), pp. 3–27
7. M.J. Abraham, T. Murtola, R. Schulz, S. Páll, J.C. Smith, B. Hess, E. Lindahl, SoftwareX **1-2**, 19 (2015)
8. J. Wang, R.M. Wolf, J.W. Caldwell, P.A. Kollman, D.A. Case, Journal of Computational Chemistry **25**(9), 1157 (2004)
9. A. Jakalian, B.L. Bush, D.B. Jack, C.I. Bayly, Journal of Computational Chemistry **21**(2), 132 (2000)
10. A. Jakalian, D.B. Jack, C.I. Bayly, Journal of Computational Chemistry **23**(16), 1623 (2002)
11. S. Liu, S. Cao, K. Hoang, K.L. Young, A.S. Paluch, D.L. Mobley, Journal of Chemical Theory and Computation **12**(4), 1930 (2016)
12. P.V. Klimovich, D.L. Mobley, Journal of Computer-Aided Molecular Design **24**(4), 307 (2010)
13. P.V. Klimovich, M.R. Shirts, D.L. Mobley, Journal of Computer-Aided Molecular Design **29**(5), 397 (2015)
14. S. Parameswaran, D.L. Mobley, Journal of Computer-Aided Molecular Design **28**(8), 825 (2014)
15. R. Wang, Y. Fu, L. Lai, Journal of Chemical Information and Modeling **37**(3), 615 (1997)
16. R. Wang, Y. Gao, L. Lai, Perspectives in Drug Discovery and Design **19**(1), 47 (2000)
17. A. Leo, C. Hansch, D. Elkins, Chemical Reviews **71**(6), 525 (1971)
18. D.L. Mobley, K.L. Wymer, N.M. Lim, J.P. Guthrie, Journal of Computer-Aided Molecular Design **28**(3), 135 (2014)