

UC Irvine

UC Irvine Electronic Theses and Dissertations

Title

Advancing physicochemical property predictions in computational drug discovery

Permalink

<https://escholarship.org/uc/item/40g7k7rd>

Author

Bergazin, Teresa Danielle

Publication Date

2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA,
IRVINE

Advancing physicochemical property predictions in computational drug discovery

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Pharmacological Sciences

by

Teresa Danielle Bergazin

Dissertation Committee:
Professor David L. Mobley, Chair
Assistant Professor Trina M. Norden-Krichmar
Professor Andrej Luptak

2022

Chapter 1 and 2 © 2021 Journal of Computer-Aided Molecular Design
Chapter 3 © 2021 Journal of Computer-Aided Molecular Design
All other materials © 2022 Teresa Danielle Bergazin

DEDICATION

Thank you to my family, who have always loved and supported me through everything I do. You mean the world to me.

TABLE OF CONTENTS

	Page
LIST OF FIGURES	vi
LIST OF TABLES	xviii
ACKNOWLEDGMENTS	xxiii
VITA	xxiv
ABSTRACT OF THE DISSERTATION	xxvi
1 Introduction	1
2 Assessing the accuracy of octanol-water partition coefficient predictions in the SAMPL6 Part II log P Challenge	3
2.1 Abstract	3
2.1.1 Abbreviations	4
2.2 Introduction	5
2.2.1 SAMPL Challenge History and Motivation	7
2.2.2 Common computational approaches for predicting log P	12
2.3 Challenge design and evaluation	19
2.3.1 Challenge structure	19
2.3.2 Evaluation approach	22
2.4 Methods for reference calculations	24
2.4.1 Physical Reference Calculations	24
2.4.2 Empirical Reference Calculations	26
2.4.3 Our null prediction method	27
2.5 Results and Discussion	27
2.5.1 Overview of challenge results	27
2.5.2 Lessons learned from physical reference calculations	42
2.5.3 Lessons learned from empirical reference calculations	53
2.5.4 Performance of reference methods on additional molecules	56
2.5.5 Take-away lessons from the SAMPL6 Challenge	60
2.5.6 Suggestions for the design of future challenges	63
2.6 Conclusion	65
2.7 Author Contributions	67

2.8	Acknowledgments	67
2.9	Supplementary Information	68
2.9.1	Overview of Supplementary Information	68
2.9.2	Code and Data Availability	69
2.9.3	Detailed methods	71
2.9.4	Supplementary Figures and tables	78
3	Evaluation of log P, pK_a, and log D predictions from the SAMPL7 blind challenge	88
3.1	Abstract	88
3.1.1	Keywords	89
3.1.2	Abbreviations	89
3.2	Introduction	90
3.2.1	Motivation for the log P and pK_a challenge	94
3.2.2	Historical SAMPL pK_a performance	95
3.2.3	Approaches to predicting small molecule pK_a 's	98
3.2.4	Approaches to predicting log P	103
3.3	Challenge design and evaluation	105
3.3.1	General challenge structure	105
3.3.2	log P challenge structure	107
3.3.3	pK_a challenge structure	109
3.3.4	Combining log P and pK_a predictions to estimate log D	113
3.3.5	Evaluation approach	114
3.4	Results and Discussion	116
3.4.1	Overview of log P challenge results	116
3.4.2	Overview of pK_a challenge results	124
3.4.3	Overview of log D challenge results	136
3.5	Conclusions	141
3.6	Code and Data Availability	147
3.7	Overview of Supplementary Information	148
3.8	Author Contributions	149
3.9	Acknowledgments	149
3.10	Disclaimers	150
3.11	Disclosures	150
3.12	Supplementary Information	151
3.12.1	Supplementary Figures and tables	151
4	Enhancing Water Sampling of Buried Binding Sites Using Nonequilibrium Candidate Monte Carlo	162
4.1	Abstract	162
4.1.1	Keywords	163
4.1.2	Abbreviations	163
4.2	Introduction	164
4.3	Methods	165
4.3.1	Implementation of NCMC/MD in BLUES	167

4.3.2	Translational water moves with BLUES	170
4.3.3	Comparing sampling efficiency using the number of force evaluations	172
4.3.4	Test cases and simulation details	174
4.4	Results and Discussion	177
4.5	Conclusions and Future Work	183
4.6	Code and Data Availability	185
4.7	Acknowledgments	185
4.8	Supplementary Information	186
4.8.1	Supplementary tables	186
5	Progress towards improving host-guest binding free energy calculations by refitting host force field parameters	189
5.1	Introduction	189
5.2	Methods	190
5.3	Results and Discussion	191
	Bibliography	201

LIST OF FIGURES

	Page
2.1 The desire to deconvolute the distinct sources of error contributing to the large errors observed in the SAMPL5 log D challenge motivated the separation of pK_a and log P challenges in SAMPL6. The SAMPL6 pK_a and log P challenges aim to evaluate protonation state predictions of small molecules in water and transfer free energy predictions between two solvents, isolating these prediction problems.	8
2.2 Structures of the 11 protein kinase inhibitor fragments used for the SAMPL6 log P Blind Prediction Challenge. These compounds are a subset of the SAMPL6 pK_a Challenge compound set [89] which were found to be tractable potentiometric measurements with sufficient solubility and pK_a values far from pH titration limits. Chemical identifiers of these molecules are available in Table 2.8 and experimental log P values are published [88]. Molecular structures in the figure were generated using OEDepict Toolkit [5].	20
2.3 Overall accuracy assessment for all methods participating in the SAMPL6 log P Challenge. Both root-mean squared error (RMSE) and mean absolute error (MAE) are shown, with error bars denoting 95% confidence intervals obtained by bootstrapping over challenge molecules. Submission IDs are summarized in Table 2.3. Submission IDs of the form <i>REF##</i> refer to non-blinded reference methods computed after the blind challenge submission deadline, and <i>NULL0</i> is the null prediction method; all others refer to blind, prospective predictions.	31
2.4 Overall correlation assessment for all methods participating SAMPL6 log P Challenge. Pearson’s R^2 and Kendall’s Rank Correlation Coefficient Tau (τ) are shown, with error bars denoting 95% confidence intervals obtained by bootstrapping over challenge molecules. Submission IDs are summarized in Table 2.3. Submission IDs of the form <i>REF##</i> refer to non-blinded reference methods computed after the blind challenge submission deadline, and <i>NULL0</i> is the null prediction method; all others refer to blind, prospective predictions. Overall, a large number and wide variety of methods have a statistically indistinguishable performance on ranking, in part because of the relatively small dynamic range of this set and because of the small size of the set. Roughly the top half of methods with Kendall’s Tau > 0.5 fall into this category. . . .	32

2.5	Performance statistics of physical methods. Physical methods are further classified into quantum chemical (QM) methods and molecular mechanics (MM) methods. RMSE and Kendall’s Rank Correlation Coefficient Tau are shown, with error bars denoting 95% confidence intervals obtained by bootstrapping over challenge molecules. Submission IDs are summarized in Table 2.3. Submission IDs of the form <i>REF##</i> refer to non-blinded reference methods computed after the blind challenge submission deadline; all others refer to blind, prospective predictions.	35
2.6	Predicted vs experimental value correlation plots of 8 best-performing methods and one representative average method. Dark and light green shaded areas indicate 0.5 and 1.0 units of error. Error bars indicate standard error of the mean of predicted and experimental values. Experimental log <i>P</i> SEM values are too small to be seen under the data points. EC_RISM_wet_P1w+1o method (<i>rdsnw</i>) was selected as the representative average method, as it is the method with the highest RMSE below the median.	38
2.7	Molecule-wise prediction error distribution plots show how variable the prediction accuracy was for individual molecules across all prediction methods. (A) MAE calculated for each molecule as an average of all methods shows relatively uniform MAE across the challenge set. SM14 and SM16 predictions were slightly more accurate than the rest. (B) MAE of each molecule broken out by method category shows that for each method category the most challenging molecules were different. Predictions of SM08, SM13, SM09, and SM12 log <i>P</i> values were significantly less accurate with Physical (MM) methods than the other method categories. For QM-based methods SM04 and SM02 were most challenging. Largest MAE for Empirical methods were observed for SM11 and SM15. (C) Error distribution for each SAMPL6 molecule overall prediction methods. It is interesting to note that most distributions are peaked near an error of zero, suggesting that perhaps a consensus model might outperform most individual models. However, SM15 is more significantly shifted away from zero than any other compound. SM08 has a significant tail showing probability of overestimated log <i>P</i> predictions by some methods. (D) Error distribution for each molecule calculated for only 7 methods from blind submissions that were determined to be consistently well-performing (<i>hmz0n</i> , <i>gmoq5</i> , <i>j8nwc</i> , <i>hdpuj</i> , <i>dqk4</i> , <i>vzgyt</i> , <i>qyzjx</i>).	40
2.8	Comparison of independent predictions that use seemingly identical methods (free energy calculations using GAFF and TIP3P water) shows significant systematic deviations between predictions for many compounds. Comparison of the calculated and experimental values for submissions <i>v2q0t</i> (InterX_GAFF_WET_OCTANOL), <i>6nmtd</i> (MD-AMBER-wetoct), <i>sqosi</i> (MD-AMBER-dryoct) and physical reference calculations <i>REF02</i> (YANK-GAFF-TIP3P-wet-oct) and <i>REF07</i> (YANK-GAFF-TIP3P-dry-oct). (A) compares calculations that used wet octanol, and (B) compares those that used dry octanol. Plots C to F show the methods compared to one another. The dark and light-shaded region indicates 0.5 and 1.0 units of error, respectively.	44

2.9	<p>Comparison of predictions that use free energy calculations using GAFF and TIP3P water show deviations between predictions for the challenge molecules and several alternative tautomers and resonance structures. Deviations seem to largely stem from differences in equilibration amount and choice of tautomer. A compares reference direct transfer free energy (DFE, <i>REF07</i>) and indirect solvation-based transfer free energy (IFE) protocols to experiment for the challenge provided resonance states of molecules and a couple of extra resonance states for SM14 and SM11, and extra tautomers for SM08. B compares the same exact tautomers for submission <i>sqosi</i> (MD-AMBER-dryoct) and the two reference protocols to experiment. Submission <i>sqosi</i> (MD-AMBER-dryoct) used different tautomers than the ones provided in the challenge. C-E compares the calculated log P between different methods using the same tautomers. All of the predicted values can be found in Table 2.5.</p>	47
2.10	<p>The prediction errors per molecule indicate some compounds were more difficult to predict than others for the reference calculations category. (A) MAE of each SAMPL6 molecule broken out by physical and empirical reference method category. (B) Error distribution for each molecule calculated for the reference methods. SM08 was the most difficult to predict for the physical reference calculations, due to our partial charge assignment procedure.</p>	50
2.11	<p>The tautomer and resonance structure choice resulted in discrepancies in the reference calculations. Shown here are calculated values for different input structures using the reference direct transfer free energy method. The uncertainties of the log P predictions were calculated as the standard error of the mean (SEM) of three replicate predictions. Structures labelled as SM08, SM11, and SM14 are based on input SMILES provided in SAMPL6 log P Challenge instructions. Three microstates shown for SM08 are different tautomers. SM08 (SM08_micro011) and SM08_micro010 are carboxylic acids, while SM08_micro008 is a carboxylate ion. SM08 (SM08_micro011) has a carbonyl group in the ring, while SM08_micro008 and SM08_micro010 have a hydroxyl in the ring. Structures pertaining to SM11 and SM14 are different resonance hybrids of the same tautomer (neutral microstate). Enumeration of all theoretically possible neutral tautomers of SAMPL6 molecules can be found in the SAMPL6 GitHub Repository (https://github.com/samplchallenges/SAMPL6/tree/master/physical_properties/pKa/microstates).</p>	55
2.12	<p>Structures of the 27 additional molecules that were included in follow-up assessment of the reference methods. These molecules were not included in the statistics overview.</p>	57

2.13	Distribution of reference method calculation errors by molecule on our extra set shows that a few of the molecules were more challenging than others. (A) MAE of each of the extra molecules broken out by physical and empirical reference method category. Majority of molecules have mean absolute errors below 1 log P unit for physical reference calculations. All of the mean absolute errors are well below 1 log P unit for empirical reference calculations. (B) Error distribution for each molecule calculated for the reference methods. A couple molecules have a significant tail showing probability of overestimated log P predictions.	58
2.14	Varying the amount of water in the octanol phase has no significant effect on the predicted log P in reference calculations, as discussed in section 2.5.2. Comparison of predicted log P values to the experimental values using wet (27% water) and dry octanol phases and the (A) GAFF and (B) SMIRNOFF force field, from non-blinded reference calculations performed for this paper, shows no statistically significant difference in performance of methodologies. Comparison of the calculated log P using dry and wet octanol phases for (C) the GAFF force field and (D) the SMIRNOFF force field shows a small systematic difference.	84
2.15	Shown here are the 2- and 3D structures of SM08_micro011 with the carboxylic acid in “anti” and “syn” conformation. The dihedral angle is indicated by the arrow around the carbon and oxygen atom. The calculated log P is included for comparison. The charges pertaining to each conformation are listed in Figure 2.13 and transition data is available in Figure 2.16	85
2.16	For the DFE method, the starting conformation impacts the number of C-O dihedral transitions for SM08_micro011, influencing sampling. Here is the transition data for the C-O dihedral in Figure 2.15, with charges listed in Table 2.13, for the DFE method (run in triplicate). In the “anti starting position” the torsion remains “anti” throughout the simulation, while the “syn starting position” allows transitions.	87
3.1	Structures of the 22 molecules used for the SAMPL7 physical property blind prediction challenge. Log of the partition coefficient between n-octanol and water was determined via potentiometric titrations using a Sirius T3 instrument. pK_a values were determined by potentiometric titrations using a Sirius T3 instrument. Log of the distribution coefficient between n-octanol and aqueous buffer at pH 7.4 were determined via potentiometric titrations using a Sirius T3 instrument, except for compounds SM27, SM28, SM30-SM34, SM36-SM39 which had log $D_{7.4}$ values determined via shake-flask assay. PAMPA assay data includes effective permeability, membrane retention, and log of the apparent permeability coefficient. Permeabilities for compounds SM33, SM35, and SM39 were not determined. Compounds SM35, SM36 and SM37 are single <i>cis</i> configuration isomers. All other compounds are not chiral.	108

- 3.2 **For each molecule in the SAMPL7 pK_a challenge we asked participants to predict the relative free energy between our selected neutral reference microstate and the rest of the enumerated microstates for that molecule.** In this case, we asked for the relative state free energy including the proton free energy, which could also be called the reaction free energy for the microstate transition which has the reference state as the reactant and the alternate state as the product. Using SM43 as an example, participants were asked to predict the relative free energy between SM43_micro000 (our selected neutral microstate highlighted in yellow) and all of the other enumerated microstates (SM43_micro001–SM43_micro005) for a total of 5 relative state free energies (ΔG_{BA} , ΔG_{CA} , ΔG_{DA} , ΔG_{EA} , ΔG_{FA}). Some transitions involved a change in a single protonation state (e.g. the D-A transition of Figure 3.2) or tautomer (e.g. the C-A transition of Figure 3.2). A few cases involved a change of multiple protons (e.g. the F-A transition of Figure 3.2). All transitions were defined as *away* from the neutral reference state. Distinct microstates are defined as all tautomers of each charge state. For each relative free energy prediction reported, participants also submitted the formal charge after transitioning from the selected neutral state to the other state. For example, the reported charge state after transitioning from SM43_micro000 to SM43_micro001 would be -1, SM43_micro000 to SM43_micro004 would be 0 (these are tautomers of each other), SM43_micro000 to SM43_micro005 would be +1, and SM43_micro000 to SM43_micro003 would be +2. 112
- 3.3 **Using the microstate probability to convert microscopic pK_a predictions to macroscopic pK_a 's with the titration method pK_a 's.** Blue and orange lines represent two states. Blue states have one more proton than the orange states, and thus a formal charge higher by +1. The blue state has one tautomer and the orange state has 3, denoted by the dashed lines. The solid lines are the ensemble averaged state probability for each group with a given charge. The crossing point between two ensemble lines is the macroscopic pK_a 116
- 3.4 **Overall accuracy assessment for all methods participating in the SAMPL7 log P challenge shows that many methods did not exhibit statistically significant differences in performance and there was no single clear winner; however, empirical methods tended to perform better in general.** Both root-mean-square error (RMSE) and mean absolute error (MAE) are shown, with error bars denoting 95% confidence intervals obtained by bootstrapping over challenge molecules. Empirical methods outperform the majority of the other methods. Methods that achieved a RMSE ≤ 1.0 log P units were mainly empirical based, and some were QM-based physical methods. Submitted methods are listed in Table 3.1. The submission *REF1 ChemAxon* [2] was a reference method included after the blind challenge submission deadline, and *NULL0 mean cLogP FDA* is the null prediction method; all others refer to blind predictions. 118

3.5	Predicted vs. experimental value correlation plots of 5 best performing methods and one representative average method in the SAMPL7 log <i>P</i> challenge.	
	Dark and light green shaded areas indicate 0.5 and 1.0 units of error. Error bars indicate standard error of the mean of predicted and experimental values. In some cases, log <i>P</i> SEM values are too small to be seen under the data points. The best-performing methods were made up of three empirical methods (<i>ClassicalGSG DB3</i> [52], <i>TFE MLR</i> [174], <i>Chemprop</i> [118]) and two QM-based physical methods (<i>COSMORS</i> [224], <i>TFE-NHLBI-TZVP-QM</i>). Details of the methods can be found in Section 3.4.1 and performance statistics are available in 3.2. Method <i>NES-1 (GAFF2/OPC3 G)</i> was selected as the representative average method, which has a median RMSE.	123
3.6	Molecule-wise prediction accuracy in the log <i>P</i> challenge point to isoxazoles as poorly predicted, especially by MM-based physical methods.	
	Molecules are labeled with their compound class as a reference. (A) The MAE calculated for each molecule as an average of all methods. (B) The MAE of each molecule separated by method category. (C) log <i>P</i> prediction error distribution for each molecule across all prediction methods.	125
3.7	Overall accuracy assessment for all methods participating in the SAMPL7 p<i>K</i>_a challenge shows that two methods, one a Physical (QM) method and one a QSPR/ML, performed better than other methods.	
	Both root-mean-square error (RMSE) and mean absolute error (MAE) are shown, with error bars denoting 95% confidence intervals obtained by bootstrapping over challenge molecules. <i>REF00_Chemaxon_Chemicalize</i> [2] is a reference method that was included after the blind challenge submission deadline, and all other method names refer to blind predictions. Methods are listed out in Table 3.3 and statistics calculated for all methods are available in Table 3.8.	128
3.8	Overall correlation assessment for all methods participating in the SAMPL7 p<i>K</i>_a challenge shows that one Physical (QM) method and one QSPR/ML reference method exhibited modestly better performance than others.	
	Pearson’s R^2 and Kendall’s Rank Correlation Coefficient Tau (τ) are shown, with error bars denoting 95% confidence intervals obtained by bootstrapping over challenge molecules. Submission methods are listed out in Table 3.3. <i>REF00_Chemaxon_Chemicalize</i> [2] is a reference method that was included after the blind challenge submission deadline, and all other method names refer to blind predictions. Most methods have a statistically indistinguishable performance on ranking, however, for R^2 , two methods (<i>EC_RISM</i> [209], <i>REF_Chemaxon_Chemicalize</i>), tend to have a greater ranking ability than the other methods. Evaluation statistics calculated for all methods are available in Table 3.8 of the Supplementary Information.	129

- 3.9 **Predicted vs. experimental value correlation plots of 2 best performing methods and one representative average method in the SAMPL7 pK_a challenge.** Dark and light green shaded areas indicate 0.5 and 1.0 units of error. Error bars indicate standard error of the mean of predicted and experimental values. Some SEM values are too small to be seen under the data points. Method *DFT_M05-2X_SMD* [66] was selected as the method with the median RMSE of all ranked methods analyzed in the challenge. Performance statistics of these methods is available in Table 3.4 . 132
- 3.10 **Molecule-wise prediction error distribution plots show the prediction accuracy for individual molecules across all prediction methods for the pK_a challenge.** Molecules are labeled with their compound class as a reference. **(A)** The MAE of each molecule separated by method category suggests the most challenging molecules were different for each method category. It is difficult to draw statistically significant conclusions where there are large overlapping confidence intervals. The QM+LEC method category appears to be less accurate for the majority of the molecules compared to the other method categories. QSPR/ML methods performed better for isoxazoles (SM41-SM43) and 1,2,3-triazoles (SM44-SM46) compared to the other two method categories. Physical QM-based methods performed poorly for acyl-sulfonamides (SM26 and SM25). **(B)** Error distribution for each molecule over all prediction methods. SM25 has the most spread in pK_a prediction error. 133
- 3.11 **Chemical transformations that lead to common sign disagreements among participants typically involve a protonated nitrogen in terminal nitrogen groups, 1,2,3-triazoles, and isoxazoles.** Shown are some chemical transformations that repeatedly show up as having large disagreement on the sign of the relative free energy prediction, as seen in Figure 3.13. 137
- 3.12 **The average relative microstate free energy predicted per microstate and the distribution across predictions in the SAMPL7 pK_a challenge show how varied predictions were.** Molecules are labeled with their compound class as a reference. **(A)** The average relative microstate free energy predicted per microstate. Error bars are the standard deviation of the relative microstate free energy predictions. A lower standard deviation indicates that predictions for a microstate generally agree, while a larger standard deviation means that predictions disagree. Predictions made for microstates such as SM25_micro001, SM26_micro002, SM28_micro001, SM43_micro003, SM46_micro003 widely disagree, while predictions for microstates such as SM26_micro004 are in agreement. **(B)** Distribution for each relative microstate free energy prediction over all prediction methods shows how prediction agreement among methods varied depending on the microstate. 138

3.13	The Shannon entropy (H) per microstate transition shows that participants disagree on many of the signs of the relative free energy predictions. Microstates with entropy values greater than 0 reflect increasing disagreement in the predicted sign. Microstates with an entropy of 0 are not shown here, but indicate that methods made predictions which had the same sign for the free energy change associated with a particular transition. About 44% of all microstates predictions disagreed with one another based on the sign, and the rest agreed. Roughly 5% of microstates strongly disagreed on the sign of predictions— meaning that predicted relative free energies were fairly evenly split between positive, neutral, and negative values. This indicates that these transitions were particularly challenging.	139
3.14	Structures of microstates where relative microstate free energy predictions disagree. Shown are some of the microstate transitions where participants predictions largely disagree with one another, based on Figure 3.12. The average relative free energy prediction (ΔG) along with the standard deviation are listed under each transition.	140
3.15	Overall accuracy assessment for log D estimation. Both root-mean-square error (RMSE) and mean absolute error (MAE) are shown, with error bars denoting 95% confidence intervals obtained by bootstrapping over challenge molecules. <i>REF00_ChemAxon</i> [2] is a reference method and <i>NULL0</i> is a null method that was included after the blind challenge submission deadline, and all other method names refer to blind predictions. Methods are listed out in Table 3.5 and statistics calculated for all methods are available in Table 3.9.	142
3.16	Predicted vs. experimental value correlation plots of all log D estimation methods in the SAMPL7 challenge. Dark and light green shaded areas indicate 0.5 and 1.0 units of error. Error bars indicate standard error of the mean of predicted and experimental values. Some SEM values are too small to be seen under the data points. Performance statistics of all methods is available in Table 3.9	143
3.17	log D values from a combination of the best pK_a and log P are typically superior. Shown is the RMSE in calculated log D values, with error bars denoting 95% confidence intervals from bootstrapping over challenge molecules. This plot is similar to Figure 3.4.3A, except it includes some additional pK_a and log P combinations (for log D estimation). Method <i>logP_experimental + EC_RISM</i> combines the experimental log P with the top performing pK_a method (based on RMSE). Method <i>logP_experimental + pKa_experimental</i> combines the experimental log P and pK_a value. Method <i>TFE MLR + EC_RISM</i> combines the best performing (based on RMSE) log P and pK_a methods. Method <i>TFE MLR + pKa_experimental</i> combines the best performing (based on RMSE) log P method with the experimental pK_a . Method <i>logP_experimental + DFT_M05-2X_SMD</i> combines the experimental log P with an average performing pK_a method. Method <i>NES-1 (GAFF2/OPC3) B + pKa_experimental</i> combines a log P method with average performance with the experimental pK_a . All other methods are the same as in Figure 3.4.3A.	144

- 3.18 **Distribution of molecular properties of the 22 compounds from the SAMPL7 physical property blind challenge.** (A) Histogram of $\log P$ measurements collected with Sirius T3 instrument. The ticks along the x-axis indicate the individual values. Compounds have experimental $\log P$ values in the range of 0.58-2.96. (B) Histogram of $\text{p}K_{\text{a}}$ measurements collected with Sirius T3 instrument.. Eight compounds have measured $\text{p}K_{\text{a}}$'s in the range of 4.49–6.62 and eleven in the range 9.58- 11.93. Two compounds are included here as having $\text{p}K_{\text{a}}$'s of 12, but actually had experimental values greater than 12, and were therefore outside of the experimental detection range. (C) Histogram of $\log D$ measurements between n-octanol and aqueous buffer at pH 7.4 were determined via potentiometric titrations using a Sirius T3 instrument, except for compounds SM27, SM28, SM30-SM34, SM36-SM39 which had $\log D_{7.4}$ values determined via shake-flask assay. $\log D$ measurements ranged from -0.87-2.96. (D) Histogram of molecular weights calculated for the compounds in the SAMPL7 dataset. The molecular weight ranged from 227-365 Da. (E) Histogram of the number of rotatable bonds in each molecule. The number of rotatable bonds in challenge molecules ranged from 3-6. . . 152
- 3.19 **Overall correlation assessment for all methods participating in the SAMPL7 $\log P$ challenge show that the uncertainty of each correlation statistic is quite high, not allowing a true ranking based on correlation.** Pearson's R^2 and Kendall's Rank Correlation Coefficient Tau (τ) are shown, with error bars denoting 95% confidence intervals obtained by bootstrapping over challenge molecules. Submitted methods are listed in Table 3.1. The submission *REF1 ChemAxon* was a reference method included after the blind challenge submission deadline, and *NULL0 mean cLogP FDA* is the null prediction method; all others refer to blind predictions. Most methods have a statistically indistinguishable performance on ranking because of the small dynamic range of the dataset. Evaluation statistics calculated for all methods are available in Table 3.6 of the Supplementary Information. . . 154
- 3.20 **Compound classes and structures of the molecules in the SAMPL7 physical property challenge.** SMILES of the compounds are in Table 3.20. 160

4.1	Molecular interactions between atoms are turned off and on during a NCMC move to translate a water molecule. In this cartoon, water molecules are represented here by red and white spheres for the oxygen and hydrogen atoms. The black-filled water represents a fully interacting water molecule that has been selected to be moved. Gray-filled water represents intermediate levels of interaction and white-filled represents the fully non-interacting water molecule. A) The water molecule (in black) is fully interacting with its surrounding environment, and in this case, other water molecules. B) The water's interactions are partially off, allowing the other water molecules to slightly relax. C) The water's interactions are fully turned off. D) The water is randomly translated to somewhere else in the system (indicated by a black arrow) with its interactions remaining off. E) The water's interactions are partially turned on and the propagation steps of NCMC allow relaxation of the translated water and its surroundings to resolve clashes. F) At the end of the NCMC protocol, the water molecule is once again in the fully interacting state and in a new location. This entire process comprises a proposed NCMC move, which is accepted or rejected based on the nonequilibrium work done in this process, and then followed by conventional MD.	166
4.2	Example of a user-defined radius that covers a particular area of interest. Here, the MUP-1 protein-ligand system is shown. The radius used (indicated by the black dashed line) defines a sphere around a user-selected atom (represented by a blue star) in the system, such as an atom inside the binding site of a protein.	170
4.3	Workflow of a BLUES iteration with translational water hopping move proposals. Before any water is translated to a new location, the user first selects an atom and picks a radius defining a sphere encompassing an area of interest around the position of the atom and BLUES identifies all the water and protein residues in the system. Afterward, BLUES goes through a number of BLUES iterations n number of times, where each BLUES iteration is as shown inside the dashed box. A schematic of the NCMC move process is shown in Figure 4.1.	171
4.4	Systems used to test the ability of BLUES (NCMC/MD) water hopping to allow the exchange of water. (A) A C_{60} buckyball with a single trapped water molecule. (B) The buried hydration site of the MUP-1 protein with a bound ligand. (C) The hydration site of the HSP90 protein bound to a ligand. The protein-ligand systems have internal water(s) (indicated by the black dashed line) that do not easily exchange with bulk.	173
4.5	Impermeable graphene sheets divide a box into separate regions with initially different densities, testing the ability of water hopping moves to equilibrate the density. (A) The water box system with dividing graphene sheets. (B) Shown here are the water densities between the two sheets (blue) and outside the sheets (orange). The densities in the two regions reach equilibrium and stabilize with this approach, serving to validate our implementation.	177

4.6	Increasing the amount of NCMC steps increases the rate of water transfer from bulk to the internal hydration site in MUP-1. Ten replicate simulations with different random seed numbers were run for each NCMC step value. All of the BLUES simulations were run for 10,000 BLUES iterations, with each iteration consisting of a certain number of steps of NCMC and MD. The different colors indicate various amounts of NCMC steps used. The success rate is equivalent to the ratio of the number of replicate simulations where the MUP-1 site (Figure 4.4.B) has been hydrated relative to the total number of replicate simulations. (A) shows that using a lower NCMC step amount increases the number of BLUES iterations for the cavity to become hydrated, such as 1,250 (green) and 2,500 (orange) NCMC steps. The inset, (B), zooms in on the success rate at low iteration number and shows that increasing the amount of NCMC steps decreases the number of iterations needed. 5,000 (blue) NCMC steps needed a little more than 400 BLUES iterations to hydrate the cavity and 30,000 (pink) NCMC steps needed no more than 250 BLUES iterations to hydrate the cavity.	178
5.1	The structure of the tetra-endomethyl OctaAcid (TEMOA) host.	192
5.2	The structure of the five guests used in binding free energy prediction.	192
5.3	Average prediction error in binding free energy calculations using the new host parameters (Bespoke) and original host parameters (Parsley) show there isn't a statistically significant difference between the two methods.	193
5.4	Predicted vs experimental value correlation plots of the two methods.	194
5.5	The BespokeFit host is less likely to be in the collapsed stance compared to the Parsley host when the ligand is not in the host cavity	195
5.6	Shown is the timeseries of the distance ratio of the distances of the diagonal upper cavity phenyl groups for the BespokeFit-parameterized host (blue) and the Parsley-parameterized host (orange) when guest number # 1 is not in the binding pocket.	196
5.7	Shown is the timeseries of the distance ratio of the distances of the diagonal upper cavity phenyl groups for the BespokeFit-parameterized host (blue) and the Parsley-parameterized host (orange) when guest number # 2 is not in the binding pocket.	197
5.8	Shown is the timeseries of the distance ratio of the distances of the diagonal upper cavity phenyl groups for the BespokeFit-parameterized host (blue) and the Parsley-parameterized host (orange) when guest number # 3 is not in the binding pocket.	198
5.9	Shown is the timeseries of the distance ratio of the distances of the diagonal upper cavity phenyl groups for the BespokeFit-parameterized host (blue) and the Parsley-parameterized host (orange) when guest number # 4 is not in the binding pocket.	199

5.10 Shown is the timeseries of the distance ratio of the distances of the diagonal upper cavity phenyl groups for the BespokeFit-parameterized host (blue) and the Parsley-parameterized host (orange) when guest number # 5 is not in the binding pocket. 200

LIST OF TABLES

	Page
2.1 Methods used as reference calculations for the MM-based physical methods category. Please see Section 2.9.3 in the Supplementary Information for detailed description of physical reference methods.	26
2.2 Methods used as reference calculations for the empirical log P prediction category. Please see section 2.9.3 in the Supplementary Information for a detailed description of empirical methods.	27
2.3 Submission IDs, names, category, and type for all the log P participant and reference calculation submissions. Submission IDs of methods are listed in the ID column. Reference calculations are labeled as <i>REF##</i> . The method name column lists the names provided by each participant in the submission file. The “type” column indicates if submission was or a post-deadline reference calculation, denoted by “Blind” or “Reference” respectively. The table is ordered by increasing RMSE from top to down and left to right, although many consecutively listed methods are statistically indistinguishable. All calculated error statistics are available in Tables 2.10 and 2.11.	29
2.4 Eight consistently well-performing prediction methods based on consistent ranking within the Top 20 according to various statistical metrics. Submissions were ranked according to RMSE, MAE, R^2 , and τ . Many top methods were found to be statistically indistinguishable considering uncertainties of error metrics. Moreover, sorting of methods was influenced significantly by the choice of metric chosen. We assessed top 20 methods according the each metric to determine which methods are always among the top 20 according to all four statistical metrics chosen. A set of consistently well-performing methods were determined: Four QM-based and four empirical methods. Seven of these methods are blind submissions of SAMPL6 Challenge, and one of them (<i>REF13</i>) is a non-blind reference calculation. Performance statistics are provided as mean and 95% confidence intervals.	37

2.5	Predicted log P values of free energy calculations of methods using GAFF, TIP3P water, and dry octanol. The methods listed are the reference direct transfer free energy (DFE) protocol, reference indirect solvation-based transfer free energy (IFE) protocol and submission <i>sqosi</i> (MD-AMBER-dryoct). Details of the two reference protocols can be found in Section 2.9.3. log P predictions for multiple tautomers (SM08) and resonance structures (SM11 and SM14) are listed, when available. The experimental values are provided for comparison. The same experimental log P values are stated for multiple tautomers or resonance structures. Potentiometric log P measurements do not provide information about the identity or populations of tautomers.	54
2.6	Statistics of the physical and empirical reference method predictions on the extra test of molecules. Methods were ranked according to increasing RMSE in this table. Performance statistics of MAE, R^2 , and Kendall’s Tau are also provided. Mean and 95% confidence intervals of all statistics are presented.	59
2.7	Method details of log P predictions with MM-based physical methods. Force fields, water models, and octanol phase choice are reported. A dry octanol phase indicates the octanol phase was treated as consisting of pure octanol. A wet octanol phase indicates the octanol phase was treated as a mixture of octanol and water. RMSE and Kendall’s Tau values are reported as mean and 95% confidence intervals. A CSV version of this table can be found in <i>SAMPL6-supplementary-documents.tar.gz</i>	79
2.8	SMILES and InChI identifiers of SAMPL6 log P Challenge molecules. Experimental log P values can be found in a separate paper reporting measurements [88]. A CSV version of this table can be found in <i>SAMPL6-supplementary-documents.tar.gz</i>	80
2.9	SMILES and InChI identifiers of extra molecules included in the evaluation of reference methods. A CSV version of this table can be found in <i>SAMPL6-supplementary-documents.tar.gz</i> . Experimental log P values can be found in a separate paper reporting measurements [202] and in machine readable format in https://github.com/samplchallenges/SAMPL6/blob/master/physical_properties/logP/analysis_of_extra_molecules/logP_experimental_values.csv	81
2.10	Evaluation statistics calculated for all methods. Methods are represented via their SAMPL6 submission IDs which can be cross referenced with Table 2.3 for method details. There are six error metrics reported: the root-mean-squared error (RMSE), mean absolute error (MAE), mean (signed) error (ME), coefficient of determination (R^2), linear regression slope (m), and Kendall’s Rank Correlation Coefficient (τ). This table is ranked by increasing RMSE. A CSV version of this table can be found in <i>SAMPL6-supplementary-documents.tar.gz</i>	82

2.11	[Table 2.10 continued.] Evaluation statistics calculated for all methods. Methods are represented via their SAMPL6 submission IDs which can be cross referenced with Table 2.3 for method details. There are six error metrics reported: the root-mean-squared error (RMSE), mean absolute error (MAE), mean (signed) error (ME), coefficient of determination (R^2), linear regression slope (m), and Kendall’s Rank Correlation Coefficient (τ). This table is ranked by increasing RMSE. A CSV version of this table can be found in <i>SAMPL6-supplementary-documents.tar.gz</i>	83
2.12	Comparison of force field parameters of the TIP3P, TIP3P-FB and OPC water models.	85
2.13	Comparison of the charges assigned to the syn and anti conformation of SM08_micro011 in the DFE protocol.	86
3.1	Method names, category, and submission type for all the log P calculation submissions. The “submission type” column indicates if submission was a blind submission (denoted by “Blind”) or a post-deadline reference or null calculation (denoted by “Reference”). The table is ordered from lowest to highest RMSE, although many consecutively listed methods are statistically indistinguishable. All calculated error statistics are available in Table 3.6. . .	119
3.2	Five consistently well-performing log P prediction methods based on consistent ranking within the top 10 according to various statistical metrics. Submissions were ranked according to RMSE, MAE, R^2 , and Kendall’s Tau. Many top methods were found to be statistically indistinguishable when considering the uncertainties of their error metrics. Additionally, the sorting of methods was significantly influenced by the metric that was chosen. We determined which ranked log P prediction methods were consistently the best according to all four chosen statistical metrics by assessing the top 10 methods according to each metric. A set of five consistently well-performing methods were determined— three empirical methods and two QM-based physical methods. Performance statistics are provided as mean and 95% confidence intervals. Correlation plots of the best performing methods and one average method is shown in Figure 3.5. Additional statistics are available in Table 3.6.	122
3.3	Method names, category, and submission type for all the pK_a submissions. The “submission type” column indicates if submission was a blind submission (denoted by “Blind”) or a post-deadline reference calculation (denoted by “Reference”). The table is ordered from lowest to highest RMSE, although many consecutively listed methods are statistically indistinguishable. All calculated error statistics are available in Table 3.8.	126

3.4	Two consistently well-performing pK_a prediction methods based on consistent ranking within the top three according to various statistical metrics. Ranked submissions were ranked/ordered according to RMSE, MAE, R^2 , and Kendall's Tau. Many methods were found to be statistically indistinguishable when considering the uncertainties of their error metrics. Additionally, the sorting of methods was significantly influenced by the metric that was chosen. We determined which methods are repeatedly among the top two according to all four chosen statistical metrics by assessing the top three methods according to each metric. Two QM-based methods consistently performed better than others. Performance statistics are provided as mean and 95% confidence intervals. All statistics for all methods are in Table 3.8.	131
3.5	Method names, category, and submission type for all the log D estimations. Method names are based off the submitted pK_a and log P method names, with the log P method name listed first followed by "+" and then the pK_a method name. The "submission type" column indicates if submission was a blind submission (denoted by "Blind") or a post-deadline reference calculation (denoted by "Reference"). All calculated error statistics are available in Table 3.9.	136
3.6	Evaluation statistics calculated for all methods in the log P challenge. Submitted predictions are represented by their method name. There are six error metrics reported: the root-mean-squared error (RMSE), mean absolute error (MAE), mean (signed) error (ME), coefficient of determination (R^2), linear regression slope (m), and Kendall's Rank Correlation Coefficient (τ), and error slope (ES). The mean and 95% confidence intervals of each statistic is presented. This table is ranked by increasing RMSE.	153
3.7	Details of MM-based physical methods in the log P prediction challenge. Force fields, water models, and octanol phase choice are reported. A dry octanol phase indicates the octanol phase was composed of only octanol. A wet octanol phase indicates the octanol phase was treated as a mixture of octanol and water. RMSE, MAE, R^2 , and Kendall's Tau values are reported as mean and 95% confidence intervals.	155
3.8	Evaluation statistics calculated for all methods in the pK_a challenge. Submitted predictions are represented by their method name. There are six error metrics reported: the root-mean-squared error (RMSE), mean absolute error (MAE), mean (signed) error (ME), coefficient of determination (R^2), linear regression slope (m), Kendall's Rank Correlation Coefficient (τ), and error slope (ES). The mean and 95% confidence intervals of each statistic is presented. This table is ranked by increasing RMSE.	156

3.9	Evaluation statistics calculated for all log D estimates. Predictions are represented a name based on method names participants submitted to the and log P challenges. There are six error metrics reported: the root-mean-squared error (RMSE), mean absolute error (MAE), mean (signed) error (ME), coefficient of determination (R^2), linear regression slope (m), Kendall's Rank Correlation Coefficient (τ), and error slope (ES). The mean and 95% confidence intervals of each statistic is presented. This table is ranked by increasing RMSE.	157
3.10	Additional info for microscopic pK_a predictions.	158
3.11	SMILES and compound class of SAMPL7 physical property challenge molecules. A view of the compounds and their classes can be found in Figure 3.20.	159
3.12	Number of states per charge state for the microstates used in the SAMPL7 pK_a challenge. The total number of microstates (protomers and tautomers) is listed. Some of the molecules have up to 6 microstates, while others have only 2.	161
4.1	Increasing the number of NCMC steps generally increases the acceptance rate of all moves in the MUP-1 protein-ligand system. Here is the average acceptance rate of all BLUES moves, the average number of force evaluations across 10-12 replicates for the buried cavity in the MUP-1 system to become hydrated, and the average wallclock time in hours for BLUES to hydrate MUP-1. Each simulation was run for 10,000 BLUES iterations, where each iteration consisted of a single NCMC move (consisting of n NCMC steps) and 1,000 MD steps.	179
4.2	Acceptance ratios of the replicate simulations at different NCMC step amounts for the MUP-1 system. Each simulation was run for 10,000 BLUES iterations. The number of NCMC steps were varied from 1,250 to 30,000 steps and the number of MD steps was 1,000 steps in all cases	187
4.3	Acceptance ratios of all attempted moves for each replicate simulation of the HSP90 protein-ligand system. Shown is the acceptance ratio for four replicate simulations. It took an average of 1693 BLUES iterations to hydrate the cavity.	188
4.4	Shown here are the average acceptance rate of all BLUES moves, the average number of force evaluations across 4 replicates for the buried cavity in the HSP90 system to become hydrated, and the average wallclock time in hours for BLUES to hydrate HSP90. It took an average of 1693 BLUES iterations to hydrate the HSP90 cavity, and each BLUES iteration consisted of a single NCMC move (consisting of 2,500 NCMC steps) and 1,000 MD steps.}	188

ACKNOWLEDGMENTS

I acknowledge and appreciate support from the National Institutes of Health (1R01GM124270-01A1), the National Science Foundation (CHE 1352608), and the Intel/ACM SIGHPC Graduate Fellowship.

VITA

Teresa Danielle Bergazin

EDUCATION

Doctor of Philosophy in Pharmacological Sciences **2022**
University of California – Irvine *Irvine, CA*

Masters of Science in Pharmacological Sciences **2020**
University of California – Irvine *Irvine, CA*

Bachelor of Arts in Chemistry **2016**
California Lutheran University *Thousand Oaks, CA*

RESEARCH EXPERIENCE

Graduate Researcher **June 2017–May 2022**
University of California – Irvine *Irvine, California*

Computational Drug Discovery Intern **May 2021–Oct 2021**
Genentech, Inc. *South San Francisco, California*

TEACHING EXPERIENCE

Teaching Assistant: Pharmacology **2017**
University of California – Irvine *Irvine, CA*

Teaching Assistant: Medicinal Chemistry Lab **2021**
University of California – Irvine *Irvine, CA*

PUBLICATIONS

1. Işık, M.*; **Bergazin, T.D.***; Fox, T.; Rizzi, A.; Chodera, J. D.; Mobley, D. L. *Assessing the accuracy of octanol-water partition coefficient predictions in the SAMPL6 Part II log P Challenge* Journal of Computer-Aided Molecular Design, 34(4), 335–370. 2020. * – Denotes equal contribution.
2. **Bergazin, T.D.**; Ben-Shalom, I. Y.; Lim, N. M.; Gill, S. C.; Gilson, M. K.; Mobley, D. L. *Enhancing water sampling of buried binding sites using nonequilibrium candidate Monte Carlo.* Journal of Computer Aided Molecular Design, 35(2), 167–177. 2020.
3. **Bergazin, T. D.**; Tielker, N.; Zhang, Y.; Mao, J.; Gunner, M. R.; Francisco, K.; Ballatore, C.; Kast, S. M.; Mobley, D. L. *Evaluation of log P, pKa, and log D predictions*

from the SAMPL7 blind challenge. Journal of Computer Aided Molecular Design, 35(7), 771–802. 2021.

AWARDS

Computational and Data Science Fellowship	Intel and ACM SIGHPC
PyData Los Angeles Scholarship	PyData
Pharmaceutical Sciences Travel Grant	UC-Irvine
Associated Graduate Student Travel Grant	UC-Irvine
Carl Storm Fellowship	Gordon Research Group

PROFESSIONAL MEMBERSHIPS & OTHER EXPERIENCES

Iota Sigma Phi – Professional Development Chair	UC-Irvine, CA
Southern Ca. Area Group for Informatics and Modeling Meetups	San Diego, CA
Agile for Any Project Workshop	UC-Irvine, CA
Intro to Linux on the HPC	UC-Irvine, CA
PyData Los Angeles at the University of Southern California	Los Angeles, CA
Molecular Operating Environment Workshop	CCG, San Diego, CA
Association for Computing Machinery Member	ACM group
San Diego Supercomputer Center Summer Institute	UC-San Diego, CA
Computational Chemistry Gordon Research Conference	West Dover, Vermont
Computational Chemistry Gordon Research Seminar	West Dover, Vermont
Scalable and Reproducible Structural Bioinformatics Workshop – Application of Big Data Technology	UC-San Diego, CA
MD data analysis, Markov modeling + PyEMMA workshop	UC-San Diego, CA
Competitive Edge Summer Research Program	UC-Irvine, CA
Iota Sigma Phi – Calcium Chapter, Women Chemists Honors Society	UC-Irvine, CA

ABSTRACT OF THE DISSERTATION

Advancing physicochemical property predictions in computational drug discovery

By

Teresa Danielle Bergazin

Doctor of Philosophy in Pharmacological Sciences

University of California, Irvine, 2022

Professor David L. Mobley, Chair

Computer-aided drug design aims to guide the discovery of compounds with optimal pharmaceutical properties. Computational tools can evaluate large libraries of virtual molecules to help prioritize new compounds to synthesize and test. Properties such as protein-ligand binding affinity and physicochemical properties are of interest. To learn how reliable computational models are, it's necessary to evaluate the prediction accuracy of physicochemical property prediction. In Chapter 2, I describe my work in testing the accuracy of free energy calculations through partition coefficient predictions. In Chapter 3, I assess the accuracy of pK_a and partitioning predictions in a physical property prediction challenge. Additionally, I present work in which I developed and/or applied computational chemistry tools. In Chapter 4, I discuss my work on enhancing the sampling of water rearrangements through the extension of a hybrid simulation method. In Chapter 5, I describe work towards improving host-guest binding free energy calculations by refitting host force field parameters.

Chapter 1

Introduction

Drug discovery and development is a long and time-consuming process that costs billions of dollars. On average, it takes 10 years for a drug to be developed and approved for prescription, and can cost up to several billion dollars to bring a new drug to the market. Millions to billions of dollars are allocated to the preclinical studies where researchers search for a molecule that acts against a target of interest with sufficient affinity and drug-like properties.

In the lead optimization stage of drug discovery, initial “hits” that bind to a target receptor are found and then modified to improve their physicochemical properties. This process requires synthesis and experimentation, which can take months, and in some cases, even years. Computational methods can be used in advance of synthesis to narrow down the number of leads that need to be experimentally tested, saving thousands of dollars per compound in synthesis costs and months of work.

Various computational methods are used to guide molecular design and find new potential drugs and targets. Some examples of these methods are molecular dynamics simulations, free energy calculations, and virtual screening. Although computational chemistry techniques are

widely used in industry and academia, there is still a need for improvement.

Here, I first present my work on two community-wide blind challenges in which I benchmark the accuracy of physicochemical predictions. Secondly, I present work on developing a method for enhanced water sampling. Lastly, I describe work applying a tool to enhance guest-host binding calculations.

In Chapter 2, I describe my work in testing the accuracy of free energy calculations through partition coefficient predictions. I report on the results of the SAMPL6 log P blind challenge for 11 molecules and my reference calculation procedure. In Chapter 3, I assess the accuracy of pK_a and partitioning predictions in the SAMPL7 physical property prediction challenge. In Chapter 4, I present my work on enhancing the sampling of water rearrangements through a hybrid method that combines nonequilibrium candidate Monte Carlo simulations and molecular dynamics. In Chapter 5, I describe work towards improving host-guest binding free energy calculations by refitting host force field parameters.

Chapter 2

Assessing the accuracy of octanol-water partition coefficient predictions in the SAMPL6 Part II log P Challenge

Mehtap Işık*, Teresa Danielle Bergazin*, Thomas Fox, Andrea Rizzi, John D. Chodera, and David L. Mobley.

* – Denotes equal contribution.

Journal of Computer-Aided Molecular Design volume 34, pages 335–370 (2020)

doi: 10.1007/s10822-020-00295-0

Publication Date (Web): February 27, 2020

2.1 Abstract

The SAMPL Challenges aim to focus the biomolecular and physical modeling community on issues that limit the accuracy of predictive modeling of protein-ligand binding for ratio-

nal drug design. In the SAMPL5 log D Challenge, designed to benchmark the accuracy of methods for predicting drug-like small molecule transfer free energies from aqueous to nonpolar phases, participants found it difficult to make accurate predictions due to the complexity of protonation state issues. In the SAMPL6 log P Challenge, we asked participants to make blind predictions of the octanol-water partition coefficients of neutral species of 11 compounds and assessed how well these methods performed absent the complication of protonation state effects. This challenge builds on the SAMPL6 pK_a Challenge, which asked participants to predict pK_a values of a superset of the compounds considered in this log P challenge. Blind prediction sets of 91 prediction methods were collected from 27 research groups, spanning a variety of quantum mechanics (QM) or molecular mechanics (MM)-based physical methods, knowledge-based empirical methods, and mixed approaches. There was a 50% increase in the number of participating groups and a 20% increase in the number of submissions compared to the SAMPL5 log D Challenge. Overall, the accuracy of octanol-water log P predictions in SAMPL6 Challenge was higher than cyclohexane-water log D predictions in SAMPL5, likely because modeling only the neutral species was necessary for log P and several categories of method benefited from the vast amounts of experimental octanol-water log P data. There were many highly accurate methods: 10 diverse methods achieved RMSE less than 0.5 log P units. These included QM-based methods, empirical methods, and mixed methods with physical modeling supported with empirical corrections. A comparison of physical modeling methods showed that QM-based methods outperformed MM-based methods. The average RMSE of the most accurate five MM-based, QM-based, empirical, and mixed approach methods based on RMSE were 0.92 ± 0.13 , 0.48 ± 0.06 , 0.47 ± 0.05 , and 0.50 ± 0.06 , respectively.

2.1.1 Abbreviations

SAMPL Statistical Assessment of the Modeling of Proteins and Ligands

log P \log_{10} of the organic solvent-water partition coefficient (K_{ow}) of neutral species

log D \log_{10} of organic solvent-water distribution coefficient (D_{ow})

p K_a $-\log_{10}$ of the acid dissociation equilibrium constant

SEM Standard error of the mean

RMSE Root mean squared error

MAE Mean absolute error

τ Kendall's rank correlation coefficient (Tau)

R² Coefficient of determination (R-Squared)

QM Quantum Mechanics

MM Molecular Mechanics

2.2 Introduction

The development of computational biomolecular modeling methodologies is motivated by the goal of enabling quantitative molecular design, prediction of properties and biomolecular interactions, and achieving a detailed understanding of mechanisms (chemical and biological) via computational predictions. While many approaches are available for making such predictions, methods often suffer from poor or unpredictable performance, ultimately limiting their predictive power. It is often difficult to know which method would give the most accurate predictions for a target system without extensive evaluation of methods. However, such extensive comparative evaluations are infrequent and difficult to perform, partly because no single group has expertise in or access to all relevant methods and also because of the scarcity of blind experimental data sets that would allow prospective evaluations. In addition, many

publications which report method comparisons for a target system constructs these studies with the intention of highlighting the success of a method being developed.

The SAMPL (Statistical Assessment of the Modeling of Proteins and Ligands) Challenges [<http://sAMPLchallenges.github.io>] provide a forum to test and compare methods with the following goals:

1. Determine prospective predictive power rather than accuracy in retrospective tests.
2. Allow a head to head comparison of a wide variety of methods on the same data.

Regular SAMPL challenges focus attention on modeling areas that need improvement, and sometimes revisit key test systems, providing a crowdsourcing mechanism to drive progress. Systems are carefully selected to create challenges of gradually increasing complexity spanning between prediction objectives that are tractable and that are understood to be slightly beyond the capabilities of contemporary methods. So far, most frequent SAMPL challenges have been on solvation and binding systems. Iterated blind prediction challenges have played a key role in driving innovations in the prediction of physical properties and binding. Here we report on a SAMPL6 log P Challenge on octanol-water partition coefficients, treating molecules resembling fragments of kinase inhibitors. This is a follow-on to the earlier SAMPL6 pK_a Challenge which included the same compounds.

The partition coefficient describes the equilibrium concentration ratio of the neutral state of a substance between two phases:

$$\log P = \log_{10} K_{ow} = \log_{10} \frac{[\text{unionized solute}]_{\text{octanol}}}{[\text{unionized solute}]_{\text{water}}} \tag{2.1}$$

The log P challenge examines how well we model transfer free energy of molecules between

different solvent environments in the absence of any complications coming from predicting protonation states and pK_a values. Assessing $\log P$ prediction accuracy also allows evaluating methods for modeling protein-ligand affinities in terms of how well they capture solvation effects.

2.2.1 SAMPL Challenge History and Motivation

The SAMPL blind challenges aim to focus the field of quantitative biomolecular modeling on major issues that limit the accuracy of protein-ligand binding prediction. Companion exercises such as the Drug Design Data Resource (D3R) blind challenges aim to assess the current accuracy of biomolecular modeling methods in predicting bound ligand poses and affinities on real drug discovery project data. D3R blind challenges serve as an accurate barometer for accuracy. However, due to the conflation of multiple accuracy-limiting problems in these complex test systems it is difficult to derive clear insights into how to make further progress towards better accuracy.

Instead, SAMPL seeks to isolate and focus attention on individual accuracy-limiting issues. We aim to field blind challenges just at the limit of tractability in order to identify underlying sources of error and help overcome these challenges. Working on similar model systems or the same target with new blinded datasets in multiple iterations of prediction challenges maximize our ability to learn from successes and failures. Often, these challenges focus on physical properties of high relevance to drug discovery in their own right, such as partition or distribution coefficients critical to the development of potent, selective, and bioavailable compounds.

The partition coefficient ($\log P$) and the distribution coefficient ($\log D$) are driven by the free energy of transfer from an aqueous to a nonpolar phase. Transfer free energy of only neutral species are considered for $\log P$, whereas both neutral and ionizable species con-

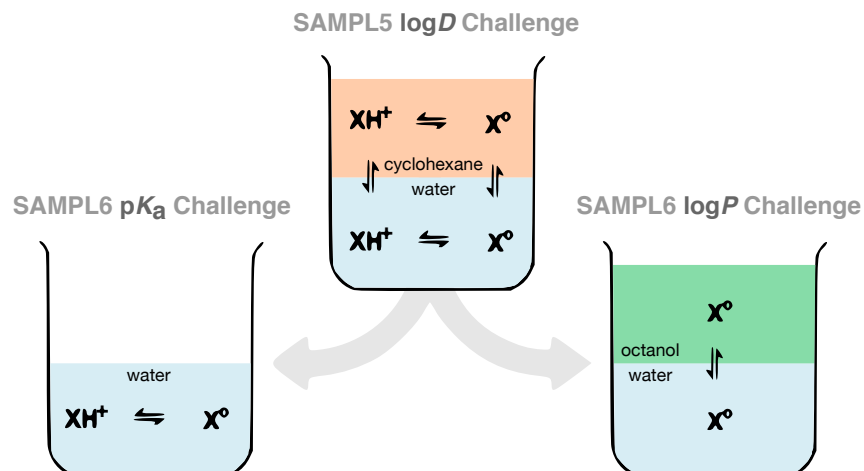


Figure 2.1: **The desire to deconvolute the distinct sources of error contributing to the large errors observed in the SAMPL5 log D challenge motivated the separation of pK_a and log P challenges in SAMPL6.** The SAMPL6 pK_a and log P challenges aim to evaluate protonation state predictions of small molecules in water and transfer free energy predictions between two solvents, isolating these prediction problems.

tribute to log D . Such solute partitioning models are a simple proxy for the transfer free energy of a drug-like molecule to a relatively hydrophobic receptor binding pocket, in the absence of specific interactions. Protein-ligand binding equilibrium is analogous to partitioning of a small molecule between two environments: protein binding site and aqueous phase. Methods that employ thermodynamic cycles, such as free energy calculations, can therefore use similar strategies for calculating binding affinities and partition coefficients, and given the similarity in technique and environment, we might expect the accuracy on log P and log D may be related to the accuracy expected from binding calculations, or at least a lower bound for the error these techniques might make in more complex protein-ligand binding phenomena. Evaluating log P or log D predictions makes it far easier to probe the accuracy of computational tools used to model protein-ligand interactions and to identify sources of error to be corrected. For physical modeling approaches, evaluation of partition coefficient predictions comes with the additional advantage of separating force field accuracy from protonation state modeling challenges.

The SAMPL5 log D challenge uncovered surprisingly large modeling errors Hydration free energies formed the basis of several previous SAMPL challenges, but distribution coefficients (log D) capture many of the same physical effects—namely, solvation in the respective solvents—and thus replaced hydration free energies in SAMPL5 [191, 19]. This choice was also driven due to a lack of ongoing experimental work with the potential to generate new hydration free energy data for blind challenges. Octanol-water log D is also a property relevant to drug discovery, often used as a surrogate for lipophilicity, further justifying its choice for a SAMPL challenge. The SAMPL5 log D Challenge allowed decoupled evaluation of small molecule solvation models (captured by the transfer free energy between environments) from other issues, such as the sampling of slow receptor conformational degrees of freedom. This blind challenge generated considerable insight into the importance of various physical effects [19, 191]; see the SAMPL5 special issue (<https://link.springer.com/journal/10822/30/11/page/1>) for more details.

The SAMPL5 log D Challenge used cyclohexane as an apolar solvent, partly to further simplify this challenge by avoiding some complexities of octanol. In particular, log D is typically measured using water-saturated octanol for the nonaqueous phase, which can give rise to several challenges in modeling accuracy such as a heterogeneous environment with potentially micelle-like bubbles [110, 29, 38, 129], resulting in relatively slow solute transitions between environments [29, 48]. The precise water content of wet octanol is unknown, as it is affected by environmental conditions such as temperature as well as the presence of solutes, the organic molecule of interest, and salts (added to control pH and ionic strength). Inverse micelles transiently formed in wet octanol create spatial heterogeneity and can have long correlation times in molecular dynamics simulations, potentially presenting a challenge to modern simulation methods [110, 29, 38, 129], resulting in relatively slow solute transitions between environments [29, 48].

Performance in the SAMPL5 log D Challenge was much poorer than the organizers initially

expected—and than would have been predicted based on past accuracy in hydration free energy predictions—and highlighted the difficulty of accurately accounting for protonation state effects [19]. In many SAMPL5 submissions, many participants treated distribution coefficients ($\log D$) as if they were asked to predict partition coefficients ($\log P$). The difference between $\log D$ (reflects the transfer free energy at a given pH including the effects of accessing all equilibrium protonation states of the solute in each phase) and $\log P$ (reflects aqueous-to-apolar phase transfer free energies of the neutral species only) proved particularly important. In some cases, other effects like the presence of a small amount of water in cyclohexane may also have played a role.

Because the SAMPL5 $\log D$ Challenge highlighted the difficulty in correctly predicting transfer free energies involving protonation states (the best methods obtained an RMSE of 2.5 log units [19]), the SAMPL6 Challenge aimed to further subdivide the assessment of modeling accuracy into two challenges: A small-molecule pK_a prediction challenge [89] and a $\log P$ challenge. The SAMPL6 pK_a Challenge asked participants to predict microscopic and macroscopic acid dissociation constants ($pK_{a,s}$) of 24 small organic molecules and concluded in early 2018. Details of the challenge are maintained on the GitHub repository (<https://github.com/samplchallenges/SAMPL6/>). pK_a prediction proved to be difficult. A large number of methods showed RMSE in the range of 1-2 pK_a units, with only a handful achieving less than 1 pK_a unit. These results were in line with expectations from the SAMPL5 Challenge about protonation state predictions being one of the major sources of error for $\log D$. But the present challenge allows us delve deeper into modeling the solvation of neutral species and focus on $\log P$.

The SAMPL6 $\log P$ Challenge focused on small molecules resembling kinase inhibitor fragments By measuring the $\log P$ of a series of compounds resembling fragments of kinase inhibitors—a subset of those used in the SAMPL6 pK_a Prediction Challenge—we

sought to assess the limitations of force field accuracy in modeling transfer free energies of drug-like molecules in binding-like processes. This time, the challenge featured octanol as the apolar medium to assess whether wet octanol presented as much of a problem as was previously suspected. Participants are asked to predict the partition coefficient ($\log P$) of the neutral species between octanol and water phases. Here we focus on different aspects of the challenge, particularly the staging, analysis, results, and lessons learned. Experimental work for collecting the $\log P$ values are described elsewhere [88]. One of the goals of this challenge is to encourage prediction of model uncertainties (an estimate of the inaccuracy with which your model predicts the physical property), since the ability to tell when methods will be successful or not would be very useful for increasing the application potential and impact of computational methods.

The SAMPL challenges aim to advance predictive quantitative models The SAMPL challenges have a key focus on lessons learned. In principle, they are a challenge or competition, but we see it as far more important to learn how to improve accuracy than to announce the top-performing methods. To aid in learning as much as possible, this overview paper provides an overall assessment of performance and some analysis of the relative performance of different methods in different categories, provides some insights into lessons we have learned (and/or other participants have learned). Additionally, this work presents our own reference calculations which provide points of comparison for participants (some relatively standard and some more recent, especially in the physical category) and also allow us to provide some additional lessons learned. The data, from all participants and all reference calculations, is made freely available (see Section 2.9.2) to allow others to compare methods retrospectively and dig into additional lessons learned.

2.2.2 Common computational approaches for predicting $\log P$

Many methods have been developed to predict octanol-water $\log P$ values of small organic molecules including physical modeling (QM and MM-based methods) and knowledge-based empirical prediction approaches (atom-contribution approaches and QSPR). There are also $\log P$ prediction methods that combine the strengths of physical and empirical approaches. Here, we briefly highlight some of the major ideas and background behind physical and empirical $\log P$ prediction methods.

Physical modeling approaches for predicting $\log P$

Physical approaches begin with a detailed atomistic model of the solute and its conformation and attempt to estimate partitioning behavior directly from that. Details depend on the approach employed.

1.2.1.1. Quantum mechanical (QM) approaches for predicting $\log P$ QM approaches to solvation modeling utilize numerical solution of the Schrödinger equation to estimate solvation free energies (and thereby partitioning) directly from first principles. There are a number of approaches for these calculations, and discussing them is outside the scope of this work. However, it is important to note that direct solution of the underlying equations, especially when coupled with dynamics, becomes impractical for large systems such as molecules in solution. So, several approximations must be made before such approaches can be applied to estimating phase transfer free energies. These typical approximations include assuming the solute has one or a small number of dominant conformations in each phase being considered, and using an implicit solvent model to represent the solvent. The basis set and level of theory can be important choices and can significantly affect accuracy of calculated values. Additionally, protonation or tautomerization state selected as an input can

also introduce errors. With QM approaches possible protonation states and tautomers can be evaluated to find the lowest energy state in each solvent. However, if these estimates are erroneous, any errors will propagate into the final transfer free energy and $\log P$ predictions.

Implicit solvent models can be used, in the context of the present SAMPL, both to represent water and octanol. Such models are often parameterized—sometimes highly so—based on experimental solvation free energy data. This means that such models perform well for solvents (and solute chemistries) where solvation free energy data is abundant (as in the present challenge) but are often less successful when far less training data is available. In this respect, QM methods, by virtue of the solvent model, have some degree of overlap with the empirical methods discussed further below.

Several solvent models are particularly common, and in the present challenge two were particularly commonly employed. One was Marenich, Cramer and Truhlar’s SMD solvation model [135], which derives its electrostatics from the widely used IEF-PCM model and was empirically trained on various solutes/solvents utilizing a total of some 2821 different solvation data points. This has been employed in various SAMPL challenges in the past in the context of calculation of hydration free energies, including the earliest SAMPL challenges [134, 183]. Others in the Cramer-Truhlar series of solvent models were also used, including the 2012 SM12 solvation model, which is based on the generalized born (GB) approximation [137]. Another set of submissions also used the reference interaction site model (RISM) integral equation approach, discussed further below.

The COSMO-RS solvation model is another method utilized in this context which covers a particularly broad range of solvents, typically quite well [127, 106, 104, 103, 107]. In the present challenge, a “Cosmoquick” variant was also applied and falls into the “Mixed” method category, as it utilizes additional empirical adjustments. The COSMO-RS implementation of COSMOtherm takes into account conformational effects to some extent; the chemical potential in each phase is computed using the Boltzmann weights of a fixed set of conformers.

In general, while choice of solvation model can be a major factor impacting QM approaches, the neglect of conformational changes means these approaches typically (though not always) neglect any possibility of significant change of conformation on transfer between phases and they simply estimate solvation by the difference in (estimated) solvation free energies for each phase of a fixed conformation. Additionally, solute entropy is often neglected, assuming the single-conformation solvation free energy plays the primary role in driving partitioning between phases. In addition to directly estimating solvation, QM approaches can also be used to drive the selection of the gas- or solution-phase tautomer, and thus can be used to drive the choice of inputs for MM approaches discussed further below.

Integral equation-based approaches

Integral equation approaches provide an alternate approach to solvation modeling (for both water and non-water solvents) and have been applied in SAMPL challenges within both the MM and QM frameworks [211, 207, 128]. In this particular challenge, however, the employed approaches were entirely QM, and utilized the reference interaction site model (RISM) approach [210, 24, 111]. Additionally, as noted above, the IEF-PCM model used by the SMD solvation model (discussed above) is also an integral equation approach. Practical implementation details mean that RISM approaches typically have one to a few adjustable parameters (e.g. four [207]) which are empirically tuned to experimental solvation free energies, in contrast to the SMD and SM-n series of solvation models which tend to have a larger number of adjustable parameters and thus require larger training sets. In this particular SAMPL challenge, RISM participation was limited to embedded cluster EC-RISM methods [109, 211, 210], which combine RISM with a quantum mechanical treatment of the solute.

1.2.1.2. Molecular mechanics (MM) approaches for predicting $\log P$ MM approaches to computing solvation and partition free energies (and thus $\log P$ values), as

typically applied in SAMPL, use a force field or energy model which gives the energy (and, usually, forces) in a system as a function of the atomic positions. These models include all-atom fixed charge additive force fields, as well as polarizable force fields. Such approaches typically (though not always) are applied in a dynamical framework, integrating the equations of motion to solve for the time evolution of the system, though Monte Carlo approaches are also possible.

MM-based methods are typically coupled with free energy calculations to estimate partitioning. Often, these are so-called alchemical methods which utilize a non-physical thermodynamic cycle to estimate transfer between phases, though pulling-based techniques which directly model phase transfer are in principle possible [47, 40]. Such free energy methods allow detailed all-atom modeling of each phase, and compute the full free energy of the system, in principle (in the limit of adequate sampling) providing the correct free energy difference given the choice of energy model (“force field”). However, adequate sampling can sometimes prove difficult to achieve.

Key additional limitations facing MM approaches are the accuracy of the force field, the fact that protonation state/tautomer is generally selected as an input and held fixed (meaning that incorrect assignment or multiple relevant states can introduce significant errors), and timescale—simulations only capture motions that are faster than simulation timescale. However, these approaches *do* capture conformational changes on phase transfer, as long as such changes occur on timescales faster than the simulation timescale.

Empirical log P predictions

Due to the importance of accurate log P predictions, ranging from pharmaceutical sciences to environmental hazard assessment, a large number of empirical models to predict this property have been developed and reviewed [164, 133, 45]. An important characteristic of

many of these methods is that they are very fast, so even large virtual libraries of molecules can be characterized.

In general, two main methodologies can be distinguished: group- or atom-contribution approaches, also called additive group methods, and quantitative structure-property relationship (QSPR) methods.

1.2.2.1 Atom- and group-contribution approaches Atom contribution methods, pioneered by Crippen in the late 1980s [71, 72], are the easiest to understand conceptually. These assume that each atom contributes a specific amount to the solvation free energy and that these contributions to $\log P$ are additive. Using a potentially large number of different atom types (typically in the order of 50-100), the $\log P$ is the sum of the individual atom types times the number of their occurrences in the molecule:

$$\log P = \sum_{i=1}^n n_i a_i \quad (2.2)$$

A number of $\log P$ calculation programs are based on this philosophy, including AlogP [73], AlogP98 [73], and moe_SlogP [225].

The assumption of independent atomic contributions fails for compounds with complex aromatic systems or stronger electronic effects. Thus correction factors and contributions from neighboring atoms were introduced to account for these shortcomings (e.g. in XlogP [222, 223, 39] and SlogP [225]).

In contrast, in group contribution approaches, $\log P$ is calculated as a sum of group contributions, usually augmented by correction terms that take into account intramolecular

interactions. Thus, the basic equation is

$$\log P = \sum_{i=1}^n a_i f_i + \sum_{j=1}^m b_j F_j \quad (2.3)$$

where the first term describes the contribution of the fragments f_i (each occurring a_i times), the second term gives the contributions of the correction factors F_j occurring b_j times in the compound. Group contribution approaches assume that the details of the electronic or intermolecular structure can be better modeled with whole fragments. However, this breaks down when molecular features are not covered in the training set. Prominent examples of group contribution approaches include clogP [122, 121, 119, 192], KlogP [108], ACD/logP [175] and KowWIN [142].

clogP is probably one of the most widely used $\log P$ calculation programs [122, 121, 119]. clogP relies on fragment values derived from measured data of simple molecules, e.g., carbon and hydrogen fragment constants were derived from measured values for hydrogen, methane, and ethane. For more complex hydrocarbons, correction factors were defined to minimize the difference to the experimental values. These can be structural correction factors taking into account bond order, bond topology (ring/chain/branched) or interaction factors taking into account topological proximity of certain functional groups, electronic effects through π -bonds, or special ortho-effects.

1.2.2.2 QSPR approaches Quantitative structure-property relationships (QSPR) provide an entirely different category of approaches. In QSPR methods, a property of a compound is calculated from molecular features that are encoded by so-called molecular descriptors. Often, these theoretical molecular descriptors are classified as 0D-descriptors (constitutional descriptors, only based on the structural formula), 1D-descriptors (i.e. list of structural fragments, fingerprints), 2D-descriptors (based on the connection table, topological descriptors), and 3D-descriptors (based on the three-dimensional structure of the

compound, thus conformation-dependent). Sometimes, this classification is extended to 4D-descriptors, which are derived from molecular interaction fields (e.g., GRID, CoMFA fields).

Over the years, a large number of descriptors have been suggested, with varying degrees of interpretability. Following the selection of descriptors, a regression model that relates the descriptors to the molecular property is derived by fitting the individual contributions of the descriptors to a dataset of experimental data; both linear and nonlinear fitting is possible. Various machine learning approaches such as random forest models, artificial neural network models, etc. also belong to this category. Consequently, a large number of estimators of this type have been proposed; some of the more well-known ones include MlogP [157] and VlogP [77].

Expectations from different prediction approaches

Octanol-water log P literature data abounds, impacting our expectations. Given this abundance of data, in contrast to cyclohexane-water log D data, e.g., for the SAMPL5 log D Challenge, we expected higher accuracy here. Some sources of public octanol-water log P values include DrugBank [226], ChemSpider [173], PubChem, the NCI CACTUS databases [4, 3], and SRC's PHYSPROP Database [6].

Our expectation was that empirical knowledge-based and other trained methods (implicit solvent QM, mixed methods) would outperform other methods in the present challenge as they are impacted directly by the availability of octanol-water data. Methods well trained to experimental octanol-water partitioning data should typically result in higher accuracy, if fitting is done well. The abundance of octanol-water data may also provide empirical and mixed approaches with an advantage over physical modeling methods. Current molecular mechanics-based methods and other methods not trained to experimental log P data ought to do worse in this challenge. Performance of strictly physical modeling based prediction

methods might generalize better across other solvent types where training data is scarce, but that will not be tested by this challenge. In principle, molecular mechanics-based methods could also be fitted using octanol-water data as one of the targets for force field optimization, but present force fields have not made broad use of this data in fitting. Thus, top methods are expected to be from empirical knowledge-based, QM-based approaches and combination of QM-based and empirical approaches because of training data availability. These categories are broken out separately for analysis.

2.3 Challenge design and evaluation

2.3.1 Challenge structure

The SAMPL6 Part II Challenge was conducted as a blind prediction challenge on predicting octanol-water partition coefficients of 11 small molecules that resemble fragments of kinase inhibitors. The challenge molecule set was composed of small molecules with limited flexibility (less than 5 non-terminal rotatable bonds) and covers limited chemical diversity. There are six 4-aminoquinazolines, two benzimidazoles, one pyrazolo[3,4-d]pyrimidine, one pyridine, one 2-oxoquinoline substructure containing compounds with $\log P$ values in the range of 1.95–4.09. Information on experimental data collection is presented elsewhere [88].

The dataset composition was announced several months before the challenge including details of the measurement technique (potentiometric $\log P$ measurement, at room temperature, using water saturated octanol phase, and ionic strength-adjusted water with 0.15 M KCl [88]), but not the identity of the small molecules. The instructions and the molecule set were released at the challenge start date (Nov 1, 2018), and then submissions were accepted until March 22, 2019.

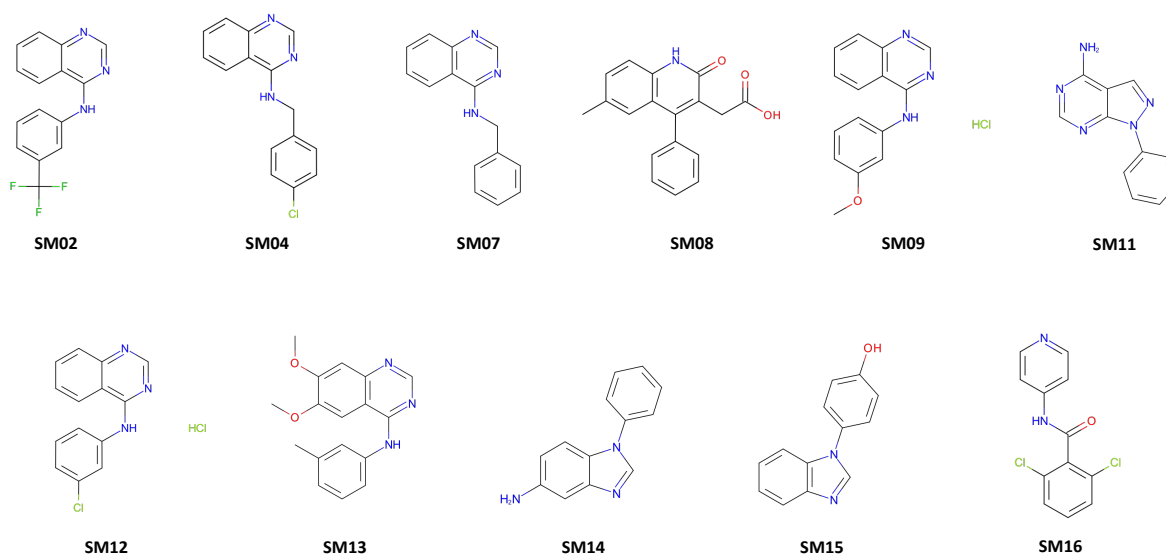


Figure 2.2: **Structures of the 11 protein kinase inhibitor fragments used for the SAMPL6 log P Blind Prediction Challenge.** These compounds are a subset of the SAMPL6 pK_a Challenge compound set [89] which were found to be tractable potentiometric measurements with sufficient solubility and pK_a values far from pH titration limits. Chemical identifiers of these molecules are available in Table 2.8 and experimental log P values are published [88]. Molecular structures in the figure were generated using OEDepict Toolkit [5].

Following the conclusion of the blind challenge, the experimental data was made public on Mar 25, 2019 and results are first discussed in a virtual workshop (on May 16, 2019) [151] then later in an in person workshop (Joint D3R/SAMPL Workshop, San Diego, Aug 22-23, 2019). The purpose of the virtual workshop was to go over a preliminary evaluation of results, begin considering analysis and lessons learned, and nucleate opportunities for follow up and additional discussion. Part of the goal was to facilitate discussion so that participants can work together to maximize lessons learned in the lead up to an in-person workshop and special issue of a journal. The SAMPL6 log P Virtual Workshop video [151] and presentation slides [152] are available, as are organizer presentation slides from the joint D3R/SAMPL Workshop 2019 [147, 94] on the SAMPL Community Zenodo page (<https://zenodo.org/communities/sampl/>).

A machine-readable submission file format was specified for blind submissions. Participants

were asked to report SAMPL6 Molecule IDs, predicted octanol-water log P values, the log P standard error of the mean (SEM), and model uncertainty. It was mandatory to submit predictions for all these values, including the estimates of uncertainty. The log P SEM captures the statistical uncertainty of the predicted method, and the model uncertainty is an estimate of how well prediction and experimental values will agree. Molecule IDs assigned in SAMPL6 pK_a Challenge were conserved in the challenge for the ease of reference.

Participants were asked to categorize their methods as belonging to one of four method categories — physical, empirical, mixed or other. The following are definitions provided to participants for selecting a method category: Empirical models are prediction methods that are trained on experimental data, such as QSPR, machine learning models, artificial neural networks etc. Physical models are prediction methods that rely on the physical principles of the system such as molecular mechanics or quantum mechanics based methods to predict molecular properties. Methods taking advantage of both kinds of approaches were asked to be reported as “Mixed”. The “other” category was for methods which do not match the previous ones. At the analysis stage, some categories were further refined, as discussed in Section 3.3.5.

The submission files also included fields for naming the method, listing the software utilized, and a free text method section for the detailed documentation of each method. Only one log P value for each molecule per submission and only full prediction sets were allowed. Incomplete submissions – such as for a subset of compounds – were not accepted. We highlighted various factors for participants to consider in their log P predictions. These included:

1. There is a significant partitioning of water into the octanol phase. The mole fraction of water in octanol was previously measured as 0.271 ± 0.003 at 25°C [117].
2. The solutes can impact the distribution of water and octanol. Dimerization or oligomer-

ization of solute molecules in one or more of the phases may also impact results [120].

3. $\log P$ measurements capture partition of neutral species which may consist of multiple tautomers with significant populations or the major tautomer may not be the one given in the input file.
4. Shifts in tautomeric state populations on transfer between phases are also possible.

Research groups were allowed to participate with multiple submissions, which allowed them to submit prediction sets to compare multiple methods or to investigate the effect of varying parameters of a single method. All blind submissions were assigned a 5-digit alphanumeric submission ID, which will be used throughout this paper and also in the evaluation papers of participants. These abbreviations are defined in Table 2.3.

2.3.2 Evaluation approach

A variety of error metrics were considered when analyzing predictions submitted to the SAMPL6 $\log P$ Challenge. Summary statistics were calculated for each submission for method comparison, as well as error metrics of predictions of each method. Both summary statistics and individual error analysis of predictions were provided to participants before the virtual workshop. Details of the analysis and scripts are maintained on the SAMPL6 Github Repository (described in section 2.9.2).

There are six error metrics reported: the root-mean-squared error (RMSE), mean absolute error (MAE), mean (signed) error (ME), coefficient of determination (R^2), linear regression slope (m), and Kendall’s Rank Correlation Coefficient (τ). In addition to calculating these performance metrics, 95% confidence intervals were computed for these values using a bootstrapping-over-molecules procedure (with 10000 bootstrap samples) as described elsewhere in a previous SAMPL overview article [155]. Due to the small dynamic range of

experimental log P values of the SAMPL6 set, it is more appropriate to use accuracy based performance metrics, such as RMSE and MAE, to evaluate methods than correlation-based statistics. This observation is also typically reflected in the confidence intervals on these metrics. Calculated errors statistics of all methods can be found in Tables 2.10 and 2.11.

Submissions were originally assigned to four method categories (physical, empirical, mixed, and other) by participants. However, when we evaluated the set of participating methods it became clear that it was going to be more informative to group them using the following categories: **physical (MM)**, **physical (QM)**, **empirical**, and **mixed**. Methods from the “other” group were reassigned to empirical or physical (QM) categories as appropriate. Methods submitted as Physical by participants included quantum mechanical (QM), molecular mechanics-based (MM) and, to a lesser extent, integral equation-based approaches (EC-RISM). We subdivided these submissions into “physical (MM)” and “physical (QM)” categories. Integral equation-based approaches were also evaluated under the Physical (QM) category. The “mixed” category includes methods that physical and empirical approaches are used in combination. Table 2.3 indicates the final category assignments in the “Category” column.

We created a shortlist of consistently well-performing methods that were ranked in the top 20 consistently according to two error and two correlation metrics: RMSE, MAE, R-Squared, and Kendall’s Tau. These are shown in Table 2.4.

We included null and reference method prediction sets in the analysis to provide perspective for performance evaluations of blind predictions. Null models or null predictions employ a model which is not expected to be useful and can provide a simple point of comparison for more sophisticated methods, as ideally, such methods should improve on predictions from a null model. We created a null prediction set (submission ID *NULL0*) by predicting a constant log P value for every compound, based on a plausible log P value for drug-like compounds. We also provide reference calculations using several physical (alchemical)

and empirical approaches as a point of comparison. The analysis is presented with and without the inclusion of reference calculations in the SAMPL6 GitHub repository. All Figures and statistics tables in this manuscript include reference calculations. As the reference calculations were not formal submissions, these were omitted from formal ranking in the challenge, but we present plots in this article which show them for easy comparison. These are labeled with submission IDs of the form *REF##* to allow easy recognition of non-blind reference calculations.

In addition to the comparison of methods we also evaluated the relative difficulty of predicting $\log P$ of each molecule in the set. For this purpose, we plotted prediction error distributions of each molecule considering all prediction methods. We also calculated MAE for each molecule’s overall predictions as well as for predictions from each category as a whole.

2.4 Methods for reference calculations

Here we highlight the null prediction method and reference methods. We have included several widely-used physical and empirical methods as reference calculations in the comparative evaluation of $\log P$ prediction methods, in addition to the blind submissions of the SAMPL6 $\log P$ Challenge. These reference calculations are not formally part of the challenge but are provided as comparison methods. They were collected after the blind challenge deadline when the experimental data was released to the public. For a more detailed description of the methods used in the reference calculations, please refer to Section 2.9.3.

2.4.1 Physical Reference Calculations

Physical reference calculations were carried out using YANK [16], an alchemical free energy calculation toolkit [220, 184]. YANK implements Hamiltonian replica exchange molecular

dynamics (H-REMD) simulations to sample multiple alchemical states and is able to explore a number of different alchemical intermediate functional forms using the OpenMM toolkit for molecular simulation [57, 56, 55].

The GAFF 1.81 [?] and SMIRNOFF (smirnoff99Frosst 1.0.7) [148] force fields were combined with three different water models. Water models are important for accuracy in modeling efforts in molecular modeling and simulation. The majority of modeling packages make use of rigid and fixed charge models due to their computational efficiency. To test how different water models can impact predictions, we combined three explicit water models TIP3P [98], TIP3P Force Balance (TIP3P-FB) [221] and the Optimal Point Charge (OPC) model [92] with the GAFF and SMIRNOFF force fields. The TIP3P and TIP3P-FB models are a part of the three-site water model class where each atom has partial atomic charges and a Lennard-Jones interaction site centered at the oxygen atom. The OPC model is a rigid 4-site, 3-charge water model that has the same molecular geometry as TIP3P, but the negative charge on the oxygen atom is located on a massless virtual site at the HOH angle bisector. This arrangement is meant to improve the water molecule’s electrostatic distribution. While TIP3P is one of the older and more common models used, OPC and TIP3P-FB are newer models that were parameterized to more accurately reproduce more of the physical properties of liquid bulk water.

Reference calculations also included wet and dry conditions for the octanol phase using the GAFF and SMIRNOFF force field with TIP3P water. The wet octanol phase was 27% water by mole fraction [117]. The methods used for physical reference calculations are summarized in Table 2.1.

Physical reference calculations (submission IDs: *REF01-REF08*) were done using a previously untested direct transfer free energy calculation protocol (DFE) which involved calculating the transfer free energy between water and octanol phases (explained in detail in Section 2.9.3), rather than a more typical protocol involving calculating a gas-to-solvent

Table 2.1: **Methods used as reference calculations for the MM-based physical methods category.** Please see Section 2.9.3 in the Supplementary Information for detailed description of physical reference methods.

Submission ID	Approach	Force Field	Water Model	Octanol Phase	Number of Replicates
<i>REF01</i>	YANK, DFE protocol	GAFF 1.81	TIP3P-FB	Wet	3
<i>REF02</i>	YANK, DFE protocol	GAFF 1.81	TIP3P	Wet	3
<i>REF03</i>	YANK, DFE protocol	GAFF 1.81	OPC	Wet	3
<i>REF04</i>	YANK, DFE protocol	smirnoff99Frosst 1.0.7	TIP3P-FB	Wet	3
<i>REF05</i>	YANK, DFE protocol	smirnoff99Frosst 1.0.7	TIP3P	Wet	3
<i>REF06</i>	YANK, DFE protocol	smirnoff99Frosst 1.0.7	OPC	Wet	3
<i>REF07</i>	YANK, DFE protocol	GAFF 1.81	TIP3P	Dry	3
<i>REF08</i>	YANK, DFE protocol	smirnoff99Frosst 1.0.7	TIP3P	Dry	3

transfer free energy for each phase – an indirect solvation-based transfer free energy (IFE) protocol. In order to check for problems caused by this error, we included additional calculations performed by the more typical IFE protocol. Method details for the IFE protocol are presented in Section 2.9.3 and results are discussed in Section 2.5.2. However, only reference calculations performed with DFE protocol were included in overall evaluation of the SAMPL6 Challenge presented in Section 2.5.1, because only these spanned the full range of force fields and solvent models we sought to explore.

2.4.2 Empirical Reference Calculations

As empirical reference models, we used a number of commercial calculation programs, with the permission of the respective vendors, who agreed to have the results included in the SAMPL6 comparison. The programs are summarized in Table 2.2 and cover several of the different methodologies described in sections 2.2.2 and 2.9.3.

Table 2.2: **Methods used as reference calculations for the empirical log P prediction category.** Please see section 2.9.3 in the Supplementary Information for a detailed description of empirical methods.

Submission ID	Name	Vendor	Approach	Website
<i>REF09</i>	clogP (BioByte)	BioByte	group contributions	www.biobyte.com
<i>REF13</i>	SlogP (MOE)	Chemical Computing Group	atomic contributions	www.chemcomp.com
<i>REF11</i>	logP(ow) (MOE)	Chemical Computing Group	atomic contributions and correction factors	www.chemcomp.com
<i>REF10</i>	h_logP (MOE)	Chemical Computing Group	QSPR, based on extended Hückel theory descriptors	www.chemcomp.com
<i>REF12</i>	MoKa_logP	Molecular Discovery	QSPR, based on Molecular Interaction Field descriptors	www.moldiscovery.com

2.4.3 Our null prediction method

This submission set is designed as a null model which predicts the log P of all molecules to be equal to the mean clogP of FDA approved oral new chemical entities (NCEs) between the years 1998 and 2017 based on the analysis of Micheal D. Shultz (2019) [199]. We show this null model with submission ID *NULL0*. The mean clogP of FDA approved oral NCEs approved between 1900-1997, 1998-2007, and 2008-2017 were reported 2.1, 2.4, and 2.9, respectively, using StarDrop clogP calculations (<https://www.optibrium.com/>). We calculated the mean of NCEs approved between 1998 – 2017, which is 2.66, to represent the average log P of contemporary drug-like molecules. We excluded the years 1900-1997 from this calculation as the early drugs tend to be much smaller and much more hydrophilic than the ones being developed at present.

2.5 Results and Discussion

2.5.1 Overview of challenge results

A large variety of methods were represented in the SAMPL6 log P Challenge. There were 91 blind submissions collected from 27 participating groups in the log P challenge (Tables of

participants and the predictions they submitted are presented in SAMPL6 GitHub Repository and its archived copy in the Supporting Information.) This represented an increase in interest over the previous SAMPL challenges. In the SAMPL5 Cyclohexane-Water log D Challenge, there were 76 submissions from 18 participating groups [19], so participation was even higher this iteration.

Out of blind submissions of the SAMPL6 log P Challenge, there were 31 in the physical (MM) category, 25 in the physical (QM) category, 18 in the empirical category, and 17 in the mixed method category (Table 2.3). We also provided additional reference calculations – five in the empirical category, and eight in the physical (MM) category.

The following sections present detailed performance evaluation of blind submissions and reference prediction methods. Performance statistics of all the methods can be found in 2.10. Methods are referred to by their submission ID’s which are provided in 2.3.

Performance statistics for method comparison

Many methods in the SAMPL6 Challenge achieved good predictive accuracy for octanol-water log P values. Figure 2.3 shows the performance comparison of methods based on accuracy with RMSE and MAE. 10 methods achieved an RMSE ≤ 0.5 log P units. These methods were QM-based, empirical, and mixed approaches (submission IDs: *hmz0n*, *gmoq5*, *3vqbi*, *sq07q*, *j8nwc*, *xxh4i*, *hdpuj*, *dqzk4*, *vzgyt*, *ypmr0*). Many of the methods had an RMSE ≤ 1.0 log P units. These 40 methods include 34 blind predictions, 5 reference calculations, and the null prediction method.

Correlation-based statistics methods only provide a rough comparison of methods of the SAMPL6 Challenge, given the small dynamic range of the experimental log P dataset. Figure 2.4 shows R^2 and Kendall’s Tau values calculated for each method, sorted from high to low performance. However, the uncertainty of each correlation statistic is quite high, not

Table 2.3: **Submission IDs, names, category, and type for all the log P participant and reference calculation submissions.** Submission IDs of methods are listed in the ID column. Reference calculations are labeled as *REF##*. The method name column lists the names provided by each participant in the submission file. The “type” column indicates if submission was or a post-deadline reference calculation, denoted by “Blind” or “Reference” respectively. The table is ordered by increasing RMSE from top to down and left to right, although many consecutively listed methods are statistically indistinguishable. All calculated error statistics are available in Tables 2.10 and 2.11.

ID	Method Name	Category	Submission Type	ID	Method Name	Category	Submission Type
<i>hzm20n</i>	cosmotherm_FINE19 [14]	Physical (QM)	Blind	<i>rdsnw</i>	EC_RISM_wet_P1w+1o [22]	Physical (QM)	Blind
<i>gmoq5</i>	Global XGBoost-Based QSPR LogP Predictor	Empirical	Blind	<i>ggm6n</i>	FS-GM (Fast switching Growth Method) [73]	Physical (MM)	Blind
<i>3vqb1</i>	cosmoquick_TZVP18+ML [14]	Mixed	Blind	<i>jjd0b</i>	MD/S-MBIS-GAFF-TIP3P/MBAR [74]	Physical (MM)	Blind
<i>sq07q</i>	Local XGBoost-Based QSPR LogP Predictor	Empirical	Blind	<i>2zb0</i>	EC_RISM_dry_P1w+1o [22]	Physical (QM)	Blind
<i>j8nwc</i>	EC_RISM_wet_P1w+2o [22]	Physical (QM)	Blind	<i>cr3hs</i>	PLS3 from NIST data and QM-generated QSAR Descriptors subset [75]	Mixed	Blind
<i>xxh4i</i>	SM12-Solvation-Trained [76]	Mixed	Blind	<i>arw58</i>	DLPNO-CCSD(T)/cc-pVTZ//B3LYP-D3/cc-pVTZ [75]	Physical (QM)	Blind
<i>hdpuj</i>	RayLogP-II, a cheminformatic QSPR model predicting the octanol/water partition coefficient, logP. [77]	Empirical	Blind	<i>ahmtf</i>	B3PW91-TZ SMD kcl-wet-oct [78]	Physical (QM)	Blind
<i>dqkx4</i>	LogP_SMD_Solvation_DFT [79]	Physical (QM)	Blind	<i>o7djk</i>	B3PW91-TZ SMD wetoct [78]	Physical (QM)	Blind
<i>vzgyt</i>	rfs-logp	Empirical	Blind	<i>fmf7r</i>	dice	Mixed	Blind
<i>ypmr0</i>	SM8-Solvation [76]	Physical (QM)	Blind	<i>4p2ph</i>	DLPNO-Solv-ccCA [75]	Physical (QM)	Blind
<i>y6ub</i>	S+logP	Empirical	Blind	<i>6fyg5</i>	Solvation-M062X [80]	Physical (QM)	Blind
<i>7egyc</i>	SMD-Solvation-Trained [76]	Mixed	Blind	<i>sqos1</i>	MD-AMBER-dryoct [81]	Physical (MM)	Blind
<i>0a7a8</i>	ML Prediction using MD Feature Vector Trained on logP_octanol_water, with Additional Meta-learner [82]	Mixed	Blind	<i>rs4ns</i>	BLYP/cc-pVTZ//B3LYP-D3/cc-pVTZ [75]	Physical (QM)	Blind
<i>7dhtp</i>	LogP-prediction-method-name	Empirical	Blind	<i>c7t5j</i>	PBE/cc-pVTZ//B3LYP-D3/cc-pVTZ [75]	Physical (QM)	Blind
<i>qyzjx</i>	EC_RISM_dry_P1w+2o [22]	Physical (QM)	Blind	<i>jc68f</i>	PW91/cc-pVTZ//B3LYP-D3/cc-pVTZ [75]	Physical (QM)	Blind
<i>REF11</i>	logP(o/w) (MOE)	Empirical	Reference	<i>03gyy</i>	Linear Regression-B3LYP/6-311G** [80]	Mixed	Blind
<i>REF13</i>	SlogP (MOE)	Empirical	Reference	<i>hsotx</i>	B3LYP/cc-pVTZ//B3LYP-D3/cc-pVTZ [75]	Physical (QM)	Blind
<i>w6jta</i>	ML Prediction using MD Feature Vector Trained on logP_octanol_water [82]	Mixed	Blind	<i>ke5gu</i>	MD/S-MBIS-GAFF-SPCE/MBAR [74]	Physical (MM)	Blind
<i>REF12</i>	MoKa_logP	Empirical	Reference	<i>mmuua</i>	MD-LigParGen-wetoc [81]	Physical (MM)	Blind
<i>jj2zm</i>	SM8-Solvation-Trained [76]	Mixed	Blind	<i>fe8ws</i>	B3PW91/cc-pVTZ//B3LYP-D3/cc-pVTZ [75]	Physical (QM)	Blind
<i>5krd1</i>	ZINC15 versus PM3 [80]	Mixed	Blind	<i>50yn</i>	PBE0/cc-pVTZ//B3LYP-D3/cc-pVTZ [75]	Physical (QM)	Blind
<i>REF10</i>	h_logP (MOE)	Empirical	Reference	<i>fyx45</i>	LogP-prediction-Drude-FEP-HuangLab	Physical (MM)	Blind
<i>gnxuu</i>	ML Prediction using MD Feature Vector Trained on logP_octanol_water [82]	Mixed	Blind	<i>6nmtt</i>	MD-AMBER-wetoc [81]	Physical (MM)	Blind
<i>tc4xa</i>	NHLBI-NN-5HL	Empirical	Blind	<i>eufcy</i>	MD-LigParGen-dryoct [81]	Physical (MM)	Blind
<i>6cayo</i>	SM12-Solvation [76]	Physical (QM)	Blind	<i>tzzb5</i>	Alchemical-CGenFF [83]	Physical (MM)	Blind
<i>dbmg3</i>	GC-LSER	Empirical	Blind	<i>3oqhx</i>	MD-CHARMM-dryoct	Physical (MM)	Blind
<i>kxsp3</i>	PLS2 from NIST data and QM-generated QSAR Descriptors [75]	Mixed	Blind	<i>bzeez</i>	FS-AGM (Fast switching Annihilation/Growth Method) [73]	Physical (MM)	Blind
<i>nh6c0</i>	Molecular-Dynamics-Expanded-Ensembles [84]	Physical (MM)	Blind	<i>ynqyk</i>	TWOVAR	Empirical	Blind
<i>kiwfv</i>	LogP-prediction-method-IEFPCM/MST [85]	Physical (QM)	Blind	<i>5svjv</i>	FS-GM (Fast switching Growth Method) [73]	Physical (MM)	Blind
<i>NULL0</i>	mean clogP of FDA approved oral drugs (1998-2017)	Empirical	Reference	<i>odex0</i>	InterX_ARROW_2017_PIMD_SOLVENT2_WET_OCTANOL	Physical (MM)	Blind
<i>ujsgv</i>	Alchemical-CGenFF [83]	Physical (MM)	Blind	<i>padym</i>	InterX_ARROW_2017_PIMD_WET_OCTANOL	Physical (MM)	Blind
<i>REF09</i>	dlogP (Biobyte)	Empirical	Reference	<i>pnc4j</i>	LogP-prediction-Drude-Umbrella-HuangLab	Physical (MM)	Blind
<i>wu52s</i>	LogP-PLS-ECFC4_CSsep-Bayer	Empirical	Blind	<i>REF02</i>	YANK-GAFF-TIP3P-wet-oct	Physical (MM)	Reference
<i>g6dwz</i>	NHLBI-NN-3HL	Empirical	Blind	<i>REF05</i>	YANK-SMIRNOFF-TIP3P-wet-oct	Physical (MM)	Reference
<i>5mahv</i>	ML Prediction using MD Feature Vector Trained on Hydration Free Energy [82]	Mixed	Blind	<i>REF08</i>	YANK-SMIRNOFF-TIP3P-dry-oct	Physical (MM)	Reference
<i>bqeuu</i>	ISIDA-LSER	Empirical	Blind	<i>REF07</i>	YANK-GAFF-tip3p-dry-oct	Physical (MM)	Reference
<i>d7vth</i>	UFZ-LSER	Empirical	Blind	<i>fcspk</i>	ARROW_2017_PIMD_SOLVENT2	Physical (MM)	Blind
<i>2mi5w</i>	Alchemical-CGenFF [83]	Physical (MM)	Blind	<i>6cm6a</i>	ARROW_2017_PIMD	Physical (MM)	Blind
<i>kuddg</i>	LogP-Pred-MTNN-GraphConv-Bayer	Empirical	Blind	<i>bq6fo</i>	Extended solvent-contact model approach	Mixed	Blind
<i>qz8d5</i>	SMD-Solvation [76]	Physical (QM)	Blind	<i>623c0</i>	MD-OPLSAA-wetoc [81]	Physical (MM)	Blind
<i>y0xdd</i>	FS-GM (Fast switching Growth Method) [73]	Physical (MM)	Blind	<i>4nfz2</i>	MD/S-HI-GAFF-TIP3P/MBAR [74]	Physical (MM)	Blind
<i>2ggir</i>	FS-AGM (Fast switching Annihilation/Growth Method) [73]	Physical (MM)	Blind	<i>eg52i</i>	ARROW_2017	Physical (MM)	Blind
<i>dxybt</i>	B3PW91-TZ SMD set1 [78]	Physical (QM)	Blind	<i>cp8kv</i>	MD-OPLSAA-dryoct [81]	Physical (MM)	Blind
<i>mm0jf</i>	LogP-prediction-SMD-HuangLab	Physical (QM)	Blind	<i>5585v</i>	Alchemical-CGenFF [83]	Physical (MM)	Blind
<i>h83sb</i>	Linear Regression with B3LYP/6-31G+ [80]	Mixed	Blind	<i>j4nb3</i>	FOURVAR	Empirical	Blind
<i>3wvjy</i>	Alchemical-CGenFF [83]	Physical (MM)	Blind	<i>REF04</i>	YANK-SMIRNOFF-TIP3P-FB-wet-oct	Physical (MM)	Reference
<i>f3dpg</i>	PLS from NIST data and QM-generated QSAR Descriptors [75]	Mixed	Blind	<i>hf4wj</i>	MD/S-HI-GAFF-SPCE/MBAR [74]	Physical (MM)	Blind
<i>25s67</i>	FS-AGM (Fast switching Annihilation/Growth Method) [73]	Physical (MM)	Blind	<i>REF01</i>	YANK-GAFF-TIP3P-FB-wet-oct	Physical (MM)	Reference
<i>zdj0j</i>	Solvation-B3LYP [80]	Physical (QM)	Blind	<i>REF03</i>	YANK-GAFF-OPC-wet-oct	Physical (MM)	Reference
<i>7gg6s</i>	MLR from NIST data and QM-generated QSAR Descriptors [75]	Mixed	Blind	<i>REF06</i>	YANK-SMIRNOFF-OPC-wet-oct	Physical (MM)	Reference
<i>hwf2k</i>	Extended solvent-contact model approach	Empirical	Blind	<i>pku5g</i>	SAMPL5_49_retro3	Empirical	Blind
<i>pcv32</i>	Solvation- WB97X-D [80]	Physical (QM)	Blind	<i>po4g2</i>	SAMPL5_49	Empirical	Blind
<i>v2q0t</i>	InterX_GAFF_WET_OCTANOL	Physical (MM)	Blind				

allowing a true ranking based on correlation. Methods with R^2 and Kendall’s Tau higher than 0.5 constitute around 50% of the methods and can be considered as the better half. However, the performance of individual methods is statistically indistinguishable. Nevertheless, it is worth noting that QM-based methods appeared marginally better at capturing the correlation and ranking of experimental $\log P$ values. These methods comprised the top four based on R^2 (≥ 0.75 ; submission IDs: *2tzb0*, *rdsnw*, *hmz0n*, *mm0jf*), and the top six based on Kendall’s Tau, (≥ 0.70 ; submission IDs: *j8nwc*, *qyzjx*, *2tzb0*, *rdsnw*, *mm0jf*, and *6fyg5*). However, due to the small dynamic range and the number of experimental $\log P$ values of the SAMPL6 set, correlation-based statistics are less informative than accuracy-based performance metrics such as RMSE and MAE.

Results from physical methods

One of the aims of the SAMPL6 $\log P$ Challenge was to assess the accuracy of physical approaches in order to potentially provide direction for improvements which could later impact accuracy in downstream applications like protein-ligand binding. Some MM-based methods used for $\log P$ predictions use the same technology applied to protein-ligand binding predictions, so improvements made to modeling of partition may in principle carry over. However, prediction of partition between two solvent phases is a far simpler test only capturing some aspects of affinity prediction – specifically, small molecule and solvation modeling – in the absence of protein-ligand interactions and protonation state prediction problems.

Figure 2.5 shows a comparison of the performance of MM- and QM-based methods in terms of RMSE and Kendall’s Tau. Both in terms of accuracy and ranking ability, QM methods resulted in better results, on average. QM methods using implicit solvation models outperformed MM-based methods with explicit solvent methods that were expected to capture the heterogeneous structure of the wet octanol phase better. Only 3 MM-based methods and 8 QM-based methods achieved RMSE less than 1 $\log P$ unit. 5 of these QM-based methods

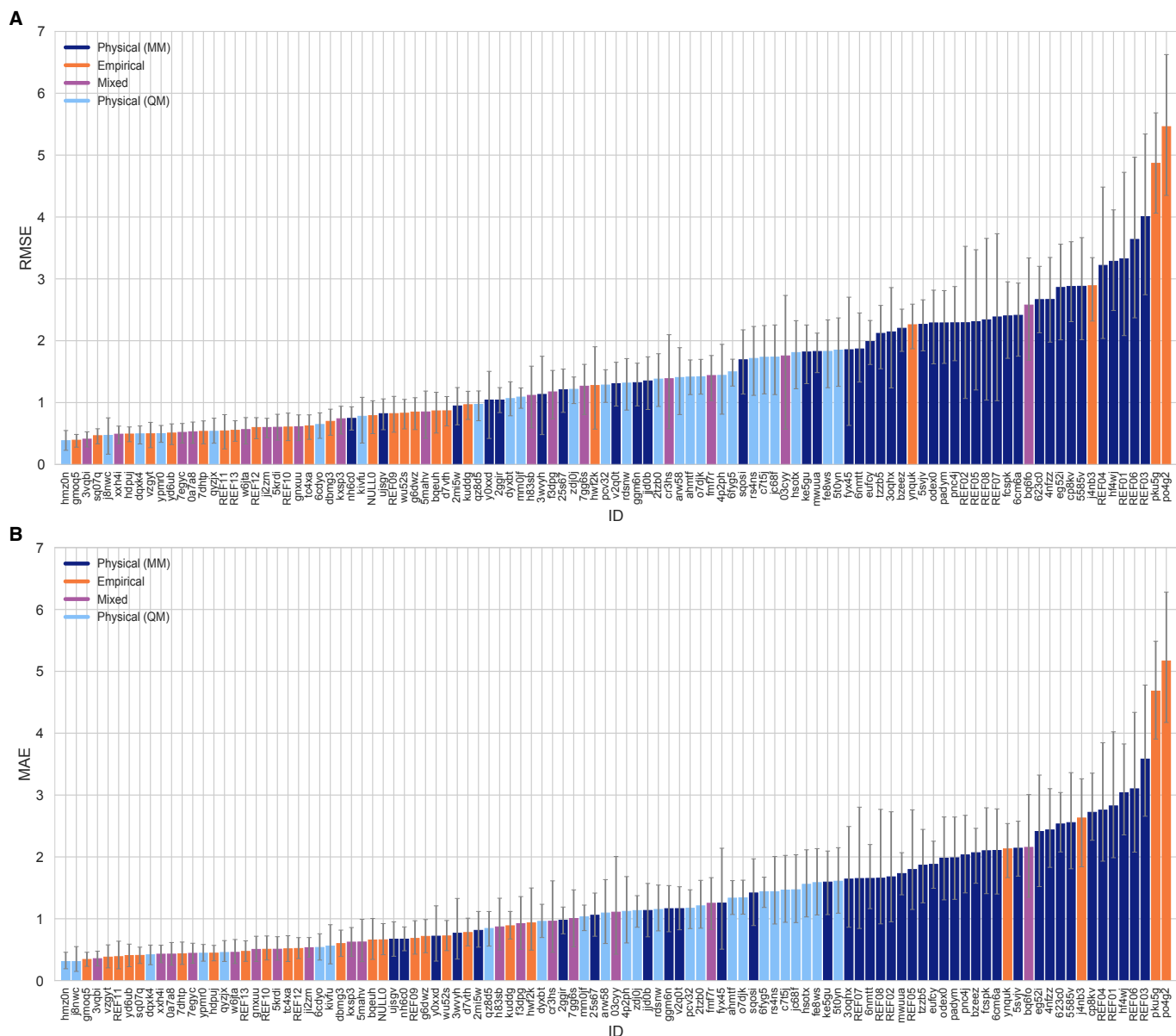


Figure 2.3: **Overall accuracy assessment for all methods participating in the SAMPL6 log P Challenge.** Both root-mean squared error (RMSE) and mean absolute error (MAE) are shown, with error bars denoting 95% confidence intervals obtained by bootstrapping over challenge molecules. Submission IDs are summarized in Table 2.3. Submission IDs of the form *REF##* refer to non-blinded reference methods computed after the blind challenge submission deadline, and *NULL0* is the null prediction method; all others refer to blind, prospective predictions.

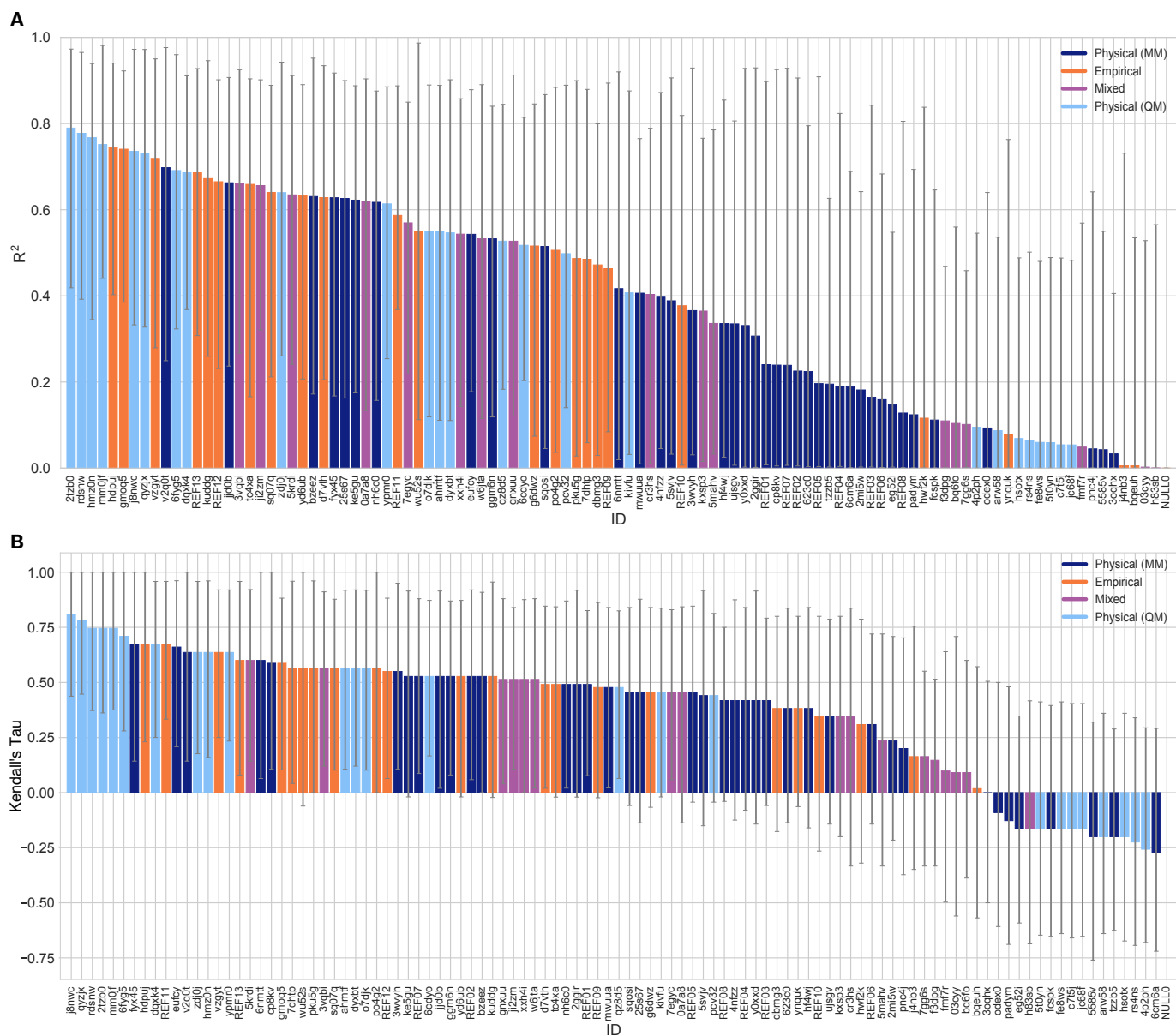


Figure 2.4: Overall correlation assessment for all methods participating SAMPL6 log P Challenge. Pearson’s R^2 and Kendall’s Rank Correlation Coefficient τ are shown, with error bars denoting 95% confidence intervals obtained by bootstrapping over challenge molecules. Submission IDs are summarized in Table 2.3. Submission IDs of the form *REF##* refer to non-blinded reference methods computed after the blind challenge submission deadline, and *NULL0* is the null prediction method; all others refer to blind, prospective predictions. Overall, a large number and wide variety of methods have a statistically indistinguishable performance on ranking, in part because of the relatively small dynamic range of this set and because of the small size of the set. Roughly the top half of methods with Kendall’s $\tau > 0.5$ fall into this category.

showed very high accuracy ($\text{RMSE} \leq 0.5 \log P$ units). The three MM-based methods with the lowest RMSE were:

- Molecular-Dynamics-Expanded-Ensembles (*nh6c0*): This submission used an AMBER/OPLS-based force field with manually adjusted parameters (following rules from the participant’s article [166]), modified Toukan-Rahman water model with Non-zero Lennard-Jones parameters [165], and modified Expanded Ensembles (EEMD) method [130] for free energy estimations .
- Alchemical-CGenFF (*ujsgv, 2mi5w*, [112]): These two submissions used Multi-Phase Boltzmann Weighting with the CHARMM Generalized Force Field (CGenFF) [214], and the TIP3P water model [98]. From the brief method descriptions submitted to the challenge we could not identify the difference between these prediction sets.

RMSE values for predictions made with MM-based methods ranged from 0.74 to $4.00 \log P$ units, with the average of the better half being $1.44 \log P$ units.

Submissions included diverse molecular simulation-based $\log P$ predictions made using alchemical approaches. These included Free Energy Perturbation (FEP) [235] and BAR estimation [27], Thermodynamic integration (TI) [102], and non-equilibrium switching (NES) [180, 96]. Predictions using YANK [16] Hamiltonian replica exchange molecular dynamics and MBAR [198] were provided as reference calculations.

A variety of combinations of force fields and water models were represented in the challenge. These included CGenFF with TIP3P or OPC3 [93] water models; OPLS-AA [51] with OPC3 and TIP4P [98] water models; GAFF [219] with TIP3P, TIP3P Force Balance [221], OPC [92], and OPC3 water models; GAFF2 [215] with the OPC3 water model; GAFF with Hirshfeld-I [32] and Minimal Basis Set Iterative Stockholder(MBIS) [216] partial charges and the TIP3P or SPCE water models [115]; the SMIRNOFF force field [148] with the

TIP3P, TIP3P Force Balance, and OPC water models; and submissions using Drude [124] and ARROW [99] polarizable force fields.

Predictions that used polarizable force fields did not show an advantage over fixed-charged force fields in this challenge. RMSEs for polarizable force field submissions range from 1.85 to 2.86 (submissions with the Drude Force Field were *fyx45*, *pnc4j*, and those with the ARROW Force Field were *odex0*, *padym* *fcspk*, and *6cm6a*).

Predictions using both dry and wet octanol phases were submitted to the log P challenge. When submissions from the same participants were compared, we find that including water in the octanol phase only slightly lowers the RMSE (0.05-0.10 log P units), as seen in Alchemical-CGenFF predictions (wet: *ujshv*, *2mi5w*, *ttzb5*; dry: *3wvyh*), YANK-GAFF-TIP3P predictions (wet: *REF02*, dry: *REF07*), MD-LigParGen predictions with OPLS and TIP4P (wet: *mwuuu*, dry: *eufcy*), and MD-OPLSAA predictions with TIP4P (wet: *623c0*, dry: *cp8kv*). However this improvement in performance with wet octanol phase was not found to be a significant effect on overall prediction accuracy. Methodological differences and choice of force field have a greater impact on prediction accuracy than the composition of the octanol phase.

Refer to Table 2.7 for a summary of force fields and water models used in MM-based submissions. For additional analysis, we refer the interested reader to the work of Piero Procacci and Guido Guarnieri, who provide a detailed comparison of MM-based alchemical equilibrium and non-equilibrium approaches in SAMPL6 Challenge in their paper [181]. Specifically, in the section “Overview on MD-based SAMPL6 submissions” of their paper, they provide comparisons subdividing submissions based force field (for CGenFF, GAFF1/2, and OPLS-AA).

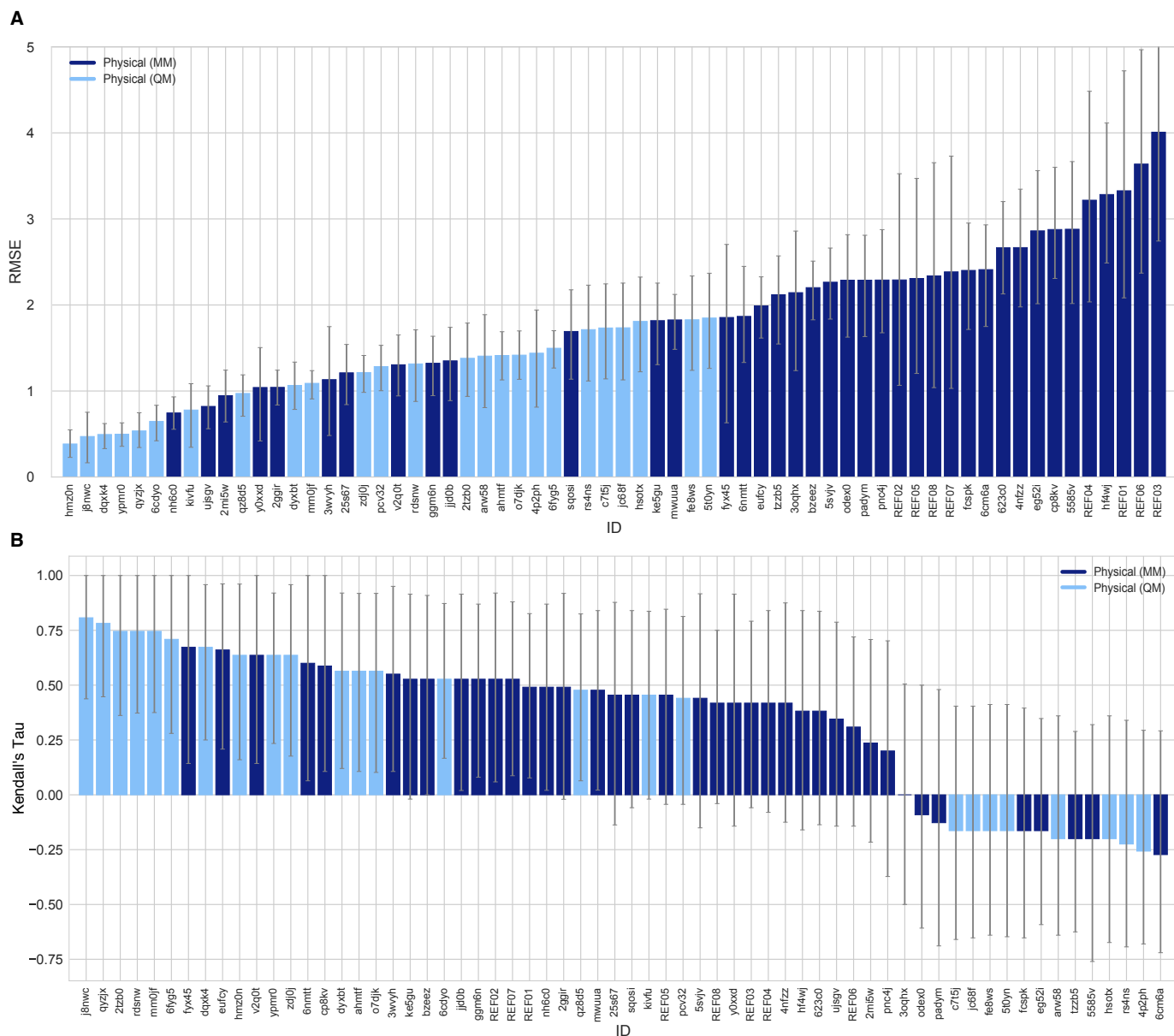


Figure 2.5: **Performance statistics of physical methods.** Physical methods are further classified into quantum chemical (QM) methods and molecular mechanics (MM) methods. RMSE and Kendall's Rank Correlation Coefficient Tau are shown, with error bars denoting 95% confidence intervals obtained by bootstrapping over challenge molecules. Submission IDs are summarized in Table 2.3. Submission IDs of the form *REF##* refer to non-blinded reference methods computed after the blind challenge submission deadline; all others refer to blind, prospective predictions.

A shortlist of consistently well-performing methods

Although there was not any single method that performed significantly better than others in the log P challenge, we identified a group of consistently well performing methods. There were many methods with good performance when judged based on RMSE, but not many methods consistently showed up at the top according to all metrics. When individual error metrics are considered, many submissions were not different from one another in a statistically significant way, and ranking typically depends on the metric chosen due overlapping confidence intervals. Instead, we identified several consistently well performing methods by looking at several different metrics – two assessing accuracy (RMSE and MAE) and two assessing correlation (Kendall’s Tau and R^2). We determined those methods which are in the top 20 by each of these metrics. This resulted in a list of eight methods which are consistently well performing. The shortlist of consistently well-performing methods are presented in Table 2.4.

The resulting eight consistently well-performing methods were QM-based physical models and empirical methods. These eight methods were fairly diverse. Traditional QM-based physical methods included log P predictions with COSMO-RS method as implemented in COSMOtherm v19 at the BP//TZVPD//FINE Single Point level (*hmz0n*, [104, 103, 107]) and the SMD solvation model with the M06 density functional family (*dqak4*, [79]). Additionally, two other top QM-based methods seen in this shortlist used EC-RISM theory with wet or dry octanol (*j8nwc* and *qyzjx*) [210]. Several empirical submissions also were among these well-performing methods – specifically, the Global XGBoost-Based QSPR LogP Predictor (*gmoq5*), the RayLogP-II (*hdpuj*) approach, and rfs-logp (*vzgyt*). Among reference calculations, SlogP calculated by MOE software (*REF13*) was the only method that was consistently well performing.

Figure 2.6 compares log P predictions with experimental values for these 8 well-performing

methods, as well as one additional method which has an average level of performance. This representative method (*rdsnw*, [210]) is the method with the highest RMSE below the median of all methods (including reference methods).

Table 2.4: Eight consistently well-performing prediction methods based on consistent ranking within the Top 20 according to various statistical metrics. Submissions were ranked according to RMSE, MAE, R^2 , and τ . Many top methods were found to be statistically indistinguishable considering uncertainties of error metrics. Moreover, sorting of methods was influenced significantly by the choice of metric chosen. We assessed top 20 methods according to each metric to determine which methods are always among the top 20 according to all four statistical metrics chosen. A set of consistently well-performing methods were determined: Four QM-based and four empirical methods. Seven of these methods are blind submissions of SAMPL6 Challenge, and one of them (*REF13*) is a non-blind reference calculation. Performance statistics are provided as mean and 95% confidence intervals.

ID	Method Name	Category	Type	RMSE	MAE	R^2	Kendall's Tau (τ)
<i>hmz0n</i>	cosmotherm_FINE19	Physical (QM)	Blind	0.38 [0.23, 0.55]	0.31 [0.19, 0.46]	0.77 [0.36, 0.94]	0.64 [0.17, 1.00]
<i>gmaq5</i>	Global XGBoost-Based QSPR LogP Predictor	Empirical	Blind	0.39 [0.28, 0.49]	0.34 [0.23, 0.46]	0.74 [0.40, 0.92]	0.59 [0.12, 0.89]
<i>j8nwc</i>	EC_RISM_wet_P1w+2o	Physical (QM)	Blind	0.47 [0.17, 0.75]	0.31 [0.15, 0.54]	0.74 [0.33, 0.97]	0.81 [0.46, 1.00]
<i>hdpuj</i>	RayLogP-II, a cheminformatic QSPR model predicting the octanol/water partition coefficient	Empirical	Blind	0.49 [0.37, 0.61]	0.44 [0.32, 0.57]	0.74 [0.40, 0.94]	0.67 [0.22, 1.00]
<i>dqk4</i>	LogP_SMD_Solvation_DFT	Physical (QM)	Blind	0.49 [0.33, 0.62]	0.42 [0.26, 0.57]	0.69 [0.35, 0.91]	0.67 [0.27, 0.96]
<i>vzgyt</i>	rfs-logp	Empirical	Blind	0.50 [0.27, 0.68]	0.38 [0.21, 0.58]	0.72 [0.29, 0.95]	0.64 [0.23, 0.92]
<i>qyzjx</i>	EC_RISM_dry_P1w+2o	Physical (QM)	Blind	0.54 [0.34, 0.75]	0.46 [0.31, 0.64]	0.73 [0.31, 0.97]	0.78 [0.44, 1.00]
<i>REF13</i>	SlogP (MOE)	Empirical	Reference	0.55 [0.38, 0.71]	0.47 [0.31, 0.65]	0.69 [0.29, 0.92]	0.60 [0.08, 0.96]

Difficult chemical properties for log P predictions

In addition to comparing method performance, we analyzed the prediction errors for each compound in the challenge set to assess whether particular compounds or chemistries are especially challenging (Figure 2.7). For this analysis, MAE is a more appropriate statistical value for following global trends, as its value is less affected by outliers than is RMSE.

Performance on individual molecules shows relatively uniform MAE across the challenge set (Figure 2.7A). Predictions of SM14 and SM16 were slightly more accurate than the rest of the molecules when averaged across all methods. Prediction accuracy on each molecule, however, is highly variable depending on method category (Figure 2.7B). Predictions of SM08, SM13, SM09, and SM12 were significantly less accurate with physical (MM) methods

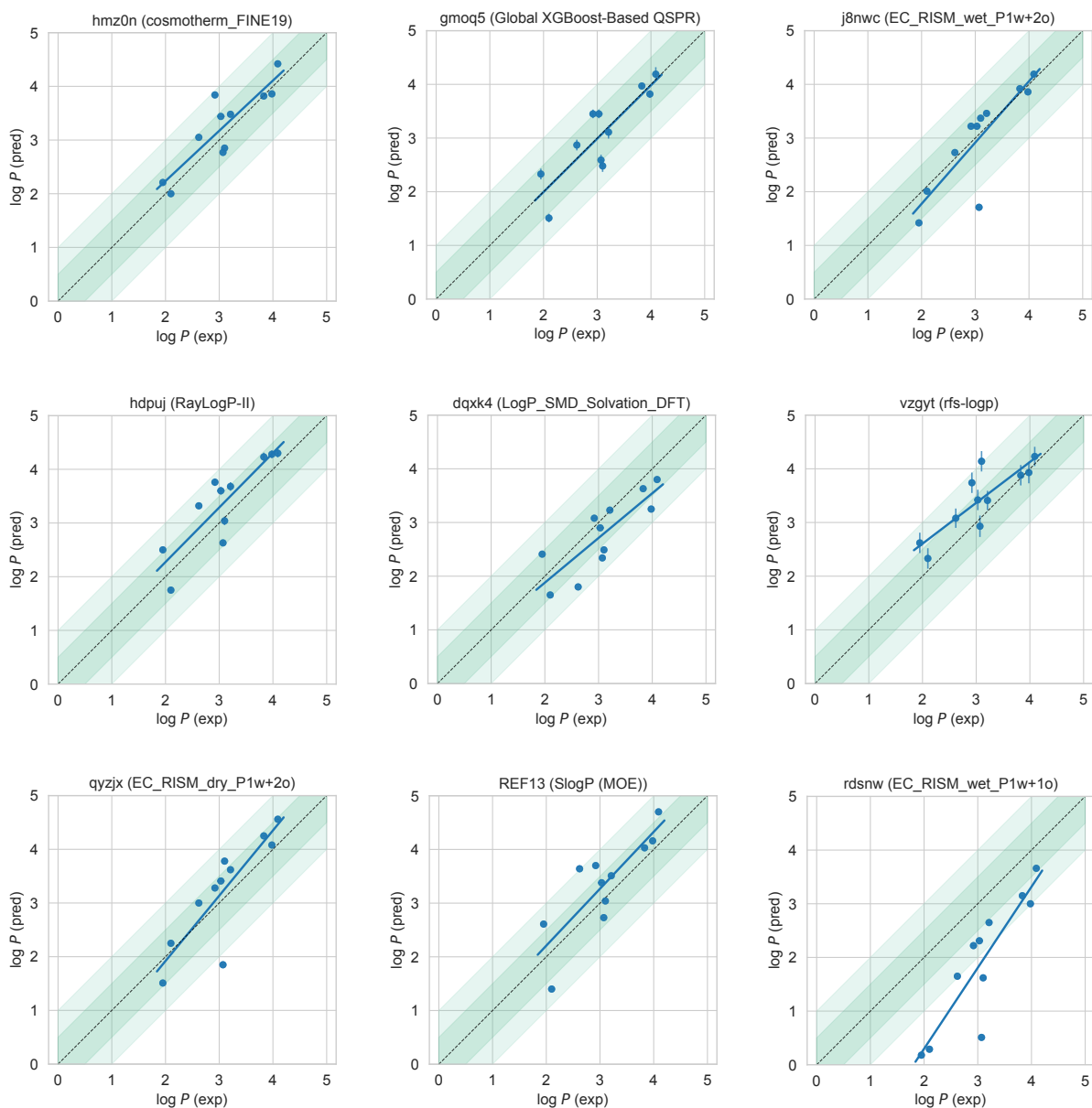


Figure 2.6: **Predicted vs experimental value correlation plots of 8 best-performing methods and one representative average method.** Dark and light green shaded areas indicate 0.5 and 1.0 units of error. Error bars indicate standard error of the mean of predicted and experimental values. Experimental $\log P$ SEM values are too small to be seen under the data points. EC_RISM_wet_P1w+1o method (*rdsnw*) was selected as the representative average method, as it is the method with the highest RMSE below the median.

than the other method categories by $2 \log P$ units in terms of MAE over all methods in each category. These molecules were not challenging for QM-based methods. Discrepancies in predictions of SM08 and SM13 are discussed in Section 2.5.2. For QM-based methods, SM04 and SM02 were most challenging. The largest MAE for empirical methods was observed for SM11 and SM15.

Figure 2.7C shows the error distribution for each SAMPL6 molecule over all prediction methods. It is interesting to note that most distributions are peaked near an error of zero, suggesting that perhaps a consensus model might outperform most individual models. However, SM15 is more significantly shifted away from zero than any other compound (ME calculated across all molecules is -0.88 ± 1.49 for SM15). SM08 had the most spread in $\log P$ prediction error.

This challenge focused on $\log P$ of neutral species, rather than $\log D$ as studied in SAMPL5, which meant that we do not see the same trends where performance is significantly worse for compounds with multiple protonation states/tautomers or where pK_a values are uncertain. However, in principle, tautomerization can still influence $\log P$ values. Multiple neutral tautomers can be present at significant populations in either solvent phase, or the major tautomer can be different in each solvent phase. However, this was not expected to be the case for any of the 11 compounds in this SAMPL6 Challenge. We do not have experimental data on the identity or ratio of tautomers, but tautomers other than those depicted in Figure 2.2 would be much higher in energy according to QM predictions [210] and, thus, very unlikely to play a significant role. Still, for most $\log P$ prediction methods, it was at least necessary for participants to select the major neutral tautomer. We do not observe statistically worse error for compounds with potential tautomer uncertainties here, suggesting it was not a major factor in overall accuracy, some participants *did* chose to run calculations on tautomers that were not provided in the challenge input files (Figure 2.11 and Table 2.5), as we discuss in Section 2.5.2.

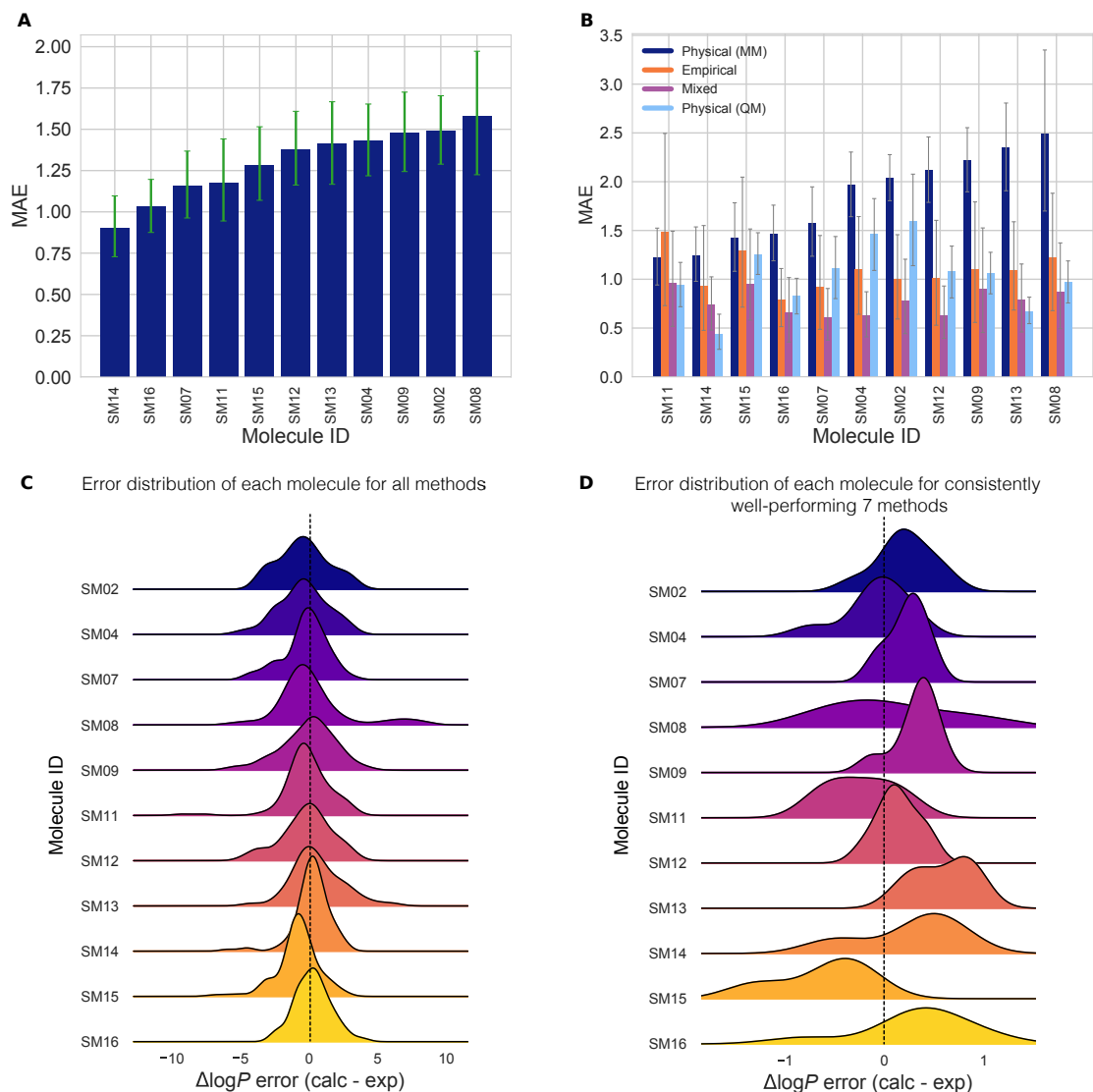


Figure 2.7: **Molecule-wise prediction error distribution plots show how variable the prediction accuracy was for individual molecules across all prediction methods.** (A) MAE calculated for each molecule as an average of all methods shows relatively uniform MAE across the challenge set. SM14 and SM16 predictions were slightly more accurate than the rest. (B) MAE of each molecule broken out by method category shows that for each method category the most challenging molecules were different. Predictions of SM08, SM13, SM09, and SM12 log P values were significantly less accurate with Physical (MM) methods than the other method categories. For QM-based methods SM04 and SM02 were most challenging. Largest MAE for Empirical methods were observed for SM11 and SM15. (C) Error distribution for each SAMPL6 molecule overall prediction methods. It is interesting to note that most distributions are peaked near an error of zero, suggesting that perhaps a consensus model might outperform most individual models. However, SM15 is more significantly shifted away from zero than any other compound. SM08 has a significant tail showing probability of overestimated log P predictions by some methods. (D) Error distribution for each molecule calculated for only 7 methods from blind submissions that were determined to be consistently well-performing (*hmz0n*, *gmoq5*, *j8nwc*, *hdpuj*, *dqak4*, *vzgyt*, *qyzjx*).

Comparison to the past SAMPL challenges

Overall, SAMPL6 log P predictions were more accurate than log D predictions in the SAMPL5 Cyclohexane-Water log D Challenge (Figure 2.3). In the log D challenge, only five submissions had an RMSE ≤ 2.5 log units, with the best having an RMSE of 2.1 log P units. A rough estimate of expected error for log P and log D is 1.54 log units. This comes from taking the mean RMSE of the top half of submissions in SAMPL4 Hydration Free Energy Prediction Challenge (1.5 kcal/mol) [155] and assuming the error in each phase is independent and equal to this value, yielding an expected error of 1.54 log P units [19]. Here, 64 log P challenge methods performed better than this threshold (58 blind predictions, 5 reference calculations, and the null prediction). However, only 10 of them were MM-based methods, with the lowest RMSE of 0.74 observed for method named Molecular-Dynamics-Expanded-Ensembles (*nh6c0*).

Challenge construction and experimental data availability are factors that contributed to the higher prediction accuracy observed in SAMPL6 compared to prior years. The log P challenge benefited from having a well-defined protonation state, especially for physical methods. Empirical methods benefited from the wealth of octanol-water training data. Accordingly, empirical methods were among the best performers here. But also, the chemical diversity represented by 11 compounds of the SAMPL6 log P challenge is very restricted and lower than the 53 small molecules in the SAMPL5 log D Challenge set. This was somewhat consistent with our expectations (discussed in Section 2.2.2)—that empirical, QM (with trained implicit solvation models), and mixed methods would outperform MM methods given their more extensive use of abundant octanol-water data in training (Figure 2.3).

2.5.2 Lessons learned from physical reference calculations

Comparison of reference calculations did not indicate a single force field or water model with dramatically better performance

As in previous SAMPL challenges, we conducted a number of reference calculations with established methods to provide a point of comparison. These included calculations with alchemical physical methods. Particularly, to see how the choice of water model affects accuracy we included three explicit solvent water models – TIP3P, TIP3P-FB and the OPC model – with the GAFF and SMIRNOFF force fields in our physical reference calculations. Deviations from experiment were significant (RMSE values ranged from 2.3 [1.1, 3.5] to 4.0 [2.7, 5.3] log units) across all the conditions used in the physical reference predictions (Figure 2.3A). In general, all the water models tend to overestimate the log P , especially for the carboxylic acid in the challenge set, SM08, though our calculations on this molecule had some specific difficulties we discuss further below. Relative to the TIP3P-FB and OPC water models, predictions which used TIP3P showed improvement in some of the error metrics, such as lower deviation from experiment with an RMSE range of 2.3 [1.1, 3.5] to 2.34 [1.0, 3.7] log units. The OPC and TIP3P-FB containing combinations had a higher RMSE range of 3.2 [2.0, 4.5] to 4.0 [2.7, 5.3] log units.

Physical reference calculations also included wet and dry conditions for the octanol phase using the GAFF and SMIRNOFF force field with TIP3P water. The wet octanol phase was composed of 27% water and dry octanol was modeled as pure octanol (0% water content). For reference calculations with the TIP3P water model the GAFF, and SMIRNOFF force fields using wet or dry octanol phases resulted in statistically indistinguishable performance. With GAFF, the dry octanol (*REF07*) RMSE was 2.4 [1.0, 3.7]. The wet octanol (*REF02*) RMSE was 2.3 [1.1, 3.5]. With SMIRNOFF, the dry octanol *REF08* RMSE was 2.4 [1.0, 3.7], with a wet octanol (*REF05*) RMSE of 2.3 [1.2, 3.5] (Table 2.10 and 2.11).

While water model and force field may have significantly impacted differences in performance across methods in some cases in this challenge, we have very few cases – aside from these reference calculations – where submitted protocols differed *only* by force field or water model, making it difficult to know the origin of performance differences for certain.

Different simulation protocols lead to different results between “equivalent” methods that use the same force field and water model

Several participants submitted predictions from physical methods which are equivalent to those used in our reference calculations and use the same force field and water model, which in principle ought to give identical results given adequate simulation time. There were three submissions which used the GAFF force field, TIP3P water model, and wet octanol phase: *6nmtt* (MD-AMBER-wetoct), *v2q0t* (InterX_GAFF_WET_OCTANOL), and *REF02* (YANK-GAFF-TIP3P-dry-oct). As can be seen in Figure 2.3A, *v2q0t* (InterX_GAFF_WET_OCTANOL) showed the best accuracy with an RMSE of 1.31 [0.94, 1.65]. *6nmtt* (MD-AMBER-wetoct) and *REF02* (YANK-GAFF-TIP3P-dry-oct) had higher RMSE values of 1.87 [1.33, 2.45] and 2.29 [1.07, 3.53], respectively. Two methods that used GAFF force field, TIP3P water model and wet octanol phase are *sqosi* (MD-AMBER-dryoct) and *REF07* (YANK-GAFF-TIP3P-dry-oct). These two also have an RMSE difference of 0.7 log P units. Although, in terms of overall accuracy there are differences, Figure 2.8 shows that in terms of individual predictions, submissions using the same force field and water model largely agree for most compounds.

Some discrepancies are observed for molecules SM13 and SM07, but are largest for SM08. For SM13 and SM07, method *v2q0t* (InterX_GAFF_WET_OCTANOL) performs over 1 log P unit better than *6nmtt* (MD-AMBER-wetoct). The rest of the predictions for these two methods differ by no more than about 1 log P unit, with the majority of the molecules differing by about 0.5 log P units or less from each other. Comparing *6nmtt* (MD-

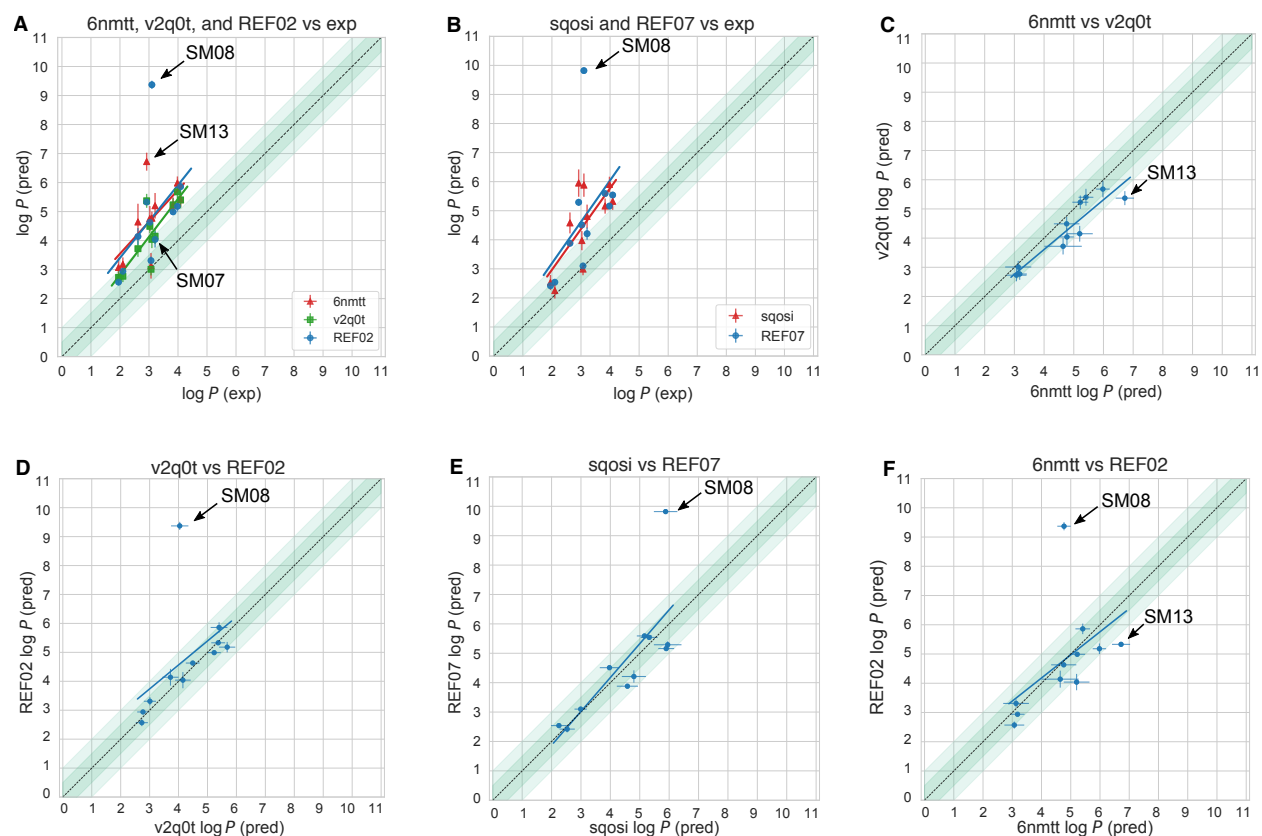


Figure 2.8: **Comparison of independent predictions that use seemingly identical methods (free energy calculations using GAFF and TIP3P water) shows significant systematic deviations between predictions for many compounds.** Comparison of the calculated and experimental values for submissions *v2q0t* (InterX_GAFF_WET_OCTANOL), *6nmtt* (MD-AMBER-wetoct), *sqosi* (MD-AMBER-dryoct) and physical reference calculations *REF02* (YANK-GAFF-TIP3P-wet-oct) and *REF07* (YANK-GAFF-TIP3P-dry-oct). (A) compares calculations that used wet octanol, and (B) compares those that used dry octanol. Plots C to F show the methods compared to one another. The dark and light-shaded region indicates 0.5 and 1.0 units of error, respectively.

AMBER-wetoct) vs *REF02* (YANK-GAFF-TIP3P-wet-oct) (Figure 2.8A), there is a substantial difference in the predicted values for molecules SM08 (4.6 log unit difference), SM13 (1.4 log unit difference), and SM07 (1.2 log unit difference). Method *v2q0t* (InterX_GAFF_WET_OCTANOL) and *6nmitt* (MD-AMBER-wetoct) perform about 5 log P units better than *REF02* (YANK-GAFF-TIP3P-wet-oct) for molecule SM08. Besides SM08, predictions from *v2q0t* (InterX_GAFF_WET_OCTANOL) and *REF02* (YANK-GAFF-TIP3P-wet-oct) differ by 0.5 log P units or less from each other. In dry octanol, *REF07* (YANK-GAFF-TIP3P-dry-oct) performs about 4 log P units worse than *sqosi* (MD-AMBER-dryoct) for SM08 (Figure 2.8B).

Submissions *6nmitt* (MD-AMBER-wetoct), *sqosi* (MD-AMBER-dryoct) and *v2q0t* (InterX_GAFF_WET_OCTANOL) used GAFF version 1.4 and the reference calculations used version 1.81, though GAFF differences are not expected to play a significant role here (i.e. only the valence parameters differ).

Selected small molecule state differences may have caused divergence between otherwise equivalent methods

In several of these approaches, users selected their own starting conformation, protonation state and tautomer, rather than those provided in the SAMPL6 challenge, so the differences here could possibly be attributed to differences in tautomer or resonance structures. Submissions *6nmitt* (MD-AMBER-wetoct) and *sqosi* (MD-AMBER-dryoct) used different tautomers for SM08 and different resonance structures for SM11 and SM14 (microstates SM08_micro010, SM11_micro005, SM14_micro001 from the previous SAMPL6 pK_a Challenge). We will discuss possible differences due to tautomer choice below in Section 2.5.2. The majority of the calculated log P values in *6nmitt* (MD-AMBER-wetoct), *sqosi* (MD-AMBER-dryoct), *v2q0t* (InterX_GAFF_WET_OCTANOL), *REF02* (YANK-GAFF-TIP3P-wet-oct), and *REF07* (YANK-GAFF-TIP3P-dry-oct) show the molecules having a

greater preference for octanol over water than the experimental measurements (Figure 2.8A, B). Methods *6nmitt* (MD-AMBER-wetoct) and *REF02* (YANK-GAFF-TIP3P-wet-oct) overestimate $\log P$ more than *v2q0t* (InterX_GAFF_WET_OCTANOL) (Figure 2.8A). Method *REF07* (YANK-GAFF-TIP3P-dry-oct) overestimates $\log P$ slightly more than *sqosi* (MD-AMBER-dryoct) (Figure 2.8B).

Three equivalent wet octanol methods and 2 equivalent dry octanol methods gave dissimilar results, and specific molecules were identified that show the major differences in predicted values (Figure 2.8C-F). GAFF and the TIP3P water model were used in all of these cases, but different simulation setups and codes were used, as well as different equilibration protocols and production methods. Submissions *6nmitt* (MD-AMBER-wetoct) and *sqosi* (MD-AMBER-dryoct), which come from the same group, used 10 ps NPT, 15 ns additional equilibration with MD, and Thermodynamic integration for production in their setup. Submission *v2q0t* (InterX_GAFF_WET_OCTANOL) used 200 ns of molecular dynamics to pre-equilibrate octanol systems, 10 ns of Temperature replica exchange in equilibration, and Isothermal-isobaric ensemble based molecular dynamics simulations in production. The reference calculations (*REF02* and *REF07*) were equilibrated for about 500 ns and used Hamiltonian replica exchange in production. Reference calculations performed with the IFE protocol and MD-AMBER-dryoct (*sqosi*) method used shorter equilibration times than the DFE protocol (*REF07*).

DFE and IFE protocols led to indistinguishable performance, except for SM08 and SM02

The direct transfer free energy (DFE) protocol was used for the physical reference calculations (*REF01-REF08*). Because the DFE protocol implemented in YANK [16] (which was also used in our reference calculations (*REF01-REF08*)) was relatively untested (see Section 2.9.3 for more details), we wanted to ensure it had not dramatically affected performance, so we

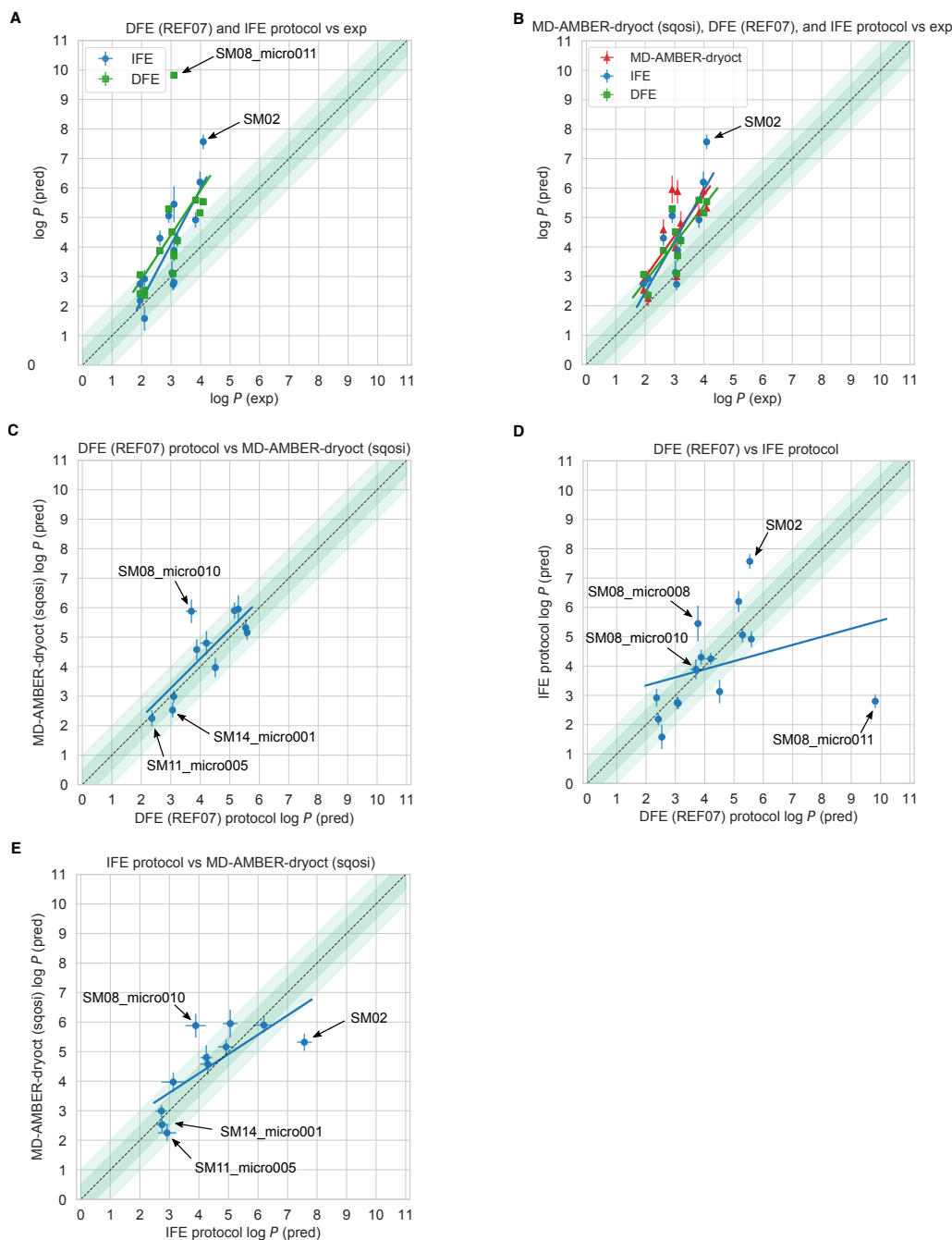


Figure 2.9: Comparison of predictions that use free energy calculations using GAFF and TIP3P water show deviations between predictions for the challenge molecules and several alternative tautomers and resonance structures. Deviations seem to largely stem from differences in equilibration amount and choice of tautomer. **A** compares reference direct transfer free energy (DFE, *REF07*) and indirect solvation-based transfer free energy (IFE) protocols to experiment for the challenge provided resonance states of molecules and a couple of extra resonance states for SM14 and SM11, and extra tautomers for SM08. **B** compares the same exact tautomers for submission *sqosi* (MD-AMBER-dryoct) and the two reference protocols to experiment. Submission *sqosi* (MD-AMBER-dryoct) used different tautomers than the ones provided in the challenge. **C-E** compares the calculated log P between different methods using the same tautomers. All of the predicted values can be found in Table 2.5.

compared it to the indirect solvation-based transfer free energy protocol (IFE) [20] protocol. The DFE protocol directly computed the transfer free energy between solvents without any gas phase calculation, whereas the IFE protocol (used in the blind submissions and some additional reference calculations labeled IFE) computed gas-to-solution solvation free energies in water and octanol separately and then subtracted to obtain the transfer free energy. The IFE protocol calculates the transfer free energy as the difference between the solvation free energy of the solute going from the gas to the octanol phase, and the hydration free energy going from the gas to the water phase. These protocols ought to yield equivalent results in the limit of sufficient sampling, but may have different convergence behavior.

Figure 2.9A shows calculations from our two different reference protocols using the DFE and IFE methods. We find that the two protocols yield similar results, with the exception of two molecules. Molecule SM08 is not substantially overestimated using the IFE protocol, where it is with the DFE protocol, and SM02 is largely overestimated by IFE, but not DFE (Figure 2.9A). The DFE (*REF07*) and IFE protocol both tend to overestimate the molecules' preference for octanol over water than in experiment, with the DFE protocol overestimating it slightly more. Figure 2.9D shows comparison of predicted $\log P$ values of the same tautomers by the DFE (*REF07*) and IFE protocols. The DFE and IFE protocols are almost within statistical error of one another, with the largest discrepancies coming from SM02 and SM08. The DFE and IFE protocols are in better agreement for some tautomers of SM08 more than others. They agree better on the predicted values for SM08_micro008 and SM08_micro010 than for SM08_micro011.

In the SAMPL6 blind submissions, there was a third putatively equivalent method to our reference predictions with the DFE protocol (*REF07*) and IFE protocol: *sqosi* (MD-AMBER-dryoct). It is identical in chosen force field, water model, and composition of octanol phase, however, different tautomers and resonance states for some molecules were used. All three predictions used free energy calculations with GAFF, TIP3P water, and a dry octanol phase.

Additionally, *sqosi* (MD-AMBER-dryoct) also used the more traditional indirect solvation free energy protocol. We chose to investigate the differences in these equivalent approaches by comparing predictions using matching tautomers and resonance structures (Figure 2.9). Figure 2.9B shows comparison of these three methods using predictions made with DFE and IFE protocols using identical tautomer and resonance input states as *sqosi* (MD-AMBER-dryoct): SM08_micro010, SM11_micro005, and SM14_micro001 (structures can be found in Figure 2.11). Except SM02, there is general agreement between these predictions. Figure 2.9C, other than the SM08_micro010 tautomer, predictions of DFE (*REF07*) and *sqosi* (MD-AMBER-dryoct) largely agree. Figure 2.9E highlights SM02 and SM08_micro010 predictions as the major differences between our predictions with IFE protocol and *sqosi* (MD-AMBER-dryoct).

Only results from the DFE protocol were assigned submission numbers (of the form *REF##*) and presented in the overall method analysis in Section 2.5.1. More details of the solvation and transfer free energy protocol can be found in section 2.9.3.

SM08 and SM13 were the most challenging for physical reference calculations

For the physical reference calculations category, some of the challenge molecules were harder to predict than others (Figure 2.10). Overall, the chemical diversity in the SAMPL6 Challenge dataset was limited. This set has 6 molecules with 4-amino quinazoline groups and 2 molecules with a benzimidazole group. The experimental values have a narrow dynamic range from 1.95 to 4.09 and the number of heavy atoms ranges from 16 to 22 (with the average being 19), and the number of rotatable bonds ranges from 1 to 4 (with most having 3 rotatable bonds). SM13 had the highest number of rotatable bonds and number of heavy atoms. This molecule was overestimated in the reference calculations. As noted earlier, molecule SM08, a carboxylic acid, was predicted poorly across all reference calculations. The origin of problems with molecule SM08 are discussed below in Section 2.5.2.

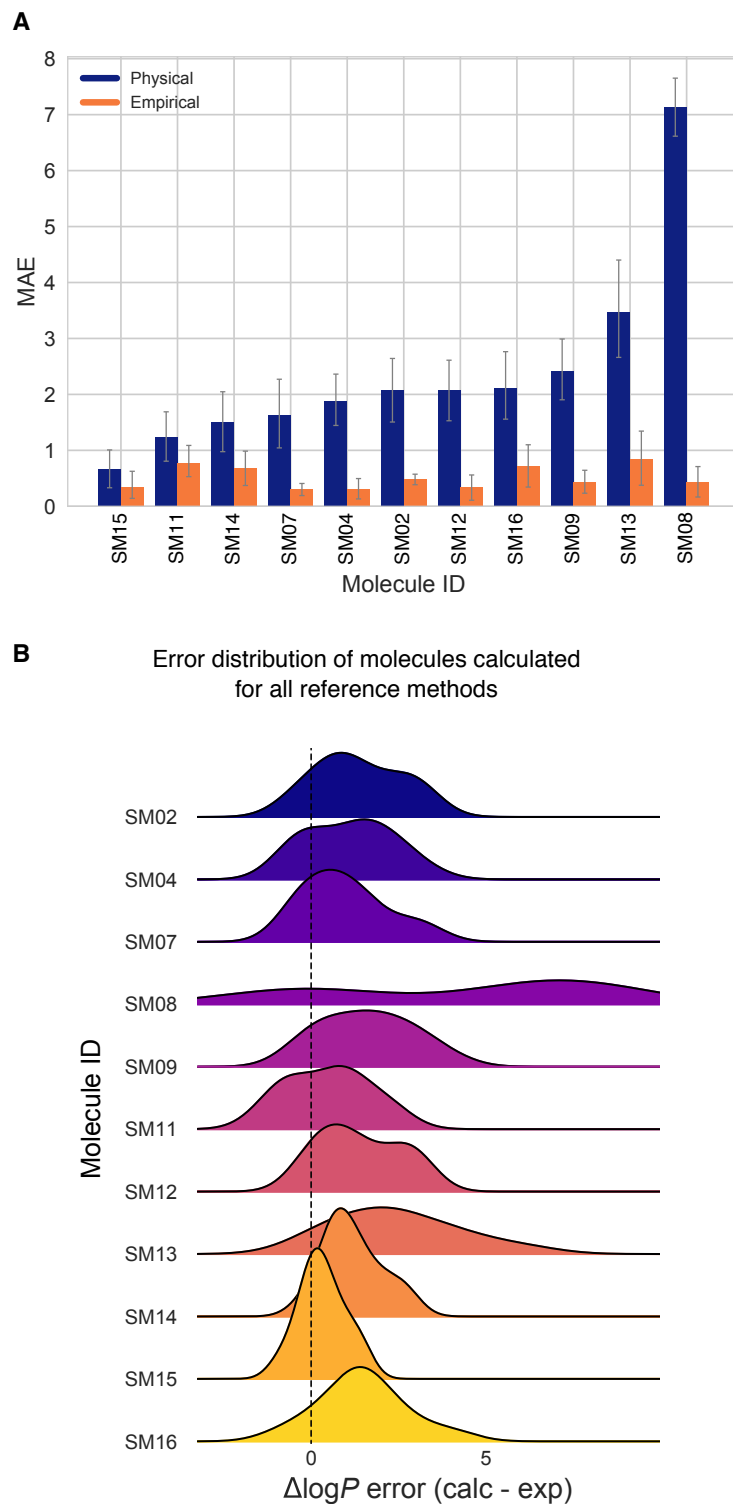


Figure 2.10: **The prediction errors per molecule indicate some compounds were more difficult to predict than others for the reference calculations category.** (A) MAE of each SAMPL6 molecule broken out by physical and empirical reference method category. (B) Error distribution for each molecule calculated for the reference methods. SM08 was the most difficult to predict for the physical reference calculations, due to our partial charge assignment procedure.

SM08 is a carboxylic acid and can potentially form an internal hydrogen bond. This molecule was greatly overestimated in the physical reference calculations. When this one molecule in the set is omitted from the analysis, log P prediction accuracy improves. For example, the average RMSE and R^2 values across all of the physical reference calculations when the carboxylic acid is included are 2.9 (2.3–4.0 RMSE range) and 0.2 (0.1–0.2 R^2 range), respectively. Excluding this molecule gives an average RMSE of 2.1 (1.3–3.3 RMSE range) and R^2 of 0.57 (0.3–0.7 R^2 range), which is still considerably worse than best-performing methods.

Choice of tautomer, resonance state, and assignment of partial charges impact log P predictions appreciably

Some physical submissions selected alternate tautomers or resonance structures for some compounds. Figure 2.11 shows three tautomers of SM08, and two alternative resonance structures of SM11 and SM14, all of which were considered by some participants. The leftmost structure of the alternate structure group of each molecule depicts the structure provided to participants.

Because some participants chose alternate structures, we explored how much variation in the selected input structures impacted the results. Particularly, for molecules SM14, SM11 and SM08, both AMBER MD protocols (submissions *sqosi* (MD-AMBER-dryoct) and *6nmtt* (MD-AMBER-wetoc)) used the SM14_micro001 microstate for SM14, SM11_micro005 for SM11 and SM08_micro010 for SM08, rather than the input structures provided as SMILES in the SAMPL6 log P Challenge instructions (See Figure 2.11 for depictions). The reference calculations and the submission from InterX (*v2q0t* (InterX_GAFF_WET_OCTANOL)) used the exact input structures provided as input SMILES for the challenge. Below, we refer to these several submissions as the MD-AMBER and InterX submissions.

To assess whether the choice of tautomer or resonance structure was important, we performed direct transfer free energy (DFE) and indirect solvation-based transfer free energy (IFE) [20] calculations for these alternate structures (please refer to section 2.5.2 for an explanation of the DFE and IFE methods). We —and the other participants utilizing these MM-based methods— assumed that the tautomers and resonance structures are fixed on transfer between phases and we did not do any assessments of how such populations might shift between octanol and water. Table 2.5 and Figure 2.9B compare $\log P$ calculations starting from the same input structures across the three methods: *sqosi* (MD-AMBER-dryoct), *REF07* (YANK-GAFF-TIP3P-dry-oct) which used the DFE protocol, and an additional set of calculations with the IFE protocol (using YANK, the GAFF force field, and TIP3P water just like *REF07*). The DFE protocol prediction set presented in Figure 2.9A is the same as *REF07* (YANK-GAFF-TIP3P-dry-oct), but includes extra tautomers for SM08, and extra resonance structures for SM11 and SM14.

From our comparison of our reference calculations and those with the InterX and MD-AMBER, we find that the choice of input tautomer has a significant effect on $\log P$ predictions. Particularly, within the traditional IFE method, our results indicate up to 2.7 log units variation between $\log P$ values for different tautomers of SM08 (between SM08_micro011, SM08_micro08 and SM08_micro10) (Table 2.5). Our exploration of these issues was prompted by the fact that the MD-AMBER protocols had utilized different tautomers than those initially employed in our physical reference calculations.

We also find that the choice of resonance structure affects calculated values, though less strongly so than the choice of tautomer. Within the IFE method we find 1.3 log units of variation between $\log P$ values calculated with different resonance structures of SM11 (SM11 and SM11_micro005) and 0.6 log units of variation between resonance structures of SM14 (SM14 and SM14_micro001) (Table 2.5).

We also find that the partial charge assignment procedure can also dramatically impact $\log P$

values for carboxylic acids (Table 2.13). Particularly, our calculations with the DFE and IFE protocols employed different partial charge assignment procedures as an unintentional feature of the protocol difference, as we detail below, and this impacted calculated $\log P$ values by up to 6.7 log units for SM08 (specifically SM08_micro011, the carboxylic acid in the set) compared to experiment. Particularly, the DFE protocol utilized antechamber for assigning AM1-BCC charges, whereas the IFE protocol used OpenEye’s quacpac. Antechamber utilizes the provided conformer (in this case, the *anti* conformation) for each molecule, whereas quacpac’s procedure computes charges for carboxylic acids in the *syn* conformation because this has been viewed as the relevant conformation, and because of concerns that the *anti* conformation might result in unusually large and inappropriate charges. Thus, because of this difference, the DFE and IFE protocols used dramatically different partial charges for these molecules (Table 2.13). Our results for SM08_micro011 (likely the dominant state) indicate that indeed, the conformer used for charging plays a major role in assigned charges and the resulting $\log P$ values (Table 2.13, Figure 2.15). We find our DFE protocol, which used the *anti* conformation for charging, overestimates the $\log P$ by about 6.7 log units, whereas the IFE protocol which used the *syn* conformation only overestimates it by about 0.3 log units. With the IFE method, we calculated a $\log P$ of 2.8 ± 0.2 for SM08, whereas with DFE method we obtained a value of 9.8 ± 0.1 (Table 2.5).

2.5.3 Lessons learned from empirical reference calculations

Empirical methods are fast and can be applied to large virtual libraries (100 000 cmps/min/CPU). This is in contrast to physical methods, which are often far more computationally demanding. Most of the empirical methods are among the top performers, with the exception of a few approaches that use descriptors and/or pre-factors that do not yield accurate $\log P$ predictions. Most empirical methods obtain RMSE and MAE values below 1 $\log P$ unit. The best empirical method achieved RMSE and MAE below 0.5 (*gmoq5*, Global XGBoost-Based

Table 2.5: **Predicted log P values of free energy calculations of methods using GAFF, TIP3P water, and dry octanol.** The methods listed are the reference direct transfer free energy (DFE) protocol, reference indirect solvation-based transfer free energy (IFE) protocol and submission *sqosi* (MD-AMBER-dryoct). Details of the two reference protocols can be found in Section 2.9.3. log P predictions for multiple tautomers (SM08) and resonance structures (SM11 and SM14) are listed, when available. The experimental values are provided for comparison. The same experimental log P values are stated for multiple tautomers or resonance structures. Potentiometric log P measurements do not provide information about the identity or populations of tautomers.

Molecule	Indirect Solvation-Based Transfer Free Energy (IFE) Protocol	REF07 Direct Transfer Free Energy (DFE) Protocol	<i>sqosi</i> (MD-AMBER-dryoct)	Experimental
SM02	7.6±0.3	5.5±0.1	5.3±0.3	4.09±0.03
SM04	6.2±0.4	5.2±0.1	5.9±0.3	3.98±0.03
SM07	4.3±0.2	4.2±0.3	4.8±0.4	3.21±0.04
SM08 ^{1,2}	2.8±0.2	9.8±0.1	-	3.10±0.03
SM08_micro008	5.5±0.6	3.8±0.1	-	3.10±0.03
SM08_micro010	3.9±0.3	3.7±0.2	5.9±0.4	3.10±0.03
SM09	3.1±0.4	4.51±0.03	4.0±0.3	3.0±0.1
SM11 ¹	1.6±0.4	2.5±0.1	-	2.10±0.04
SM11_micro005	2.9±0.3	2.36±0.01	2.3±0.3	2.10±0.04
SM12	4.9±0.3	5.6±0.1	5.2±0.3	3.83±0.03
SM13	5.1±0.3	5.3±0.1	6.0±0.5	2.92±0.04
SM14 ¹	2.2±0.2	2.4±0.1	-	1.95±0.03
SM14_micro001	2.8±0.2	3.1±0.1	2.5±0.3	1.95±0.03
SM15	2.7±0.2	3.1±0.1	3.0±0.2	3.07±0.03
SM16	4.3±0.3	3.9±0.1	4.6±0.4	2.62±0.01

The tautomer or resonance structure presented as the input SMILES for the SAMPL6 log P Challenge. It corresponds to the microstate SM08_micro011 of the SAMPL6 pK_a Challenge.

QSPR LogP Predictor). In all these cases, using a relatively large training set (>1000-10000 compounds) seems to be key.

The exact choice of method or descriptors seems to be less critical. Predictions based on atom or group contributions perform as well as those using either a small set of EHT-derived descriptors or a large set of diverse descriptors, sometimes additionally including fingerprint descriptors. A possible explanation could be that log P is, to first order, primarily an additive property so that empirical methods can do well since a wealth of octanol-water data is available for training. This is also reflected in the success of the simple methods summing up atom contributions. This approach may become problematic, however, when a functional group is present that was underrepresented or missing in the training set. In such cases, higher errors are expected.

As is true for the physical methods, empirical methods depend on the tautomeric state of

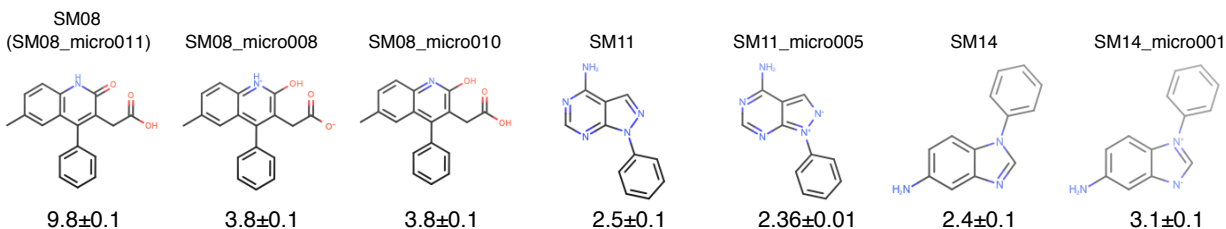


Figure 2.11: **The tautomer and resonance structure choice resulted in discrepancies in the reference calculations.** Shown here are calculated values for different input structures using the reference direct transfer free energy method. The uncertainties of the log *P* predictions were calculated as the standard error of the mean (SEM) of three replicate predictions. Structures labelled as SM08, SM11, and SM14 are based on input SMILES provided in SAMPL6 log *P* Challenge instructions. Three microstates shown for SM08 are different tautomers. SM08 (SM08_micro011) and SM08_micro010 are carboxylic acids, while SM08_micro008 is a carboxylate ion. SM08 (SM08_micro011) has a carbonyl group in the ring, while SM08_micro008 and SM08_micro010 have a hydroxyl in the ring. Structures pertaining to SM11 and SM14 are different resonance hybrids of the same tautomer (neutral microstate). Enumeration of all theoretically possible neutral tautomers of SAMPL6 molecules can be found in the SAMPL6 GitHub Repository (https://github.com/samplchallenges/SAMPL6/tree/master/physical_properties/pKa/microstates).

the compound. Here we have observed that clogP is particularly sensitive. clogP shifts of more than one log unit upon change of the tautomer are not uncommon. h_logP is much less sensitive to tautomers with shifts usually below 0.5 log *P* units. This is also true for molecule SM08, for which different tautomeric forms are possible (as seen in Figure 2.11). For the pyridone form of SM08 (SM08_micro011), clogP predicts a log *P* of 2.17, whereas the hydroxy-pyridine form (SM08_micro010) yields a log *P* of 3.63. For h_logP, the respective values are 3.09 and 3.06.

Despite the small training sets of the MOE models, good prediction for kinase inhibitor fragments and the extra compounds was achieved. This is possibly because the training set for this model was biased towards drug-like compounds, with substantial similarity to the SAMPL6 Challenge set.

Other studies have found that some empirical methods tend to overestimate log *P* when

molecular weight increases [163, 205]. In this challenge, this was less of an issue as molecular size remained relatively constant.

According to in-house experience at Boehringer-Ingelheim, different experimental $\log P$ measurement methods produce values that are correlated with one another with an R^2 value of around 0.7 (T. Fox, P. Sieger, unpublished results), indicating that experimental methods themselves can disagree with one another significantly. This is especially true when it comes to more approximate methods of estimating $\log P$ experimentally, such as HPLC-based methods [218, 192]. A dataset composed of 400 compounds from Boehringer-Ingelheim measured both with GLpKa and HPLC assays covering a range from 0-7 $\log P$ units had R^2 of 0.56, though in some cases these methods may have higher correlations with potentiometric approaches [76]. Thus, if an empirical model is trained on $\log P$ data from one particular method, testing it on data collected via another method may not yield performance as high as expected.

Here, all of the analyzed empirical reference methods achieved absolute error <2.0 , and often <1.5 calculated for each molecule in the SAMPL6 $\log P$ Challenge set. This is a sign of more consistent accuracy of the predictions across different molecules compared to physical methods. However, it is difficult to draw general conclusions given the small size of the data set, and many hypotheses being based on only one example.

2.5.4 Performance of reference methods on additional molecules

To broaden the analysis with a larger set with more chemical diversity and larger dynamic range of $\log P$ values, an extra set of 27 compounds were included in the analysis of reference calculations (Figure 2.12). These compounds had literature experimental $\log P$ values collected using the same method as the SAMPL6 dataset. This set is composed of substituted phenols, substituted quinolines, barbiturate derivatives and other pharmaceutically relevant

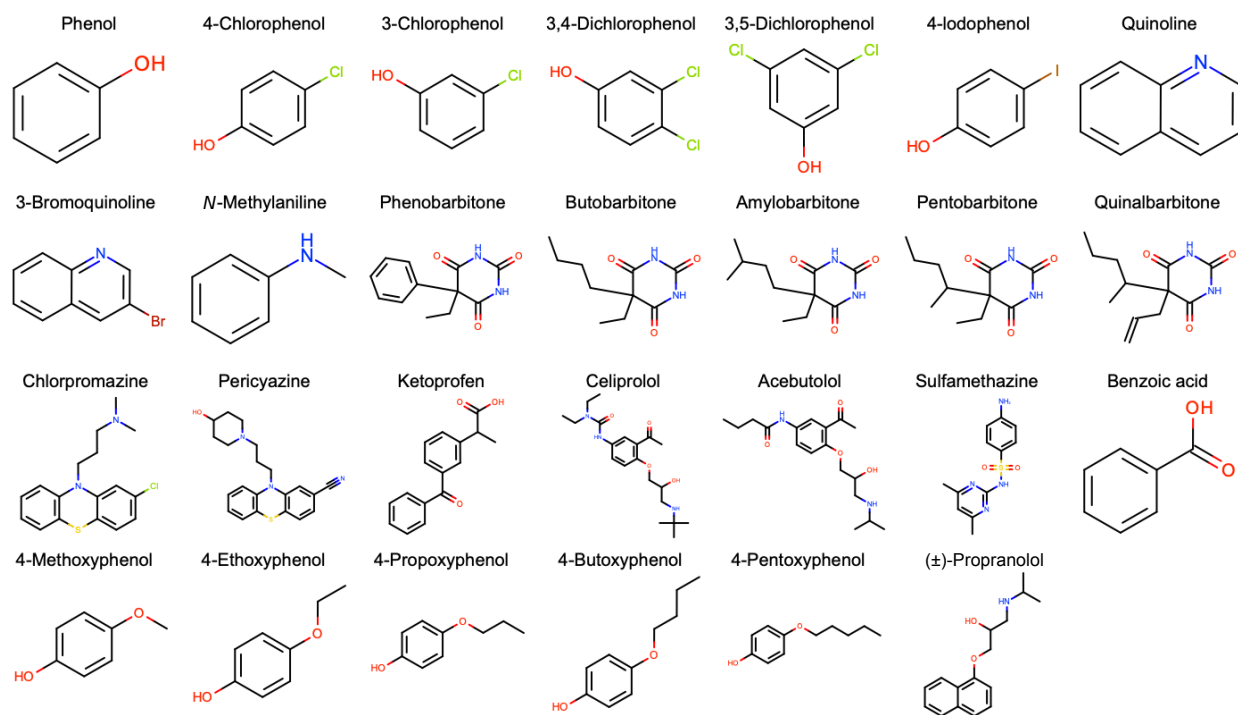


Figure 2.12: **Structures of the 27 additional molecules that were included in follow-up assessment of the reference methods.** These molecules were not included in the statistics overview.

compounds [202].

This set of molecules is larger and more diverse than the SAMPL6 challenge set, spanning a range of 4.5 log units compared to the challenge set which had a range of 2.1 log units. For this set, the number of rotatable bonds ranges from 0 to 12, with an average of 3 per compound. The number of heavy atoms ranges from 7 to 27, and the average per compound is 14. Most of the worst-performing compounds for the physical reference calculations had a higher number of heavy atoms – celiprolol (27), acebutolol (24) and pericyazine (26). Celiprolol and acebutolol both have the highest number of rotatable bonds in the set, 12 and 11 respectively. Chlorpromazine, pericyazine, and sulfamethazine all contain sulfur. Sulfur can in some cases pose particular challenges for force fields, especially hypervalent sulfur [149], which may account for the poor performance of pericyazine, chlorpromazine, and sulfamethazine. Pericyazine, one of the worst performing compounds, is also the only

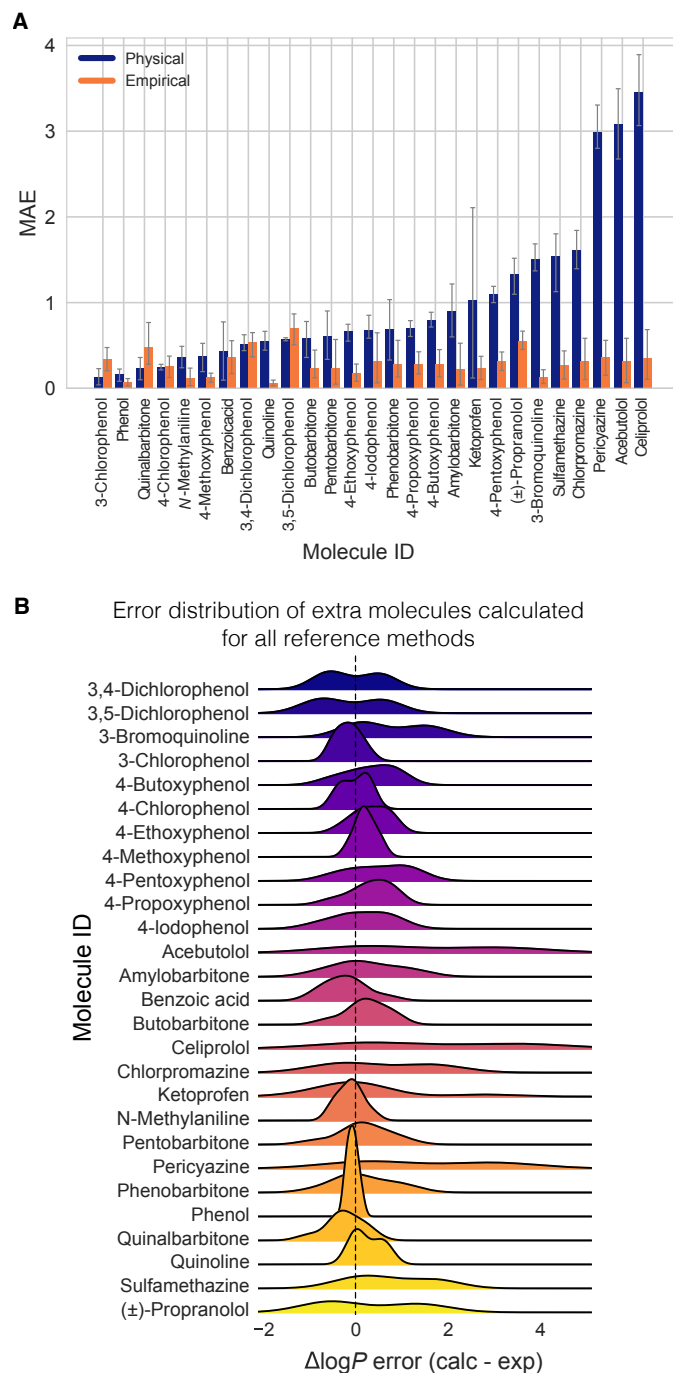


Figure 2.13: **Distribution of reference method calculation errors by molecule on our extra set shows that a few of the molecules were more challenging than others.** (A) MAE of each of the extra molecules broken out by physical and empirical reference method category. Majority of molecules have mean absolute errors below 1 $\log P$ unit for physical reference calculations. All of the mean absolute errors are well below 1 $\log P$ unit for empirical reference calculations. (B) Error distribution for each molecule calculated for the reference methods. A couple molecules have a significant tail showing probability of overestimated $\log P$ predictions.

Table 2.6: **Statistics of the physical and empirical reference method predictions on the extra test of molecules.** Methods were ranked according to increasing RMSE in this table. Performance statistics of MAE, R^2 , and Kendall’s Tau are also provided. Mean and 95% confidence intervals of all statistics are presented.

ID	Method Name	Category	Type	RMSE	MAE	R^2	Kendall’s Tau (τ)
EXT09	clogP (Biobyte)	Empirical	Reference	0.23 [0.16, 0.29]	0.17 [0.12, 0.23]	0.94 [0.86, 0.98]	0.85 [0.75, 0.93]
EXT12	MoKa_logP	Empirical	Reference	0.28 [0.20, 0.35]	0.22 [0.16, 0.28]	0.91 [0.82, 0.97]	0.83 [0.73, 0.91]
EXT11	logP(o/w) (MOE)	Empirical	Reference	0.32 [0.22, 0.41]	0.24 [0.17, 0.33]	0.90 [0.81, 0.95]	0.80 [0.66, 0.91]
EXT10	h_logP (MOE)	Empirical	Reference	0.43 [0.34, 0.51]	0.35 [0.26, 0.45]	0.83 [0.62, 0.93]	0.74 [0.55, 0.90]
EXT13	SlogP (MOE)	Empirical	Reference	0.59 [0.48, 0.69]	0.49 [0.36, 0.61]	0.71 [0.41, 0.87]	0.55 [0.33, 0.73]
EXT08	YANK-SMIRNOFF-TIP3P-dry-oct	Physical (MM)	Reference	1.26 [0.88, 1.60]	0.97 [0.69, 1.28]	0.56 [0.16, 0.83]	0.50 [0.23, 0.73]
EXT07	YANK-GAFF-TIP3P-dry-oct	Physical (MM)	Reference	1.27 [0.69, 1.74]	0.88 [0.57, 1.26]	0.55 [0.19, 0.88]	0.60 [0.34, 0.81]
EXT02	YANK-GAFF-TIP3P-wet-oct	Physical (MM)	Reference	1.38 [0.94, 1.78]	1.03 [0.70, 1.40]	0.58 [0.26, 0.83]	0.58 [0.35, 0.78]
EXT05	YANK-SMIRNOFF-TIP3P-wet-oct	Physical (MM)	Reference	1.50 [0.96, 1.98]	1.11 [0.75, 1.52]	0.50 [0.13, 0.81]	0.54 [0.29, 0.75]

molecule in the set that has a nitrile.

In the physical reference calculations, the mean absolute errors are below 1 log P unit for dry octanol conditions and below 1 log P units for wet octanol conditions (Table 2.6). The calculated log P values had an average RMSE of 1.4 (RMSE range of 1.3 [0.9, 1.6] to 1.5 [1.0, 2.0]), and an average R^2 of 0.5 (with a correlation range of 0.5 [0.1, 0.8] to 0.6 [0.3, 0.8]). Physical methods are on par with empirical ones for the smaller, less flexible compounds, but in general are worse, especially for compounds with long flexible hydrophilic tails. The exception is chlorpromazine, but the smaller error seen in this molecule might be due to an error compensation caused by the presence of the sulfur atom since force fields have challenges with sulfur-containing compounds. [149].

Empirical methods are more stable in the sense that there are no gross outliers found in the extended set. For the empirical reference calculations, the absolute errors for the 27 extra compounds are all below 1 log P unit. For clogP, most compounds have errors below 0.4 log P unit, with only (\pm)-propranolol a bit higher. Compound 3,5-dichlorophenol and 3,4-dichlorophenol consistently had a slightly higher error; there is no obvious correlation between method performance and size or complexity of the compounds. Figure 2.13A shows that 3,5-dichlorophenol, 3,4-dichlorophenol, and (\pm)-propranolol were the most challenging

compounds for empirical reference methods. The MAE calculated for these three molecules as the average of five methods (*EXT09*, *EXT10*, *EXT11*, *EXT12*, *EXT13*) was higher than $0.5 \log P$ units (Table 2.6). RMSE overall compounds between 0.23 for clogP and 0.59 for MOE_SlogP, significantly below the best physical model. This is mirrored in the Kendall tau values, where the best empirical method (clogP) achieves 0.85, whereas the best physical methods are comparable to the worst empirical method with a value of 0.55.

When prediction performance of empirical prediction methods for this dataset and for the SAMPL6 Challenge set are compared, we observe better prediction accuracy for this set, with an RMSE range of 0.2 [0.2, 0.3] to 0.6 [0.5,0.7] for the extra molecules and 0.5 [0.2, 0.8] to 0.8 [0.5, 1.1] for the challenge molecules. The average R^2 was 0.6 (with a correlation range of 0.38 [0.01, 0.82] to 0.7 [0.3, 0.9]). This may be due to SAMPL6 compounds being more challenging, or it may be that these extra molecules appear in the training sets used in developing empirical methods.

2.5.5 Take-away lessons from the SAMPL6 Challenge

Empirical and QM-based prediction methods represented in SAMPL6 Challenge in general performed better than MM-based methods. Ten empirical and QM-based methods achieved an RMSE $< 0.5 \log P$ units. The lowest RMSE observed for MM-based physical methods was 0.74 and the average RMSE of the better half of MM-based methods was $1.44 \log P$ units. However, the RMSE of the best two MM-based methods was similar to the null model, which simply guessed that all compounds had a constant, typical $\log P$ value.

For MM approaches, prediction accuracy varied based on methodological choices such as simulation method, equilibration protocol, free energy estimation method, force field, and water model. Only a small number of MM-based physical models achieved an accuracy similar to the null model, which had an RMSE of 0.8. Some MM methods outperformed the

null model, but such performance was variable across approaches and not clearly linked to a single choice of force field, water model, etc. Polarizable force fields also did not provide an advantage for $\log P$ predictions, possibly due to solute simplicity and the absence of formal charge, or because other sources of error dominated.

Analysis of the MM-based reference calculations highlighted equilibration and charging protocols, sampling challenges, identification of the dominant neutral tautomer, and selection of input resonance states as confounding factors. Comparison of equivalent calculations from independent participants (identical methods such as free energy calculations using GAFF and TIP3P water with different setups and code) showed significant systematic deviations between predicted values for some compounds. The comparison of identical methods also showed that the tautomer and resonance state choice for some molecules resulted in discrepancies in calculated $\log P$ values. In one case, conformation selected for a carboxylic acid before charging was important. We have also noticed differences in equilibration protocols, which could be particularly important for the octanol phase, though the present challenge does not conclusively demonstrate that differences in equilibration made a significant difference.

Fast empirical methods showed greater consistency of prediction accuracy across test molecules compared to physical methods. Most of the empirical methods were among the better-performing methods. The size of the training sets seems to be more important for accuracy than the exact methods and descriptors used in each model. Although not observed in the SAMPL6 Challenge set, empirical methods may experience problems if a functional group is underrepresented in training sets. Just like the physical methods, the choice of tautomer makes a difference. For example, shifts greater than 1 log unit in the calculated $\log P$ of different tautomers are common.

Performance in the SAMPL6 $\log P$ challenge was generally better than in the SAMPL5 $\log D$ Challenge. The change of partition solvent from cyclohexane to octanol, absence of

protonation state effects, and smaller chemical diversity represented in the challenge are likely reasons. In the SAMPL5 log D Challenge, only five submissions had an RMSE below 2.5 log units, while here, 10 methods achieved an RMSE ≤ 0.5 log P units and many of the submissions had an RMSE ≤ 1.0 log P units. The design of the SAMPL6 log P Challenge removed some of the factors confounding accuracy in the earlier log D challenge, namely pK_a prediction and cyclohexane (a challenging solvent for empirical methods).

Compared to expected accuracy for partition coefficients based on SAMPL4 Challenge performance, many QM-based methods were better while only a small number of MM-based methods achieved slightly better results. In SAMPL4, the top-performing hydration free energy predictions had an error of about 1.5 kcal/mol, which would yield an expected error here (assuming independent errors/no error cancellation) of about 1.54 log units [19], if log P values were estimated from a difference in solvation free energies. Many physical methods achieved roughly this accuracy or slightly better.

Partition coefficient predictions can also serve, for physical calculations, as a model system that reflects how well solvation effects can be captured by the same techniques developed for protein-ligand binding predictions – where solvation also plays a role in calculations. Relative binding free energy calculations tend to achieve errors, in the best-case scenario, in the 1–2 kcal/mol range [44], or about 1.03–2.06 log units if similar accuracy were achieved here for solvation in each phase (with independent errors). Many methods did better than 2 log P units of error in this challenge, which is in agreement with the expectation that partition coefficients present an easier model system compared to protein-ligand binding affinities.

Performance of empirical methods far surpassed these thresholds taking advantage of the available octanol-water experimental data, however, these empirical techniques are specifically oriented towards predicting partitioning and cannot be applied to the binding problem.

2.5.6 Suggestions for the design of future challenges

In the SAMPL6 Challenge, the log P focus proved helpful to allow a focus on modeling of solvation effects without the complexities of modeling different protonation states present in a log D challenge. Challenges which focus on specific aspects of modeling help isolate methodological problems, making challenges like log P and log D modeling particularly helpful. We believe the largest benefits to the field will be achieved from iterated challenges, as seen from the progress achieved in predicting hydration free energies over multiple SAMPL challenges [155].

As MM-based physical methods struggled with octanol-water log P predictions in SAMPL6, we recommend additional SAMPL iterations focused on log P with larger datasets and more chemical diversity to facilitate progress. The conclusions of SAMPL6 pK_a and log P Challenges indicate that, if this had been posed as a log D challenge rather than a log P challenge, larger pK_a prediction errors would have masked underlying issues in predicting equilibrium partitioning of neutral solutes between solvent phases. The fact that performance for physical methods was still relatively poor illustrates the potential benefit of future log P challenges.

For near-term challenges, we would like to keep the level of difficulty reasonable by keeping the focus on smaller and fragment-like compounds and limiting the number of non-terminal rotatable bonds (maximum of 6) similar to SAMPL5. The SAMPL5 Challenge suggested that molecules with many rotatable bonds still pose challenges for contemporary methods, suggesting this is a criterion for difficulty. However, in later challenges we hope to gradually increase the difficulty of the compounds considered to provide a more diverse set that includes more difficult compounds including varying numbers of rotatable bonds.

Ideally, a more diverse combination of functional groups in the compounds should be included in future sets, with improved chemical diversity posing more challenges and also helping

provide additional lessons learned. For example, a dataset could include matched molecular pairs which differ by only a single functional group, helping to isolate which functional groups pose particular challenges. Current MM-based methods are known to often have difficulty modeling sulfonyl and sulfonamide groups, but a challenge utilizing matched molecular pairs could reveal other such challenging functional groups. In addition, expanding partition coefficient challenges with a diverse set of solvent phases would be beneficial for improving solute partitioning models.

The statistical power of the SAMPL6 log P Challenge for comparative method evaluation was limited due to the narrow experimental data set with only 2 log P units of dynamic range and 11 data points, both of which were driven by limitations of the experimental methodology chosen for this challenge [88]. Future log P challenges would benefit from larger blind datasets with a broader dynamic range. We recommend at least a log P range of 1–5. The potentiometric log P measurement method used for the collection for SAMPL6 data was rather low throughput, requiring method optimization for each molecule. High-throughput log D measurement methods performed at pHs that would ensure neutral states of the analytes may provide a way to collect larger datasets of log P measurements. However, this approach poses some challenges. First, it is necessary to measure pK_a values of the molecules first. Second, partitioning measurements need to be done at a pH that guarantees that the compound has neutral charge, in which case solubility will be lower than if it is charged and may become a limitation for the experiment.

SAMPL6 log P Challenge molecules were not expected to have multiple tautomers affecting log P predictions (based on QM predictions). The choice of the challenge set also ensured participants did not have to calculate contributions of multiple relevant tautomerization states or shifts in tautomerization states during transfer between phases. However, participants still had to select a major tautomer for each compound. To evaluate the tautomer predictions in the future, experimental measurement of tautomer populations in each sol-

vent phase would provide valuable information. However, such experimental measurements are difficult and low throughput. If measuring tautomers is not a possibility, the best approach may be to exclude compounds that present potential tautomerization issues from the challenge, unless the challenge focus is specifically on tautomer prediction.

Overall, for future solute partitioning challenges, we would like to focus on fragment-like compounds, matched molecular pairs, larger dynamic range, larger set size, and functional group diversity.

2.6 Conclusion

Several previous SAMPL challenges focused on modeling solvation to help address this key accuracy-limiting component of protein-ligand modeling. Thus, the SAMPL0 through SAMPL4 challenges included hydration free energy prediction as a component, followed by cyclohexane-water distribution coefficient in SAMPL5.

Here, a community-wide blind partition coefficient prediction challenge was fielded for the first time, and participants were asked to predict octanol-water partition coefficients for small molecules resembling fragments of kinase inhibitors. As predicting $\log D$ in the previous challenge was quite challenging due to issues with pK_a prediction, the present challenge focused on $\log P$, avoiding these challenges and placing it at roughly the right level of complexity for evaluating contemporary methods and issues they face regarding the modeling of small molecule solvation in different liquid phases. The set of molecules selected for the challenge were small and relatively rigid fragment-like compounds without tautomerization issues which further reduces the difficulty of the prospective prediction challenge.

Participation in the challenge was much higher than in SAMPL5, and included submissions from many diverse methods. A total of 27 research groups participated, submitting 91 blind

submissions in total. The best prospective prediction performance observed in SAMPL6 log P Challenge came from QM-based physical modeling methods and empirical knowledge-based methods, with 10 methods achieving an RMSE below 0.5 log P units. On the other hand, only a small number of MM-based physical models achieved an accuracy similar to the null model (which predicted a constant, typical log P value), which had an RMSE of 0.8. Empirical predictions showed performance which was less dependent on the compound/dataset than physical methods in this study. For empirical methods, the size and chemical diversity of the training set employed in developing the method seems to be more important than the exact methods and descriptors employed. We expected many of the empirical methods to be the top performers, given the wealth of octanol-water log P training data available, and this expectation was borne out.

Better prediction performance was seen for octanol-water log P challenge than the SAMPL5 cyclohexane-water log D challenge. In addition to absence of pK_a prediction problem for the partition system, the molecules in the SAMPL6 log P Challenge were considerably less diverse than in the SAMPL5 log D Challenge, which may have also affected relative performance in the two challenges. Physical methods fared slightly better in this challenge than previous cyclohexane-water log D challenge, likely because of the elimination of the need to consider protonation state effects. However, MM-based physical methods with similar approaches did not necessarily agree on predicted values, with occasionally large discrepancies resulting from apparently relatively modest variations in protocol.

All information regarding the challenge structure, experimental data, blind prediction submission sets, and evaluation of methods is available in the SAMPL6 GitHub Repository to allow follow up analysis and additional method testing.

Overall, high participation and clear lessons learned pave the way forward for improving solute partitioning and biomolecular binding models for structure-based drug design.

2.7 Author Contributions

Conceptualization, MI, TDB, JDC, DLM ; Methodology, MI, TDB, DM, JDC ; Software, MI, TDB, AR ; Formal Analysis, MI, TDB ; Investigation, MI, TDB, DLM, TF; Resources, JDC, DLM; Data Curation, MI, TDB ; Writing-Original Draft, MI, TDB, DLM, TF; Writing - Review and Editing, MI, TDB, DLM, TF, JDC, AZ; Visualization, MI, TDB ; Supervision, DLM, JDC ; Project Administration, MI ; Funding Acquisition, DLM, JDC, MI, TDB.

2.8 Acknowledgments

We would like to thank OpenEye, especially Gaetano Calabró, for help with Orion, and for constructing the Orion workflows partially utilized here. We would like to thank experimental collaborators Timothy Rhodes (ORCID: 0000-0001-7534-9221), Dorothy Levorse, and Brad Sherborne (ORCID: 0000-0002-0037-3427).

MI and JDC acknowledge support from the Sloan Kettering Institute. JDC acknowledges partial support from NIH grant P30 CA008748. MI, TDB, JDC, and DLM gratefully acknowledge support from NIH grant R01GM124270 supporting the SAMPL Blind Challenges. MI acknowledges support from a Doris J. Hutchinson Fellowship during the collection of experimental data. TDB acknowledges support from the ACM SIGHPC/Intel Fellowship. DLM appreciates financial support from the National Institutes of Health (1R01GM108889-01) and the National Science Foundation (CHE 1352608). We acknowledge contributions from Caitlin Bannan who provided feedback on experimental data collection and structure of log P challenge from a computational chemist’s perspective. MI and JDC are grateful to OpenEye Scientific for providing a free academic software license for use in this work. TF thanks BioByte, MOE, and Molecular Discovery for allowing us to include log P predictions calculated by their software in this work as empirical reference calculations.

2.9 Supplementary Information

2.9.1 Overview of Supplementary Information

Contents of Supplementary Information

- **Code and Data Availability**
- **Detailed methods section:**
 1. Physical reference calculations - Direct Transfer Free Energy Approach
 2. Physical reference calculations - Indirect Solvation-Based Transfer Free Energy Approach
 3. Empirical reference calculations
- **Table 2.7** Method details of log P predictions with MM-based physical methods.
- **Table 2.8** SMILES and InChI identifiers of SAMPL6 log P Challenge molecules.
- **Table 2.9** SMILES and InChI identifiers of extra molecules included in the evaluation of reference methods.
- **Table 2.10 and Table 2.11** Evaluation statistics calculated for all methods.
- **Table 2.12** Comparison of force field parameters of the TIP3P, TIP3P-FB and OPC water models
- **Table 2.13** Comparison of the charges assigned to the syn and anti conformation of SM08_micro011 in the DFE protocol
- **Figure 2.14:** Varying the amount of water in the octanol phase has no significant effect on the predicted log P in reference calculations.

- **Figure 2.15** 2D and 3D structures of SM08_micro011 with the carboxylic acid in “anti” and “syn” conformation.
- **Figure 2.16** For the DFE method, the starting conformation impacts the number of C-O dihedral transitions for SM08_micro011.

Additional supplementary files *SAMPL6-supplementary-documents.tar.gz* file includes:

- An archive copy of the log *P* Challenge directory of SAMPL6 GitHub Repository (*SAMPL6-repository-logP-directory.zip*)
- SAMPL6 log *P* Challenge Instructions (*logP_challenge_instructions.md*)
- Table 2.7 in CSV format (*SI-table-MM-method-details.csv*)
- Table 2.8 in CSV format (*SAMPL6-logP-chemical-identifiers-table.csv*)
- Table 2.9 in CSV format (*extra-chemical-identifiers-table.csv*)
- Table 2.10 and Table 2.11 in CSV format (*statistics.csv*)
- The free energy and enthalpy values of each phase in triplicate and comparisons of calculated solvation free energies across trials for the physical reference calculations (*analysis-of-physical-reference-calculations.zip*)
- Scripts related to the physical reference calculations (*physical-reference-calculation-scripts.zip*)

2.9.2 Code and Data Availability

All SAMPL6 log *P* Challenge instructions, submissions, experimental data and analysis are available at

https://github.com/samplchallenges/SAMPL6/tree/master/physical_properties/logP.

An archive copy of SAMPL6 GitHub Repository log P Challenge directory is also available in the Supplementary Documents bundle (*SAMPL6-supplementary-documents.tar.gz*).

Some useful files from this repository are highlighted below.

- Table of participants and their submission filenames:

https://github.com/samplchallenges/SAMPL6/blob/master/physical_properties/logP/predictions/SAMPL6-user-map-logP.csv

- Table of methods including submission IDs, method names, participant assigned method category, and reassigned method categories:

https://github.com/samplchallenges/SAMPL6/blob/master/physical_properties/logP/predictions/SAMPL6-logP-method-map.csv

- Submission files of prediction sets:

https://github.com/samplchallenges/SAMPL6/tree/master/physical_properties/logP/predictions/submission_files

- Python analysis scripts and outputs:

https://github.com/samplchallenges/SAMPL6/blob/master/physical_properties/logP/analysis_with_reassigned_categories/

- Table of performance statistics calculated for all methods:

https://github.com/samplchallenges/SAMPL6/blob/master/physical_properties/logP/analysis_with_reassigned_categories/analysis_outputs_withrefs/StatisticsTables/statistics.csv

2.9.3 Detailed methods

Physical reference calculations - Direct Transfer Free Energy Approach

$\log P$ can be estimated directly from the transfer free energy of a solute moving from the organic to the aqueous layer. Specifically, we calculate the transfer free energy from the difference in solvation free energy into octanol and hydration free energy. $\log P$ is directly proportional to the difference between the solvation free energy for the solute into each solvent

$$\log P = -\frac{\Delta G_{transfer}}{RT \ln 10} = \frac{\Delta G_{solvation} - \Delta G_{hydration}}{RT \ln 10} \quad (2.4)$$

where $\Delta G_{transfer}$ is the transfer free energy, $\Delta G_{solvation}$ is the solvation free energy of the solute going from the gas to the octanol phase, $\Delta G_{hydration}$ is the hydration free energy going from the gas to the water phase, R is the gas constant (8.314 J / mol · K) and T is the temperature (298.15 K).

The direct transfer free energy (DFE) protocol that was used for the physical reference calculations (*REF01-REF08*, *EXT02*, *EXT05*, *EXT07*, *EXT08*) directly computed the transfer free energy between solvents without any gas phase calculation, whereas the IFE protocol (discussed in Section ??) computed gas-to-solution phase solvation free energies in water and octanol separately and then subtracted to obtain the transfer free energy.

To explore how solvent mixing would effect predicted values, water was included in the octanol phase for the majority of the reference calculations. A portion of the calculations treated the octanol and water phase as completely immiscible for comparison. The experimental mole fraction of water in octanol was measured as 0.2705 [117]. The solutions and

calculations modeled each phase at infinite dilution, with only a single solute molecule in each solvent.

The initial input files were made using the Solvation Toolkit (<https://github.com/MobleyLab/SolvationToolkit>), which converts SMILES strings to parameterized molecules and builds topology and coordinate files for use in molecular dynamics software packages. Solvation Toolkit is a driver utility that utilizes the OpenEye toolkits (version 2018.10.1) for cheminformatics (specifically file conversion and handling of molecular identities), and OEChem for reading and writing files. AmberTools [34] was used to parameterize systems with the General AMBER Force Field for organic molecules (version 2017.1.81) and water was parameterized with the TIP3P water model, AM1-BCC charges were assigned via Antechamber, Packmol (version 18.169) [138] was used to build boxes, and lastly AMBER topology and coordinate files were made with LEaP.

The SMILES string and the mole fraction of each compound in the system were used as input. The “wet” octanol systems were generated using a mole fraction of 0.7295 for octanol and 0.2705 for water, producing systems with about 200 octanol molecules and 74 water molecules, depending on the solute size. The “dry” octanol systems had no water component and about 211 octanol molecules. All of the water systems had 1497 molecules. The box dimensions were about 40x40 Å in all cases.

The following equilibration stages were carried out using the GAFF forcefield (version 2017.1.81), the TIP3P water model and OpenMM (version 7.3.1) [55, 58], a molecular simulation toolkit.

For minimization, an energy tolerance of 10 kilojoules/mole was used and the systems were minimized until convergence was reached. A Langevin integrator was used with a 0.5 fs timestep. Minimization was followed by 100 ps of NVT using a Langevin integrator and 1.0 fs timestep, 100 ps of NPT using a Langevin integrator and 2.0 fs timestep, and lastly 500

ns of NPT using a Langevin integrator and 2.0 fs timestep.

Three independent equilibrations were run starting from water and octanol phase systems of the initial setup, in order to obtain three different sets of starting coordinates for replicate transfer free energy calculations with YANK (version 0.24.0 [184]). The protocol for creating systems with different force field and/or water model conditions (2.1) is detailed below.

Following equilibration, the resulting systems were saved to PDBs. For each solvent system, a ParmEd Structure was created using the topology and positions from the equilibrated PDB, with parameters coming from the original GAFF/TIP3P OpenMM System. The ParmEd structure of the system was split into individual components or structures and then used to create newly parameterized OpenMM systems. The water was parameterized with either the TIP3P, TIP3P-FB or OPC water model, and the solute and solvent were parameterized with the SMIRNOFF force field (smirnoff99Frosst version 1.0.7) or remained parameterized with GAFF. In just the OPC case, a dummy atom was added to the water component structure. After parameterization, the OpenMM systems of the solute-octanol and water were converted back to ParmEd structures which maintained their new parameters. The final OpenMM system was created using the particle mesh Ewald (PME) method for periodic boundary conditions, an error tolerance of $1e-4$ and a cutoff for nonbonded interactions was set to 11 Å.

The resulting OpenMM Systems were saved as XMLs for use later on. Prior to using YANK, the new systems were briefly equilibrated using the same setup described previously, excluding the 500 ns of NPT. The final equilibrated PDB and system XML files were used as input files for solvation and transfer free energy calculations with YANK [16], a toolkit that uses Hamiltonian replica exchange and can compute solvation free energies. For the YANK simulations, hydrogen mass repartitioning (HMR) was used to allow a 3 fs timestep. HMR works by slowing down the fastest motions in the simulation by reallocating mass from the connected heavy atom to the hydrogens [84]. The temperature was set to 298.15 K (the

experimental temperature), the pressure to 1.0 atm, and an anisotropic dispersion cutoff of 12.0 Å was used. There were 5000 iterations total and 335 steps per iteration. The overall length of the YANK simulations were 5 ns for each replica.

In the octanol and water phase the electrostatic interactions of the solute with the solvent were scaled off through a λ (lambda) parameter using the following lambda values where $\lambda = [1.00, 0.75, 0.50, 0.25, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00]$, and steric interactions were scaled using $\lambda = [1.00, 1.00, 1.00, 1.00, 1.00, 0.95, 0.90, 0.80, 0.70, 0.60, 0.50, 0.40, 0.35, 0.30, 0.25, 0.20, 0.15, 0.10, 0.05, 0.00]$.

The direct transfer free energy was obtained from YANK [16], where the $\Delta G_{transfer}$ was equivalent to $\Delta G_{octanol} - \Delta G_{water}$. This was then converted to $\log P$ using Equation 2.4. The uncertainties of the $\log P$ predictions were calculated as the standard error of the mean (SEM) of three replicate predictions. The SEM was estimated as $SEM = \sigma/\sqrt{N}$ where σ is the sample standard deviation and N is the size of the sample (in this case the number of replicate predictions made). The model uncertainty was reported as 1.6 log units, based on similar previous work [20].

Physical Reference Calculations - Indirect Solvation-Based Transfer Free Energy Approach

We ran an additional set of reference calculations using a more traditional indirect solvation-based transfer free energy method to see how it would compare to the direct transfer free energy method (described in Section 2.9.3). Specifically, the IFE protocol calculates the transfer free energy as the difference between the solvation free energy of the solute going from the gas to the octanol phase, and the hydration free energy going from the gas to the water phase. The direct transfer free energy method that was run with YANK had not computed gas-to-water transfer free energies as previous work had done when computing

$\log P$ and $\log D$ values and, while in principle this should be an unimportant methodological detail, we wanted to assess whether this choice had negatively impacted results. Thus, we ran the indirect solvation-based transfer free energy protocol described below.

The set of indirect solvation-based alchemical free energy calculations were run using OpenEye’s Orion cloud computing platform, also with the YANK software but with an alternate, fully automated workflow. The Orion workflow utilizes a very similar approach to that utilized above, except that it employs solvation free energy calculations for each molecule in each phase, rather than computing transfer free energies. Details of equilibration and simulation length are also different, as described below – with the largest difference being equilibration protocol.

On Orion, the input for each calculation was the target solute (SMILES) and the target solvent (SMILES), along with the temperature (298.15 K) and a guessed initial density for each solution (here 1.0 g/mL for solutes in water, 0.83 g/mL for octanol, to match experiment roughly). These settings are used on Orion, to prepare initial simulations via an internal Orion workflow based on that used in SolvationToolkit. The GAFF version used during parameterization on Orion was 1.8. In this Orion workflow, we also tested several additional potential tautomers for some molecules. For each molecule, we conducted a solvation free energy calculation of the solute in pure water and another in octanol. After parameterization, equilibration stages were run with OpenMM (version 7.2.2.dev-32bc79a) and free energy calculations were done with YANK (version 0.23.7 [185]). A cutoff for nonbonded interactions was set to 9 Å, electrostatic interactions were computed using PME, bonds involving hydrogen were constrained and HMR was used to allow for a 4 fs timestep.

The equilibration was carried out with OpenMM on Orion. The first step was 200 ps of NVT simulation with the solute heavy atoms harmonically restrained with 2.0 kcal/(mol·Å²) spring constants. The second step of equilibration was 200 ps of NPT simulation with harmonic heavy atom restraints with a 0.1 kcal/(mol·Å²) spring constant. These equilibrated

structures were then used in YANK [16] simulations. The length of the YANK simulations were 5 ns for each replica. The electrostatic and steric interactions of the solute with the solvent were scaled using the same λ parameters listed in the transfer free energy protocol previously. The OpenEye workflow was also different in that it employed the ELF10 AM1-BCC charging engine (<https://docs.eyesopen.com/toolkits/python/quacpactk/OEProtonClasses/OEAM1BCCELf10Charges.html>, https://docs.eyesopen.com/applications/quacpac/theory/molcharge_theory.html), and only *syn* conformers of neutral carboxylic acids were retained for charging because, in OpenEye’s view, *anti* conformers result in incorrect charges dominated by strong internal interactions which are not well suited for MM applications. The only carboxylic acid studied was SM08, but the modification of the charging procedure in this case (relative to that employed in our direct solvation free energy approach) appears to have significantly impacted employed partial charges, likely for the better, as performance on SM08 was markedly different with this protocol.

Empirical reference calculations

For all empirical calculations, the compounds were stripped of counter ions and neutralized. The pyridone tautomer of SM08 was used, as given, and as it is assumed to be the most stable tautomer.

The MOE/logP(o/w) model, the MOE/h_logP model, and the MOE/S_logP model are all available within the graphical modeling program MOE (MOE, available from the Chemical Computing Group, Montreal, www.chemcomp.com). The MOE/logP(o/w) model is based on 95 atom types, plus a few corrections for geminal halogens, 1-4 aromatic nitrogens, ethylene-glycol ethers, alkane carbons, and amino acids. The individual contributions were obtained from fitting to a data set of 1827 measurements, yielding an R^2 of 0.931 and an RMSE of 0.393 (P. Labute, logP(o/w) model, unpublished).

The MOE/h_logP model uses 8 2D-descriptors derived from Extended Huckel Theory (the descriptors used are the sum of atomic EHT donor and acceptor strengths, the sum over $\log(1 + \pi$ -bond order), the sum over $\log(1 + d$ -orbital bond order), the Gerber ring number and Gerber atomic surface area [70], and the number of hydrogens and number of hydrophobic carbons (carbons with no heteroatom within 3 bonds). The contributions of these descriptors were obtained by fitting to 1836 molecules yielding a model with an R^2 of 0.084 and an RMSE of 0.59 (P. Labute, MOE h_mr, h_logP, and h_logS models, unpublished). The MOE program is available from the Chemical Computing Group, Montreal (www.chemcomp.com). The MOE/S_logP model is described in this reference [225]. In this model, 68 different atom types were defined based on element and nearest neighbors, e.g. 27 different carbon types or 14 different nitrogen types. Then the atomic contributions were determined by fitting to a training set of almost 10000 molecules.

The MoKa/logP methodology [MoKa-3.2.2, Molecular Discovery Ltd, London, www.moldiscovery.com] builds on a similar approach as the corresponding pK_a prediction [145]. The procedure starts by calculating molecular interaction fields based on the GRID force field on a large number of molecular fragments. The 3D energy fields of these fragments are then stored and used to recompute any molecule as a summation of appropriate 3D fragments. Therefore any molecule can be quickly approximated by 3D fields describing polar and hydrophobic interaction with water and n-octanol. From these fields, VolSurf descriptors are computed and used in a training scheme using a database of about 20000 compounds. From the training model, a final model is computed to make external predictions (G. Cruciani, personal communication).

References for the Supplementary Information

- 111 Case D, Berryman J, Betz R, Cerutti D, Cheatham Iii T, Darden T, Duke R, Giese T, Gohlke H, Goetz A et al (2015) AMBER 2015. University of California, San Francisco.

- 112 Martínez L, Andrade R, Birgin EG, Martínez JM (2009) PACKMOL: a package for building initial configurations for molecular dynamics simulations. *J Comput Chem* 30(13):2157–2164. <https://doi.org/10.1002/jcc.21224>
- 113 Eastman P, Swails J, Chodera JD, McGibbon RT, Zhao Y, Beauchamp KA, Wang LP, Simmonett AC, Harrigan MP, Stern CD, Wiewiora RP, Brooks BR, Pande VS (2017) OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput Biol* 13(7):e1005659. <https://doi.org/10.1371/journal.pcbi.1005659>
- 114 Hopkins CW, Le Grand S, Walker RC, Roitberg AE (2015) Longtime- step molecular dynamics through hydrogen mass repartitioning. *J Chem Theory Comput* 11(4):1864–1874. <https://doi.org/10.1021/ct5010406>
- 115 Rizzi A, Chodera J, Naden L, Beauchamp K, Albanese S, Grinaway P, Rustenburg B, Saladi S, Boehm K (2018) choderalab/ yank: Bugix release. Zenodo. <https://doi.org/10.5281/zenodo.1447109>.
- 116 Gerber PR (1998) Charge distribution from a simple molecular orbital type calculation and non-bonding interaction terms in the force field MAB. *J Comput Aided Mol Des* 12(1):37–51. <https://doi.org/10.1023/A:1007902804814>
- 117 Milletti F, Storchi L, Sforna G, Cruciani G (2007) New and original pKa prediction method using grid molecular interaction fields. *J Chem Inf Model* 47(6):2172–2181. <https://doi.org/10.1021/ci700018y>

2.9.4 Supplementary Figures and tables

Table 2.7: **Method details of log P predictions with MM-based physical methods.** Force fields, water models, and octanol phase choice are reported. A dry octanol phase indicates the octanol phase was treated as consisting of pure octanol. A wet octanol phase indicates the octanol phase was treated as a mixture of octanol and water. RMSE and Kendall’s Tau values are reported as mean and 95% confidence intervals. A CSV version of this table can be found in *SAMPL6-supplementary-documents.tar.gz*.

Submission ID	Method Name	Force Field	Water Model	Octanol Phase	RMSE	Kendall's Tau (τ)
<i>nh6c0</i>	Molecular-Dynamics-Expanded-Ensembles	AMBER/OPLS like force field with manually adjusted parameters	modified Toukan-Rahman	wet	0.74 [0.56, 0.93]	0.49 [0.02, 0.87]
<i>ujsgv</i>	Alchemical-CGenFF	CGenFF	TIP3P	wet	0.82 [0.56, 1.06]	0.35 [-0.14, 0.79]
<i>2mi5w</i>	Alchemical-CGenFF	CGenFF	TIP3P	wet	0.95 [0.64, 1.24]	0.24 [-0.22, 0.71]
<i>y0xxd</i>	FS-GM (Fast switching Growth Method)	CGenFF	OPC3	dry	1.04 [0.42, 1.50]	0.42 [-0.14, 0.91]
<i>2ggir</i>	FS-AGM (Fast switching Annihilation/Growth Method)	CGenFF	OPC3	dry	1.04 [0.84, 1.24]	0.49 [-0.02, 0.92]
<i>3wyyh</i>	Alchemical-CGenFF	CGenFF	TIP3P	dry	1.13 [0.48, 1.75]	0.55 [0.11, 0.95]
<i>25s67</i>	FS-AGM (Fast switching Annihilation/Growth Method)	OPLS-AA	OPC3	dry	1.21 [0.84, 1.54]	0.45 [-0.14, 0.88]
<i>v2q0t</i>	InterX_GAFF_WET_OCTANOL	GAFF	TIP3P	wet	1.31 [0.94, 1.65]	0.64 [0.14, 1.00]
<i>ggm6n</i>	FS-GM (Fast switching Growth Method)	OPLS-AA	OPC3	dry	1.32 [0.95, 1.64]	0.53 [0.08, 0.87]
<i>jjd0b</i>	MD/S-MBIS-GAFF-TIP3P/MBAR/	GAFF (parameters refined w.r.t to QM)	TIP3P	dry	1.35 [0.89, 1.74]	0.53 [0.02, 0.91]
<i>sqosi</i>	MD-AMBER-dryoct	GAFF	TIP3P	dry	1.69 [1.14, 2.18]	0.45 [-0.06, 0.84]
<i>ke5gu</i>	MD/S-MBIS-GAFF-SPCE/MBAR/	GAFF (parameters refined w.r.t to QM)	SPCE	dry	1.82 [1.31, 2.25]	0.53 [-0.02, 0.91]
<i>mwwua</i>	MD-LigParGen-wetoct	OPLS-AA	TIP4P	wet	1.83 [1.48, 2.12]	0.48 [0.02, 0.84]
<i>fyx45</i>	LogP-prediction-Drude-FEP-HuangLab	Drude	unknown	unknown	1.85 [0.63, 2.70]	0.67 [0.14, 1.00]
<i>6nmtt</i>	MD-AMBER-wetoct	GAFF	TIP3P	wet	1.87 [1.33, 2.45]	0.60 [0.06, 1.00]
<i>eufcy</i>	MD-LigParGen-dryoct	OPLS-AA	TIP4P	dry	1.99 [1.62, 2.33]	0.66 [0.21, 0.96]
<i>tzzb5</i>	Alchemical-CGenFF	CGenFF (parameters refined w.r.t to QM)	TIP3P	wet	2.12 [1.55, 2.57]	-0.20 [-0.63, 0.29]
<i>3oqhx</i>	MD-CHARMM-dryoct	CGenFF	TIP3P	dry	2.14 [1.24, 2.86]	0.00 [-0.5, 0.51]
<i>bzeez</i>	FS-AGM (Fast switching Annihilation/Growth Method)	GAFF2	OPC3	dry	2.20 [1.83, 2.51]	0.53 [0.00, 0.91]
<i>5svjv</i>	FS-GM (Fast switching Growth Method)	GAFF2	OPC3	dry	2.26 [1.84, 2.66]	0.44 [-0.15, 0.92]
<i>odex0</i>	InterX_ARROW_2017_PIMD_SOLVENT2_WET_OCTANOL	ARROW FF	PIMD	wet	2.29 [1.63, 2.82]	-0.09 [-0.61, 0.50]
<i>padym</i>	InterX_ARROW_2017_PIMD_WET_OCTANOL	ARROW FF	PIMD	wet	2.29 [1.63, 2.81]	-0.13 [-0.69, 0.48]
<i>pnc4j</i>	LogP-prediction-Drude-Umbrella-HuangLab	Drude	unknown	unknown	2.29 [1.68, 2.88]	0.20 [-0.37, 0.70]
<i>REF02</i>	YANK-GAFF-tip3p-wet-oct	GAFF	TIP3P	wet	2.29 [1.07, 3.53]	0.53 [0.06, 0.92]
<i>REF05</i>	YANK-SMIRNOFF-tip3p-wet-oct	SMIRNOFF	TIP3P	wet	2.31 [1.20, 3.47]	0.45 [-0.04, 0.85]
<i>REF08</i>	YANK-SMIRNOFF-tip3p-dry-oct	SMIRNOFF	TIP3P	dry	2.34 [1.04, 3.65]	0.42 [-0.04, 0.75]
<i>REF07</i>	YANK-GAFF-tip3p-dry-oct	GAFF	TIP3P	dry	2.38 [1.03, 3.73]	0.53 [0.09, 0.88]
<i>fcspk</i>	ARROW_2017_PIMD_SOLVENT2	ARROW FF	ARROW FF	dry	2.40 [1.72, 2.95]	-0.16 [-0.65, 0.40]
<i>6cm6a</i>	ARROW_2017_PIMD	ARROW FF	ARROW FF	dry	2.41 [1.75, 2.93]	-0.27 [-0.72, 0.29]
<i>623c0</i>	MD-OPLSAA-wetoct	OPLS-AA	TIP4P	wet	2.67 [2.13, 3.20]	0.38 [-0.14, 0.84]
<i>4nfzz</i>	MD/S-HI-GAFF-TIP3P/MBAR/	GAFF (parameters refined w.r.t to QM)	TIP3P	dry	2.67 [1.98, 3.35]	0.42 [-0.13, 0.88]
<i>eg52i</i>	ARROW_2017	ARROW FF	ARROW FF	dry	2.86 [2.01, 3.56]	-0.16 [-0.59, 0.35]
<i>cp8kv</i>	MD-OPLSAA-dryoct	OPLS-AA	TIP4P	dry	2.88 [2.31, 3.60]	0.59 [0.11, 1.00]
<i>5585v</i>	Alchemical-CGenFF	CGenFF (parameters refined w.r.t to QM)	TIP3P	wet	2.88 [2.02, 3.67]	-0.2 [-0.76, 0.32]
<i>REF04</i>	YANK-SMIRNOFF-TIP3P-FB-wet-oct	SMIRNOFF	TIP3P-FB	wet	3.22 [2.04, 4.48]	0.42 [-0.08, 0.84]
<i>hf4wj</i>	MD/S-HI-GAFF-SPCE/MBAR/	GAFF (parameters refined w.r.t to QM)	SPCE	dry	3.28 [2.49, 4.11]	0.38 [-0.16, 0.84]
<i>REF01</i>	YANK-GAFF-TIP3P-FB-wet-oct	GAFF	TIP3P-FB	wet	3.33 [2.08, 4.72]	0.49 [0.08, 0.83]
<i>REF06</i>	YANK-SMIRNOFF-OPC-wet-oct	SMIRNOFF	OPC	wet	3.64 [2.37, 4.97]	0.31 [-0.14, 0.72]
<i>REF03</i>	YANK-GAFF-opc-wet-oct	GAFF	OPC	wet	4.01 [2.74, 5.34]	0.42 [-0.06, 0.79]

Table 2.8: **SMILES and InChI identifiers of SAMPL6 log P Challenge molecules.** Experimental log P values can be found in a separate paper reporting measurements [88]. A CSV version of this table can be found in *SAMPL6-supplementary-documents.tar.gz*.

SAMPL6 Molecule ID	isomeric SMILES	InChI
SM02	<chem>c1ccc2c(c1)c(ncn2)Nc3cccc(c3)C(F)(F)F</chem>	InChI=1S/C15H10F3N3/c16-15(17,18)10-4-3-5-11(8-10)21-14-12-6-1-2-7-13(12)19-9-20-14/h1-9H,(H,19,20,21)
SM04	<chem>c1ccc2c(c1)c(ncn2)Nc3ccc(cc3)Cl</chem>	InChI=1S/C15H12ClN3/c16-12-7-5-11(6-8-12)9-17-15-13-3-1-2-4-14(13)18-10-19-15/h1-8,10H,9H2,(H,17,18,19)
SM07	<chem>c1ccc(cc1)CNc2c3cccc3ncn2</chem>	InChI=1S/C15H13N3/c1-2-6-12(7-3-1)10-16-15-13-8-4-5-9-14(13)17-11-18-15/h1-9,11H,10H2,(H,16,17,18)
SM08	<chem>Cc1ccc2c(c1)c(c(=O)[nH]2)CC(=O)Oc3cccc3</chem>	InChI=1S/C18H15NO3/c1-11-7-8-15-13(9-11)17(12-5-3-2-4-6-12)14(10-16(20)21)18(22)19-15/h2-9H,10H2,1H3,(H,19,22)(H,20,21)
SM09	<chem>COc1cccc(c1)Nc2c3cccc3ncn2.Cl</chem>	InChI=1S/C15H13N3O.ClH/c1-19-12-6-4-5-11(9-12)18-15-13-7-2-3-8-14(13)16-10-17-15/h2-10H,1H3,(H,16,17,18);1H
SM11	<chem>c1ccc(cc1)n2c3c(cn2)c(ncn3)N</chem>	InChI=1S/C11H9N5/c12-10-9-6-15-16(11(9)14-7-13-10)8-4-2-1-3-5-8/h1-7H,(H2,12,13,14)
SM12	<chem>c1ccc2c(c1)c(ncn2)Nc3cccc(c3)Cl.Cl</chem>	InChI=1S/C14H10ClN3.ClH/c15-10-4-3-5-11(8-10)18-14-12-6-1-2-7-13(12)16-9-17-14/h1-9H,(H,16,17,18);1H
SM13	<chem>Cc1cccc(c1)Nc2c3cc(c(cc3ncn2)OC)OC</chem>	InChI=1S/C17H17N3O2/c1-11-5-4-6-12(7-11)20-17-13-8-15(21-2)16(22-3)9-14(13)18-10-19-17/h4-10H,1-3H3,(H,18,19,20)
SM14	<chem>c1ccc(cc1)n2cnc3c2ccc(c3)N</chem>	InChI=1S/C13H11N3/c14-10-6-7-13-12(8-10)15-9-16(13)11-4-2-1-3-5-11/h1-9H,14H2
SM15	<chem>c1ccc2c(c1)ncn2c3ccc(cc3)O</chem>	InChI=1S/C13H10N2O/c16-11-7-5-10(6-8-11)15-9-14-12-3-1-2-4-13(12)15/h1-9,16H
SM16	<chem>c1cc(c(c1)Cl)C(=O)Nc2ccncc2Cl</chem>	InChI=1S/C12H8Cl2N2O/c13-9-2-1-3-10(14)11(9)12(17)16-8-4-6-15-7-5-8/h1-7H,(H,15,16,17)

Table 2.9: **SMILES and InChI identifiers of extra molecules included in the evaluation of reference methods.** A CSV version of this table can be found in *SAMPL6-supplementary-documents.tar.gz*. Experimental log *P* values can be found in a separate paper reporting measurements [202] and in machine readable format in https://github.com/samplchallenges/SAMPL6/blob/master/physical_properties/logP/analysis_of_extra_molecules/logP_experimental_values.csv.

Extra Molecule ID	Isomeric SMILES	InChI
3,4-Dichlorophenol	<chem>c1cc(cc1O)Cl</chem>	InChI=1S/C6H4Cl2O/c7-5-2-1-4(9)3-6(5)8/h1-3,9H
3,5-Dichlorophenol	<chem>c1cc(cc1Cl)Cl</chem>	InChI=1S/C6H4Cl2O/c7-4-1-5(8)3-6(9)2-4/h1-3,9H
3-Bromoquinoline	<chem>c1ccc2c(c1)cc(cn2)Br</chem>	InChI=1S/C9H6BrN/c10-8-5-7-3-1-2-4-9(7)11-6-8/h1-6H
3-Chlorophenol	<chem>c1cc(cc(c1)Cl)O</chem>	InChI=1S/C6H5ClO/c7-5-2-1-3-6(8)4-5/h1-4,8H
4-Butoxyphenol	<chem>CCCCOc1ccc(cc1)O</chem>	InChI=1S/C10H14O2/c1-2-3-8-12-10-6-4-9(11)5-7-10/h4-7,11H,2-3,8H2,1H3
4-Chlorophenol	<chem>c1cc(ccc1O)Cl</chem>	InChI=1S/C6H5ClO/c7-5-1-3-6(8)4-2-5/h1-4,8H
4-Ethoxyphenol	<chem>CCOc1ccc(cc1)O</chem>	InChI=1S/C8H10O2/c1-2-10-8-5-3-7(9)4-6-8/h3-6,9H,2H2,1H3
4-Iodophenol	<chem>c1cc(ccc1O)I</chem>	InChI=1S/C6H5IO/c7-5-1-3-6(8)4-2-5/h1-4,8H
4-Methoxyphenol	<chem>COc1ccc(cc1)O</chem>	InChI=1S/C7H8O2/c1-9-7-4-2-6(8)3-5-7/h2-5,8H,1H3
4-Pentoxyphenol	<chem>CCCCOc1ccc(cc1)O</chem>	InChI=1S/C11H16O2/c1-2-3-4-9-13-11-7-5-10(12)6-8-11/h5-8,12H,2-4,9H2,1H3
4-Propoxyphenol	<chem>CCCOc1ccc(cc1)O</chem>	InChI=1S/C9H12O2/c1-2-7-11-9-5-3-8(10)4-6-9/h3-6,10H,2,7H2,1H3
Acebutolol	<chem>CCCC(=O)Nc1ccc(c(c1)C(=O)C)OCC(CNC(C)C)O</chem>	InChI=1S/C18H28N2O4/c1-5-6-18(23)20-14-7-8-17(16(9-14)13(4)21)24-11-15(22)10-19-12(2)3/h7-9,12,15,19,22H,5-6,10-1
Amylobarbitone	<chem>CCC1(C(=O)NC(=O)NC1=O)CCC(C)C</chem>	InChI=1S/C11H18N2O3/c1-4-11(6-5-7(2)3)8(14)12-10(16)13-9(11)15/h7H,4-6H2,1-3H3,(H2,12,13,14,15,16)
Benzoicacid	<chem>c1ccc(cc1)C(=O)O</chem>	InChI=1S/C7H6O2/c8-7(9)6-4-2-1-3-5-6/h1-5H,(H,8,9)
Butobarbitone	<chem>CCCC1(C(=O)NC(=O)NC1=O)CC</chem>	InChI=1S/C10H16N2O3/c1-3-5-6-10(4-2)7(13)11-9(15)12-8(10)14/h3-6H2,1-2H3,(H2,11,12,13,14,15)
Celiprolol	<chem>CN(C)C(C(=O)Nc1ccc(c(c1)C(=O)C)OCC(CNC(C)C)C)O</chem>	InChI=1S/C20H33N3O4/c1-7-23(8-2)19(26)22-15-9-10-18(17(11-15)14(3)24)27-13-16(25)12-21-20(4,5)6/h9-11,16,21,25H,7-8,12-13H2,1-6H3,(H,22,26)
Chlorpromazine	<chem>CN(C)CCCN1c2cccc2Sc3c1cc(cc3)Cl</chem>	InChI=1S/C17H19ClN2S/c1-19(2)10-5-11-20-14-6-3-4-7-16(14)21-17-9-8-13(18)12-15(17)20/h3-4,6-9,12H,5,10-11H2,1-2H3
Ketoprofen	<chem>CC(c1cccc(c1)C(=O)c2ccccc2)C(=O)O</chem>	InChI=1S/C16H14O3/c1-11(16(18)19)13-8-5-9-14(10-13)15(17)12-6-3-2-4-7-12/h2-11H,1H3,(H,18,19)
<i>N</i> -Methylaniline	<chem>CNc1ccccc1</chem>	InChI=1S/C7H9N/c1-8-7-5-3-2-4-6-7/h2-6,8H,1H3
Pentobarbitone	<chem>CCCC(C)C1(C(=O)NC(=O)NC1=O)CC</chem>	InChI=1S/C11H18N2O3/c1-4-6-7(3)11(5-2)8(14)12-10(16)13-9(11)15/h7H,4-6H2,1-3H3,(H2,12,13,14,15,16)
Pericyazine	<chem>c1ccc2c(c1)N(c3cc(ccc3S2)C#N)CCCN4CCC(CC4)O</chem>	InChI=1S/C21H23N3OS/c22-15-16-6-7-21-19(14-16)24(18-4-1-2-5-20(18)26-21)11-3-10-23-12-8-17(25)9-13-23/h1-2,4-7,14,17,25H,3,8-13H2
Phenobarbitone	<chem>CCC1(C(=O)NC(=O)NC1=O)c2ccccc2</chem>	InChI=1S/C12H12N2O3/c1-2-12(8-6-4-3-5-7-8)9(15)13-11(17)14-10(12)16/h3-7H,2H2,1H3,(H2,13,14,15,16,17)
Phenol	<chem>c1ccc(cc1)O</chem>	InChI=1S/C6H6O/c7-6-4-2-1-3-5-6/h1-5,7H
Quinalbarbitone	<chem>CCCC(C)C1(C(=O)NC(=O)NC1=O)CC=C</chem>	InChI=1S/C12H18N2O3/c1-4-6-8(3)12(7-5-2)9(15)13-11(17)14-10(12)16/h5,8H,2,4,6-7H2,1,3H3,(H2,13,14,15,16,17)
Quinoline	<chem>c1ccc2c(c1)cccn2</chem>	InChI=1S/C9H7N/c1-2-6-9-8(4-1)5-3-7-10-9/h1-7H
(±)-Propranolol	<chem>CC(C)NCC(CO)c1ccc2c1ccccc2O</chem>	InChI=1S/C16H21NO2/c1-12(2)17-10-14(18)11-19-16-9-5-7-13-6-3-4-8-15(13)16/h3-9,12,14,17-18H,10-11H2,1-2H3
Sulfamethazine	<chem>Cc1cc(nc(n1)NS(=O)(=O)c2ccc(cc2)N)C</chem>	InChI=1S/C12H14N4O2S/c1-8-7-9(2)15-12(14-8)16-19(17,18)11-5-3-10(13)4-6-11/h3-7H,13H2,1-2H3,(H,14,15,16)

Table 2.10: **Evaluation statistics calculated for all methods.** Methods are represented via their SAMPL6 submission IDs which can be cross referenced with Table 2.3 for method details. There are six error metrics reported: the root-mean-squared error (RMSE), mean absolute error (MAE), mean (signed) error (ME), coefficient of determination (R^2), linear regression slope (m), and Kendall’s Rank Correlation Coefficient (τ). This table is ranked by increasing RMSE. A CSV version of this table can be found in *SAMPL6-supplementary-documents.tar.gz*.

Submission ID	RMSE	MAE	ME	R^2	m	Kendall’s Tau
<i>hmz0n</i>	0.38 [0.23,0.55]	0.31 [0.19,0.46]	-0.17 [-0.38,0.03]	0.77 [0.34,0.94]	0.94 [0.59,1.15]	0.64 [0.16,0.96]
<i>gmoq5</i>	0.39 [0.28,0.49]	0.34 [0.23,0.46]	0.01 [-0.21,0.25]	0.74 [0.39,0.92]	0.99 [0.66,1.33]	0.59 [0.10,0.88]
<i>3uqbi</i>	0.41 [0.28,0.53]	0.36 [0.24,0.48]	-0.08 [-0.30,0.17]	0.66 [0.27,0.93]	0.78 [0.50,1.10]	0.56 [0.12,0.91]
<i>sq07q</i>	0.47 [0.33,0.58]	0.41 [0.28,0.54]	0.03 [-0.24,0.31]	0.64 [0.21,0.89]	0.92 [0.51,1.30]	0.56 [0.10,0.88]
<i>j8nwc</i>	0.47 [0.17,0.75]	0.31 [0.15,0.55]	0.07 [-0.16,0.38]	0.74 [0.33,0.97]	1.14 [0.84,1.38]	0.81 [0.44,1.00]
<i>xzh4i</i>	0.49 [0.34,0.62]	0.43 [0.29,0.57]	0.18 [-0.09,0.43]	0.54 [0.14,0.86]	0.60 [0.29,1.03]	0.51 [0.00,0.88]
<i>hdpuj</i>	0.49 [0.37,0.61]	0.44 [0.32,0.57]	-0.29 [-0.51,-0.05]	0.74 [0.40,0.94]	1.02 [0.69,1.35]	0.67 [0.23,1.00]
<i>dqzk4</i>	0.49 [0.33,0.62]	0.42 [0.26,0.57]	0.30 [0.06,0.53]	0.69 [0.37,0.91]	0.83 [0.50,1.26]	0.67 [0.25,0.96]
<i>vzgyt</i>	0.50 [0.27,0.68]	0.38 [0.21,0.58]	-0.35 [-0.57,-0.15]	0.72 [0.28,0.95]	0.76 [0.48,0.98]	0.64 [0.25,0.92]
<i>ypmr0</i>	0.50 [0.36,0.63]	0.44 [0.31,0.58]	0.07 [-0.23,0.35]	0.61 [0.25,0.89]	0.93 [0.54,1.52]	0.64 [0.23,0.92]
<i>yd6ub</i>	0.51 [0.32,0.66]	0.41 [0.23,0.59]	0.09 [-0.21,0.38]	0.63 [0.21,0.89]	0.99 [0.47,1.41]	0.53 [-0.02,0.87]
<i>7egyc</i>	0.52 [0.35,0.66]	0.44 [0.28,0.60]	0.27 [0.01,0.52]	0.57 [0.22,0.85]	0.50 [0.32,0.77]	0.45 [0.06,0.83]
<i>0a7a8</i>	0.53 [0.34,0.69]	0.43 [0.25,0.62]	0.32 [0.07,0.56]	0.62 [0.13,0.90]	0.74 [0.34,1.02]	0.45 [-0.14,0.84]
<i>7dhtp</i>	0.54 [0.33,0.70]	0.44 [0.26,0.62]	0.06 [-0.27,0.36]	0.49 [0.06,0.88]	0.73 [0.26,1.15]	0.56 [0.04,0.96]
<i>qyzjr</i>	0.54 [0.34,0.75]	0.46 [0.31,0.65]	-0.15 [-0.41,0.19]	0.73 [0.33,0.97]	1.22 [0.89,1.50]	0.78 [0.45,1.00]
<i>REF11</i>	0.54 [0.25,0.80]	0.39 [0.19,0.64]	0.19 [-0.09,0.50]	0.59 [0.37,0.89]	0.90 [0.37,1.62]	0.67 [0.33,0.96]
<i>REF13</i>	0.55 [0.37,0.71]	0.47 [0.31,0.64]	-0.27 [-0.55,0.02]	0.69 [0.31,0.93]	1.06 [0.55,1.55]	0.60 [0.08,0.96]
<i>w6jta</i>	0.56 [0.33,0.76]	0.46 [0.28,0.66]	0.32 [0.06,0.61]	0.53 [0.12,0.89]	0.62 [0.34,0.86]	0.51 [0.02,0.88]
<i>REF12</i>	0.60 [0.42,0.76]	0.52 [0.36,0.70]	-0.08 [-0.43,0.26]	0.67 [0.23,0.90]	1.21 [0.76,1.53]	0.55 [0.06,0.88]
<i>ji2zm</i>	0.60 [0.43,0.75]	0.53 [0.38,0.70]	0.45 [0.22,0.67]	0.66 [0.32,0.90]	0.66 [0.43,0.96]	0.51 [0.11,0.84]
<i>5krdi</i>	0.60 [0.39,0.81]	0.51 [0.33,0.71]	-0.30 [-0.60,0.01]	0.63 [0.24,0.91]	1.03 [0.59,1.51]	0.60 [0.14,0.92]
<i>REF10</i>	0.60 [0.39,0.83]	0.51 [0.33,0.72]	-0.04 [-0.42,0.30]	0.38 [0.01,0.82]	0.65 [-0.03,1.21]	0.35 [-0.27,0.8]
<i>gnxuu</i>	0.61 [0.39,0.80]	0.51 [0.31,0.72]	0.40 [0.13,0.68]	0.53 [0.12,0.91]	0.57 [0.34,0.79]	0.51 [0.04,0.88]
<i>tc4xa</i>	0.62 [0.41,0.80]	0.51 [0.31,0.73]	0.17 [-0.18,0.53]	0.66 [0.17,0.90]	1.21 [0.52,1.65]	0.49 [-0.02,0.84]
<i>6cdyo</i>	0.65 [0.42,0.83]	0.54 [0.33,0.76]	-0.24 [-0.60,0.10]	0.52 [0.20,0.81]	0.93 [0.48,1.70]	0.53 [0.17,0.87]
<i>dbmq3</i>	0.70 [0.47,0.89]	0.60 [0.39,0.82]	0.42 [0.09,0.74]	0.47 [0.03,0.80]	0.75 [0.12,1.29]	0.38 [-0.18,0.80]
<i>kasp3</i>	0.74 [0.49,0.94]	0.62 [0.39,0.86]	0.48 [0.14,0.80]	0.36 [0.02,0.77]	0.54 [0.04,1.15]	0.35 [-0.20,0.80]
<i>nh6c0</i>	0.74 [0.56,0.93]	0.67 [0.48,0.87]	0.09 [-0.35,0.53]	0.62 [0.16,0.88]	1.34 [0.52,1.91]	0.49 [0.02,0.87]
<i>kujfu</i>	0.78 [0.35,1.08]	0.56 [0.27,0.90]	-0.03 [-0.51,0.40]	0.41 [0.03,0.88]	0.97 [0.30,1.43]	0.45 [-0.02,0.84]
<i>NULL0</i>	0.79 [0.50,1.03]	0.66 [0.42,0.92]	0.42 [0.02,0.81]	0.00 [0.00,0.00]	0.00 [0.00,0.00]	0.00 [0.00,0.00]
<i>ujsgv</i>	0.82 [0.56,1.06]	0.67 [0.39,0.95]	-0.31 [-0.76,0.15]	0.33 [0.01,0.81]	0.80 [-0.02,1.45]	0.35 [-0.14,0.79]
<i>REF09</i>	0.82 [0.52,1.10]	0.68 [0.43,0.97]	-0.26 [-0.73,0.20]	0.46 [0.08,0.89]	1.09 [0.44,1.78]	0.48 [-0.02,0.86]
<i>wu52s</i>	0.83 [0.57,1.05]	0.72 [0.49,0.97]	0.70 [0.43,0.97]	0.55 [0.11,0.99]	0.54 [0.24,0.88]	0.56 [-0.06,1.00]
<i>g6duz</i>	0.85 [0.56,1.08]	0.72 [0.45,0.99]	0.35 [-0.11,0.80]	0.52 [0.07,0.85]	1.18 [0.47,1.70]	0.45 [-0.07,0.84]
<i>5mahv</i>	0.85 [0.42,1.19]	0.62 [0.31,0.99]	-0.02 [-0.53,0.47]	0.34 [0.03,0.79]	0.90 [0.28,1.37]	0.24 [-0.33,0.72]
<i>bgeuh</i>	0.87 [0.51,1.17]	0.66 [0.34,1.01]	0.25 [-0.24,0.73]	0.01 [0.00,0.53]	-0.05 [-0.42,0.49]	0.02 [-0.57,0.57]
<i>d7vth</i>	0.87 [0.62,1.10]	0.78 [0.56,1.01]	-0.65 [-0.96,-0.30]	0.63 [0.21,0.93]	1.11 [0.73,1.39]	0.49 [0.02,0.85]
<i>2mi5w</i>	0.95 [0.64,1.24]	0.81 [0.54,1.12]	-0.3 [-0.83,0.23]	0.18 [0.00,0.64]	0.61 [-0.12,1.25]	0.24 [-0.22,0.71]
<i>kuddg</i>	0.97 [0.73,1.18]	0.89 [0.67,1.12]	0.89 [0.67,1.12]	0.67 [0.26,0.95]	0.71 [0.44,1.04]	0.53 [-0.02,0.96]
<i>qz8d5</i>	0.97 [0.71,1.19]	0.84 [0.56,1.12]	0.77 [0.42,1.10]	0.53 [0.18,0.84]	0.93 [0.49,1.58]	0.48 [0.06,0.82]
<i>y0xxd</i>	1.04 [0.42,1.50]	0.72 [0.32,1.21]	0.37 [-0.18,1.00]	0.33 [0.00,0.93]	1.03 [-0.20,2.00]	0.42 [-0.14,0.91]
<i>2ggir</i>	1.04 [0.84,1.24]	0.98 [0.76,1.19]	-0.36 [-0.88,0.27]	0.31 [0.00,0.93]	0.98 [-0.33,1.88]	0.49 [-0.02,0.92]
<i>dyxbt</i>	1.07 [0.79,1.34]	0.96 [0.70,1.23]	0.96 [0.70,1.23]	0.55 [0.11,0.9]	0.68 [0.22,1.15]	0.56 [0.12,0.92]
<i>mm0jf</i>	1.09 [0.91,1.24]	1.03 [0.81,1.22]	1.03 [0.81,1.22]	0.75 [0.44,0.98]	0.60 [0.39,0.82]	0.75 [0.38,1.00]
<i>h83sb</i>	1.12 [0.59,1.59]	0.87 [0.50,1.33]	-0.21 [-0.91,0.40]	0.00 [0.00,0.57]	-0.02 [-1.06,0.84]	-0.16 [-0.69,0.42]
<i>3wvyh</i>	1.13 [0.48,1.75]	0.77 [0.35,1.33]	0.26 [-0.32,0.99]	0.37 [0.03,0.93]	1.24 [0.32,2.29]	0.55 [0.11,0.95]
<i>f3dpg</i>	1.17 [0.74,1.52]	0.92 [0.50,1.36]	-0.85 [-1.33,-0.38]	0.11 [0.00,0.47]	0.36 [-0.18,0.85]	0.15 [-0.33,0.51]
<i>25s67</i>	1.21 [0.84,1.54]	1.06 [0.72,1.42]	-0.97 [-1.39,-0.55]	0.63 [0.16,0.90]	1.33 [0.43,2.34]	0.45 [-0.14,0.88]
<i>zdj0j</i>	1.21 [0.98,1.41]	1.13 [0.86,1.37]	1.13 [0.86,1.37]	0.64 [0.26,0.94]	0.86 [0.41,1.31]	0.64 [0.18,0.96]
<i>7gg6s</i>	1.27 [0.81,1.62]	1.00 [0.55,1.47]	-1.00 [-1.47,-0.55]	0.10 [0.00,0.46]	0.31 [-0.17,0.77]	0.16 [-0.33,0.55]
<i>hwf2k</i>	1.28 [0.57,1.90]	0.93 [0.49,1.50]	-0.09 [-0.92,0.57]	0.12 [0.00,0.84]	0.68 [-0.77,1.60]	0.31 [-0.32,0.79]
<i>pcv32</i>	1.28 [1.00,1.53]	1.17 [0.84,1.47]	1.17 [0.84,1.47]	0.50 [0.14,0.89]	0.75 [0.26,1.38]	0.44 [-0.04,0.81]
<i>v2q0t</i>	1.31 [0.94,1.65]	1.16 [0.82,1.52]	-1.15 [-1.52,-0.79]	0.70 [0.25,0.98]	1.31 [0.92,1.57]	0.64 [0.14,1.00]

Table 2.11: [Table 2.10 continued.] Evaluation statistics calculated for all methods. Methods are represented via their SAMPL6 submission IDs which can be cross referenced with Table 2.3 for method details. There are six error metrics reported: the root-mean-squared error (RMSE), mean absolute error (MAE), mean (signed) error (ME), coefficient of determination (R^2), linear regression slope (m), and Kendall’s Rank Correlation Coefficient (τ). This table is ranked by increasing RMSE. A CSV version of this table can be found in *SAMPL6-supplementary-documents.tar.gz*.

Submission ID	RMSE	MAE	ME	R^2	m	Kendall’s Tau
<i>rdsnw</i>	1.32 [0.88,1.71]	1.15 [0.80,1.54]	1.15 [0.80,1.54]	0.78 [0.39,0.97]	1.51 [1.16,1.77]	0.75 [0.37,1.00]
<i>ggm6n</i>	1.32 [0.95,1.64]	1.16 [0.79,1.54]	-1.15 [-1.53,-0.76]	0.53 [0.12,0.84]	1.04 [0.46,1.67]	0.53 [0.08,0.87]
<i>jjd0b</i>	1.35 [0.89,1.74]	1.13 [0.71,1.57]	-1.09 [-1.56,-0.63]	0.66 [0.24,0.91]	1.51 [0.81,2.04]	0.53 [0.02,0.91]
<i>2tzb0</i>	1.38 [0.94,1.79]	1.21 [0.85,1.62]	1.21 [0.85,1.62]	0.79 [0.42,0.97]	1.58 [1.21,1.86]	0.75 [0.36,1.00]
<i>cr3hs</i>	1.39 [0.58,2.10]	0.96 [0.46,1.61]	0.80 [0.21,1.52]	0.40 [0.01,0.79]	1.36 [-0.19,2.63]	0.35 [-0.33,0.84]
<i>arw58</i>	1.41 [0.81,1.89]	1.09 [0.60,1.63]	1.01 [0.45,1.61]	0.09 [0.00,0.54]	-0.24 [-0.75,0.26]	-0.20 [-0.64,0.36]
<i>ahmtf</i>	1.41 [1.13,1.69]	1.33 [1.07,1.62]	1.33 [1.07,1.62]	0.55 [0.11,0.89]	0.70 [0.23,1.16]	0.56 [0.11,0.92]
<i>o7djk</i>	1.42 [1.14,1.70]	1.34 [1.07,1.62]	1.34 [1.07,1.62]	0.55 [0.12,0.89]	0.70 [0.24,1.16]	0.56 [0.10,0.92]
<i>fnf7r</i>	1.44 [1.03,1.76]	1.25 [0.83,1.66]	1.44 [1.03,1.76]	0.05 [0.00,0.57]	0.47 [-0.89,2.08]	0.10 [-0.5,0.64]
<i>4p2ph</i>	1.44 [0.81,1.94]	1.12 [0.61,1.68]	1.04 [0.47,1.64]	0.09 [0.00,0.55]	-0.26 [-0.77,0.25]	-0.26 [-0.68,0.29]
<i>6fyg5</i>	1.50 [1.27,1.70]	1.44 [1.18,1.67]	1.44 [1.18,1.67]	0.69 [0.32,0.96]	0.93 [0.50,1.51]	0.71 [0.28,1.00]
<i>sqosi</i>	1.69 [1.14,2.18]	1.42 [0.89,1.97]	-1.40 [-1.97,-0.86]	0.51 [0.05,0.87]	1.40 [0.39,2.02]	0.45 [-0.06,0.84]
<i>rs4ns</i>	1.71 [1.12,2.23]	1.44 [0.92,2.01]	1.44 [0.92,2.01]	0.06 [0.00,0.50]	-0.19 [-0.71,0.29]	-0.22 [-0.69,0.34]
<i>c7t5j</i>	1.73 [1.14,2.24]	1.47 [0.94,2.02]	1.47 [0.94,2.02]	0.05 [0.00,0.49]	-0.18 [-0.72,0.30]	-0.16 [-0.66,0.40]
<i>jc68f</i>	1.74 [1.13,2.25]	1.47 [0.94,2.03]	1.47 [0.94,2.03]	0.05 [0.00,0.48]	-0.18 [-0.71,0.30]	-0.16 [-0.65,0.40]
<i>03cyy</i>	1.75 [0.57,2.73]	1.11 [0.43,2.01]	0.03 [-0.89,1.16]	0.00 [0.00,0.53]	0.12 [-1.15,1.49]	0.09 [-0.56,0.71]
<i>hsotx</i>	1.81 [1.22,2.32]	1.56 [1.03,2.11]	1.56 [1.03,2.11]	0.07 [0.00,0.49]	-0.19 [-0.66,0.25]	-0.20 [-0.67,0.36]
<i>ke5gu</i>	1.82 [1.31,2.25]	1.59 [1.07,2.09]	-1.59 [-2.09,-1.07]	0.62 [0.17,0.89]	1.54 [0.74,2.16]	0.53 [-0.02,0.91]
<i>mwuuu</i>	1.83 [1.48,2.12]	1.73 [1.39,2.07]	-1.73 [-2.07,-1.39]	0.41 [0.01,0.77]	0.67 [0.07,1.13]	0.48 [0.02,0.84]
<i>fe8ws</i>	1.83 [1.24,2.34]	1.58 [1.06,2.13]	1.58 [1.06,2.13]	0.06 [0.00,0.48]	-0.18 [-0.67,0.26]	-0.16 [-0.64,0.41]
<i>5t0yn</i>	1.85 [1.26,2.37]	1.61 [1.09,2.15]	1.61 [1.09,2.15]	0.06 [0.00,0.49]	-0.18 [-0.67,0.27]	-0.16 [-0.65,0.41]
<i>fy445</i>	1.85 [0.63,2.70]	1.25 [0.51,2.14]	0.65 [-0.3,1.74]	0.63 [0.17,0.92]	2.63 [1.09,3.88]	0.67 [0.14,1.00]
<i>6nmtt</i>	1.87 [1.33,2.45]	1.65 [1.16,2.20]	-1.65 [-2.20,-1.16]	0.42 [0.02,0.92]	1.10 [0.23,1.56]	0.60 [0.06,1.00]
<i>eufcy</i>	1.99 [1.62,2.33]	1.88 [1.49,2.25]	-1.77 [-2.25,-1.17]	0.54 [0.18,0.88]	1.43 [0.49,2.41]	0.66 [0.21,0.96]
<i>tzzb5</i>	2.12 [1.55,2.57]	1.87 [1.26,2.44]	1.43 [0.50,2.31]	0.20 [0.00,0.63]	-0.76 [-1.61,0.17]	-0.20 [-0.63,0.29]
<i>3oqhx</i>	2.14 [1.24,2.86]	1.64 [0.86,2.49]	1.11 [0.06,2.22]	0.03 [0.00,0.41]	-0.44 [-1.90,1.03]	0.00 [-0.50,0.51]
<i>bzeez</i>	2.20 [1.83,2.51]	2.07 [1.57,2.46]	-2.07 [-2.46,-1.57]	0.63 [0.17,0.95]	1.39 [0.77,2.03]	0.53 [0.00,0.91]
<i>ynquk</i>	2.26 [1.87,2.59]	2.13 [1.67,2.54]	2.13 [1.67,2.54]	0.08 [0.00,0.76]	0.25 [-0.25,0.61]	0.38 [-0.06,0.80]
<i>5svju</i>	2.26 [1.84,2.66]	2.14 [1.69,2.58]	-2.03 [-2.57,-1.36]	0.39 [0.03,0.91]	1.20 [0.44,1.77]	0.44 [-0.15,0.92]
<i>odex0</i>	2.29 [1.63,2.82]	1.98 [1.31,2.65]	1.73 [0.82,2.57]	0.09 [0.00,0.64]	-0.53 [-1.76,0.68]	-0.09 [-0.61,0.50]
<i>padym</i>	2.29 [1.63,2.81]	1.99 [1.31,2.64]	1.72 [0.78,2.57]	0.12 [0.00,0.69]	-0.60 [-1.92,0.73]	-0.13 [-0.69,0.48]
<i>pnc4j</i>	2.29 [1.68,2.88]	2.03 [1.42,2.67]	2.03 [1.42,2.67]	0.04 [0.00,0.64]	0.31 [-0.81,1.30]	0.20 [-0.37,0.70]
<i>REF02</i>	2.29 [1.07,3.53]	1.68 [0.95,2.73]	-1.68 [-2.73,-0.95]	0.23 [0.00,0.91]	1.26 [0.02,2.29]	0.53 [0.06,0.92]
<i>REF05</i>	2.31 [1.20,3.47]	1.80 [1.15,2.76]	-1.80 [-2.76,-1.15]	0.20 [0.00,0.91]	1.07 [-0.08,2.18]	0.45 [-0.04,0.85]
<i>REF08</i>	2.34 [1.04,3.65]	1.66 [0.92,2.77]	-1.66 [-2.77,-0.92]	0.13 [0.00,0.81]	0.95 [-0.39,2.05]	0.42 [-0.04,0.75]
<i>REF07</i>	2.38 [1.03,3.73]	1.65 [0.84,2.80]	-1.65 [-2.80,-0.84]	0.24 [0.01,0.93]	1.43 [0.07,2.65]	0.53 [0.09,0.88]
<i>fcspk</i>	2.40 [1.72,2.95]	2.10 [1.41,2.79]	1.97 [1.12,2.76]	0.11 [0.00,0.65]	-0.50 [-1.60,0.61]	-0.16 [-0.65,0.40]
<i>6cm6a</i>	2.41 [1.75,2.93]	2.10 [1.40,2.78]	1.94 [1.04,2.74]	0.19 [0.00,0.69]	-0.66 [-1.77,0.32]	-0.27 [-0.72,0.29]
<i>bq6fo</i>	2.58 [1.68,3.34]	2.15 [1.35,3.01]	1.55 [0.30,2.74]	0.10 [0.00,0.56]	1.05 [-0.88,2.73]	0.09 [-0.39,0.60]
<i>623c0</i>	2.67 [2.13,3.20]	2.53 [2.08,3.04]	-2.53 [-3.04,-2.08]	0.22 [0.00,0.80]	0.64 [-0.05,1.09]	0.38 [-0.14,0.84]
<i>4nfzz</i>	2.67 [1.98,3.35]	2.44 [1.83,3.10]	-2.44 [-3.10,-1.83]	0.40 [0.05,0.87]	1.30 [0.56,1.85]	0.42 [-0.13,0.88]
<i>eg52i</i>	2.86 [2.01,3.56]	2.41 [1.52,3.32]	2.06 [0.88,3.21]	0.15 [0.00,0.55]	-0.94 [-2.15,0.19]	-0.16 [-0.59,0.35]
<i>cp8kv</i>	2.88 [2.31,3.60]	2.72 [2.27,3.35]	-2.72 [-3.35,-2.27]	0.24 [0.01,0.93]	0.78 [-0.01,1.47]	0.59 [0.00,1.11]
<i>5585v</i>	2.88 [2.02,3.67]	2.55 [1.81,3.36]	2.40 [1.46,3.31]	0.04 [0.00,0.55]	-0.41 [-1.97,0.62]	-0.2 [-0.76,0.32]
<i>j4nb3</i>	2.89 [2.32,3.34]	2.63 [1.84,3.26]	2.63 [1.84,3.26]	0.01 [0.00,0.73]	0.12 [-0.74,0.90]	0.16 [-0.35,0.76]
<i>REF04</i>	3.22 [2.04,4.48]	2.76 [1.93,3.85]	-2.76 [-3.84,-1.93]	0.19 [0.00,0.82]	1.20 [0.01,2.22]	0.42 [-0.08,0.84]
<i>hf4wj</i>	3.28 [2.49,4.11]	3.04 [2.36,3.83]	-3.04 [-3.82,-2.36]	0.34 [0.03,0.85]	1.31 [0.48,1.95]	0.38 [-0.16,0.84]
<i>REF01</i>	3.33 [2.08,4.72]	2.82 [1.99,4.02]	-2.82 [-4.02,-1.99]	0.24 [0.01,0.90]	1.46 [0.05,2.63]	0.49 [0.08,0.83]
<i>REF06</i>	3.64 [2.37,4.97]	3.10 [2.08,4.34]	-3.10 [-4.33,-2.08]	0.16 [0.00,0.68]	1.24 [-0.50,2.68]	0.31 [-0.14,0.72]
<i>REF03</i>	4.01 [2.74,5.34]	3.58 [2.66,4.78]	-3.58 [-4.78,-2.66]	0.17 [0.00,0.84]	1.20 [-0.53,2.54]	0.42 [-0.06,0.79]
<i>pku5g</i>	4.87 [4.06,5.68]	4.68 [3.90,5.49]	4.68 [3.90,5.49]	0.49 [0.03,0.90]	1.80 [0.28,2.99]	0.56 [0.00,0.96]
<i>po4g2</i>	5.46 [4.35,6.63]	5.17 [4.17,6.28]	5.17 [4.17,6.28]	0.51 [0.04,0.88]	2.33 [0.36,3.75]	0.56 [0.00,1.00]

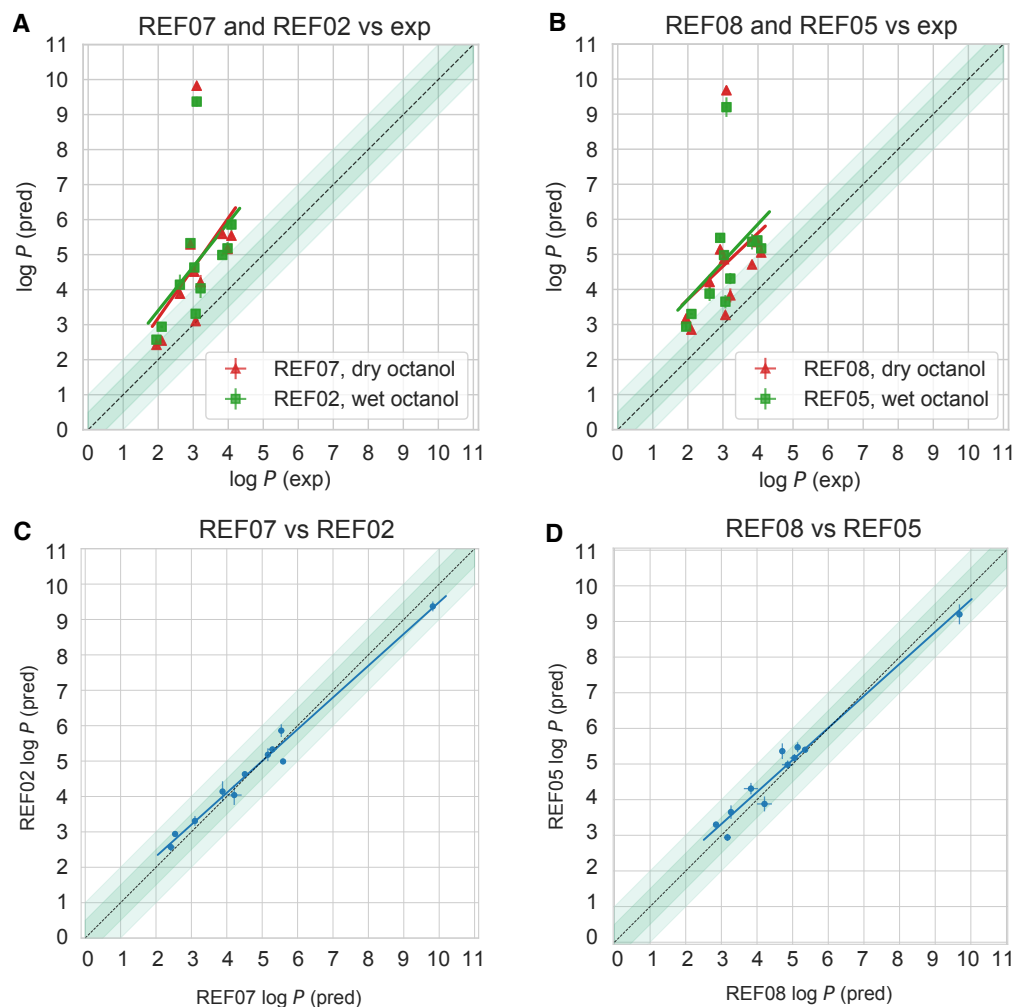


Figure 2.14: **Varying the amount of water in the octanol phase has no significant effect on the predicted $\log P$ in reference calculations, as discussed in section 2.5.2.** Comparison of predicted $\log P$ values to the experimental values using wet (27% water) and dry octanol phases and the (A) GAFF and (B) SMIRNOFF force field, from non-blinded reference calculations performed for this paper, shows no statistically significant difference in performance of methodologies. Comparison of the calculated $\log P$ using dry and wet octanol phases for (C) the GAFF force field and (D) the SMIRNOFF force field shows a small systematic difference.

Table 2.12: Comparison of force field parameters of the TIP3P, TIP3P-FB and OPC water models.

Water model	$q_H(e)^1$	$q_O(e)^1$	\angle HOH (deg)	$l_1(\text{\AA})^2$	$l_2(\text{\AA})^3$	$\sigma_{LJ}(\text{\AA})^4$	$\epsilon_{LJ}(\text{kJ/mol})^4$
TIP3P	0.417	-0.834	104.52	0.9572	-	3.151	0.636
TIP3P-FB	0.424	-0.848	108.15	1.0118	-	3.178	0.652
OPTIMAL POINT CHARGE	0.679	-1.358	103.6	0.8724	0.1594	3.167	0.89

¹ Corresponds to the hydrogen and oxygen charges.

² Corresponds to the bond length between the oxygen and hydrogen atoms.

³ Corresponds to the length between the oxygen atom and virtual site.

⁴ Corresponds to the Lennard-Jones (LJ) parameters of the oxygen.

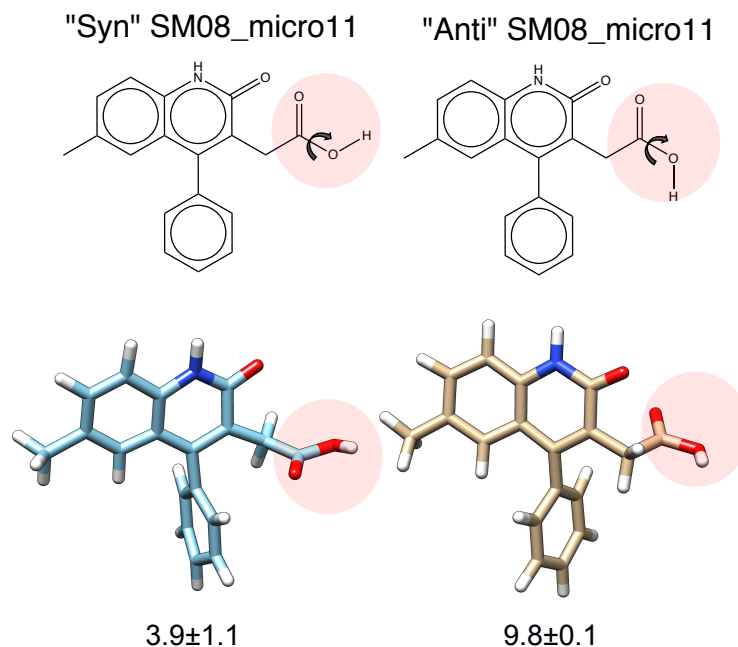


Figure 2.15: Shown here are the 2- and 3D structures of SM08_micro011 with the carboxylic acid in “anti” and “syn” conformation. The dihedral angle is indicated by the arrow around the carbon and oxygen atom. The calculated $\log P$ is included for comparison. The charges pertaining to each conformation are listed in Figure 2.13 and transition data is available in Figure 2.16

Table 2.13: Comparison of the charges assigned to the syn and anti conformation of SM08_micro011 in the DFE protocol.

anti conformation				syn conformation			
atom number	atom type	atom name	charge	atom number	atom type	atom name	charge
1	C1	C1	0.1252	1	C1	C1	-0.1205
2	C2	C2	-0.0073	2	C1	C2	-0.1322
3	C2	C3	0.0000	3	C1	C3	-0.1322
4	C2	C4	0.0000	4	C1	C4	-0.1112
5	C2	C5	-0.0074	5	C1	C5	-0.1112
6	C2	C6	-0.0118	6	C1	C6	-0.0742
7	C2	C7	0.0000	7	C1	C7	-0.1856
8	C2	C8	0.0236	8	C1	C8	-0.0653
9	C2	C9	-0.0196	9	C1	C9	-0.0839
10	C2	C10	0.3597	10	C1	C10	-0.1199
11	O1	O1	-0.2755	11	C1	C11	-0.1160
12	N1	N1	-0.1461	12	C1	C12	0.0947
13	C1	C11	0.0310	13	C1	C13	0.1024
14	C2	C12	0.3291	14	C1	C14	-0.1731
15	O1	O2	-0.1890	15	C1	C15	0.6987
16	O2	O3	-0.2911	16	C1	C16	0.6458
17	C2	C13	-0.0118	17	C2	C17	-0.0498
18	C2	C14	0.0000	18	C2	C18	-0.0780
19	C2	C15	0.0000	19	N1	N1	-0.4371
20	C2	C16	0.0000	20	O1	O1	-0.6386
21	C2	C17	0.0000	21	O1	O2	-0.5474
22	C2	C18	0.0000	22	O2	O3	-0.6145
23	H1	H1	-0.0393	23	H1	H1	0.1352
24	H1	H2	-0.0393	24	H1	H2	0.1377
25	H1	H3	-0.0393	25	H1	H3	0.1377
26	H2	H4	0.0000	26	H1	H4	0.1459
27	H2	H5	0.0000	27	H1	H5	0.1459
28	H2	H6	0.0000	28	H1	H6	0.1381
29	H3	H7	0.0865	29	H1	H7	0.1406
30	H1	H8	-0.0393	30	H1	H8	0.1469
31	H1	H9	-0.0393	31	H2	H9	0.0455
32	H4	H10	0.2010	32	H2	H10	0.0455
33	H2	H11	0.0000	33	H2	H11	0.0455
34	H2	H12	0.0000	34	H2	H12	0.1028
35	H2	H13	0.0000	35	H2	H13	0.1028
36	H2	H14	0.0000	36	H3	H14	0.3362
37	H2	H15	0.0000	37	H4	H15	0.4428



Figure 2.16: For the DFE method, the starting conformation impacts the number of C-O dihedral transitions for SM08_micro011, influencing sampling. Here is the transition data for the C-O dihedral in Figure 2.15, with charges listed in Table 2.13, for the DFE method (run in triplicate). In the “anti starting position” the torsion remains “anti” throughout the simulation, while the “syn starting position” allows transitions.

Chapter 3

Evaluation of $\log P$, pK_a , and $\log D$ predictions from the SAMPL7 blind challenge

Teresa Danielle Bergazin, Nicolas Tielker, Yingying Zhang, Junjun Mao, M. R. Gunner, Karol Francisco, Carlo Ballatore, Stefan M. Kast, and David L. Mobley.

Journal of Computer-Aided Molecular Design volume 35, pages771–802 (2021)

doi: 10.1007/s10822-021-00397-3

Publication Date (Web): June 24, 2021

3.1 Abstract

The Statistical Assessment of Modeling of Proteins and Ligands (SAMPL) challenges focuses the computational modeling community on areas in need of improvement for rational drug design. The SAMPL7 physical property challenge dealt with prediction of octanol-water

partition coefficients and pK_a for 22 compounds. The dataset was composed of a series of N-acylsulfonamides and related bioisosteres. 17 research groups participated in the $\log P$ challenge, submitting 33 blind submissions total. For the pK_a challenge, 7 different groups participated, submitting 9 blind submissions in total. Overall, the accuracy of octanol-water $\log P$ predictions in the SAMPL7 challenge was lower than octanol-water $\log P$ predictions in SAMPL6, likely due to a more diverse dataset. Compared to the SAMPL6 pK_a challenge, accuracy remains unchanged in SAMPL7. Interestingly, here, though macroscopic pK_a values were often predicted with reasonable accuracy, there was dramatically more disagreement among participants as to which microscopic transitions produced these values (with methods often disagreeing even as to the sign of the free energy change associated with certain transitions), indicating far more work needs to be done on pK_a prediction methods.

3.1.1 Keywords

octanol-water partition coefficient · $\log P$ · blind prediction challenge · SAMPL · free energy calculations · solvation modeling · pK_a · Macroscopic pK_a · Microscopic pK_a · Macroscopic protonation state · Microscopic protonation state · Relative free energy

3.1.2 Abbreviations

SAMPL Statistical Assessment of the Modeling of Proteins and Ligands

$\log P$ \log_{10} of the organic solvent-water partition coefficient (K_{ow}) of neutral species

$\log D$ \log_{10} of organic solvent-water distribution coefficient (D_{ow})

pK_a $-\log_{10}$ of the acid dissociation equilibrium constant

SEM Standard error of the mean

RMSE Root mean squared error

MAE Mean absolute error

τ Kendall's rank correlation coefficient (Tau)

R² Coefficient of determination (R-Squared)

QM Quantum Mechanics

MM Molecular Mechanics

DL Database lookup

LFER Linear free energy relationship

QSPR Quantitative structure-property relationship

ML Machine learning

LEC Linear empirical correction

3.2 Introduction

Computational modeling aims to enable molecular design, property prediction, prediction of biomolecular interactions, and provide a detailed understanding of chemical and biological mechanisms. Methods for making these types of predictions can suffer from poor or unpredictable performance, thus hindering their predictive power. Without a large scale evaluation of methods, it can be difficult to know what method would yield the most accurate predictions for a system of interest. Large scale comparative evaluations of methods are rare and difficult to perform because no individual group has expertise in or access to all relevant methods. Thus, methodological studies typically focus on introducing new methods, without extensive comparisons to other methods.

The Statistical Assessment of Modeling of Proteins and Ligands (SAMPL) challenges tackle modeling areas in need of improvement, focusing the community on one accuracy-limiting problem at a time. In SAMPL challenges, participants predict a target property such as solvation free energy, given a target set of molecules. Then the corresponding experimental data remains inaccessible to the public until the challenge officially closes. By focusing on specific areas in need of improvement, SAMPL helps drive progress in computational modeling.

Here, we report on a SAMPL7 physical property challenge that focused on octanol-water partition coefficients ($\log P$) and $\text{p}K_{\text{a}}$. The $\text{p}K_{\text{a}}$ of a molecule, or the negative logarithm of the acid–base dissociation constant, is related to the equilibrium constant for the dissociation of a particular acid into its conjugate base and a free proton. The $\text{p}K_{\text{a}}$ also corresponds to the pH at which the corresponding acid and its conjugate base each are populated equally in solution. Given that the $\text{p}K_{\text{a}}$ corresponds to a transition between specific protonation states, a given molecule may have multiple $\text{p}K_{\text{a}}$ values.

The $\text{p}K_{\text{a}}$ is an important physical property to take into account in drug development. The $\text{p}K_{\text{a}}$ value is used to indicate the strength of an acid. A lower $\text{p}K_{\text{a}}$ value indicates a stronger acid, indicating the acid more fully dissociates in water. Molecules with multiple ionizable centers have multiple $\text{p}K_{\text{a}}$ values, and knowledge of the $\text{p}K_{\text{a}}$ of each of the ionizable moieties allows for the percentage of ionised/neutral species to be calculated at a given pH (if activity coefficients are known/assumed). $\text{p}K_{\text{a}}$ plays a particularly important role in drug development because the ionization state of molecules at physiological pH can have important ramifications in terms of drug-target interactions (e.g., ionic interactions) and/or by influencing other key determinants of drug absorption, distribution, metabolism and excretion (ADME) [132], such as lipophilicity, solubility, membrane permeability and plasma protein binding [36].

Accurate $\text{p}K_{\text{a}}$ predictions play a critical role in molecular design and discovery as well since

pK_a comes up in so many contexts. For example, inaccurate protonation state predictions impair the accuracy of predicted distribution coefficients such as those from free energy calculations. Similarly, binding calculations can be affected by a change in protonation state [11]. If a ligand in a protein-ligand system has a different protonation state in the binding pocket compared to when the molecule is in the aqueous phase, then this needs to be taken into account in the thermodynamic cycle when computing protein-ligand binding affinities.

Multiprotic molecules, and those with multiple tautomeric states, have two types of pK_a , microscopic and macroscopic. The *microscopic* pK_a applies to a specific transition or equilibrium between microstates, i.e. for a transition between a specific tautomer at one formal charge and that at another formal charge (e.g. two states at different formal charges in Figure 3.2). It relates to the acid dissociation constant associated with that specific transition. As a special case, a microscopic pK_a sometimes refers to the pK_a of deprotonation of a single titratable group while all the other titratable and tautomerizable functional groups of the same molecule are held fixed, but this might possibly not reflect the dominant deprotonation pathway of a given acidic tautomer if the base state possesses energetically favored alternate tautomers. There is no pK_a between two tautomers with the same formal charge because they have the same number of protons so their relative probability is independent of pH. The pH-independent free energy difference between them determines their relative population [190].

At some level, the macroscopic pK_a can be thought of as describing the acid dissociation constant related to the loss of a proton from a molecule regardless of which functional group the proton is dissociating from, but it may be more helpful to think of it (in the case of polyprotic molecules) as a macroscopic observable describing the collective behavior of various tautomeric states as the dominant formal charge of the molecule shifts. In cases where a molecule has only a single location for a titratable proton, the microscopic pK_a

becomes equal to the macroscopic pK_a .

In the current challenge, we explored how well methods could predict macroscopic pK_a 's through microscopic pK_a calculations.

The partition coefficient ($\log P$) and the distribution coefficient ($\log D$) are relevant to drug discovery, as they are used to describe lipophilicity. Lipophilicity influences drug-target and off-target interactions through hydrophobic interactions, and relatively high lipophilicity results in reduced aqueous solubility and increased likelihood of metabolic instability [140].

Prediction of partitioning and distribution has some relevance to drug distribution. Particularly, partitioning and distribution experiments involve a biphasic system with separated aqueous and organic phases, such as water and octanol, so such experiments have some of the features of the interface between blood or cytoplasm and the cell membrane [74, 117] and thus improved predictive power for partitioning and distribution may pay off with an improved understanding of such *in vivo* events.

Methods to predict $\log P/\log D$ may also use (and test) some of the same techniques which can be applied to binding predictions. Both types of calculations can use solvation free energies and partitioning between environments (though this could be avoided by computing the transfer free energy). Such solute partitioning models are simple test systems for the transfer free energy of a molecule to a hydrophobic environment of a protein binding pocket, without having to account for additional specific interactions which are present in biomolecular binding sites. Thus partitioning and distribution calculations allow separating force-field accuracy from errors related to conformational sampling of proteins and protonation state predictions of proteins and ligands.

The $\log P$ is usually defined as the equilibrium concentration ratio of the neutral state of a

substance between two phases:

$$\log P = \log_{10} K_{ow} = \log_{10} \frac{[\text{unionized solute}]_{\text{octanol}}}{[\text{unionized solute}]_{\text{water}}} \quad (3.1)$$

Strictly speaking, this definition of the partition coefficient P as a thermodynamic equilibrium constant is independent of total solute concentration in the infinite dilution limit only. This reference state is commonly assumed in physics-based prediction models. The log P prediction challenge explores how well current methods are able to model the transfer free energy of molecules between different solvent environments without any complications coming from predicting protonation states.

3.2.1 Motivation for the log P and pK_a challenge

Previous SAMPL challenges have looked at the prediction of solvation free energies [162, 81, 153, 69, 155], guest-host [159, 229, 186, 187, 14, 158, 100] and protein-ligand binding affinities [154, 28, 67, 83, 113, 114, 49], pK_a [90, 196, 21, 234, 207, 177, 179], distribution coefficients [19, 191, 99, 106], and partition coefficients [87, 61, 233, 97]. These challenges have helped uncover sources of error, pinpoint the reasons various methods performed poorly or well and their strengths and weaknesses, and facilitate dissemination of lessons learned after each challenge ends, ultimately leading to improved methods and algorithms.

Several past challenges focused on solvation modeling in order to help address this accuracy-limiting component of protein-ligand modeling. The SAMPL0 through SAMPL4 challenges included hydration free energy prediction, followed by cyclohexane-water distribution coefficient prediction in SAMPL5, and octanol-water distribution coefficient prediction in SAMPL6. Large errors were observed in the SAMPL5 cyclohexane-water log D prediction challenge due to tautomers and protonation states not being taken into account [? 176] or adequately handled. Many participants reported log P predictions in place of log D pre-

dictions, in part because the different ionization states of the molecules were thought not to be particularly relevant in the challenge, but this proved not to be the case. Methods that treated multiple protonation and tautomeric states and incorporated pK_a corrections (which relies on accurate pK_a prediction) in their predictions performed better [176].

In order to pinpoint sources of error in $\log D$ predictions, separate $\log P$ and pK_a challenges were organized for SAMPL6 [87, 91, 90, 88]. Better prediction performance was seen in the SAMPL6 octanol-water $\log P$ challenge compared to the SAMPL5 cyclohexane-water $\log D$ challenge. Performance improved in SAMPL6 for several reasons. First, the latter challenge avoided the pK_a prediction problem. Second, far more experimental training data was available (aiding empirical and implicit QM methods). Finally, the more narrow chemical diversity in SAMPL6 may have helped participants. For the present SAMPL7 physical properties challenge, we focused on assessing the accuracy of $\log P$ and pK_a predictions, and then combined pK_a and $\log P$ predictions to obtain $\log D$ predictions.

3.2.2 Historical SAMPL pK_a performance

During the SAMPL6 challenge a broad range of conceptually different empirical and physics-based computational methods were used to predict pK_a values, as discussed in the overview paper [91]. To provide some context for the results of the SAMPL7 challenge the main results are summarized here.

The empirical approaches used during SAMPL6 can be divided into three categories, Database Lookup (DL), Linear Free Energy Relationship (LFER), and Quantitative Structure–Property/Machine Learning (QSPR/ML) approaches [155]. The physical approaches can be divided into pure quantum–mechanical (QM) methods, QM with a linear empirical correction (QM+LEC) to account for the free energy of the proton in solution or potential systematic errors caused by the chosen method, and QM in combination with molecular mechanics (QM+MM). Gen-

erally speaking, the empirical methods require significantly less computational effort than their physics-based counterparts once they are parameterized.

The best-performing models included four empirical methods and one QM-based model. These five methods were able to predict the acidity constants of the challenge compounds to within 1 pK_a unit. In fact, while most empirical models – except for the DL and two of the five QSPR/ML approaches – were able to predict the acidity constants to within about 1.5 pK_a units, the range of predictions was much wider for the QM-based models.

In SAMPL6, many groups submitted multiple predictions to test the performance of different variations using the same basic methodology, such as exploring different levels of theory, model parameters, or conformational ensembles.

Well-performing empirical models included both LFER methods, such as ACD/pKa Classic (submission ID *xmyhm*) and Epik Scan (*nb007*), and QSPR/ML methods such as MoKa (*nb017*) and S+pKa (*gyuhx*), all performing with root mean square errors (RMSE) between 0.73 and 0.95 pK_a units [1, 197, 156, 200]. These well-established tools thus demonstrated their reliability and quality.

Among the physics-based models, the most straightforward approach involved calculation of the acidity constants without any empirical corrections, including the experimental value for the free energy of solvation of the proton [212]. One group applied different calculation schemes to the compounds of the SAMPL6 challenge that differed in the use of gas phase and/or solution phase geometries as well as additional high-level single point gas phase calculations [234]. While the results achieved by this method were quite promising, with an initial RMSE of 1.77 pK_a units (*ryzue*) that could be improved to 1.40 by including a standard state correction and a different value for the free energy of the proton, the authors also showed the effectiveness of a simple linear regression scheme to correct the raw acidity constants. In this case the RMSE of the best-performing model decreased further from 1.40

to 0.73 $\text{p}K_{\text{a}}$ units after regression.

This type of empirical correction was used by most QM-based approaches, including the best-performing method of the SAMPL6 challenge [91], improving some systematic deficiencies of the QM level of theory and basis sets and accounting for the proton’s solvation free energy. The best-performing QM+LEC method, *xvxzd*, achieved an RMSE of 0.68 $\text{p}K_{\text{a}}$ units during the challenge using the COSMO-RS solvation model. This also made it the best-performing model overall, with two other methods using the same solvation model only slightly worse (*yqkga* and *8xt50*, with RMSEs of 1.01 and 1.07 $\text{p}K_{\text{a}}$ units, respectively [91, 105, 177]).

A QM+LEC method using a different solvation approach, EC-RISM, only achieved an RMSE of 1.70 $\text{p}K_{\text{a}}$ units for the submitted model (*nb001*), but a post-submission optimization of the conformer generation workflow and the electrostatic interactions improved the RMSE to 1.13, which is more in line with the other well-performing QM+LEC methods [207]. The CPCM implicit solvation model was used by one group [91, 196] and performed only slightly worse than COSMO-RS (RMSEs from the paper do not agree with official numbers. Only officially submitted ones are discussed here). For these two models, differing only by training either a single LEC for all compounds (*35bdm*) or two separate LECs for deprotonations of neutral compounds to anions and deprotonations of cations to neutral compounds (*p0jba*), the RMSEs were 1.72 and 1.31 $\text{p}K_{\text{a}}$ units, respectively. These results show that accurate $\text{p}K_{\text{a}}$ values can be predicted when using the QM+LEC approach with different solvation models.

A slightly different approach was used by one participant (*0wfsz0*) where QM calculations of the free energy of deprotonation and thermodynamic integration, an MM method, were combined to calculate the difference of the solvation free energies between the acid and its conjugate base [179]. This approach yielded an average level of performance, with an RMSE of 2.89 for the macroscopic acidity constants calculated from the submitted microscopic acidity constants, excluding two compounds (SM14 and SM18) from the analysis as they

exhibited multiple pK_a values too close to each other.

3.2.3 Approaches to predicting small molecule pK_a 's

Calculations of aqueous pK_a values have a long history in computational chemistry, with methods ranging from direct quantum-mechanical approaches for determining the free energy of protonated and deprotonated species in solution using explicit, implicit, or hybrid solvation models, to continuum electrostatics-based computations of relative pK_a shifts, and empirical or rule-based algorithms, as summarized in a number of review articles, e.g. Alongi et al. [12], and Liao et al. [126] and in the SAMPL6 overview papers [90, 91].

Computational methods typically designate tautomeric states (“microstates”) for acid and base forms of a compound separated by a unit charge upon (de-)protonation. Their free energies can be linked individually in a pair-wise manner (“microstate transitions”) to yield so-called microstate pK_a values from which the macroscopic pK_a can be determined [30]. Alternatively, the tautomer free energies, combined across the underlying conformational states, contribute to the ratio of partition functions representing acid and base forms, allowing the direct calculation of macroscopic acidity constants [206]. A complication arises if, as is common practice with quantum-mechanical approaches, the difference of solution-state (standard) free energies for differently charged species, $G(A_{\text{aq}}^-)$ and $G(\text{HA}_{\text{aq}})$ for a general reaction



are scaled by a “slope” factor m and augmented by an intercept parameter b to account for the free energy of the proton, yielding a regression equation, given here for microstate j of

the base and k of the acid form, respectively,

$$\text{p}K_{a,jk} = b + \frac{m}{RT \ln 10} [G_j(\text{A}^-) - G_k(\text{HA})] \quad (3.3)$$

where slope and intercept are typically adjusted with respect to databases of experimental $\text{p}K_a$ values [206] and RT has the usual thermodynamic meaning. Here G denotes the Gibbs free energy, but a similar expression would hold for Helmholtz free energy depending on the choice of ensemble.

As derived in Tielker et al. [206], statistics over all connected microstates (in the “state transition” (ST) approach) and *a priori* partition function summation (in the “partition function” (PF) approach) are identical if and only if $m = 1$, though in practice the difference is usually negligible.

For the SAMPL7 $\text{p}K_a$ challenge, participants were required to submit predictions in a novel format, reporting transition free energies between microstates as in the “ ΔG^0 ” formalism outlined in Gunner et al. [80] (and similar to the work of Selwa et al. [196]). Here, the pH-dependent free energy change between “states” k and j is defined by rewriting the well-known Henderson-Hasselbalch equation for, e.g., the general reaction (Eq. 3.3) in the form

$$\Delta G_{jk}(\text{pH}) = \Delta m_{jk} C_{\text{units}} (\text{pH} - \text{p}K_{a,jk}) \quad (3.4)$$

with $C_{\text{units}} = RT \ln 10$ and, for a transition away from the reference state which involves loss of a proton, $\Delta m_{jk} = -1$, denoting the charge difference between the “reference state” k (second index, usually taken as a selected neutral microstate, in this case HA_{aq}) and the

target state j .

For the thermodynamic standard state at $\text{pH} = 0$ we can write

$$\Delta G_{jk}^0 = -\Delta m_{jk} C_{\text{units}} \text{p}K_{a,jk} \quad (3.5)$$

which shows that ΔG_{jk}^0 can be identified with a formal free energy of reaction. An advantage of this approach is that closed thermodynamic cycles by summing over ΔG_{jk}^0 with identical reference k would add to zero for consistent computational methods, which can serve as an added value for testing theoretical frameworks [80].

The macroscopic $\text{p}K_a$ is obtained by computing the total fraction of all microstates with charge q and $j \in q$ via

$$x_{j \in q}(\text{pH}) = \frac{\exp[-\Delta G_{j \in q,k}(\text{pH})/RT]}{\sum_i \exp[-\Delta G_{ik}(\text{pH})/RT]} \quad (3.6)$$

and solving, usually numerically, for the pH at which

$$x_{j \in q(1)}(\text{pH}) = x_{j \in q(2)}(\text{pH}) \quad (3.7)$$

for adjacent net charges $q(1)$ and $q(2)$. At this pH , $\text{p}K_a = \text{pH}$ for these particular charge states, and this approach constitutes a formal “titration”.

Outlining the connection between the ΔG^0 and the ST and PF formalisms [206] is useful

for practitioners who directly compute microstate free energies (including corresponding tautomerization free energies for which no $\text{p}K_{\text{a}}$ is defined) or microstate transition $\text{p}K_{\text{a}}$ values for single deprotonation reactions where a specific reaction direction is by definition implied. The general algorithm is as follows, with subscript order $\text{p}K_{a,jk}$ implying the reaction $j \rightarrow k^- + \text{H}^+$ for any total charge on j and subscript order ΔG_{jk}^0 meaning the reaction $k(+m\text{H}^+) \rightarrow j(+n\text{H}^+)$ with neutral k . For all states i not equal to the neutral reference microstate k we have

- a) If $q(i) = 0$, $\Delta G_{ik}^0 = m\Delta G^0(k \rightarrow i)$
- b) If $q(i) - q(k) = +1$ (the reaction is $k + \text{H}^+ \rightarrow i^+$), then $\Delta G_{ik}^0 = -C_{\text{units}}\text{p}K_{a,ik}$
- c) If $q(i) - q(k) = -1$ (the reaction is $k \rightarrow i^- + \text{H}^+$), then $\Delta G_{ik}^0 = +C_{\text{units}}\text{p}K_{a,ki}$
- d) If $q(i) - q(k) = +2$ (the reaction is $k + 2\text{H}^+ \rightarrow i^{2+}$ via the individual reactions $k + \text{H}^+ \rightarrow j^+$ and $j^+ + \text{H}^+ \rightarrow i^{2+}$), then $\Delta G_{ik}^0 = -C_{\text{units}}(\text{p}K_{a,jk} + \text{p}K_{a,ij})$
- e) If $q(i) - q(k) = -2$ (the reaction is $k \rightarrow i^{2-} + 2\text{H}^+$ via the individual reactions $k \rightarrow j^- + \text{H}^+$ and $j^- \rightarrow i^{2-} + \text{H}^+$), then $\Delta G_{ik}^0 = +C_{\text{units}}(\text{p}K_{a,kj} + \text{p}K_{a,ji})$

This scheme is readily generalized to changes of more than two unit charges. The scaling by the factor m in (a) guarantees consistency over closed thermodynamic cycles in the common case of non-zero slope parameter for QM-based models.

To demonstrate how macroscopic $\text{p}K_{\text{a}}$ values computed this way relate to ST and PF results it is instructive to treat the simple example of a two-tautomer acid in equilibrium with a single-tautomer base, i.e.



for which Eq. (3.3) yields [206]

$$K_a^{ST} = \left(\frac{1}{K_{a,1}} + \frac{1}{K_{a,2}} \right)^{-1} = 10^{-b} \frac{\exp[-mG(A^-)/RT]}{\exp[-mG(HA_1)/RT] + \exp[-mG(HA_2)/RT]} \quad (3.9)$$

Following the algorithm for ΔG_{jk}^0 above with HA_1 assumed as neutral reference and augmenting the pH dependence according to Eq. (3.4) we have

$$\Delta G(HA_1) = 0 \quad (3.10)$$

$$\Delta G(HA_2) = m[G(HA_2) - G(HA_1)] \quad (3.11)$$

$$\Delta G(A^-) = -C_{\text{units}}(\text{pH} - \text{p}K_{a,1}) = m[G(A^-) - G(HA_1)] - C_{\text{units}}(\text{pH} - b) \quad (3.12)$$

From Eq. 3.5 and equating neutral and charged molar fractions it follows from $x(\text{HA}) = x(\text{A}^-)$

$$1 + \exp \{ -m [G(\text{HA}_2) - G(\text{HA}_1)] / RT \} = 10^{-b} \exp \{ +m [G(\text{HA}_1) - G(\text{A}^-)] / RT \} / K_a \quad (3.13)$$

which, upon rearrangement and comparison with (3.9), yields

$$K_a = K_a^{\text{ST}} \quad (3.14)$$

Generalization to more complex tautomeric mixtures and arbitrary reference states is possible, the latter by recognizing that these would only imply cancelling additive constants. The ΔG^0 and ST formalisms are therefore equivalent, as is the PF approach for $m = 1$.

3.2.4 Approaches to predicting $\log P$

Approaches for predicting octanol-water $\log P$ values include physical modeling methods, such as quantum mechanics (QM) and molecular mechanics (MM) approaches, and empirical knowledge-based prediction methods, such as contribution-type approaches. We give some brief background on these prediction methods.

QM approaches use a numerical solution of the Schrödinger equation to estimate solvation free energies and partitioning. These approaches are not practical for larger systems, so certain approximations need to be made so that they can be used for calculating transfer free energies. Methods typically represent the solvent using an implicit solvent model and make the assumption that the solute has a single or a small number of dominant conformations in the aqueous and non-aqueous phase. The accuracy of predictions can be influenced by the basis set, level of theory, and the tautomer used as input. Implicit solvent models are used to represent both octanol and water, and these models are often highly parameterized on experimental solvation free energy data. The abundance of training data contributes to the success of QM methods, much like empirical prediction methods. Solvent models such as SMD [136], the SM-n series of models [137], and COSMO-RS [127, 106, 104, 103, 107] are frequently used by SAMPL participants.

MM approaches use a force field which gives the energy of a system as a function of the atomic positions and are usually used by SAMPL participants to compute solvation free energies and $\log P$ values. Force fields can be fixed charge and additive, or polarizable [124, 99], and typically include all atoms, though this need not always be the case. These approaches are usually applied by integrating the equations of motion to solve for the time evolution of the system. Force fields such as GAFF [219], GAFF2 [215], CGenFF [214], and OPLS-AA [51], and water models such as TIP3P [98], TIP4P [98], OPC3 [93] are frequently used in SAMPL challenges [87]. Free energy calculations can be combined with MM methods to give a partitioning estimate. These types of calculations often use alchemical free energy methods to estimate phase transfer via a non-physical thermodynamic cycle. Some examples of alchemical approaches include non-equilibrium switching [180, 96] and equilibrium alchemical free energy calculations [235] analyzed via thermodynamic integration [102] or BAR/MBAR estimation [27, 198]. Such simulations can also use techniques like Hamiltonian replica exchange molecular dynamics.

Some limitations of MM approaches include the accuracy of the force field and the limitation that motions can only be captured in simulations that are faster than simulation timescales. The state of the molecule that is used as input is also important— usually, a single tautomer/protonation state is selected and held fixed throughout the simulation, which can introduce errors if the wrong state was selected or if there are multiple relevant states.

Empirical prediction models are trained on experimental data and can be used to quickly characterize large virtual libraries. These include additive group methods, such as fragment- or atom-contribution approaches, and quantitative structure-property relationship (QSPR) methods. In atom contribution approaches, the $\log P$ is equal to the sum of contributions from the individual atom types multiplied by the number of occurrences of each in the molecule. These methods make the assumption that each atom contributes a certain amount to the solvation free energy and that these contributions are additive to the $\log P$

. In fragment (or group) contribution approaches, the $\log P$ is equivalent to the sum of the contributions from the fragment groups (more than a single atom), and typically uses correction terms that consider intramolecular interactions. These approaches are generally calculated by adding together the sum of the fragment contributions times the number of occurrences and the sum of the correction contributions times the number of occurrences in the molecule. The other class of empirical $\log P$ prediction approaches relies on QSPR. In QSPR, molecular descriptors are calculated and then used to make $\log P$ predictions. Descriptors can vary in complexity- some rely on simple counts of heteroatoms and carbon, while others are derived from correlating the 3D shape, electrostatic, and hydrogen bonding characteristics with the $\log P$ of the molecule. To find the $\log P$, a regression model gets derived by fitting the descriptor contributions to experimental data. Machine learning approaches such as random forest models, deep neural network models, Gaussian processes, support vector machines, and ridge regression [178, 195] belong under this category.

Empirical methods tend to benefit from a large and diverse training set, especially when there’s a large body of experimental data to train on, such as octanol-water data like in the present and previous $\log P$ challenge [87]. However, empirical methods can experience problems if a training set has an underrepresented functional group. Additionally, these techniques are geared towards partitioning predictions, and, unlike physical-based methods, are not able to be applied to protein-ligand binding.

3.3 Challenge design and evaluation

3.3.1 General challenge structure

The SAMPL7 physical property challenge focused on pK_a , partitioning, and permeability. As reported separately, KF and CB collected a set of measured water-octanol $\log P$, $\log D$,

and pK_a values for 22 compounds, along with PAMPA permeability values [64]. Since this was our first time hosting a permeability challenge, and these calculations remain challenging for many methods, we did not have enough participants to form meaningful conclusions (one participant submitted two sets of predictions in total) so the challenge is not discussed in this paper, but we provide a link to the challenge’s GitHub page (https://github.com/samplchallenges/SAMPL7/tree/master/physical_property/permeability).

The SAMPL7 challenge molecules had weights that ranged from 227 to 365 Da, and varied in flexibility (the number of non-terminal rotatable bonds ranged from 3-6). The dataset had experimental $\log P$ values in the range of 0.58–2.96, pK_a values in the range of 4.49–11.93, and $\log D$ values in the range of -0.87–2.96. Information on experimental data collection is presented elsewhere [64].

The physical properties challenge was announced on June 29th, 2020 and the molecules and experimental details were made available at this time. Additional input files, instructions, and submission templates were made available afterward and participant submissions were accepted until October 8th, 2020. Following the conclusion of the blind challenge, the experimental data was made public on October 9th, 2020, and results were discussed in a virtual workshop (on November 2-5, 2020) (SAMPL Community Zenodo page <https://zenodo.org/communities/sampl/?page=1&size=20>)

A machine-readable submission file format was specified for blind submissions. The submission files included fields for naming the method of the computational protocol, listing the average compute time across all of the molecules, detailing the computing and hardware used, listing the major software packages and the versions that were used, and a free text method section for providing the detailed documentation of each method, the values of key parameters with units, and to explain how statistical uncertainties were estimated. There was also a field where participants indicated whether or not they wanted their submission formally evaluated. In addition to their predictions, participants were asked to estimate the

statistical error (expressed as a standard error of the mean (SEM)) associated with their predictions, and the uncertainty of their model. The SEM captures the statistical uncertainty of a method’s predictions, and the model uncertainty corresponds to the method’s expected prediction accuracy, which estimates how well a participant expects their predicted values will agree with experiment. Historically, model uncertainty estimates have received relatively little attention from participants, but we retain hope that participants may eventually predict useful model uncertainties since users benefit from knowing the accuracy of a predicted value.

Participants had the option of submitting predictions from multiple methods, and were asked to fill out separate template files for each different method. Each participant or organization could submit predictions from multiple methods, but could only have one ranked submission. Allowing multiple submissions gave participants the opportunity to submit prediction sets to compare multiple methods or to investigate the effect of varying parameters of a single method. All of the submissions were assigned a short descriptive method name based on the name they provided for their protocol in their submission file. This descriptive method name was used in the analysis and throughout this paper and is presented in Tables 3.1, 3.3, and 3.5.

3.3.2 log P challenge structure

The SAMPL7 log P challenge consisted of predicting the water-octanol partition coefficients of 22 molecules. Our goal was to evaluate how well current models can capture the transfer free energy of small molecules between different solvent environments through blind predictions. challenge participants were asked to predict the difference in free energy for the neutral form of each molecule between water and octanol. For the log P challenge, participants were required to report, for each molecule, the SAMPL7 molecule ID tag (the

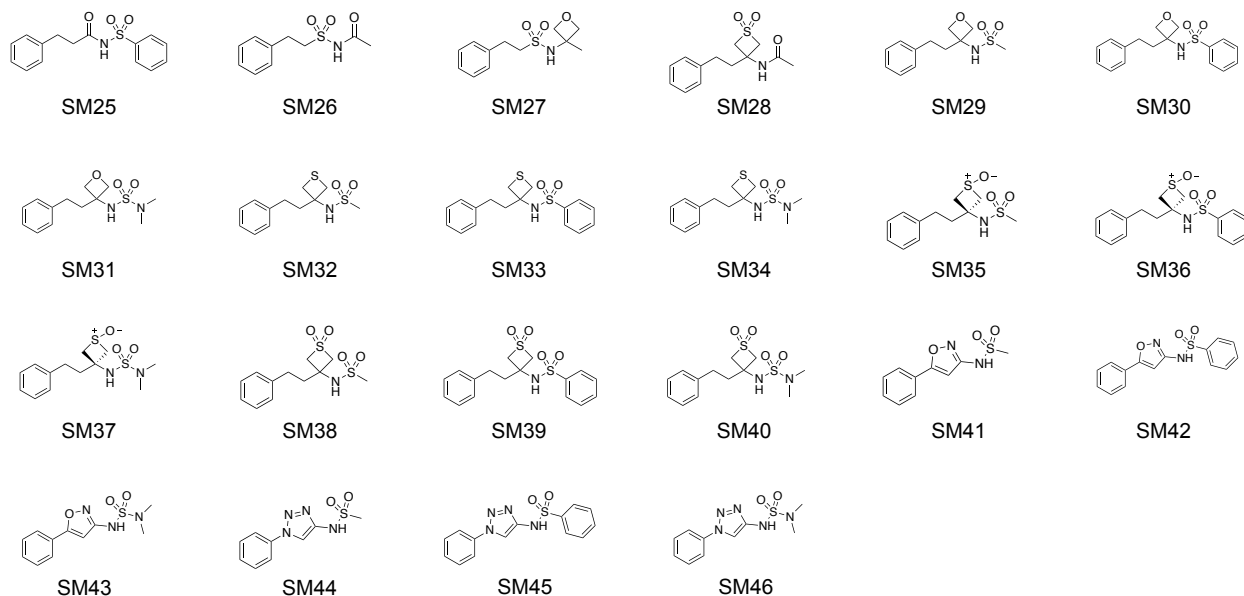


Figure 3.1: **Structures of the 22 molecules used for the SAMPL7 physical property blind prediction challenge.** Log of the partition coefficient between n-octanol and water was determined via potentiometric titrations using a Sirius T3 instrument. pK_a values were determined by potentiometric titrations using a Sirius T3 instrument. Log of the distribution coefficient between n-octanol and aqueous buffer at pH 7.4 were determined via potentiometric titrations using a Sirius T3 instrument, except for compounds SM27, SM28, SM30-SM34, SM36-SM39 which had $\log D_{7.4}$ values determined via shake-flask assay. PAMPA assay data includes effective permeability, membrane retention, and log of the apparent permeability coefficient. Permeabilities for compounds SM33, SM35, and SM39 were not determined. Compounds SM35, SM36 and SM37 are single *cis* configuration isomers. All other compounds are not chiral.

challenge provided neutral microstate), the microstate ID or IDs that were considered, and the predicted transfer free energy, transfer free energy SEM, and model uncertainty.

Participants were asked to categorize their methods as one of the five method categories—physical (QM), physical (MM), empirical, or mixed. Participants were asked to indicate their method based on the following definitions: Empirical models are prediction methods that are trained on experimental data, such as QSPR, machine learning models, artificial neural networks, etc. Physical models are prediction methods that rely on the physical principles of the system such as MM or QM based physical methods to predict molecular properties. Participants were asked to indicate whether their physical method was QM or MM based.

Methods taking advantage of both kinds of approaches were asked to be reported as “Mixed”. If a participant chose the “Mixed” category, they were asked to explain their decision in the method description section in their submission file.

We highlighted that octanol may be found in the aqueous phase, in case participants wanted to consider this in their predictions. The mole fraction of water in octanol was measured as 0.271 ± 0.003 at 25°C [117]

3.3.3 pK_a challenge structure

The SAMPL7 pK_a challenge consisted of predicting relative free energies between microstates (microscopic pK_a 's) to determine the macroscopic pK_a of 22 molecules. Our goal for the SAMPL7 pK_a challenge was to assess how well current pK_a prediction methods perform for the 22 challenge molecules through blind predictions.

We chose to have participants report relative free energies of microstates for simplicity of analysis. Particularly, for each molecule, participants were asked to predict the relative free energy, including the proton free energy, between our selected neutral reference microstate and the rest of the enumerated microstates for that molecule at a reference pH of 0 (see Section 3.2.3 on approaches to calculating pK_a). This can also be thought of as a reaction free energy for the microstate transition where the reference state is the reactant and the other microstate the product (though a proton may also be a product, depending on the direction of the transition). As an example for one molecule, we asked for the reaction free energy (relative free energy) associated with each of the reactions as seen in Figure 3.2. This approach differs from that used in past pK_a challenges, which typically focused on macroscopic pK_a predictions. The shift, here, helps resolve several key problems:

1. A macroscopic pK_a can be reported for the wrong microstates, leading to predictions

that are accidentally correct, but fundamentally wrong because the titration referred to a different states of the molecule.

2. Analysis of pK_a predictions requires pairing calculated macroscopic pK_a values with corresponding experimental macroscopic pK_a values [91] and such pairing can be very complex without information on which states are being predicted; while pairing is still required when specific transitions are predicted, it is aided by knowing *which* transitions are predicted (e.g. a -1 to 0 prediction from one participant can no longer accidentally be compared with a 0 to +1 transition from another participant)
3. Ultimately, populations and free energy differences between states drive the experimental measurements, so analysis ought to focus on state populations

In this work, all possible tautomers of each ionization (charge) state are defined as distinct protonation microstates. For the pK_a challenge, participants were required to report, for each molecule and each microstate they considered, the microstate ID of the reference state (selected by challenge organizers), the microstate ID of the microstate they were considering a transition to, the formal charge for the target microstate, and the predicted free energy change associated with a transition to the target microstate (Figure 3.2), the relative free energy SEM, and the relative free energy model uncertainty. In many cases, the transitions to be considered were a particular physical reaction involving a change in a single protonation state or tautomer. However, in some cases transitions involved a change of multiple protons (e.g. the F-A transition of Figure 3.2) and thus did not involve a single protonation or deprotonation event. Additionally, all transitions were defined as *away* from the reference state (and thus some involve gaining a proton, the opposite of a typical acid dissociation event), a point which caused confusion for a number of participants.

All predictions were required to use free energy units, in kcal/mol, which was another point which caused confusion for participants, as we received predictions in several different sets

of units and had to handle unit conversion after the challenge close.

Participants were asked to define and categorize their methods based on the following six method categories- experimental database lookup (DL), linear free energy relationship (LFER) [155], quantitative structure-property relationship or machine learning (QSPR/ML) [155], quantum mechanics without empirical correction (QM) models, quantum mechanics with linear empirical correction (QM+LEC), and combined quantum mechanics and molecular mechanics (QM+MM), or “Other”. If the “Other” category was chosen, participants were asked to explain their decision in the beginning of the method description section in their submission file.

Microstate enumeration

The SAMPL7 pK_a challenge participants were asked to predict relative free energies between microstates to determine the pK_a of molecules. We define distinct protonation microstates as all possible tautomers of each ionization (charge) state. Participants could consider any of these microstates in their predictions, and had the option of submitting others. Participants were provided a reference microstate for each compound, and asked to predict transition free energies to all microstates they viewed as relevant, relative to this reference state.

Here, we provided some enumeration of potential microstates that participants might want to consider. To do so, we used more than one toolkit to try and ensure all reasonable tautomers and protomers were included. Our microstates were generated using RDKit [182] and OpenEye QUACPAC [170] for protonation state/tautomer enumeration, and then cross checked with ChemAxon Chemicalize [2] and Schrodinger Epik [197, 78] to ensure we had not missed states. We also allowed participants to submit additional microstates they might view as important, and received one set of such submissions, which resulted in us adding a microstate with a +1 formal charge to molecules SM31 (SM31_micro002) and SM34

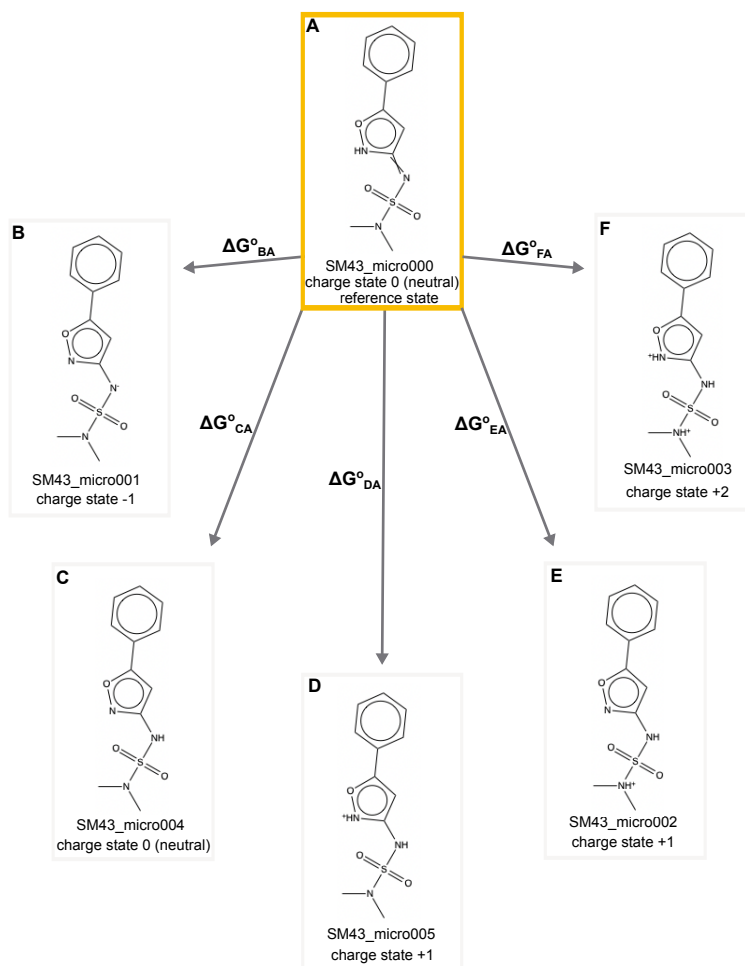


Figure 3.2: For each molecule in the SAMPL7 pK_a challenge we asked participants to predict the relative free energy between our selected neutral reference microstate and the rest of the enumerated microstates for that molecule. In this case, we asked for the relative state free energy including the proton free energy, which could also be called the reaction free energy for the microstate transition which has the reference state as the reactant and the alternate state as the product. Using SM43 as an example, participants were asked to predict the relative free energy between SM43_micro000 (our selected neutral microstate highlighted in yellow) and all of the other enumerated microstates (SM43_micro001–SM43_micro005) for a total of 5 relative state free energies (ΔG_{BA}° , ΔG_{CA}° , ΔG_{DA}° , ΔG_{EA}° , ΔG_{FA}°). Some transitions involved a change in a single protonation state (e.g. the D-A transition of Figure 3.2) or tautomer (e.g. the C-A transition of Figure 3.2). A few cases involved a change of multiple protons (e.g. the F-A transition of Figure 3.2). All transitions were defined as *away* from the neutral reference state. Distinct microstates are defined as all tautomers of each charge state. For each relative free energy prediction reported, participants also submitted the formal charge after transitioning from the selected neutral state to the other state. For example, the reported charge state after transitioning from SM43_micro000 to SM43_micro001 would be -1, SM43_micro000 to SM43_micro004 would be 0 (these are tautomers of each other), SM43_micro000 to SM43_micro005 would be +1, and SM43_micro000 to SM43_micro003 would be +2.

(SM34_micro002). It is unclear why this state was not identified by the tools we used to enumerate microstates.

We provided participants CSV (.csv) tables which included microstate IDs and their corresponding canonical isomeric SMILES string, as well as individual MOL2 (.mol2) and SDF (.sdf) files for each individual microstate. These are available in the SAMPL7 GitHub repository.

3.3.4 Combining $\log P$ and pK_a predictions to estimate $\log D$

In the SAMPL7 challenge, $\log P$ and pK_a predictions were combined in order to estimate $\log D$. The relationship between partition and distribution coefficients at a given pH can be computed via [211, 208]

$$\log D_{\text{pH}} = \log P - \log (1 + 10^{\text{p}K_a - \text{pH}}) \quad (3.15)$$

for bases (if no deprotonation site is present or if $\text{p}K_b < \text{p}K_a$) and

$$\log D_{\text{pH}} = \log P - \log (1 + 10^{\text{pH} - \text{p}K_a}) \quad (3.16)$$

for acidic compounds. The $\log D$ was calculated under the assumption that the ionic species cannot partition into the organic phase [19], which may be important in some cases (e.g. in compounds with high lipophilicity or in cases where pH is so extreme that partitioning of a charged species might become important).

3.3.5 Evaluation approach

We considered a variety of error metrics when analyzing predictions submitted to the SAMPL7 physical property set of challenges. We report the following 6 error metrics: the root-mean-squared error (RMSE), mean absolute error (MAE), mean (signed) error (ME), coefficient of determination (R^2), linear regression slope (m), and Kendall’s Tau rank correlation coefficient (τ). Additionally, 95% confidence intervals were computed for these values using a bootstrapping-over-molecules procedure (with 10,000 bootstrap samples), as in prior SAMPL challenges [155].

Accuracy based performance metrics, such as RMSE and MAE, are more appropriate than correlation-based statistics to evaluate methods because of the small dynamic range of experimental $\log P$ values (0.6-3.0). This is usually reflected in the confidence intervals on these metrics. Calculated error statistics of all methods can be found in Tables 3.6, 3.8, and 3.9. Summary statistics were calculated for each submission for method comparison. Details of the analysis and scripts are preserved on the SAMPL7 GitHub repository (described in the “Code and data availability” section).

For each challenge we included a reference and/or null method set of predictions in the analysis to provide perspective for performance evaluations of blind predictions. Null models or null predictions employ a model that is not expected to be useful and can provide a simple point of comparison for more sophisticated methods, as ideally, such methods should improve on predictions from a null model. Reference methods are not formally part of the challenge, but are provided as comparison methods. For the $\log P$ challenge we included a null prediction set which predicts a constant $\log P$ value of 2.66 for every compound, as described in a previous SAMPL paper [87]. For $\log D$ evaluation we included a set of null predictions that all of the molecules partition equally between the water and octanol phase.

For the $\log P$ and pK_a challenge and the $\log D$ evaluation, we provide reference calculations

using ChemAxon’s Chemicalize [2], a commercially available empirical toolkit, as a point of comparison. These include *REF#* in the method name in all of the Figures so that they are easily recognized as non-blind reference calculations. The analysis is presented with and without the inclusion of reference and/or null calculations in the SAMPL7 GitHub repository. The Figures and statistics tables pertaining to the log P and pK_a challenges and the log D evaluation in this manuscript include reference calculations.

For the log P and pK_a challenge, we list consistently well-performing methods that were ranked in the top consistently according to two error and two correlation metrics: RMSE, MAE, R^2 , and Kendall’s Tau. These are shown in Table 3.2 and 3.4.

For each challenge, we also evaluated the relative difficulty of predicting the physical property of interest of each molecule in the set. We plotted the distributions of errors in prediction for each molecule considering all prediction methods. We also calculated the MAE for each molecule as an average of all methods, as well as for predictions from each method category.

Converting relative free energies between microstates to macroscopic pK_a

In the pK_a challenge, participants submitted predictions consisting of the free energy changes between a reference microstate and every other relevant microstate for each compound. Specifically, participants were asked to predict the relative free energy between a selected neutral reference microstate and the rest of the enumerated microstates for that molecule at a reference pH of 0. In order to compare participants’ predictions to experimental pK_a values, these predicted relative free energies had to be converted to macroscopic pK_a values.

Here, we analyzed submissions using the titration method discussed above (Section 3.2.3). This approach computes the population of each charge state as a function of pH and finds the pH at which the population of one charge state crosses that of another (Figure 3.3); as noted above this approach is equivalent to the transition and free energy approaches detailed

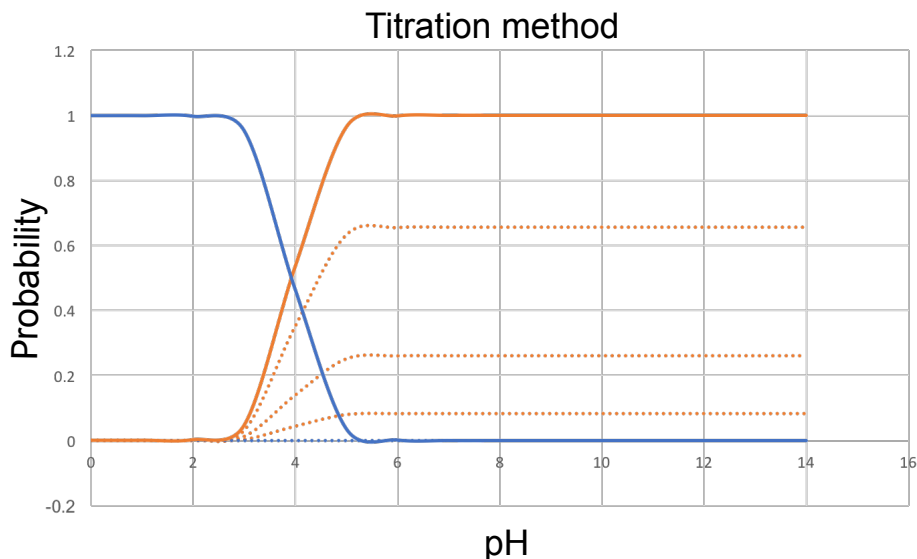


Figure 3.3: **Using the microstate probability to convert microscopic pK_a predictions to macroscopic pK_a 's with the titration method pK_a 's.** Blue and orange lines represent two states. Blue states have one more proton than the orange states, and thus a formal charge higher by +1. The blue state has one tautomer and the orange state has 3, denoted by the dashed lines. The solid lines are the ensemble averaged state probability for each group with a given charge. The crossing point between two ensemble lines is the macroscopic pK_a .

previously.

In our analysis Python code used in the present challenge we work from Equation 3.6 and Equation 3.7 to find the pH at which populations of the two charge states are equal. Here, we do this using `fsolve` from `scipy` in Python.

3.4 Results and Discussion

3.4.1 Overview of log P challenge results

A variety of methods were used in the log P challenge. There were 33 blind submissions collected from 17 groups (Tables of participants and their predictions can be found in the SAMPL7 GitHub Repository and in the Supporting Information.). In the SAMPL6 octanol-

water $\log P$ challenge there were 91 blind submissions collected from 27 participating groups. In the SAMPL5 Cyclohexane-Water $\log D$ challenge, there were 76 submissions from 18 participating groups [19], so participation was lower than previous iterations. This modestly decreased participation (by one group) was likely in part because of COVID-19-related disruptions and because this challenge had to be conducted on a short timescale with relatively limited publicity because the experimental data was not generated specifically for SAMPL, and thus staging of the SAMPL7 challenge required delaying submission of an experimental study which was already complete.

Out of blind submissions of the SAMPL7 $\log P$ challenge, there were 10 in the physical (MM) category, 10 in the physical (QM) category, and 12 in the empirical category. An additional null and reference method were included in the empirical method category.

The following sections evaluate the performance of $\log P$ prediction methods. Performance statistics of all the methods can be found in Table 3.6. Methods are referred to by their method names, which are provided in Table 3.1.

Performance statistics to compare $\log P$ prediction methods

Some methods in the challenge achieved a good octanol–water $\log P$ prediction accuracy. Figure 3.4 shows the performance comparison of methods based on accuracy with RMSE and MAE. The uncertainty in the correlation statistics was too high to rank method performance based on correlation, but we provide an overall correlation assessment for all methods in the SI in Figure 3.19. 16 submissions achieved a $\text{RMSE} \leq 1.0 \log P$ units, but no method achieved a $\text{RMSE} \leq 0.5 \log P$ units. Methods that achieved a $\text{RMSE} \leq 1.0 \log P$ units were mainly empirical, but some were QM-based. Prediction methods include 15 blind predictions and one reference method.

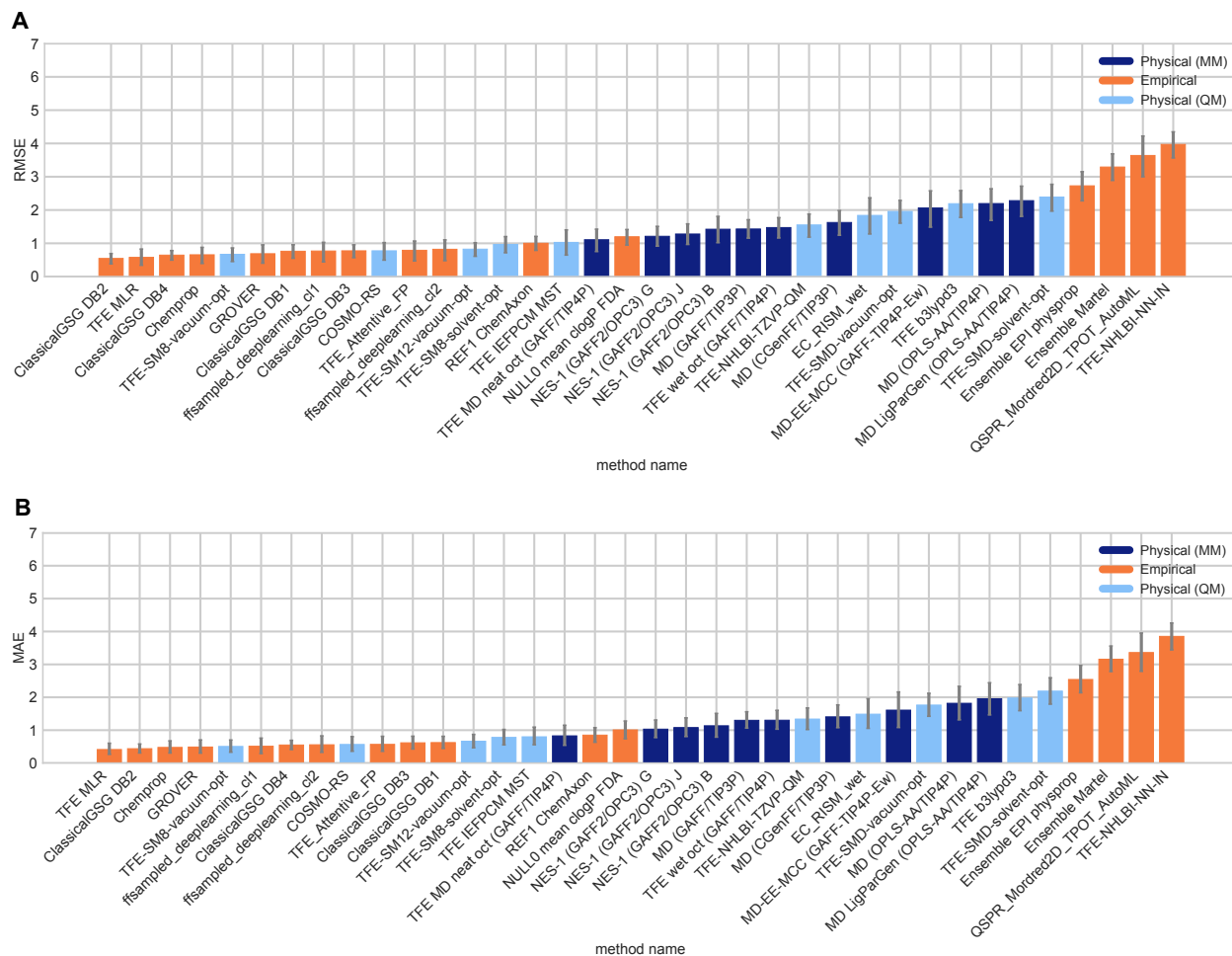


Figure 3.4: Overall accuracy assessment for all methods participating in the SAMPL7 log P challenge shows that many methods did not exhibit statistically significant differences in performance and there was no single clear winner; however, empirical methods tended to perform better in general. Both root-mean-square error (RMSE) and mean absolute error (MAE) are shown, with error bars denoting 95% confidence intervals obtained by bootstrapping over challenge molecules. Empirical methods outperform the majority of the other methods. Methods that achieved a RMSE ≤ 1.0 log P units were mainly empirical based, and some were QM-based physical methods. Submitted methods are listed in Table 3.1. The submission *REF1 ChemAxon* [2] was a reference method included after the blind challenge submission deadline, and *NULL0 mean cLogP FDA* is the null prediction method; all others refer to blind predictions.

Table 3.1: **Method names, category, and submission type for all the log P calculation submissions.** The “submission type” column indicates if submission was a blind submission (denoted by “Blind”) or a post-deadline reference or null calculation (denoted by “Reference”). The table is ordered from lowest to highest RMSE, although many consecutively listed methods are statistically indistinguishable. All calculated error statistics are available in Table 3.6.

Method Name	Category	Submission Type
<i>ClassicalGSG DB2</i> [68, 52, 53]	Empirical	Blind
<i>TFE MLR</i> [174]	Empirical	Blind
<i>ClassicalGSG DB4</i> [68, 52, 53]	Empirical	Blind
<i>Chemprop</i> [118]	Empirical	Blind
<i>TFE-SM8-vacuum-opt</i>	Physical (QM)	Blind
<i>GROVER</i>	Empirical	Blind
<i>ClassicalGSG DB1</i> [68, 52, 53]	Empirical	Blind
<i>ffsampled_deeplearning_cl1</i>	Empirical	Blind
<i>ClassicalGSG DB3</i> [68, 52, 53]	Empirical	Blind
<i>COSMO-RS</i> [224]	Physical (QM)	Blind
<i>TFE_Attentive_FP</i>	Empirical	Blind
<i>ffsampled_deeplearning_cl2</i>	Empirical	Blind
<i>TFE-SM12-vacuum-opt</i>	Physical (QM)	Blind
<i>TFE-SM8-solvent-opt</i>	Physical (QM)	Blind
<i>REF1 ChemAxon</i> [2]	Empirical	Reference
<i>TFE IEFPCM MST</i> [217]	Physical (QM)	Blind
<i>TFE MD neat oct (GAFF/TIP4P)</i>	Physical (MM)	Blind
<i>NULL0 mean clogP FDA</i> [87]	Empirical	Reference
<i>NES-1 (GAFF2/OPC3) G</i>	Physical (MM)	Blind
<i>NES-1 (GAFF2/OPC3) J</i>	Physical (MM)	Blind
<i>NES-1 (GAFF2/OPC3) B</i>	Physical (MM)	Blind
<i>MD (GAFF/TIP3P)</i> [62]	Physical (MM)	Blind
<i>TFE wet oct (GAFF/TIP4P)</i>	Physical (MM)	Blind
<i>MD (CGenFF/TIP3P)</i> [62]	Physical (MM)	Blind
<i>EC_RISM_wet</i> [209]	Physical (QM)	Blind
<i>TFE-SMD-vacuum-opt</i>	Physical (QM)	Blind
<i>MD-EE-MCC (GAFF-TIP4P-Ew)</i> [60]	Physical (MM)	Blind
<i>TFE b3lypd3</i> [66]	Physical (QM)	Blind
<i>MD (OPLS-AA/TIP4P)</i> [62]	Physical (MM)	Blind
<i>MD LigParGen (OPLS-AA/TIP4P)</i> [62]	Physical (MM)	Blind
<i>TFE-SMD-solvent-opt</i>	Physical (QM)	Blind
<i>TFE-NHLBI-TZVP-QM</i>	Physical (QM)	Blind
<i>Ensemble EPI physprop</i>	Empirical	Blind
<i>Ensemble Martel</i>	Empirical	Blind
<i>QSPR_Mordred2D_TPOT_AutoML</i>	Empirical	Blind
<i>TFE-NHLBI-NN-IN</i>	Empirical	Blind

A shortlist of consistently well-performing methods in the log P challenge

Here, many performance differences are not statistically significant, but we identified five consistently well-performing ranked methods that appear in the top 10 according to two accuracy based (RMSE and MAE) and two correlation based metrics (Kendall’s Tau and

R^2), as shown in Table 3.2. The resulting 5 best-performing methods were made up of three empirical methods and two QM-based physical methods.

Method *TFE MLR* [174] was an empirical method that used a multi-linear regression (MLR) made from experimental $\log P$ values from 60 sulfonamides obtained from PubChem [101] and DrugBank [54]. The dataset was mainly composed of sulfonamide drugs and smaller molecules with other classical functional groups. The following descriptors were used to create the MLR: the frequency of functional groups, hydrogen bond acceptors, hydrogen bond donors, molar refractivity, and topological polar surface area. The functional group frequency was calculated with an in-house script from a modified function of Open Babel [169], the rest was obtained from supplied Open Babel properties.

Method *Chemprop* was an empirical method which used the $\log P$ dataset of the OPERA models in their approach [118]. Molecules from the Opera set were compared with the challenge molecules and those with an ECFP_6 fingerprint (extended connectivity fingerprint) tanimoto coefficient (TC) greater than 0.25 were flagged as test molecules for a total of 233 testing molecules. The training set was created from the rest of the Opera data set by filtering out molecules with a ECFP_6 TC >0.4 to test set molecules. Several models were built using a Directed-Message Passing Neural Network (D-MPNN) [37, 228] to predict the $\log P$, which was then used to get the transfer free energy.

Submission *ClassicalGSG DB3* is an empirical method that employed neural networks (NNs) where the inputs are molecular features generated using a method called Geometric Scattering for Graphs (GSG) [68, 52, 53]. In GSG, atomic features are transformed into molecular features using the graph molecular structure. For atomic features, predictions used 4 physical quantities from classical molecular dynamics forcefields: partial charge, Lennard-Jones well depth, Lennard-Jones radius and atomic type. A training dataset was built from 7 datasets for a total of 44,595 unique molecules. Open Babel was used to convert RDKit generated canonical SMILES to MOL2 files, which were then used as input into CGenFF to

determine partial charges and Lennard-Jones parameters for all atoms in each molecule. The generation of CGenFF atomic attributes failed for some molecules, so the final dataset had 41,409 molecules, and is referred to as the “full dataset”. A training set of 2,379 molecules was obtained by filtering the full training set and keeping only those with sulfonyl functional groups. This was done using the `HasSubstructMatch` function of the RDKit toolkit. The $\log P$ values were predicted by the model trained on this training set.

Method *COSMO-RS* was a QM-based physical prediction approach [224]. First, this approach used COSMOquick [42] to generate tautomers and discarded irrelevant states due to an internal energy threshold implemented in COSMOquick. The participants conducted a conformational search of every microstate with COSMOconf [41] using up to 150 conformers. Second, for each conformer they performed a geometry optimization using the BP86 functional with a TZVP basis set and the COSMO solvation scheme, followed by a single point energy calculation using the BP86 functional with a def2-TZVPD basis set and the FINE COSMO cavity. All density functional theory calculations were carried out with the TURBOMOLE 7.5 program package [17, 213]. Third, a conformer selection was done by applying COSMOconf (using internally COSMOtherm) to reduce the number of conformers and tautomers for the neutral molecule sets. The final set of the neutral state contained only those conformers and states that are relevant in liquid solutions. Fourth, the COSMOtherm software (version 2020) [43] was used to calculate the free energy difference for each molecule set (from the second step described here) and to calculate the relative weight of the microstates in water. All free energy calculations were carried out using the BP-TZVPD-FINE 20 level of COSMO-RS in COSMOtherm. Within the used COSMO-RS, an ensemble of conformers and microstates is automatically used and weighted according to the total free energy in the respective liquid phase, i.e. different weights are used in water and octanol.

Submission *TFE-NHLBI-TZVP-QM* was a QM-based physical method that used the Def2-TZVP basis set for all calculations. Calculations were performed in either Gaussian 09

or Gaussian 16. Structures were optimized with the B3LYP density functional and were verified to be local minima via frequency calculations on an integration grid with harmonic frequencies. Details of solvation handling were not included in the method description.

Figure 3.5 show predicted $\log P$ vs experimental $\log P$ value comparison plots of these 5 well-performing methods and also a method that represents average performance in this challenge. Representative method *NES-1 (GAFF2/OPC3) G* was selected because it has the median RMSE of all ranked methods analyzed in the challenge.

Table 3.2: Five consistently well-performing $\log P$ prediction methods based on consistent ranking within the top 10 according to various statistical metrics. Submissions were ranked according to RMSE, MAE, R^2 , and Kendall’s Tau. Many top methods were found to be statistically indistinguishable when considering the uncertainties of their error metrics. Additionally, the sorting of methods was significantly influenced by the metric that was chosen. We determined which ranked $\log P$ prediction methods were consistently the best according to all four chosen statistical metrics by assessing the top 10 methods according to each metric. A set of five consistently well-performing methods were determined— three empirical methods and two QM-based physical methods. Performance statistics are provided as mean and 95% confidence intervals. Correlation plots of the best performing methods and one average method is shown in Figure 3.5. Additional statistics are available in Table 3.6.

Method Name	Category	RMSE	MAE	R^2	Kendall’s Tau
<i>TFE MLR</i> [174]	Empirical	0.58 [0.34, 0.83]	0.41 [0.26, 0.60]	0.43 [0.06, 0.80]	0.56 [0.23, 0.83]
<i>Chemprop</i> [118]	Empirical	0.66 [0.39, 0.89]	0.48 [0.30, 0.69]	0.41 [0.11, 0.76]	0.54 [0.25, 0.82]
<i>ClassicalGSG DB3</i> [68, 52, 53]	Empirical	0.77 [0.57, 0.96]	0.62 [0.43, 0.82]	0.51 [0.18, 0.77]	0.48 [0.14, 0.75]
<i>COSMO-RS</i> [224]	Physical (QM)	0.78 [0.49, 1.01]	0.57 [0.36, 0.80]	0.49 [0.17, 0.80]	0.53 [0.25, 0.78]
<i>TFE-NHLBI-TZVP-QM</i>	Physical (QM)	1.55 [1.19, 1.87]	1.34 [1.02, 1.76]	0.52 [0.19, 0.78]	0.51 [0.19, 0.78]

Difficult chemical properties for $\log P$ predictions

To learn about chemical properties that are challenging for $\log P$ predictions, we analyzed the prediction errors of the molecules (Figure 3.6). We chose to use MAE for this analysis because it is less affected by outliers compared to RMSE and is therefore more appropriate for following global trends. Although methods varied in performance, as indicated by large and overlapping confidence intervals, the MAE calculated for each molecule as an average

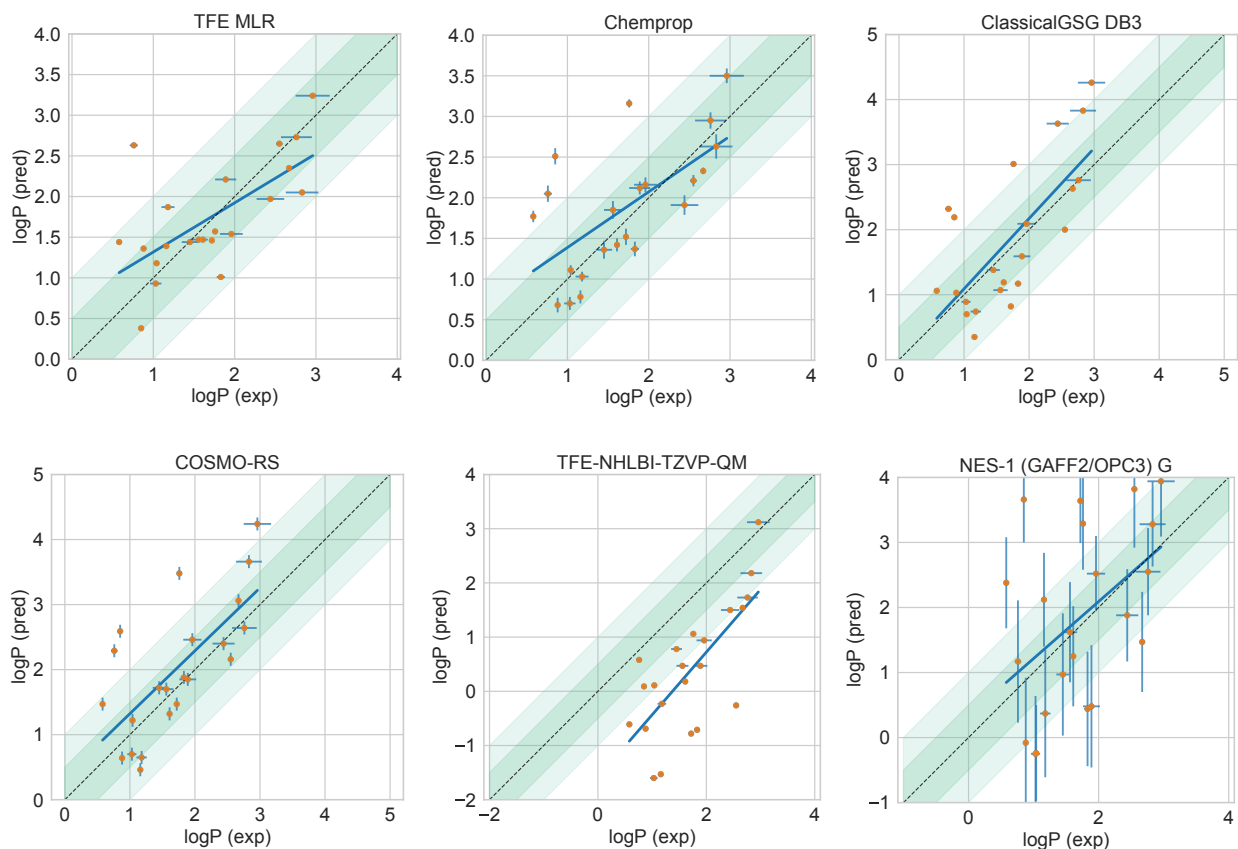


Figure 3.5: **Predicted vs. experimental value correlation plots of 5 best performing methods and one representative average method in the SAMPL7 log P challenge.** Dark and light green shaded areas indicate 0.5 and 1.0 units of error. Error bars indicate standard error of the mean of predicted and experimental values. In some cases, log P SEM values are too small to be seen under the data points. The best-performing methods were made up of three empirical methods (*ClassicalGSG DB3* [52], *TFE MLR* [174], *Chemprop* [118]) and two QM-based physical methods (*COSMO-RS* [224], *TFE-NHLBI-TZVP-QM*). Details of the methods can be found in Section 3.4.1 and performance statistics are available in 3.2. Method *NES-1 (GAFF2/OPC3) G* was selected as the representative average method, which has a median RMSE.

across all methods indicates that some of the molecules were better predicted than others (Figure 3.6A). For reference, compound classes and structures of the molecules are available in Figure 3.20. Molecules such as SM26, SM27, and SM28 were well predicted on average. Molecules such as SM42, SM43, and SM36 were not well predicted on average.

Certain groups of molecules seem to be more challenging for log P predictions. Two of the most poorly predicted molecules, SM42 and SM43, are isoxazoles. Isoxazoles are oxygen and nitrogen-containing heteroaromatics. When we consider the calculated MAE of each molecule separated out by method category, we find that predictions for 2 out of the 3 molecules (SM41 and SM43) belonging to the isoxazole compound class are less accurate with MM-based physical methods than with QM-based physical and empirical method categories (Figure 3.6B).

Figure 3.6C shows error distribution for each challenge molecule over all prediction methods. Molecules such as SM33, SM36, SM41, SM42, and SM43 are shifted to the right, indicating that methods likely had a tendency to overestimate how much these molecules favored the octanol phase.

3.4.2 Overview of pK_a challenge results

In the SAMPL7 pK_a challenge there were 9 blind submissions from 7 different groups. Blind submissions included 7 QM-based physical methods, 1 QM+LEC method, and 1 QSPR/ML method. An additional reference prediction method was included in the QSPR/ML method category.

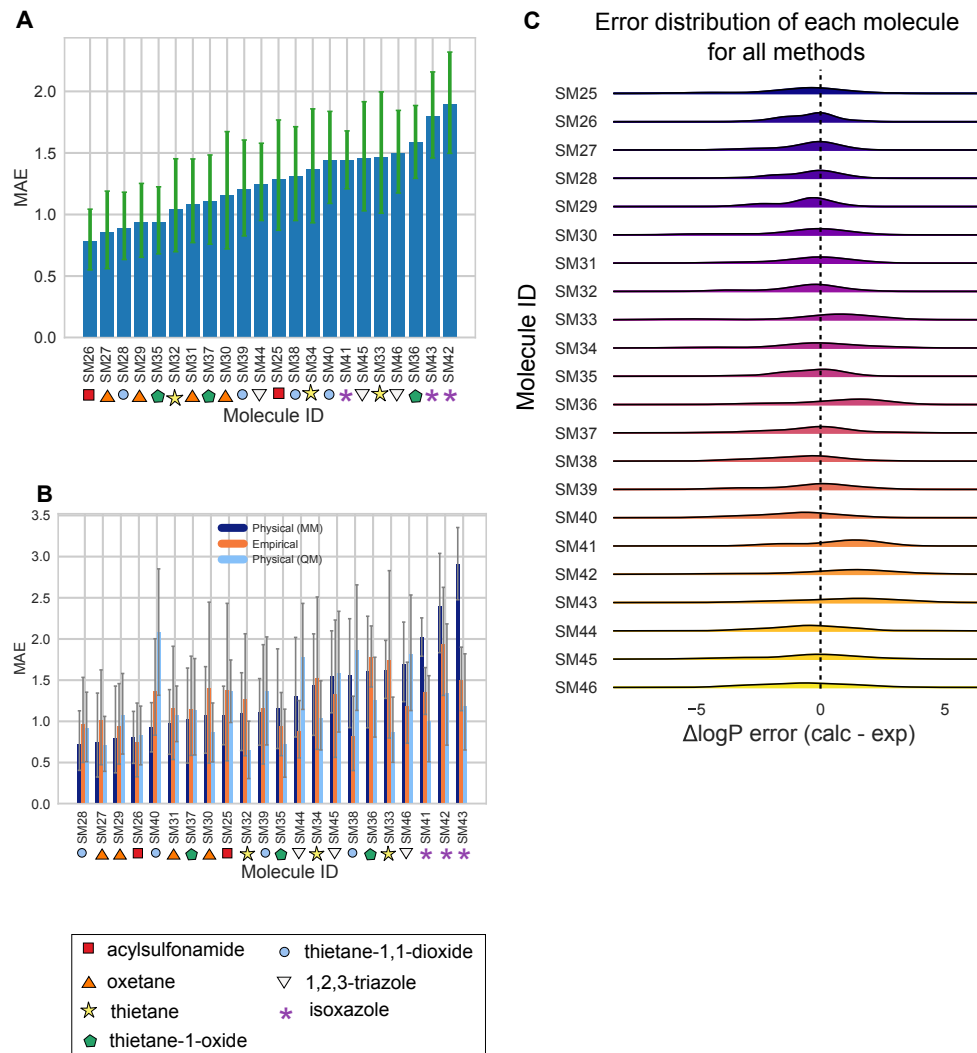


Figure 3.6: Molecule-wise prediction accuracy in the log P challenge point to isoxazoles as poorly predicted, especially by MM-based physical methods. Molecules are labeled with their compound class as a reference. (A) The MAE calculated for each molecule as an average of all methods. (B) The MAE of each molecule separated by method category. (C) log P prediction error distribution for each molecule across all prediction methods.

Table 3.3: **Method names, category, and submission type for all the pK_a submissions.** The “submission type” column indicates if submission was a blind submission (denoted by “Blind”) or a post-deadline reference calculation (denoted by “Reference”). The table is ordered from lowest to highest RMSE, although many consecutively listed methods are statistically indistinguishable. All calculated error statistics are available in Table 3.8.

Method Name	Category	Submission Type
<i>REF00_Chemaxon_Chemicalize</i> [2]	QSPR/ML	Reference
<i>EC_RISM</i> [209]	QM	Blind
<i>IEFPCM/MST</i> [217]	QM	Blind
<i>DFT_M05-2X_SMD</i> [66]	QM	Blind
<i>TZVP-QM</i>	QM	Blind
<i>Standard Gaussian Process</i>	QSPR/ML	Blind
<i>DFT_M06-2X_SMD_implicit</i>	QM	Blind
<i>DFT_M06-2X_SMD_implicit_SAS</i>	QM	Blind
<i>DFT_M06-2X_SMD_explicit_water</i>	QM	Blind
<i>Gaussian_corrected</i>	QM+LEC	Blind

pK_a performance statistics for method comparison

Some methods in the SAMPL7 challenge achieved a good prediction accuracy for pK_a 's. Figure 3.7 shows the performance comparison of methods based on accuracy with RMSE and MAE. Two submissions achieved a $RMSE < 1.0$ pK_a units, no methods achieved a $RMSE \leq 0.5$ pK_a units. One of the methods that achieved a $RMSE < 1.0$ pK_a units was a QM-based physical prediction method (*EC_RISM* [209]), and the other was a QSPR/ML method that was submitted as a reference method (*REF00_Chemaxon_Chemicalize* [2]).

Correlation-based statistics methods provide a rough comparison of methods. Figure 3.8 shows R^2 and Kendall's Tau values calculated for each method, sorted from high to low performance. It is not possible to truly rank these methods based on correlation due to the high uncertainty of each correlation statistic. Over half of the methods have R^2 and Kendall's Tau values equal to or greater than 0.5 and can be considered as the better half, however individual performance is largely indistinguishable from one another. For R^2 , two methods (*EC_RISM*, *REF00_Chemaxon_Chemicalize*), seem to have a greater ranking ability than

the other methods.

There were six methods with an $R^2 \geq 0.5$ — four of the methods were QM methods, one was a QM+LEC method, and one was a QSPR/ML method. Seven methods had a Kendall’s Tau ≥ 0.50 . Of these, five were QM methods, one was a QM+LEC method, and one was a QSPR/ML method.

A shortlist of consistently well-performing methods in the pK_a challenge

We determined a group of consistently well-performing methods in the pK_a challenge. When looking at individual error metrics, many submissions are not different from one another in a way that is statistically significant. Ranking among methods changes based on the chosen statistical metric and does not necessarily lead to strong conclusions due to confidence intervals that often overlap with one another. Here, we determined consistently well-performing methods according to two accuracy (RMSE and MAE) and two correlation metrics (Kendall’s Tau and R^2). For ranked submissions, we identified two consistently well-performing methods that were ranked in the top three according to these statistical metrics. The list of consistently well-performing methods are presented in Table 3.4. The resulting two best-performing methods were both QM-based physical methods.

Submission *EC_RISM* was a QM-based physical method [209]. In this approach, multiple geometries were generated for each microstate using the `EmbedMultipleConfs` function of RDKit. These structures were pre-optimized with Amber 12 using GAFF 1.7 parameters and AM1-BCC charges with an ALPB model to represent the dielectric environment of water. Conformations with an energy of more than 20 kcal/mol than the minimum structure of that microstate were discarded and the remaining structures clustered with a structural RMSD of 0.5 Angstrom. The cluster representatives were then optimized using Gaussian 16revC01 with IEF-PCM using default settings for water at the B3LYP/6-311+G(d,p) level of theory.

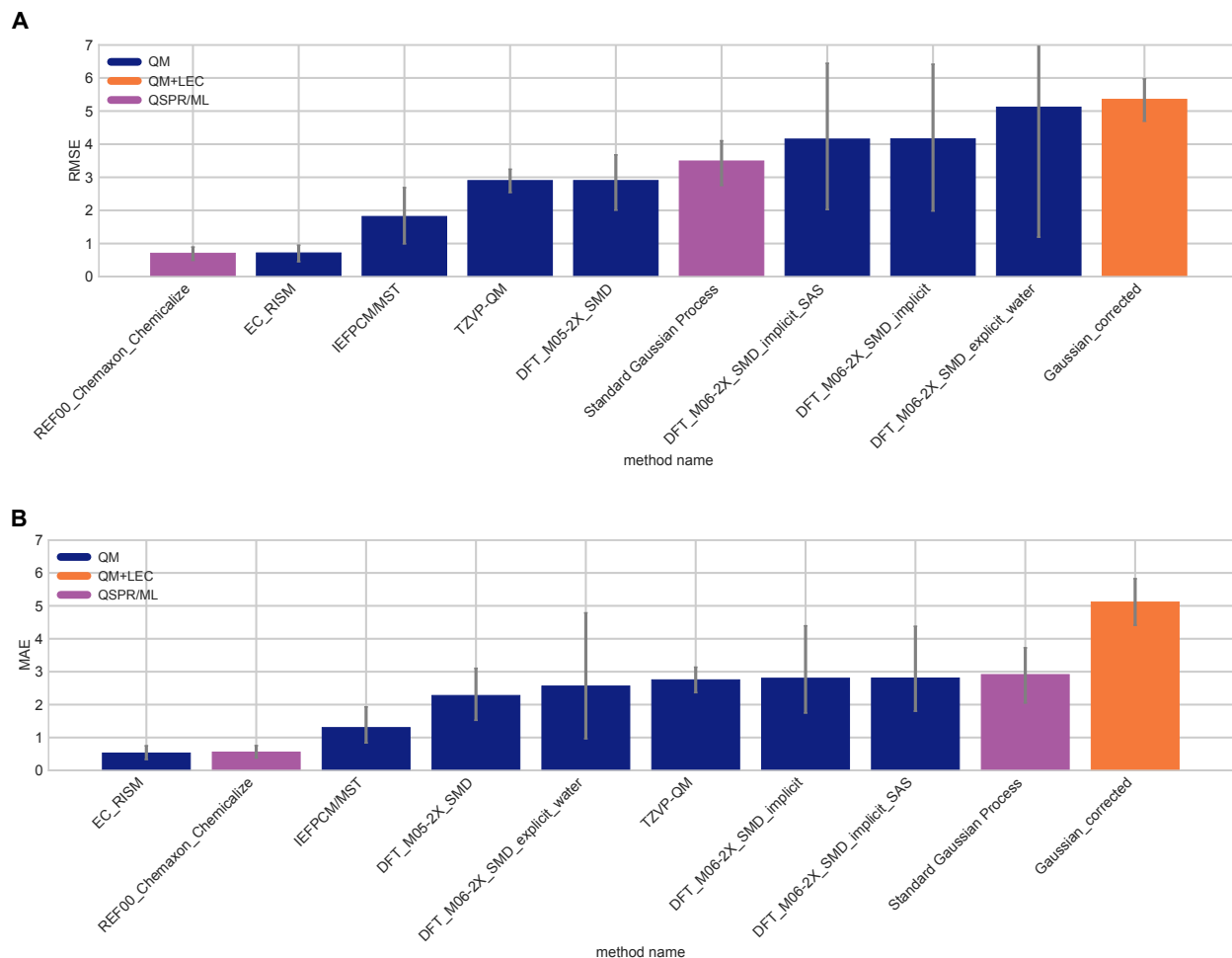


Figure 3.7: Overall accuracy assessment for all methods participating in the SAMPL7 pK_a challenge shows that two methods, one a Physical (QM) method and one a QSPR/ML, performed better than other methods. Both root-mean-square error (RMSE) and mean absolute error (MAE) are shown, with error bars denoting 95% confidence intervals obtained by bootstrapping over challenge molecules. *REF00_Chemaxon_Chemicalize* [2] is a reference method that was included after the blind challenge submission deadline, and all other method names refer to blind predictions. Methods are listed out in Table 3.3 and statistics calculated for all methods are available in Table 3.8.

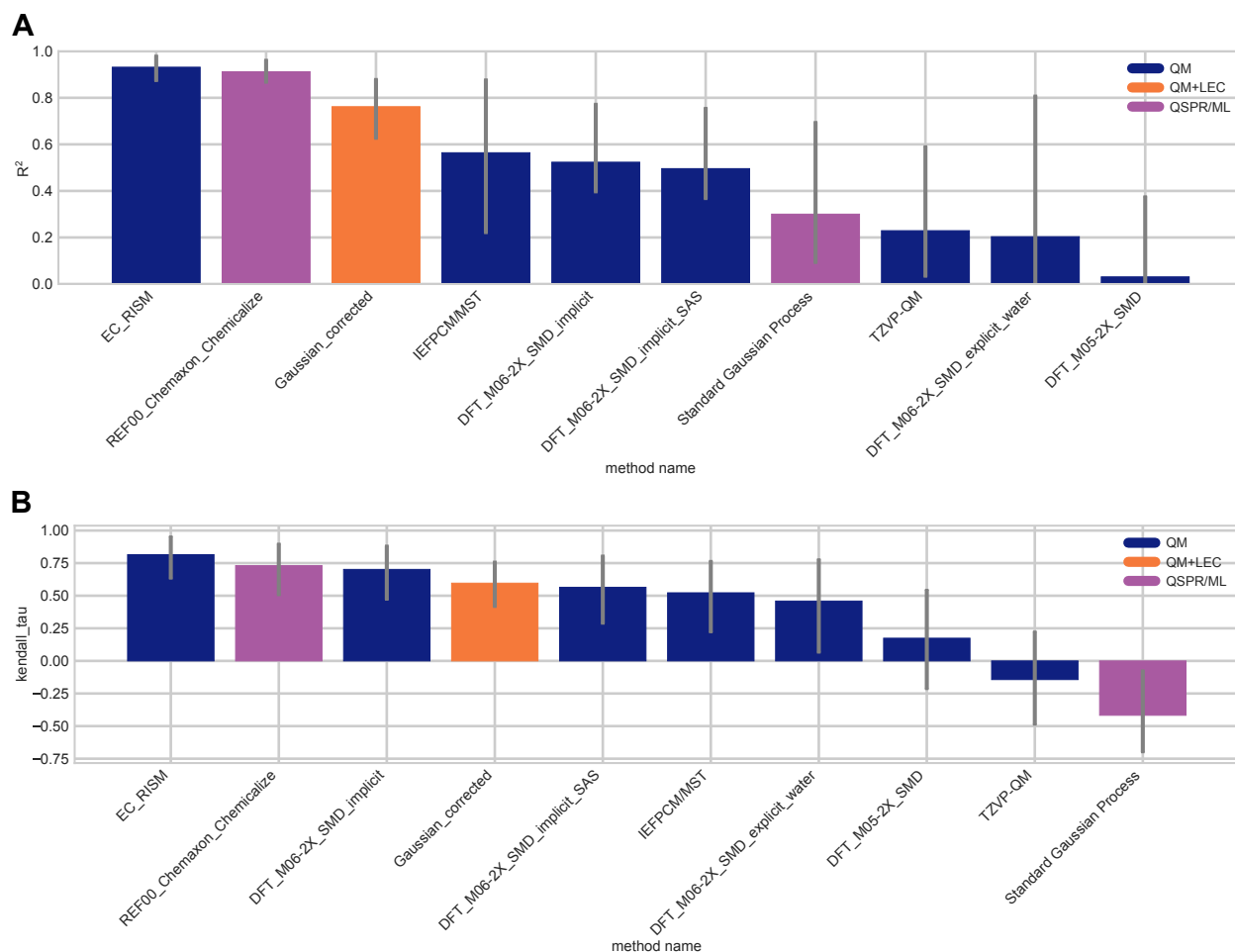


Figure 3.8: **Overall correlation assessment for all methods participating in the SAMPL7 pK_a challenge shows that one Physical (QM) method and one QSPR/ML reference method exhibited modestly better performance than others.** Pearson's R^2 and Kendall's Rank Correlation Coefficient Tau (τ) are shown, with error bars denoting 95% confidence intervals obtained by bootstrapping over challenge molecules. Submission methods are listed out in Table 3.3. *REF00_Chemaxon_Chemicalize* [2] is a reference method that was included after the blind challenge submission deadline, and all other method names refer to blind predictions. Most methods have a statistically indistinguishable performance on ranking, however, for R^2 , two methods (*EC_RISM* [209], *REF_Chemaxon_Chemicalize*), tend to have a greater ranking ability than the other methods. Evaluation statistics calculated for all methods are available in Table 3.8 of the Supplementary Information.

Additional stereoisomers were treated as if they were additional conformational states of the same microstate so that for each microstate only up to 5 conformations with the lowest PCM energies for each solvent were treated with EC-RISM/MP2/6-311+G(d,p) using the PSE2 closure [206] and the resulting EC-RISM energies were corrected. To calculate the relative free energies with respect to each neutral reference state, 4 different formulas were used, depending on the difference in the protonation state. Macrostate pK_a values were calculated using the partition function approach of equation 5 found elsewhere [206].

Submission *IEFPCM/MST* was a QM-based physical method [217]. This approach used the Frog 2.14 software [146, 65] to explore microstate conformations. The molecular geometries of the compounds were fully optimized at the B3LYP/6-31G(d) level of theory, taking into account the solvation effect of water on the geometrical parameters of the solutes, using the IEFPCM version of the MST model. The resulting minima were verified by vibrational frequency analysis, which gave positive frequencies in all cases. The relative energies of the whole set of conformational species were refined from single-point computations performed at the MP2/aug-cc-pVDZ levels of theory. In addition, the gas phase estimate of the free energy difference for all microstates was derived by combining the MP2 energies with zero point energy corrections. Finally, solvation effects were added by using the B3LYP/6-31G(d) version of the IEFPCM/MST model, which is a quantum mechanical self-consistent continuum solvation method. The pK_a was determined using both the experimental hydration free energy of the proton (-270.28 kcal/mol) and a Boltzmann's weighting scheme to the relative stabilities of the conformational species determined for the microstates involved in the equilibrium constant for the dissociation reaction following the thermodynamic cycle reported in previous studies [31].

Figure 3.9 show predicted pK_a vs experimental pK_a value comparison plots of the two well-performing methods and also a method that represents average performance. Representative average method *DFT_M05-2X_SMD* [66] was selected as the method with the median

RMSE of all ranked methods analyzed in the challenge.

Table 3.4: **Two consistently well-performing pK_a prediction methods based on consistent ranking within the top three according to various statistical metrics.** Ranked submissions were ranked/ordered according to RMSE, MAE, R^2 , and Kendall’s Tau. Many methods were found to be statistically indistinguishable when considering the uncertainties of their error metrics. Additionally, the sorting of methods was significantly influenced by the metric that was chosen. We determined which methods are repeatedly among the top two according to all four chosen statistical metrics by assessing the top three methods according to each metric. Two QM-based methods consistently performed better than others. Performance statistics are provided as mean and 95% confidence intervals. All statistics for all methods are in Table 3.8.

Method Name	Category	RMSE	MAE	R^2	Kendall’s Tau
<i>EC_RISM</i> [209]	QM	0.72 [0.45, 0.95]	0.53 [0.33, 0.75]	0.93 [0.87, 0.98]	0.81 [0.63, 0.96]
<i>IEFPCM/MST</i> [217]	QM	1.82 [1.00, 2.69]	1.30 [0.84, 1.92]	0.56 [0.22, 0.87]	0.52 [0.22, 0.76]

Difficult chemical properties for pK_a predictions

To learn about chemical properties that pose challenges for pK_a predictions, we analyzed the prediction errors of the molecules (Figure 3.10). For reference, compound classes and structures of the molecules are available in Figure 3.20. We chose to use MAE for molecular analysis because it is less affected by outliers compared to RMSE and is, therefore, more appropriate for following global trends. When we consider the calculated MAE of each molecule separated out by method category the prediction accuracy of each molecule varies based on method category (Figure 3.10A). The MAE calculated for each molecule as an average of all methods shows that SM25 was the most poorly predicted molecule. The QM+LEC method category appears to be less accurate for the majority of the molecules compared to the other method categories. Compared to the other two method categories, QSPR/ML methods performed better for molecules SM41-SM43, which are isoxazoles (oxygen and nitrogen containing heteroaromatics), and molecule SM44-SM46, which are 1,2,3-triazoles (nitrogen containing heteroaromatics). Physical QM methods performed poorly for molecules SM25 and SM26 (acylsulfonamide compound class). Figure 3.10B shows error distribution for each

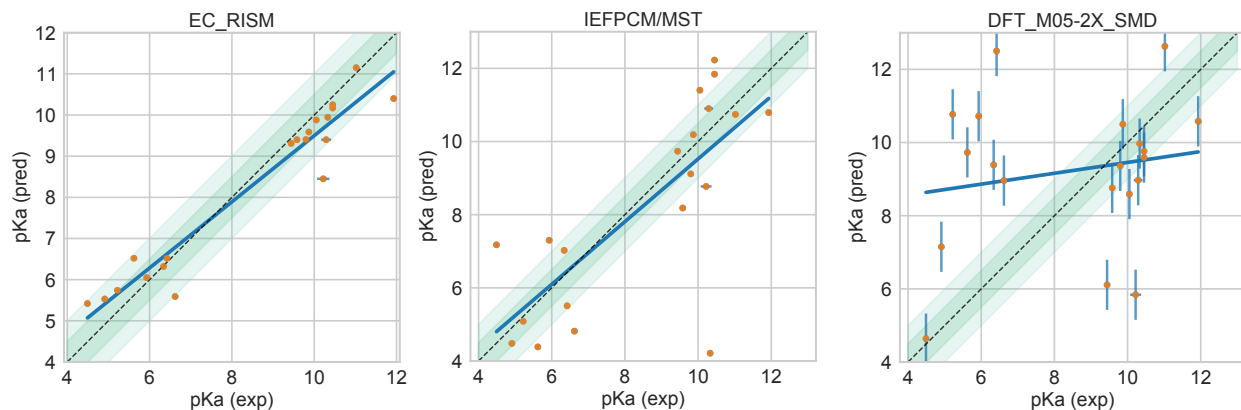


Figure 3.9: **Predicted vs. experimental value correlation plots of 2 best performing methods and one representative average method in the SAMPL7 pK_a challenge.** Dark and light green shaded areas indicate 0.5 and 1.0 units of error. Error bars indicate standard error of the mean of predicted and experimental values. Some SEM values are too small to be seen under the data points. Method *DFT_M05-2X_SMD* [66] was selected as the method with the median RMSE of all ranked methods analyzed in the challenge. Performance statistics of these methods is available in Table 3.4

challenge molecule over all the prediction methods. Molecule SM25 has the most spread in pK_a prediction error.

Microscopic pK_a performance

SAMPL7 challenge pK_a participants were asked to report the relative free energy between microstates, using a provided neutral microstate as reference. Microstates are defined as the enumerated protomers and tautomers of a molecule. Details of how microstates were found can be found in Section 3.3.3. Some molecules had 2 microstates, while others had as many as 6 (Table 3.12).

Figure 3.12 shows the predicted free energy change between the reference state and each microstate, on average, for all transitions across all predictions. Molecules are labeled with their compound class as a reference. Predictions disagree widely for some transitions, like those from the reference state to SM26_micro002, SM28_micro001, SM43_micro003,

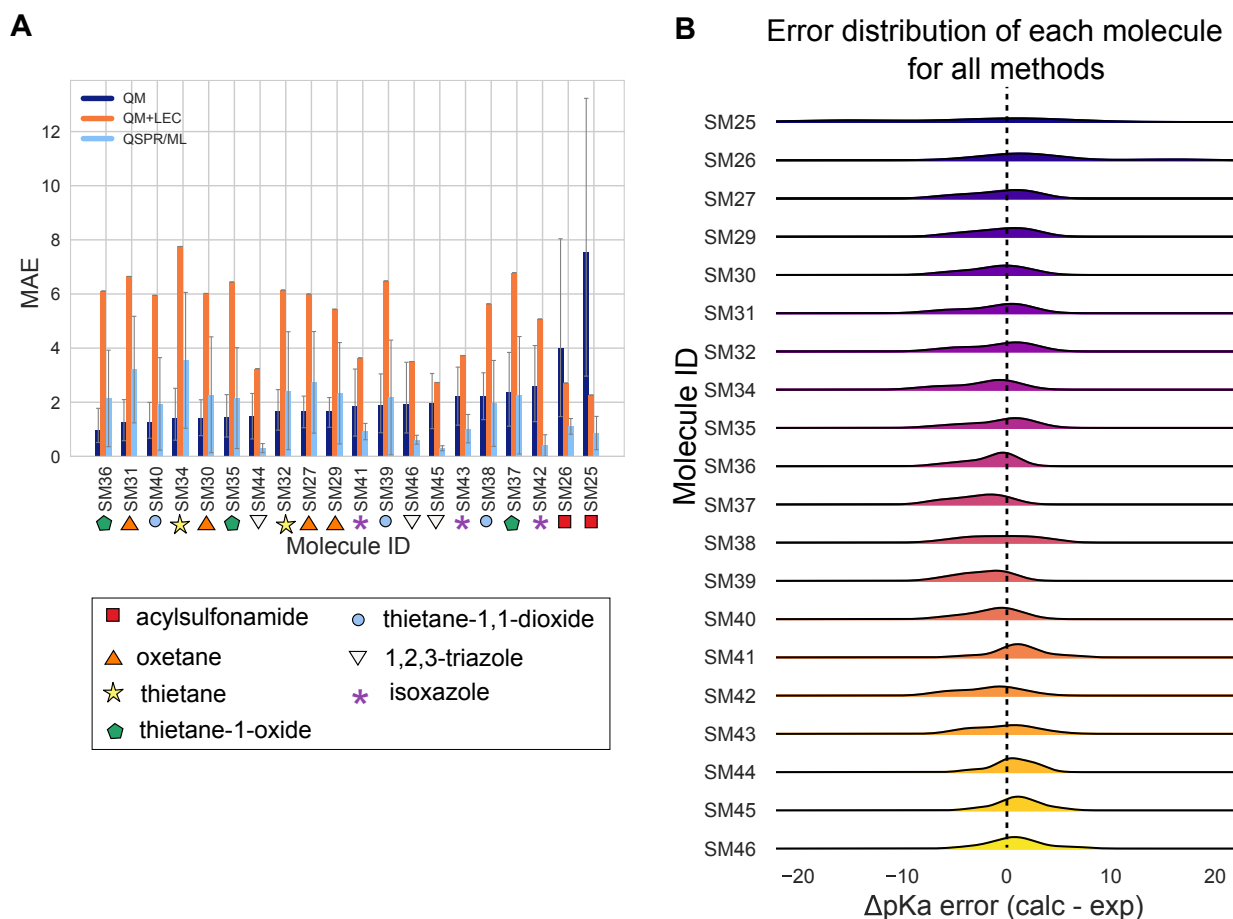


Figure 3.10: Molecule-wise prediction error distribution plots show the prediction accuracy for individual molecules across all prediction methods for the pK_a challenge. Molecules are labeled with their compound class as a reference. (A) The MAE of each molecule separated by method category suggests the most challenging molecules were different for each method category. It is difficult to draw statistically significant conclusions where there are large overlapping confidence intervals. The QM+LEC method category appears to be less accurate for the majority of the molecules compared to the other method categories. QSPR/ML methods performed better for isoxazoles (SM41-SM43) and 1,2,3-triazoles (SM44-SM46) compared to the other two method categories. Physical QM-based methods performed poorly for acylsulfonamides (SM26 and SM25). (B) Error distribution for each molecule over all prediction methods. SM25 has the most spread in pK_a prediction error.

SM46_micro003, while predictions for other transitions such as that from the reference microstate to SM26_micro004 are in agreement (as shown by small error bars in Figure 3.12A, 3.14).

Figure 3.14 shows examples of some microstate transitions where participants' predicted transition free energies disagree. We also examined how the microstate transition free energies (relative to the reference state) are distributed across predictions (Figure 3.12B). We find that some transitions are much more consistently predicted than others, but in some cases there is broad disagreement even about the sign of the free energy change associated with the particular transition – so methods disagree as to which protonation state or tautomer is preferred at the reference pH.

To further analyze which transitions were difficult, we focused on how consistently methods agreed as to the sign of the free energy change for each transition. Particularly, we calculated the Shannon Entropy (H) for the transition *sign* for each transition, shown in Figure 3.13. For each microstate, we calculated H via:

$$H = - \sum_i P_i \ln(P_i) \tag{3.17}$$

where P_i is the probability of a particular outcome i ; here, we use i to indicate a positive sign or a negative sign for the predicted free energy change. So P_{positive} is the fraction of positive sign predictions, P_{negative} is the fraction of negative sign predictions, and P_{neutral} is the fraction of neutral sign predictions (which were somewhat frequent as a few participants predicted a free energy change of exactly 0 for some transitions). For example, for SM25_micro001, given the predictions we received, the P_{positive} is 0.5, the P_{negative} is 0.4 and the P_{neutral} is 0 (no neutral sign predictions). The Shannon entropy H is then $-(0.5 \ln(0.5) + 0.4 \ln(0.4) + 0)$,

which is roughly 0.7 and indicates predictions had difficulty agreeing on the sign.

While the Shannon entropy may not be a perfect tool for analyzing this issue, we find it helpful here. For a particular transition, a value of 0 indicates all predictions agreed as to the sign of the free energy change (whether positive, negative, or neutral), while values greater than 0 reflect an increasing level of disagreement in the sign of the prediction. 32 of the microstates had a H value of 0, 21 had a values that ranged from 0.5-0.7, and 3 microstates had values greater than 0.9 (the highest level of disagreement). The 3 microstates with the most disagreement belong to the thietane-1-oxide compound class (one from SM35, one from SM36 and one from SM37).

Transitions that pose difficulty for participants involve a protonated nitrogen and keto-enol neutral state tautomerism. Chemical transformations involving a protonated nitrogen in terminal nitrogen groups, 1,2,3-triazoles, and isoxazoles were all found to occur in molecules that have high levels of disagreement in sign prediction. Depictions of some of these types of transitions are presented in Figure 3.11. Predictions for these transitions were substantially divided on the predicted sign – roughly half of the methods predict a positive sign, while the other half predict a negative sign. This means methods could not agree on the preferred state at the reference pH. The number of positive, negative, and neutral sign predictions per microstate is available in Table 3.10

In several cases, the SAMPL input files provided a reference microstate with unspecified stereochemistry, then a separate but otherwise equivalent microstate with specified stereochemistry (SM35_micro002, SM36_micro002, SM37_micro003). Experiments were done on the compound with specified stereochemistry, so participants were instructed to assume that the reference microstate (which had unspecified stereochemistry) had the same free energy as the microstate with specified stereochemistry. However, many participants didn't use the microstate with specified stereochemistry as the reference state, and most ended up predicting a nonzero relative free energy between the reference state and the microstate with

Table 3.5: **Method names, category, and submission type for all the log D estimations.** Method names are based off the submitted pK_a and log P method names, with the log P method name listed first followed by “+” and then the pK_a method name. The “submission type” column indicates if submission was a blind submission (denoted by “Blind”) or a post-deadline reference calculation (denoted by “Reference”). All calculated error statistics are available in Table 3.9.

Method Name	Category	Submission Type
<i>REF0 ChemAxon</i>	Empirical	Reference
<i>TFE IEFPCM MST + IEFPCM/MST</i>	Physical (QM)	Standard
<i>NULL0</i>	Empirical	Reference
<i>EC_RISM_wet + EC_RISM</i>	Physical (QM)	Standard
<i>TFE-NHLB1-TZVP-QM + TZVP-QM</i>	Physical (QM)	Standard
<i>TFE b3lypd3 + DFT_M05-2X_SMD</i>	Physical (QM)	Standard
<i>MD (CGenFF/TIP3P) + Gaussian_corrected</i>	Physical (MM) + QM+LEC	Standard
<i>TFE-SMD-solvent-opt + DFT_M06-2X_SMD_explicit_water</i>	Physical (QM)	Standard

specified stereochemistry, despite instructions.

3.4.3 Overview of log D challenge results

In the SAMPL7 physical property prediction challenge, log P and pK_a predictions were combined in order to estimate log D , as described in Section 3.3.4.

There were 6 log D estimates and 2 reference methods. Methods are listed in Table 3.5 and statistics for all log D prediction methods are available in Table 3.9. There were 5 methods that belonged to the physical (QM) category, and 1 in the Physical (MM) + QM+LEC category (this category used a MM-based physical method in the log P challenge, and a QM+LEC method in the pK_a challenge). The null and reference method were included in the empirical method category.

log D performance statistics for method comparison

Figure 3.15 compares the accuracy of methods based on RMSE and MAE. No method achieved a RMSE ≤ 1.0 log D units, and the overall RMSE ranged from 1.1 to 4.5 log D

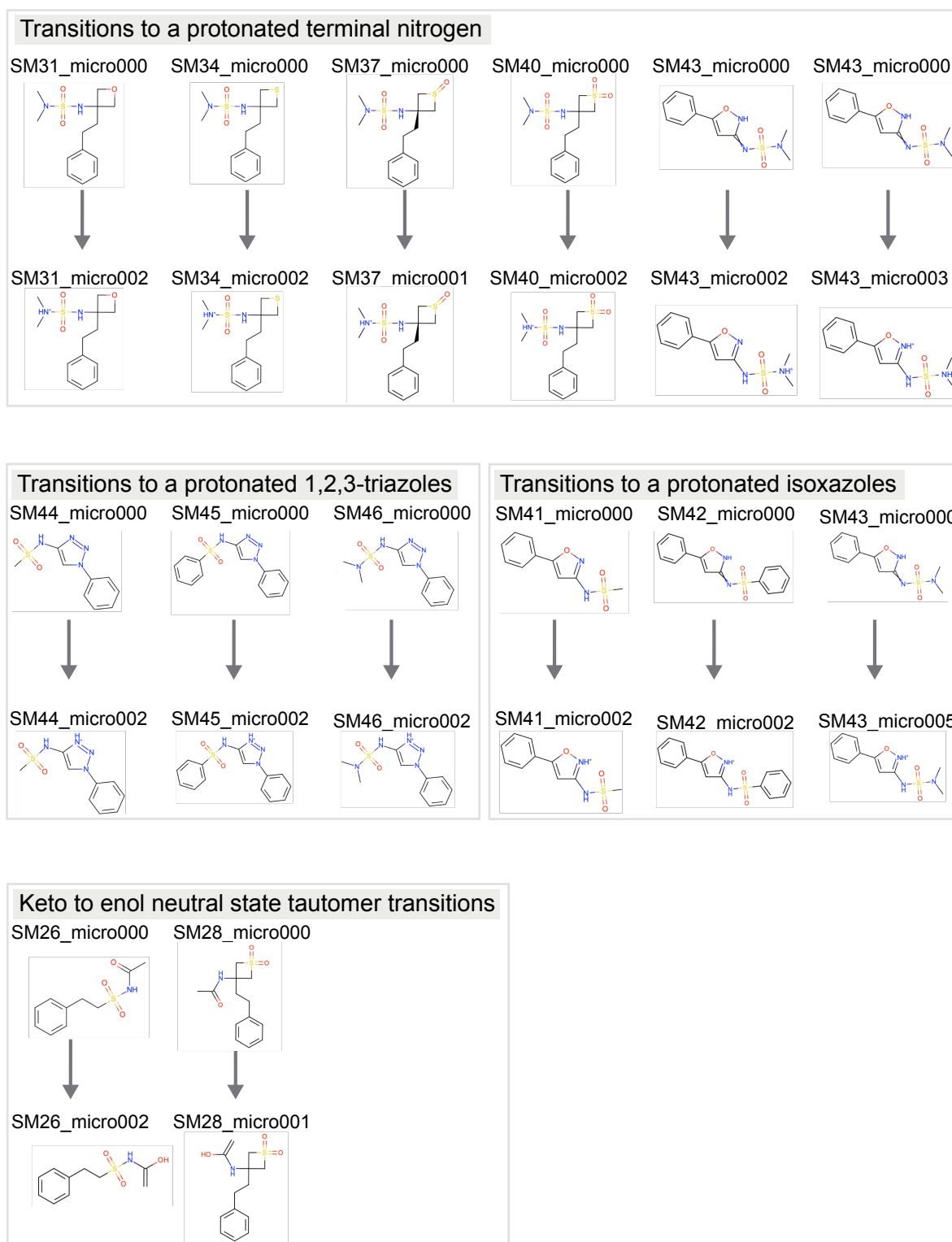


Figure 3.11: **Chemical transformations that lead to common sign disagreements among participants typically involve a protonated nitrogen in terminal nitrogen groups, 1,2,3-triazoles, and isoxazoles.** Shown are some chemical transformations that repeatedly show up as having large disagreement on the sign of the relative free energy prediction, as seen in Figure 3.13.

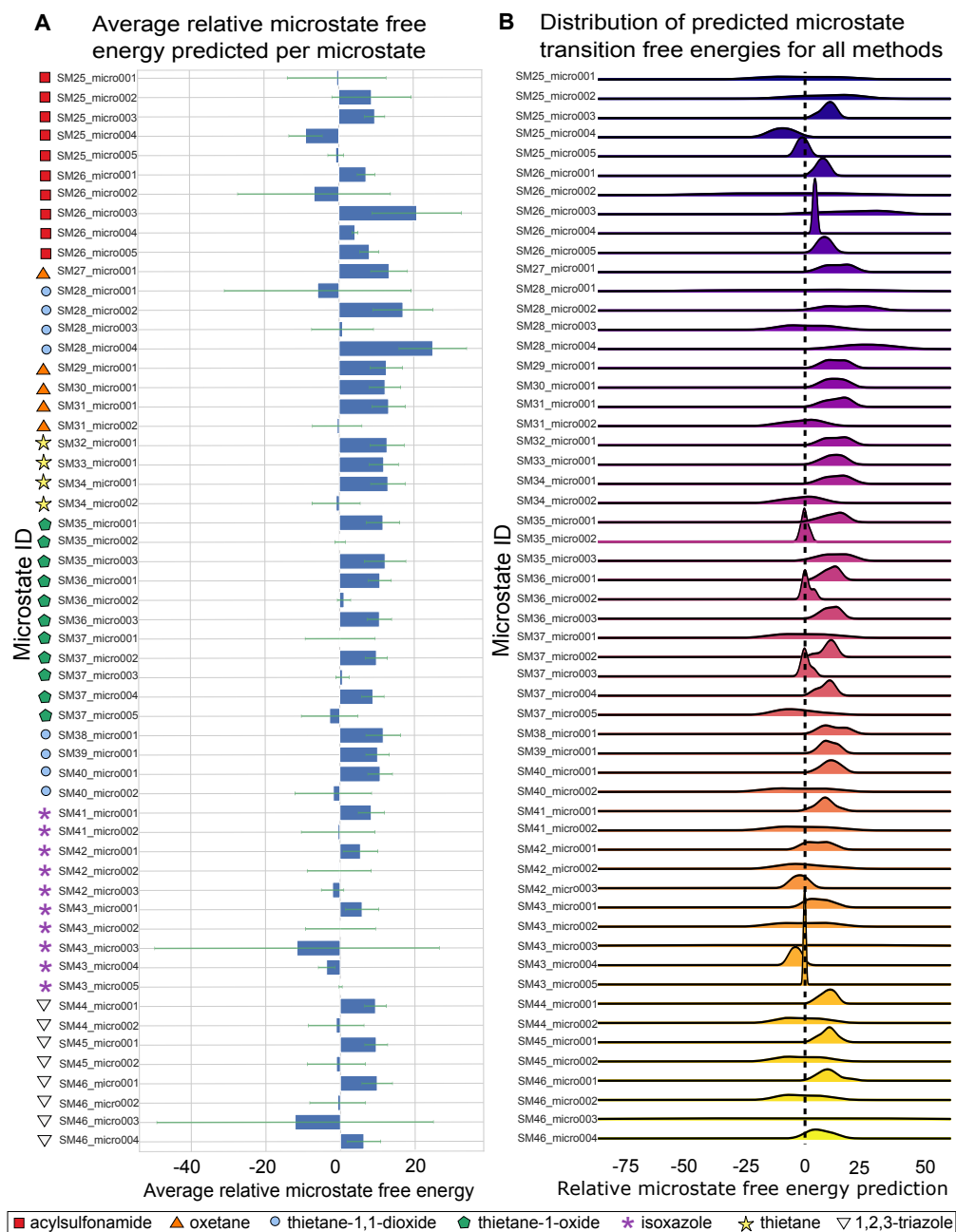


Figure 3.12: **The average relative microstate free energy predicted per microstate and the distribution across predictions in the SAMPL7 pK_a challenge show how varied predictions were.** Molecules are labeled with their compound class as a reference. (A) The average relative microstate free energy predicted per microstate. Error bars are the standard deviation of the relative microstate free energy predictions. A lower standard deviation indicates that predictions for a microstate generally agree, while a larger standard deviation means that predictions disagree. Predictions made for microstates such as SM25_micro001, SM26_micro002, SM28_micro001, SM43_micro003, SM46_micro003 widely disagree, while predictions for microstates such as SM26_micro004 are in agreement. (B) Distribution for each relative microstate free energy prediction over all prediction methods shows how prediction agreement among methods varied depending on the microstate.

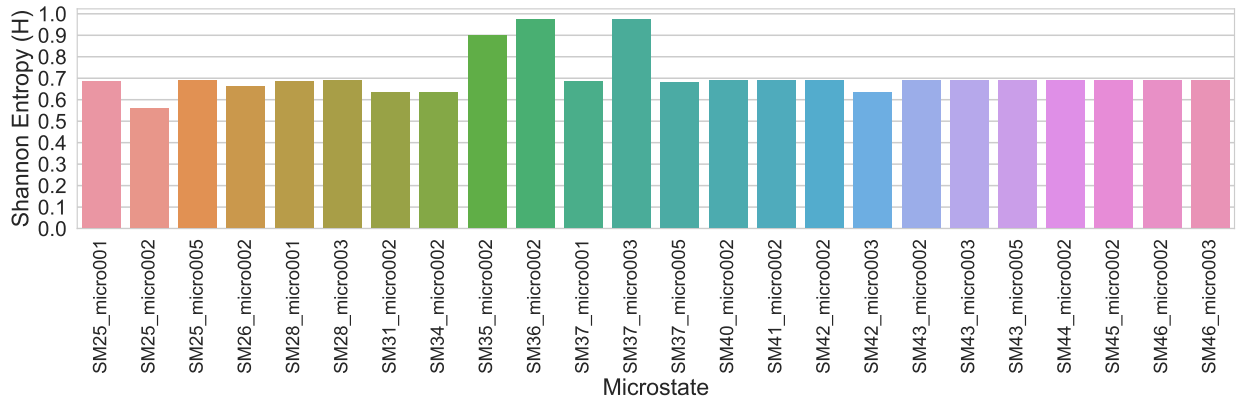


Figure 3.13: **The Shannon entropy (H) per microstate transition shows that participants disagree on many of the signs of the relative free energy predictions.** Microstates with entropy values greater than 0 reflect increasing disagreement in the predicted sign. Microstates with an entropy of 0 are not shown here, but indicate that methods made predictions which had the same sign for the free energy change associated with a particular transition. About 44% of all microstates predictions disagreed with one another based on the sign, and the rest agreed. Roughly 5% of microstates strongly disagreed on the sign of predictions— meaning that predicted relative free energies were fairly evenly split between positive, neutral, and negative values. This indicates that these transitions were particularly challenging.

units. Four methods had a RMSE between 1 and 2, and three methods had an RMSE between 2 and 3. Accuracy is better than the previous $\log D$ challenge. In the SAMPL5 $\log D$ challenge, out of 63 submissions, no submissions had a RMSE below $2 \log D$ units. Here, eight methods were submitted and half of them achieved a RMSE below $2 \log D$ units. Overall, $\log D$ prediction accuracy has improved since SAMPL5.

When the best $\log P$ and pK_a prediction methods are combined we find that the resulting composite approach outperforms most of the other ranked methods, achieving a RMSE of 0.6 (see Figure3.17, method name *TFE MLR + EC_RISM*).

When the experimental $\log P$ and pK_a are combined to yield a $\log D$ (as in Section 3.3.4), the resulting $\log D$ values do not perfectly match with the reported experimental $\log D$ values, an inconsistency that requires further investigation.

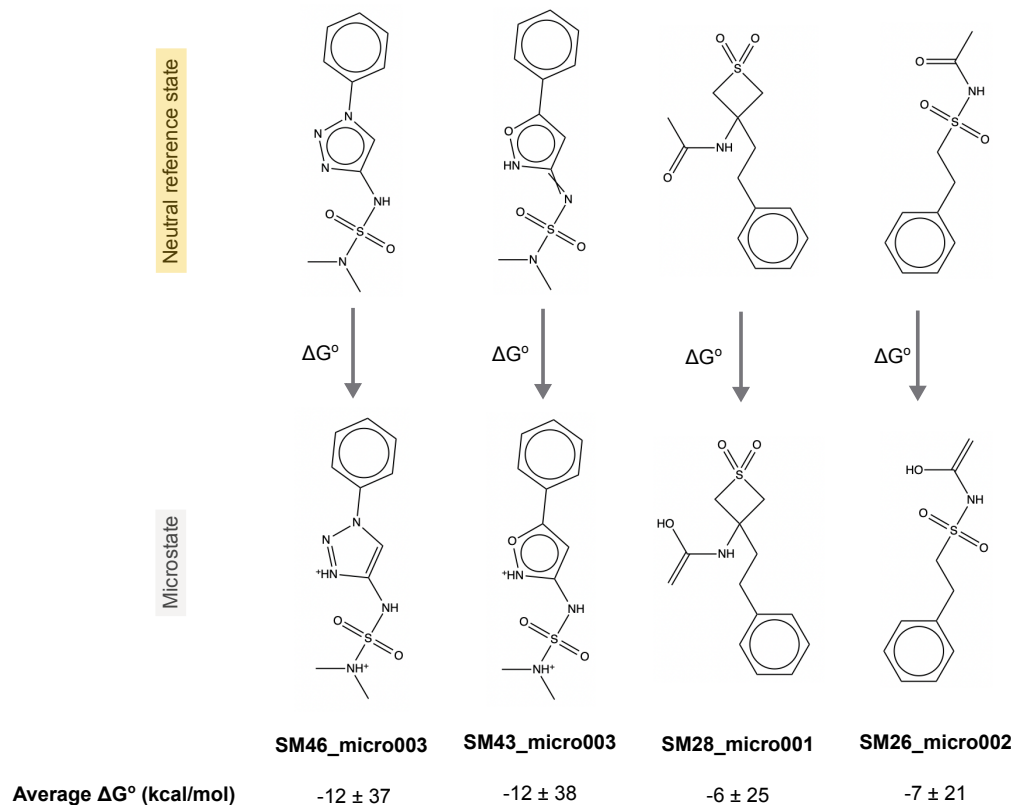


Figure 3.14: **Structures of microstates where relative microstate free energy predictions disagree.** Shown are some of the microstate transitions where participants predictions largely disagree with one another, based on Figure 3.12. The average relative free energy prediction (ΔG) along with the standard deviation are listed under each transition.

A consistently well performing method in log D estimation

For ranked submissions, we identified a single consistently well-performing method that was ranked in the top three according to RMSE, MAE, Kendall’s Tau, and R^2 (all statistics are available in Table 3.9). The best-performing method was *TFE IEFPCM MST + IEFPCM/MST*, which used a QM-based physical method for pK_a and log P predictions [217]. The *IEFPCM/MST* model has previously been used to predict the log D of over 35 ionizable drugs, where it achieved a RMSE of 1.6 [232], all little worse than a RMSE of 1.3 in SAMPL7. The pK_a prediction protocol used in the challenge is described in Section 3.4.2, where it was ranked among the consistently well performing pK_a methods.

3.5 Conclusions

Here, a community-wide blind prediction challenge was held that focused on partitioning and pK_a for 22 compounds composed of a series of N-acylsulfonamides and related bioisosteres. Participants had the option of submitting predictions for both, or either, challenge.

In the SAMPL7 log P challenge, participants were asked to predict a partition coefficient for each compound between octanol and water and report the result as a transfer free energy. A total of 17 research groups participated, submitting 33 blind submissions total. Many submissions achieved a RMSE around 1.0 or lower for log P predictions, but none were below 0.5 log P units. RMSEs ranged from 0.6 to 4 log P units– 15 methods achieved a RMSE of 1.0 or lower, while a RMSE between 1 and 4 log units was observed for the majority of methods. Many methods achieved an accuracy similar to the null model which had a RMSE of 1.2 and predicted that each compound had a constant log P value of 2.66. A few methods outperformed the null model (4 were empirical and 1 was an QM based method). In general, empirical methods tended to perform better than other methods, which makes sense given

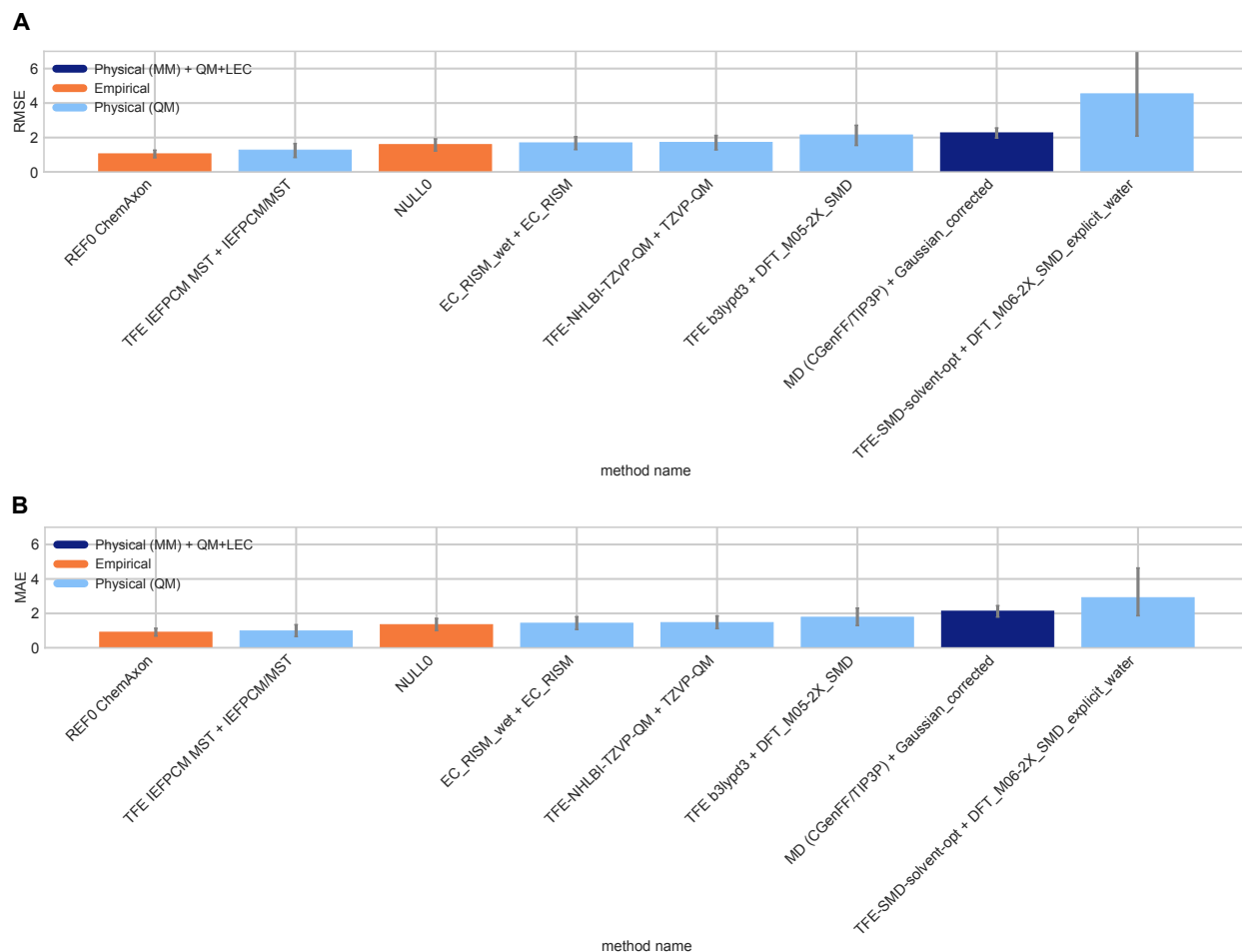


Figure 3.15: **Overall accuracy assessment for $\log D$ estimation.** Both root-mean-square error (RMSE) and mean absolute error (MAE) are shown, with error bars denoting 95% confidence intervals obtained by bootstrapping over challenge molecules. *REF00_ChemAxon* [2] is a reference method and *NULL0* is a null method that was included after the blind challenge submission deadline, and all other method names refer to blind predictions. Methods are listed out in Table 3.5 and statistics calculated for all methods are available in Table 3.9.

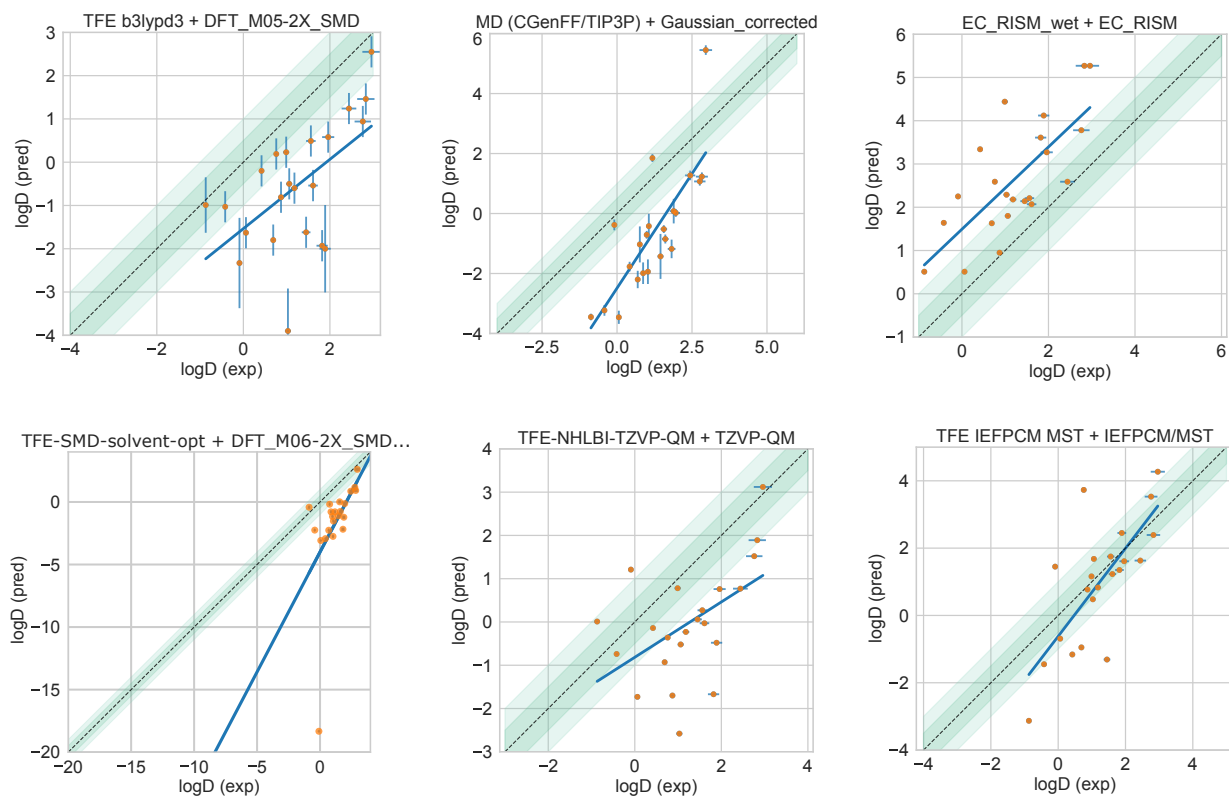


Figure 3.16: **Predicted vs. experimental value correlation plots of all $\log D$ estimation methods in the SAMPL7 challenge.** Dark and light green shaded areas indicate 0.5 and 1.0 units of error. Error bars indicate standard error of the mean of predicted and experimental values. Some SEM values are too small to be seen under the data points. Performance statistics of all methods is available in Table 3.9

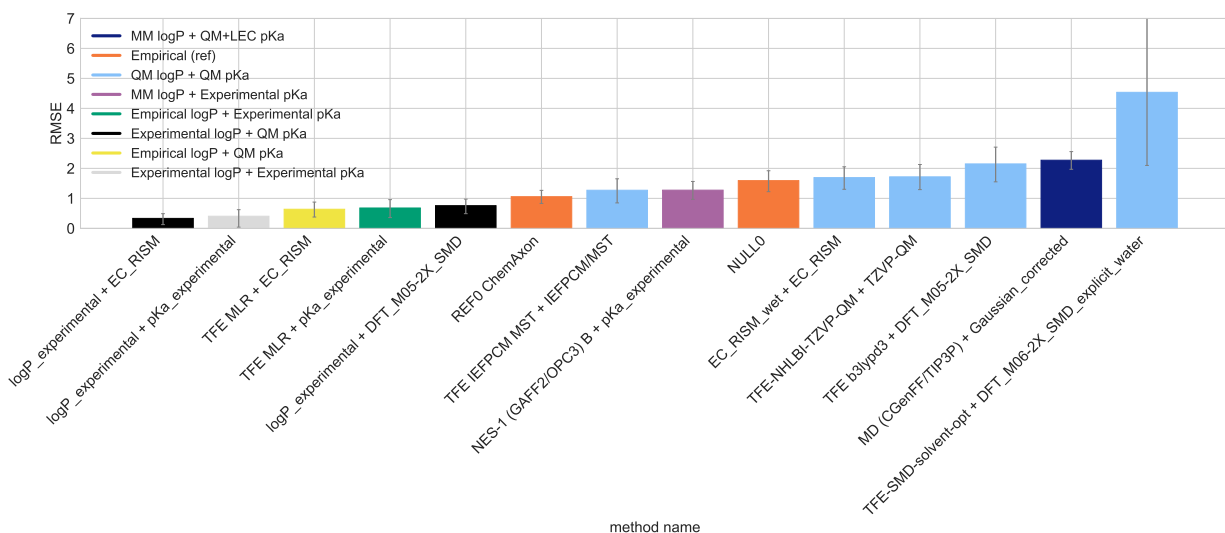


Figure 3.17: **log D values from a combination of the best pK_a and log P are typically superior.** Shown is the RMSE in calculated log D values, with error bars denoting 95% confidence intervals from bootstrapping over challenge molecules. This plot is similar to Figure 3.4.3A, except it includes some additional pK_a and log P combinations (for log D estimation). Method *logP_experimental + EC_RISM* combines the experimental log P with the top performing pK_a method (based on RMSE). Method *logP_experimental + pKa_experimental* combines the experimental log P and pK_a value. Method *TFE MLR + EC_RISM* combines the best performing (based on RMSE) log P and pK_a methods. Method *TFE MLR + pKa_experimental* combines the best performing (based on RMSE) log P method with the experimental pK_a . Method *logP_experimental + DFT_M05-2X_SMD* combines the experimental log P with an average performing pK_a method. Method *NES-1 (GAFF2/OPC3) B + pKa_experimental* combines a log P method with average performance with the experimental pK_a . All other methods are the same as in Figure 3.4.3A.

the availability of octanol-water $\log P$ training data.

Performance in the SAMPL7 $\log P$ challenge was poorer than in the SAMPL6 $\log P$ challenge. In the SAMPL6 $\log P$ challenge, 10 methods achieved a RMSE $\leq 0.5 \log P$ units, while here, none did. In general, the SAMPL7 molecules were more flexible, which may have contributed to this accuracy difference. The chemical diversity in the SAMPL6 challenge dataset was limited to 6 molecules with 4-amino quinazoline groups and 2 molecules with a benzimidazole group. The SAMPL7 set was larger and more diverse, thus possibly more challenging.

For ranked submissions, we identified 5 consistently well-performing methods for $\log P$ evaluations based on several statistical metrics. These particularly well performing methods included three empirical methods, and two QM-based physical methods.

To see if any molecules posed particular challenges, we looked at $\log P$ prediction accuracy for each molecule across all methods. Compounds belonging to the isoxazole compound class had higher $\log P$ prediction errors. MM-based physical methods tended to make predictions that were less accurate for molecules belonging to the isoxazole compound class compared to QM-based physical and empirical method categories.

In the SAMPL7 pK_a challenge, participants predicted free energies for transitions between microstates. Predicted relative free energies were then converted to macroscopic pK_a values in order to compare participants' predictions to experimental pK_a values and calculate performance statistics of predictions. This format allowed us to avoid some of the challenges of matching microscopic transitions to macroscopic pK_a values [91], making analysis more straightforward. As noted above, some matching is still required, but this approach eliminates uncertainty about which transitions are predicted.

Macroscopic pK_a evaluations relied on accuracy and correlation metrics. No method achieved a RMSE around 0.5 or lower for macroscopic pK_a predictions for the challenge molecules

which means methods did not achieve experimental accuracy, which is likely around 0.5 pK_a units [63]. Methods had RMSE values between 0.7 to 5.4 pK_a units. Compared to the previous SAMPL6 pK_a challenge, accuracy remains roughly the same. Out of all submitted methods in SAMPL7, two methods achieved a RMSE lower than 1 pK_a unit (one of which was a commercially available method that we used as a reference method), while a RMSE between 1.8 and 5.4 log units was observed for the majority of methods. In terms of correlation, predictions had R^2 values ranging from 0.03 to 0.93 and only two methods achieved an R^2 greater than 0.9.

We tested ChemAxon’s Chemicalize toolkit [2] as an empirical reference method to make macroscopic pK_a predictions and it performed better than other methods. Excluding the reference method, the two best performing methods across several performance statistics were both QM-based physical methods.

For microscopic pK_a , we find that some transitions are much more consistently predicted than others, but in some cases there is broad disagreement even about the sign of the free energy change associated with a particular transition – so methods disagree as to which protonation state or tautomer is preferred at the reference pH. Participants agreed on the sign of predictions for roughly 56% of all microstates, while 38% disagreed on sign (predictions were negative or positive). Certain chemical transformations were found to have a high level of disagreement, especially protonation of nitrogens in 1,2,3-triazoles, isoxazoles, as well as those in terminal nitrogen groups. Transitions involving keto-enol neutral state tautomerism also often lead to sign disagreement.

The current challenge combined log P and pK_a submissions in order to evaluate the current state of log D predictions. In general we find that the accuracy of octanol-water log P predictions in SAMPL7 is higher than that of cyclohexane-water log D predictions in SAMPL5. Half of the methods in the current challenge achieved a RMSE below 2 log D units, while no submissions achieved this in the SAMPL5 challenge. Given the abundance of octanol-

water partitioning and distribution data (compared to cyclohexane-water data in SAMPL5) it makes sense that accuracy would be higher here in SAMPL7 since trained methods (i.e. empirical methods and implicit solvent QM) are impacted by availability of training data.

3.6 Code and Data Availability

All SAMPL7 physical property instructions, submissions, experimental data and analysis are available at

https://github.com/samplchallenges/SAMPL7/tree/master/physical_property.

Figures and supporting material for this paper can be found at

<https://github.com/MobleyLab/sampl7-physical-property-challenge-manuscript>. This repository contains graphs and plots from the paper, some of which are available in the main SAMPL7 physical property repository listed directly above, but also includes:

- A graph that shows the distribution of molecular properties of the 22 compounds from the SAMPL7 physical property blind challenge.
- Details of MM-based physical methods that made log P predictions.
- A table that lists additional info for microscopic pK_a predictions. The table lists the: microstate, total number of relative free energy predictions, average relative free energy prediction, average relative free energy prediction STD, Minimum relative free energy prediction, maximum relative free energy prediction, number of (+) sign predictions, number of (-) sign predictions, number of neutral (0) sign predictions, and Shannon entropy (H).
- A table of the number of states per charge state for the microstates used in the SAMPL7 pK_a challenge.

- A table of the SAMPL7 molecule ID, compound class, and isomeric SMILES of SAMPL7 physical property challenge molecules.
- Structures of the molecules in the SAMPL7 physical property challenge grouped by compound class.
- A figure showing an example of a relative free energy network.
- A figure showing chemical transformations that repeatedly show up as having large disagreement on the sign of the relative free energy prediction in the pK_a challenge.
- Structures of microstates where relative microstate free energy predictions disagree for the pK_a challenge.
- A figure showing the Shannon entropy per microstate transition in the pK_a challenge.

3.7 Overview of Supplementary Information

Contents of Supplementary Information

- **Table 3.18** Distribution of molecular properties of the 22 compounds from the SAMPL7 physical property blind challenge.
- **Table 3.6** Evaluation statistics calculated for all methods in the $\log P$ challenge.
- **Table 3.19** Overall correlation assessment for all methods participating in the SAMPL7 $\log P$ challenge.
- **Table 3.7** Details MM-based physical methods that made $\log P$ predictions.
- **Table 3.8** Evaluation statistics calculated for all methods in the pK_a challenge.
- **Table 3.10** Additional info for microscopic pK_a predictions.

- **Table 3.12** Number of states per charge state for the microstates used in the SAMPL7 pK_a challenge.
- **Table 3.9** Evaluation statistics calculated for all log D estimates.
- **Figure 3.20** SMILES and compound class of SAMPL7 physical property challenge molecules.
- **Table 3.11** Compound classes and structures of the molecules in the SAMPL7 physical property challenge.

3.8 Author Contributions

Conceptualization, TDB, DLM; Methodology, TDB, DLM, MRG, NT, SMK; Software, TDB, JM, YZ, NT; Formal Analysis, TDB; Investigation, TDB, DLM; Resources, DLM; Data Curation, TDB; Writing-Original Draft, TDB, DLM; Writing - Review and Editing, TDB, DLM, JM, SMK; Visualization, TDB, YZ; Supervision, DLM; Project Administration, DLM; Funding Acquisition, DLM, TDB.

3.9 Acknowledgments

We appreciate Michael Gilson at the University of California of San Diego (UCSD) for making the introduction which made this challenge possible. TDB and DLM gratefully acknowledge support from NIH Grant R01GM124270 supporting the SAMPL Blind challenges. TDB acknowledges and appreciates support from the Association for Computing Machinery’s Special Interest Group on High Performance Computing (ACM SIGHPC) and Intel Fellowship. DLM appreciates financial support from the National Institutes of Health (1R01GM124270-01A1) and the National Science Foundation (CHE 1352608). MRG and YZ acknowledge

support from the National Science Foundation (MCB-1519640). SMK and NK acknowledge support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-2033 – Projektnummer 390677874, and under the Research Unit FOR 1979. SMK and NK thank the IT and Media Center (ITMC) of the TU Dortmund for computational support.

3.10 Disclaimers

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

3.11 Disclosures

David L. Mobley serves on the Scientific Advisory Board of OpenEye Scientific Software and is an Open Science Fellow with Silicon Therapeutics, a subsidiary of Ruyvant.

3.12 Supplementary Information

3.12.1 Supplementary Figures and tables

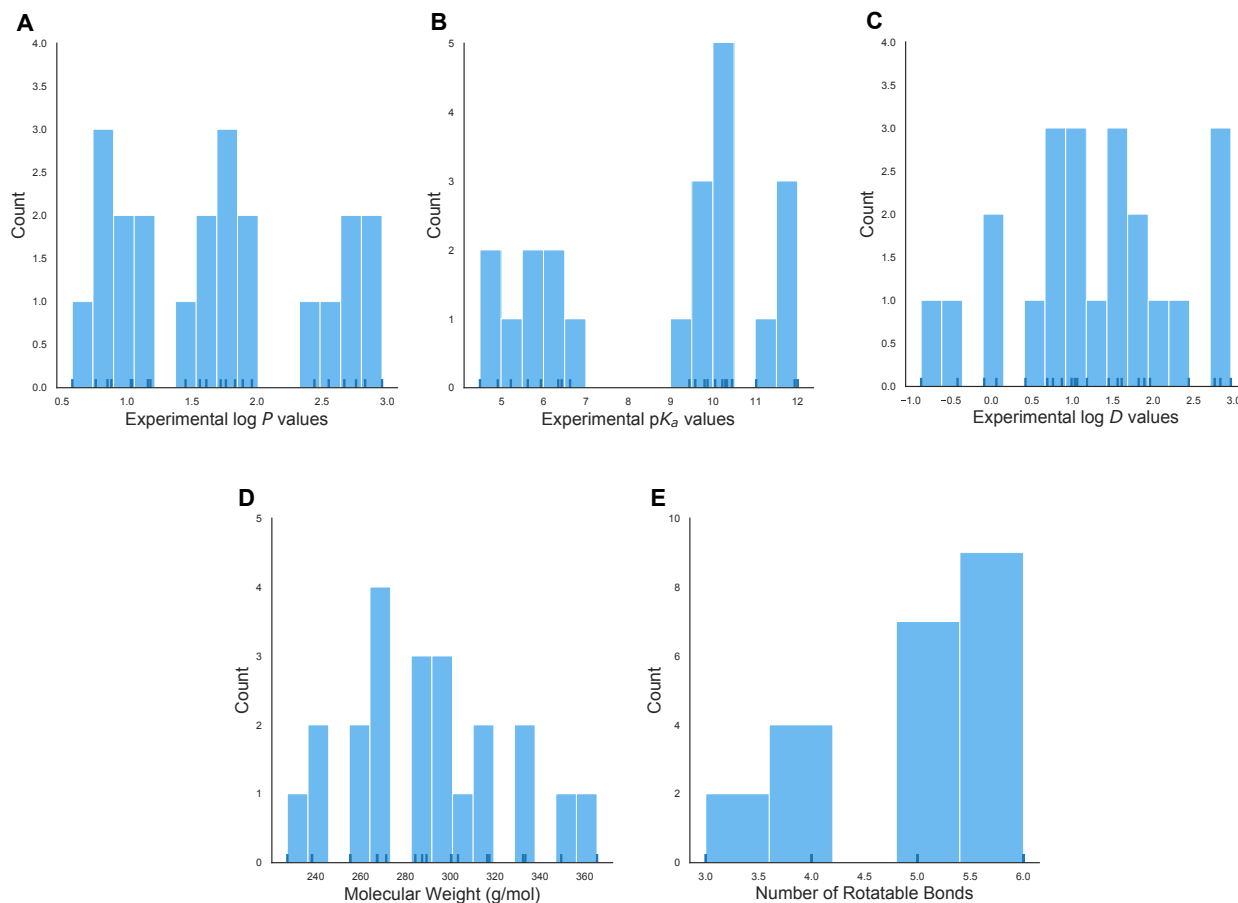


Figure 3.18: **Distribution of molecular properties of the 22 compounds from the SAMPL7 physical property blind challenge.** (A) Histogram of log P measurements collected with Sirius T3 instrument. The ticks along the x-axis indicate the individual values. Compounds have experimental log P values in the range of 0.58-2.96. (B) Histogram of pK_a measurements collected with Sirius T3 instrument.. Eight compounds have measured pK_a 's in the range of 4.49–6.62 and eleven in the range 9.58– 11.93. Two compounds are included here as having pK_a 's of 12, but actually had experimental values greater than 12, and were therefore outside of the experimental detection range. (C) Histogram of log D measurements between n-octanol and aqueous buffer at pH 7.4 were determined via potentiometric titrations using a Sirius T3 instrument, except for compounds SM27, SM28, SM30-SM34, SM36-SM39 which had log $D_{7.4}$ values determined via shake-flask assay. log D measurements ranged from -0.87-2.96. (D) Histogram of molecular weights calculated for the compounds in the SAMPL7 dataset. The molecular weight ranged from 227-365 Da. (E) Histogram of the number of rotatable bonds in each molecule. The number of rotatable bonds in challenge molecules ranged from 3-6.

Table 3.6: **Evaluation statistics calculated for all methods in the log P challenge.** Submitted predictions are represented by their method name. There are six error metrics reported: the root-mean-squared error (RMSE), mean absolute error (MAE), mean (signed) error (ME), coefficient of determination (R^2), linear regression slope (m), and Kendall’s Rank Correlation Coefficient (τ), and error slope (ES). The mean and 95% confidence intervals of each statistic is presented. This table is ranked by increasing RMSE.

Method Name	Category	Submission Type	RMSE	MAE	ME	R^2	m	Kendall’s Tau	ES
<i>ClassicalGSG DB2</i>	Empirical	Blind	0.55 [0.38, 0.69]	0.44 [0.31, 0.58]	0.05 [-0.20, 0.26]	0.51 [0.18, 0.82]	0.71 [0.36, 1.06]	0.51 [0.18, 0.78]	0.81 [0.62, 1.03]
<i>TFE MLR</i>	Empirical	Blind	0.58 [0.34, 0.83]	0.41 [0.26, 0.60]	-0.04 [-0.30, 0.19]	0.43 [0.06, 0.80]	0.60 [0.21, 0.95]	0.56 [0.23, 0.82]	1.38 [1.27, 1.45]
<i>ClassicalGSG DB4</i>	Empirical	Blind	0.65 [0.50, 0.78]	0.55 [0.41, 0.69]	0.25 [0.01, 0.50]	0.51 [0.19, 0.76]	0.82 [0.39, 1.22]	0.45 [0.15, 0.71]	0.57 [0.46, 0.85]
<i>Chemprop</i>	Empirical	Blind	0.66 [0.39, 0.88]	0.48 [0.30, 0.68]	-0.17 [-0.44, 0.08]	0.41 [0.11, 0.76]	0.69 [0.31, 1.08]	0.54 [0.25, 0.82]	1.03 [0.79, 1.21]
<i>TFE-SM8-vacuum-opt</i>	Physical (QM)	Blind	0.67 [0.45, 0.86]	0.51 [0.33, 0.69]	0.15 [-0.13, 0.42]	0.45 [0.11, 0.75]	0.80 [0.33, 1.23]	0.50 [0.18, 0.76]	0.99 [0.75, 1.20]
<i>GROVER</i>	Empirical	Blind	0.69 [0.41, 0.96]	0.49 [0.31, 0.71]	-0.21 [-0.50, 0.05]	0.33 [0.04, 0.70]	0.56 [0.18, 0.92]	0.37 [0.05, 0.66]	0.87 [0.62, 1.09]
<i>ClassicalGSG DB1</i>	Empirical	Blind	0.76 [0.56, 0.96]	0.62 [0.45, 0.82]	0.10 [-0.23, 0.40]	0.28 [0.06, 0.60]	0.61 [0.26, 0.99]	0.36 [0.04, 0.63]	0.63 [0.43, 0.85]
<i>ffsampled_deeplearning_cl1</i>	Empirical	Blind	0.77 [0.44, 1.04]	0.51 [0.29, 0.77]	-0.25 [-0.58, 0.04]	0.31 [0.05, 0.70]	0.63 [0.24, 1.05]	0.42 [0.06, 0.74]	0.99 [0.72, 1.19]
<i>ClassicalGSG DB3</i>	Empirical	Blind	0.77 [0.57, 0.96]	0.62 [0.43, 0.82]	-0.15 [-0.46, 0.16]	0.51 [0.18, 0.78]	1.08 [0.55, 1.59]	0.48 [0.15, 0.75]	0.60 [0.42, 0.89]
<i>COSMO-RS</i>	Physical (QM)	Blind	0.78 [0.49, 1.01]	0.57 [0.36, 0.80]	-0.30 [-0.61, -0.01]	0.49 [0.17, 0.79]	0.97 [0.49, 1.45]	0.53 [0.25, 0.78]	0.97 [0.74, 1.18]
<i>TFE_Attentive_FP</i>	Empirical	Blind	0.79 [0.47, 1.07]	0.57 [0.36, 0.82]	-0.18 [-0.53, 0.12]	0.19 [0.00, 0.61]	0.44 [0.04, 0.87]	0.34 [-0.02, 0.69]	0.93 [0.69, 1.13]
<i>ffsampled_deeplearning_cl2</i>	Empirical	Blind	0.82 [0.48, 1.11]	0.56 [0.32, 0.83]	-0.37 [-0.69, -0.08]	0.36 [0.07, 0.72]	0.73 [0.31, 1.16]	0.40 [0.07, 0.69]	0.94 [0.67, 1.15]
<i>TFE-SM12-vacuum-opt</i>	Physical (QM)	Blind	0.82 [0.61, 1.02]	0.66 [0.47, 0.87]	0.28 [-0.06, 0.60]	0.41 [0.08, 0.72]	0.90 [0.36, 1.42]	0.39 [0.05, 0.67]	0.88 [0.65, 1.09]
<i>TFE-SM8-solvent-opt</i>	Physical (QM)	Blind	0.97 [0.71, 1.20]	0.78 [0.55, 1.02]	0.65 [0.35, 0.94]	0.42 [0.10, 0.70]	0.83 [0.35, 1.31]	0.44 [0.13, 0.69]	0.71 [0.47, 0.94]
<i>REF1 ChemAxon</i>	Empirical	Reference	1.00 [0.79, 1.20]	0.85 [0.63, 1.08]	0.46 [0.08, 0.83]	0.39 [0.10, 0.70]	0.98 [0.45, 1.53]	0.40 [0.09, 0.68]	0.13 [-0.00, 0.29]
<i>TFE IEFFCM MST</i>	Physical (QM)	Blind	1.03 [0.65, 1.41]	0.80 [0.56, 1.10]	-0.07 [-0.53, 0.33]	0.27 [0.01, 0.68]	0.85 [0.12, 1.50]	0.42 [0.10, 0.70]	1.07 [0.88, 1.23]
<i>TFE MD neat oct (GAFF/TIP4P)</i>	Physical (MM)	Blind	1.11 [0.74, 1.43]	0.83 [0.52, 1.15]	-0.74 [-1.10, -0.40]	0.56 [0.24, 0.82]	1.25 [0.64, 1.83]	0.58 [0.27, 0.82]	1.30 [1.19, 1.40]
<i>NULL0 mean clogP FDA</i>	Empirical	Reference	1.20 [0.94, 1.42]	1.01 [0.73, 1.28]	-0.96 [-1.26, -0.64]	0.00 [0.00, 0.00]	0.00 [-0.00, 0.00]	nan [nan, nan]	0.18 [0.04, 0.32]
<i>NES-1 (GAFF2/OPC3) G</i>	Physical (MM)	Blind	1.21 [0.92, 1.51]	1.03 [0.78, 1.31]	-0.13 [-0.63, 0.37]	0.22 [0.01, 0.59]	0.88 [0.15, 1.59]	0.34 [0.02, 0.63]	1.23 [1.11, 1.33]
<i>NES-1 (GAFF2/OPC3) J</i>	Physical (MM)	Blind	1.28 [0.97, 1.58]	1.08 [0.81, 1.38]	0.01 [-0.54, 0.53]	0.21 [0.01, 0.63]	0.92 [0.09, 1.76]	0.33 [0.00, 0.64]	1.21 [1.08, 1.33]
<i>NES-1 (GAFF2/OPC3) B</i>	Physical (MM)	Blind	1.42 [1.02, 1.81]	1.13 [0.79, 1.51]	-0.51 [-1.08, 0.05]	0.27 [0.02, 0.65]	1.11 [0.30, 1.91]	0.36 [0.05, 0.65]	1.17 [1.01, 1.31]
<i>MD (GAFF/TIP3P)</i>	Physical (MM)	Blind	1.43 [1.15, 1.71]	1.30 [1.06, 1.56]	-1.30 [-1.56, -1.06]	0.48 [0.22, 0.79]	0.77 [0.45, 1.12]	0.55 [0.28, 0.80]	0.94 [0.80, 1.09]
<i>TFE wet oct (GAFF/TIP4P)</i>	Physical (MM)	Blind	1.47 [1.16, 1.77]	1.30 [1.03, 1.60]	-1.30 [-1.60, -1.03]	0.42 [0.10, 0.75]	0.80 [0.30, 1.30]	0.47 [0.14, 0.75]	1.15 [1.03, 1.27]
<i>TFE-NHLBI-TZVP-QM</i>	Physical (QM)	Blind	1.55 [1.19, 1.88]	1.34 [1.02, 1.67]	1.32 [1.00, 1.67]	0.52 [0.19, 0.78]	1.16 [0.59, 1.65]	0.51 [0.19, 0.78]	0.05 [-0.00, 0.17]
<i>0.05 MD (CGenFF/TIP3P)</i>	Physical (MM)	Blind	1.63 [1.25, 1.98]	1.41 [1.08, 1.76]	-1.38 [-1.74, -1.02]	0.54 [0.26, 0.82]	1.26 [0.81, 1.76]	0.52 [0.26, 0.76]	0.90 [0.70, 1.07]
<i>[HTML]EFEFEF EC_RISM_wet</i>	Physical (QM)	Blind	1.84 [1.31, 2.36]	1.49 [1.07, 1.96]	-1.49 [-1.96, -1.06]	0.29 [0.05, 0.68]	0.96 [0.37, 1.57]	0.38 [0.08, 0.67]	0.67 [0.45, 0.90]
<i>TFE-SMD-vacuum-opt</i>	Physical (QM)	Blind	1.96 [1.60, 2.30]	1.76 [1.42, 2.13]	1.76 [1.42, 2.13]	0.44 [0.12, 0.68]	1.04 [0.46, 1.59]	0.41 [0.03, 0.70]	0.68 [0.50, 0.87]
<i>[HTML]EFEFEF MD-EE-MCC (GAFF-TIP4P-Ew)</i>	Physical (MM)	Blind	2.06 [1.48, 2.59]	1.61 [1.09, 2.17]	-0.93 [-1.70, -0.17]	0.03 [0.00, 0.28]	0.47 [-0.53, 1.49]	0.11 [-0.16, 0.38]	0.76 [0.51, 1.03]
<i>MD (OPLS-AA/TIP4P)</i>	Physical (MM)	Blind	2.19 [1.69, 2.65]	1.82 [1.31, 2.34]	-1.35 [-2.03, -0.60]	0.28 [0.06, 0.58]	1.47 [0.58, 2.55]	0.36 [0.07, 0.62]	0.73 [0.48, 0.97]
<i>[HTML]EFEFEF TFE b3typd3</i>	Physical (QM)	Blind	2.19 [1.76, 2.57]	1.98 [1.59, 2.37]	1.98 [1.59, 2.37]	0.40 [0.10, 0.67]	1.06 [0.47, 1.64]	0.45 [0.11, 0.72]	0.22 [0.09, 0.41]
<i>MD LigParGen (OPLS-AA/TIP4P)</i>	Physical (MM)	Blind	2.28 [1.80, 2.71]	1.95 [1.46, 2.44]	0.35 [-0.60, 1.29]	0.07 [0.00, 0.37]	0.83 [-0.51, 2.26]	0.19 [-0.14, 0.50]	0.65 [0.42, 0.88]
<i>[HTML]EFEFEF TFE-SMD-solvent-opt</i>	Physical (QM)	Blind	2.39 [1.97, 2.78]	2.19 [1.79, 2.60]	2.19 [1.79, 2.60]	0.40 [0.09, 0.67]	1.09 [0.45, 1.67]	0.42 [0.09, 0.68]	0.51 [0.34, 0.68]
<i>Ensemble EPI physprop</i>	Empirical	Blind	2.73 [2.27, 3.16]	2.54 [2.13, 2.98]	2.54 [2.13, 2.98]	0.33 [0.04, 0.64]	-0.30 [-0.49, -0.10]	-0.35 [-0.60, -0.03]	-0.00 [-0.00, -0.00]
<i>Ensemble Martel</i>	Empirical	Blind	3.29 [2.89, 3.68]	3.16 [2.78, 3.55]	3.16 [2.78, 3.55]	0.39 [0.05, 0.73]	-0.25 [-0.40, -0.09]	-0.46 [-0.72, -0.14]	-0.00 [-0.00, -0.00]
<i>QSPR_Mordred2D_TPOT_AutoML</i>	Empirical	Blind	3.64 [3.01, 4.24]	3.36 [2.80, 3.96]	3.36 [2.80, 3.96]	0.39 [0.10, 0.71]	-0.72 [-1.12, -0.33]	-0.37 [-0.65, -0.04]	-0.00 [-0.00, -0.00]
<i>TFE-NHLBI-NN-IN</i>	Empirical	Blind	3.97 [3.57, 4.34]	3.85 [3.45, 4.25]	3.85 [3.45, 4.25]	0.00 [0.00, 0.15]	0.02 [-0.30, 0.34]	0.02 [-0.23, 0.27]	0.01 [-0.00, 0.02]

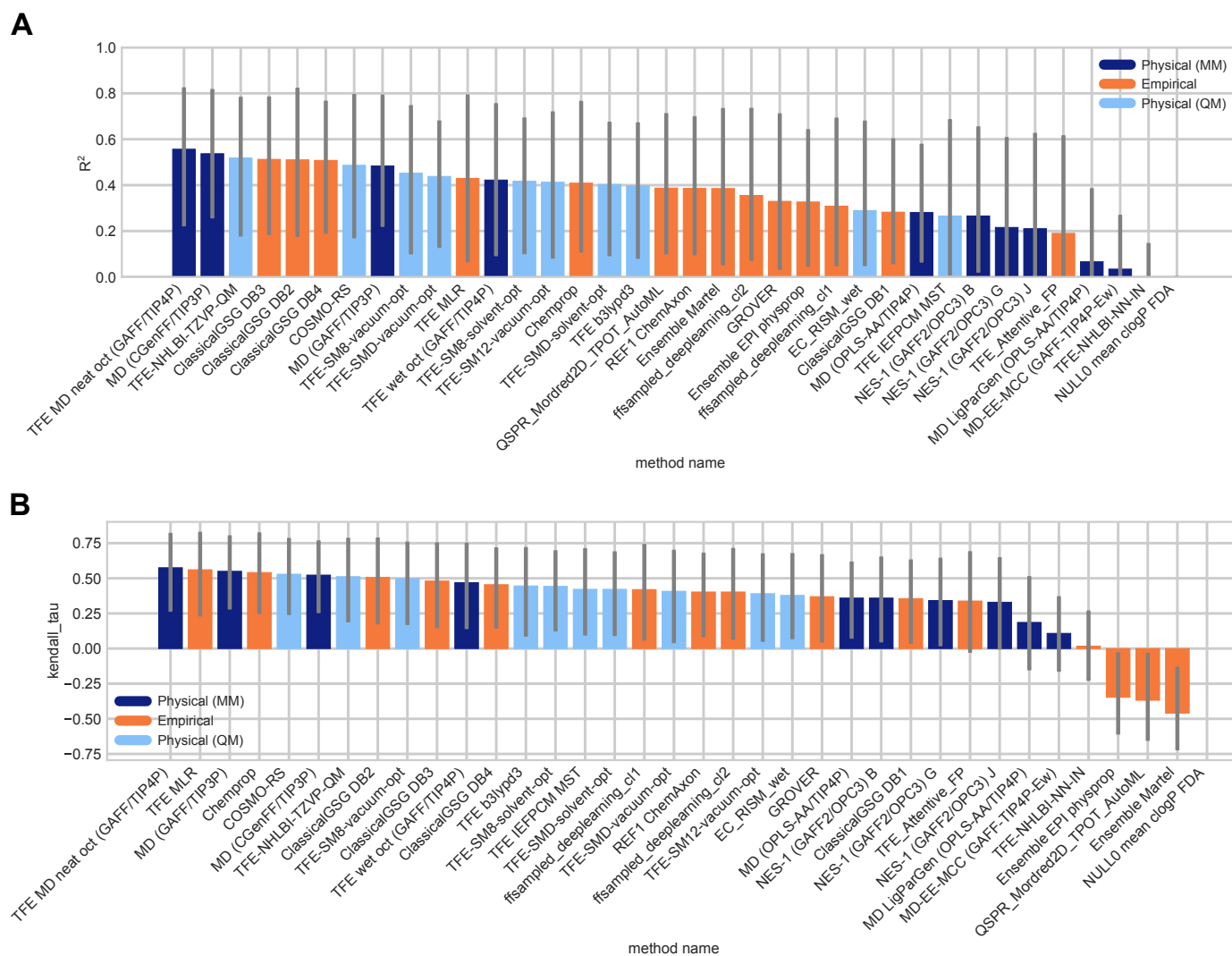


Figure 3.19: Overall correlation assessment for all methods participating in the SAMPL7 log P challenge show that the uncertainty of each correlation statistic is quite high, not allowing a true ranking based on correlation. Pearson's R^2 and Kendall's Rank Correlation Coefficient τ are shown, with error bars denoting 95% confidence intervals obtained by bootstrapping over challenge molecules. Submitted methods are listed in Table 3.1. The submission *REF1 ChemAxon* was a reference method included after the blind challenge submission deadline, and *NULL0 mean cLogP FDA* is the null prediction method; all others refer to blind predictions. Most methods have a statistically indistinguishable performance on ranking because of the small dynamic range of the dataset. Evaluation statistics calculated for all methods are available in Table 3.6 of the Supplementary Information.

Table 3.7: **Details of MM-based physical methods in the log P prediction challenge.** Force fields, water models, and octanol phase choice are reported. A dry octanol phase indicates the octanol phase was composed of only octanol. A wet octanol phase indicates the octanol phase was treated as a mixture of octanol and water. RMSE, MAE, R^2 , and Kendall’s Tau values are reported as mean and 95% confidence intervals.

Method Name	Force Field	Water Model	Octanol Phase	RMSE	MAE	R^2	Kendall’s Tau
<i>TFE MD neat oct (GAFF/TIP4P)</i>	GAFF	TIP4P	dry	1.11 [0.74, 1.43]	0.83 [0.52, 1.15]	0.56 [0.24, 0.82]	0.58 [0.27, 0.82]
<i>NES-1 (GAFF2/OPC3) G</i>	GAFF2	OPC3	dry	1.21 [0.92, 1.51]	1.03 [0.78, 1.31]	0.22 [0.01, 0.59]	0.34 [0.02, 0.63]
<i>NES-1 (GAFF2/OPC3) J</i>	GAFF2	OPC3	dry	1.28 [0.97, 1.58]	1.08 [0.81, 1.38]	0.21 [0.01, 0.63]	0.33 [0.00, 0.64]
<i>NES-1 (GAFF2/OPC3) B</i>	GAFF2	OPC3	dry	1.42 [1.02, 1.81]	1.13 [0.79, 1.51]	0.27 [0.02, 0.65]	0.36 [0.05, 0.65]
<i>MD (GAFF/TIP3P)</i>	GAFF	TIP3P	dry	1.43 [1.15, 1.71]	1.30 [1.06, 1.56]	0.48 [0.22, 0.79]	0.55 [0.28, 0.80]
<i>TFE wet oct (GAFF/TIP4P)</i>	GAFF	TIP4P	wet	1.47 [1.16, 1.77]	1.30 [1.03, 1.60]	0.42 [0.10, 0.75]	0.47 [0.14, 0.75]
<i>MD (CGenFF/TIP3P)</i>	CGenFF	TIP3P	dry	1.63 [1.25, 1.98]	1.41 [1.08, 1.76]	0.54 [0.26, 0.82]	0.52 [0.26, 0.76]
<i>MD-EE-MCC (GAFF-TIP4P-Ew)</i>	GAFF	TIP4P-eW	dry	2.06 [1.48, 2.59]	1.61 [1.09, 2.17]	0.03 [0.00, 0.28]	0.11 [-0.16, 0.38]
<i>MD (OPLS-AA/TIP4P)</i>	OPLS-AA	TIP4P	dry	2.19 [1.69, 2.65]	1.82 [1.31, 2.34]	0.28 [0.06, 0.58]	0.36 [0.07, 0.62]
<i>MD LigParGen (OPLS-AA/TIP4P)</i>	OPLS-AA	TIP4P	dry	2.28 [1.80, 2.71]	1.95 [1.46, 2.44]	0.07 [0.00, 0.37]	0.19 [-0.14, 0.50]

Table 3.8: **Evaluation statistics calculated for all methods in the pK_a challenge.** Submitted predictions are represented by their method name. There are six error metrics reported: the root-mean-squared error (RMSE), mean absolute error (MAE), mean (signed) error (ME), coefficient of determination (R^2), linear regression slope (m), Kendall’s Rank Correlation Coefficient (τ), and error slope (ES). The mean and 95% confidence intervals of each statistic is presented. This table is ranked by increasing RMSE.

Method Name	Category	Submission Type	RMSE	MAE	ME	R^2	m	Kendall’s Tau	ES
<i>REF00_Chemaxon_Chemicalize</i>	QSPR/ML	Reference	0.71 [0.50, 0.90]	0.56 [0.38, 0.76]	0.09 [-0.23, 0.38]	0.91 [0.86, 0.96]	0.88 [0.72, 1.02]	0.73 [0.51, 0.90]	0.83 [0.58, 1.04]
<i>EC_RISM</i>	QM	Blind	0.72 [0.45, 0.95]	0.53 [0.33, 0.75]	0.20 [-0.10, 0.50]	0.93 [0.87, 0.98]	0.80 [0.72, 0.91]	0.81 [0.63, 0.96]	1.32 [1.19, 1.42]
<i>IEFPCM/MST</i>	QM	Blind	1.82 [1.00, 2.69]	1.30 [0.84, 1.92]	0.25 [-0.46, 1.09]	0.56 [0.22, 0.87]	0.86 [0.53, 1.18]	0.52 [0.22, 0.76]	1.00 [0.80, 1.17]
<i>DFT_M05-2X_SMD</i>	QM	Blind	2.90 [2.04, 3.69]	2.28 [1.53, 3.10]	-0.78 [-2.02, 0.41]	0.03 [0.00, 0.37]	0.15 [-0.32, 0.53]	0.17 [-0.22, 0.54]	0.55 [0.31, 0.81]
<i>TZVP-QM</i>	QM	Blind	2.90 [2.52, 3.25]	2.75 [2.34, 3.14]	1.20 [0.02, 2.33]	0.23 [0.03, 0.60]	-0.11 [-0.20, -0.04]	-0.14 [-0.49, 0.23]	-0.00 [-0.00, -0.00]
<i>Standard Gaussian Process</i>	QSPR/ML	Blind	3.49 [2.76, 4.12]	2.91 [2.06, 3.75]	2.47 [1.38, 3.55]	0.30 [0.10, 0.69]	-0.05 [-0.09, -0.02]	-0.42 [-0.70, -0.08]	1.11 [0.96, 1.24]
<i>DFT_M06-2X_SMD_implicit</i>	QM	Blind	4.16 [2.00, 6.38]	2.80 [1.76, 4.33]	-0.07 [-1.61, 1.95]	0.52 [0.39, 0.78]	1.70 [0.80, 2.77]	0.70 [0.48, 0.88]	0.50 [0.30, 0.70]
<i>DFT_M06-2X_SMD_implicit_SAS</i>	QM	Blind	4.16 [2.03, 6.44]	2.81 [1.80, 4.36]	-0.20 [-1.71, 1.85]	0.50 [0.36, 0.77]	1.64 [0.72, 2.72]	0.56 [0.28, 0.81]	0.14 [0.02, 0.31]
<i>DFT_M06-2X_SMD_explicit_water</i>	QM	Blind	5.12 [1.19, 7.92]	2.56 [0.96, 4.76]	-0.35 [-2.62, 1.93]	0.20 [0.00, 0.81]	1.10 [-0.39, 2.50]	0.46 [0.06, 0.78]	0.52 [0.29, 0.77]
<i>Gaussian_corrected</i>	QM+LEC	Blind	5.36 [4.70, 5.95]	5.12 [4.42, 5.79]	5.12 [4.42, 5.79]	0.76 [0.63, 0.88]	0.35 [0.27, 0.45]	0.60 [0.42, 0.76]	0.00 [-0.00, 0.00]

Table 3.9: **Evaluation statistics calculated for all log D estimates.** Predictions are represented a name based on method names participants submitted to the and log P challenges. There are six error metrics reported: the root-mean-squared error (RMSE), mean absolute error (MAE), mean (signed) error (ME), coefficient of determination (R^2), linear regression slope (m), Kendall’s Rank Correlation Coefficient (τ), and error slope (ES). The mean and 95% confidence intervals of each statistic is presented. This table is ranked by increasing RMSE.

Method Name	Category	Submission Type	RMSE	MAE	ME	R^2	m	Kendall’s Tau	ES
<i>REF0 ChemAxon</i>	Empirical	Reference	1.06 [0.82, 1.27]	0.91 [0.68, 1.14]	0.28 [-0.14, 0.70]	0.27 [0.01, 0.58]	0.54 [0.10, 0.90]	0.31 [-0.02, 0.61]	0.12 [-0.00, 0.28]
<i>TFE IEFPCM MST + IEFPCM/MST</i>	Physical (QM)	Standard	1.27 [0.85, 1.64]	0.98 [0.67, 1.33]	0.24 [-0.28, 0.75]	0.55 [0.17, 0.87]	1.31 [0.71, 1.70]	0.57 [0.27, 0.82]	1.16 [0.89, 1.25]
<i>NULL0</i>	Empirical	Reference	1.59 [1.22, 1.93]	1.35 [1.00, 1.71]	1.23 [0.81, 1.65]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	nan [nan, nan]	0.65 [0.44, 0.87]
<i>EC_RISM</i>	Physical (QM)	Standard	1.69 [1.30, 2.05]	1.43 [1.07, 1.82]	-1.43 [-1.81, -1.07]	0.53 [0.20, 0.77]	0.95 [0.54, 1.29]	0.51 [0.21, 0.74]	0.84 [0.64, 1.02]
<i>TFE-NHLBI-TZVP-QM + TZVP-QM</i>	Physical (QM)	Standard	1.72 [1.30, 2.12]	1.47 [1.12, 1.86]	1.26 [0.78, 1.75]	0.25 [0.01, 0.64]	0.64 [0.08, 1.25]	0.38 [0.02, 0.70]	0.05 [-0.00, 0.18]
<i>TFE b3typd3 + DFT_M05-2X_SMD</i>	Physical (QM)	Standard	2.15 [1.56, 2.71]	1.78 [1.31, 2.31]	1.78 [1.31, 2.31]	0.32 [0.04, 0.66]	0.80 [0.27, 1.30]	0.41 [0.05, 0.72]	0.42 [0.27, 0.70]
<i>MD (CGenFF/TIP3P) + Gaussian_corrected</i>	Physical (MM) + QM+LEC	Standard	2.27 [1.97, 2.55]	2.13 [1.80, 2.45]	1.84 [1.21, 2.35]	0.62 [0.35, 0.84]	1.53 [0.93, 2.18]	0.62 [0.36, 0.82]	0.88 [0.75, 1.00]
<i>TFE-SMD-solvent-opt + DFT_M06-2X_SMD_explicit_water</i>	Physical (QM)	Standard	4.54 [2.09, 7.15]	2.92 [1.88, 4.57]	2.88 [1.80, 4.55]	0.25 [0.11, 0.76]	1.92 [0.53, 4.45]	0.55 [0.22, 0.80]	0.55 [0.38, 0.73]

Table 3.10: Additional info for microscopic pK_a predictions.

Microstate	Total number of relative free energy predictions	Average relative free energy prediction	Average relative free energy prediction STD	Minimum relative free energy prediction	Maximum relative free energy prediction	Number of (+) sign predictions	Number of (-) sign predictions	Number of neutral (0) sign predictions	Shannon entropy (H)
SM25_micro001	9	-0.6	13.2	-15.6	16.3	4	5	0	0.7
SM25_micro002	8	8.8	10.6	-7.5	20.4	6	2	0	0.6
SM25_micro003	8	9.6	2.7	4.5	12.6	8	0	0	0.0
SM25_micro004	2	-8.9	4.5	-12.1	-5.8	0	2	0	0.0
SM25_micro005	2	-0.8	2.1	-2.3	0.7	1	1	0	0.7
SM26_micro001	9	7.3	2.4	3.0	10.7	9	0	0	0.0
SM26_micro002	8	-6.7	20.5	-31.7	22.1	3	5	0	0.7
SM26_micro003	8	20.9	12.0	0.9	32.4	8	0	0	0.0
SM26_micro004	2	4.3	0.7	3.8	4.8	2	0	0	0.0
SM26_micro005	2	8.1	2.6	6.3	10.0	2	0	0	0.0
SM27_micro001	9	13.4	4.9	6.1	19.0	9	0	0	0.0
SM28_micro001	9	-5.7	25.0	-39.0	23.5	4	5	0	0.7
SM28_micro002	8	17.1	8.0	8.2	26.5	8	0	0	0.0
SM28_micro003	8	0.9	8.3	-10.0	12.6	4	4	0	0.7
SM28_micro004	2	25.1	9.1	18.7	31.5	2	0	0	0.0
SM29_micro001	9	12.6	4.3	6.3	18.7	9	0	0	0.0
SM30_micro001	9	12.3	4.2	5.9	17.7	9	0	0	0.0
SM31_micro001	9	13.2	4.4	6.0	18.1	9	0	0	0.0
SM31_micro002	3	-0.6	6.6	-8.1	4.5	2	1	0	0.6
SM32_micro001	9	12.8	4.6	5.9	18.9	9	0	0	0.0
SM33_micro001	9	11.9	3.9	5.2	17.1	9	0	0	0.0
SM34_micro001	9	13.0	4.6	5.7	19.7	9	0	0	0.0
SM34_micro002	3	-0.9	6.4	-8.1	4.4	2	1	0	0.6
SM35_micro001	9	11.7	4.5	3.2	16.2	9	0	0	0.0
SM35_micro002	8	0.2	1.4	-1.9	2.5	5	2	1	0.9
SM35_micro003	8	12.2	5.6	3.2	18.1	8	0	0	0.0
SM36_micro001	9	10.8	3.1	5.2	14.9	9	0	0	0.0
SM36_micro002	8	1.2	1.8	0.0	4.4	4	1	3	1.0
SM36_micro003	8	10.7	3.3	5.2	14.7	8	0	0	0.0
SM37_micro001	9	0.1	9.4	-11.7	13.7	5	4	0	0.7
SM37_micro002	8	9.8	2.9	3.7	12.7	8	0	0	0.0
SM37_micro003	8	0.7	1.8	-1.5	4.2	4	3	1	1.0
SM37_micro004	8	8.9	3.0	3.8	12.4	8	0	0	0.0
SM37_micro005	7	-2.7	7.6	-10.6	11.0	3	4	0	0.7
SM38_micro001	9	11.6	4.6	5.2	17.5	9	0	0	0.0
SM39_micro001	9	10.1	3.1	5.1	14.6	9	0	0	0.0
SM40_micro001	9	10.8	3.3	5.0	15.7	9	0	0	0.0
SM40_micro002	8	-1.8	10.3	-15.5	11.8	4	4	0	0.7
SM41_micro001	9	8.4	3.5	2.2	14.8	9	0	0	0.0
SM41_micro002	8	-0.5	9.9	-12.9	13.9	4	4	0	0.7
SM42_micro001	9	5.5	4.6	0.2	12.3	9	0	0	0.0
SM42_micro002	8	-0.2	8.6	-10.8	14.3	4	4	0	0.7
SM42_micro003	3	-2.0	3.0	-5.1	1.0	1	2	0	0.6
SM43_micro001	9	5.9	4.4	0.5	13.4	9	0	0	0.0
SM43_micro002	8	0.1	9.4	-11.0	11.0	4	4	0	0.7
SM43_micro003	8	-11.6	38.1	-60.9	38.2	4	4	0	0.7
SM43_micro004	2	-3.6	2.2	-5.2	-2.1	0	2	0	0.0
SM43_micro005	2	0.1	0.4	-0.2	0.4	1	1	0	0.7
SM44_micro001	9	9.5	2.9	4.3	12.9	9	0	0	0.0
SM44_micro002	8	-1.1	7.4	-10.3	9.9	4	4	0	0.7
SM45_micro001	9	9.6	3.1	4.4	14.7	9	0	0	0.0
SM45_micro002	8	-1.0	7.8	-11.0	9.6	4	4	0	0.7
SM46_micro001	9	9.9	4.1	4.0	18.4	9	0	0	0.0
SM46_micro002	8	-0.7	7.5	-9.6	10.5	4	4	0	0.7
SM46_micro003	8	-12.2	37.1	-63.5	39.0	4	4	0	0.7
SM46_micro004	3	6.3	4.5	2.4	11.3	3	0	0	0.0

Table 3.11: SMILES and compound class of SAMPL7 physical property challenge molecules. A view of the compounds and their classes can be found in Figure 3.20.

SAMPL7 Molecule ID	Compound Class	Isomeric SMILES
<i>SM25</i>	acylsulfonamide	<chem>O=C(NS(C1=CC=CC=C1)(=O)=O)CCC2=CC=CC=C2</chem>
<i>SM26</i>	acylsulfonamide	<chem>O=S(CCC1=CC=CC=C1)(NC(C)=O)=O</chem>
<i>SM27</i>	oxetane	<chem>O=S(CCC1=CC=CC=C1)(NC2(C)COC2)=O</chem>
<i>SM28</i>	thietane-1,1-dioxide	<chem>O=S(CC1(NC(C)=O)CCC2=CC=CC=C2)(C1)=O</chem>
<i>SM29</i>	oxetane	<chem>CS(NC1(COC1)CCC2=CC=CC=C2)(=O)=O</chem>
<i>SM30</i>	oxetane	<chem>O=S(NC1(COC1)CCC2=CC=CC=C2)(C3=CC=CC=C3)=O</chem>
<i>SM31</i>	oxetane	<chem>O=S(NC1(COC1)CCC2=CC=CC=C2)(N(C)C)=O</chem>
<i>SM32</i>	thietane	<chem>CS(NC1(CSC1)CCC2=CC=CC=C2)(=O)=O</chem>
<i>SM33</i>	thietane	<chem>O=S(NC1(CSC1)CCC2=CC=CC=C2)(C3=CC=CC=C3)=O</chem>
<i>SM34</i>	thietane	<chem>O=S(NC1(CSC1)CCC2=CC=CC=C2)(N(C)C)=O</chem>
<i>SM35</i>	thietane-1-oxide	<chem>CS(N[C@@]1(C[S+](O-))C1)CCC2=CC=CC=C2)(=O)=O</chem>
<i>SM36</i>	thietane-1-oxide	<chem>O=S(N[C@@]1(C[S+](O-))C1)CCC2=CC=CC=C2)(C3=CC=CC=C3)=O</chem>
<i>SM37</i>	thietane-1-oxide	<chem>O=S(N[C@@]1(C[S+](O-))C1)CCC2=CC=CC=C2)(N(C)C)=O</chem>
<i>SM38</i>	thietane-1,1-dioxide	<chem>CS(NC1(CS(C1)=O)=O)CCC2=CC=CC=C2)(=O)=O</chem>
<i>SM39</i>	thietane-1,1-dioxide	<chem>O=S(NC1(CS(C1)=O)=O)CCC2=CC=CC=C2)(C3=CC=CC=C3)=O</chem>
<i>SM40</i>	thietane-1,1-dioxide	<chem>O=S(NC1(CS(C1)=O)=O)CCC2=CC=CC=C2)(N(C)C)=O</chem>
<i>SM41</i>	isoxazole	<chem>O=S(NC1=NOC(C2=CC=CC=C2)=C1)(C)=O</chem>
<i>SM42</i>	isoxazole	<chem>O=S(NC1=NOC(C2=CC=CC=C2)=C1)(C3=CC=CC=C3)=O</chem>
<i>SM43</i>	isoxazole	<chem>O=S(NC1=NOC(C2=CC=CC=C2)=C1)(N(C)C)=O</chem>
<i>SM44</i>	1,2,3-triazole	<chem>O=S(NC(N=N1)=CN1C2=CC=CC=C2)(C)=O</chem>
<i>SM45</i>	1,2,3-triazole	<chem>O=S(NC(N=N1)=CN1C2=CC=CC=C2)(C3=CC=CC=C3)=O</chem>
<i>SM46</i>	1,2,3-triazole	<chem>O=S(NC(N=N1)=CN1C2=CC=CC=C2)(N(C)C)=O</chem>

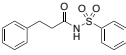
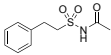
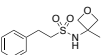
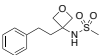
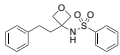
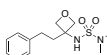
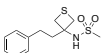
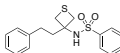
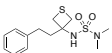
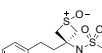
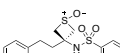
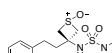
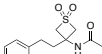
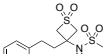
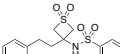
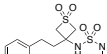
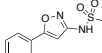
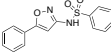
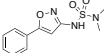
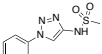
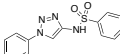
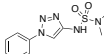
Compound Classes	Structures
acylsulfonamide	  SM25 SM26
oxetane	    SM27 SM29 SM30 SM31
thietane	   SM32 SM33 SM34
thietane-1-oxide	   SM35 SM36 SM37
thietane-1,1-dioxide	    SM28 SM38 SM39 SM40
isoxazole	   SM41 SM42 SM43
1,2,3-triazole	   SM44 SM45 SM46

Figure 3.20: Compound classes and structures of the molecules in the SAMPL7 physical property challenge. SMILES of the compounds are in Table 3.20.

Table 3.12: **Number of states per charge state for the microstates used in the SAMPL7 pK_a challenge.** The total number of microstates (protomers and tautomers) is listed. Some of the molecules have up to 6 microstates, while others have only 2.

	Charge State				Total #
	+2	+1	0	-1	
<i>SM25</i>	0	1	3	2	6
<i>SM26</i>	0	1	3	2	6
<i>SM27</i>	0	0	1	1	2
<i>SM28</i>	0	1	2	2	5
<i>SM29</i>	0	0	1	1	2
<i>SM30</i>	0	0	1	1	2
<i>SM31</i>	0	1	1	1	3
<i>SM32</i>	0	0	1	1	2
<i>SM33</i>	0	0	1	1	2
<i>SM34</i>	0	1	1	1	3
<i>SM35</i>	0	0	2	3	5
<i>SM36</i>	0	0	2	3	5
<i>SM37</i>	0	2	2	2	6
<i>SM38</i>	0	0	1	1	2
<i>SM39</i>	0	0	1	1	2
<i>SM40</i>	0	1	1	1	3
<i>SM41</i>	0	1	1	1	3
<i>SM42</i>	0	1	2	1	4
<i>SM43</i>	1	2	2	1	6
<i>SM44</i>	0	1	1	1	3
<i>SM45</i>	0	1	1	1	3
<i>SM46</i>	1	2	1	1	5

Chapter 4

Enhancing Water Sampling of Buried Binding Sites Using Nonequilibrium Candidate Monte Carlo

Teresa Danielle Bergazin, Ido Y. Ben-Shalom, Nathan M. Lim, Sam C. Gill, Michael K. Gilson,
and David L. Mobley.

Journal of Computer-Aided Molecular Design volume 35, pages 167–177 (2021)

doi: 10.1007/s10822-020-00344-8

Publication Date (Web): September 24, 2020

4.1 Abstract

Water molecules can be found interacting with the surface and within cavities in proteins. However, water exchange between bulk and buried hydration sites can be slow compared to simulation timescales, thus leading to the inefficient sampling of the locations of water.

This can pose problems for free energy calculations for computer-aided drug design. Here, we apply a hybrid method that combines nonequilibrium candidate Monte Carlo (NCCMC) simulations and molecular dynamics (MD) to enhance sampling of water in specific areas of a system, such as the binding site of a protein. Our approach uses NCCMC to gradually remove interactions between a selected water molecule and its environment, then translates the water to a new region, before turning the interactions back on. This approach of gradual removal of interactions, followed by a move and then reintroduction of interactions, allows the environment to relax in response to the proposed water translation, improving acceptance of moves and thereby accelerating water exchange and sampling. We validate this approach on several test systems including the ligand-bound MUP-1 and HSP90 proteins with buried crystallographic waters removed. We show that our BLUES (NCCMC/MD) method enhances water sampling relative to normal MD when applied to these systems. Thus, this approach provides a strategy to improve water sampling in molecular simulations which may be useful in practical applications in drug discovery and biomolecular design.

4.1.1 Keywords

Molecular Dynamics simulations · Monte Carlo · NCCMC · nonequilibrium candidate Monte Carlo · enhanced sampling · water sampling · buried binding sites · buried cavity · buried water · Major Urinary Protein · Heat Shock Protein 90

4.1.2 Abbreviations

BLUES Binding modes of Ligands Using Enhanced Sampling

MD Molecular Dynamics

NCCMC Nonequilibrium Candidate Monte Carlo

MUP-1 Major Urinary Protein

HSP90 Heat Shock Protein 90

4.2 Introduction

Proteins are found in aqueous environments where water plays a major role in determining their structure, function, and dynamics [123, 18]. Water molecules can also be found in cavities in proteins [141, 171, 59] where they play a variety of roles, such as facilitating receptor-ligand recognition and contributing to the stability of proteins [204, 171, 25, 123, 22].

Classical molecular dynamics (MD) simulations can be used to understand the motions and interactions of biomolecular systems, including how proteins interact with water. However, water exchange between bulk and buried hydration sites can be slow compared to simulation timescales [44, 150, 139]. This leads to the inefficient sampling of the locations of water and water's role in binding events [50]. Simulations that do not account for these water motions will give an incomplete picture of the binding process and any downstream predictions will thus risk being in error [150, 50].

Several methods may better sample water occupancy and rearrangements in the cavities of proteins. Monte Carlo (MC) methods can substantially accelerate water sampling via large translational water moves around a system, but these MC moves can be difficult to get accepted due to steric clashes in the system. For example, grand canonical Monte Carlo [9, 10], which works by insertion and deletion of water to maintain a specific chemical potential, has been applied to sample water configurations and accelerate occupancy of buried sites [227, 116, 188]. However, this approach has been shown to be inefficient due to steric clashes which results in a high rejection of the proposed moves [143, 189]. Another approach integrates

Metropolis MC translational water moves with traditional MD to equilibrate water across steric barriers and into buried hydration sites that are not accessible with pure MD [26].

Here, we seek to enhance the sampling of water rearrangements through extension of our Binding Modes of Ligands Using Enhanced Sampling (BLUES) approach [75], which combines hybrid nonequilibrium candidate Monte Carlo (NCMC) [167] with MD simulations. BLUES has been shown to enhance ligand sampling efficiency by more than two orders of magnitude compared to classical MD when applied to a model test system [75]. In BLUES, NCMC alchemically scales off the electrostatic and steric interactions until a water molecule is no longer interacting with its environment and then translates it to a new location before scaling the interactions back on. This results in a proposed NCMC move which is either accepted or rejected based on the integrated work during this process. After this, the NCMC move is followed by traditional molecular dynamics. By mixing NCMC translational water sampling moves with classical MD simulations, we improve water sampling in a selected region, such as a binding site of a protein, where water motions are known to be challenging or slow to sample and likely to pose problems for calculations of interest, such as free energy calculations [50]. In this work, we use the BLUES framework to exchange waters around a specified region of a system. Here, we focus on testing it in specific contexts where water rearrangements can pose challenges for MD sampling, such as buried binding sites in proteins.

4.3 Methods

We introduce a method that integrates NCMC translational water moves with classical MD, allowing water molecules to hydrate buried sites. Here, we detail how this approach is implemented and tested.

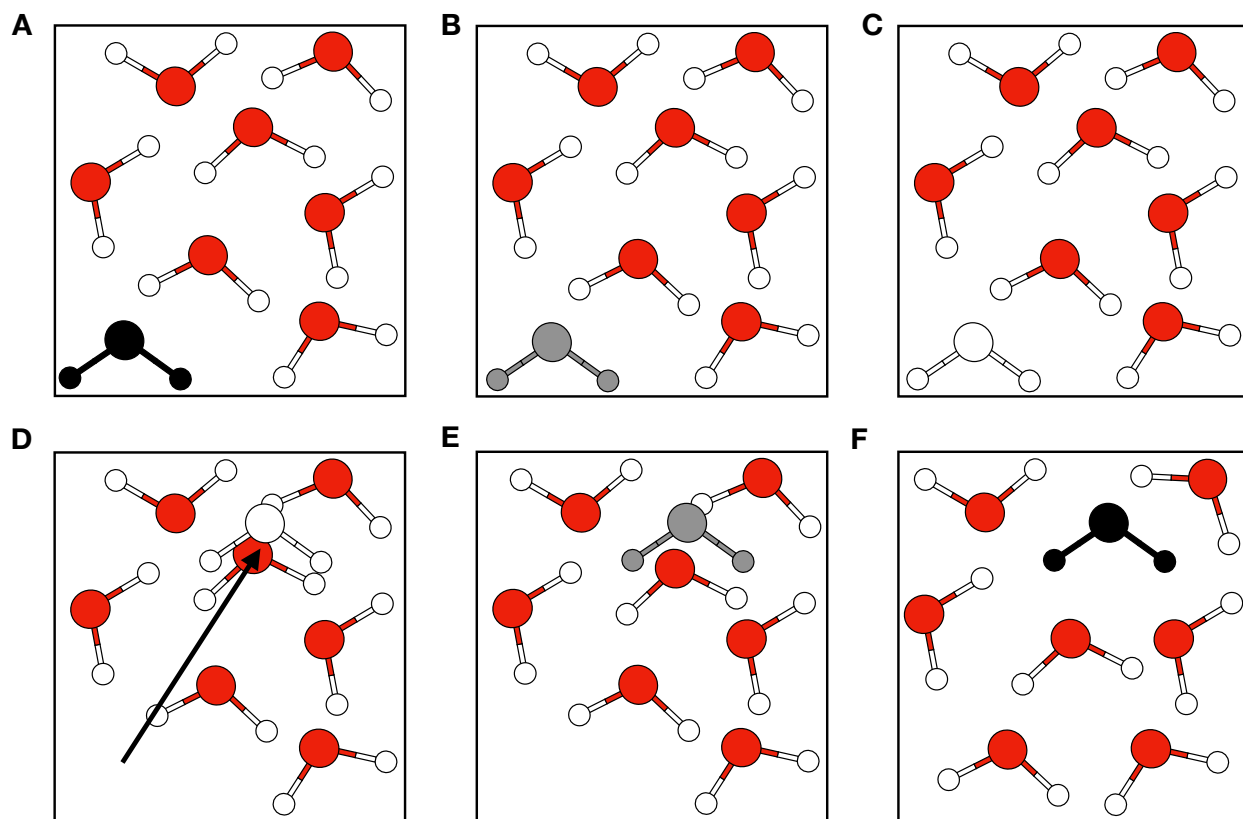


Figure 4.1: **Molecular interactions between atoms are turned off and on during a NEMC move to translate a water molecule.** In this cartoon, water molecules are represented here by red and white spheres for the oxygen and hydrogen atoms. The black-filled water represents a fully interacting water molecule that has been selected to be moved. Gray-filled water represents intermediate levels of interaction and white-filled represents the fully non-interacting water molecule. A) The water molecule (in black) is fully interacting with its surrounding environment, and in this case, other water molecules. B) The water's interactions are partially off, allowing the other water molecules to slightly relax. C) The water's interactions are fully turned off. D) The water is randomly translated to somewhere else in the system (indicated by a black arrow) with its interactions remaining off. E) The water's interactions are partially turned on and the propagation steps of NEMC allow relaxation of the translated water and its surroundings to resolve clashes. F) At the end of the NEMC protocol, the water molecule is once again in the fully interacting state and in a new location. This entire process comprises a proposed NEMC move, which is accepted or rejected based on the nonequilibrium work done in this process, and then followed by conventional MD.

4.3.1 Implementation of NCMC/MD in BLUES

BLUES (Binding Modes of Ligands Using Enhanced Sampling), which combines NCMC with classical MD, was originally created to enhance the sampling of ligand binding modes [75], but has begun applying the same techniques to enhance sampling of other degrees of freedom also important in ligand binding, such as sidechain rearrangements [33] and, here, water motions.

A BLUES iteration consists of a NCMC move followed by regular MD. A NCMC move consists of a series of NCMC steps sandwiching a perturbation to the system, such as a translational water move. The NCMC steps are a series of alchemical steps where the electrostatic/steric interactions are gradually turned off and then back on. While the interactions are completely turned off, a perturbation provided by a translational water move occurs.

Some of our key terminology here is as follows:

- *BLUES iteration* — an NCMC move followed by a series of regular MD steps.
- *NCMC move* — a series of NCMC steps sandwiching a perturbation to the system.
- *NCMC steps* — a series of alchemical steps where the electrostatic/steric interactions are gradually turned off and then scaled back on.
- *MD steps* — a number of steps to advance the MD simulation.

In BLUES, NCMC moves are executed through a switching protocol that is comprised of a series of perturbation and propagation/relaxation steps involving structural and dynamic degrees of freedom [167]. This process helps lower possible steric or electrostatic clashes by allowing the environment surrounding the perturbed region to relax around the proposed state.

NCMC moves are implemented by alchemically “turning off” the interactions between an object in the system and its surrounding environment before the move, followed by turning the interactions back on, as detailed in Figure 4.1. First, the electrostatic and then the steric interactions are turned off (and then later back on) by scaling λ , a variable that controls the strength of nonbonded interactions, from 1 (fully interacting state) to 0 (noninteracting state) over a user-determined number of n NCMC steps (Figure 4.1.A-C). At the point where the object is noninteracting (Figure 4.1.C), the target object’s atoms are repositioned (Figure 4.1.D) and then the interactions are scaled back on (Figure 4.1.D-F) until $\lambda=1$ in reverse order (first sterics and then electrostatics). When the target object’s atoms are repositioned the internal coordinates/conformation remain the same during the move.

The total work done during this process is summed and used to either accept or reject the proposed move (following a modified Metropolis-Hastings acceptance criterion [141] to maintain detailed balance). The NCMC move is then followed by a user-determined number of MD steps. Additional details of BLUES are described in the work of Gill et al. [75].

A proposed NCMC move is either accepted or rejected based on the total work $w[X]$ done during the nonequilibrium process X , estimated as

$$w[X] \equiv \sum_{t=1}^T [u_t(x_t) - u_{t-1}(x_t)] + w_{\text{shadow}}[X] \tag{4.1}$$

where x_t is a microstate at a simulation step t and u_t is the reduced potential energy.

The total work includes both “protocol work” and “shadow work” [201]. In the equation above, the first term is the protocol work and the second term is the shadow work which

accounts for errors introduced by the use of finite-time-step Langevin integrators [193, 201].

The protocol work is computed every time there is a perturbation to the system, so after changing lambda, we track the potential energy change between the states before and after lambda is changed and add this difference to the protocol work. Accumulation would also happen during translational moves, except that the water is non-interacting and the proposed move is just a rigid body translation move, so the system’s energy does not change and thus no protocol work is accumulated during the translational move. The shadow work can be tracked in a similar fashion, except the total energy differences (potential and kinetic) would be taken into account during the propagation phase. However, use of a BAOAB integrator allows us to neglect the shadow work contribution without introducing large errors (the explanation for this is in the original BLUES paper [75]).

To maintain detailed balance, the acceptance probability $A[X]$ is determined using a modified Metropolis-Hastings criterion [82]

$$A[X] = \min \{1, e^{-w_{\text{protocol}}(X)}\} \tag{4.2}$$

After each accepted or rejected NCMC move, velocities are randomly reassigned based on the Maxwell–Boltzmann distribution in order to maintain detailed balance [75]. The amount of relaxation used does not affect whether this procedure preserves the correct distribution. The NCMC move is followed by a series of conventional MD steps, using a Langevin integrator to relax the entire system. This process of proposing (and accepting or rejecting) a NCMC move then conducting a series of MD steps is then repeated many times. This process of a NCMC move, followed up by traditional MD, is what we refer to as a BLUES iteration.

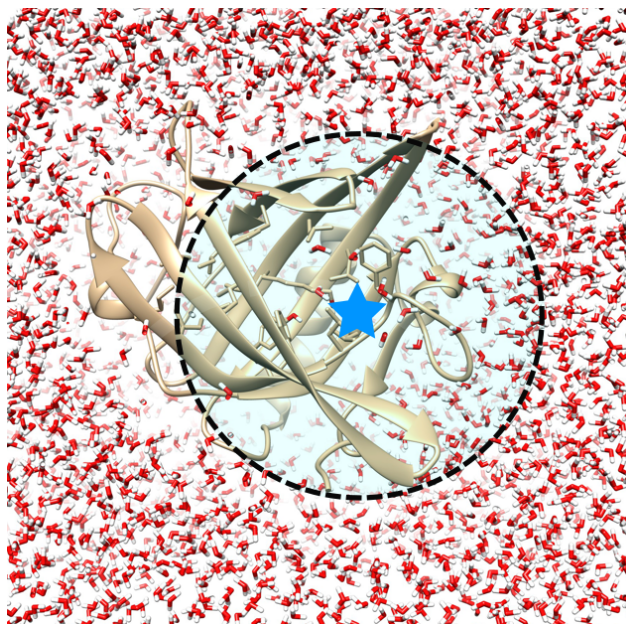


Figure 4.2: **Example of a user-defined radius that covers a particular area of interest.** Here, the MUP-1 protein-ligand system is shown. The radius used (indicated by the black dashed line) defines a sphere around a user-selected atom (represented by a blue star) in the system, such as an atom inside the binding site of a protein.

4.3.2 Translational water moves with BLUES

Here, we build upon the BLUES framework by incorporating “water hopping” moves where random water molecules can be translated between bulk and within a region via NCMC move proposals. Water hopping moves were created in order to enhance sampling of key hydration sites such as in water bridging locations between a protein and ligand, and particularly in buried cavities inaccessible from bulk water.

To define a region within which the water hops occur, the user selects an atom as the center and defines a radius to generate a sphere which encompasses the area of interest (Figure 4.2). Additionally, the sampling region can be set to automatically span from the center of mass of a protein or ligand, rather than manually defining a specific atom. This area of interest must be large enough to include some bulk water to allow water exchange. Our algorithm will subsequently use this radius to select a random water molecule and

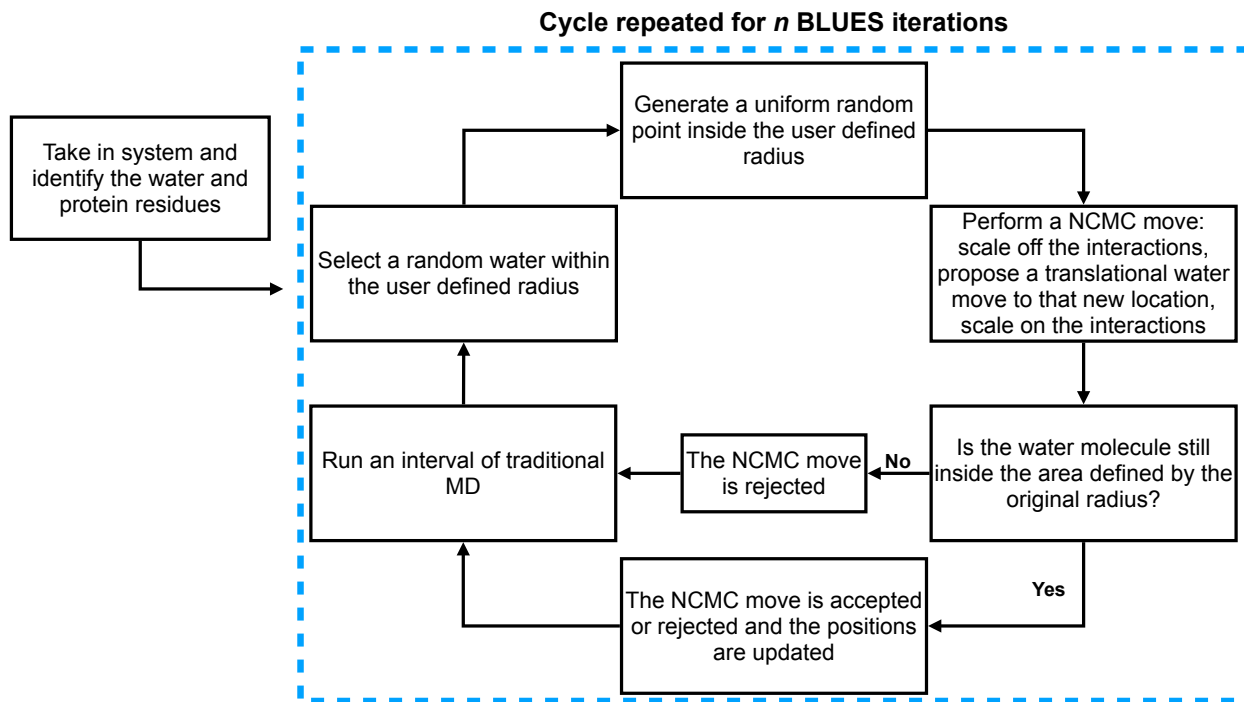


Figure 4.3: **Workflow of a BLUES iteration with translational water hopping move proposals.** Before any water is translated to a new location, the user first selects an atom and picks a radius defining a sphere encompassing an area of interest around the position of the atom and BLUES identifies all the water and protein residues in the system. Afterward, BLUES goes through a number of BLUES iterations n number of times, where each BLUES iteration is as shown inside the dashed box. A schematic of the NCMC move process is shown in Figure 4.1.

propose moving it to a random new position within this region. During a BLUES iteration, a random water molecule is selected, a random point in this region is generated, then a NCMC move proposal is performed. During the NCMC move proposal the interactions are scaled off between the atoms of the water molecule and its surrounding environment, then the water molecule is translated to the new location defined by the random point, and then the interactions are scaled back on. This the NCMC move proposal process is depicted in Figure 4.1. The work done during this NCMC move proposal is accumulated and the move is either accepted or rejected using the Metropolis-Hastings acceptance criterion [141]. Afterwards, an interval of regular MD is run. A workflow a BLUES iteration is depicted in Figure 4.3. Further water hopping implementation details used in this work are available in python scripts deposited in the Supporting Information. More documentation, details, and the full BLUES package are available on GitHub at <https://github.com/MobleyLab/blues>, in the BLUES documentation (<https://mobleylab-blues.readthedocs.io>), and detailed in the work of Gill et al. [75].

4.3.3 Comparing sampling efficiency using the number of force evaluations

BLUES simulations consist of intervals of both classical MD and NCMC moves, so comparing a BLUES simulations to classical MD simulation requires accounting for the cost of the switching protocol that occurs during the NCMC move. We account for the additional cost from NCMC by considering the number of force evaluations rather than the aggregated simulation time in nanoseconds or microseconds.

NCMC carries out a single force evaluation for each perturbation or propagation/relaxation step. The perturbation steps are the instantaneous perturbation of the water molecules coordinates (or for turning off/on the alchemical parameters), and this is combined with

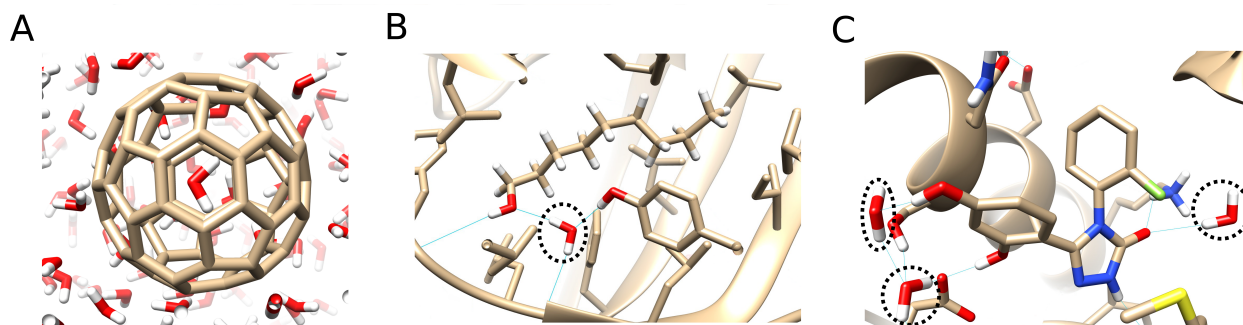


Figure 4.4: **Systems used to test the ability of BLUES (NCMC/MD) water hopping to allow the exchange of water.** (A) A C_{60} buckyball with a single trapped water molecule. (B) The buried hydration site of the MUP-1 protein with a bound ligand. (C) The hydration site of the HSP90 protein bound to a ligand. The protein-ligand systems have internal water(s) (indicated by the black dashed line) that do not easily exchange with bulk.

propagation steps via Langevin dynamics [75]. In other words, perturbation steps modify the system or its potential, and propagation steps propagate the dynamics. A BLUES simulation consists of NCMC and MD, so a BLUES simulation will have a total cost in force evaluations of:

$$\text{Total force evaluations} = (nSteps_{MD} + nSteps_{NCMC}) \times nIter \quad (4.3)$$

where $nSteps_{MD}$ is the number of MD steps per BLUES iteration, $nSteps_{NCMC}$ is the number of NCMC steps per BLUES iteration and $nIter$ is the number of BLUES iterations, which consists of a NCMC move proposal followed by a series of regular MD. The total cost in force evaluations for classical MD is equivalent to the total number of MD steps.

4.3.4 Test cases and simulation details

We used a C_{60} buckyball, a water box system with dividing graphene sheets, Major Urinary Protein (MUP-1) and Heat Shock Protein 90 (HSP90) as systems to test the ability of the BLUES (NCMC/MD) water hopping moves to enhance the sampling of water molecules in desired regions. Many of these systems were also used in a similar study to validate Metropolis MC translational water moves with traditional MD [26].

The first system was a C_{60} buckyball with a water molecule trapped inside (Figure 4.4.A). This water molecule is unable to interact with bulk water and cannot form any hydrogen bonds with the buckyball's carbon atoms. Hence, it is in an energetically unfavorable environment, but it is unable to diffuse out. We chose a sampling region that was centered on a carbon atom in the buckyball and extended 12 Å out, such that the region included the entire buckyball and some bulk water. The box size was $\sim 44 \times 44 \times 44$ Å³, and had a total of 213 water molecules.

The second system was a rectangular water box divided into two regions by impermeable planar graphene sheets (Figure 4.5.A). These two regions had initially different water densities where the outer and inner region had densities of about 21.5 water/nm³ and 18.5 water/nm³, respectively. The rectangular box was $\sim 32 \times 32 \times 85$ Å³ and the system had a total of 1915 water molecules. The initially differing densities between the outer and inner region tested the ability of the BLUES (NCMC/MD) water hopping method to equalize the water densities between the sheets. We chose a sampling region that was centered on a carbon atom in the middle of one of the sheets and extended 15 Å out so that the sampling region covered the same amount of area in the inner and outer regions. This choice was important to ensure that we didn't make dramatically more move proposals to one region relative to the other. Additionally, we chose our sampling region so that it did not extend outside of the simulation box, thus avoiding issues where we might place waters in the same

region more than once due to periodic boundary conditions, leading to artifacts.

The third and fourth systems tested the method’s ability to exchange water between bulk and buried sites in two proteins. The third system was the MUP-1 protein [203] which contains a buried crystallographic water molecule that bridges between the ligand and the protein (Figure 4.4.B). The crystallographic water molecule was removed in order to test the ability of our water hopping moves to hydrate the buried cavity and reform the water bridging interaction. We chose a sampling region that was centered on a carbon atom in the ligand and extended 20 Å out to include some bulk water (Figure 4.2). The box was $\sim 70 \times 70 \times 70 \text{ \AA}^3$ the system had a total of 8,678 water molecules. The fourth system was the HSP90 protein (PDBID:5J64) [13] bound to a ligand which forms interactions with the protein through three bridging water molecules, as shown in Figure 4.4.C. The box was $\sim 82 \times 82 \times 82 \text{ \AA}^3$ and the system had a total of 13,831 water molecules. We chose a sampling region that was centered on a carbon atom in the ligand, and extended 15 Å out to include some bulk water.

The simulation boxes were built using `tLeap` from AmberTools [35]. All of the systems used, where appropriate, the protein and ligand force field parameters from AMBER ff14SB [86, 131] and GAFF [219], respectively. The water molecules were parameterized using the TIP3P water model [98] in all cases. MD and BLUES simulations were performed using OpenMM (version 7.1.1) [55, 58]. The systems were minimized until forces were below a tolerance of 10 kJ/mol. Long-range electrostatics were calculated using Particle Mesh Ewald [46]. Simulations were run using the hydrogen mass repartitioning scheme with 4 femtosecond timesteps [85].

To focus on water exchange the α -Carbons and ligands in the protein ligand systems were restrained with a force constant of $5 \text{ kcal/mol} \cdot \text{\AA}^2$, thus keeping the protein cavities from quickly collapsing. The carbon atoms in the buckyball and graphene walls in the water box system were also restrained with the same force constant as the protein-ligand systems,

which held the buckyball in place and kept the graphene walls from collapsing/folding.

The temperature was set to 300 K in all cases except the water box with graphene sheets, which was set to 500 K so that the water in the system was less dense than liquid water and wouldn't form water droplets; thus, increasing the NCMC move acceptance rate so that any errors due to the method would be obvious because the density in the two boxes would not reach equilibrium. For the Buckyball system, equilibration consisted of 250 ps of NVT MD and 10 ns of NPT MD of equilibration. For the water box with dividing graphene sheets, equilibration consisted of 5 ns NVT MD. The MUP-1 system was equilibrated for 1 ns of NVT MD and 10ns NPT MD. The MD production run for the water box with dividing graphene sheets and the MUP-1 system was for 40 ns in the NPT ensemble. The HSP90 system was equilibrated for 1 ns of NVT MD and 80 ns NPT MD. The MD production run for HSP90 was for 285 ns in the NPT ensemble.

A BLUES simulation consists of a number of BLUES iterations, where each iteration of BLUES is composed of a NCMC move and traditional MD. Each NCMC move is comprised of a certain number of NCMC perturbation and propagation/relaxation steps (wherein the electrostatic and steric interactions are alchemically scaled off/on, as depicted in Figure 4.1). Here, we used the same amount of NCMC steps for all of the systems (except MUP-1, detailed below). For the water box system with dividing graphene sheets, BLUES with translational water moves was executed for 240,000 BLUES iterations, with each iteration consisting of 2,500 NCMC steps and 1,000 MD steps. The buckyball system was simulated for a total of 1,000 BLUES iterations, using 2,500 NCMC steps and 1,000 MD steps per iteration. Both of the solvated MUP-1 and HSP90 systems were simulated for a total of 10,000 BLUES iterations. For the MUP-1 system, 1,250, 2,500, 5,000 and 30,000 NCMC steps per iteration were tested to see how the number of NCMC steps affects the rate of water transfer from from bulk to the internal hydration site. The number of MD steps in all cases was 1,000 MD steps per iteration. For the HSP90 system, each BLUES iteration consisted of 2,500 NCMC

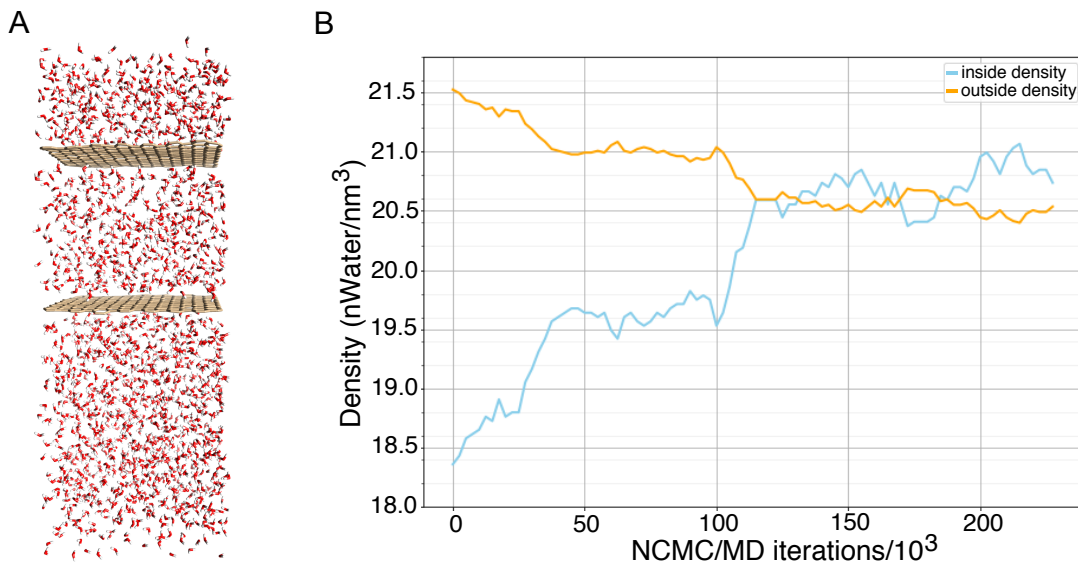


Figure 4.5: **Impermeable graphene sheets divide a box into separate regions with initially different densities, testing the ability of water hopping moves to equilibrate the density.** (A) The water box system with dividing graphene sheets. (B) Shown here are the water densities between the two sheets (blue) and outside the sheets (orange). The densities in the two regions reach equilibrium and stabilize with this approach, serving to validate our implementation.

steps and 1,000 MD steps. Further simulation details are available in scripts deposited in the SI.

4.4 Results and Discussion

The hybrid BLUES (NCMC/MD) approach described here accelerates water sampling during simulations by incorporating translational water moves during the NCMC component of each BLUES iteration. We refer to these translational water moves in BLUES as “water hopping”. Here, we tested these water hopping moves in a range of systems. Particularly, we use a C_{60} buckyball, water box system with dividing graphene sheets, MUP-1 and HSP90 protein-ligand systems to validate the water hopping methodology. Across all of the systems tested, we find that BLUES water hopping moves allowed water exchange between regions, while

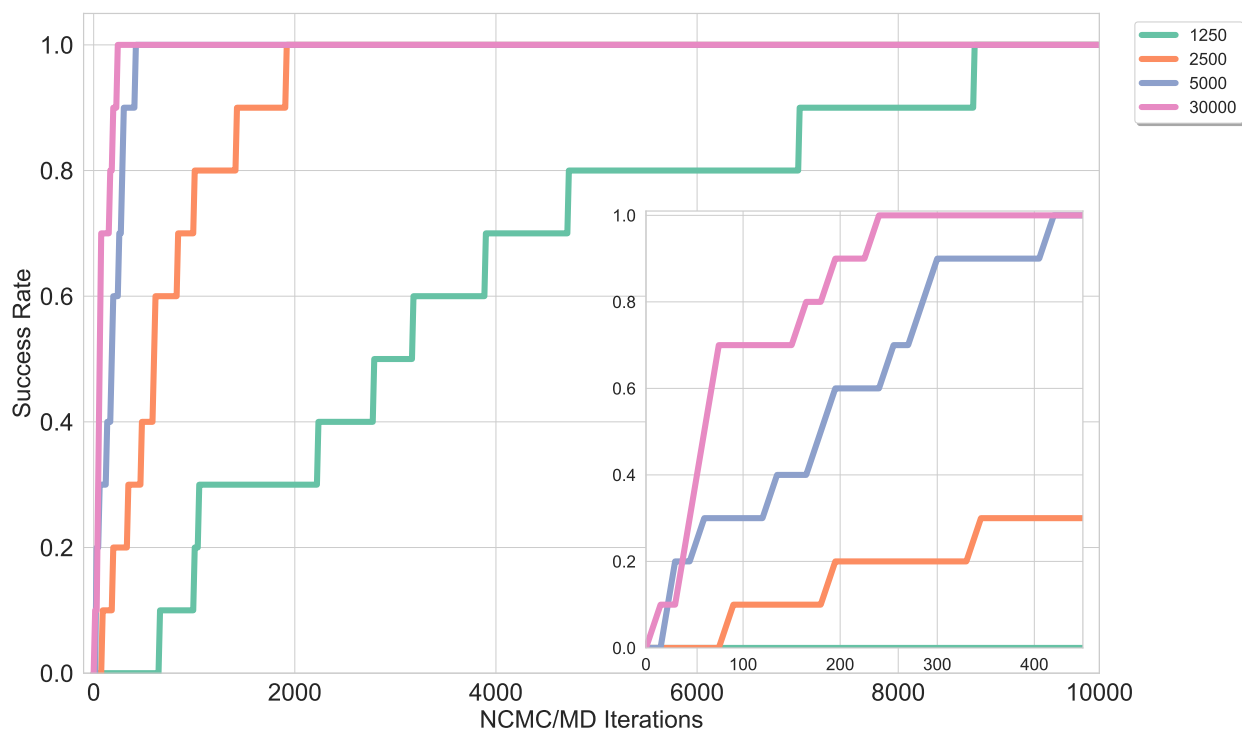


Figure 4.6: **Increasing the amount of NCMC steps increases the rate of water transfer from bulk to the internal hydration site in MUP-1.** Ten replicate simulations with different random seed numbers were run for each NCMC step value. All of the BLUES simulations were run for 10,000 BLUES iterations, with each iteration consisting of a certain number of steps of NCMC and MD. The different colors indicate various amounts of NCMC steps used. The success rate is equivalent to the ratio of the number of replicate simulations where the MUP-1 site (Figure 4.4.B) has been hydrated relative to the total number of replicate simulations. (A) shows that using a lower NCMC step amount increases the number of BLUES iterations for the cavity to become hydrated, such as 1,250 (green) and 2,500 (orange) NCMC steps. The inset, (B), zooms in on the success rate at low iteration number and shows that increasing the amount of NCMC steps decreases the number of iterations needed. 5,000 (blue) NCMC steps needed a little more than 400 BLUES iterations to hydrate the cavity and 30,000 (pink) NCMC steps needed no more than 250 BLUES iterations to hydrate the cavity.

Table 4.1: **Increasing the number of NCMC steps generally increases the acceptance rate of all moves in the MUP-1 protein-ligand system.** Here is the average acceptance rate of all BLUES moves, the average number of force evaluations across 10-12 replicates for the buried cavity in the MUP-1 system to become hydrated, and the average wallclock time in hours for BLUES to hydrate MUP-1. Each simulation was run for 10,000 BLUES iterations, where each iteration consisted of a single NCMC move (consisting of n NCMC steps) and 1,000 MD steps.

n NCMC steps	Average acceptance rate of all BLUES moves	Average number of force evaluations to hydrate the MUP-1 cavity	Average wallclock time to hydrate the MUP-1 cavity
1,250	0.1%	7.9×10^6	50 hours
2,500	0.3%	2.6×10^6	12 hours
5,000	1.1%	1.1×10^6	3 hours
30,000	2.8%	2.5×10^6	4 hours

plain MD did not.

The first test system was a C_{60} buckyball simulated in bulk water, with a single water molecule housed inside (Figure 4.4.A). For the buckyball, it is very unfavorable to have the water inside the buckyball because the water molecule is in an energetically unstable environment relative to a water molecule in bulk. Having a water molecule inside of the buckyball is a state which should not be sampled (to any significant degree) at equilibrium, and we deliberately started with the water in this state to test if BLUES would allow it to escape relatively efficiently. As expected, we find that water hopping moves can relocate the water molecule from the inside of the buckyball to bulk water. Since the trapped water molecule is unable to interact with bulk water or form hydrogen bonds with the buckyball’s carbon shell, it is thermodynamically favorable for it to escape, but it is unable to do so with conventional MD. We chose a sampling region centered on a carbon atom in the buckyball so that the sampling region encompassed the buckyball and some bulk water. While the water molecule is not able to escape the buckyball with plain MD [26], water hopping allowed the water molecule to escape, returning it to the surrounding bulk water after 2.1×10^5 force evaluations. The buckyball remains unoccupied after the water molecule leaves. Since we expect unidirectional transitions, we did not explore how the amount of relaxation affects the acceptance rate.

The second test system was a water box system divided into two regions by impermeable graphene sheets (Figure 4.5.A), with each region having different initial water densities. We find that water hopping successfully equalizes the water between the two regions (Figure 4.5.B). We chose a sampling region centered on a carbon atom in the middle of one of the graphene sheets, such that the sampling region encompassed equivalent amounts of both the inner and outer regions. The relative densities of each region initially differed, but should become uniform over time if BLUES is allowing waters to hop between the two regions. Standard MD does not allow water to enter the inner region between the graphene sheets because the sheets act as barriers that prevent water from passing through them. However, we find that translational water moves in BLUES allow water molecules to hop across the sheets, causing the densities to gradually equalize in both regions (Figure 4.5.B). Here we found this took 4.2×10^8 force evaluations.

Next, we examined a buried hydration site in MUP-1, which has a buried crystallographic water molecule that bridges between the ligand and the protein (Figure 4.4.B). The crystallographic water molecule was removed from the buried site and water hopping successfully rehydrated it. We chose a sampling region that was centered on an atom in the ligand and extended out to include some bulk water (such as in Figure 4.2), such that the sampling region encompassed the buried hydration site and had access to bulk. With plain MD the water did not resume its crystallographic bridging position even after $1.5 \mu s$, equivalent to 3.8×10^8 force evaluations and 120 wallclock hours. However, BLUES was able to recover the crystallographic water. On average (across 11 replicates), it took BLUES 2.6×10^6 force evaluations and 12 wallclock hours to hydrate the site (using 2,500 NCMC steps and 1,000 MD steps per BLUES iteration, as shown in Table 4.1), and no BLUES moves were accepted that dehydrated the site. Additionally, we tested how the number of NCMC steps per BLUES iteration affects the rate of water transfer to the hydration site by simulating with 1,250, 2,500, 5,000, and 30,000 NCMC steps per BLUES iteration, and used 1,000 MD steps per BLUES iteration for each. As expected, increasing the number of NCMC steps per

BLUES iteration increases the rate of water transfer from bulk to the buried hydration site in MUP-1, as shown in Figure 4.6. Here, 30,000 NCMC steps is worse than 5,000 NCMC steps because it will take 6x the number of NCMC steps, but the success rate is certainly not 6x higher (it's only about 2x higher). On the other hand, running 2,500 NCMC steps per BLUES iteration is certainly better than 1,250 NCMC steps. Although it takes 2x the number of NCMC steps, the success rate ends up being more than 2X higher- it's roughly 4.7X higher. Similarly, running 5,000 NCMC steps per BLUES iteration is better than 2,500 NCMC steps because the success rate is about 4x higher.

Although increasing the number of NCMC steps per BLUES iteration decreases the number of BLUES iterations required for the site to become hydrated, we find that increasing the number of NCMC steps per BLUES iteration can also start to negatively effect the efficiency in terms of force evaluations of the water hopping in hydrating the cavity (Table 4.1). Eventually, the increase in efficiency from allowing more relaxation is swamped by the associated increase in computational cost. However, relatively small amounts of relaxation have considerable payoff, resulting in a sort of sweet spot in terms of amount of relaxation. To ensure water hopping is as efficient as possible in terms of force evaluations, we recommend keeping the number of NCMC steps in the lower range, such as 1,250, 2,500, or 5,000.

In terms of wallclock time, 5,000 NCMC steps takes roughly the same amount of time to hydrate the cavity as 30,000 NCMC steps. 2,500 NCMC steps requires 4x less wallclock time to hydrate the cavity compared to 1,250 NCMC steps, and using 5,000 NCMC steps takes 4x less wallclock time to hydrate the cavity compared to 2,500 NCMC steps. Based on this, 5,000 NCMC steps seems to be the most efficient in terms of wallclock time.

Lastly, we examined three hydration sites in the binding site region of the HSP90 protein-ligand system (Figure 4.4.C). All three crystallographic water molecules were removed from the hydration site in the HSP90 system and water hopping successfully rehydrated each hydration site. We chose a sampling region that was centered on a ligand atom and extended

out to encompass the buried hydration site, ligand and some bulk water. With plain MD, only one out of the three water molecules were able to resume the crystallographic bridging positions within 285 ns, which is equivalent to 7.1×10^7 force evaluations. This water molecule moved in from a starting position in bulk water. It took BLUES 5.9×10^6 force evaluations on average (across 4 replicates) to occupy all three of the hydration sites. After the buried cavity had been hydrated, no NCMC moves were accepted that removed any of the water molecules, indicating that the occupancy of these sites is favorable. We did not explore how the amount of relaxation would affect the acceptance rate as we already explored this in the MUP-1 system and found that, in general, increasing the amount of NCMC increases the acceptance rate of all moves (Table 4.1). In terms of wallclock time, the 285 ns MD simulation took about 54 hours and was unable to completely fill the cavity. However, BLUES only took 31 hours to completely rehydrate the cavity.

In both of the protein-ligand systems studied, we restrained the proteins and ligands with a force constant of $5 \text{ kcal/mol} \cdot \text{\AA}^2$ and artificially removed the crystallographic water, which is highly favorable in its place. Therefore, once the water returned to its crystallographic position, it did not transition out of the binding site again.

The sampling region used for the protein-ligand systems encapsulates the binding pocket and some bulk water. Relative to MD, we find that we can increase efficiency by making the area of interest the focal point of NCMC move attempts. Making the sampling region just large enough to cover a specific ligand-binding site and bulk water allows us to speed up the equilibration of water between these two regions, and this strategy has been successfully used elsewhere [26]. If the sampling region covered a greater amount of bulk water in these cases, the efficiency would decrease because the equilibration of water between regions would be slower as more water moves would move water molecules around in just bulk water. In general, we recommend setting the radius to be as small as possible while ensuring that the particular area of interest and some bulk is covered, thus increasing efficiency. In some cases,

a larger sampling region may be more desirable, such as a protein with multiple hydration cavities, and this would simply require defining a larger sampling region which covers all of the cavities. Additionally, the user must be careful when defining the sampling region when using periodic boundary conditions. If the radius is set to encompass any area outside of the box, and periodic boundary conditions are used, there could be overlapping regions in the sampling area and this will result in more water moves being proposed to those areas, creating problems as noted above.

Water hopping could be used to discover important hydration sites in proteins. Crystallography does not always provide an accurate view of water positions and occupancies [168]. Only relatively highly ordered waters can be resolved in crystal structures, which may be a small subset of all waters which are present. Additionally, partial and weak density can obscure determination of where water molecules are present. At the same time, waters can be critical in protein dynamics [194, 231] and for the thermodynamics of ligand binding [230, 8, 161, 160, 144, 7, 172, 23, 125], meaning that treatment of such waters — even when not obvious from experimental data — can be critical. Our method could explore such feasible hydration sites as well as the orientation of critical water molecules in cases where structural data is ambiguous.

4.5 Conclusions and Future Work

In this study, we implemented water hopping moves within our BLUES (NCMC+MD) framework to enhance the sampling of water rearrangements relative to traditional MD for systems that have buried hydration sites.

We validated BLUES with translational water moves on a water box with dividing graphene sheets, a buckyball with an energetically unfavorable water trapped inside, and both the

MUP-1 and HSP90 proteins bound to a ligand with crystallographic bridging water removed. We then evaluate the efficiency of BLUES in hydrating the sites in the protein-ligand systems, based on the number of force evaluations. Overall, we demonstrate that NCMC enhances sampling relative to normal MD.

This water hopping approach can be used to find areas that are likely to be populated by waters in protein binding sites and sample water rearrangements potentially more efficiently than traditional MD. Water hopping moves could be combined with additional types of BLUES moves such as ligand [75, 193] or sidechain [33] rotational moves for broader applications.

The size of the sampling region is an important parameter in our method, and one we intend to optimize in the future. In the future, additional work could be done to help improve the acceptance of water hopping moves. To improve the acceptance and increase the efficiency of BLUES translational water moves, move proposals could be made to be more selective. In the current work, the move proposals can be made anywhere that is encompassed by the radius. To make the move more efficient, water hopping could be redesigned to help reduce move proposals that only move water molecules around in bulk, thus focusing on move proposals to the interior of the protein using methods like those detailed in the work of Ben-Shalom et al. [26]. Additional work could also include comparisons of BLUES (NCMC/MD) water hopping to MC/MD water hopping, allowing us to test whether or not NCMC enhances sampling relative to MC; here, we compared only with traditional MD.

Previous work from Gill et al. compared the speed of non-equilibrium relaxation and MC for ligand rearrangements and found that NCMC provided benefits over doing large numbers of pure MC attempts [75]. We speculate that the same may be true here. Compared to previous work from Ben-Shalom et al. [26], where MC/MD was run on the same MUP-1 protein-ligand system to hydrate the site, we found that BLUES (NCMC/MD) more efficiently hydrates the crystallographic site. There seems to be a 3-4x increase in efficiency using BLUES

(NCCM/MD) based on average number of force evaluations. We believe this increase in efficiency will extend out to other systems, but this needs exploring. Within our current framework, direct comparisons to MC are not feasible (there is a the low acceptance rate and the need to run a large number of trials) because MC evaluations with OpenMM need to be done off-GPU, making the MC move proposals unreasonably slow. This is something that can be explored in future work.

Overall, here, we introduced and validated our new water hopping approach to enhanced sampling of water rearrangements in BLUES, and find it is more efficient than standard MD on a by-force-evaluation basis for the systems considered here.

4.6 Code and Data Availability

The Supporting Information is available free of charge on <https://github.com/MobleyLab/blues-water-hopping-paper> and includes the code, scripts and input files used in this work.

4.7 Acknowledgments

TDB acknowledges support from the ACM SIGHPC/Intel Fellowship. DLM appreciates financial support from the National Institutes of Health (1R01GM108889-01) and the National Science Foundation (CHE 1352608). MKG acknowledges funding from the National Institute of General Medical Sciences (GM61300). The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

4.8 Supplementary Information

4.8.1 Supplementary tables

Table 4.2: **Acceptance ratios of the replicate simulations at different NCMC step amounts for the MUP-1 system.** Each simulation was run for 10,000 BLUES iterations. The number of NCMC steps were varied from 1,250 to 30,000 steps and the number of MD steps was 1,000 steps in all cases

<i>n</i> NCMC steps	Replicate 1	Replicate 2	Replicate 3	Replicate 4	Replicate 5	Replicate 6	Replicate 7	Replicate 8	Replicate 9	Replicate 10	Replicate 11	Replicate 12
1,250	0.001	0.002	0.001	0.001	0.002	0.001	0.001	0.001	0.001	0.002		
2,500	0.004	0.004	0.002	0.002	0.005	0.003	0.003	0.004	0.004	0.003	0.003	
5,000	0.011	0.012	0.009	0.010	0.011	0.010	0.012	0.010	0.011	0.012		
7,500	0.016	0.028	0.020	0.018	0.012	0.024	0.028	0.032	0.032	0.024	0.028	0.012
10,000	0.036	0.028	0.008	0.036	0.020	0.020	0.016	0.036	0.024	0.028	0.028	
15,000	0.012	0.040	0.020	0.048	0.044	0.016	0.028	0.032	0.036	0.040		
20000	0.028	0.040	0.030	0.024	0.044	0.044	0.020	0.016	0.036	0.044	0.032	0.024
30,000	0.036	0.036	0.032	0.020	0.032	0.036	0.016	0.016	0.028	0.024	0.036	0.024

Table 4.3: **Acceptance ratios of all attempted moves for each replicate simulation of the HSP90 protein-ligand system.** Shown is the acceptance ratio for four replicate simulations. It took an average of 1693 BLUES iterations to hydrate the cavity.

n NCMC steps	Replicate 1	Replicate 2	Replicate 3	Replicate 4
2,500	0.004	0.002	0.002	0.002

Table 4.4: **Shown here are the average acceptance rate of all BLUES moves, the average number of force evaluations across 4 replicates for the buried cavity in the HSP90 system to become hydrated, and the average wallclock time in hours for BLUES to hydrate HSP90.** It took an average of 1693 BLUES iterations to hydrate the HSP90 cavity, and each BLUES iteration consisted of a single NCMC move (consisting of 2,500 NCMC steps) and 1,000 MD steps.}

n NCMC steps	Average acceptance rate of all BLUES moves	Average number of force evaluations to hydrate the HSP90 cavity	Average wallclock time to hydrate the HSP90 cavity
2,500	0.2%	5.9×10^6	31 hours

Chapter 5

Progress towards improving host-guest binding free energy calculations by refitting host force field parameters

5.1 Introduction

There are three main areas that the statistical assessment of modelling of proteins and ligands (SAMPL) series of challenges focuses on; physical property challenges, guest-host binding, and protein-ligand binding. This section focuses on guest-host binding. Molecular recognition in host-guest complexes is a particularly simple case of molecular recognition, where the complex is held together through non-covalent bonding.

Hosts have fewer conformations to sample than proteins since they are smaller and have far fewer degrees of freedom, reducing potential error from inadequate conformational sampling. This allows the evaluation of other possible sources of error. It has been found that the results of host-guest binding free energy calculations are very sensitive to torsion parameters, which

can change binding free energy predictions by 3–4 kcal/mol [14].

Recently, the OpenFF group (<https://openforcefield.org/>) has developed the BespokeFit toolkit, which builds custom parameters for individual molecules. Here, I use the OpenFF BespokeFit toolkit to refit the tetra-endomethyl OctaAcid (TEMOA)(Figure 5.1) hosts' torsion force field parameters. Afterward, I run host-guest binding free energy calculations with and without the updated host parameters. I then benchmark the results to see what difference the custom torsions made and how they impacted host-guest binding predictions.

5.2 Methods

The BespokeFit toolkit is a tool that automates the workflow for creating new parameters for individual molecules. More information on the OpenFF BespokeFit toolkit can be found here <https://openff-bespokefit.readthedocs.io/en/latest/> and here <https://github.com/openforcefield/openff-toolkit>.

OpenFF BespokeFit can only refit torsions for small molecules because conformer and charge generation is difficult for larger macrocycles. Therefore BespokeFit was not able to be run on an entire host. Instead, the host of interest (TEMOA) needed to be split into its repeating units, and then BespokeFit was applied to the unit. Afterward, the parameters were applied to the entire host system.

To break TEMOA up, the host was visualized and modified in Chimera (<https://www.rbvi.ucsf.edu/chimera/>). Once the host was cleaved into a repeating unit, the cleaved bonds were capped with methyl groups (so that it would be able to be run through the toolkit). Afterward, BespokeFit was applied to the fragment, and custom parameters were made for the host fragment. To make sure the new parameters correctly mapped back to the whole host molecule, the generated SMIRKS were manually modified and cross-

checked against the host structure. The SMIRKS were also cross-checked against the guest structures (Figure 5.2) to ensure the guest molecules would not be parameterized with any of the custom made parameters that were only meant to be applied to the host structure.

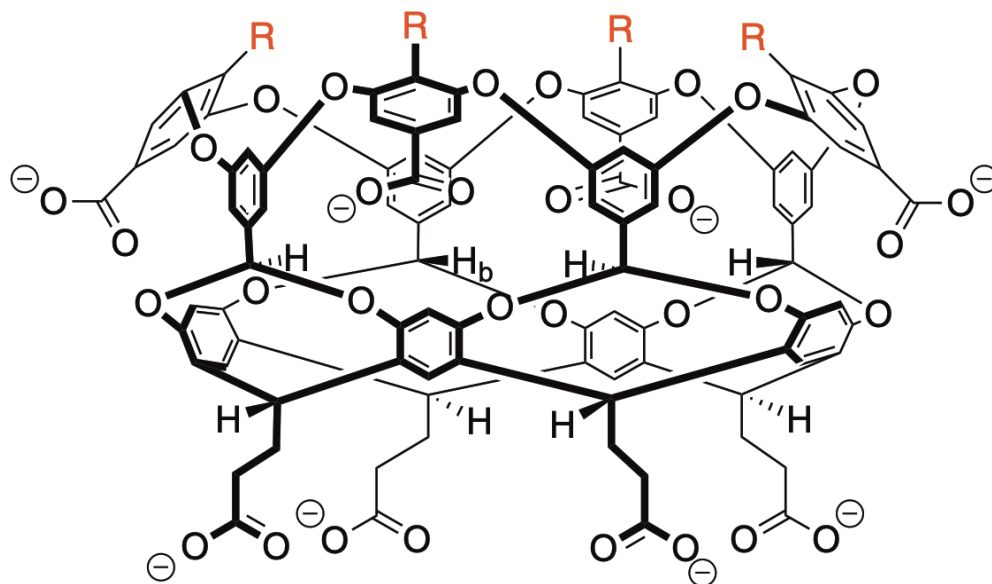
The preparation of host-guest binding free energy calculations and the calculations for running the calculations were carried out as described in [15]. To see how BespokeFit would impact binding free energy results, calculations using a BespokeFit-parameterized host were compared to calculations where the host was parameterized with the OpenFF Parsley (version 1.2.0) force field, which is a general small molecule force field [95]. Specifically, one set of calculations used the Bespoke-parameterized TEMOA with Parsley-parameterized SAMPL8 guest molecules, and the other set of calculation used a Parsley-parameterized TEMOA host and SAMPL8 guest molecules.

Files for the SAMPL8 TEMOA host and five guest files are available here: https://github.com/samplchallenges/SAMPL8/tree/master/host_guest/GDCC.

5.3 Results and Discussion

To evaluate the performance of the custom host parameters on guest-host binding free energy calculations, results were benchmarked against calculations where the host did not have custom torsion parameters.

In term of overall accuracy, the methods don't have a statistically significant difference in RMSE (Figure 5.3), however, the custom parameters do appear to reduce the amount of systematic error in calculations and make predictions that tend to be closer to experimental values on a point-by-point basis compared to the calculations using the Parsley force field, as shown in Figure 5.4.



TEMOA (R = Me)

Figure 5.1: **The structure of the tetra-endomethyl OctaAcid (TEMOA) host.**
 Host files are available here: https://github.com/samplchallenges/SAMPL8/tree/master/host_guest/GDCC.

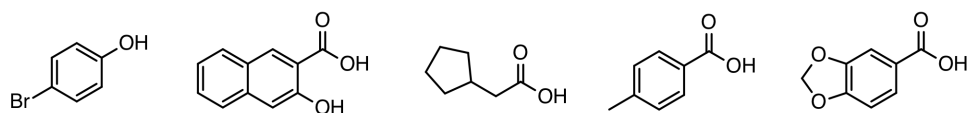


Figure 5.2: **The structure of the five guests used in binding free energy prediction.**
 Files for the guest molecule are available here: https://github.com/samplchallenges/SAMPL8/tree/master/host_guest/GDCC.

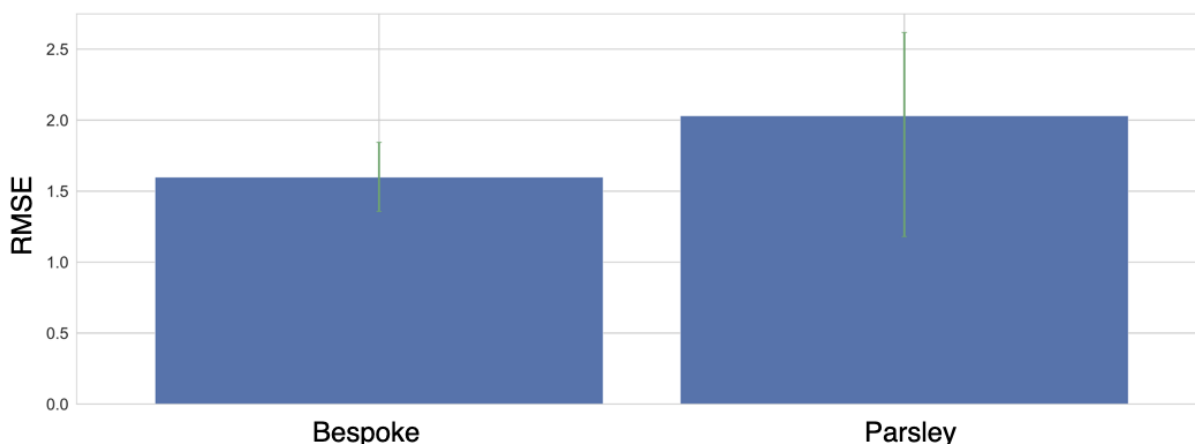


Figure 5.3: **Average prediction error in binding free energy calculations using the new host parameters (Bespoke) and original host parameters (Parsley) show there isn't a statistically significant difference between the two methods.**

Besides the host parameters, everything between the two methods was kept identical (setup, equilibration, production, etc.). Error bars denote 95% confidence intervals obtained by bootstrapping over guest molecules.

One way of looking at the difference between the two methods is by comparing the motion of the host. When the ligand is not inside the host, which happens when the ligand is in the “release” phase of the free energy calculations, it’s been noted that the host can make “breathing motions” where the basket shape of the host collapses. To measure this collapsing, the distance ratio was calculated by measuring the distance between opposite (diagonal) phenyl groups, then dividing the distance between one pair by that between the other pair. A distance ratio less than or greater than 1 indicates the host is exhibiting a breathing or collapsing motion. A value of about 1 means the host is not exhibiting a breathing or collapsed motion and is symmetrical. We find that the BespokeFit-parameterized host forms a collapsed stance less often than the Parsley-parameterized host (as shown in figures 5.6, 5.7, 5.8, and 5.9, suggesting BespokeFit tends to “soften” the breathing motion of the host.

For future work it would be interesting to test BespokeFit on additional hosts to see how the movements of the host are effected, and to see what changes it makes to guest-host binding free energy predictions.

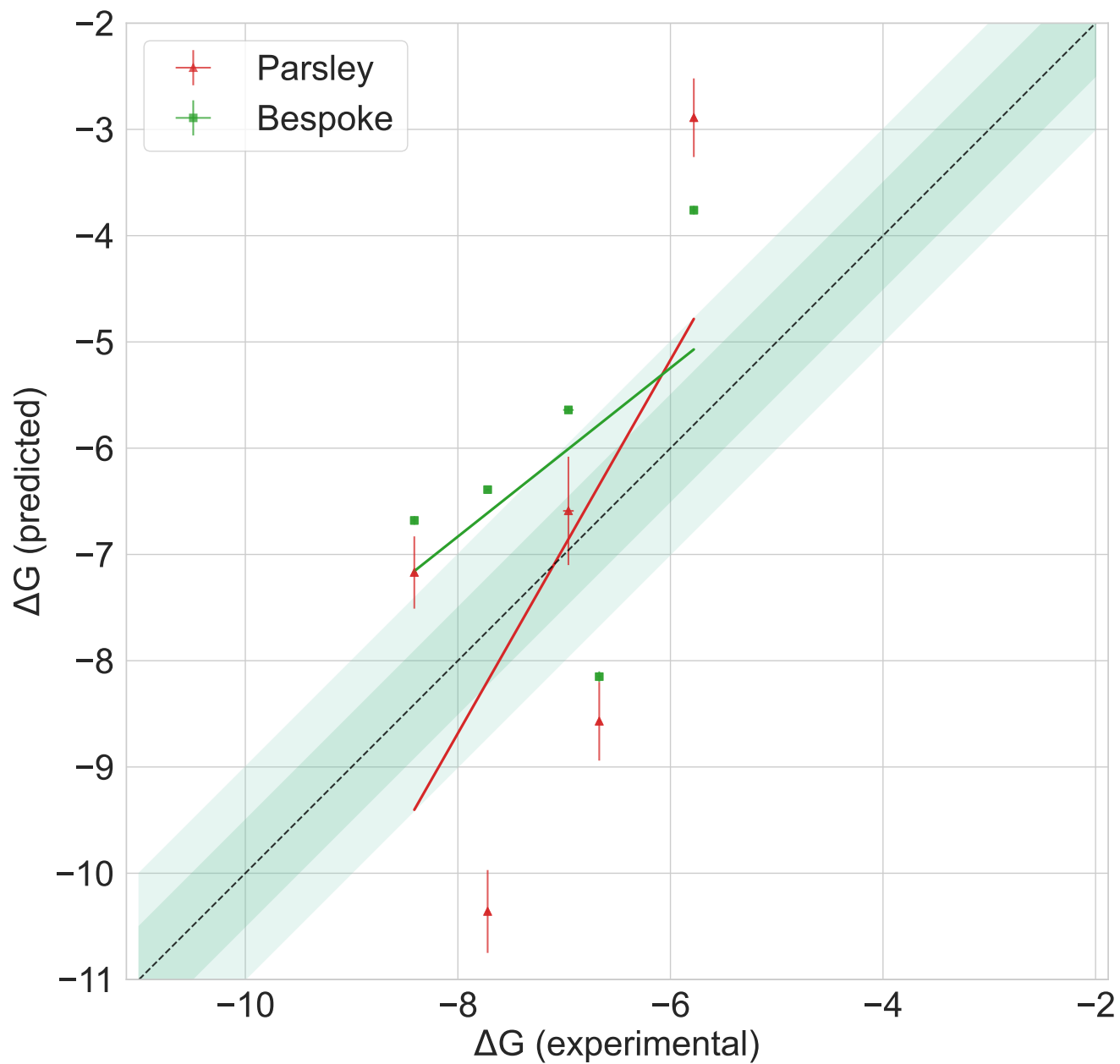


Figure 5.4: **Predicted vs experimental value correlation plots of the two methods.** The set of calculations that used the updated parameters (Bespoke/BespokeFit, green square) seems to have less systematic error than calculations using the original force field parameters for the host (Parsley, red triangle).

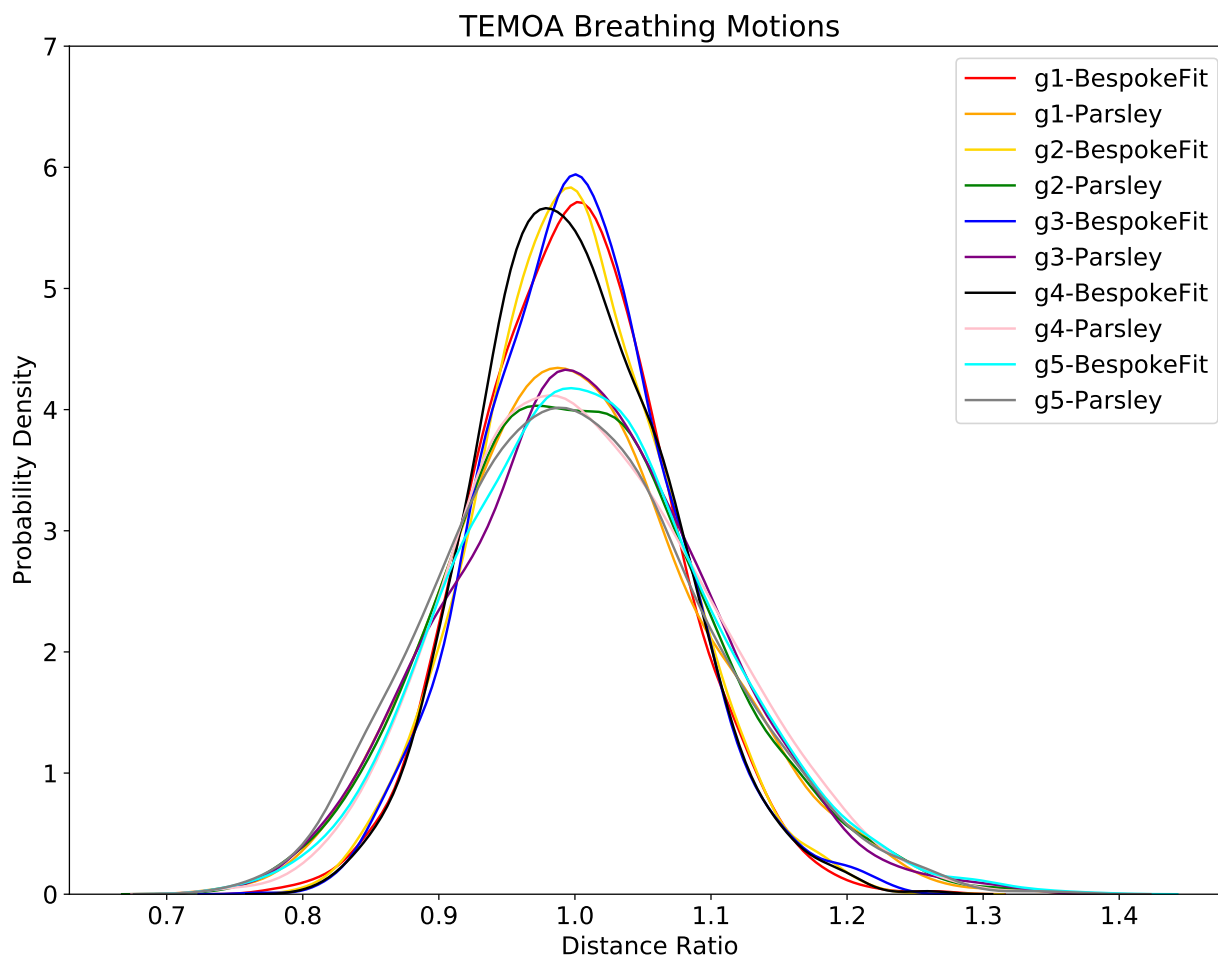


Figure 5.5: The BespokeFit host is less likely to be in the collapsed stance compared to the Parsley host when the ligand is not in the host cavity

Here, the distance ratio is a measure of the collapsing or “the breathing motions” of the host, where the distance of the diagonal upper cavity phenyl groups are measured. The distance ratio was calculated by measuring the distance of opposite (diagonal) phenyl groups, then dividing them. A value around 1 indicates the host does not collapse/remains symmetrical. In other words, it doesn’t exhibit breathing motions. A distance ratio greater than or less than 1 indicates the host collapsed/exhibiting the breathing motion. We find that BespokeFit prevents the breathing motion of the host when the ligand is out of the cavity for all guests, except guest 5 (It’s unknown why the host exhibits breathing motions in this case). The timeseries of the distance ratio of the distances of the diagonal upper cavity phenyl groups for the BespokeFit-parameterized and Parsley-parameterized host are available in figures 5.6, 5.7, 5.8, 5.9, and 5.10.

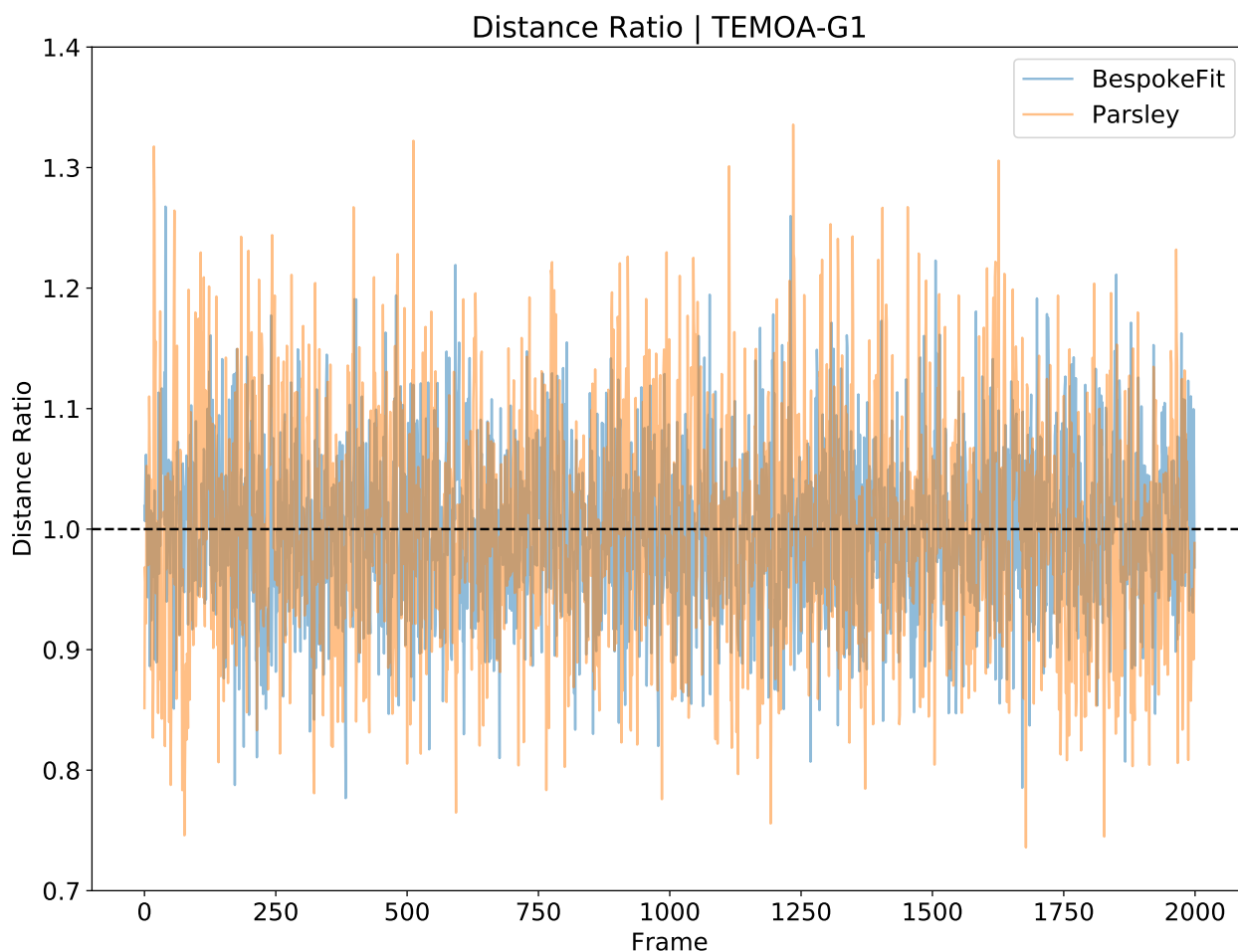


Figure 5.6: Shown is the timeseries of the distance ratio of the distances of the diagonal upper cavity phenyl groups for the BespokeFit-parameterized host (blue) and the Parsley-parameterized host (orange) when guest number # 1 is not in the binding pocket.

The Parsley-parameterized host shows a distance ratio that indicates the host frequents a collapsed position (a distance ratio around 1.2–1.3 and 0.7–0.8) more often than the BespokeFit-parameterized host. The distance ratio between the phenyl groups was calculated for each frame in the trajectory of the “release” phase of the free energy calculations.

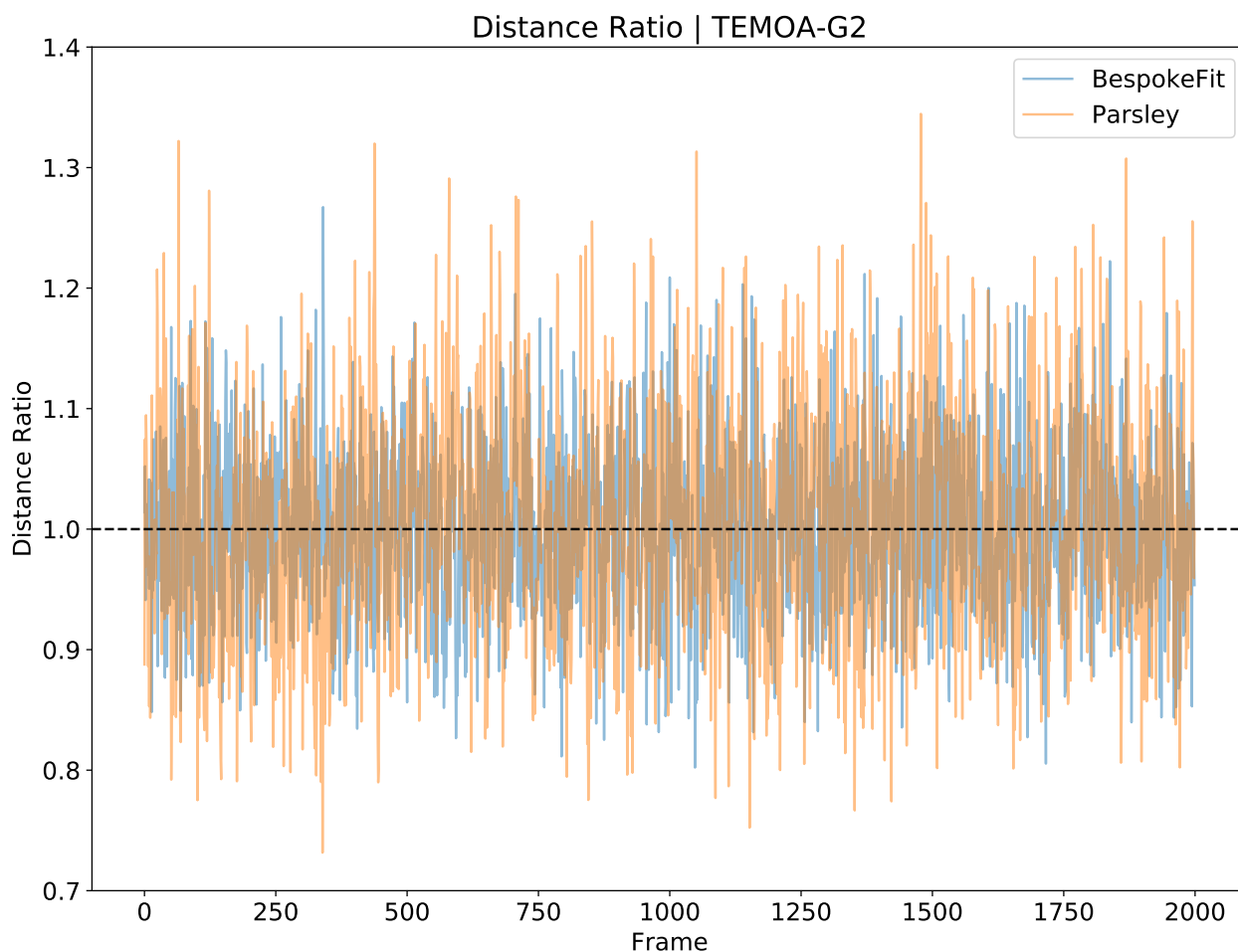


Figure 5.7: Shown is the timeseries of the distance ratio of the distances of the diagonal upper cavity phenyl groups for the BespokeFit-parameterized host (blue) and the Parsley-parameterized host (orange) when guest number # 2 is not in the binding pocket.

The Parsley-parameterized host shows a distance ratio that indicates the host frequents a collapsed position (a distance ratio around 1.2–1.3 and 0.7–0.8) more often than the BespokeFit-parameterized host. The distance ratio between the phenyl groups was calculated for each frame in the trajectory of the “release” phase of the free energy calculations.

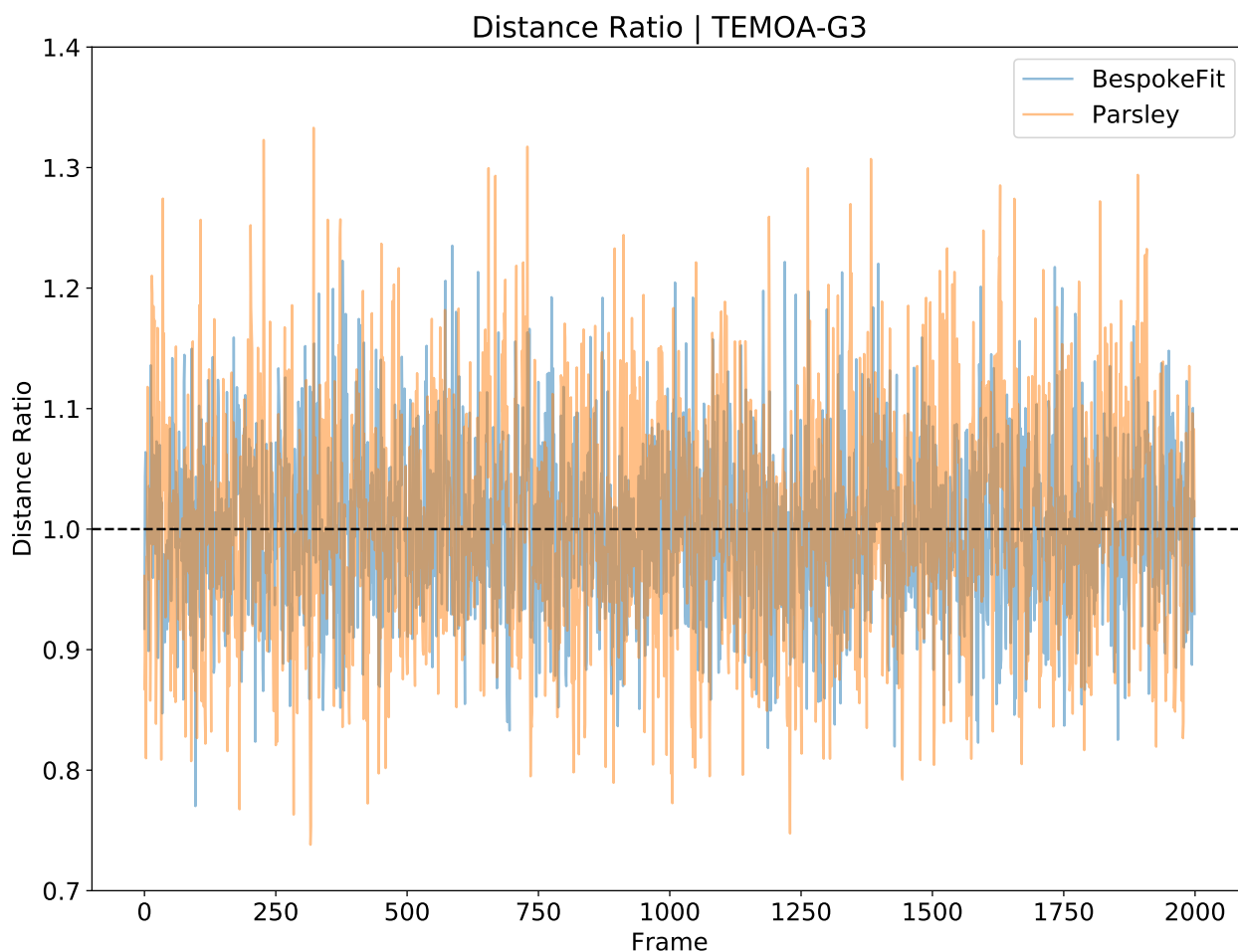


Figure 5.8: Shown is the timeseries of the distance ratio of the distances of the diagonal upper cavity phenyl groups for the BespokeFit-parameterized host (blue) and the Parsley-parameterized host (orange) when guest number # 3 is not in the binding pocket.

The Parsley-parameterized host shows a distance ratio that indicates the host frequents a collapsed position (a distance ratio around 1.2–1.4 and 0.7–0.9) more often than the BespokeFit-parameterized host. The distance ratio between the phenyl groups was calculated for each frame in the trajectory of the “release” phase of the free energy calculations.

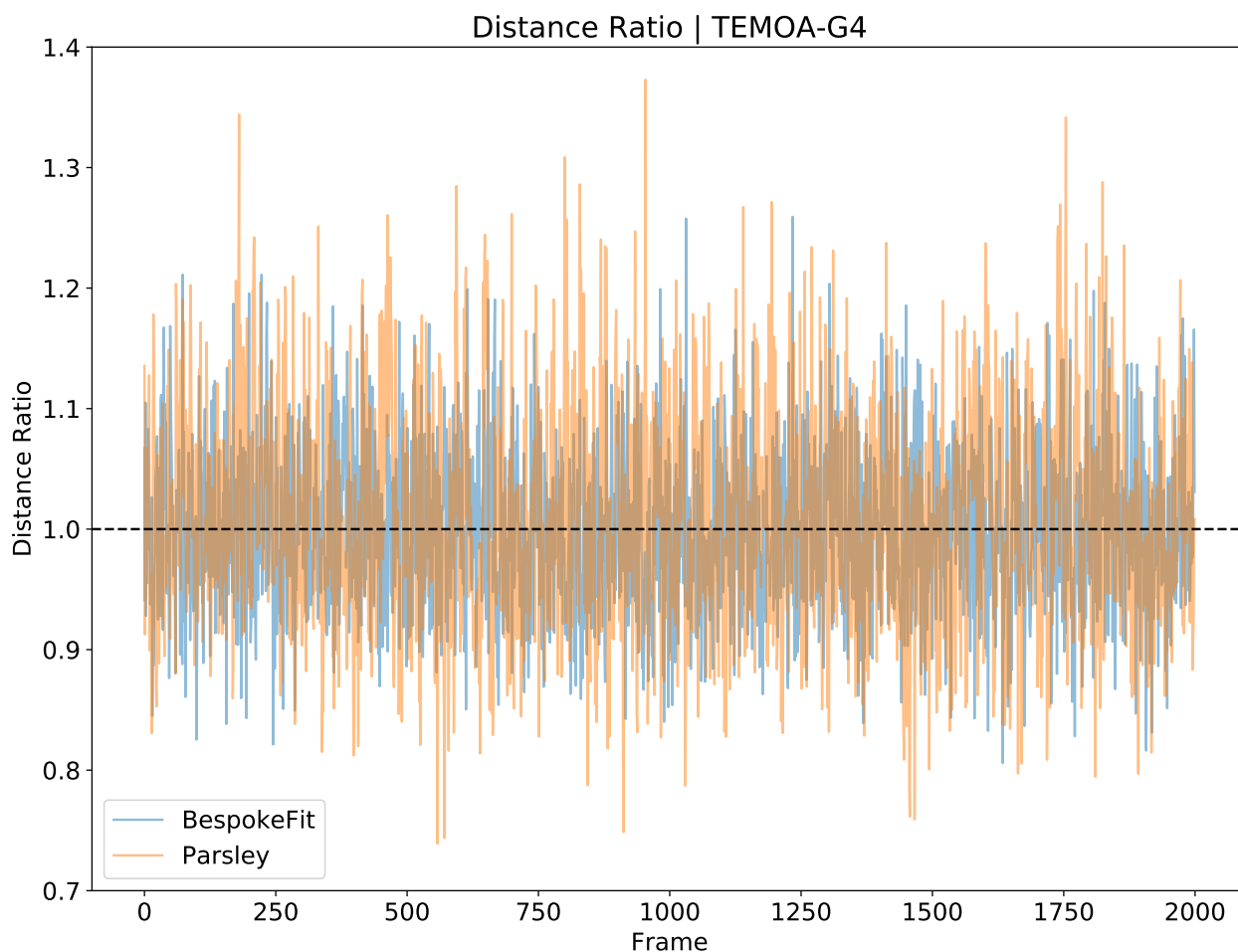


Figure 5.9: Shown is the timeseries of the distance ratio of the distances of the diagonal upper cavity phenyl groups for the BespokeFit-parameterized host (blue) and the Parsley-parameterized host (orange) when guest number # 4 is not in the binding pocket.

The Parsley-parameterized host shows a distance ratio that indicates the host frequents a collapsed position (a distance ratio around 1.1–1.4 and 0.7–0.8) more often than the BespokeFit-parameterized host. The distance ratio between the phenyl groups was calculated for each frame in the trajectory of the “release” phase of the free energy calculations.

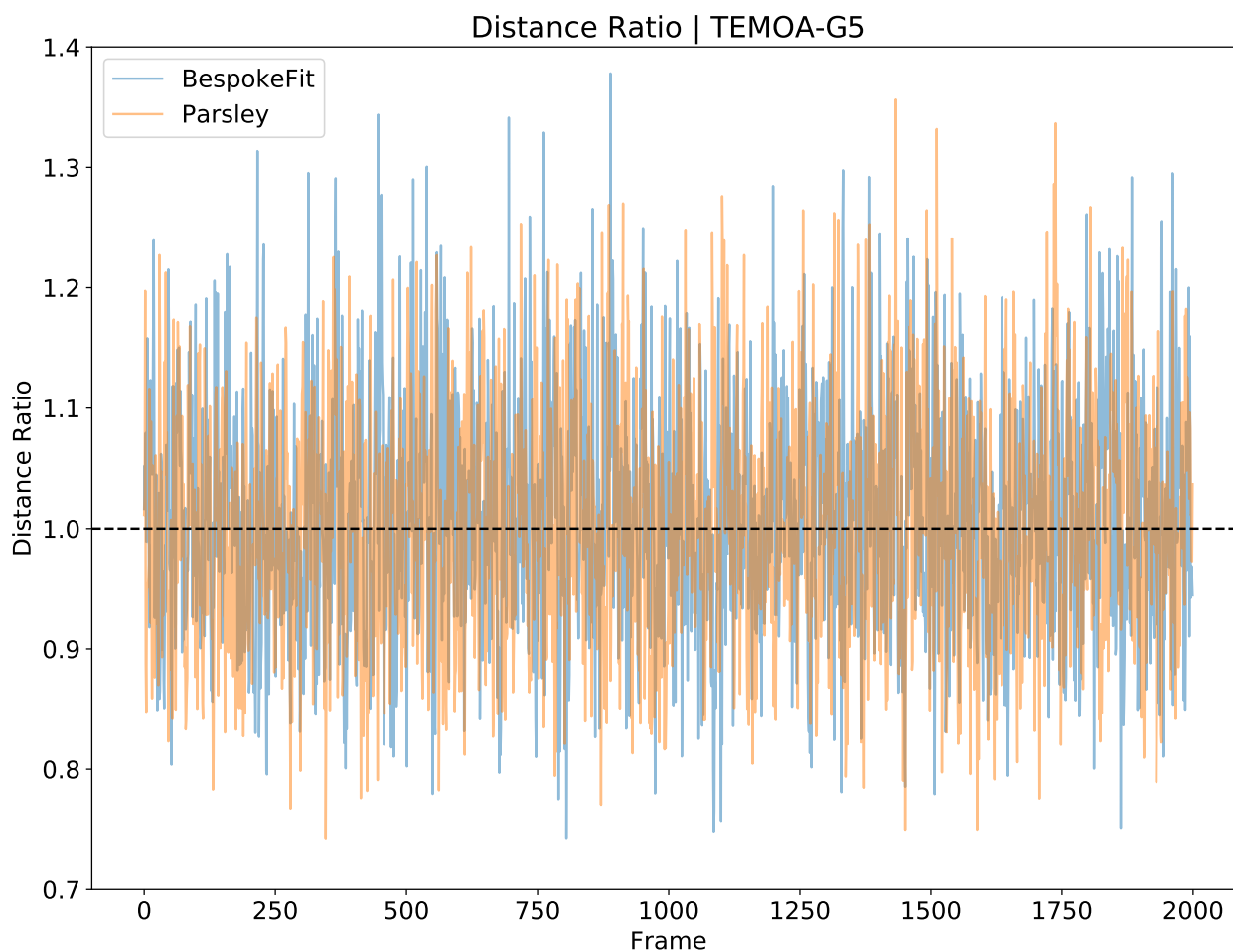


Figure 5.10: Shown is the timeseries of the distance ratio of the distances of the diagonal upper cavity phenyl groups for the BespokeFit-parameterized host (blue) and the Parsley-parameterized host (orange) when guest number # 5 is not in the binding pocket.

The BespokeFit-parameterized host shows a distance ratio that indicates the host frequents a collapsed position about as frequently as the Parsley-parameterized host. It is unknown why the distance ratio differs from the distance ratio's seen in the timeseries in figures 5.6, 5.7, 5.8, and 5.9. The distance ratio between the phenyl groups was calculated for each frame in the trajectory of the “release” phase of the free energy calculations.

Bibliography

- [1] Acd/pka classic (acd/percepta kernel v1.6);. advanced chem-istry development, inc., toronto, on, canada, 2018. <https://www.acdlabs.com/products/percepta/predictors/pKa/>.
- [2] Chemicalize toolkit: Property and structure calculator, accessed 2020. developed by chemaxon. <https://chemicalize.com/>.
- [3] Enhanced nci database browser 2.2. <https://cactus.nci.nih.gov/ncidb2.2/>.
- [4] Nci open database, august 2006 release. <https://cactus.nci.nih.gov/download/nci/>.
- [5] Oedepict toolkit 2017.feb.1. <http://www.eyesopen.com>. OpenEye Scientific Software, Santa Fe, NM.
- [6] Src's physprop database. <https://www.srcinc.com/what-we-do/environmental/scientific-databases.html>.
- [7] R. Abel, N. K. Salam, J. Shelley, R. Farid, R. A. Friesner, and W. Sherman. Contribution of Explicit Solvent Effects to the Binding Affinity of Small-Molecule Inhibitors in Blood Coagulation Factor Serine Proteases. *ChemMedChem*, 6(6):1049–1066, June 2011.
- [8] R. Abel, T. Young, R. Farid, B. J. Berne, and R. A. Friesner. Role of the Active-Site Solvent in the Thermodynamics of Factor Xa Ligand Binding. *J. Am. Chem. Soc.*, 130(9):2817–2831, Mar. 2008.
- [9] D. Adams. Chemical potential of hard-sphere fluids by Monte Carlo methods. *Molecular Physics*, 28(5):1241–1252, Nov. 1974.
- [10] D. Adams. Grand canonical ensemble Monte Carlo for a Lennard-Jones fluid. *Molecular Physics*, 29(1):307–311, Jan. 1975.
- [11] B. Aguilar, R. Anandakrishnan, J. Z. Ruscio, and A. V. Onufriev. Statistics and Physical Origins of pK and Ionization State Changes upon Protein-Ligand Binding. *Biophysical Journal*, 98(5):872–880, Mar. 2010.

- [12] K. S. Alongi and G. C. Shields. Theoretical Calculations of Acid Dissociation Constants: A Review Article. In *Annual Reports in Computational Chemistry*, volume 6, pages 113–138. Elsevier, 2010.
- [13] M. Amaral, D. B. Kokh, J. Bomke, A. Wegener, H. P. Buchstaller, H. M. Eggenweiler, P. Matias, C. Sirrenberg, R. C. Wade, and M. Frech. Protein conformational flexibility modulates kinetics and thermodynamics of drug binding. *Nat Commun*, 8(1):2276, Dec. 2017.
- [14] M. Amezcua, L. El Khoury, and D. L. Mobley. SAMPL7 Host–Guest Challenge Overview: Assessing the reliability of polarizable and non-polarizable methods for binding free energy calculations. *Journal of Computer-Aided Molecular Design*, 35(1):1–35, Jan. 2021.
- [15] M. Amezcua, J. Setiadi, Y. Ge, and D. Mobley. An Overview of the SAMPL8 Host–Guest Binding Challenge. Preprint, Chemistry, Apr. 2022.
- [16] A. Andrea, P. Grinaway, D. Parton, M. Shirts, K. Wang, P. Eastman, M. Friedrichs, V. Pande, K. Branson, D. Mobley, and J. Chodera. YANK: A GPU-accelerated platform for alchemical free energy calculations.
- [17] S. G. Balasubramani, G. P. Chen, S. Coriani, M. Diedenhofen, M. S. Frank, Y. J. Franzke, F. Furche, R. Grotjahn, M. E. Harding, C. Hättig, A. Hellweg, B. Helmich-Paris, C. Holzer, U. Huniar, M. Kaupp, A. Marefat Khah, S. Karbalaeei Khani, T. Müller, F. Mack, B. D. Nguyen, S. M. Parker, E. Perlt, D. Rappoport, K. Reiter, S. Roy, M. Rückert, G. Schmitz, M. Sierka, E. Tapavicza, D. P. Tew, C. van Wüllen, V. K. Voora, F. Weigend, A. Wodyński, and J. M. Yu. TURBOMOLE: Modular program suite for *ab initio* quantum-chemical and condensed-matter simulations. *J. Chem. Phys.*, 152(18):184107, May 2020.
- [18] P. Ball. Water as an Active Constituent in Cell Biology. *Chem. Rev.*, 108(1):74–108, Jan. 2008.
- [19] C. C. Bannan, K. H. Burley, M. Chiu, M. R. Shirts, M. K. Gilson, and D. L. Mobley. Blind prediction of cyclohexane–water distribution coefficients from the SAMPL5 challenge. *Journal of Computer-Aided Molecular Design*, 30(11):927–944, Nov. 2016.
- [20] C. C. Bannan, G. Calabró, D. Y. Kyu, and D. L. Mobley. Calculating Partition Coefficients of Small Molecules in Octanol/Water and Cyclohexane/Water. *Journal of Chemical Theory and Computation*, 12(8):4015–4024, Aug. 2016.
- [21] C. C. Bannan, D. L. Mobley, and A. G. Skillman. SAMPL6 challenge results from pka predictions based on a general Gaussian process model. *Journal of Computer-Aided Molecular Design*, 32(10):1165–1177, Oct. 2018.
- [22] R. Baron, P. Setny, and J. A. McCammon. Water in Cavity-Ligand Recognition. *J. Am. Chem. Soc.*, 132(34):12091–12097, Sept. 2010.

- [23] A. S. Bayden, D. T. Moustakas, D. Joseph-McCarthy, and M. L. Lamb. Evaluating Free Energies of Binding and Conservation of Crystallographic Waters Using SZMAP. *J. Chem. Inf. Model.*, 55(8):1552–1565, Aug. 2015.
- [24] D. Beglov and B. Roux. An Integral Equation To Describe the Solvation of Polar Molecules in Liquid Water. *J. Phys. Chem. B*, 101(39):7821–7826, Sept. 1997.
- [25] M.-C. Bellissent-Funel, A. Hassanali, M. Havenith, R. Henchman, P. Pohl, F. Sterpone, D. van der Spoel, Y. Xu, and A. E. Garcia. Water Determines the Structure and Dynamics of Proteins. *Chem. Rev.*, 116(13):7673–7697, July 2016.
- [26] I. Y. Ben-Shalom, C. Lin, T. Kurtzman, R. C. Walker, and M. K. Gilson. Simulating Water Exchange to Buried Binding Sites. *Journal of Chemical Theory and Computation*, 15(4):2684–2691, Apr. 2019.
- [27] C. H. Bennett. Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics*, 22(2):245–268, Oct. 1976.
- [28] M. L. Benson, J. C. Faver, M. N. Ucisik, D. S. Dashti, Z. Zheng, and K. M. Merz. Prediction of trypsin/molecular fragment binding affinities by free energy decomposition and empirical scores. *Journal of Computer-Aided Molecular Design*, 26(5):647–659, May 2012.
- [29] S. A. Best, K. M. Merz, and C. H. Reynolds. Free Energy Perturbation Study of Octanol/Water Partition Coefficients: Comparison with Continuum GB/SA Calculations. *The Journal of Physical Chemistry B*, 103(4):714–726, Jan. 1999.
- [30] A. D. Bochevarov, M. A. Watson, J. R. Greenwood, and D. M. Philipp. Multiconformation, Density Functional Theory-Based pK_a Prediction in Application to Large, Flexible Organic Molecules with Diverse Functional Groups. *Journal of Chemical Theory and Computation*, 12(12):6001–6019, Dec. 2016.
- [31] T. N. Brown and N. Mora-Diez. Computational Determination of Aqueous pK_a Values of Protonated Benzimidazoles (Part 2). *J. Phys. Chem. B*, 110(41):20546–20554, Oct. 2006.
- [32] P. Bultinck, C. Van Alsenoy, P. W. Ayers, and R. Carbo-Dorca. Critical analysis and extension of the hirshfeld atoms in molecules. *The Journal of Chemical Physics*, 126(14):144111, 2007.
- [33] K. H. Burley, S. C. Gill, N. M. Lim, and D. L. Mobley. Enhancing Side Chain Rotamer Sampling Using Nonequilibrium Candidate Monte Carlo. *Journal of Chemical Theory and Computation*, 15(3):1848–1862, Mar. 2019.
- [34] D. Case, J. Berryman, R. Betz, D. Cerutti, T. Cheatham Iii, T. Darden, R. Duke, T. Giese, H. Gohlke, A. Goetz, et al. AMBER 2015. *University of California, San Francisco*, 2015.

- [35] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods. The Amber biomolecular simulation programs. *J. Comput. Chem.*, 26(16):1668–1688, Dec. 2005.
- [36] P. S. Charifson and W. P. Walters. Acidic and Basic Drugs in Medicinal Chemistry: A Perspective. *J. Med. Chem.*, 57(23):9701–9717, Dec. 2014.
- [37] Chemprop: Directed message passing neural network. <https://chemprop.readthedocs.io/en/latest/>.
- [38] B. Chen and J. I. Siepmann. Microscopic Structure and Solvation in Dry and Wet Octanol. *The Journal of Physical Chemistry B*, 110(8):3555–3563, Mar. 2006.
- [39] T. Cheng, Y. Zhao, X. Li, F. Lin, Y. Xu, X. Zhang, Y. Li, R. Wang, and L. Lai. Computation of Octanol-Water Partition Coefficients by Guiding an Additive Model with Knowledge. *Journal of Chemical Information and Modeling*, 47(6):2140–2148, Nov. 2007.
- [40] J. Comer, J. C. Gumbart, J. Hénin, T. Lelièvre, A. Pohorille, and C. Chipot. The Adaptive Biasing Force Method: Everything You Always Wanted To Know but Were Afraid To Ask. *The Journal of Physical Chemistry B*, 119(3):1129–1151, Jan. 2015.
- [41] Cosmoconf: A flexible conformer generator for cosmo-rs. <https://www.3ds.com/products-services/biovia/products/molecular-modeling-simulation/solvation-chemistry/cosmoconf/>.
- [42] Cosmoquick: Cosmo-rs based toolbox. <https://www.3ds.com/products-services/biovia/products/molecular-modeling-simulation/solvation-chemistry/cosmoquick/>.
- [43] Biovia cosmotherm: Tool for predictive property calculation of liquids. version 2020. dassault systemes. <https://www.3ds.com/products-services/biovia/products/molecular-modeling-simulation/solvation-chemistry/cosmotherm/>.
- [44] Z. Cournia, B. Allen, and W. Sherman. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *Journal of Chemical Information and Modeling*, 57(12):2911–2937, Dec. 2017.
- [45] L. O. K. T.-N. G. A. D. Eros, I. Kovcsdi and G. Keri. Reliability of logp predictions based on calculated molecular descriptors: A critical review. *Current Medicinal Chemistry*, 9(20):1819–1829, 2002.
- [46] T. Darden, D. York, and L. Pedersen. Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089–10092, June 1993.
- [47] E. Darve and A. Pohorille. Calculating free energies using average force. *The Journal of Chemical Physics*, 115(20):9169–9183, 2001.

- [48] S. E. DeBolt and P. A. Kollman. Investigation of Structure, Dynamics, and Solvation in 1-Octanol and Its Water-Saturated Solution: Molecular Dynamics and Free-Energy Perturbation Studies. *J. Am. Chem. Soc.*, 117(19):5316–5340, May 1995.
- [49] N. Deng, S. Forli, P. He, A. Perryman, L. Wickstrom, R. S. K. Vijayan, T. Tiefenbrunn, D. Stout, E. Gallicchio, A. J. Olson, and R. M. Levy. Distinguishing Binders from False Positives by Free Energy Calculations: Fragment Screening Against the Flap Site of HIV Protease. *J. Phys. Chem. B*, 119(3):976–988, Jan. 2015.
- [50] Y. Deng and B. Roux. Computation of binding free energy with molecular dynamics and grand canonical Monte Carlo simulations. *The Journal of Chemical Physics*, 128(11):115103, Mar. 2008.
- [51] L. S. Dodda, I. Cabeza de Vaca, J. Tirado-Rives, and W. L. Jorgensen. LigParGen web server: An automatic OPLS-AA parameter generator for organic ligands. *Nucleic Acids Research*, 45(W1):W331–W336, July 2017.
- [52] N. Donyapour and A. Dickson. Predicting partition coefficients for the sampl7 physical property challenge using the classicalgsg method. *Journal of Computer-Aided Molecular Design*, 2021.
- [53] N. Donyapour, M. Hirn, and A. Dickson. ClassicalGSG: Prediction of log P using classical molecular force fields and geometric scattering for graphs. *Journal of Computational Chemistry*, 42(14):1006–1017, 2021.
- [54] Drugbank: Online database of drug and drug target information. <https://www.drugbank.com/>.
- [55] P. Eastman, M. S. Friedrichs, J. D. Chodera, R. J. Radmer, C. M. Bruns, J. P. Ku, K. A. Beauchamp, T. J. Lane, L.-P. Wang, D. Shukla, T. Tye, M. Houston, T. Stich, C. Klein, M. R. Shirts, and V. S. Pande. OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *Journal of Chemical Theory and Computation*, 9(1):461–469, Jan. 2013.
- [56] P. Eastman and V. Pande. OpenMM: A Hardware-Independent Framework for Molecular Simulations. *Comput. Sci. Eng.*, 12(4):34–39, July 2010.
- [57] P. Eastman and V. S. Pande. Efficient nonbonded interactions for molecular dynamics on a graphics processing unit. *Journal of Computational Chemistry*, pages NA–NA, 2009.
- [58] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, and V. S. Pande. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Computational Biology*, 13(7):e1005659, July 2017.
- [59] J. Ernst, R. Clubb, H. Zhou, A. Gronenborn, and G. Clore. Demonstration of positionally disordered water within a protein hydrophobic cavity by NMR. *Science*, 267(5205):1813–1817, Mar. 1995.

- [60] F. Falcioni, J. Kalayan, and R. Henchman. Energy-entropy prediction of octanol-water $\log P$ of sampl7 n-acyl sulfonamidebioisomers. *Journal of Computer-Aided Molecular Design*, 2021.
- [61] S. Fan, B. I. Iorga, and O. Beckstein. Prediction of octanol-water partition coefficients for the SAMPL6- $\log P$ molecules using molecular dynamics simulations with OPLS-AA, AMBER and CHARMM force fields. *Journal of Computer-Aided Molecular Design*, 34(5):543–560, May 2020.
- [62] S. Fan, H. Nedev, R. Vijayan, B. I. Iorga, and O. Beckstein. Precise force-field-based calculations of octanol-water partition coefficients for the sampl7 molecules. *Journal of Computer-Aided Molecular Design*, 2021.
- [63] R. Fraczkiwicz. In Silico Prediction of Ionization. In *Comprehensive Medicinal Chemistry II*, pages 603–626. Elsevier, 2007.
- [64] K. R. Francisco, C. Varricchio, T. J. Paniak, M. C. Kozlowski, A. Brancale, and C. Ballatore. Structure Property Relationships of N-Acylsulfonamides and Related Bioisosteres. *European Journal of Medicinal Chemistry*, page 113399, Mar. 2021.
- [65] Frog v2.14: Free on line drug conformation generation. <https://bioserv.rpbs.univ-paris-diderot.fr/services/Frog2/>.
- [66] B. K. Findik, Z. P. Haslak, E. Arslan, and V. Aviyente. Sampl7 blind challenge: Quantum-mechanical prediction of partition coefficients and acid dissociation constants for small drug-like molecules. *Journal of Computer-Aided Molecular Design*, 2021.
- [67] E. Gallicchio, N. Deng, P. He, L. Wickstrom, A. L. Perryman, D. N. Santiago, S. Forli, A. J. Olson, and R. M. Levy. Virtual screening of integrase inhibitors by large scale binding free energy calculations: The SAMPL4 challenge. *Journal of Computer-Aided Molecular Design*, 28(4):475–490, Apr. 2014.
- [68] F. Gao, G. Wolf, and M. Hirn. Geometric Scattering for Graph Data Analysis. In *International Conference on Machine Learning*, pages 2122–2131. PMLR, May 2019.
- [69] M. T. Geballe, A. G. Skillman, A. Nicholls, J. P. Guthrie, and P. J. Taylor. The SAMPL2 blind prediction challenge: Introduction and overview. *Journal of Computer-Aided Molecular Design*, 24(4):259–279, Apr. 2010.
- [70] P. R. Gerber. Charge distribution from a simple molecular orbital type calculation and non-bonding interaction terms in the force field MAB. *Journal of Computer-Aided Molecular Design*, 12(1):37–51, Jan. 1998.
- [71] A. K. Ghose and G. M. Crippen. Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships I. Partition Coefficients as a Measure of Hydrophobicity. *Journal of Computational Chemistry*, 7(4):565–577, 1986.

- [72] A. K. Ghose and G. M. Crippen. Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions. *J. Chem. Inf. Comput. Sci.*, 27(1):21–35, Feb. 1987.
- [73] A. K. Ghose, V. N. Viswanadhan, and J. J. Wendoloski. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods. *The Journal of Physical Chemistry A*, 102(21):3762–3772, May 1998.
- [74] C. Giaginis and A. Tsantili-Kakoulidou. Alternative Measures of Lipophilicity: From Octanol–Water Partitioning to IAM Retention. *Journal of Pharmaceutical Sciences*, 97(8):2984–3004, Aug. 2008.
- [75] S. C. Gill, N. M. Lim, P. B. Grinaway, A. S. Rustenburg, J. Fass, G. A. Ross, J. D. Chodera, and D. L. Mobley. Binding Modes of Ligands Using Enhanced Sampling (BLUES): Rapid Decorrelation of Ligand Binding Modes via Nonequilibrium Candidate Monte Carlo. *The Journal of Physical Chemistry B*, Feb. 2018.
- [76] A. Glomme, J. März, and J. Dressman. Comparison of a Miniaturized Shake-Flask Solubility Method with Automated Potentiometric Acid/Base Titrations and Calculated Solubilities. *Journal of Pharmaceutical Sciences*, 94(1):1–16, Jan. 2005.
- [77] V. K. Gombar and K. Enslein. Assessment of n-Octanol/Water Partition Coefficient: When Is the Assessment Reliable? *J. Chem. Inf. Comput. Sci.*, 36(6):1127–1134, Jan. 1996.
- [78] J. R. Greenwood, D. Calkins, A. P. Sullivan, and J. C. Shelley. Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. *Journal of Computer-Aided Molecular Design*, 24(6-7):591–604, June 2010.
- [79] D. Guan, R. Lui, and S. Matthews. LogP prediction performance with the SMD solvation model and the M06 density functional family for SAMPL6 blind prediction challenge molecules. *Journal of Computer-Aided Molecular Design*, Jan. 2020.
- [80] M. R. Gunner, T. Murakami, A. S. Rustenburg, M. Işık, and J. D. Chodera. Standard state free energies, not pK_as, are ideal for describing small molecule protonation and tautomeric states. *Journal of Computer-Aided Molecular Design*, 34(5):561–573, May 2020.
- [81] J. P. Guthrie. A Blind Challenge for Computational Solvation Free Energies: Introduction and Overview. *J. Phys. Chem. B*, 113(14):4501–4507, Apr. 2009.
- [82] W. K. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57:97, 1970.

- [83] H. Hogues, T. Sulea, and E. O. Purisima. Exhaustive docking and solvated interaction energy scoring: Lessons learned from the SAMPL4 challenge. *Journal of Computer-Aided Molecular Design*, 28(4):417–427, Apr. 2014.
- [84] C. W. Hopkins, S. Le Grand, R. C. Walker, and A. E. Roitberg. Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *Journal of Chemical Theory and Computation*, 11(4):1864–1874, Apr. 2015.
- [85] C. W. Hopkins, S. Le Grand, R. C. Walker, and A. E. Roitberg. Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *Journal of Chemical Theory and Computation*, 11(4):1864–1874, Apr. 2015.
- [86] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins*, 65(3):712–725, Nov. 2006.
- [87] M. Işık, T. D. Bergazin, T. Fox, A. Rizzi, J. D. Chodera, and D. L. Mobley. Assessing the accuracy of octanol–water partition coefficient predictions in the SAMPL6 Part II log P Challenge. *Journal of Computer-Aided Molecular Design*, 34(4):335–370, Apr. 2020.
- [88] M. Işık, D. Levorse, D. L. Mobley, T. Rhodes, and J. D. Chodera. Octanol–water partition coefficient measurements for the SAMPL6 blind prediction challenge. *Journal of Computer-Aided Molecular Design*, Dec. 2019.
- [89] M. Işık, D. Levorse, A. S. Rustenburg, I. E. Ndukwe, H. Wang, X. Wang, M. Reibarkh, G. E. Martin, A. A. Makarov, D. L. Mobley, T. Rhodes, and J. D. Chodera. pKa measurements for the SAMPL6 prediction challenge for a set of kinase inhibitor-like fragments. *Journal of Computer-Aided Molecular Design*, 32(10):1117–1138, Oct. 2018.
- [90] M. Işık, D. Levorse, A. S. Rustenburg, I. E. Ndukwe, H. Wang, X. Wang, M. Reibarkh, G. E. Martin, A. A. Makarov, D. L. Mobley, T. Rhodes, and J. D. Chodera. pKa measurements for the SAMPL6 prediction challenge for a set of kinase inhibitor-like fragments. *Journal of Computer-Aided Molecular Design*, 32(10):1117–1138, Oct. 2018.
- [91] M. Işık, A. S. Rustenburg, A. Rizzi, M. R. Gunner, D. L. Mobley, and J. D. Chodera. Overview of the SAMPL6 pKa challenge: Evaluating small molecule microscopic and macroscopic pKa predictions. *Journal of Computer-Aided Molecular Design*, 35(2):131–166, Feb. 2021.
- [92] S. Izadi, R. Anandakrishnan, and A. V. Onufriev. Building Water Models: A Different Approach. *The Journal of Physical Chemistry Letters*, 5(21):3863–3871, Nov. 2014.
- [93] S. Izadi and A. V. Onufriev. Accuracy limit of rigid 3-point water models. *The Journal of Chemical Physics*, 145(7):074501, Aug. 2016.
- [94] M. Işık. SAMPL6 Part II Partition Coefficient Challenge Overview, Sept. 2019.

- [95] H. Jang, J. Maat, Y. Qiu, D. G. Smith, S. Boothroyd, J. Wagner, C. C. Bannan, T. Gokey, V. T. Lim, X. Lucas, B. Tjanaka, M. R. Shirts, M. K. Gilson, J. D. Chodera, C. I. Bayly, D. L. Mobley, and L.-P. Wang. openforcefield/openforcefields: Version 1.2.0 "Parsley" Update, June 2020.
- [96] C. Jarzynski. Nonequilibrium Equality for Free Energy Differences. *Physical Review Letters*, 78(14):2690–2693, Apr. 1997.
- [97] M. R. Jones and B. R. Brooks. Quantum chemical predictions of water–octanol partition coefficients applied to the SAMPL6 logP blind challenge. *Journal of Computer-Aided Molecular Design*, 34(5):485–493, May 2020.
- [98] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935, July 1983.
- [99] G. Kamath, I. Kurnikov, B. Fain, I. Leontyev, A. Illarionov, O. Butin, M. Olevanov, and L. Pereyaslavets. Prediction of cyclohexane-water distribution coefficient for SAMPL5 drug-like compounds with the QMPFF3 and ARROW polarizable force fields. *Journal of Computer-Aided Molecular Design*, 30(11):977–988, Nov. 2016.
- [100] Y. Khalak, G. Tresadern, B. L. de Groot, and V. Gapsys. Non-equilibrium approach for binding free energies in cyclodextrins in SAMPL7: Force fields and software. *Journal of Computer-Aided Molecular Design*, 35(1):49–61, Jan. 2021.
- [101] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, and E. E. Bolton. PubChem in 2021: New data content and improved web interfaces. *Nucleic Acids Research*, 49(D1):D1388–D1395, Jan. 2021.
- [102] J. G. Kirkwood. Statistical Mechanics of Fluid Mixtures. *The Journal of Chemical Physics*, 3(5):300–313, May 1935.
- [103] A. Klamt. Conductor-like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena. *The Journal of Physical Chemistry*, 99(7):2224–2235, Feb. 1995.
- [104] A. Klamt, F. Eckert, and M. Diedenhofen. Prediction of the Free Energy of Hydration of a Challenging Set of Pesticide-Like Compounds [†]. *The Journal of Physical Chemistry B*, 113(14):4508–4510, Apr. 2009.
- [105] A. Klamt, F. Eckert, M. Diedenhofen, and M. E. Beck. First Principles Calculations of Aqueous p K_a Values for Organic and Inorganic Acids Using COSMO-RS Reveal an Inconsistency in the Slope of the p K_a Scale. *J. Phys. Chem. A*, 107(44):9380–9386, Nov. 2003.
- [106] A. Klamt, F. Eckert, J. Reinisch, and K. Wichmann. Prediction of cyclohexane-water distribution coefficients with COSMO-RS on the SAMPL5 data set. *Journal of Computer-Aided Molecular Design*, 30(11):959–967, Nov. 2016.

- [107] A. Klamt, V. Jonas, T. Bürger, and J. C. W. Lohrenz. Refinement and Parametrization of COSMO-RS. *The Journal of Physical Chemistry A*, 102(26):5074–5085, June 1998.
- [108] G. Klopman, J.-Y. Li, S. Wang, and M. Dimayuga. Computer Automated log P Calculations Based on an Extended Group Contribution Approach. *Journal of Chemical Information and Modeling*, 34(4):752–781, July 1994.
- [109] T. Kloss, J. Heil, and S. M. Kast. Quantum Chemistry in Solution by Combining 3D Integral Equation Theory with a Cluster Embedding Approach. *The Journal of Physical Chemistry B*, 112(14):4337–4343, Apr. 2008.
- [110] P. A. Kollman. Advances and Continuing Challenges in Achieving Realistic and Predictive Simulations of the Properties of Organic and Biological Molecules. *Accounts of Chemical Research*, 29(10):461–469, Jan. 1996.
- [111] A. Kovalenko and F. Hirata. Three-dimensional density profiles of water in contact with a solute of arbitrary shape: A RISM approach. *Chemical Physics Letters*, 290(1-3):237–244, June 1998.
- [112] A. Krämer, P. S. Hudson, M. R. Jones, and B. R. Brooks. Multi-Phase Boltzmann Weighting: Accounting for Local Inhomogeneity in Molecular Simulations of Water-Octanol Partition Coefficients. *Journal of Computer-Aided Molecular Design*, (SAMPL6 Part II Special Issue), 2020.
- [113] J. L. Kulp, S. N. Blumenthal, Q. Wang, R. L. Bryan, and F. Guarnieri. A fragment-based approach to the SAMPL3 Challenge. *Journal of Computer-Aided Molecular Design*, 26(5):583–594, May 2012.
- [114] A. Kumar and K. Y. J. Zhang. Computational fragment-based screening using RosettaLigand: The SAMPL3 challenge. *Journal of Computer-Aided Molecular Design*, 26(5):603–616, May 2012.
- [115] P. G. Kusalik and I. M. Svishchev. The Spatial Structure in Liquid Water. *Science*, 265(5176):1219–1221, Aug. 1994.
- [116] S. K. Lakkaraju, E. P. Raman, W. Yu, and A. D. MacKerell. Sampling of Organic Solutes in Aqueous and Heterogeneous Environments Using Oscillating Excess Chemical Potentials in Grand Canonical-like Monte Carlo-Molecular Dynamics Simulations. *Journal of Chemical Theory and Computation*, 10(6):2281–2290, June 2014.
- [117] B. E. Lang. Solubility of Water in Octan-1-ol from (275 to 369) K. *Journal of Chemical & Engineering Data*, 57(8):2221–2226, Aug. 2012.
- [118] E. B. Lenselink and P. F. W. Stouten. Multitask machine learning models for predicting lipophilicity (logp). *Journal of Computer-Aided Molecular Design*, 2021.
- [119] A. Leo. The octanol–water partition coefficient of aromatic solutes: The effect of electronic interactions, alkyl chains, hydrogen bonds, and ortho-substitution. *J. Chem. Soc., Perkin Trans. 2*, (6):825–838, 1983.

- [120] A. Leo, C. Hansch, and D. Elkins. Partition coefficients and their uses. *Chem. Rev.*, 71(6):525–616, Dec. 1971.
- [121] A. J. Leo. Calculating log p(oct) from structures. *Chemical Reviews*, 93(4):1281–1306, 1993.
- [122] A. J. Leo and D. Hoekman. Calculating log P(oct) with no missing fragments; The problem of estimating new interaction parameters. *Perspectives in Drug Discovery and Design*, 18(1):19–38, June 2000.
- [123] Y. Levy and J. N. Onuchic. WATER MEDIATION IN PROTEIN FOLDING AND MOLECULAR RECOGNITION. *Annu. Rev. Biophys. Biomol. Struct.*, 35(1):389–415, June 2006.
- [124] H. Li, J. Chowdhary, L. Huang, X. He, A. D. MacKerell, and B. Roux. Drude Polarizable Force Field for Molecular Dynamics Simulations of Saturated and Unsaturated Zwitterionic Lipids. *Journal of Chemical Theory and Computation*, 13(9):4535–4552, Sept. 2017.
- [125] Z. Li and T. Lazaridis. Computing the Thermodynamic Contributions of Interfacial Water. In R. Baron, editor, *Computational Drug Discovery and Design*, volume 819, pages 393–404. Springer New York, New York, NY, 2012.
- [126] C. Liao and M. C. Nicklaus. Comparison of Nine Programs Predicting p K_a Values of Pharmaceutical Substances. *J. Chem. Inf. Model.*, 49(12):2801–2812, Dec. 2009.
- [127] C. Loschen, J. Reinisch, and A. Klamt. COSMO-RS based predictions for the SAMPL6 logP challenge. *Journal of Computer-Aided Molecular Design*, Nov. 2019.
- [128] T. Luchko, N. Blinov, G. C. Limon, K. P. Joyce, and A. Kovalenko. SAMPL5: 3D-RISM partition coefficient calculations with partial molar volume corrections and solute conformational sampling. *Journal of Computer-Aided Molecular Design*, 30(11):1115–1127, Nov. 2016.
- [129] A. P. Lyubartsev, S. P. Jacobsson, G. Sundholm, and A. Laaksonen. Solubility of Organic Compounds in Water/Octanol Systems. A Expanded Ensemble Molecular Dynamics Simulation Study of log P Parameters. *The Journal of Physical Chemistry B*, 105(32):7775–7782, Aug. 2001.
- [130] A. P. Lyubartsev, A. A. Martsinovski, S. V. Shevkunov, and P. N. Vorontsov-Velyaminov. New approach to Monte Carlo calculation of the free energy: Method of expanded ensembles. *Journal of chemical physics.*, 96(3):1776–1783, 1992.
- [131] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation*, 11(8):3696–3713, Aug. 2015.

- [132] D. T. Manallack. The pK_a Distribution of Drugs: Application to Drug Discovery. *Perspect Medicin Chem*, 1:1177391X0700100, Jan. 2007.
- [133] R. Mannhold, G. I. Poda, C. Ostermann, and I. V. Tetko. Calculation of Molecular Lipophilicity: State-of-the-Art and Comparison of LogP Methods on more than 96,000 Compounds. *Journal of Pharmaceutical Sciences*, 98(3):861–893, Mar. 2009.
- [134] A. V. Marenich, C. J. Cramer, and D. G. Truhlar. Performance of SM6, SM8, and SMD on the SAMPL1 Test Set for the Prediction of Small-Molecule Solvation Free Energies. *J. Phys. Chem. B*, 113(14):4538–4543, Apr. 2009.
- [135] A. V. Marenich, C. J. Cramer, and D. G. Truhlar. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B*, 113(18):6378–6396, May 2009.
- [136] A. V. Marenich, C. J. Cramer, and D. G. Truhlar. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B*, 113(18):6378–6396, May 2009.
- [137] A. V. Marenich, C. J. Cramer, and D. G. Truhlar. Generalized Born Solvation Model SM12. *Journal of Chemical Theory and Computation*, 9(1):609–620, Jan. 2013.
- [138] L. Martínez, R. Andrade, E. G. Birgin, and J. M. Martínez. PACKMOL: A package for building initial configurations for molecular dynamics simulations. *Journal of Computational Chemistry*, 30(13):2157–2164, Oct. 2009.
- [139] M. Maurer, S. de Beer, and C. Oostenbrink. Calculation of Relative Binding Free Energy in the Water-Filled Active Site of Oligopeptide-Binding Protein A. *Molecules*, 21(4):499, Apr. 2016.
- [140] N. A. Meanwell. Improving Drug Candidates by Design: A Focus on Physicochemical Properties As a Means of Improving Compound Disposition and Safety. *Chem. Res. Toxicol.*, 24(9):1420–1456, Sept. 2011.
- [141] E. Meyer. Internal water molecules and H-bonding in biological macromolecules: A review of structural features with functional implications. *Protein Sci.*, 1(12):1543–1562, Dec. 1992.
- [142] W. M. Meylan and P. H. Howard. Atom/Fragment Contribution Method for Estimating Octanol–Water Partition Coefficients. *Journal of Pharmaceutical Sciences*, 84(1):83–92, Jan. 1995.
- [143] M. Mezei. A cavity-biased (T, V, μ) Monte Carlo method for the computer simulation of fluids. *Molecular Physics*, 40(4):901–906, July 1980.
- [144] J. Michel, J. Tirado-Rives, and W. L. Jorgensen. Prediction of the Water Content in Protein Binding Sites. *J. Phys. Chem. B*, 113(40):13337–13346, Oct. 2009.

- [145] F. Milletti, L. Storchi, G. Sforna, and G. Cruciani. New and Original pK_a Prediction Method Using Grid Molecular Interaction Fields. *Journal of Chemical Information and Modeling*, 47(6):2172–2181, Nov. 2007.
- [146] M. A. Miteva, F. Guyon, and P. Tuffery. Frog2: Efficient 3D conformation ensemble generator for small compounds. *Nucleic Acids Research*, 38(Web Server):W622–W627, July 2010.
- [147] D. L. Mobley. SAMPL: Its present and future, and some work on the logP challenge, Aug. 2019.
- [148] D. L. Mobley, C. C. Bannan, A. Rizzi, C. I. Bayly, J. D. Chodera, V. T. Lim, N. M. Lim, K. A. Beauchamp, M. R. Shirts, M. K. Gilson, and P. K. Eastman. Open Force Field Consortium: Escaping atom types using direct chemical perception with SMIRNOFF v0.1. *bioRxiv*, Mar. 2018.
- [149] D. L. Mobley, C. I. Bayly, M. D. Cooper, and K. A. Dill. Predictions of Hydration Free Energies from All-Atom Molecular Dynamics Simulations. *J Phys Chem B*, 113:4533–4537, Jan. 2009.
- [150] D. L. Mobley and M. K. Gilson. Predicting Binding Free Energies: Frontiers and Benchmarks. *Annu. Rev. Biophys.*, 46(1):531–558, May 2017.
- [151] D. L. Mobley, M. Işık, A. Paluch, C. Loschen, N. Tielker, E. Vöhringer-Martinez, and A. Nikitin. The sampl6 logp virtual workshop, May 2019.
- [152] D. L. Mobley, M. Işık, A. Paluch, C. Loschen, N. Tielker, E. Vöhringer-Martinez, and A. Nikitin. The sampl6 logp virtual workshop github repository for presentation slides, May 2019.
- [153] D. L. Mobley, S. Liu, D. S. Cerutti, W. C. Swope, and J. E. Rice. Alchemical prediction of hydration free energies for SAMPL. *Journal of Computer-Aided Molecular Design*, 26(5):551–562, May 2012.
- [154] D. L. Mobley, S. Liu, N. M. Lim, K. L. Wymer, A. L. Perryman, S. Forli, N. Deng, J. Su, K. Branson, and A. J. Olson. Blind prediction of HIV integrase binding from the SAMPL4 challenge. *Journal of Computer-Aided Molecular Design*, 28(4):327–345, Apr. 2014.
- [155] D. L. Mobley, K. L. Wymer, N. M. Lim, and J. P. Guthrie. Blind prediction of solvation free energies from the SAMPL4 challenge. *Journal of Computer-Aided Molecular Design*, 28(3):135–150, Mar. 2014.
- [156] Moka;. molecular discovery, hertfordshire, uk, 2018. <https://www.moldiscovery.com/software/moka/>.
- [157] I. Moriguchi, S. Hirono, Q. Liu, I. Nakagome, and Y. Matsushita. Simple Method of Calculating Octanol/Water Partition Coefficient. *Chem. Pharm. Bull.*, 40(1):127–130, 1992.

- [158] H. S. Muddana, C. Daniel Varnado, C. W. Bielawski, A. R. Urbach, L. Isaacs, M. T. Geballe, and M. K. Gilson. Blind prediction of host–guest binding affinities: A new SAMPL3 challenge. *Journal of Computer-Aided Molecular Design*, 26(5):475–487, May 2012.
- [159] H. S. Muddana, A. T. Fenley, D. L. Mobley, and M. K. Gilson. The SAMPL4 host–guest blind prediction challenge: An overview. *Journal of Computer-Aided Molecular Design*, 28(4):305–317, Apr. 2014.
- [160] C. N. Nguyen, A. Cruz, M. K. Gilson, and T. Kurtzman. Thermodynamics of Water in an Enzyme Active Site: Grid-Based Hydration Analysis of Coagulation Factor Xa. *Journal of Chemical Theory and Computation*, 10(7):2769–2780, July 2014.
- [161] C. N. Nguyen, T. Kurtzman Young, and M. K. Gilson. Grid inhomogeneous solvation theory: Hydration structure and thermodynamics of the miniature receptor cucurbit[7]uril. *The Journal of Chemical Physics*, 137(4):044101, July 2012.
- [162] A. Nicholls, D. L. Mobley, J. P. Guthrie, J. D. Chodera, C. I. Bayly, M. D. Cooper, and V. S. Pande. Predicting Small-Molecule Solvation Free Energies: An Informal Blind Test for Computational Chemistry. *J. Med. Chem.*, 51(4):769–779, Feb. 2008.
- [163] G. J. Niemi, S. C. Basak, G. Grunwald, and G. D. Veith. Prediction of octanol/water partition coefficient (K_{ow}) with algorithmically derived variables. *Environmental Toxicology and Chemistry*, 11(7):893–900, July 1992.
- [164] C. Nieto-Draghi, G. Fayet, B. Creton, X. Rozanska, P. Rotureau, J.-C. de Hemptinne, P. Ungerer, B. Rousseau, and C. Adamo. A General Guidebook for the Theoretical Prediction of Physicochemical Properties of Chemicals for Regulatory Purposes. *Chemical Reviews*, 115(24):13093–13164, Dec. 2015.
- [165] A. Nikitin. Non-zero Lennard-Jones parameters for the Toukan–Rahman water model: More accurate calculations of the solvation free energy of organic substances. *Journal of Computer-Aided Molecular Design*, Nov. 2019.
- [166] A. Nikitin, Y. Milchevskiy, and A. Lyubartsev. AMBER-ii: New Combining Rules and Force Field for Perfluoroalkanes. *The Journal of Physical Chemistry B*, 119(46):14563–14573, Nov. 2015.
- [167] J. P. Nilmeier, G. E. Crooks, D. D. L. Minh, and J. D. Chodera. Nonequilibrium candidate Monte Carlo is an efficient tool for equilibrium simulation. *Proceedings of the National Academy of Sciences*, 108(45):E1009–E1018, Nov. 2011.
- [168] E. Nittinger, N. Schneider, G. Lange, and M. Rarey. Evidence of Water Molecules—A Statistical Evaluation of Water Molecules Based on Electron Density. *J. Chem. Inf. Model.*, 55(4):771–783, Apr. 2015.
- [169] N. M. O’Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison. Open Babel: An open chemical toolbox. *J. Cheminformatics*, 3:33, Oct. 2011.

- [170] Quacpac toolkit 2020.2.0 openeye scientific software, santa fe, nm. <http://www.eyesopen.com>.
- [171] S. Park and J. G. Saven. Statistical and molecular dynamics studies of buried waters in globular proteins. *Proteins*, 60(3):450–463, June 2005.
- [172] R. A. Pearlstein, W. Sherman, and R. Abel. Contributions of water transfer energy to protein-ligand association and dissociation barriers: Watermap analysis of a series of p38 α MAP kinase inhibitors: Water Transfer in Structure-Kinetic Relationships. *Proteins*, 81(9):1509–1526, Sept. 2013.
- [173] H. E. Pence and A. Williams. ChemSpider: An Online Chemical Information Resource. *J. Chem. Educ.*, 87(11):1123–1124, Nov. 2010.
- [174] K. L. Perez, S. Pinheiro, and W. Zamora. Multiple linear regression models for predicting the n-octanol/water partition coefficients in the sampl7 blind challenge. *Journal of Computer-Aided Molecular Design*, 2021.
- [175] A. A. Petrauskas and E. A. Kolovanov. ACD/Log P method description. *Persect. Drug. Discov.*, 19(1):99–116, Sept. 2000.
- [176] F. C. Pickard, G. König, F. Tofoleanu, J. Lee, A. C. Simmonett, Y. Shao, J. W. Ponder, and B. R. Brooks. Blind prediction of distribution in the SAMPL5 challenge with QM based protomer and pK_a corrections. *Journal of Computer-Aided Molecular Design*, 30(11):1087–1100, Nov. 2016.
- [177] P. Pracht, R. Wilcken, A. Udvarhelyi, S. Rodde, and S. Grimme. High accuracy quantum-chemistry-based calculation and blind prediction of macroscopic pK_a values in the context of the SAMPL6 challenge. *Journal of Computer-Aided Molecular Design*, 32(10):1139–1149, Oct. 2018.
- [178] S. Prasad and B. R. Brooks. A deep learning approach for the blind logP prediction in SAMPL6 challenge. *Journal of Computer-Aided Molecular Design*, 34(5):535–542, May 2020.
- [179] S. Prasad, J. Huang, Q. Zeng, and B. R. Brooks. An explicit-solvent hybrid QM and MM approach for predicting pK_a of small molecules in SAMPL6 challenge. *Journal of Computer-Aided Molecular Design*, 32(10):1191–1201, Oct. 2018.
- [180] P. Procacci and C. Cardelli. Fast Switching Alchemical Transformations in Molecular Dynamics Simulations. *Journal of Chemical Theory and Computation*, 10(7):2813–2823, July 2014.
- [181] P. Procacci and G. Guarnieri. SAMPL6 blind predictions of water-octanol partition coefficients using nonequilibrium alchemical approaches. *Journal of Computer-Aided Molecular Design*, Oct. 2019.
- [182] RDKit: Open-source cheminformatics. <http://www.rdkit.org>.

- [183] R. F. Ribeiro, A. V. Marenich, C. J. Cramer, and D. G. Truhlar. Prediction of SAMPL2 aqueous solvation free energies and tautomeric ratios using the SM8, SM8AD, and SMD solvation models. *Journal of Computer-Aided Molecular Design*, 24(4):317–333, Apr. 2010.
- [184] A. Rizzi, J. Chodera, L. Naden, K. Beauchamp, S. Albanese, P. Grinaway, B. Rustenburg, ajsilveira, S. Saladi, and K. Boehm. choderalab/yank: 0.24.0 - Experimental support for online status files, Feb. 2019.
- [185] A. Rizzi, J. Chodera, L. Naden, K. Beauchamp, S. Albanese, P. Grinaway, B. Rustenburg, S. Saladi, and K. Boehm. choderalab/yank: Bugfix release, Oct. 2018.
- [186] A. Rizzi, T. Jensen, D. R. Slochower, M. Aldeghi, V. Gapsys, D. Ntekoumes, S. Bosisio, M. Papadourakis, N. M. Henriksen, B. L. de Groot, Z. Cournia, A. Dickson, J. Michel, M. K. Gilson, M. R. Shirts, D. L. Mobley, and J. D. Chodera. The SAMPL6 SAMPLing challenge: Assessing the reliability and efficiency of binding free energy calculations. *Journal of Computer-Aided Molecular Design*, 34(5):601–633, May 2020.
- [187] A. Rizzi, S. Murkli, J. N. McNeill, W. Yao, M. Sullivan, M. K. Gilson, M. W. Chiu, L. Isaacs, B. C. Gibb, D. L. Mobley, and J. D. Chodera. Overview of the SAMPL6 host–guest binding affinity prediction challenge. *Journal of Computer-Aided Molecular Design*, 32(10):937–963, Oct. 2018.
- [188] G. A. Ross, M. S. Bodnarchuk, and J. W. Essex. Water Sites, Networks, And Free Energies with Grand Canonical Monte Carlo. *J. Am. Chem. Soc.*, 137(47):14930–14943, Dec. 2015.
- [189] G. A. Ross, H. E. Bruce Macdonald, C. Cave-Ayland, A. I. Cabedo Martinez, and J. W. Essex. Replica-Exchange and Standard State Binding Free Energies with Grand Canonical Monte Carlo. *Journal of Chemical Theory and Computation*, 13(12):6373–6381, Dec. 2017.
- [190] M. Rupp, R. Korner, and I. V. Tetko. Predicting the pKa of Small Molecules. *CCHTS*, 14(5):307–327, June 2011.
- [191] A. S. Rustenburg, J. Dancer, B. Lin, J. A. Feng, D. F. Ortwine, D. L. Mobley, and J. D. Chodera. Measuring experimental cyclohexane-water distribution coefficients for the SAMPL5 challenge. *Journal of Computer-Aided Molecular Design*, 30(11):945–958, Nov. 2016.
- [192] J. Sangster. Octanol-water partition coefficients of simple organic compounds. *Journal of Physical and Chemical Reference Data*, 18(3):1111–1229, 1989.
- [193] S. Sasmal, S. C. Gill, N. M. Lim, and D. L. Mobley. Sampling conformational changes of bound ligands using Nonequilibrium Candidate Monte Carlo. *Journal of Chemical Theory and Computation*, page acs.jctc.9b01066, Feb. 2020.

- [194] J. L. Schlessman, C. Abe, A. Gittis, D. A. Karp, M. A. Dolan, and B. García-Moreno. Crystallographic Study of Hydration of an Internal Cavity in Engineered Proteins with Buried Polar or Ionizable Groups. *Biophysical Journal*, 94(8):3208–3216, Apr. 2008.
- [195] T. S. Schroeter, A. Schwaighofer, S. Mika, A. Ter Laak, D. Suelzle, U. Ganzer, N. Heinrich, and K.-R. Müller. Predicting Lipophilicity of Drug-Discovery Molecules using Gaussian Process Models. *ChemMedChem*, 2(9):1265–1267, Sept. 2007.
- [196] E. Selwa, I. M. Kenney, O. Beckstein, and B. I. Iorga. SAMPL6: Calculation of macroscopic pKa values from ab initio quantum mechanical free energies. *Journal of Computer-Aided Molecular Design*, 32(10):1203–1216, Oct. 2018.
- [197] J. C. Shelley, A. Cholleti, L. L. Frye, J. R. Greenwood, M. R. Timlin, and M. Uchimaya. Epik: A software program for pK a prediction and protonation state generation for drug-like molecules. *Journal of Computer-Aided Molecular Design*, 21(12):681–691, Dec. 2007.
- [198] M. R. Shirts and J. D. Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of Chemical Physics*, 129(12):124105, Sept. 2008.
- [199] M. D. Shultz. Two Decades under the Influence of the Rule of Five and the Changing Properties of Approved Oral Drugs: Miniperspective. *J. Med. Chem.*, 62(4):1701–1714, Feb. 2019.
- [200] Simulations plus admet predictor v8.5;. simulations plus, lancaster, ca, 2018. <https://www.simulations-plus.com/software/admetpredictor/physicochemical-biopharmaceutical/>.
- [201] D. A. Sivak, J. D. Chodera, and G. E. Crooks. Using Nonequilibrium Fluctuation Theorems to Understand and Correct Errors in Equilibrium and Nonequilibrium Simulations of Discrete Langevin Dynamics. *Phys. Rev. X*, 3(1):011007, Jan. 2013.
- [202] B. Slater, A. McCormack, A. Avdeef, and J. E. Comer. PH-Metric logP.4. Comparison of Partition Coefficients Determined by HPLC and Potentiometric Methods to Literature Values. *Journal of Pharmaceutical Sciences*, 83(9):1280–1283, Sept. 1994.
- [203] H. Stöckmann, A. Bronowska, N. R. Syme, G. S. Thompson, A. P. Kalverda, S. L. Warriner, and S. W. Homans. Residual Ligand Entropy in the Binding of *p*-Substituted Benzenesulfonamide Ligands to Bovine Carbonic Anhydrase II. *J. Am. Chem. Soc.*, 130(37):12420–12426, Sept. 2008.
- [204] K. Takano, Y. Yamagata, and K. Yutani. Buried water molecules contribute to the conformational stability of a protein. *Protein Engineering, Design and Selection*, 16(1):5–9, Jan. 2003.
- [205] I. V. Tetko, V. Y. Tanchuk, and A. E. P. Villa. Prediction of n-Octanol/Water Partition Coefficients from PHYSPROP Database Using Artificial Neural Networks and E-State

- Indices. *Journal of Chemical Information and Computer Sciences*, 41(5):1407–1421, Sept. 2001.
- [206] N. Tielker, L. Eberlein, C. Chodun, S. Güssregen, and S. M. Kast. pKa calculations for tautomerizable and conformationally flexible molecules: Partition function vs. state transition approach. *J Mol Model*, 25(5):139, May 2019.
- [207] N. Tielker, L. Eberlein, S. Güssregen, and S. M. Kast. The SAMPL6 challenge on predicting aqueous pKa values from EC-RISM theory. *Journal of Computer-Aided Molecular Design*, 32(10):1151–1163, Oct. 2018.
- [208] N. Tielker, L. Eberlein, G. Hessler, K. F. Schmidt, S. Güssregen, and S. M. Kast. Quantum–mechanical property prediction of solvated drug molecules: What have we learned from a decade of SAMPL blind prediction challenges? *Journal of Computer-Aided Molecular Design*, 35(4):453–472, Apr. 2021.
- [209] N. Tielker, S. Güssregen, and S. M. Kast. Sampl7 physical property prediction from ec-rism theory. *Journal of Computer-Aided Molecular Design*, 2021.
- [210] N. Tielker, D. Tomazic, L. Eberlein, S. Güssregen, and S. M. Kast. The SAMPL6 challenge on predicting octanol-water partition coefficients from ECRISM theory. *Journal of Computer-Aided Molecular Design*, (SAMPL6 Part II Special Issue), 2020.
- [211] N. Tielker, D. Tomazic, J. Heil, T. Kloss, S. Ehrhart, S. Güssregen, K. F. Schmidt, and S. M. Kast. The SAMPL5 challenge for embedded-cluster integral equation theory: Solvation free energies, aqueous pK a, and cyclohexane–water log D. *Journal of Computer-Aided Molecular Design*, 30(11):1035–1044, Nov. 2016.
- [212] M. D. Tissandier, K. A. Cowen, W. Y. Feng, E. Gundlach, M. H. Cohen, A. D. Earhart, J. V. Coe, and T. R. Tuttle. The Proton’s Absolute Aqueous Enthalpy and Gibbs Free Energy of Solvation from Cluster-Ion Solvation Data. *J. Phys. Chem. A*, 102(40):7787–7794, Oct. 1998.
- [213] Turbomole v7.5. university of karlsruhe and forschungszentrum karlsruhe gmbh, 1989-2007, turbomole gmbh, since 2007. <https://www.turbomole.org>.
- [214] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. D. Mackerell. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of Computational Chemistry*, pages NA–NA, 2009.
- [215] D. Vasseti, M. Pagliai, and P. Procacci. Assessment of GAFF2 and OPLS-AA General Force Fields in Combination with the Water Models TIP3P, SPCE, and OPC3 for the Solvation Free Energy of Druglike Organic Molecules. *Journal of Chemical Theory and Computation*, 15(3):1983–1995, Mar. 2019.
- [216] T. Verstraelen, S. Vandenbrande, F. Heidar-Zadeh, L. Vanduyfhuys, V. Van Speybroeck, M. Waroquier, and P. W. Ayers. Minimal Basis Iterative Stockholder: Atoms

- in Molecules for Force-Field Development. *Journal of Chemical Theory and Computation*, 12(8):3894–3912, Aug. 2016.
- [217] A. Viayna, S. Pinheiro, C. Curutchet, F. J. Luque, and W. J. Zamora. Prediction of n-octanol/water partition coefficients and acidity constants (pka) in the sampl7 blind challenge with the iefpcm-mst model. *Journal of Computer-Aided Molecular Design*, 2021.
- [218] C. Vraka, L. Nics, K.-H. Wagner, M. Hacker, W. Wadsak, and M. Mitterhauser. Log P , a yesterday’s value? *Nuclear Medicine and Biology*, 50:1–10, July 2017.
- [219] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case. Development and testing of a general amber force field. *Journal of Computational Chemistry*, 25(9):1157–1174, July 2004.
- [220] K. Wang, J. D. Chodera, Y. Yang, and M. R. Shirts. Identifying ligand binding sites and poses using gpu-accelerated hamiltonian replica exchange molecular dynamics. *Journal of computer-aided molecular design*, 27(12):989–1007, 2013.
- [221] L.-P. Wang, T. J. Martinez, and V. S. Pande. Building Force Fields: An Automatic, Systematic, and Reproducible Approach. *J. Phys. Chem. Lett.*, 5(11):1885–1891, June 2014.
- [222] R. Wang, Y. Fu, and L. Lai. A New Atom-Additive Method for Calculating Partition Coefficients. *Journal of Chemical Information and Computer Sciences*, 37(3):615–621, May 1997.
- [223] R. Wang, Y. Gao, and L. Lai. Calculating partition coefficient by atom-additive method. *Perspect. Drug. Disc.*, 19(1):47–66, 2000.
- [224] J. Warnau, K. Wichmann, and J. Reinisch. Cosmo-rs predictions of logp in the sampl7 blind challenge. *Journal of Computer-Aided Molecular Design*, 2021.
- [225] S. A. Wildman and G. M. Crippen. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.*, 39(5):868–873, Sept. 1999.
- [226] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey. DrugBank: A comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.*, 34(Database issue):D668–672, Jan. 2006.
- [227] H.-J. Woo, A. R. Dinner, and B. Roux. Grand canonical Monte Carlo simulations of water in protein environments. *The Journal of Chemical Physics*, 121(13):6392–6400, Oct. 2004.
- [228] K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen, and R. Barzilay. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.*, 59(8):3370–3388, Aug. 2019.

- [229] J. Yin, N. M. Henriksen, D. R. Slochower, M. R. Shirts, M. W. Chiu, D. L. Mobley, and M. K. Gilson. Overview of the SAMPL5 host–guest challenge: Are we doing better? *Journal of Computer-Aided Molecular Design*, 31(1):1–19, Jan. 2017.
- [230] T. Young, R. Abel, B. Kim, B. J. Berne, and R. A. Friesner. Motifs for molecular recognition exploiting hydrophobic enclosure in protein–ligand binding. *PNAS*, 104(3):808–813, Jan. 2007.
- [231] B. Yu, M. Blaber, A. M. Gronenborn, G. M. Clore, and D. L. D. Caspar. Disordered water within a hydrophobic protein cavity visualized by x-ray crystallography. *Proceedings of the National Academy of Sciences*, 96(1):103–108, Jan. 1999.
- [232] W. J. Zamora, C. Curutchet, J. M. Campanera, and F. J. Luque. Prediction of pH-Dependent Hydrophobic Profiles of Small Molecules from Miertus–Scrocco–Tomasi Continuum Solvation Calculations. *J. Phys. Chem. B*, 121(42):9868–9880, Oct. 2017.
- [233] W. J. Zamora, S. Pinheiro, K. German, C. Ràfols, C. Curutchet, and F. J. Luque. Prediction of the n-octanol/water partition coefficients in the SAMPL6 blind challenge from MST continuum solvation calculations. *Journal of Computer-Aided Molecular Design*, 34(4):443–451, Apr. 2020.
- [234] Q. Zeng, M. R. Jones, and B. R. Brooks. Absolute and relative pKa predictions via a DFT approach applied to the SAMPL6 blind challenge. *Journal of Computer-Aided Molecular Design*, 32(10):1179–1189, Oct. 2018.
- [235] R. W. Zwanzig. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *The Journal of Chemical Physics*, 22(8):1420–1426, Aug. 1954.