

# Accuracy of macroscopic and microscopic $pK_a$ predictions of small molecules evaluated by the SAMPL6 blind prediction challenge

Mehtap Işık (ORCID: [0000-0002-6789-952X](#))<sup>1,2\*</sup>, Ariën S. Rustenburg (ORCID: [0000-0002-3422-0613](#))<sup>1,3</sup>, Andrea Rizzi (ORCID: [0000-0001-7693-2013](#))<sup>1,4</sup>, M. R. Gunner<sup>6</sup>, David L. Mobley (ORCID: [0000-0002-1083-5533](#))<sup>5</sup>, John D. Chodera (ORCID: [0000-0003-0542-119X](#))<sup>1</sup>

<sup>1</sup>Computational and Systems Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY 10065, United States; <sup>2</sup>Tri-Institutional PhD Program in Chemical Biology, Weill Cornell Graduate School of Medical Sciences, Cornell University, New York, NY 10065, United States; <sup>3</sup>Graduate Program in Physiology, Biophysics, and Systems Biology, Weill Cornell Medical College, New York, NY 10065, United States; <sup>4</sup>Tri-Institutional PhD Program in Computational Biology and Medicine, Weill Cornell Graduate School of Medical Sciences, Cornell University, New York, NY 10065, United States; <sup>5</sup>Department of Pharmaceutical Sciences and Department of Chemistry, University of California, Irvine, Irvine, California 92697, United States; <sup>6</sup>Department of Physics, City College of New York, New York NY 10031

**\*For correspondence:**  
[mehtap.isik@choderalab.org](mailto:mehtap.isik@choderalab.org) (MI)

## Abstract

Complete abstract.

- number of submissions [1]
- summary of analysis
- difficulties observed

## 0.1 Keywords

SAMPL · blind prediction challenge · acid dissociation constant ·  $pK_a$  · small molecule · macroscopic  $pK_a$  · microscopic  $pK_a$  · macroscopic protonation state · microscopic protonation state

## 0.2 Abbreviations

**SAMPL** Statistical Assessment of the Modeling of Proteins and Ligands

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

**SEM** Standard error of the mean

**RMSE** Root mean squared error

**MAE** Mean absolute error

$\tau$  Kendall's rank correlation coefficient (Tau)

**R<sup>2</sup>** Coefficient of determination (R-Squared)

# 1 Introduction

Complete introduction section: - Importance of small molecule pKa prediction for pharmaceutical efforts. - Definition of pKa - Acid dissociation equilibrium constant - Add pKa equation - Add free energy of protonation state equation - Definition of microscopic and macroscopic pKas - Introduce linear protonation state free energy diagram [Cite Gunner et al 2019 paper] FIGURE: linear plot of free energy vs pH

Importance of small molecule pKa prediction for pharmaceutical efforts.

Explain why we are doing a pKa challenge and connect to past and previous challenges

SAMPL (Statistical Assessment of the Modeling of Proteins and Ligands). About SAMPL challenges: Collectively, these challenges have assessed the effects of force field accuracy, solvation models, pKa and tautomer predictions.

During the SAMPL5 challenge, log D predictions experienced difficulties predicting log D values accurately, unless protonation states and tautomers were taken into account.

For this iteration of the SAMPL challenge, we have taken one step back and isolated just the problem of predicting solvent protonation states.

This is the first time a blind pKa prediction challenge has been fielded as part of SAMPL. In this first iteration of the challenge, we aimed to assess the performance of current pKa prediction methods and isolate potential causes of inaccurate pKa estimates, with the aim of determining how pKa prediction inaccuracies might impact predicted affinities for drug-like molecules. For example, for both logD and binding affinity predictions, any error in predicting the free energy of accessing a minor protonation state in solution that becomes dominant in the complex will directly add to the error in the predicted transfer or binding free energy.

Challenge goal: determining how pKa prediction inaccuracies might impact predicted affinities for drug-like molecules. For example, for both logD and binding affinity predictions, any error in predicting the free energy of accessing a minor protonation state in solution that becomes dominant in the complex will directly add to the error in the predicted transfer or binding free energy.

Reason for blind pKa challenge: - Impact on binding affinity predictions - Impact on logD predictions (SAMPL6) - Drug-like molecules are especially challenging.

Protonation state effects were a dominant accuracy-limiting factor for logD from SAMPL5, and should also be accuracy-limiting in binding free energy predictions. Errors in pKa predictions can cause modeling the wrong charge, protonation and tautomerization states which affect hydrogen bonding opportunities and overall dipole moment of the ligand.

Explain the physics of the predicted property

EQUATION: pKa equation

EQUATION: free energy of protonation state equation

Introducing linear protonation state free energy diagram

FIGURE: linear plot of free energy vs pH

FIGURE: a diagram illustrating the ways in which the pKa errors can influence prediction errors for binding affinities

Overview of kinds of pKa prediction methods available (ML, QM, empirical methods ...)

Explain challenge design.

Experimental macroscopic pKa values were measured using a UV-metric assay performed using a Sirius T3 [cite exp. paper] supported by Merck, MRL, Rahway NJ.

Communicate concepts behind challenge design and why we made specific choices: Explain why we have types I, II, III Explain why we preenumerated microstates

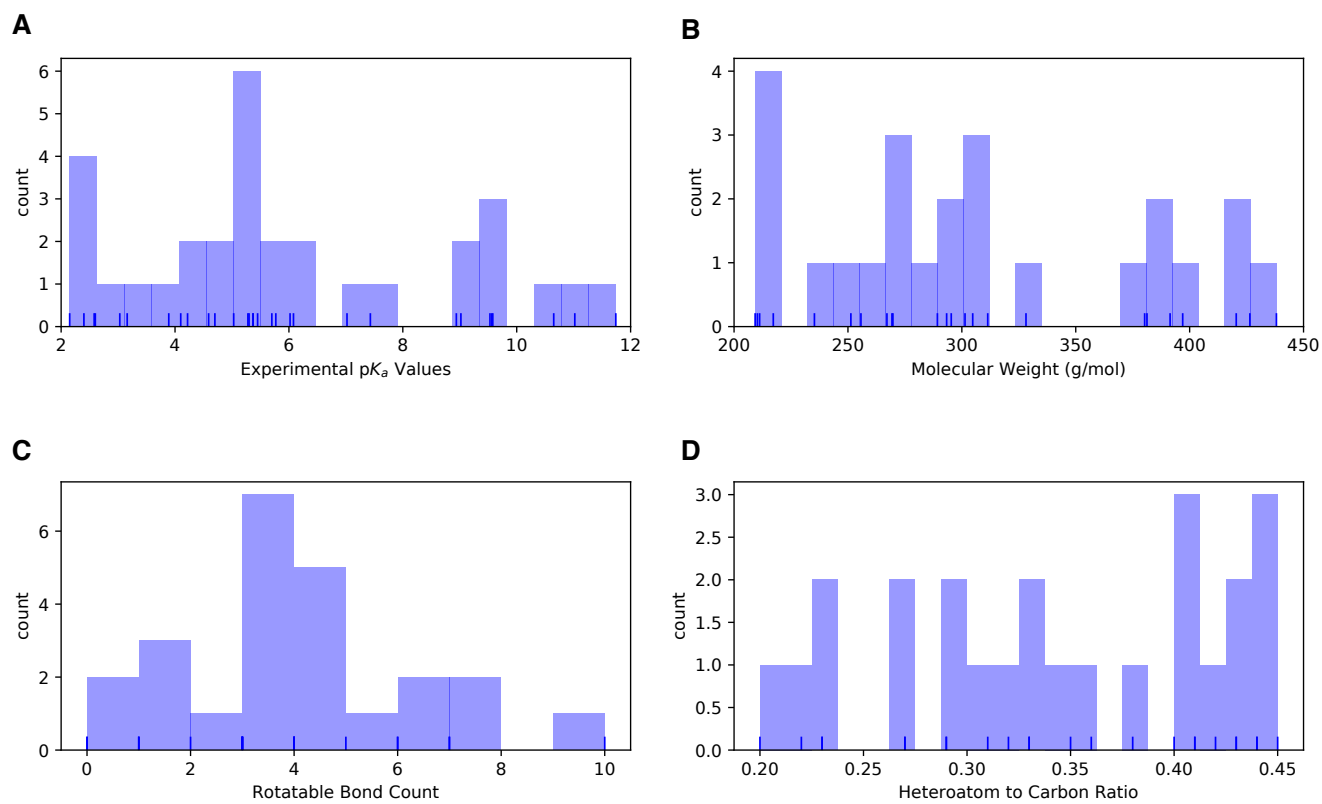
Participants had the option to submit predictions in one of 3 categories: Microscopic pKa values (type I), microscopic state populations (type II), or macroscopic pKa values (type III).

The comparison between macroscopic and microscopic pKa values is not always a straightforward one.

Overview of available pKa prediction methods and methods that participated in SAMPL6. [Reminder to cite all papers here.]

Explain future direction for this challenge

Challenge path: predict pKas, give people pKas to predict logDs on same molecules, then predict for new set of compounds logDs without provided pKas.



**Figure 1. Distribution of molecular properties of 24 compounds in SAMPL6  $pK_a$  Challenge.** **A** Histogram of spectrophotometric  $pK_a$  measurements collected with Sirius T3 [1]. Overlaid carpet plot indicates the actual values. Five compounds have multiple measured  $pK_a$ s in the range of 2-12. **B** Histogram of molecular weights of compounds in SAMPL6 set. Molecular weights were calculated by neglecting counter ions. **C** Histogram of the number of non-terminal rotatable bonds in each molecule. **D** The histogram of the ratio of heteroatom (non-carbon heavy atom) count to the number of carbon atoms.

Explain potential benefits of these challenge

Improving computational methods...

## 1.1 Motivation for a blind pKa challenge

why we are doing a pKa challenge and connect to past and previous challenge?

SAMPL (Statistical Assessment of the Modeling of Proteins and Ligands). About SAMPL challenges: Collectively, these challenges have assessed the effects of force field accuracy, solvation models, pKa and tautomer predictions.

During the SAMPL5 challenge, log D predictions experienced difficulties predicting log D values accurately, unless protonation states and tautomers were taken into account.

For this iteration of the SAMPL challenge, we have taken one step back and isolated just the problem of predicting solvent protonation states.

This is the first time a blind pKa prediction challenge has been fielded as part of SAMPL. In this first iteration of the challenge, we aimed to assess the performance of current pKa prediction methods and isolate potential causes of inaccurate pKa estimates, with the aim of determining how pKa prediction inaccuracies might impact predicted affinities for drug-like molecules. For example, for both logD and binding affinity predictions, any error in predicting the free energy of accessing a minor protonation state in solution that becomes dominant in the complex will directly add to the error in the predicted transfer or binding free energy.

Challenge goal: determining how pKa prediction inaccuracies might impact predicted affinities for drug-like molecules. For example, for both logD and binding affinity predictions, any error in predicting the free energy of accessing a minor protonation state in solution that becomes dominant in the complex will directly add to the error in the predicted transfer or binding free energy.

Reason for blind pKa challenge: 1. Impact on binding affinity predictions 2. Impact on logD predictions (SAMPL6) 3. Drug-like molecules are especially challenging.

Future challenge direction Challenge path: predict pKas, give people pKas to predict logDs on same molecules, then predict for new set of compounds logDs without provided pKas. Potential benefits of these challenges: 1. Improving computational methods 2. Detecting hidden contributors to error

## 1.2 Approaches to predict pKas

Overview of kinds of pKa prediction methods available (ML, QM, empirical methods ...)

# 2 Methods

## 2.1 Structure and logistics of the SAMPL6 pKa prediction challenge

Describe the structure of SAMPL6 pKa challenge

Experimental macroscopic pKa values were measured using a UV-metric assay performed using a Sirius T3 [cite exp. paper] supported by Merck, MRL, Rahway NJ.

Communicate concepts behind challenge design and why we made specific choices: 1. Explain why we have types I, II, III 2. Explain why we pre-enumerated microstates

Participants had the option to submit predictions in one of 3 categories: Microscopic pKa values (type I), microscopic state populations (type II), or macroscopic pKa values (type III).

The comparison between macroscopic and microscopic pKa values is not always a straightforward one.

- When instructions and input files were made available

- Challenge dates

- Input files

- What to predict? Three type of submissions.

- Multiple submissions allowed

- Predicting the pKa values of the whole set wasn't a requirement.

- 2nd D3R/SAMPL Workshop took place in La Jolla, San Diego on Feb 22-23, 2018.

Referece Figure ?? Drug-like molecules are often larger and more complex than the ones used in this study.

## 124 2.2 Enumeration of requested prediction microscopic protonation states

125 1. OpenEye (filter out resonance structures), Epik

126 2. Participant supplied structures

127 Microstate pairs: Only +/-1 charge change transitions are allowed. List of allowed transitions. +2 transitions are not consid-  
128 ered.

## 129 2.3 Evaluation approaches

### 130 2.3.1 Statistical metrics for submission performance

131 - Root mean squared error (RMSE)

132 - Mean absolute error (MAE)

133 - Mean Error (ME)

134 - Square of Pearson Correlation Coefficient ( $R^2$ )

135 - Slope of prediction vs. experimental value linear fit

136 Uncertainty in each performance statistic was calculated by bootstrapping (10,000) to estimate 95% confidence intervals.

### 137 2.3.2 Matching algorithms for pairing predicted and experimental pKas

138 Explain why it is necessary due to lacking structural information. Cite recommendations from article such as preserving sequence.  
139 Experimental data doesn't inform protonation site and overall charge of species. Experimental data doesn't capture the whole  
140 picture. We don't know charge and we don't know tautomers. We don't know the charge state of macrostates, this causes a  
141 matching problem

142 Explain Hungarian method for matching experimental and predicted pKas

143 Explain Closest method for matching experimental and predicted pKas

144 Explain microstate based matching.

## 145 2.4 Reference calculations

146 Schrodinger Epik Schrodinger Jaguar Chemicalize MoKa

## 147 3 Results and Discussion

148 A paragraph to explain the submission methods. Define method categories: DL, LFER, QSPR/ML, QM, QM+LEC, and QM+MM, Blind pre-  
149 dictions, Reference calculations, Null model (pKa prospector lookup)

149 Submissions spanning different method categories were made to the SAMPL6 pK<sub>a</sub> Challenge: database lookup (DL), linear  
150 free energy relationship (LFER), quantitative structure property relationship (QSPR), machine learning (ML), quantum mechanics  
151 (QM) models with and without linear empirical correction (LEC), and combined quantum mechanics and molecular mechanics  
152 (QM+MM). Unique submission IDs were assigned to each submission. Table 1 matches method names with submission IDs.  
153 Unique IDs were also assigned when multiple submissions exists for different submission types of the same method such as  
154 microscopic pK<sub>a</sub>(type I) and macroscopic pK<sub>a</sub> (type III).

### 155 3.1 Analysis of macroscopic pK<sub>a</sub> predictions (Type III)

156 MI: Methods are indicated by submission IDs.

157 Refer to SI TABLE: Error statistics for all participants. Refer to SI FIGURE: Error distribution ridge plots for each method  
158 (exp-pred macroscopic pK<sub>a</sub>). Which methods tend to overestimate and which methods tend to underestimate?

159 MI: SI TABLE: Error statistics for all participants

160 Describe number of missing and extra pKa for each method. Report in total for all molecules how many predicted pKas are  
161 there and how many experimental pKas. Refer to FIGURE: missing and extra pKa counts.

162 MI: SI TABLE: Missing and extra pKa counts

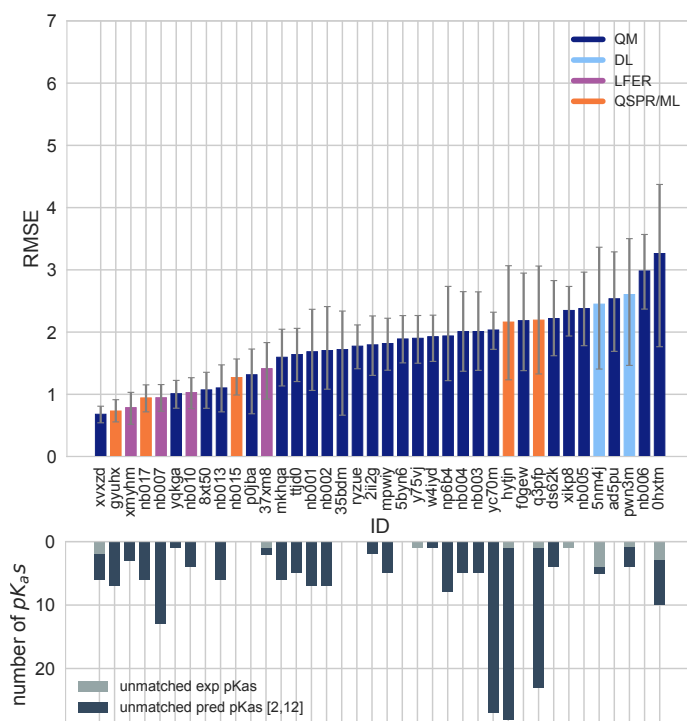
163 Describe overall performance comparison of different methods, grouped by methods class.

164 Explain rationale behind how we analyze the data and determine success/failure

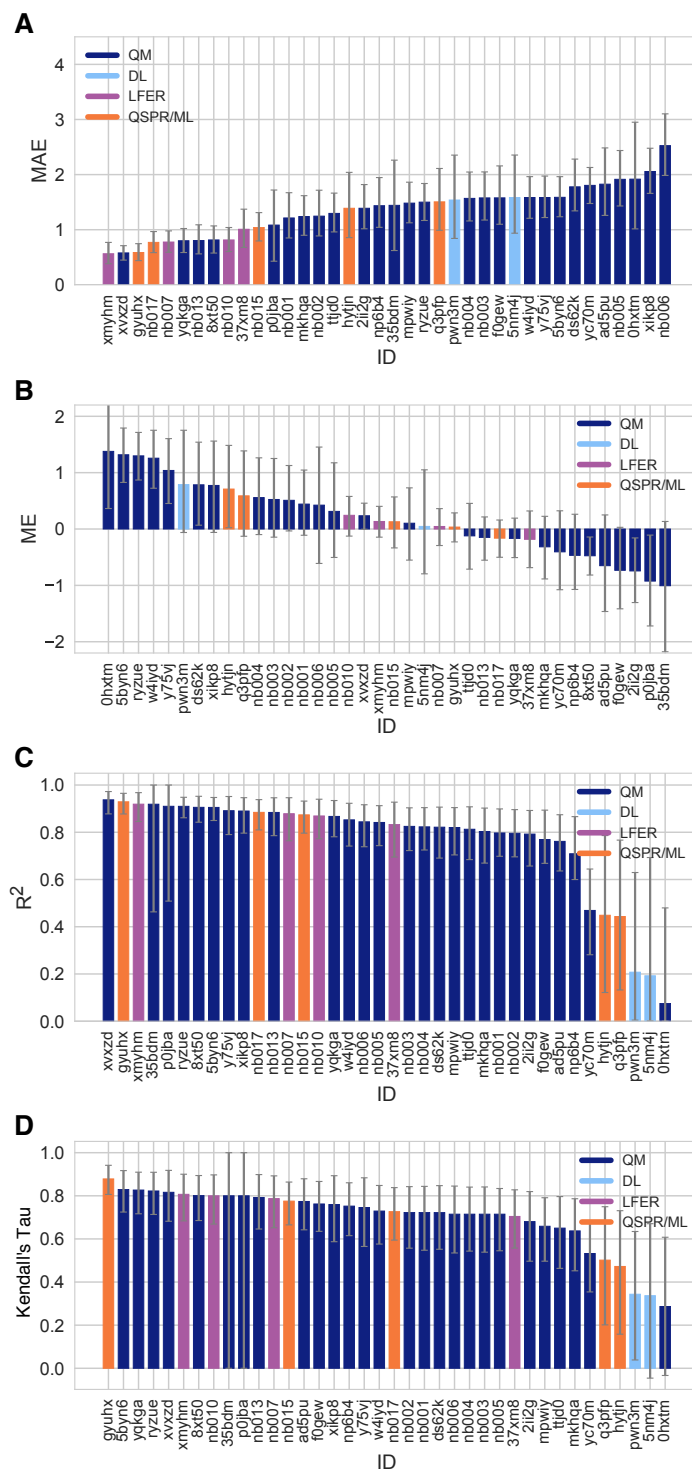
**Table 1. Submission IDs, names, category, and type for all the  $pK_a$  prediction sets.** Reference calculations are labeled as *nb###*. The method name column lists the names provided by each participant in the submission file. The “type” column indicates if submission was or a post-deadline reference calculation, denoted by “Blind” or “Reference” respectively. The table is not ordered by performance.

Method Category	Method	Microscopic $pK_a$ (Type I) Submission ID	Macroscopic $pK_a$ (Type III) Submission ID	Submission Type	Ref.
DL	Substructure matches to experimental data in pKa OpenEye pKa Prospector Database v1.0		<i>5nm4j</i>	Null	[2]
DL	OpenEye pKa-Prospector 1.0.0.3 with Analog Search ion identification algorithm		<i>pwn3m</i>	Null	[2]
LFER	ACD/pKa GALAS (ACD/Percepta Kernel v1.6)	<i>v8qph</i>	<i>37xm8</i>	Blind	[3]
LFER	ACD/pKa Classic (ACD/Percepta Kernel, v1.6)		<i>xmyhm</i>	Blind	[4]
LFER	Epik Scan (Schrodinger v2017-4)		<i>nb007</i>	Reference	[5]
LFER	Epik Microscopic (Schrodinger v2017-4)	<i>nb008</i>	<i>nb010</i>	Reference	[5]
QSPR/ML	OpenEye Gaussian Process	<i>6tvf8</i>	<i>hytjn</i>	Blind	[6]
QSPR/ML	OpenEye Gaussian Process Resampled		<i>q3pfp</i>	Blind	[6]
QSPR/ML	S+pKa (ADMET Predictor v8.5, Simulations Plus)	<i>hdiyq</i>	<i>gyuhx</i>	Blind	[7]
QSPR/ML	Chemicalize v18.23 (ChemAxon MarvinSketch v18.23)		<i>nb015</i>	Reference	[8]
QSPR/ML	MoKa v3.1.3	<i>nb016</i>	<i>nb017</i>	Reference	[9, 10]
QM	Adiabatic scheme with single point correction: SMD/M06-2X//6-311++G(d,p)//M06-2X/6-31+G(d) for bases and SMD/M06-2X//6-311++G(d,p)//M06-2X/6-31G(d) for acids + thermal corrections	<i>ko8yx</i>	<i>ryzue</i>	Blind	[11]
QM	Direct scheme with single point correction: SMD/M06-2X//6-311++G(d,p)//M06-2X/6-31+G(d) for bases and SMD/M06-2X//6-311++G(d,p)//M06-2X/6-31G(d) for acids + thermal corrections	<i>w4z0e</i>	<i>xikp8</i>	Blind	[11]
QM	Adiabatic scheme: thermodynamic cycle that uses gas phase optimized structures for gas phase free energy and solution phase geometries for solvent phase free energy. SMD/M06-2X/6-31+G(d) for bases and SMD/M06-2X/6-31G(d) for acids + thermal corrections	<i>wcvnu</i>	<i>5byn6</i>	Blind	[11]
QM	Vertical scheme: thermodynamic cycle that uses only gas phase optimized structures to compute gas phase and solvation free energy. SMD/M06-2X/6-31+G(d) for bases and SMD/M06-2X/6-31G(d) for acids + Thermal corrections	<i>arcko</i>	<i>w4iyd</i>	Blind	[11]
QM	Direct scheme: solution phase free energy is determined by solution phase geometries without thermodynamic cycle SMD/M06-2X/6-31+G(d) for bases and SMD/M06-2X/6-31G(d) for acids + thermal corrections	<i>wexjs</i>	<i>y75vj</i>	Blind	[11]
QM + LEC	Jaguar (Schrodinger v2017-4)	<i>nb011</i>	<i>nb013</i>	Reference	[12]
QM + LEC	CPCM/B3LYP/6-311+G(d,p) and global fitting	<i>y4wws</i>	<i>35bdm</i>	Blind	[13]
QM + LEC	CPCM/B3LYP/6-311+G(d,p) and separate fitting for neutral to negative and for positive to neutral transformations	<i>qsicn</i>	<i>p0jba</i>	Blind	[13]
QM + LEC	EC-RISM/MP2/6-311+G(d,p)-P3NI-q-noThiols-2par	<i>kxzt</i>	<i>ds62k</i>	Blind	[14]
QM + LEC	EC-RISM/MP2/cc-pVTZ-P2-q-noThiols-2par	<i>ftc8w</i>	<i>2ii2g</i>	Blind	[14]
QM + LEC	EC-RISM/MP2/6-311+G(d,p)-P2-phi-all-2par	<i>ktpj5</i>	<i>nb001</i>	Blind*	[14]
QM + LEC	EC-RISM/MP2/6-311+G(d,p)-P2-phi-noThiols-2par	<i>wuuvv</i>	<i>nb002</i>	Blind*	[14]
QM + LEC	EC-RISM/MP2/6-311+G(d,p)-P3NI-phi-all-2par	<i>2umai</i>	<i>nb003</i>	Blind*	[14]
QM + LEC	EC-RISM/MP2/6-311+G(d,p)-P3NI-phi-noThiols-2par	<i>cm2yq</i>	<i>nb004</i>	Blind*	[14]
QM + LEC	EC-RISM/MP2/6-311+G(d,p)-P2-phi-all-1par	<i>z7fhp</i>	<i>nb005</i>	Blind*	[14]
QM + LEC	EC-RISM/MP2/6-311+G(d,p)-P3NI-phi-all-1par	<i>8toyp</i>	<i>nb006</i>	Blind*	[14]
QM + LEC	EC-RISM/MP2/cc-pVTZ-P2-phi-noThiols-2par	<i>epvmk</i>	<i>tjd0</i>	Blind	[14]
QM + LEC	EC-RISM/MP2/cc-pVTZ-P2-phi-all-2par	<i>xnoe0</i>	<i>mkhqa</i>	Blind	[14]
QM + LEC	EC-RISM/MP2/cc-pVTZ-P3NI-phi-noThiols-2par	<i>4o0ia</i>	<i>mpwiy</i>	Blind	[14]
QM + LEC	EC-RISM/B3LYP/6-311+G(d,p)-P3NI-q-noThiols-2par	<i>nxaaw</i>	<i>ad5pu</i>	Blind	[14]
QM + LEC	EC-RISM/B3LYP/6-311+G(d,p)-P3NI-phi-noThiols-2par	<i>0xi4b</i>	<i>f0gew</i>	Blind	[14]
QM + LEC	EC-RISM/B3LYP/6-311+G(d,p)-P2-phi-noThiols-2par	<i>cywyk</i>	<i>np6b4</i>	Blind	[14]
QM + LEC	PCM/B3LYP/6-311+G(d,p)	<i>gdqeg</i>	<i>yc70m</i>	Blind	[14]
QM + LEC	COSMOtherm_FINE17 (COSMOtherm C30_1701, BP/TZVPD/FINE//BP/TZVP/COSMO)	<i>t8ewk</i>	<i>0hxtm</i>	Blind	[15, 16]
QM + LEC	DSD-BLYP-D3(BJ)/def2-TZVPD//PBEh-3c[DCOSMO-RS] + RRHO(GFN-xTB[GBSA]) + Gsolv(COSMO-RS[TZVPD]) and linear fit		<i>xvxzd</i>	Blind	[17]
QM + LEC	ReSCoSS conformations // DSD-BLYP-D3 reranking // COSMOtherm pKa: DSD-BLYP-D3(BJ)/def2-TZVPD// PBE-D3(BJ)/def2-TZVP/COSMO + RRHO(GFN-xTB + GBSA-water) + Gsolv[COSMO-RS(FINE17/TZVPD)] level and COSMOtherm pKa applied at the single conformer pair level (COSMOthermX17.0.5 release and BP-TZVPD-FINE-C30-1701 parameterization)	<i>eyetm</i>	<i>8xt50</i>	Blind	[17]
QM + LEC	ReSCoSS conformations // COSMOtherm pKa: DSD-BLYP-D3(BJ)/def2-TZVPD// PBE-D3(BJ)/def2-TZVP/COSMO + RRHO(GFN-xTB + GBSA-water) + Gsolv[COSMO-RS(FINE17/TZVPD)] level and COSMOtherm pKa applied directly on the resulting conformer sets with at least 5% Boltzmann weights for each microspecies (COSMOthermX17.0.5 release and BP-TZVPD-FINE-C30-1701 parameterization)	<i>ccpmw</i>	<i>yqkga</i>	Blind	[17]
QM + MM	M06-2X/6-31G*(for bases) or 6-31+G*(for acids) for gas phase, solvation free energy using TI with explicit solvent and GAFF, solvation free energy of proton -265.6 kcal/mol	<i>0wfzo</i>		Blind	[18]
QM + MM	M06-2X/6-31G*(for bases) or 6-31+G*(for acids) for gas phase, solvation free energy using TI with explicit solvent and GAFF, solvation free energy of proton -271.88 kcal/mol	<i>z3btx</i>		Blind	
QM + MM	M06-2X/6-31G*(for bases) or 6-31+G*(for acids) + thermal state correction for gas phase, solvation free energy using TI with explicit solvent and GAFF, solvation free energy of proton -265.6 kcal/mol	<i>758j8</i>		Blind	
QM + MM	M06-2X/6-31G*(for bases) or 6-31+G*(for acids) + thermal state correction for gas phase, solvation free energy using TI with explicit solvent and GAFF, solvation free energy of proton -271.88 kcal/mol	<i>hgn83</i>		Blind	

\* Microscopic  $pK_a$  submissions were blind, however, participant requested a correction after blind submission deadline for macroscopic  $pK_a$  submissions. Therefore, these were assigned submission IDs in the form of *nb###*.



**Figure 2. RMSE and unmatched  $pK_a$  counts vs. submission ID plots for macroscopic  $pK_a$  predictions based on Hungarian matching.** Methods are indicated by submission IDs. RMSE is shown with error bars denoting 95% confidence intervals obtained by bootstrapping over challenge molecules. Lower bar plots show the number of unmatched experimental  $pK_a$ s (light grey, missing predictions) and the number of unmatched  $pK_a$  predictions (dark grey, extra predictions) for each method between pH 2 and 12. Submission IDs are summarized in Table 1. Submission IDs of the form *nb###* refer to non-blinded reference methods computed after the blind challenge submission deadline. All others refer to blind, prospective predictions. Submissions are colored by their method categories. Light blue colored database look up methods are utilized as the null prediction method.



**Figure 3. Additional performance statistics for macroscopic  $pK_a$  predictions based on Hungarian matching.** Methods are indicated by submission IDs. Mean absolute error (MAE), mean error (ME), Pearson's  $R^2$ , and Kendall's Rank Correlation Coefficient Tau ( $\tau$ ) are shown, with error bars denoting 95% confidence intervals obtained by bootstrapping over challenge molecules. Refer to Table 1 for submission IDs and method names. Submissions are colored by their method categories. Light blue colored database look up methods are utilized as the null prediction method.



## Performance comparison of different methods, grouped by methods class

Method comparison based on statistical metrics. Explain the numerical matching methods used. Explain rationale behind how we analyze the data and determine success/failure. Method comparison according to different statistics: RMSE, MAE, ME, R2, m, Kendall's tau.

### 3.1.1 Consistently well performing methods for macroscopic $pK_a$ prediction

**Table 2. Four consistently well-performing prediction methods for macroscopic  $pK_a$  prediction based on consistent ranking within the Top 10 according to various statistical metrics.** Submissions were ranked according to RMSE, MAE,  $R^2$ , and  $\tau$ . Consistently well-performing methods were selected as the ones that rank in the Top 10 in each of these statistical metrics. These methods also have less than 2 unmatched experimental  $pK_a$ s and less than 7 unmatched predicted  $pK_a$ s according to Hungarian matching. Performance statistics are provided as mean and 95% confidence intervals.

Submission ID	Method Name	RMSE	MAE	$R^2$	Kendall's Tau ( $\tau$ )	Unmatched Exp. $pK_a$ Count	Unmatched Pred. $pK_a$ Count [2,12]
xvxzd	Full quantum chemical calculation of free energies and fit to experimental $pK_a$	0.68 [0.54, 0.81]	0.58 [0.45, 0.71]	0.94 [0.88, 0.97]	0.82 [0.68, 0.92]	2	4
gyuhx	S+pKa	0.73 [0.55, 0.91]	0.59 [0.44, 0.74]	0.93 [0.88, 0.96]	0.88 [0.8, 0.94]	0	7
xmyhm	ACD/pKa Classic	0.79 [0.52, 1.03]	0.56 [0.38, 0.77]	0.92 [0.85, 0.97]	0.81 [0.68, 0.9]	0	3
8xt50	ReSCoSS conformations // DSD-BLYP-D3 reranking // COSMOtherm $pK_a$	1.07 [0.78, 1.36]	0.81 [0.58, 1.07]	0.91 [0.84, 0.95]	0.80 [0.68, 0.89]	0	0

Check if top few performing methods are consistent between error metrics.

### 3.1.2 Which chemicals are harder to predict?

check amide next to aromatic heterocycles case

For physical prediction methods sulfur containing heterocycles, amide next to aromatic heterocycles, compounds with iodo and bromo domains have lower  $pK_a$  prediction accuracy.

Prediction performance of individual molecules

Which chemical structures make  $pK_a$  predictions more difficult?

SAMPL6  $pK_a$  set consisted of only 24 small molecules which limits our ability to do statistical analysis to determine which chemical substructures contribute to greater errors in  $pK_a$  predictions.

Illustration/explanation of effects where microscopic  $pK_a$ s and macroscopic  $pK_a$ s can differ

Are there any correlations between molecular descriptors and  $pK_a$  errors?

What can we learn from failures? Which physical effects are driving failures?

MI: FIGURE: Molecular MAE comparison across methods.

Does molecular descriptors explain errors/performance? We looked for correlation with descriptors, and potential explanation for errors. Keep spurious correlations in mind if we have many descriptors. No correlation observed. Reference the SI Figure of correlations.

Comparison of errors/performance against molecular descriptors. Look for correlation with descriptors, and potential explanation for errors. Keep spurious correlations in mind if we have many descriptors.

MI: Figure SI: correlation between prediction error and molecular descriptors

Are  $pK_a$  predictions better in middle region? No correlation between  $pK_a$  value and error was seen. Reference the SI Figure.

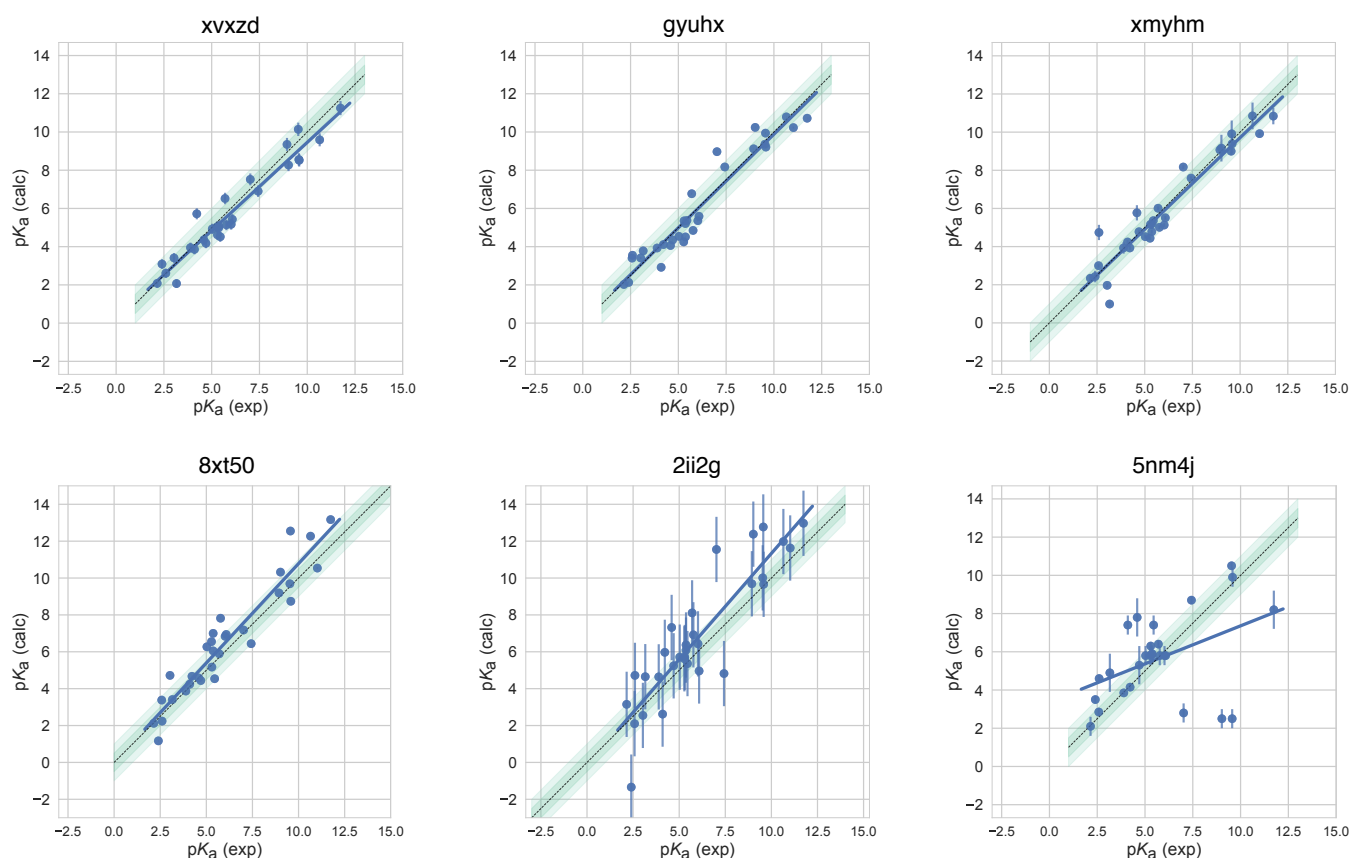
MI: Figure: Ridge plots of Delta  $pK_a$  error to identify compounds that were frequently mispredicted

Compare ME of molecules across methods. Are there molecules often overestimated or underestimated?

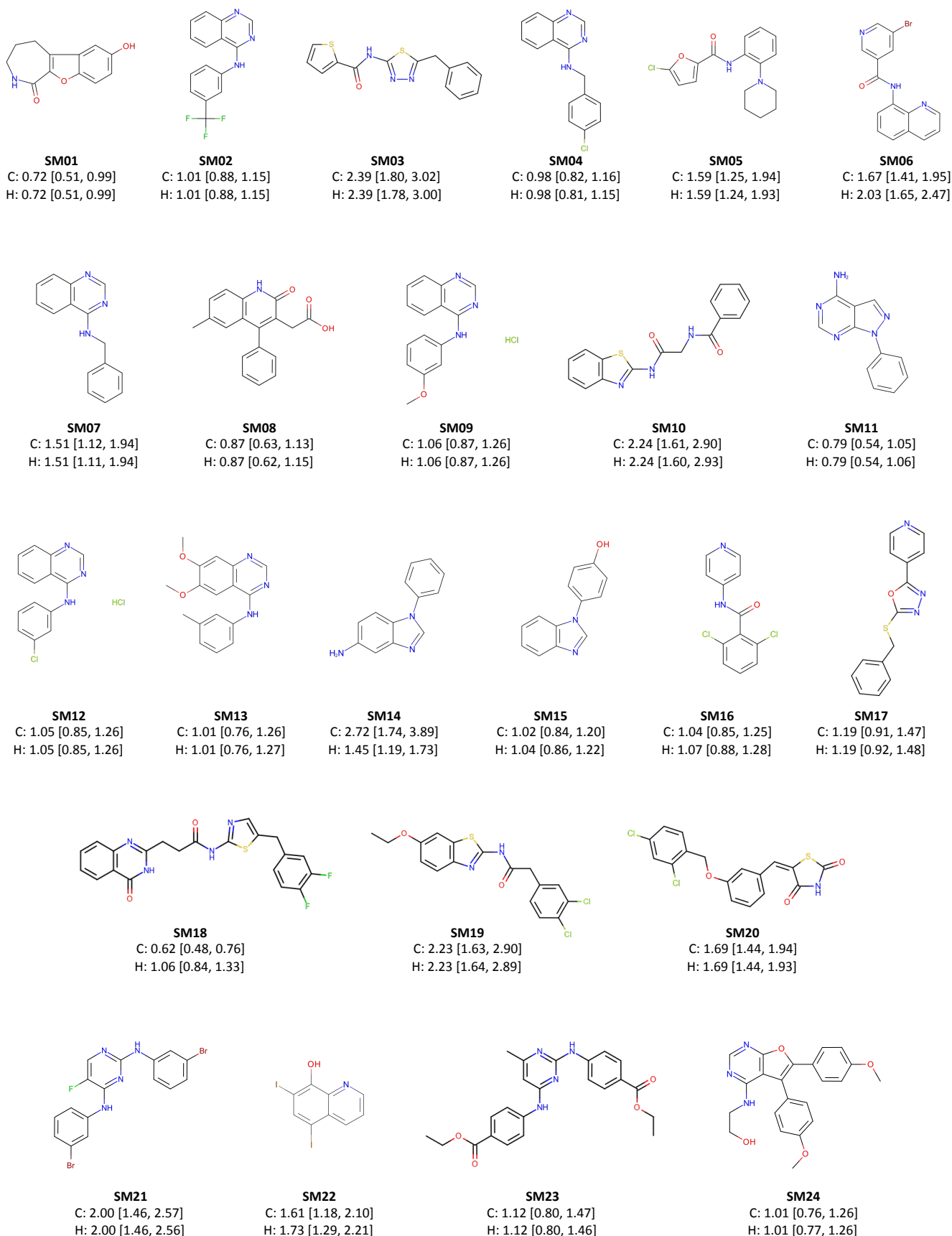
No correlation of macroscopic  $pK_a$  number to the errors? But we have low representation of multiprotic compounds

### 3.1.3 Comparing microscopic $pK_a$ predictions directly to macroscopic experimental $pK_a$ values leads to underestimation of errors

Discussion of matching experimental and predicted values



**Figure 4. Predicted vs. experimental value correlation plots of 4 consistently well-performing methods, a representative method with average performance (*2ii2g*), and the null method (*5nm4j*).** 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  $pK_a$  SEM values are too small to be seen under the data points. EC-RISM/MP2/cc-pVTZ-P2-q-noThiols-2par method (*2ii2g*) was selected as the representative method with average performance because it is the method with the highest RMSE below the median.



**Figure 5. Molecules of SAMPL6 pKa challenge with MAE calculated for all macroscopic (type III) predictions.** MAE calculated considering all prediction methods indicate which molecules had the lowest prediction accuracy in SAMPL6 challenge. MAE values calculated for each molecule include all the matched pKa values, which could be more than one per method for multiprotic molecules (SM06, SM14, SM15, SM16, SM18, SM22). "C:" and "H:" indicate results based on Closest and Hungarian matching algorithms employed for pairing experimental and predicted pKa values. Calculated MAE values were observed to differ for molecules with multiple experimental pKas. MAE values are reported with 95% confidence intervals.

Difficulty of assessing predicted pK<sub>a</sub>s using experimental data: matching problem

Explain rationale behind how we analyze the data and determine success/failure

Compare experimental data to microscopic pK<sub>a</sub> predictions, assuming experimental pK<sub>a</sub>s are titrations of distinguishable sides and therefore equal to microscopic pK<sub>a</sub>s. Molecules with only 1 pK<sub>a</sub> or well separated multiple pK<sub>a</sub>s (more than 3 pK<sub>a</sub> units apart) SM14 and SM18 were excluded from this analysis, since their experimental pK<sub>a</sub> values don't satisfy these criteria.

Errors computed by microstate-based matching are larger compared to numerical matching algorithms. Microscopic pK<sub>a</sub> analysis with numerical matching algorithms may mask errors due to higher number of guesses made.

MI: FIGURE Type I analysis, comparing analysis of 22 molecules (Hungarian vs Microstate matching)

## 3.2 Analysis of microscopic pK<sub>a</sub> predictions using microstates determined by NMR (8 molecules)

MI: FIGURE: Assign experimental pK<sub>a</sub> to microscopic transitions observed by NMR.

Conclusions will only be about 4-aminoquinazoline series and benzimidazole (8 molecules, 10 pK<sub>a</sub>s) Refer to SI figure of dominant microstates.

Choosing molecules with right protonation state is important. Do people predict the correct sequence of dominant microstates? " Even if your pK<sub>a</sub> prediction is correct, protonation state prediction can be wrong." Analyze which state has lowest free energy for each charge group ( The sequence of "experimentally visible states")

### 3.2.1 Accuracy of predicted pK<sub>a</sub> values when microstate matching is used

Assessment of individual methods by each of our analysis methods

Performance comparison of different methods, grouped by methods class

MI: FIGURE: Ranking of microscopic pK<sub>a</sub> prediction error statistics for all participants (8 mol, microstate match).

MI: FIGURE: Violin plots of Delta pK<sub>a</sub> error to identify compounds that were frequently mispredicted (microstate match)

### 3.2.2 Dominant microstate prediction accuracy of methods

Calculate relative free energy of microstates to determine dominant microstate of each charge Compare predicted and experimental dominant microstates and calculate accuracy of each method

MI: FIGURE: Dominant microstate accuracy vs method plot. Charges together and separate.

What percent of the time predictions capture the dominant protonation state correctly? Match by microstate and calculate RMSE and MAE. If you know the microstates, can you predict the value of the pK<sub>a</sub> right?

Does top 3 methods predict the same dominant microstate sequence? How differently do different methods predict microscopic transitions? (method vs method correlation plot to see if methods predict the same microstate pairs or not)

### 3.2.3 Which molecules caused lower dominant microstate prediction accuracy?

Which molecule has more errors in predicting the major microstates?

MI: FIGURE: Dominant microstate Accuracy vs Molecule ID plot, all charges and separate charges. Also think about plotting accuracy of QM and empirical methods separately.

Comment on consensus prediction accuracy. Comparison of predicted microstates using consensus set of transitions of high accuracy prediction methods

### 3.2.4 Demonstrate how numerical matching often masks the error

Match by Hungarian and calculate accuracy of microstate prediction overall. When matched by pK<sub>a</sub> value, do people come with the same transition pairs?

MI: FIGURE: [accuracy-of-microstates-based-on-numeric-matching] For most methods the microstate pair of Hungarian predicted pK<sub>a</sub> does not match experimentally determined microstate pair.

### 3.3 Analyzing microscopic pKa prediction from the perspective of thermodynamics

Explain linearity relative free energy of protonation states with respect to pH. Free energy perspective simplifies data capturing and analysis. Reference Marilyn's paper.

Thermodynamic cycle closure checking allows evaluation of microscopic pKas without experimental data. Checking for thermodynamic consistency

#### 3.3.1 Cycle closure error

Marilyn observed very good cycle closure results and very bad one that are up to 10 kcal/mol

She suggesting checking the cycle with maximum cycle closure error for each method and reporting that for each method. An histogram of max cycle closure error will help us bin these results into 3 categories: 1. good agreement 2. moderate 3. severe

"We think thermodynamic cycles of protonation states need to be closed" Message: Methods need to be checked for cycle closure errors. There can be information there that can be used to correct pKa predictions. When cycles are not closed it may be used as an indicator of prediction uncertainty.

### 3.4 How would pKa errors affect protein-ligand binding affinity predictions?

How do accuracy limitations in small molecule pKa prediction translate into modeling errors in ligand affinity prediction?

MI: FIGURE: a diagram illustrating the ways in which the pKa errors can influence prediction errors for binding affinities (A) When minor aqueous protonation state binds (B) When multiple protonation states can bind the complex

### 3.5 Lessons learned from SAMPL6 pKa Challenge

Do any methods predict within experimental accuracy (how is the field doing overall)?

Common challenging factors for accurate pKa predictions. Tautomers, Heterocycles etc.

Overall results: Do any methods predict within experimental accuracy (how is the field doing overall)? Common challenging factors for accurate pKa predictions. Tautomers, Heterocycles etc.

Discussion of matching problem between experimental and predicted values. Difficulty of assessing predicted pKas using experimental data: matching problem Explain rationale behind how we analyze the data and determine success/failure.

Conclusion about prediction performance of individual molecules: SAMPL6 pKa set consisted of only 24 small molecules which limits our ability to do statistical analysis to determine which chemical substructures contribute to greater errors in pKa predictions. Which chemical structures make pKa predictions more difficult?

What can we learn from failures? Which physical effects are driving failures? Cycle closure errors

### 3.6 Suggestions for future challenges

Discuss what can be done to further improve future challenges

How can we maximize what we learn? What should we have people predict? How should we select compounds / measure pKas?

Suggestions about challenge construction

Enumeration of protonation states before predictions (which states does one need to consider?)

Suggestions about challenge analysis

NMR experimental techniques could be used to validate microstate information in future challenges

Reporting microscopic pKa predictions with charges, microstate free energies is better Experimental dataset with microstate information is more helpful.

What can be done to further improve future challenges How can we maximize what we learn? What should we have people predict? How should we select compounds / measure pKas? NMR experimental techniques could be used to validate microstate information in future challenges

Suggestions about challenge construction Enumeration of protonation states before predictions (which states does one need to consider?) Suggestions about challenge analysis

## 271 4 Conclusion

## 272 5 Code and data availability

- 273 • SAMPL6 pK<sub>a</sub> challenge instructions, submissions, experimental data and analysis is available at <https://github.com/samplchallenges/SAMPL6>

## 274 6 Overview of supplementary information

275 Organized in SI document:

- 276 • TABLE SI 1: ???

277 Extra files:

- 278 • Any extra files

## 279 7 Author Contributions

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

## 283 8 Acknowledgments

284 Complete acknowledgments section. Caitlin Bannan, Thomas Fox

285 MI, ASR, and JDC acknowledge support from the Sloan Kettering Institute. JDC acknowledges support from NIH grant P30  
286 CA008748. MI acknowledges Doris J. Hutchinson Fellowship. We thank Brad Sherborne for his valuable insights at the conception  
287 of the pK<sub>a</sub> challenge and connecting us with Timothy Rhodes and Dorothy Levorse who were able to provide resources and  
288 expertise for experimental measurements performed at MRL. We acknowledge Paul Czodrowski who provided feedback on  
289 multiple stages of this work: challenge construction, purchasable compound selection and manuscript. MI, ASR, AR and JDC are  
290 grateful to OpenEye Scientific for providing a free academic software license for use in this work.

291 Mike Chui

## 292 9 Disclosures

293 JDC is a member of the Scientific Advisory Board for Schrödinger, LLC. DLM is a member of the Scientific Advisory Board of  
294 OpenEye Scientific Software.

295 Table ref: [3, 4, 7, 8, 10] trial: [], +, -, \*, #

## 296 References

- 297 [1] Işık M, Levorse D, Rustenburg AS, Ndukwe IE, Wang H, Wang X, Reibarkh M, Martin GE, Makarov AA, Mobley DL, Rhodes T, Chodera JD.  
298 pKa Measurements for the SAMPL6 Prediction Challenge for a Set of Kinase Inhibitor-like Fragments. Journal of Computer-Aided Molecular  
299 Design. 2018 Oct; 32(10):1117–1138. doi: 10.1007/s10822-018-0168-0.
- 300 [2] OpenEye pKa Prospector;. OpenEye Scientific Software, Santa Fe, NM. Accessed on Jan 23, 2018. <https://www.eyesopen.com/pka-prospector>.
- 301 [3] ACD/pKa GALAS (ACD/Percepta Kernel v1.6);. Advanced Chemistry Development, Inc., Toronto, ON, Canada, 2018. <https://www.acdlabs.com/products/percepta/predictors/pKa/>.
- 302
- 303 [4] ACD/pKa Classic (ACD/Percepta Kernel v1.6);. Advanced Chemistry Development, Inc., Toronto, ON, Canada, 2018. <https://www.acdlabs.com/products/percepta/predictors/pKa/>.
- 304
- 305 [5] Shelley JC, Cholleti A, Frye LL, Greenwood JR, Timlin MR, Uchimaya M. Epik: A Software Program for pK<sub>a</sub> Prediction and Protonation State  
306 Generation for Drug-like Molecules. Journal of Computer-Aided Molecular Design. 2007 Dec; 21(12):681–691. doi: 10.1007/s10822-007-  
307 9133-z.
- 308 [6] Bannan CC, Mobley DL, Skillman AG. SAMPL6 Challenge Results from \$\$pK<sub>a</sub> Predictions Based on a General Gaussian Process Model.  
309 Journal of Computer-Aided Molecular Design. 2018 Oct; 32(10):1165–1177. doi: 10.1007/s10822-018-0169-z.

- 310 [7] Simulations Plus ADMET Predictor v8.5;. Simulations Plus, Lancaster, CA, 2018. [https://www.simulations-plus.com/software/admetpredictor/](https://www.simulations-plus.com/software/admetpredictor/physicochemical-biopharmaceutical/)  
311 [physicochemical-biopharmaceutical/](https://www.simulations-plus.com/software/admetpredictor/physicochemical-biopharmaceutical/).
- 312 [8] Chemicalize v18.23 (ChemAxon MarvinSketch v18.23); ChemAxon, Budapest, Hungary, 2018. [https://docs.chemaxon.com/display/docs/](https://docs.chemaxon.com/display/docs/pKa+Plugin)  
313 [pKa+Plugin](https://docs.chemaxon.com/display/docs/pKa+Plugin).
- 314 [9] Milletti F, Storch L, Sforza G, Cruciani G. New and Original  $pK_a$  Prediction Method Using Grid Molecular Interaction Fields. Journal of  
315 Chemical Information and Modeling. 2007 Nov; 47(6):2172–2181. doi: 10.1021/ci700018y.
- 316 [10] MoKa;. Molecular Discovery, Hertfordshire, UK, 2018. <https://www.moldiscovery.com/software/moka/>.
- 317 [11] Zeng Q, Jones MR, Brooks BR. Absolute and Relative  $pK_a$  Predictions via a DFT Approach Applied to the SAMPL6 Blind Challenge. Journal  
318 of Computer-Aided Molecular Design. 2018 Oct; 32(10):1179–1189. doi: 10.1007/s10822-018-0150-x.
- 319 [12] Bochevarov AD, Harder E, Hughes TF, Greenwood JR, Braden DA, Philipp DM, Rinaldo D, Halls MD, Zhang J, Friesner RA. Jaguar: A High-  
320 Performance Quantum Chemistry Software Program with Strengths in Life and Materials Sciences. International Journal of Quantum  
321 Chemistry. 2013 Sep; 113(18):2110–2142. doi: 10.1002/qua.24481.
- 322 [13] Selwa E, Kenney IM, Beckstein O, Iorga BI. SAMPL6: Calculation of Macroscopic  $pK_a$  Values from Ab Initio Quantum Mechanical Free  
323 Energies. Journal of Computer-Aided Molecular Design. 2018 Oct; 32(10):1203–1216. doi: 10.1007/s10822-018-0138-6.
- 324 [14] Tielker N, Eberlein L, Güssregen S, Kast SM. The SAMPL6 Challenge on Predicting Aqueous  $pK_a$  Values from EC-RISM Theory. Journal of  
325 Computer-Aided Molecular Design. 2018 Oct; 32(10):1151–1163. doi: 10.1007/s10822-018-0140-z.
- 326 [15] Klamt A, Eckert F, Diedenhofen M, Beck ME. First Principles Calculations of Aqueous  $pK_a$  Values for Organic and Inorganic Acids Using  
327 COSMO-RS Reveal an Inconsistency in the Slope of the  $pK_a$  Scale. The Journal of Physical Chemistry A. 2003 Nov; 107(44):9380–9386. doi:  
328 10.1021/jp034688o.
- 329 [16] Eckert F, Klamt A. Accurate Prediction of Basicity in Aqueous Solution with COSMO-RS. Journal of Computational Chemistry. 2006 Jan;  
330 27(1):11–19. doi: 10.1002/jcc.20309.
- 331 [17] Pracht P, Wilcken R, Udvarhelyi A, Rodde S, Grimme S. High Accuracy Quantum-Chemistry-Based Calculation and Blind Prediction of  
332 Macroscopic  $pK_a$  Values in the Context of the SAMPL6 Challenge. Journal of Computer-Aided Molecular Design. 2018 Oct; 32(10):1139–  
333 1149. doi: 10.1007/s10822-018-0145-7.
- 334 [18] Prasad S, Huang J, Zeng Q, Brooks BR. An Explicit-Solvent Hybrid QM and MM Approach for Predicting  $pK_a$  of Small Molecules in SAMPL6  
335 Challenge. Journal of Computer-Aided Molecular Design. 2018 Oct; 32(10):1191–1201. doi: 10.1007/s10822-018-0167-1.

## 10 Supplementary Information

MI: Figure [typeIII-error-dist-by-method] Distribution of prediction errors for each method in SAMPL6 Challenge. Analyses was performed based on Hungarian matching algorithm. Y-axis labels indicate submission IDs of each method.

MI: [pKa-error-vs-pKa-value]. Error in pKa predictions does not correlate with the true value of pKa. Left figure was constructed using closest match between experimental and predicted pKas. Y-axis is absolute residuals of the pKa prediction.

MI: FIGURE [desc-vs-MAE-correlation]. There is no clear correlation between molecular descriptors and mean absolute error for each molecule when calculated for all methods.

MI: SI Table: Type I collection

MI: SI Table: Type III collection

MI: SI Figure: type I correlation plots of each method

MI: SI Figure: type III correlation plots of each method

MI: TABLE: InChI and SMILES for chemicals

MI: TABLE: Statistics based on hungarian matching

MI: TABLE: Statistics based on microstate matching

MI: TABLE: NMR determined microstates of 8 molecules