

¹ Overview of the SAMPL6 pK_a Challenge: ² Evaluating macroscopic and microscopic pK_a ³ predictions of small molecules

⁴ Mehtap Işık (ORCID: [0000-0002-6789-952X](#))^{1,2*}, Ariën S. Rustenburg (ORCID: [0000-0002-3422-0613](#))^{1,3}, Andrea
⁵ Rizzi (ORCID: [0000-0001-7693-2013](#))^{1,4}, M. R. Gunner (ORCID: [0000-0003-1120-5776](#))⁶, David L. Mobley (ORCID:
⁶ [0000-0002-1083-5533](#))⁵, John D. Chodera (ORCID: [0000-0003-0542-119X](#))¹

⁷ ¹Computational and Systems Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center,
⁸ New York, NY 10065, United States; ²Tri-Institutional PhD Program in Chemical Biology, Weill Cornell Graduate
⁹ School of Medical Sciences, Cornell University, New York, NY 10065, United States; ³Graduate Program in
¹⁰ Physiology, Biophysics, and Systems Biology, Weill Cornell Medical College, New York, NY 10065, United States;
¹¹ ⁴Tri-Institutional PhD Program in Computational Biology and Medicine, Weill Cornell Graduate School of Medical
¹² Sciences, Cornell University, New York, NY 10065, United States; ⁵Department of Pharmaceutical Sciences and
¹³ Department of Chemistry, University of California, Irvine, Irvine, California 92697, United States; ⁶Department of
¹⁴ Physics, City College of New York, New York NY 10031

¹⁵ *For correspondence:
¹⁶ mehtap.isik@choderlab.org (MI)

17

¹⁸ Abstract

The prediction of acid dissociation constants (pK_a) is a prerequisite for predicting many other properties of a small molecule, such as its protein-ligand binding affinity, distribution coefficient ($\log D$), membrane permeability, and solubility. The prediction of each of these properties requires knowledge of the relevant protonation states and solution free energy penalties of each state. The SAMPL6 pK_a Challenge was the first time that a separate challenge was conducted for evaluating pK_a predictions as a part of the Statistical Assessment of Modeling of Proteins and Ligands (SAMPL). This challenge was motivated by the inaccuracies observed in prior physical property prediction challenges, such as the SAMPL5 $\log D$ Challenge, caused by protonation state and pK_a prediction issues. The goal of the pK_a challenge was to assess the performance of contemporary pK_a prediction methods for drug-like molecules. The challenge set was composed of 24 small molecules that resembled fragments of kinase inhibitors, a number of which were multiprotic. Eleven research groups contributed blind predictions for a total of 37 pK_a distinct prediction methods. In addition to blinded submissions, four widely used pK_a prediction methods were included in the analysis as reference methods. Collecting both microscopic and macroscopic pK_a predictions allowed in-depth evaluation of pK_a prediction performance. This article highlights deficiencies of typical pK_a prediction evaluation approaches when the distinction between microscopic and macroscopic pK_a s is ignored; in particular, we suggest more stringent evaluation criteria for microscopic and macroscopic pK_a predictions guided by the available experimental data. Top-performing submissions for macroscopic pK_a predictions achieved RMSE of 0.7–1.0 pK_a units and included both quantum chemical and empirical approaches, where the total number of extra or missing macroscopic pK_a s predicted by these submissions were fewer than 8 for 24 molecules. A large number of submissions had RMSE spanning 1–3 pK_a units. Molecules with sulfur-containing heterocycles or iodo and bromo groups were less accurately predicted on average considering all methods evaluated. For a subset of molecules, we utilized experimentally-determined microstates based on NMR to evaluate the dominant tautomer predictions for each macroscopic state. Prediction of dominant tautomers was a major source of error for microscopic pK_a predictions, especially errors in charged tautomers. The inaccuracy of pK_a predictions observed in this challenge is detrimental to the protein-ligand binding affinity predictions due to errors in dominant protonation state predictions and calculation of free energy corrections for multiple protonation states. Underestimation of ligand pK_a by 1 unit can lead to errors in binding free energy errors up to 1.2

42 kcal/mol. The SAMPL6 pK_a Challenge demonstrated the need for improving pK_a prediction methods for drug-like molecules,
43 especially for challenging moieties and multiprotic molecules.

44

45 0.1 Keywords

46 SAMPL · blind prediction challenge · acid dissociation constant · pK_a · small molecule · macroscopic pK_a · microscopic pK_a · macro-
47 scopic protonation state · microscopic protonation state

48 0.2 Abbreviations

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

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

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

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

53 **SEM** Standard error of the mean

54 **RMSE** Root mean squared error

55 **MAE** Mean absolute error

56 **τ** Kendall's rank correlation coefficient (Tau)

57 **R²** Coefficient of determination (R-Squared)

58 **MPSC** Multiple protonation states correction for binding free energy

59 **DL** Database Lookup

60 **LFER** Linear Free Energy Relationship

61 **QSPR** Quantitative Structure-Property Relationship

62 **ML** Machine Learning

63 **QM** Quantum Mechanics

64 **LEC** Linear Empirical Correction

65 1 Introduction

66 The acid dissociation constant (K_a) describes the protonation state equilibrium of a molecule given pH. More commonly, we
67 refer to $pK_a = -\log_{10} K_a$, its negative logarithmic form. Predicting pK_a is a prerequisite for predicting many other properties of
68 small molecules such as their protein binding affinity, distribution coefficient ($\log D$), membrane permeability, and solubility. As a
69 major aim of computer-aided drug design (CADD) is to aid in the assessment of pharmaceutical and physicochemical properties
70 of virtual molecules prior to synthesis to guide decision-making, accurate computational pK_a predictions are required in order
71 to accurately model numerous properties of interest to drug discovery programs.

72 Ionizable sites are found often in drug molecules and influence their pharmaceutical properties including target affinity,
73 ADME/Tox, and formulation properties [1]. It has been reported that most drugs are ionized in the range of 60-90% at physiolog-
74 ical pH [2]. Drug molecules with titratable groups can exist in many different charge and protonation states based on the pH of
75 the environment. Given that experimental data of protonation states and pK_a are often not available, we rely on predicted pK_a
76 values to determine which charge and protonation states the molecules populate and the relative populations of these states,
77 so that we can assign the appropriate dominant protonation state(s) in fixed-state calculations or the appropriate solvent state
78 weights/protonation penalty to calculations considering multiple states.

79 The pH of the human gut ranges between 1-8, and 74% of approved drugs can change ionization state within this physio-
80 logical pH range [3]. Because of this, pK_a values of drug molecules provide essential information about their physicochemical
81 and pharmaceutical properties. A wide distribution of acidic and basic pK_a values, ranging from 0 to 12, have been observed in
82 approved drugs [1, 3].

83 Drug-like molecules present difficulties for pK_a prediction compared to simple monoprotic molecules. Drug-like molecules
84 are frequently multiprotic, have large conjugated systems, often contain heterocycles, and can tautomerize. In addition, drug-
85 like molecules with significant conformational flexibility can form intramolecular hydrogen bonding, which can significantly shift
86 their pK_a values compared to molecules that cannot form intramolecular hydrogen bonds. This presents further challenges for
87 modeling methods, where deficiencies in solvation models may mispredict the propensity for intramolecular hydrogen bond

88 formation.

89 Accurately predicting pK_a s of drug-like molecules accurately is a prerequisite for computational drug discovery and design.
90 Small molecule pK_a predictions can influence computational protein-ligand binding affinities in multiple ways. Errors in pK_a
91 predictions can cause modeling the wrong charge and tautomerization states which affect hydrogen bonding opportunities
92 and charge distribution within the ligand. The dominant protonation state and relative populations of minor states in aqueous
93 medium is dictated by the molecule's pK_a values. The relative free energy of different protonation states in the aqueous state
94 is a function of pH, and contributes to the overall protein-ligand affinity in the form of a free energy penalty of reaching higher
95 energy protonation states [4]. Any error in predicting the free energy of a minor aqueous protonation state of a ligand that
96 dominates the complex binding free energy will directly add to the error in the predicted binding free energy. Similarly for log D
97 predictions, an inaccurate prediction of protonation states and their relative free energies will be detrimental to the accuracy of
98 transfer free energy predictions.

99 For a monoprotic weak acid (HA) or base (B) dissociation equilibria shown in Equation 1, the acid dissociation constant is
100 expressed as in Equation 2, or, commonly, in its negative base-10 logarithmic form as in Equation 3. The ratio of ionization
101 states can be calculated with Henderson-Hasselbalch equations shown in Equation 4.



$$K_a = \frac{[A^-][H^+]}{[HA]} ; K_a = \frac{[B][H^+]}{[B^+]} \quad (2)$$

$$pK_a = -\log_{10} K_a \quad (3)$$

$$pH = pK_a + \log_{10} \frac{[A^-]}{[HA]} ; pH = pK_a + \log_{10} \frac{[B]}{[BH^+]} \quad (4)$$

102 For multiprotic molecules, the definition of pK_a diverges into macroscopic pK_a and microscopic pK_a [5-7]. Macroscopic pK_a
103 describes the equilibrium dissociation constant between different charged states of the molecule. Each charge state can be
104 composed of multiple tautomers. Macroscopic pK_a is about the deprotonation of the molecule, rather than the location of the
105 titratable group. A microscopic pK_a describes the acid dissociation equilibrium between individual tautomeric states of different
106 charges. (There is no pK_a defined between tautomers of the same charge as they have the same number of protons and their
107 relative populations are independent of pH.) The microscopic pK_a determines the identity and distribution of tautomers within
108 each charge state. Thus, each macroscopic charge state of a molecule can be composed of multiple microscopic tautomeric
109 states. The microscopic pK_a value defined between two microstates captures the deprotonation of a single titratable group with
110 a fixed background protonation state of other titratable groups. In molecules with multiple titratable groups, the protonation
111 state of one group can affect the proton dissociation propensity of another functional group, therefore the same titratable group
112 may have different proton affinities (microscopic pK_a values) based on the protonation state of the rest of the molecule.

113 Different experimental methods are sensitive to changes in the total charge or the location of individual protons, so they
114 measure different definitions of pK_a s, as explained in more detail in prior work [8]. Most common pK_a measurement techniques
115 such as potentiometric and spectrophotometric methods measure macroscopic pK_a s, while NMR measurements can determine
116 microscopic pK_a s by measuring microstate populations with respect to pH. Therefore, it is important to pay attention to the
117 source and definition of pK_a values to interpret their meaning correctly.

118 Many computational methods can predict both microscopic and macroscopic pK_a s. While experimental measurements more
119 often provide only macroscopic pK_a s, microscopic pK_a predictions are more informative for determining relevant microstates
120 (tautomers) of a molecule and their relative free energies. Predicted microstate populations can be converted to predicted
121 macroscopic pK_a s for direct comparison with experimentally obtained macroscopic pK_a s. In this paper, we explore approaches
122 to assess the performance of both macroscopic and microscopic pK_a predictions, taking advantage of available experimental
123 data.

Microscopic pK_a predictions can be converted to macroscopic pK_a predictions either directly with Equation 5 [9],

$$K_a^{\text{macro}} = \sum_{j=1}^{N_{\text{deprot}}} \frac{1}{\sum_{i=1}^{N_{\text{prot}}} \frac{1}{K_{ij}^{\text{micro}}}} , \quad (5)$$

124 or through computing the macroscopic free energy of deprotonation between ionization states with charges N and $N - 1$ via
125 Boltzmann-weighted sum of the relative free energy of microstates (G_i) as in Equations 6 and 7 [10].

$$\Delta G_{N-1,N} = RT \ln \frac{\sum_i e^{-G_i/RT} \delta_{N_i, N-1}}{\sum_i e^{-G_i/RT} \delta_{N_i, N}} \quad (6)$$

$$pK_a = pH - \frac{\Delta G_{N-1,N}}{RT \ln 10} \quad (7)$$

126 In Equation 6 $\Delta G_{N-1,N}$ is the effective macroscopic protonation free energy. $\delta_{N_i, N-1}$ is equal to 1 when the microstate i has a
127 total charge of $N-1$ and null otherwise. $\delta_{N_i, N}$ is equal to 1 when the microstate i has a total charge of N and null otherwise. RT is
128 the ideal gas constant times the temperature.

129 1.1 Motivation for a blind pK_a challenge

130 SAMPL (Statistical Assessment of the Modeling of Proteins and Ligands) is a series of annual computational prediction challenges
131 for the computational chemistry community. The goal of SAMPL community is to evaluate the current performance of the models
132 and to bring the attention of the quantitative biomolecular modeling field on problems that limit the accuracy of protein-ligand
133 binding models. We want to enable computer-aided drug discovery to make sustained progress toward higher accuracy by
134 focusing the community on critical challenges that isolate one accuracy-limiting problem at a time. By conducting a series of
135 blind challenges—which often feature the computation of specific physical properties critical for protein-ligand modeling—and
136 encouraging rapid sharing of lessons learned, SAMPL aims to accelerate progress toward quantitative accuracy in modeling.

137 SAMPL Challenges that focus on different physical properties so far have assessed intermolecular binding models of various
138 protein-ligand and host-guest systems, as well as the prediction of hydration free energies and distribution coefficients. These
139 blind challenges motivate improvements in computational methods by revealing unexpected sources of error, identifying fea-
140 tures of methods that perform well or poorly, and enabling the participants to share information after each successive challenge.
141 Previous SAMPL Challenges have focused on the limitations of force field accuracy, finite sampling, solvation modeling defects,
142 and tautomer/protonation state predictions on protein-ligand binding predictions.

143 During the SAMPL5 log D Challenge, the performance of models in predicting cyclohexane-water log D was worse than ex-
144 pected and accuracy suffered when protonation states and tautomers were not taken into account [11, 12]. Many participants
145 simply submitted log P predictions as if they were equivalent to log D , and many were not prepared to account for the con-
146 tributions of different ionization states to the distribution coefficient in their models. Challenge results highlighted that log P
147 predictions were not an accurate approximation of log D without capturing protonation state effects. The calculations were
148 improved by including free energy penalty of the neutral state which relies on obtaining an accurate pK_a prediction [11]. With
149 the goal of deconvoluting the different sources of error contributing to the large errors observed in the SAMPL5 log D Challenge,
150 we organized separate pK_a and log P challenges in SAMPL6 [8, 13, 14]. For this iteration of the SAMPL challenge, we isolated the
151 problem of predicting aqueous protonation states.

152 This is the first time a blind pK_a prediction challenge has been fielded as part of SAMPL. In this challenge, we aimed to
153 assess the performance of current pK_a prediction methods for drug-like molecules, investigate potential causes of inaccurate
154 pK_a estimates, and determine how much current level of accuracy might impact protein binding affinity predictions.

155 1.2 Approaches to predict small molecule pK_a s

156 There are a large variety of pK_a prediction methods developed for the prediction of aqueous pK_a s of small molecules. Broadly,
157 we can divide pK_a predictions as knowledge-based empirical methods and physical methods. Empirical methods include the
158 following categories: Database Lookup (DL) [15], Linear Free Energy Relationship (LFER) [16–18], Quantitative Structure-Property
159 Relationship (QSPR) [19–22], and Machine Learning (ML) approaches [23, 24]. DL methods rely on the principle that structurally
160 similar compounds have similar pK_a values and utilize an experimental database of complete structures or fragments. The pK_a
161 values of the most similar database entries are reported as the predicted pK_a of the query molecule. In the QSPR approach,
162 the pK_a values are predicted as a function of various quantitative molecular descriptors, and the parameters of the function are
163 trained on experimental datasets. A function in the form of multiple linear regression is common, although more complex forms
164 can also be used such as the artificial neural networks in ML methods. The LFER approach is the oldest pK_a prediction strategy.
165 They use Hammett-Taft type equations to predict pK_a based on classification of the molecule to a parent class (associated with

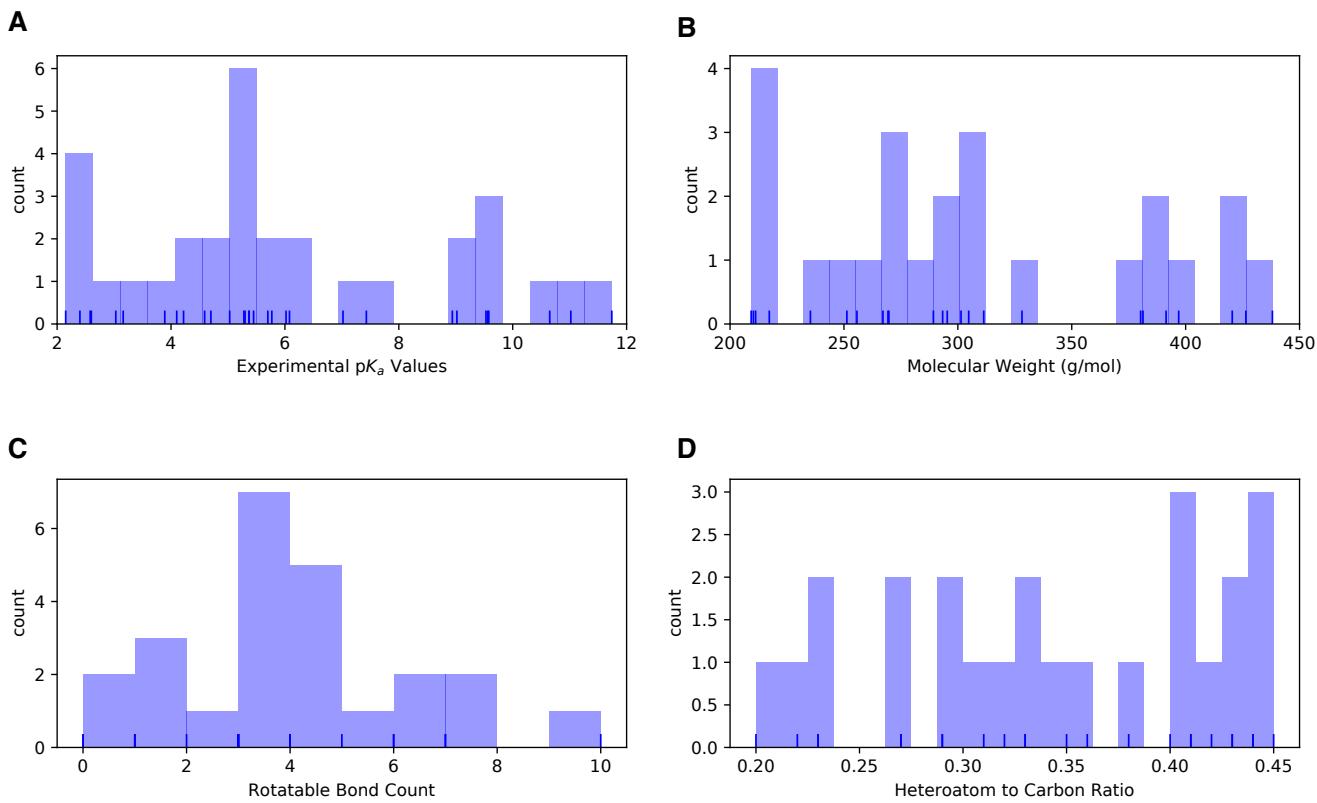


Figure 1. Distribution of molecular properties of 24 compounds in SAMPL6 pK_a Challenge. **A** Histogram of spectrophotometric pK_a measurements collected with Sirius T3 [8]. The overlayed rug plot indicates the actual values. Five compounds have multiple measured pK_a s in the range of 2–12. **B** Histogram of molecular weights calculated for the neutral state of the 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 atoms including O, N, F, S, Cl, Br, I) count to the number of carbon atoms.

166 a base pK_a value) and two parameters that describe how the base pK_a value must be modified given its substituents. Physical
 167 modeling of pK_a predictions requires Quantum Mechanics (QM) models. QM methods are often utilized together with linear
 168 empirical corrections (LEC) that are designed to rescale and unbias QM predictions for better accuracy. Classical molecular
 169 mechanics-based pK_a prediction methods are not feasible as deprotonation is a covalent bond breaking event that can only
 170 be captured by QM. Constant-pH molecular dynamics methods can calculate pK_a shifts in large biomolecular systems where
 171 there is low degree of coupling between protonation sites and linear summation of protonation energies can be assumed [25].
 172 However, this approach can not be applied to small organic molecule due to the high degree of coupling between protonation
 173 sites.

174 2 Methods

175 2.1 Design and logistics of the SAMPL6 pK_a Challenge

176 The SAMPL6 pK_a Challenge was conducted as a blind prediction challenge and focused on predicting aqueous pK_a values of 24
 177 small molecules not previously reported in the literature. The challenge set was composed of molecules that resemble fragments
 178 of kinase inhibitors. Heterocycles that are frequently found in FDA-approved kinase inhibitors were represented in this set. The
 179 compound selection process was described in depth in the prior publication reporting SAMPL6 pK_a Challenge experimental data
 180 collection [8]. The distribution of molecular weights, experimental pK_a values, number of rotatable bonds, and heteroatom to
 181 carbon ratio are depicted in Fig. 1. The challenge molecule set was composed of 17 small molecules with limited flexibility (less

than 5 non-terminal rotatable bonds) and 7 molecules with 5–10 non-terminal rotatable bonds. The distribution of experimental pK_a values was roughly uniform between 2–12. 2D representations of all compounds are provided in Fig. 5. Drug-like molecules are often larger and more complex than the ones used in this study. We limited the size and the number of rotatable bonds of compounds to create molecule set of intermediate difficulty.

The dataset composition and details of the pK_a measurement technique without the identity of the small molecules were announced approximately one month before the challenge start date. Experimental macroscopic pK_a measurements were collected with the spectrophotometric method of Sirius T3, at room temperature, in ionic strength-adjusted water with 0.15 M KCl [8]. The instructions for participation and the identity of the challenge molecules were released on the challenge start date (October 25, 2017). A table of molecule IDs (in the form of SM##) and their canonical isomeric SMILES was provided as input. Blind prediction submissions were accepted until January 22, 2018.

Following the conclusion of the blind challenge, the experimental data was made public on January 23, 2018. The SAMPL organizers and participants gathered at the Second Joint D3R/SAMPL Workshop, at UC San Diego, La Jolla, CA on February 22–23, 2018 to share results. The workshop aimed to create an opportunity for participants to have discussions, evaluate the results and lessons of the challenge together. Participants reported their results and their own evaluations in a special issue of the Journal of Computer-Aided Molecular Design [26].

While designing this first pK_a prediction challenge, we did not know the optimal format to capture pK_a predictions of participants. We wanted to capture all necessary information that will aid the evaluation of pK_a predictions at the submission stage. Our strategy was to directly evaluate macroscopic pK_a predictions comparing them to experimental macroscopic pK_a values and to use collected microscopic pK_a prediction data for more in-depth diagnostics of method performance. Therefore, we asked participants to submit their predictions in three different submission types:

- **Type I:** microscopic pK_a values and related microstate pairs
- **Type II:** fractional microstate populations as a function of pH in 0.1 pH increments
- **Type III:** macroscopic pK_a values

For each submission type, a machine-readable submission file template was specified. For type I submissions, participants were asked to report the microstate ID of the protonated state, the microstate ID of deprotonated state, the microscopic pK_a , and the predicted microscopic pK_a standard error of the mean (SEM). The method of microstate enumeration and why it was needed are discussed further in Section 2.2 "Enumeration of Microstates". The SEM aims to capture the statistical uncertainty of the prediction method. Microstate IDs were preassigned identifiers for each microstate in the form of SM##_micro##. For type II submissions, the submission format included a table that started with a microstate ID column and a set of columns reporting the natural logarithm of fractional microstate population values of each predicted microstate for 0.1 pH increments between pH 2 and 12. For type III submissions participants were asked to report molecule ID, macroscopic pK_a , and macroscopic pK_a SEM.

It was mandatory to submit predictions for all fields for each prediction, but it was not mandatory to submit predictions for all the molecules or all the submission types. Although we accepted submissions with partial sets of molecules, it would have been a better choice to require predictions for all the molecules for a better comparison of overall method performance. 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.

Participants were allowed to submit predictions for multiple methods as long as they created separate submission files. While anonymous participation was allowed, all participants opted to make their submissions public. Blind submissions were assigned a unique 5-digit alphanumeric submission ID, which will be used throughout this paper. Unique IDs were also assigned when multiple submissions exist for different submissions types of the same method such as microscopic pK_a (type I) and macroscopic pK_a (type III). These submission IDs were also reported in the evaluation papers of participants and allow cross-referencing. Submission IDs, participant provided method names, and method categories are presented in Table 1. In many cases, multiple types of submissions of the same method were provided by participants as challenge instructions requested. Although each prediction set was assigned a separate submission ID, we have matched the submissions that originated from the same method according to the reports of the participants, for cases where multiple sets of predictions came from a given method. Submission IDs for both macroscopic (type III) and microscopic (type I) pK_a predictions for each method are shown in Table 1.

2.2 Enumeration of microstates

To capture both the pK_a value and titration position of microscopic pK_a predictions, we needed microscopic pK_a values to be reported together with a pair of microstates which describe the protonated and deprotonated states of each microscopic tran-

sition. String representations of molecules such as canonical SMILES with explicit hydrogens can be written, however, there can be inconsistencies between the interpretation of canonical SMILES written by different software and algorithms. To avoid complications while reading microstate structure files from different sources, we decided that the safest route was pre-enumerating all possible microstates of challenge compounds, assigning microstate IDs to each in the form of SM##_micro##, and requiring participants to report microscopic pK_a values along with microstate pairs specified by the provided microstates IDs.

We created initial sets of microstates with Epik [27] and OpenEye QUACPAC [28] and took the union of results. Microstates with Epik were generated using Schrödinger Suite v2016-4, running Epik to enumerate all tautomers within 20 pK_a units of pH 7. For enumerating microstates with OpenEye QUACPAC, we had to first enumerate formal charges and for each charge enumerate all possible tautomers using the settings of maximum tautomer count 200, level 5, with carbonyl hybridization set to False. Then we created a union of all enumerated states written as canonical isomeric SMILES generated by OpenEye OEChem [29]. Even though resonance structures correspond to different canonical isomeric SMILES they are not different microstates, therefore it was necessary to remove resonance structures that were replicates of the same tautomer. To detect equivalent resonance structures, we converted canonical isomeric SMILES to InChI hashes with explicit and fixed hydrogen layer. Structures that describe the same tautomer but different resonance states lead to explicit hydrogen InChI hashes that are identical, allowing replicates to be removed. The Jupyter Notebook used for the enumeration of microstates is provided in supplementary information. We provided microstate ID tables with canonical SMILES and 2D-depictions to aid participants in matching predicted structures to microstate IDs. Canonical SMILES representation was selected over canonical isomeric SMILES, because resonance and geometric isomerism do not lead to different microstates according to our working microstate definition. The only exception was for molecules SM20 which should be consistently modeled as the E-isomer.

During the course of the SAMPL6 Challenge, participants identified new microstates that were not present in the initial list that we provided. Despite combining enumerated charge states and tautomers generated by both Epik and OpenEye QUACPAC, to our surprise, the microstate lists were still incomplete. Based on participant requests for new microstates, we iteratively had to update the list of microstates and assign new microstate IDs. Every time we received a request, we shared the updated microstate ID lists with all challenge participants. Some participants updated their pK_a prediction by including the newly added microstates in their calculations. In the future, developing a better algorithm that can enumerate all possible microstates (not just the ones with significant populations) would be very beneficial for anticipating microstates that may be predicted by pK_a prediction methods.

A microscopic pK_a definition was provided in challenge instructions for clarity as follows: Physically meaningful microscopic pK_a s are defined between microstate pairs that can interconvert by single protonation/deprotonation event of only one titrable group. So, microstate pairs should have total charge (absolute) difference of 1 and only one heavy atom that differs in the number of bound hydrogens, regardless of resonance state or geometric isomerism. All geometric isomer and resonance structure pairs that have the same number of hydrogens bound to equivalent heavy atoms are grouped into the same microstate. Pairs of resonance structures and geometric isomers (cis/trans, stereo) are not considered as different microstates, as long as there is no change in the number of hydrogens bound to each heavy atom. Transitions where there are shifts in the position of protons coupled to changes in the number of protons were also not considered as microscopic pK_a s [30]. Since we wanted participants to report only microscopic pK_a s that describe single deprotonation events (in contrast to transitions between microstates that are different in terms of two or more titratable protons), we have also provided a pre-enumerated list of allowed microstate pairs.

Provided microstate ID and microstate pair lists were intended to be used for reporting microstate IDs and to aid parsing of submissions. The enumerated lists of microstates were not created with the intent to guide computational predictions. This was clearly stated in the challenge instructions. However, we noticed that some participants still used the microstate lists as an input for their pK_a predictions as we received complaints from participants that due to our updates to microstate lists they needed to repeat their calculations. This would not have been an issue if participants used pK_a prediction protocols that did not rely on an external pre-enumerated list of microstates as an input. None of the participants reported this dependency in their method descriptions explicitly, so it was also not obvious how participants were using the provided states in their predictions. We could not identify which submissions used these enumerated microstate lists as input for predictions and which have followed the challenge instructions and relied only on their prediction method to generate microstates.

2.3 Evaluation approaches

Since the experimental data for the challenge was mainly composed of macroscopic pK_a values of both monoprotic and multiprotic compounds, evaluation of macroscopic and microscopic pK_a predictions was not straightforward. For a subset of 8

molecules, the dominant microstate sequence could be inferred from NMR. For the rest of the molecules, the only experimental information available was the macroscopic pK_a value. The experimental data in the form of macroscopic pK_a s did not provide any information on which group(s) are being titrated, the microscopic pK_a values, the identity of the associated macrostates (which total charge) or microstates (which tautomers). Also, experimental data did not provide any information about the charge state of protonated and deprotonated species associated with each macroscopic pK_a . Typically charges of states associated with experimental pK_a values are assigned based on pK_a predictions, not experimental evidence, but we did not utilize such computational charge assignment. For a fair performance comparison between methods, we avoided relying on any particular pK_a prediction to assist the interpretation of the experimental reference data. This choice complicated the pK_a prediction analysis, especially regarding how to pair experimental and predicted pK_a s for error analysis. We adopted various evaluation strategies guided by the experimental data. To compare macroscopic pK_a predictions to experimental values we had to utilize numerical matching algorithms before we could calculate performance statistics. For the subset of molecules with experimental data about microstates, we used microstate based matching. These matching methods were described further in the next section.

Three types of submissions were collected during the SAMPL6 pK_a Challenge. We have only utilized the type I (microscopic pK_a value and microstate IDs) and the type III (macroscopic pK_a value) predictions in this article. Type I submissions contained the same prediction information as the type II submissions which reported the fractional population of microstates with respect to pH. We collected type II submissions in order to capture relative populations of microstates, not realizing they were redundant. The microscopic pK_a predictions collected in type I submissions capture all the information necessary to calculate type II submissions. Therefore, we did not use type II submissions for challenge evaluation. In theory, type III (macroscopic pK_a) predictions can also be calculated from type I submissions, but collecting type III submissions allowed the participation of pK_a prediction methods that directly predict macroscopic pK_a s without considering microspeciation and methods that apply special empirical corrections for macroscopic pK_a predictions.

2.3.1 Matching algorithms for pairing predicted and experimental pK_a s

Macroscopic pK_a predictions can be calculated from microscopic pK_a s for direct comparison to experimental macroscopic pK_a values. One major question must be answered to allow this comparison: How should we match predicted macroscopic pK_a s to experimental macroscopic pK_a s when there could multiple pK_a values reported for a given molecule? For example, experiments on SM18 showed three macroscopic pK_a s, but prediction of *xvxzd* method reported two macroscopic pK_a values. There were also examples of the opposite situation with more predicted pK_a values than experimentally determined macroscopic pK_a s: One experimental pK_a was measured for SM02, but two macroscopic pK_a s were predicted by *xvxzd* method. The experimental and predicted values must be paired before any prediction error can be calculated, even though there was not any experimental information regarding underlying tautomer and charge states.

Knowing the charges of macrostates would have guided the pairing between experimental and predicted macroscopic pK_a s, however, not all experimental pK_a measurements can determine determine the charge of protonation states. The potentiometric pK_a measurements just captures the relative charge change between macrostates, but not the absolute value of the charge. Thus, our experimental data did not provide any information that would indicate the titration site, the overall charge, or the tautomer composition of macrostate pairs that are associated with each measured macroscopic pK_a that can guide the matching between predicted and experimental pK_a values.

For evaluating macroscopic pK_a predictions taking the experimental data as reference, Fraczkiewicz et al. delineated recommendations for fair comparative analysis of computational pK_a predictions [23]. They recommended that, in the absence of any experimental information that would aid in matching, experimental and computational pK_a s should be matched preserving the order of pK_a values and minimizing the sum of absolute errors.

We picked the Hungarian matching algorithm [31, 32] to match experimental and predicted macroscopic pK_a s with a squared error cost function as suggested by Kiril Lanevskij via personal communication. The algorithm is available in the SciPy package (`scipy.optimize.linear_sum_assignment`) [33]. This matching algorithm provides optimum global assignment that minimizes the linear sum of squared errors of all pairwise matches. We selected the squared error cost function instead of the absolute error cost function to avoid misordered matches, For instance, for a molecule with experimental pK_a values of 4 and 6, and predicted pK_a s of 7 and 8, Hungarian matching with absolute error cost function would match 6 to 7 and 4 to 9. Hungarian matching with squared error cost would match 4 to 7 and 6 to 8, preserving the increasing pK_a value order between experimental and predicted values. A weakness of this approach would be failing to match the experimental value of 6 to predicted value of 7 if that was the correct match based on underlying macrostates. But the underlying pair of states were unknown to us both because the experimental data did not determine which charge states the transitions were happening between and also because

331 we did not collect the pair of macrostates associated with each pK_a predictions in submissions. Requiring this information
332 for macroscopic pK_a predictions in future SAMPL challenges would allow for better comparison between predictions, even if
333 experimental assignment of charges is not possible. There is no perfect solution to the numerical pK_a assignment problem, but
334 we tried to determine the fairest way to penalize predictions based on their numerical deviation from the experimental values.

335 For the analysis of microscopic pK_a predictions we adopted a different matching approach. For the eight molecules for which
336 we had the requisite data for this analysis, we utilized the dominant microstate sequence inferred from NMR experiments to
337 match computational predictions and experimental pK_a s. We will refer to this assignment method as microstate matching,
338 where the experimental pK_a value is matched to the computational microscopic pK_a value which was reported for the dominant
339 microstate pair observed for each transition. We have compared the results of Hungarian matching and microstate matching.

340 Inevitably the choice of matching algorithms to assign experimental and predicted values has an impact on the calculation of
341 performance statistics. We believe the Hungarian algorithm for numerical matching of unassigned pK_a values and microstate-
342 based matching when experimental microstates are known were the best choices, providing the most unbiased matching with-
343 out introducing assumptions outside of the experimental data.

344 2.3.2 Statistical metrics for submission performance

345 A variety of accuracy and correlation statistics were considered for analyzing and comparing the performance of prediction
346 methods submitted to the SAMPL6 pK_a Challenge. Calculated performance statistics of predictions were provided to partici-
347 pants before the workshop. Details of the analysis and scripts are maintained on the SAMPL6 GitHub Repository (described in
348 Section 5).

349 There are six error metrics reported for the numerical error of the pK_a values: the root-mean-squared error (RMSE), mean ab-
350 solute error (MAE), mean error (ME), coefficient of determination (R^2), linear regression slope (m), and Kendall's Rank Correlation
351 Coefficient (τ). Uncertainty in each performance statistic was calculated as 95% confidence intervals estimated by bootstrapping
352 over predictions with 10 000 bootstrap samples. Calculated errors statistics of all methods can be found in Table S2 for macro-
353 scopic pK_a predictions and Tables S4 and S4 for microscopic pK_a predictions.

354 In addition to the numerical error aspect of the pK_a values, we also evaluated predictions in terms of their ability to capture
355 the correct macrostates (ionization states) and microstates (tautomers of each ionization state) to the extent possible from the
356 available experimental data. For macroscopic pK_a s, experiments did not provide any evidence of the identity of the ionization
357 states. However, the number of ionization states indicates the number of macroscopic pK_a s that exists between the experimental
358 range of 2.0–12.0. For instance, SM14 has two experimental pK_a s and therefore three different charge states observed between
359 pH 2.0 and 12.0. If a prediction reported 4 macroscopic pK_a s, it is clear that this method predicted an extra ionization state. With
360 this perspective we reported the number of unmatched experimental pK_a s (the number of missing pK_a predictions, i.e. missing
361 ionization states) and the number of unmatched predicted pK_a s (the number of extra pK_a predictions, i.e., extra ionization states)
362 after Hungarian matching. The later count was restricted to only predictions with pK_a values between 2 and 12 because that was
363 the range of the experimental method. Errors in extra or missing pK_a prediction errors highlight failure to predict the correct
364 number of ionization states within a pH range.

365 For the evaluation of microscopic pK_a predictions, taking advantage of the available dominant microstate sequence data
366 for a subset of 8 compounds, we calculated the dominant microstate prediction accuracy which is the ratio of correct dominant
367 tautomer predictions for each charge state divided by the total number of dominant tautomer predictions. Dominant microstate
368 prediction accuracy was calculated over all experimentally detected ionization states of each molecule which were part of this
369 analysis. In order to extract the sequence of dominant microstates from the microscopic pK_a predictions sets, we calculated
370 the relative free energy of microstates selecting a neutral tautomer and pH 0 as reference following Equation 8. Calculation of
371 relative microstate free energies was explained in more detail in a previous publication [30].

372 The relative free energy of a state with respect to reference state B at pH 0.0 (arbitrary pH value selected as reference) can
373 be calculated as follows:

$$\Delta G_{AB} = \Delta m_{AB} RT \ln 10(pH - pK_a) \quad (8)$$

374 Δm_{AB} is equal to the number protons in state A minus that in state B. R and T indicate the molar gas constant and temperature,
375 respectively. By calculating relative free energies of all predicted microstates with respect to the same reference state and pH,
376 we were able to determine the sequence of predicted dominant microstates. The dominant tautomer of each charge state
377 was determined as the microstate with the lowest free energy in the subset of predicted microstates of each ionization state.
378 This approach is feasible because the relative free energy of tautomers of the same ionization state is independent of pH and
379 therefore the choice of reference pH is arbitrary.

380 We created a shortlist of top-performing methods for macroscopic and microscopic pK_a predictions. The top macroscopic pK_a
381 predictions were selected if they ranked in the top 10 consistently according to two error metrics (RMSE, MAE) and two correlation
382 metrics (R-Squared, and Kendall's Tau), while also having fewer than eight missing or extra macroscopic pK_a s for the entire
383 molecule set (eight macrostate errors correspond to macrostate prediction mistake in roughly one third of the 24 compounds).
384 These methods are presented in Table 2. A separate list of top-performing methods was constructed for microscopic pK_a with
385 the following criteria: ranking in the top 10 methods when ranked by accuracy statistics (RMSE and MAE) and perfect dominant
386 microstate prediction accuracy. These methods are presented in Table 3.

387 In addition to comparing the performance of methods, we also wanted to compare pK_a prediction performance for each
388 molecule to determine which molecules were the most challenging for pK_a predictions considering all the methods in the chal-
389 lenger. For this purpose, we plotted prediction error distributions of each molecule calculated over all prediction methods. We
390 also calculated MAE for each molecule over all prediction sets as well as for predictions from each method category separately.

391 2.4 Reference calculations

392 Including a null model is helpful in comparative performance analysis of predictive methods to establish what the performance
393 statistics look like for a baseline method for the specific dataset. Null models or null predictions employ a simple prediction
394 model which is not expected to be particularly successful, but it provides a simple point of comparison for more sophisticated
395 methods. The expectation or goal is for more sophisticated or costly prediction methods to outperform the predictions from a
396 null model, otherwise the simpler null model would be preferable. In SAMPL6 pK_a Challenge there were two blind submissions
397 using database lookup methods that were submitted to serve as null predictions. These methods, with submission IDs 5nm4j and
398 5nm4j both used OpenEye pKa-Prospector database to find the most similar molecule to query molecule and simply reported its
399 pK_a as the predicted value. Database lookup methods with a rich experimental database do present a challenging null model to
400 beat, however, due to the accuracy level needed from pK_a predictions for computer-aided drug design we believe such methods
401 provide an appropriate performance baseline that physical and empirical pK_a prediction methods should strive to outperform.

402 We also included additional reference calculations in the comparative analysis to provide more perspective. Some widely
403 used methods by academia and industry were missing from the blind challenge submission. Therefore, we included those meth-
404 ods as reference calculations: Schrödinger/Epik (nb007, nb008, nb010), Schrödinger/Jaguar (nb011, nb013), Chemaxon/Chemicalize
405 (nb015), and Molecular Discovery/MoKa (nb016, nb017). Epik and Jaguar pK_a predictions were collected by Bas Rustenburg, Chem-
406 icalize predictions by Mehtap Isik, and MoKa predictions by Thomas Fox. All were done after the challenge deadline avoiding
407 any alterations to their respective standard procedures and any guidance from experimental data. Experimental data was pub-
408 licly available before these calculations were complete, therefore reference calculations were not formally considered as blind
409 submissions.

410 All figures and statistics tables in this manuscript include reference calculations. As the reference calculations were not formal
411 submissions, these were omitted from formal ranking in the challenge, but we present plots in this article which show them for
412 easy comparison. These are labeled with submission IDs of the form nb### to clearly indicate non-blind reference calculations.

413 3 Results and Discussion

414 Participation in SAMPL6 pK_a Challenge was high with 11 research groups contributing pK_a prediction sets for 37 methods. A
415 large variety of pK_a prediction methods were represented in the SAMPL6 Challenge. We categorized these submissions into
416 four method classes: database lookup (DL), linear free energy relationship (LFER), quantitative structure-property relationship
417 or machine learning (QSPR/ML), and quantum mechanics (QM). Quantum mechanics models were subcategorized into QM
418 methods with and without linear empirical correction (LEC), and combined quantum mechanics and molecular mechanics (QM
419 + MM). Table 1 presents method names, submission IDs, method categories, and also references for each approach. Integral
420 equation-based approaches (e.g. EC-RISM) were also evaluated under the Physical (QM) category. There were 2 DL, 4 LFER, and
421 5 QSPR/ML methods represented in the challenge, including the reference calculations. The majority of QM calculations include
422 linear empirical corrections (22 methods in QM + LEC category), and only 5 QM methods were submitted without any empirical
423 corrections. There were 4 methods that used a mixed physical modeling approach of QM + MM.

424 The following sections present a detailed performance evaluation of blind submissions and reference prediction methods
425 for macroscopic and microscopic pK_a predictions. Performance statistics of all the methods can be found in Tables S2 and S4.
426 Methods are referred to by their submission ID's which are provided in Table 1.

Table 1. Submission IDs, names, category, and type for all the pKa 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 a submission was or a post-deadline reference calculation, denoted by “Blind” or “Reference” respectively. The methods in the table are grouped by method category and not ordered by performance.

Method Category	Method	Microscopic pKa (Type I) Submission ID	Macroscopic pKa (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	[34]	
DL	OpenEye pKa-Prospector 1.0.0.3 with Analog Search ion identification algorithm	<i>pwn3m</i>	Null	[34]	
LFER	ACD/pKa GALAS (ACD/Percepta Kernel v1.6)	<i>v8qph</i>	<i>37xm8</i>	Blind	[35]
LFER	ACD/pKa Classic (ACD/Percepta Kernel, v1.6)		<i>xmyhm</i>	Blind	[36]
LFER	Epik Scan (Schrödinger v2017-4)		<i>nb007</i>	Reference	[27]
LFER	Epik Microscopic (Schrödinger v2017-4)	<i>nb008</i>	<i>nb010</i>	Reference	[27]
QSPR/ML	OpenEye Gaussian Process	<i>6tvf8</i>	<i>hytjn</i>	Blind	[12]
QSPR/ML	OpenEye Gaussian Process Resampled		<i>q3pfj</i>	Blind	[12]
QSPR/ML	S+pKa (ADMET Predictor v8.5, Simulations Plus)	<i>hdijq</i>	<i>gyuhx</i>	Blind	[24]
QSPR/ML	Chemicalize v18.23 (ChemAxon MarvinSketch v18.23)		<i>nb015</i>	Reference	[37]
QSPR/ML	Moka v3.1.3	<i>nb016</i>	<i>nb017</i>	Reference	[22, 38]
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	[39]
QM	Direct scheme with single point correction: SMD/M06-2X//6-311++G(d,p)//M06-2X/6-31G(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	[39]
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	[39]
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	[39]
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	[39]
QM + LEC	Jaguar (Schrödinger v2017-4)	<i>nb011</i>	<i>nb013</i>	Reference	[40]
QM + LEC	CPCM/B3LYP/6-311+G(d,p) and global fitting	<i>y4wws</i>	<i>35bdm</i>	Blind	[10]
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	[10]
QM + LEC	EC-RISM/MP2/6-311+G(d,p)-P3NI-q-noThiols-2par	<i>kxzt</i>	<i>ds62k</i>	Blind	[41]
QM + LEC	EC-RISM/MP2/cc-pVTZ-P2-q-noThiols-2par	<i>ftc8w</i>	<i>2ii2g</i>	Blind	[41]
QM + LEC	EC-RISM/MP2/6-311+G(d,p)-P2-phi-all-2par	<i>ktpj5</i>	<i>nb001</i>	Blind*	[41]
QM + LEC	EC-RISM/MP2/6-311+G(d,p)-P2-phi-noThiols-2par	<i>wuuvc</i>	<i>nb002</i>	Blind*	[41]
QM + LEC	EC-RISM/MP2/6-311+G(d,p)-P3NI-phi-all-2par	<i>2umai</i>	<i>nb003</i>	Blind*	[41]
QM + LEC	EC-RISM/MP2/6-311+G(d,p)-P3NI-phi-noThiols-2par	<i>cm2yq</i>	<i>nb004</i>	Blind*	[41]
QM + LEC	EC-RISM/MP2/6-311+G(d,p)-P2-phi-all-1par	<i>z7fhp</i>	<i>nb005</i>	Blind*	[41]
QM + LEC	EC-RISM/MP2/6-311+G(d,p)-P3NI-phi-all-1par	<i>8toyp</i>	<i>nb006</i>	Blind*	[41]
QM + LEC	EC-RISM/MP2/cc-pVTZ-P2-phi-noThiols-2par	<i>epvmk</i>	<i>tjjd0</i>	Blind	[41]
QM + LEC	EC-RISM/MP2/cc-pVTZ-P2-phi-all-2par	<i>xnoe0</i>	<i>mkhqa</i>	Blind	[41]
QM + LEC	EC-RISM/MP2/cc-pVTZ-P3NI-phi-noThiols-2par	<i>4o0ia</i>	<i>mpwiy</i>	Blind	[41]
QM + LEC	EC-RISM/B3LYP/6-311+G(d,p)-P3NI-q-noThiols-2par	<i>nxaaw</i>	<i>ad5pu</i>	Blind	[41]
QM + LEC	EC-RISM/B3LYP/6-311+G(d,p)-P3NI-phi-noThiols-2par	<i>0xi4b</i>	<i>f0gew</i>	Blind	[41]
QM + LEC	EC-RISM/B3LYP/6-311+G(d,p)-P2-phi-noThiols-2par	<i>cwyk</i>	<i>np6b4</i>	Blind	[41]
QM + LEC	PCM/B3LYP/6-311+G(d,p)	<i>gdqeg</i>	<i>yc70m</i>	Blind	[41]
QM + LEC	COSMOtherm_FINE17 (COSMOtherm C30_1701, BP/TZVPD/FINE//BP/TZVP/COSMO)	<i>t8ewk</i>	<i>0hxtm</i>	Blind	[42, 43]
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	[44]
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	[44]
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 was 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	[44]
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	[45]
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 pKa submissions were blind, however, participant requested a correction after blind submission deadline for macroscopic pKa submissions. Therefore, these were assigned submission IDs in the form of *nb##*.

427 3.1 Analysis of macroscopic pK_a predictions

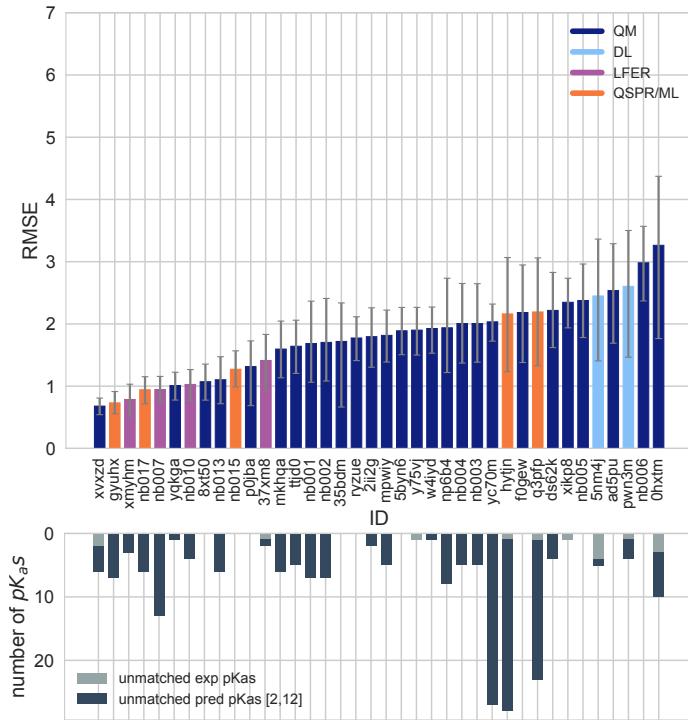


Figure 2. RMSE and unmatched pK_a counts vs. submission ID plots for macroscopic pK_a predictions based on Hungarian matching.

Methods are indicated by submission IDs. RMSE is shown with error bars denoting 95% confidence intervals obtained by bootstrapping over challenge molecules. Submissions are colored by their method categories. Light blue colored database lookup methods are utilized as the null prediction method. QM methods category (navy) includes pure QM, QM+LEC, and QM+MM approaches. 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.

The performance of macroscopic pK_a predictions were analyzed by comparison to experimental pK_a values collected by the spectrophotometric method via numerical matching following the Hungarian method. Overall pK_a prediction performance was worse than we hoped. Fig. 2 shows RMSE calculated for each prediction method represented by their submission IDs. Other performance statistics are depicted in Fig. 3. In both figures, method categories were indicated by the color of the error bars. Statistics depicted in these figures can be found in Table S2. Prediction error ranged between 0.7 to 3.2 pK_a units in terms of RMSE, while an RMSE between 2-3 log units was observed for the majority of methods (20 out of 38 methods). Only five methods achieved RMSE less than 1 pK_a unit. One is QM method with COSMO-RS approach for solvation and linear empirical correction (*xvxzd* (DSD-BLYP-D3(B)/def2-TZVPD//PBEH-3c[DCOSMO-RS] + RRHO(GFN-xTB[GBSA]) + Gsolv(COSMO-RS[TZVPD]) and linear fit)), and the remaining four are empirical prediction methods of LFER (*xmyhm* (ACD/pKa Classic), *nb007* (Schrödinger/EpiK Scan)) and QSPR/ML categories (*gyuhx* (Simulations Plus), *nb017* (MoKa)). These five methods with RMSE less than 1 pK_a unit are also the methods that have the lowest MAE. *xmyhm* and *xvxzd* were the only two methods for which the upper 95% confidence interval of RMSE was lower than 1 pK_a unit.

In terms of correlation statistics, many methods have good performance, although the ranking of methods changes according to R^2 and Kendall's Tau. Therefore, many methods are indistinguishable from one another, considering the uncertainty of the correlation statistics. 32 out of 38 methods have R and Kendall's Tau higher than 0.7 and 0.6, respectively. 8 methods have R^2 higher than 0.9 and 6 methods have Kendall's Tau higher than 0.8. The overlap of these two sets are the following: *gyuhx* (Simulations Plus), *vvxzd* (DSD-BLYP-D3(BJ)/def2-TZVPD//PBEh-3c[DCOSMO-RS] + RRHO(GFN-xTB[GBSA]) + Gsolv(COSMO-RS[TZVPD]) and linear fit), *xmyhm* (ACD/pKa Classic), *ryzue* (Adiabatic scheme with single point correction: MD/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), and *5byn6* (Adiabatic scheme: thermodynamic cycle that uses gas phase optimized structures for gas phase free energy and solution phase geometry).

tries 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). It is worth noting that the *ryzue* and *5byn6* are QM predictions without any empirical correction. Their high correlation and rank correlation coefficient scores signal that with an empirical correction their accuracy based performance could improve. Indeed, the participants have shown that this is the case in their own challenge analysis paper and achieved RMSE of 0.73 p*K_a* units after the challenge [39].

Null prediction methods based on database lookup (*5nm4j* and *pwn3m*) had similar performance, with an RMSE of roughly 2.5 p*K_a* units, an MAE of 1.5 p*K_a* units, R² of 0.2 and Kendall's Tau of 0.3. Many methods were observed to have a prediction performance advantage over the null predictions shown in light blue in Fig. 2 and Fig. 3 considering all the performance metrics as a whole. In terms of correlation statistics, the null methods are the worst performers, except *0hxtm*. From the perspective of accuracy-based statistics (RMSE and MAE), only the top 10 methods were observed to have significantly lower errors than the null methods considering the uncertainty of error metrics expressed as 95% confidence intervals.

The distribution of macroscopic p*K_a* prediction signed errors observed in each submission was plotted in Fig. 7A as ridge plots based on Hungarian matching. *2ii2g*, *f0gew*, *np64b*, *p0jba*, and *yc70m* tend to overestimate and *5byn6*, *ryzue*, and *w4iyd* tend to underestimate macroscopic p*K_a* values.

There were four submissions of QM+LEC category that used the COSMO-RS implicit solvation model. While three of these achieved the lowest RMSE among QM-based methods (*xvxzd*, *yqkga*, and *8xt50*) [44], one of them showed the highest RMSE (*0hxtm* (COSMOtherm_FINE17)) in SAMPL6 Challenge macroscopic p*K_a* predictions. All four methods used COSMO-RS/FINE17 level to compute solvation free energies. The major difference between the three low-RMSE methods and the *0hxtm* seems to be the protocol for determining relevant conformations for each microstate. *xvxzd*, *yqkga*, and *8xt50* methods used semi-empirical tight binding (GFN-xTB) method and GBSA continuum solvation model for geometry optimization of conformers and followed up with high level single point energy calculations with solvation free energy (COSMO-RS(FINE17/TZVPD)) and rigid rotor harmonic oscillator (RRHO(GFN-xTB(GBSA)) corrections. *yqkga*, and *8xt50* methods selected conformations for each microstate with Relevant Solution Conformer Sampling and Selection (ReSCoSS) workflow. Conformations were clustered according to shape and lowest energy conformations from each cluster according to BP86/TZVP/COSMO single point energies in any of the 10 different COSMO-RS solvents were considered as relevant conformers. The ReSCoSS workflow was described more in detail by Pracht et al [44]. *yqkga* method further filtered out conformers that have less than 5% Boltzmann weights at the DSD-BLYP-D3/def2-TZVPD + RRHO(GFNxTB) + COSMO-RS(fine) level. *xvxzd* method used MF-MD-GC//GFN-xTB workflow and used energy thresholds of 6 kcal/mol and 10 kcal/mol, for conformer and microstate selection. On the other hand, the conformational ensemble captured for each microstate seems to be more limited for *0hxtm* method, judging by the method description provided in the submission file (this participant did not publish an analysis of the results that they obtained for SAMPL6). *0hxtm* method reported that relevant conformations were computed with the COSMOconf 4.2 workflow which produced multiple relevant conformers for only the neutral states of SM18 and SM22. In contrast to *xvxzd*, *yqkga*, and *8xt50* methods, the *0hxtm* method also did not include a RRHO correction. Participants of the three low-RMSE methods report that capturing the chemical ensemble for each molecule including conformers and tautomers and high level QM calculations led to more successful macroscopic p*K_a* prediction results and RRHO correction provided a minor improvement [44]. Comparing these results to other QM approaches in SAMPL Challenge also points to the advantage of the COSMO-RS solvation approach compared to other implicit solvent models.

In addition to the statistics related to the p*K_a* value, we also analyzed missing or extra p*K_a* predictions. Analysis of the p*K_a* values with accuracy- and correlation-based error metrics was only possible after the assignment of predicted macroscopic p*K_a*s to experimental p*K_a*s through Hungarian matching, although this approach masks p*K_a* prediction issues in the form of extra or missing macroscopic p*K_a* predictions. To capture this form of prediction errors we reported the number of unmatched experimental p*K_a*s (missing p*K_a* predictions) and the number of unmatched predicted p*K_a*s (extra p*K_a* predictions) after Hungarian matching for each method. Both missing and extra p*K_a* prediction counts were only considered for the pH range of 2-12 which was the limits of experimental assay. The lower subplot of Fig. 2 shows the total count of unmatched experimental or predicted p*K_a*s for all the molecules in each prediction set. The order of submission IDs in the x-axis follows the RMSD based ranking so that the performance of each method from both p*K_a* value accuracy and the number of p*K_a*s can be viewed together. Presence of missing or extra macroscopic p*K_a* predictions is a critical error because inaccuracy in predicting the correct number of macroscopic transitions shows that methods are failing to predict the correct set of charge states, i.e. failing to predict the correct number of ionization states that can be observed between the specified pH range.

In the analysis of these challenge results, extra macroscopic p*K_a* predictions were found to be more common than missing p*K_a* predictions. In p*K_a* prediction evaluations, the accuracy of predicted ionization states within a pH range is usually neglected. When predictions are only evaluated for the accuracy of the p*K_a* value with numerical matching algorithms, larger number of

499 predicted pK_a s lead to greater underestimation of prediction errors. Therefore, it is not surprising that methods are biased to
500 predict extra pK_a values. The SAMPL6 pK_a Challenge experimental data consists of 31 macroscopic pK_a s in total, measured for 24
501 molecules (6 molecules in the set have multiple pK_a s). Within the 10 methods with lowest RMSE, only the *xvxzd* method predicts
502 too few pK_a values (2 unmatched out of 31 experimental pK_a s). All other methods that rank in the top 10 by RMSE have extra
503 predicted pK_a s ranging from 1 to 13. Two prediction sets without any extra pK_a predictions and low RMSE are *8xt50* (ReSCoSS
504 conformations // DSD-BLYP-D3 reranking // COSMOtherm pKa) and *nb015* (ChemAxon/Chemicalize).

505 3.1.1 Consistently well-performing methods for macroscopic pK_a prediction

506 Methods ranked differently when ordered by different error metrics, although there were a couple of methods that consistently
507 ranked at the top fraction. By using combinatorial criteria that take all multiple statistical metrics and unmatched pK_a counts into
508 account, we identified a shortlist of consistently well-performing methods for macroscopic pK_a predictions, shown in Table 2.
509 The criteria for selection were ranking in Top 10 according to RMSE, MAE, R^2 , and Kendall's Tau and also having a combined
510 unmatched pK_a (extra and missing pK_a s) count less than 8 (a third of the number of compounds). This resulted in a list of four
511 methods that are consistently well-performing across all criteria.

512 Consistently well performing methods for macroscopic pK_a prediction included methods from all categories. Two methods of
513 the QM+LEC category were *xvxzd* (DSD-BLYP-D3(BJ)/def2-TZVPD//PBEh-3c[DCOSMO-RS] + RRHO(GFN-xTB[GBSA]) + Gsolv(COSMO-
514 RS[TZVPD])) and linear fit) and (*8xt50*) (ReSCoSS conformations // DSD-BLYP-D3 reranking // COSMOtherm pKa) and both used
515 COSMO-RS approach. Empirical pK_a predictions with top performance were both proprietary software. From QSPR and LFER
516 categories, *gyuhx* (Simulation Plus) and *xmyhm* (ACD/pKa Classic) were consistently well-performing methods. The Simulation
517 Plus pK_a prediction method consisted of 10 artificial neural network ensembles trained on 16,000 compounds for 10 classes of
518 ionizable atoms. Atom type and the local molecular environment was how the ionization class of each atom was determined
519 [46]. ACD/pKa Classic method was trained on 17,000 compounds, uses Hammet-type equations and captures effects related to
520 tautomeric equilibria, covalent hydration, resonance effects, and α , β -unsaturated systems [36].

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 τ . 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 (τ)	Unmatched Exp. pK_a Count	Unmatched Pred. pK_a Count [2,12]
<i>xvxzd</i>	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
<i>gyuhx</i>	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
<i>xmyhm</i>	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
<i>8xt50</i>	ReSCoSS conformations // DSD-BLYP-D3 reranking // COSMOtherm pKa	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

521 Figure 4 shows predicted vs. experimental value correlation plots of macroscopic pK_a predictions of four consistently well-
522 performing methods, a representative average method, and the null method(*5nm4j*). We selected the method with the highest
523 RMSE below the median of all methods as the representative method with average performance: *2ii2g* (EC-RISM/MP2/cc-pVTZ-
524 P2-q-noThiols-2par).

525 3.1.2 Which chemical properties are driving macroscopic pK_a prediction failures?

526 In addition to comparing the performance of methods that participated in the SAMPL6 Challenge, we also wanted to analyze
527 macroscopic pK_a predictions from the perspective of challenge molecules and determine whether particular compounds suf-
528 ffer from larger inaccuracy in pK_a predictions. The goal of this analysis is to provide insight on which molecular properties or
529 moieties might be causing larger pK_a prediction errors. In Fig. 5, 2D depictions of the challenge molecules are presented with
530 MAE calculated for their macroscopic pK_a predictions over all methods, based on Hungarian match. For multiprotic molecules,
531 MAE was averaged over all the pK_a s. For the analysis of pK_a prediction accuracy observed for each molecule, MAE is a more
532 appropriate statistical value than RMSE for following global trends. This is because MAE value less sensitive to outliers than is
533 RMSE.

534 A comparison of the prediction accuracy of individual molecules is shown in Fig. 6. In Fig. 6A MAE each molecule is shown

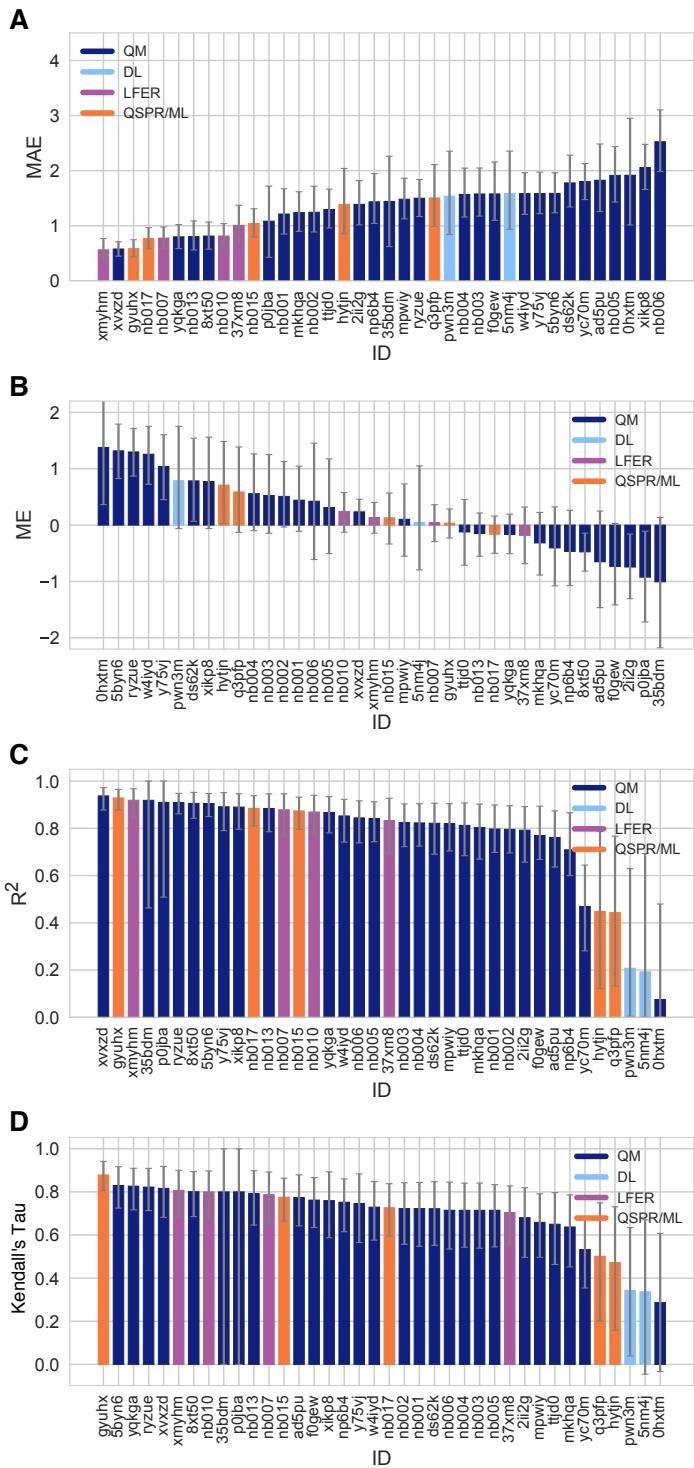


Figure 3. Additional performance statistics for macroscopic pK_a predictions based on Hungarian matching. Methods are indicated by submission IDs. Mean absolute error (MAE), mean error (ME), Pearson's R^2 , and Kendall's Rank Correlation Coefficient τ are shown, with error bars denoting 95% confidence 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 lookup methods are utilized as the null prediction method.

535 considering all blind predictions and reference calculations. A cluster of molecules marked orange and red have higher than
 536 average MAE. Molecules marked red (SM06, SM21, and SM22) are the only compounds in the SAMPL6 dataset with bromo or

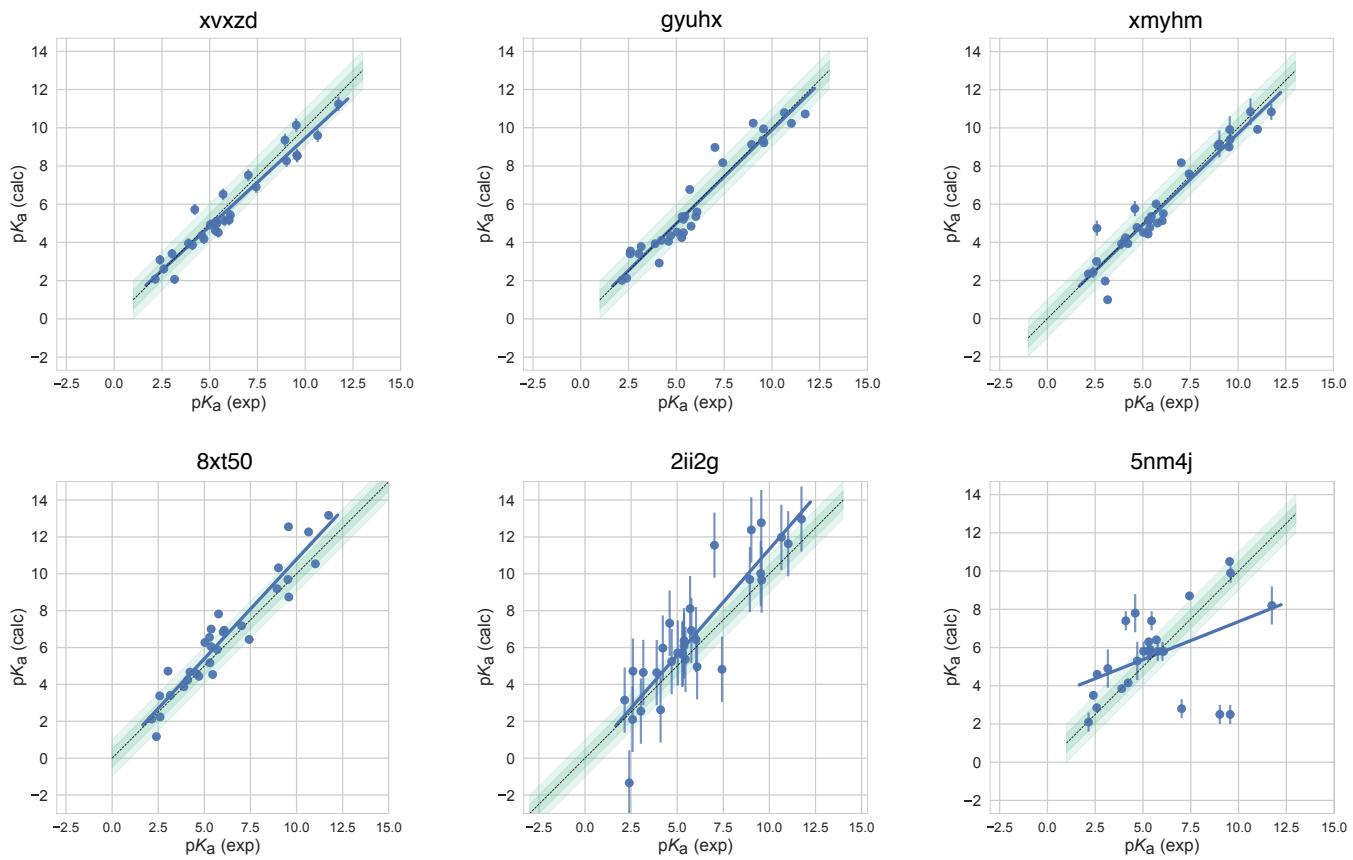


Figure 4. Predicted vs. experimental value correlation plots of four consistently well-performing methods, a representative method with average performance (2ii2g), and the null method (5nm4j). When submissions were ranked according to RMSE, MAE, R^2 , and τ , four methods ranked in the Top 10 consistently in each of these metrics. 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.

537 iodo groups and they suffered a macroscopic pK_a prediction error in the range of 1.7-2.0 pK_a units in terms of MAE. Molecules
 538 marked orange (SM03, SM10, SM18, SM19, and SM20) have sulfur-containing heterocycles, and all these molecules except SM18
 539 have MAE larger than 1.6 pK_a unit. SM18 despite containing thiazole group has a low MAE. SM18 is the only compound with
 540 three experimental pK_a s and we suspect the presence of multiple experimental pK_a s could have a masking effect on the errors
 541 captured by MAE with Hungarian matching due to more pairing choices.

542 We analyzed MAE of each molecule for empirical (LFER and QSPR/ML) and QM-based physical methods (QM, QM+LEC, and
 543 QM+MM) separately for more insight. Fig. 6B shows that the difficulty of predicting pK_a s of the same subset of molecules was
 544 a trend conserved in the performance of physical methods. For QM-based methods too sulfur-containing heterocycles, amide
 545 next to aromatic heterocycles, compounds with iodo and bromo substitutions have lower pK_a prediction accuracy.

546 SAMPL6 pK_a set consists of only 24 small molecules and lacks repeating instances of many moieties which limits our ability
 547 to determine with statistical significance which chemical substructures cause greater errors in pK_a predictions. Still, the trends
 548 observed in this challenge point to molecules with iodo, bromo, and sulfur-containing heterocycles with larger prediction
 549 errors of macroscopic pK_a value. We hope that reporting this observation will lead to the improvement of methods for similar
 550 compounds with such moieties.

551 We have also looked for correlation with molecular descriptors for finding other potential explanations for why macroscopic
 552 pK_a prediction errors were larger for some molecules. While testing the correlation between errors and many molecular de-
 553 scriptors, it is important to keep the possibility of spurious correlations in mind. We haven't observed any significant correlation

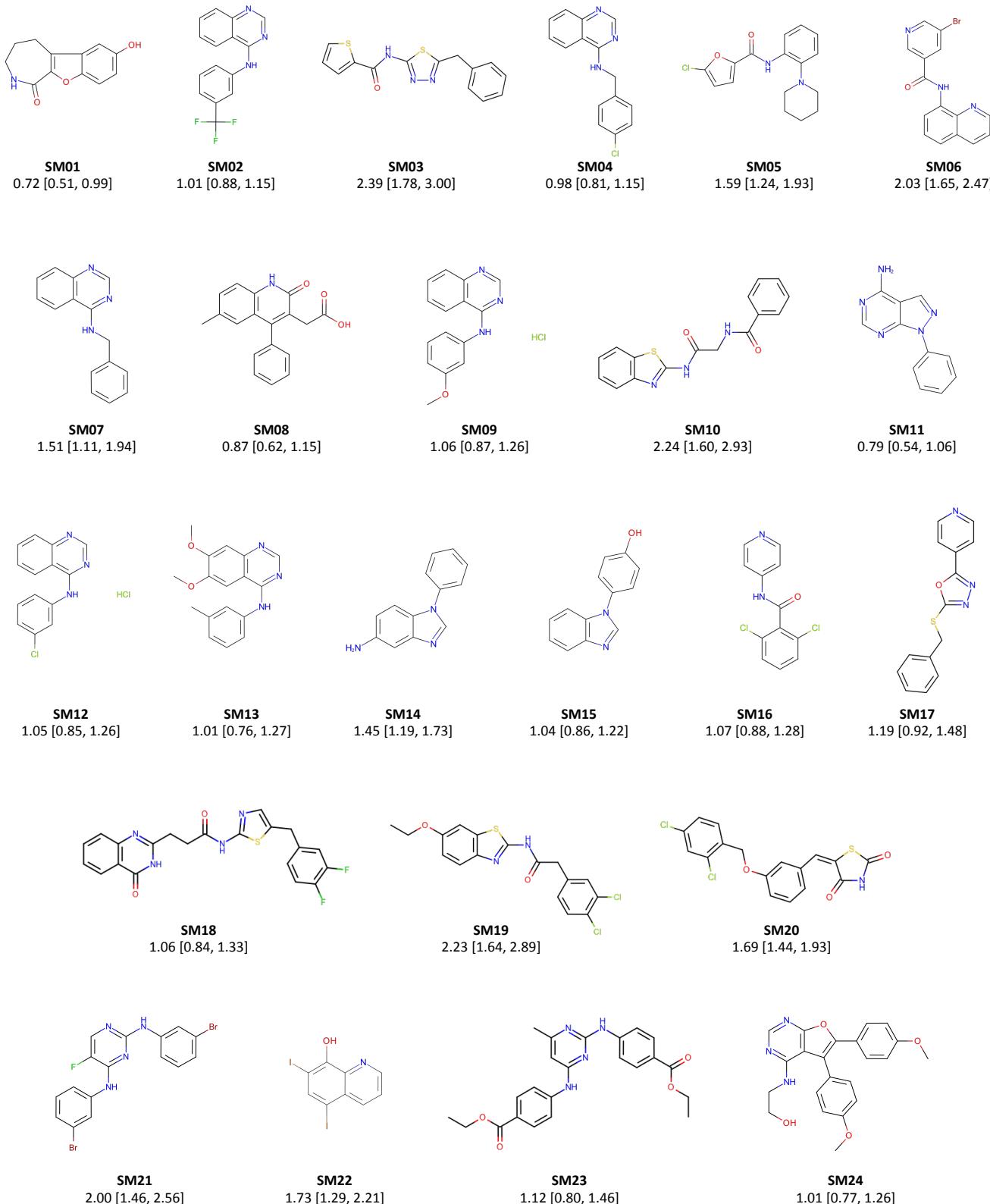
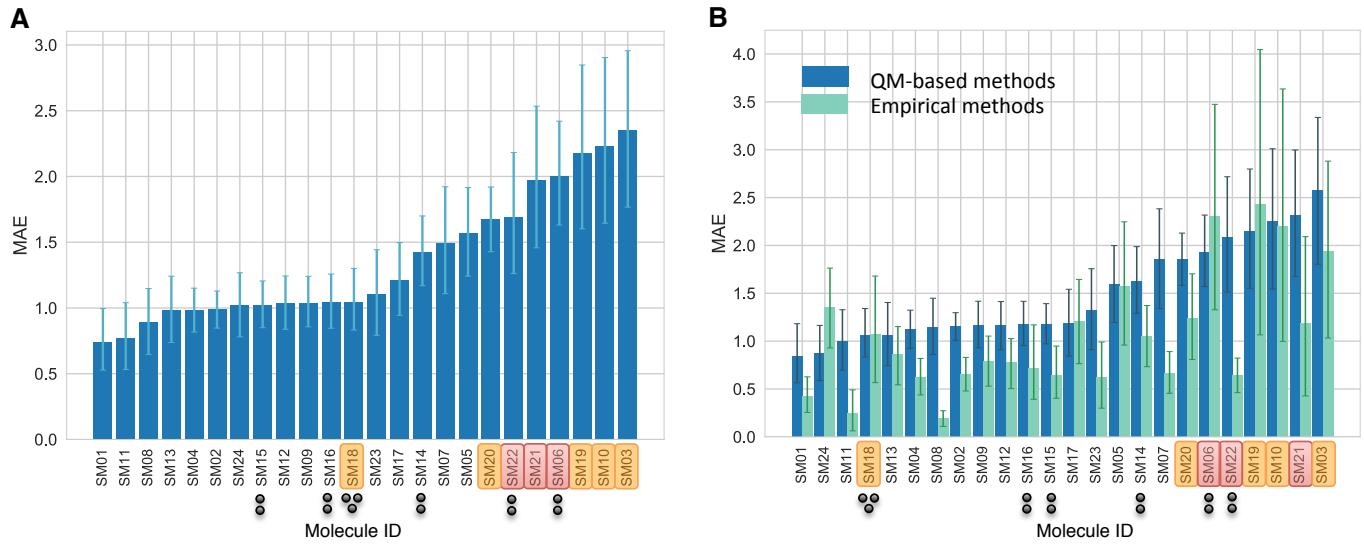
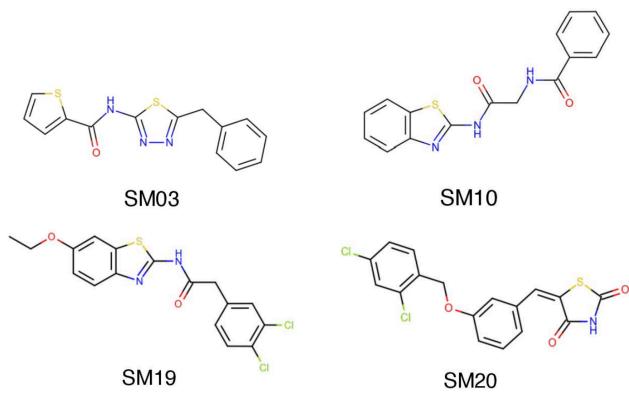


Figure 5. Molecules of SAMPL6 Challenge with MAE calculated for all macroscopic pK_a predictions. MAE calculated considering all prediction methods indicate which molecules had the lowest prediction accuracy in SAMPL6 Challenge. MAE values calculated for each molecule include all the matched pK_a values. SM06, SM14, SM15, SM16, SM18, and SM22 were multiprotic. Hungarian matching algorithm was employed for pairing experimental and predicted pK_a values. MAE values are reported with 95% confidence intervals.



C SAMPL6 molecules with sulfur-containing heterocycles



● 3 experimental pK_a values
● 2 experimental pK_a values
■ Sulfur-containing heterocycles
■ Bromo and iodo groups

D SAMPL6 molecules with bromo and iodo groups

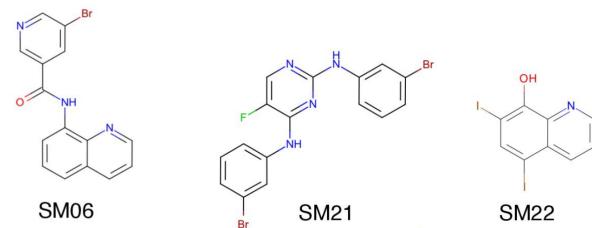


Figure 6. Average prediction accuracy calculated over all prediction methods was lower for molecules with sulfur-containing heterocycles, bromo, and iodo groups. (A) MAE calculated for each molecule as an average of all methods. (B) MAE of each molecule broken out by method category. QM-based methods (blue) include QM predictions with or without linear empirical correction. Empirical methods (green) include QSAR, ML, DL, and LFER approaches. (C) Depiction of SAMPL6 molecules with sulfur-containing heterocycles. (D) Depiction of SAMPL6 molecules with iodo and bromo groups.

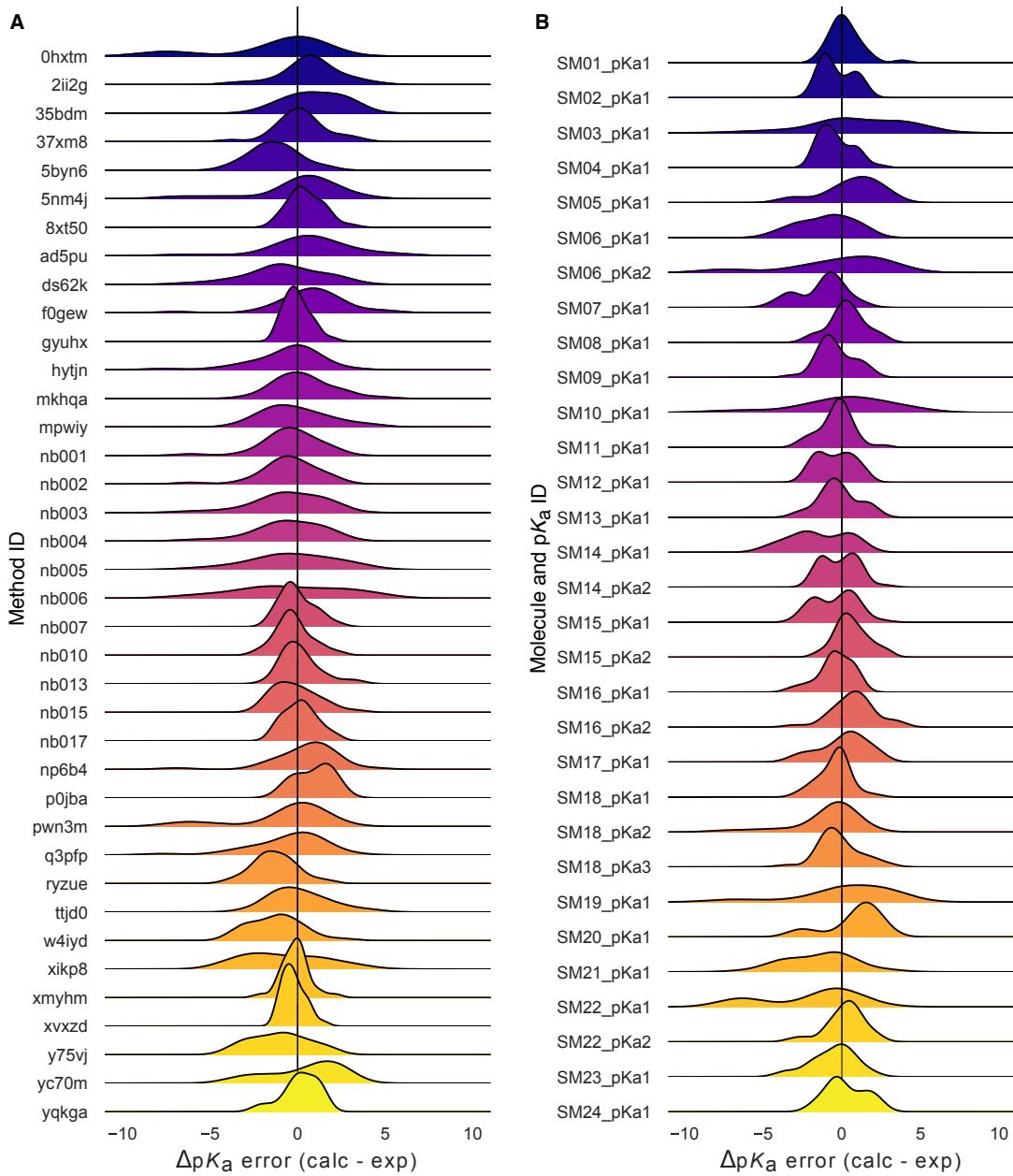


Figure 7. Macroscopic pK_a prediction error distribution plots show how prediction accuracy varies across methods and individual molecules. (A) pK_a prediction error distribution for each submission for all molecules according to Hungarian matching. (B) Error distribution for each SAMPL6 molecule for all prediction methods according to Hungarian matching. For multiprotic molecules, pK_a ID numbers (pKa1, pKa2, and pKa3) were assigned in the direction of increasing experimental pK_a value.

554 between numerical pK_a predictions and the descriptors we have tested. First, having more experimental pK_a s (Fig. 6A) did not
555 seem to associate with worse pK_a prediction performance. But we need to keep in mind that there was a low representation
556 of multiprotic compounds in the SAMPL6 set (5 molecules with 2 macroscopic pK_a s and one with 3 macroscopic pK_a). Second,
557 we checked the following other descriptors: amide group presence, molecular weight, heavy atom count, rotatable bond count,
558 heteroatom count, heteroatom to carbon ratio, ring system count, maximum ring size, and the number of microstates (as enu-
559 merated for the challenge). Correlation plots and R^2 values can be seen in Fig. S2. We had suspected that pK_a prediction methods
560 may be trained better for moderate values (4-10) than extreme values as molecules with extreme pK_a s are less likely to change
561 ionization states close to physiological pH. To test this we look at the distribution of absolute errors calculated for all molecules
562 and challenge predictions binned by experimental pK_a value 2 pK_a unit increments. As can be seen in Fig. S3B, the value of true
563 macroscopic pK_a s was not a factor affecting prediction error seen in SAMPL6 Challenge.

564 Fig. 7B is helpful to answer the question of "Are there molecules with consistently overestimated or underestimated pK_a s?".
565 This ridge plots show the error distribution of each experimental pK_a . SM02_pKa1, SM04_pKa1, SM14_pKa1, and SM21_pKa1
566 were underestimated, predicting lower protein affinity by more than 1 pK_a unit by majority of the prediction methods. SM03_pKa1,
567 SM06_pKa2, SM19_pKa1, and SM20_pKa1 were overestimated by the majority of the prediction methods by more than 1 pK_a unit.
568 SM03_pKa1, SM06_pKa2, SM10_pKa1, SM19_pKa1, and SM22_pKa1 have the highest spread of errors and were less accurately
569 predicted overall.

570 3.2 Analysis of microscopic pK_a predictions using microstates determined by NMR for 8 molecules

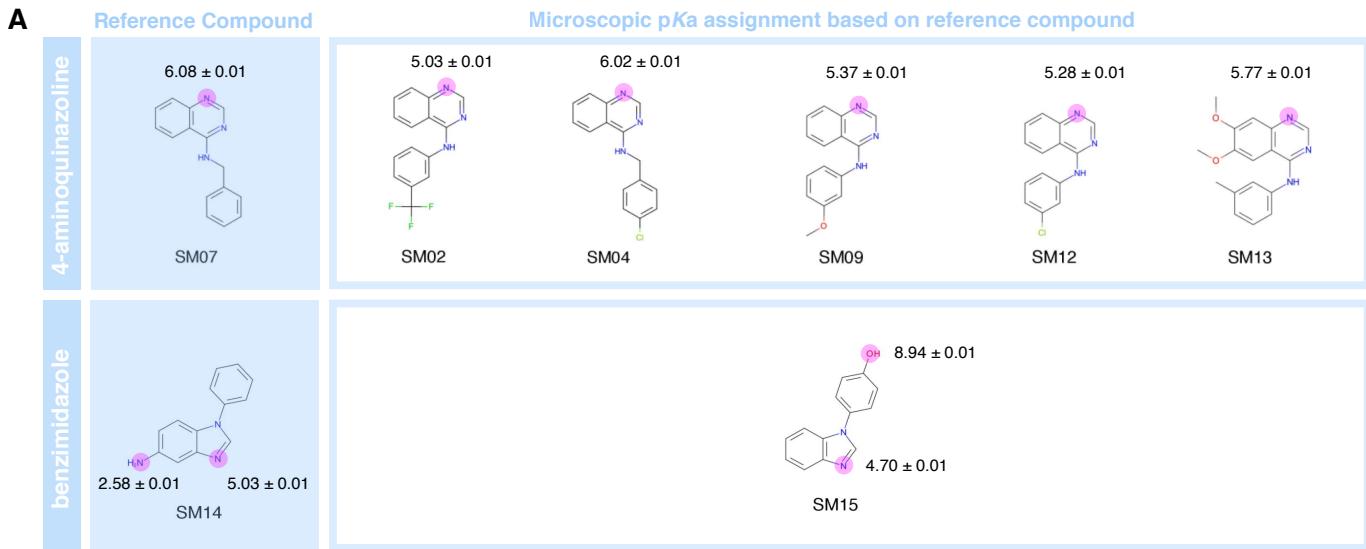
571 The common approach for analyzing microscopic pK_a prediction accuracy has been to compare it to experimental macroscopic
572 pK_a data, assuming experimental pK_a s describe titrations of distinguishable sides and, therefore, equal to microscopic pK_a s. But
573 this typical approach fails to evaluate the methods at the microscopic level.

574 Analysis of microscopic pK_a predictions of the SAMPL6 Challenge was not straightforward due to the lack of experimental
575 data with microscopic detail. For 24 molecules macroscopic pK_a s were determined with the spectrophotometric method. For 18
576 molecules a single macroscopic titration was observed and for 6 molecules multiple experimental pK_a s were reported. For 18
577 molecules with single experimental pK_a , it is probable that the molecules are monoprotic and therefore macroscopic pK_a value
578 is equal to the microscopic pK_a . There is, however, no direct experimental evidence supporting this hypothesis but only the
579 support from computational predictions. There is always the possibility that the macroscopic pK_a observed is the result of two
580 different titrations overlapping closely with respect to pH if any charge state has more than one tautomer. We did not want to
581 bias the blind challenge analysis with any prediction method. Therefore, we believe analyzing the microscopic pK_a predictions
582 via Hungarian matching to experimental values with the assumption that the 18 molecules have a single titratable site is not the
583 best approach. Instead, analysis at the level of macroscopic pK_a s is much more appropriate when a numerical matching scheme
584 is the only option to evaluate predictions using macroscopic experimental data.

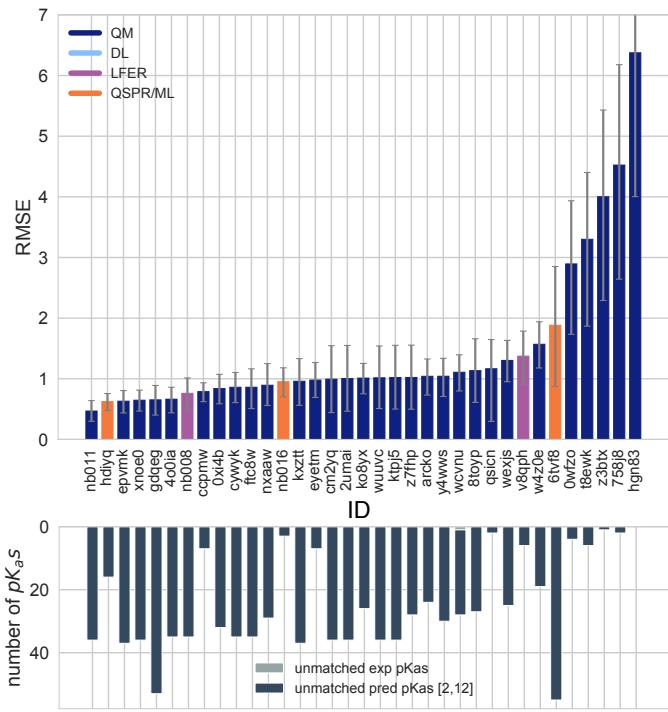
585 For a subset of the molecules in the dataset of 8 molecules, dominant microstates were inferred from NMR experiments.
586 This dataset was extremely useful for guiding the assignment between experimental and predicted pK_a values based on mi-
587 crostates. In this section, we present the performance evaluations of microscopic pK_a predictions for only the 8 compounds
588 with experimentally determined dominant microstates.

589 3.2.1 Microstate-based matching revealed errors masked by pK_a value-based matching between experimental 590 and predicted pK_a s

591 Comparing microscopic pK_a predictions directly to macroscopic experimental pK_a values with numerical matching can lead to
592 underestimation of errors. To demonstrate how numerical matching often masks pK_a prediction errors we compared the perfor-
593 mance analysis done by Hungarian matching to that from microstate-based matching for 8 molecules presented in Fig. 8A. RMSE
594 calculated for microscopic pK_a predictions matched to experimental values via Hungarian matching is shown in Fig. 8B, while
595 Fig. 8C shows RMSE calculated via microstate-based matching. The Hungarian matching incorrectly leads to significantly lower
596 RMSE compared to microstate-based matching. The reason is that the Hungarian matching assigns experimental pK_a values to
597 predicted pK_a values only based on the closeness of the numerical values, without consideration of the relative population of
598 microstates and microstate identities. Because of that, a microscopic pK_a value that describes a transition between very low
599 population microstates (high energy tautomers) can be assigned to the experimental pK_a if it has the closest pK_a value. This is
600 not helpful because, in reality, the microscopic pK_a s that influence the observable macroscopic pK_a the most are the ones with
601 higher populations (transitions between low energy tautomers).



B Hungarian matching



C Microstate-based matching

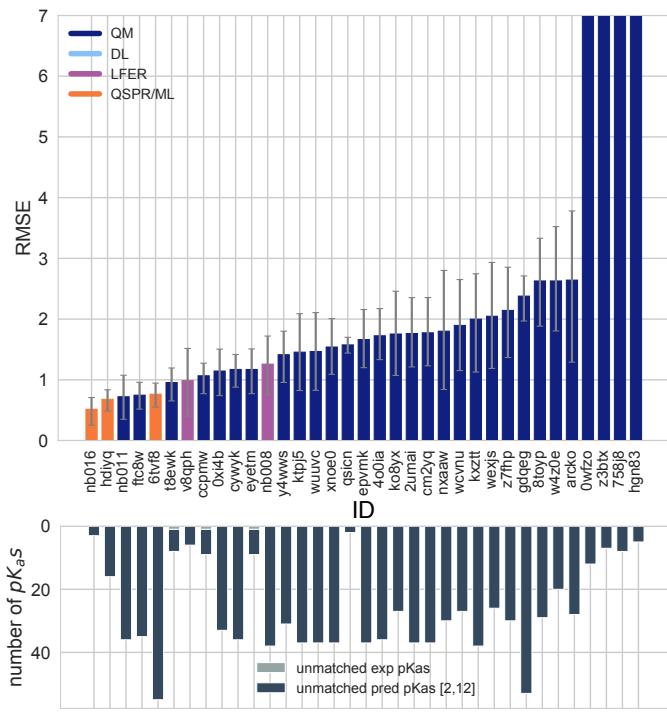


Figure 8. NMR determination of dominant microstates allowed in-depth evaluation of microscopic pK_a predictions of 8 compounds.

A Dominant microstate sequence of two compounds (SM07 and SM14) were determined by NMR [8]. Based on these reference compounds dominant microstates of 6 other derivative compounds were inferred and experimental pK_a values were assigned to titratable groups with the assumption that only the dominant microstates have significant contributions to the experimentally observed pK_a . **B** RMSE vs. submission ID and unmatched pK_a vs. submission ID plots for the evaluation of microscopic pK_a predictions of 8 molecules by Hungarian matching to experimental macroscopic pK_a s. **C** RMSE vs. submission ID and unmatched pK_a vs. submission ID plots showing the evaluation of microscopic pK_a predictions of 8 molecules by microstate-based matching between predicted microscopic pK_a s and experimental macroscopic pK_a values. Submissions *0wfzo*, *z3btx*, *758j8*, and *hgn83* have RMSE values bigger than 10 pK_a units which are beyond the y-axis limits of subplot **C** and **B**. 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.

602 The number of unmatched predicted microscopic pK_a s is shown in the lower bar plots of Fig. 8B and C, to emphasize the large
603 number of microscopic pK_a predictions submitted by many methods. In the case of microscopic pK_a , the number of unmatched
604 predictions does not indicate an error in the form of an extra predicted pK_a , because the spectrophotometric experiments do
605 not capture all microscopic pK_a s theoretically possible (transitions between all pairs of microstates that are 1 proton apart).
606 pK_a s of transitions to and from very high energy tautomers are very hard to measure by experimental methods, including the
607 most sensitive methods like NMR. Prediction of extra microscopic pK_a s can cause underestimation of prediction errors when
608 numerical matching algorithms such as Hungarian matching are used. We also checked how often Hungarian matching led to
609 the correct matches between predicted and experimental pK_a in terms of the microstate pairs, i.e. how often the microstate
610 pair of the Hungarian match recapitulates the dominant microstate pair of the experiment. The overall accuracy of microstate
611 pair matching was found to be low for SAMPL6 Challenge submission. Fig. S4 shows that for most methods the predicted
612 microstate pair selected by the Hungarian match did not correspond to the experimentally determined microstate pair. This
613 means the lower RMSE results obtained from Hungarian matching are low for the wrong reason. This problem could be avoided
614 by matching experimental and predicted values on the basis of microstate IDs.

615 Unfortunately, we are only able to perform this more reliable microstate-based matching analysis for a subset of com-
616 pounds. The conclusions in this section are only about eight compounds with limited diversity. This subset is composed of six
617 4-aminoquinazoline molecules and two with benzimidazole scaffolds, for a total of 10 pK_a values. The sequences of dominant
618 microstates for SM07 and SM14 were determined by NMR experiments directly [8] and dominant microstates of their derivatives
619 were inferred by taking them as reference (Fig. 8). Although we believe that microstate-based evaluation is more informative, the
620 lack of a large experimental dataset limits the conclusions to a very narrow chemical diversity. Still, microstate-based matching
621 revealed errors masked by pK_a value-based matching between experimental and predicted pK_a s.

622 3.2.2 Accuracy of pK_a predictions evaluated by microstate-based matching

623 Both accuracy and correlation based statistics were calculated for predicted microscopic pK_a values after microstate-based
624 matching. RMSE, MAE, ME, R^2 , and Kendall's Tau results of each method are shown in Fig. 8C and Fig. 9. A table of the calculated
625 statistics can be found in Table S4. Due to small number of data points in this set, correlation-based statistics have large uncer-
626 tainties and thus have less utility for distinguishing better performing methods. Therefore, we focused more on accuracy based
627 metrics for the analysis of microscopic pK_a s than correlation based metrics. In terms of accuracy of microscopic pK_a value, all
628 three QSPR/ML based methods (*nb016* (MoKa), *hdlyq* (Simulations Plus), *6tvf8* (OE Gaussian Process)), three QM-based methods
629 (*nb011* (Jaguar), *ftc8w* (EC-RISM/MP2/cc-pVTZ-P2-q-noThiols-2par), *t8ewk* (COSMOlogic_FINE17)), and one LFER method (*v8qph*
630 (ACD/pKa GALAS)) achieved RMSE lower than 1 pK_a unit. The same 6 methods also have the lowest MAE.

631 3.2.3 Evaluation of dominant microstate prediction accuracy

632 For many computational chemistry approaches including structure-based modeling of protein-ligand interactions, predicting
633 the ionization state and the exact position of protons is needed to establish what to include in the modeled system. This is
634 why in addition to being able to predict pK_a values accurately, we need pK_a prediction methods to be able to capture micro-
635scopic protonation states accurately. Even when the predicted pK_a value is very accurate, the predicted protonation site can be
636 wrong. Therefore, we assessed if methods participating in the SAMPL6 pK_a Challenge were predicting correctly the sequence of
637 dominant microstates, i.e. dominant tautomers of each charge state observed between pH 2 and 12.

638 Fig. 10 shows how well methods perform for predicting the dominant microstate, as analyzed for eight compounds with
639 required experimental data. The dominant microstate sequence is essentially the sequence of states that are most visible
640 experimentally, due to their higher fractional population and relative free energy within the tautomers at each charge. To extract
641 the dominant tautomers predicted for the sequence of ionization states of each method, the relative free energy of microstates
642 were first calculated at reference pH 0 [30]. Then to determine the dominant microstate at each formal charge, we have selected
643 the lowest energy tautomer for each ionization state based on the relative microstate free energies calculated at pH 0. The
644 choice of reference pH is not important, as relative free energy difference between tautomers of the same charge is always
645 constant with respect to pH. This analysis was done only for the charges -1, 0, 1, and 2 which was the charge range captured
646 by NMR experiments. Then predicted and experimental dominant microstates were compared for each charge to calculate the
647 fraction of correctly predicted dominant tautomers. This value is reported as the dominant microstate accuracy for all charges
648 (Fig. 10A). Many of the methods which participated the challenge made errors in predicting the dominant microstate. 10 QM
649 and 3 QSPR/ML methods did not make any mistakes in dominant microstate predictions, although, they are expected to make
650 mistakes in the relative population of tautomers (free energy difference between microstates) as reflected by the pK_a value errors.

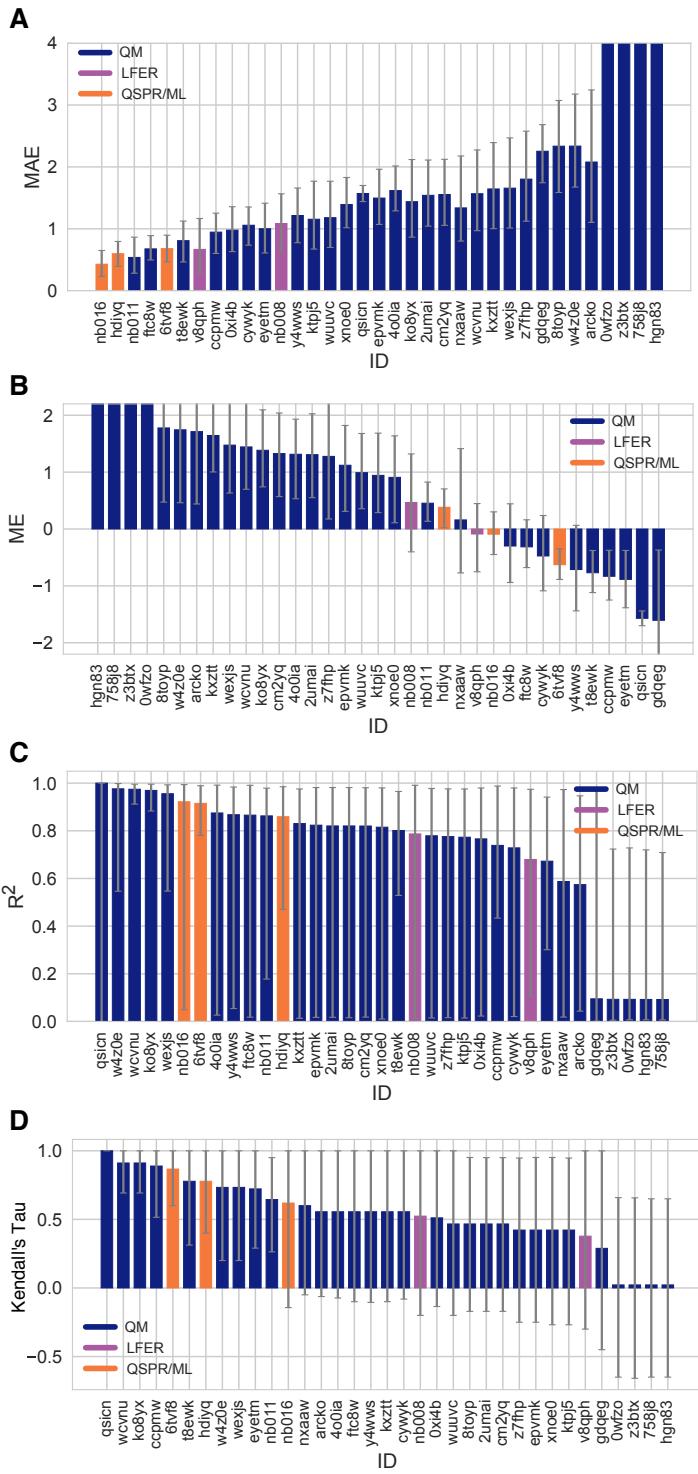


Figure 9. Additional performance statistics for microscopic pK_a predictions for 8 molecules with experimentally determined dominant microstates. Microstate-based matching was performed between experimental pK_a values and predicted microscopic pK_a s. Mean absolute error (MAE), mean error (ME), Pearson's R², and Kendall's Rank Correlation Coefficient Tau (τ) are shown, with error bars denoting 95% confidence intervals obtained by bootstrapping over challenge molecules. Methods are indicated by their submission IDs. Submissions are colored by their method categories. Refer to Table 1 for submission IDs and method names. Submissions 0wfzo, z3btz, 758j8, and hgn83 have MAE and ME values bigger than 10 pK_a units which are beyond the y-axis limits of subplots **A** and **B**. A large number and wide variety of methods have a statistically indistinguishable performance based on correlation statistics (**C** and **D**), in part because of the relatively small dynamic range the small size of the set of 8 molecules.

651 While all the participating QSPR/ML methods showed good performance in dominant microstate prediction, LFER and some QM
652 methods made mistakes. The accuracy of the predicted dominant neutral tautomers was perfect for all methods, except *qsicn*
653 (Fig. 10B). But errors in predicting the major tautomer of charge +1 were much more frequent. 22 out of 35 prediction sets made
654 at least one error in prediction the lowest energy tautomer with +1 charge. We didn't include ionization states with charges -
655 1 and +2 in this assessment because we had only one compound with these charges in the dataset. Nevertheless, errors in
656 predicting the dominant tautomers seem to be a bigger problem for charged tautomers than the neutral tautomer.

657 Only eight compounds had data on the sequence of dominant microstates. Therefore conclusions on the performance of
658 methods in terms of dominant tautomer prediction are limited to this limited chemical diversity (benzimidazole and 4-aminoquinazoline
659 derivatives). We present this analysis as a prototype of how microscopic pK_a predictions should be evaluated. Hopefully, future
660 evaluation can be done with larger experimental datasets following the strategy we demonstrated here in order to reach broad
661 conclusions about which methods are better for capturing dominant microstates and ratios of tautomers. Even if experimental
662 microscopic pK_a measurement data is not available, experimental dominant tautomer determinations are still informative for
663 assessing computational predictions.

664 The most frequent misprediction was the major tautomer of the SM14 cationic form, as shown in Fig. 10. This figure shows
665 the accuracy of the predicted dominant microstate calculated for individual molecules and for charge states 0 and +1, averaged
666 over all prediction methods. SM14, the molecule that exhibits the most frequent error in the predicted dominant microstate,
667 has two experimental pK_a values that were 2.4 pK_a units apart, and we suspect that could be a contributor to the difficulty of
668 predicting microstates accurately. Other molecules are monoprotic (4-aminoquinazolines) or their experimental pK_a values are
669 very well separated (SM14, 4.2 pK_a units). It would be very interesting to expand this assessment to a larger variety of drug-
670 like molecules to discover for which structures tautomer predictions are more accurate and for which structure computational
671 predictions are not as reliable.

672 3.2.4 Consistently well-performing methods for microscopic pK_a predictions

673 We have determined different criteria for determine consistently top-performing predictions of microscopic pK_a than macro-
674 scopic pK_a : having perfect dominant microstate prediction accuracy, unmatched pK_a count of 0, and ranking in the top 10
675 according to RMSE and MAE. Correlation statistics were not found to have utility for discriminating performance due to large un-
676 certainties in these statistics for a small dataset of 10 pK_a values. Unmatched predicted pK_a count was also not a consideration,
677 since experimental data was only informative for the pK_a between dominant microstates and did not capture all the possible
678 theoretical transitions between microstate pairs. Table 3 reports six methods that have consistent good performance according
679 to many metrics, although evaluated only for the 8 molecule set due to limitations of the experimental dataset. Six methods
680 were divided evenly between methods of QSPR/ML category and QM category. *nb016* (MoKa), *hdlyq* (Simulations Plus), and *6tvf8*
681 (OE Gaussian Process) were QSPR and ML methods that performed well. *nb011* (Jaguar), *0xi4b*(EC-RISM/B3LYP/6-311+G(d,p)-P2-
682 phi-noThiols-2par), and *cwyk* (EC-RISM/B3LYP/6-311+G(d,p)-P2-phi-noThiols-2par) were QM predictions with linear empirical
683 corrections with good performance with microscopic pK_a predictions.

684 The Simulations Plus pK_a prediction method is the only method that appeared to be consistently well performing in both the
685 assessment for macroscopic and microscopic pK_a prediction (*gyuhx* and *hdlyq*). However it is worth noting that two methods that
686 were consistently in top-performing methods list for macroscopic pK_a predictions lacked equivalent submissions of their underlying
687 microscopic pK_a predictions and therefore could not be evaluated at the microstate level. These methods were (ACD/Classic
688 pK_a) and *xvxzd*(DSD-BLYP-D3(BJ)/def2-TZVPD//PBEh-3c[DCOSMO-RS] + RRHO(GFN-xTB[GBSA]) + Gsolv(COSMO-RS[TZVPD]) and
689 linear fit).

690 3.3 How do pK_a prediction errors impact protein-ligand binding affinity predictions?

691 pK_a predictions provide a key input for computational modeling of protein-ligand binding with physical methods. The SAMPL6
692 pK_a Challenge focused focused only on small molecule pK_a prediction and pK_a prediction accuracy observed can effect modeling
693 of ligands. Many affinity prediction methods such as docking, MM/PBSA, MM/GBSA, absolute or alchemical relative free energy
694 calculation methods predict the affinity of the ligand to a receptor in a fixed protonation state. These models strictly depend on
695 pK_a predictions for determining possible protonation states of the ligand in the aqueous environment and in a protein complex,
696 as well as the free energy penalty to reach those states [4]. The accuracy of pK_a predictions can become a limitation for the
697 performance of physical models that try to capture molecular association.

698 In terms of the ligand protonation states, there are two ways in which pK_a prediction errors can influence the prediction
699 accuracy for protein-ligand binding free energies as depicted in Fig. 11. The first scenario is when a ligand is present in aqueous

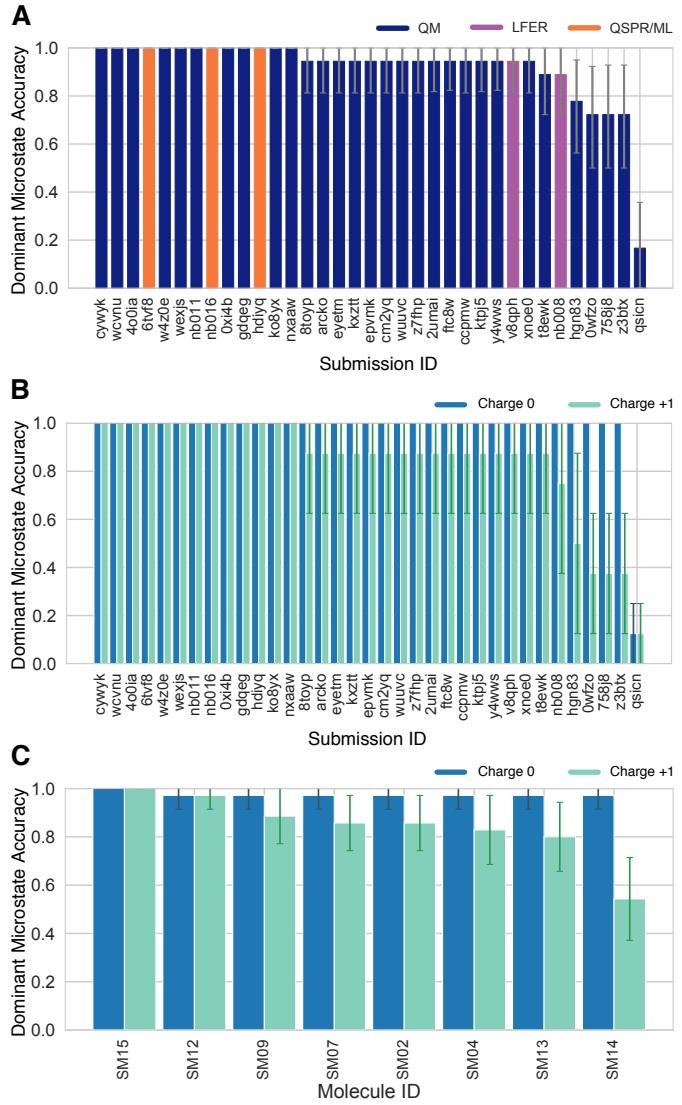


Figure 10. Some methods predicted the sequence of dominant tautomers inaccurately. Prediction accuracy of the dominant microstate of each charged state was calculated using the dominant microstate sequence determined by NMR for 8 molecules as reference. **(A)** Dominant microstate accuracy vs. submission ID plot was calculated considering all the dominant microstates seen in the experimental microstate dataset of 8 molecules. **(B)** Dominant microstate accuracy vs. submission ID plot was generated considering only the dominant microstates of charge 0 and +1 seen in the 8 molecule dataset. The accuracy of each molecule is broken out by the total charge of the microstate. **(C)** Dominant microstate prediction accuracy calculated for each molecule averaged over all methods. In **(B)** and **(C)**, the accuracy of predicting the dominant neutral tautomer is shown in blue and the accuracy of predicting the dominant +1 charged tautomer is shown in green. Error bars denoting 95% confidence intervals obtained by bootstrapping.

700 solution in multiple protonation states (Fig. 11A). When only the minor aqueous protonation state contributes to protein-ligand
 701 complex formation, the overall binding free energy (ΔG_{bind}) needs to be calculated as the sum of binding free energy of the minor
 702 state and the protonation penalty of that state (ΔG_{prot}). ΔG_{prot} is a function of pH and pK_a . A 1 unit of error in pK_a value would
 703 lead to 1.36 kcal/mol error in overall binding free energy if the protonation state with the minor population binds the protein.
 704 The equations in Fig. 11A show the calculation of overall free energy.

705 In addition to multiple protonation states being present in the aqueous environment, multiple charge states can contribute to
 706 complex formation (Fig. 11B). Then, the overall free energy of binding needs to include a Multiple Protonation States Correction
 707 (MPSC) term (ΔG_{corr}). MPSC is a function of pH, aqueous pK_a of the ligand, and the difference between the binding free energy
 708 of charged and neutral species ($\Delta G_{bind}^C - \Delta G_{bind}^N$) as shown in Fig. 11B.

709 Using the equations in Fig. 11B we can model the true MPSC (ΔG_{corr}) value with respect to the difference between pH and

Table 3. Top-performing methods for microscopic pK_a predictions based on consistent ranking within the Top 10 according to various statistical metrics calculated for 8 molecule dataset. Performance statistics are provided as mean and 95% confidence intervals. Submissions that rank in the Top 10 according to RMSE and MAE, and have perfect dominant microstate prediction accuracy were selected as consistently well-performing methods. Correlation-based statistics (R^2 , and Kendall's Tau), although reported in the table, were excluded from the statistics used for determining top-performing methods. This was because correlation-based statistics were not very discriminating due to narrow dynamic range and the small number of data points in the 8 molecule dataset with NMR-determined dominant microstates.

Submission ID	Method Name	Dominant Microstate Accuracy	RMSE	MAE	R ²	Kendall's Tau	Unmatched Exp. pK _a Count	Unmatched Pred. pK _a Count [2,12]
nb016	MoKa	1.0 [1.0, 1.0]	0.52 [0.25, 0.71]	0.43 [0.23, 0.65]	0.92 [0.05, 0.99]	0.62 [-0.14, 1.00]	0	3
hd1yq	S+pKa	1.0 [1.0, 1.0]	0.68 [0.49, 0.83]	0.60 [0.39, 0.80]	0.86 [0.47, 0.98]	0.78 [0.40, 1.00]	0	16
nb011	Jaguar	1.0 [1.0, 1.0]	0.72 [0.35, 1.07]	0.54 [0.28, 0.86]	0.86 [0.18, 0.98]	0.64 [0.26, 0.95]	0	36
6tvf8	OE Gaussian Process	1.0 [1.0, 1.0]	0.76 [0.55, 0.95]	0.68 [0.46, 0.90]	0.92 [0.78, 0.99]	0.87 [0.6, 1.00]	0	55
0xi4b	EC-RISM/B3LYP/6-311+G(d,p)-P3NI-phi-noThiols-2par	1.0 [1.0, 1.0]	1.15 [0.75, 1.50]	0.98 [0.63, 1.36]	0.77 [0.02, 0.98]	0.51 [-0.14, 1.00]	0	33
cywyk	EC-RISM/B3LYP/6-311+G(d,p)-P2-phi-noThiols-2par	1.0 [1.0, 1.0]	1.17 [0.88, 1.41]	1.06 [0.74, 1.35]	0.73 [0.02, 0.98]	0.56 [-0.08, 1.00]	0	36

the pK_a of the ligand, to see when this value has significant impact to the overall binding free energy. In Fig. 12, the true MPSC value that needs to be added to ΔG_{bind}^N is shown for ligands with varying binding affinity difference between protonation states ($\Delta\Delta G = \Delta G_{bind}^C - \Delta G_{bind}^N$) and varying free energy of binding difference between the protonation states. Fig. 12A shows the case of a monoprotic base in which the charged state has a lower affinity than the neutral state. Solid lines show the true correction. In cases where the pK_a is lower than the pH, the correction factor disappears as the ligand fully populates the neutral state ($\Delta G_{bind} = \Delta G_{bind}^N$). As the pK_a value gets larger than the pH, the charged state is populated more and ΔG_{corr} increases to approach $\Delta\Delta G$. It is interesting to note that the pH-pK_a range over which ΔG_{corr} changes. It is often assumed that for a basic ligand if the pK_a of a ligand is more than 2 units higher than the pH, then only 1% of the population is in the neutral state and it is safe to approximate the overall binding affinity with ΔG_{bind}^C . Based on the magnitude of the relative free energy difference between ligand protonation states, this assumption is not always correct. As seen in Fig. 12A, responsive region of ΔG_{corr} can span 3 pH units for a system with $\Delta\Delta G = 1\text{kcal/mol}$ or 5 pH units for a system with $\Delta\Delta G = 4\text{kcal/mol}$. This highlights that the range of pK_a values that impact binding affinity predictions is wider than previously appreciated. Molecules with pK_as several units away from the physiological pH can still impact the overall binding affinity significantly due to the MPSC.

Despite the need to capture the contributions of multiple protonations states by including MPSC in binding affinity calculations, inaccurate pK_a predictions can lead to errors in ΔG_{corr} and overall free energy of binding prediction. In Fig. 12A dashed lines show predicted ΔG_{corr} based on pK_a error of -1 units. We have chosen a pK_a error of 1 unit as this is the average performance expected from the pK_a prediction methods based on the SAMPL6 Challenge. Underestimated pK_a causes underestimated ΔG_{corr} and overestimated affinities (i.e. too negative binding free energy) for a varying range of pH - pK_a values depending on binding affinity difference between protonation states($\Delta\Delta G$). In Fig. 12B dashed lines show how the magnitude of the absolute error caused by calculating ΔG_{corr} with an inaccurate pK_a varies with respect to pH. Different colored lines show simulated results with varying binding free energy differences between protonation states. For a system whose charged state has higher binding free energy than the neutral state ($\Delta\Delta G = 2\text{kcal/mol}$), the absolute error caused by underestimated pK_a by 1 unit can be up to 0.9 kcal/mol. For a system whose charged state has even lower affinity (more positive binding free energy) than the neutral state ($\Delta\Delta G = 4\text{kcal/mol}$), the absolute error caused by underestimated pK_a by 1 unit can be up to 1.2 kcal/mol. The magnitude of errors contributing to overall binding affinity is too large to be neglected. Improving the accuracy of small molecule pK_a prediction methods can help to minimize the error in predicted MPSC.

With the current level of pK_a prediction accuracy as observed in SAMPL6 Challenge, is it advantageous to include MPSC in affinity predictions that may include errors caused by pK_a predictions? We provide a comparison of the two choices to answer this question: (1) Neglecting MPSC completely and assuming overall binding affinity is captured by ΔG_{bind}^N , (2) including MPSC with potential error in overall affinity calculation. The magnitude of error caused by Choice 1 (ignoring MPSC) is depicted as a solid line in Fig. 12B and the magnitude of error caused by MPSC computed with inaccurate pK_a is depicted as dashed lines. What is the best strategy? Error due to choice 1 is always larger than error due to choice 2 for all pH-pK_a values. In this scenario including MPSC improves overall binding affinity prediction. The error caused by the inaccurate pK_a is smaller than the error caused by neglecting MPSC.

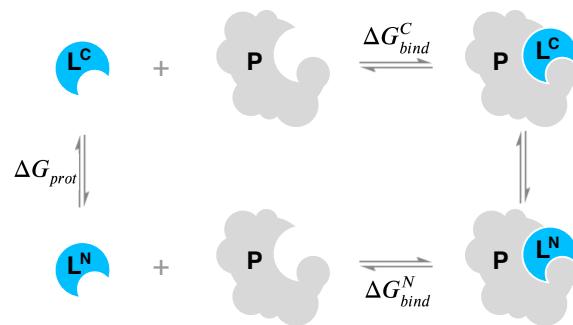
A When only the minor protonation state can bind to the protein



$$\Delta G_{bind} = \Delta G_{bind}^C + \Delta G_{prot}$$

$$\Delta G_{bind} = \Delta G_{bind}^C + RT(pH - pK_a) \ln(10)$$

B When multiple protonation states can bind to the protein



$$\Delta G_{bind} = \Delta G_{bind}^N + \Delta G_{corr}$$

$$\Delta G_{bind} = \Delta G_{bind}^N - RT \ln \frac{1 + e^{-\frac{\Delta G_{bind}^C - \Delta G_{bind}^N}{RT}} 10^{pK_a - pH}}{1 + 10^{pK_a - pH}}$$

Figure 11. Aqueous pK_a of the ligand can influence overall protein-ligand binding affinity. **A** When only the minor aqueous protonation state contributes to protein-ligand complex formation, overall binding free energy (ΔG_{bind}) needs to be calculated as the sum of binding affinity of the minor state and the protonation penalty of that state. **B** When multiple charge states contribute to complex formation, the overall free energy of binding includes a multiple protonation states correction (MPSC) term (ΔG_{corr}). MPSC is a function of pH, aqueous pK_a of the ligand, and the difference between the binding free energy of charged and neutral species ($\Delta G_{bind}^C - \Delta G_{bind}^N$).

We can also ask ourselves whether or not an MPSC calculated based on an inaccurate pK_a should be included in binding affinity predictions in different circumstances such as underestimated or overestimated pK_a values and charged states with higher or lower affinities than the neutral states. We tried to capture these 4 circumstances in four quadrants of Fig. 12. In the case of overestimated pK_a values (Fig. 12E-H) it can be seen that for the most of the pH- pK_a range it is more advantageous to include the predicted MPSC in affinity calculations, except a smaller window where the opposite choice would be more advantageous. For instance, for the system with $\Delta \Delta G = 2$ kcal/mol and overestimated pK_a (Fig. 12E) for the pH- pK_a region between -0.5 and 2, including predicted ΔG_{corr} causes more error than ignoring MPSC.

In practice, we normally do not know the exact magnitude or the direction of the error of our predicted pK_a . Therefore, using simulated MPSC error plots to decide when to include MPSC in binding affinity predictions is not possible. However, based on the analysis of a case with 1 unit of pK_a error, including MPSC correction would be more often than not helpful in improving binding affinity predictions. The detrimental effect of pK_a inaccuracy is still significant. Hopefully, future improvements in pK_a prediction methods can improve the accuracy of MPSC and binding affinity predictions of ligands which have multiple protonation states that contribute to aqueous or complex populations. Being able to predict pK_a s with 0.5 units accuracy would significantly help the binding affinity models to incorporate more accurate MPSC terms.

3.4 Take-away lessons from SAMPL6 pK_a Challenge

The SAMPL6 pK_a Challenge showed that in general pK_a prediction performance of computational methods is lower than expected for drug-like molecules. Our expectation prior to the blind challenge was that well-developed methods would achieve prediction errors as low as 0.5 pK_a units and reliable predictions of charge and tautomer states. There are many factors that complicate predicting pK_a values of drug-like molecules: multiple titratable sites, tautomerization, frequent presence of heterocycles, and extended conjugation patterns, as well as a high number of rotatable bonds, and the possibility of intramolecular hydrogen bonds. Macroscopic pK_a value predictions have not yet reached experimental accuracy (Inter-method variability of macroscopic pK_a measurements can be around 0.5 pK_a units [23]). There was not a single method in the SAMPL6 Challenge that achieved RMSE around 0.5 or lower for macroscopic pK_a predictions for the 24 molecule set of kinase inhibitor fragment-like molecules. Lower RMSE values were observed in the microscopic pK_a evaluation section of this study for some methods; however, the 8 molecule set used for that analysis poses a very limited dataset to reach conclusions about general expectations for drug-like molecules.

As the majority of experimental data was in the form of macroscopic pK_a values, we had to adopt a numerical matching

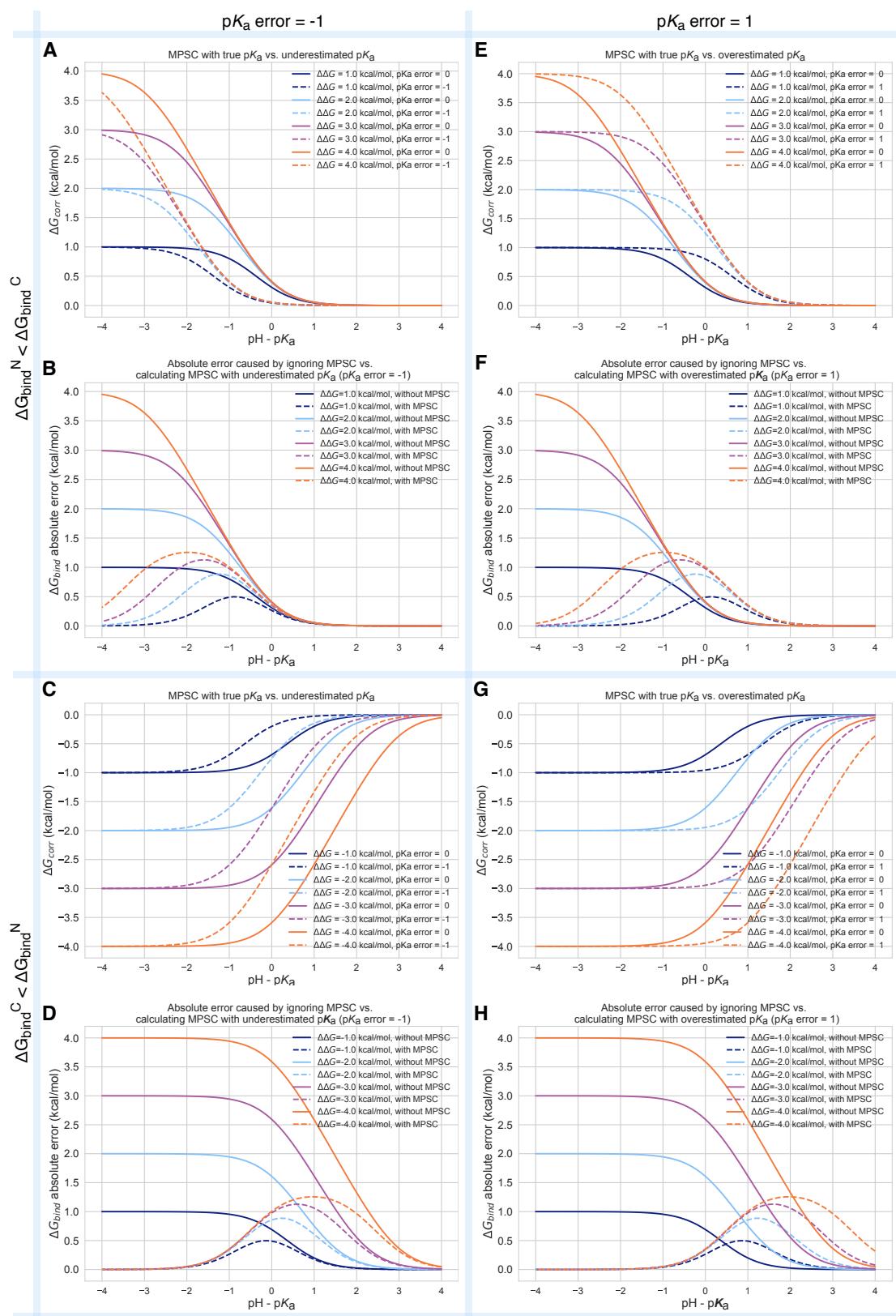


Figure 12. Inaccuracy of pK_a prediction (± 1 unit) affects the accuracy of MPSC and overall protein-ligand binding free energy calculation in varying amounts based on aqueous pK_a value and relative binding affinity of individual protonation states ($\Delta\Delta G = \Delta G_{bind}^C - \Delta G_{bind}^N$). All calculations are made for 25°C, and a ligand with a single basic titratable group. **A, C, E, and G show MPSC (ΔG_{corr}) calculated with true vs. inaccurate pK_a . **B, D, F, and H** show the comparison of the absolute error to ΔG_{bind} caused by ignoring the MPSC completely (solid lines) vs. calculating MPSC based in inaccurate pK_a value (dashed lines). These plots provide guidance on when it is beneficial to include MPSC correction based on pK_a error, $pH - pK_a$, and $\Delta\Delta G$.**

771 algorithm (Hungarian matching) to pair predicted and experimental values to calculate performance statistics of macroscopic
772 pK_a predictions. Accuracy, correlation, and extra/missing pK_a prediction counts were the main metrics for macroscopic pK_a
773 evaluations. An RMSE range of 0.7 to 3.2 pK_a units was observed. Only five methods achieved RMSE between 0.7-1 pK_a units,
774 while an RMSE between 1.5-3 log units was observed for the majority of methods. All four methods of the LFER category and three
775 out of 5 QSPR/ML methods achieved RMSE less than 1.5 pK_a units. All the QM methods that achieved this level of performance
776 included linear empirical corrections to rescale and unbias their pK_a predictions.

777 Based on the consideration of multiple error metrics, we compiled a shortlist of consistently-well performing methods for
778 macroscopic pK_a evaluations. Two methods from QM+LEC methods, one QSPR/ML, two empirical methods achieved consistent
779 performance according to many metrics. The common features of the two empirical methods were their large training sets
780 (16000-17000 compounds) and being commercial prediction models.

781 There were four submissions of QM-based methods that utilized COSMO-RS implicit solvation model. While three of these
782 achieved the lowest RMSE among QM-based methods (*xvxzd*, *yqkga*, and *8xt50*) [44], one of them showed the highest RMSE
783 (*0hxtm* (COSMOtherm_FINE17)). The comparison of these methods indicates that capturing conformational ensemble of mi-
784 crostates, high level QM calculations, and RRHO corrections contribute to better macroscopic pK_a predictions. Linear empirical
785 corrections applied QM calculations improved results, especially when the linear correction is calibrated for an experimental
786 dataset using the same level of theory as the deprotonation free energy predictions (as in *xvxzd*). This challenge also points to
787 the advantage of COSMO-RS solvation approach compared to other implicit solvent models.

788 Molecules that posed greatest difficulty for pK_a predictions were determined by comparing the macroscopic pK_a prediction
789 accuracy of each molecules averaged over all methods submitted to the challenge. pK_a prediction errors were higher for com-
790 pounds with sulfur-containing heterocycles, iodo, and bromo groups. This trend was also conserved when only QM-based
791 methods were analyzed. SAMPL6 pK_a dataset consisted of only 24 small molecules which limited our ability to statistically con-
792 firm this conclusion, however, we believe it is worth reporting molecular features that coincided with larger errors even if we
793 can not evaluate the reason for these failures.

794 Utilizing a numerical matching algorithm to pair experimental and predicted macroscopic pK_a values was a necessity, how-
795 ever, this approach did not capture all aspects of prediction errors. Computing the number of missing or extra pK_a predictions
796 remaining after Hungarian matching, provided a window of observing macroscopic pK_a prediction errors such as the number
797 of macroscopic transitions or ionization states expected in a pH interval. In pK_a evaluation studies it is very typical to just focus
798 on pK_a value errors evaluated after matching, and to ignore pK_a prediction errors that the matching protocol can not capture.
799 Ignored prediction errors include predicting missing or extra pK_a s and failing to predict the correct charge states. SAMPL6 pK_a
800 Challenge results showed sporadic presence of missing pK_a predictions and very frequent case of extra pK_a predictions. Both
801 indicate failures to capture the correct ionization states. The traditional way of evaluating pK_a s that only focuses on the pK_a
802 value error after some sort of numerical match between predictions and experimental values may have motivated these types
803 of errors as there would be no penalty for missing a macroscopic deprotonation and predicting an extra one. This problem does
804 not seem to be specific to any method category.

805 We used the eight molecule subset of SAMPL6 compounds with NMR-based dominant microstate sequence information to
806 demonstrate the advantage of evaluating pK_a prediction on the level of microstates. Comparison of statistics computed for the
807 8 molecule dataset by Hungarian matching and microstate-based matching showed how Hungarian matching, despite being the
808 optimal matching algorithm, can mask errors in pK_a predictions. Errors computed by microstate-based matching were larger
809 compared to numerical matching algorithms in terms of RMSE. Microscopic pK_a analysis with numerical matching algorithms
810 may mask errors due to the higher number of guesses made. Numerical matching based on pK_a values also ignores information
811 regarding the relative population of states. Therefore, it can lead to pK_a s defined between very low energy microstate pairs to
812 be matched to the experimentally observable pK_a between microstates of higher populations. Of course, the predicted pK_a
813 value could be correct however the predicted microstates would be wrong. Such mistakes caused by Hungarian matching were
814 observed frequently in SAMPL6 results and therefore we decided microstate-based matching of pK_a values provides a more
815 realistic picture of method performance.

816 Some QM and LFER methods made mistakes in predicting the dominant tautomers of the ionization states. Dominant tau-
817 tomer prediction seemed to be a more prominent problem for charged tautomers than the neutral tautomer. The easiest way
818 to extract dominant microstate sequence from predictions is to calculate the relative free energy of microstates at any reference
819 pH, and determining the lowest energy state in each ionization state. Errors in dominant microstate predictions were very rare
820 for neutral tautomers but more frequent in cationic tautomers with +1 charge of the 8 molecule set. SM14 was the molecule
821 with the lowest dominant microstate prediction accuracy, while dominant microstates predictions for SM15 were perfect for all

822 molecules. SM14 and SM15 both have two experimental pK_a s and benzimidazole scaffold. The difference between them is the
823 distance between the experimental pK_a values which is smaller for SM14. These results make sense from the perspective of
824 relative free energies of microstates. Closer pK_a values mean that the free energy difference between different microstates are
825 smaller for SM14, and therefore any error in predicting the relative free energy of tautomers is more likely to cause reordering of
826 relative populations of microstates and impact the accuracy of dominant microstate predictions. It would have been extremely
827 informative to evaluate the tautomeric ratios and relative free energy predictions of microstates, however, experimental data
828 was missing for this approach.

829 The overall assessment of SAMPL6 pK_a Challenge captured non-stellar performance for microscopic and macroscopic pK_a
830 predictions which can be detrimental to the accuracy of protein-ligand affinity predictions and other pH-dependent physico-
831 chemical property predictions such as distribution coefficients, membrane permeability, and solubility. Protein-ligand binding
832 affinity predictions rely on pK_a predictions in two ways: determination of relevant aqueous microstates and the free energy
833 penalty to reach these states. Microscopic pK_a predictions with better accuracy are needed for accurate incorporation of mul-
834 tiple protonation state correction (MPSC) to overall binding affinity calculations. We simulated the effect of overestimating or
835 underestimating pK_a of a ligand by one unit on overall binding affinity prediction for a ligand where both cation and neutral
836 states contribute to binding affinity. pK_a prediction error of this magnitude (assuming dominant tautomers were predicted cor-
837 rectly) could cause up to 0.9 and 1.2 kcal/mol error in overall binding affinity when the binding affinity of protonation states
838 are 2 or 4 kcal/mol different, respectively. For the case of 4 kcal/mol binding affinity difference between protonation states the
839 pH- pK_a range that the error would be larger than 0.5 kcal/mol surprisingly spans around 3.5 pH units. We demonstrated that the
840 range of pH- pK_a value that MPSC needs to be incorporated in binding affinity predictions can be wider than the widely assumed
841 range of 2 pH units, based on the affinity difference between protonation states. At the level of 1 unit pK_a error incorporating
842 MPSC would improve binding affinity predictions more often than not. If microscopic pK_a could be predicted with 0.5 pK_a units
843 of accuracy, MPSC calculations would be much more reliable.

844 There are multiple factors to consider when deciding which pK_a prediction method to utilize. These factors include the
845 accuracy of microscopic and macroscopic pK_a values, the accuracy of the number and the identity of ionization states predicted
846 within the experimental pH interval, the accuracy of microstates predicted within the experimental pH interval, the accuracy of
847 tautomeric ratio (i.e. relative free energy between microstates), how costly is the calculation in terms of time and resources, and
848 whether one has access to software licenses that might be required.

849 All of the top-performing empirical methods were developed as commercial software that require licenses to run, and there
850 were not any open-source alternatives for empirical pK_a predictions. Since then two publications reported open-source machine
851 learning-based pK_a prediction methods, however, one can only predict the most acidic or most basic macroscopic pK_a values
852 of a molecule [47] and the second one is only trained for predicting pK_a values of monoprotic molecules [48]. Recently a pK_a
853 prediction methodology was published that describes a mixed approach of semi-empirical QM calculations and machine learn-
854 ing that can predict macroscopic pK_a s of both mono-and polyprotic species [49]. The authors reported RMSE of 0.85 for the
855 retrospective analysis performed on the SAMPL6 dataset.

856 3.5 Suggestions for future blind challenge design and evaluation of pK_a predictions

857 This analysis helped us understand the current state of the field and led to many lessons. We believe the highest benefit can
858 be achieved if further iteration of small molecule pK_a prediction challenges can be organized, creating motivation for improving
859 protonation state prediction methods for drug-like molecules. In future challenges, it is desirable to increase chemical diversity
860 to cover more of common scaffolds [50] and functional groups [51] seen in drug-like molecules, and gradually increasing the
861 complexity of molecules.

862 Future challenges should promote stringent evaluation for pK_a prediction methods from the perspective of microscopic pK_a
863 and microstate predictions. It is necessary to assess the capability of pK_a prediction methods to capture the free energy profile of
864 microstates of multiprotic molecules. This is critical because pK_a predictions are often utilized to determine relevant protonation
865 states and tautomers of small molecules that must be captured in other physical modeling approaches, such as protein-ligand
866 binding affinity or distribution coefficient predictions. Different tautomers can have different binding affinities and partition
867 coefficients.

868 In this paper, we demonstrated how experimental microstate information can guide the analysis further than the typical pK_a
869 evaluation approach that has been used so far. The traditional pK_a evaluation approach focuses solely on the numerical error
870 of the pK_a values and neglects the difference between macroscopic and microscopic pK_a definitions. This is mainly caused by
871 the lack of pK_a datasets with microscopic detail. To improve pK_a and protonation state predictions of multiprotic molecules it

872 is necessary to embrace the difference between macroscopic and microscopic pK_a definitions and select strategies for experi-
873 mental data collection and prediction evaluation accordingly. In SAMPL6 Challenge the analysis was limited by the availability
874 of experimental microscopic data as well. As usual macroscopic pK_a values were abundant (24 molecules) and limited data
875 on microscopic states was available (8 molecules), although the latter opened new avenues for evaluation. For future blind
876 challenges for multiprotic compounds, striving to collect experimental datasets with microscopic pK_a s would be very beneficial.
877 Benchmark datasets of microscopic pK_a s are currently missing. This limits the improvement of pK_a and tautomer prediction
878 methods for multiprotic molecules. If the collection of experimental microscopic pK_a s is not possible due to time and resource
879 cost of such NMR experiments, at least supplementing the more automated macroscopic pK_a measurements with NMR-based
880 determination of the dominant microstate sequence or tautomeric ratios of each ionization state can create very useful bench-
881 mark datasets. This supplementary information can allow microstate-based assignment between experimental and predicted
882 pK_a s and a more realistic assessment of method performance.

883 If the only available experimental data is in the form of macroscopic pK_a values, the best way to evaluate computational
884 predictions is by calculating predicted macroscopic pK_a from microscopic pK_a predictions. With the conversion of microscopic
885 pK_a to macroscopic pK_a s all the structural information about the titration site is lost and only remaining information is the total
886 charge of macroscopic ionization states. Unfortunately, most macroscopic pK_a measurements including potentiometric and
887 spectrophotometric methods do not capture the absolute charge of the macrostates. The spectrophotometric method does
888 not measure charge at all. The potentiometric method can only capture the relative charge change between macrostates. Only
889 pH-dependent solubility based pK_a estimations can differentiate the neutral and charged states from one another. So it is very
890 common to have experimental datasets of macroscopic pK_a without any charge or protonation position information regarding
891 the macrostates. This causes an issue of assigning predicted and experimental pK_a values before any error statistics can be
892 calculated. As delineated by Fraczkiewicz et. al. the fairest and reasonable solution for pK_a matching problem involves an
893 assignment algorithm that preserves the order of predicted and experimental microstates and uses the principle of smallest
894 differences to pair values [23]. We recommend Hungarian matching with the squared error cost function. The algorithm is
895 available in SciPy package (scipy.optimize.linear_sum_assignment) [33]. In addition to the analysis of numerical error statistics
896 after Hungarian matching, at the very least number of missing and extra pK_a predictions must be reported based on unmatched
897 pK_a values. Missing or extra pK_a predictions point to a problem with capturing the right number of ionization states within
898 the pH interval of the experimental measurements. We have demonstrated that for microscopic pK_a predictions performance
899 analysis based in Hungarian matching results in overly optimistic and misleading results, instead the employed microstate-based
900 matching provided a more realistic assessment.

901 We allowed three different submission types in SAMPL6 to capture all the necessary information related to pK_a predictions.
902 These were (1) macroscopic pK_a values, (2) microscopic pK_a values and microstate pair identities, (3) fractional population of
903 microstates with respect to pH. We realized later that collecting fractional populations of microstates was redundant since mi-
904 croscopic pK_a values and microstate pairs capture all the necessary information to construct fractional population vs. pH curves.
905 Only microscopic and macroscopic pK_a values were used for the challenge analysis presented in this paper. While exploring ways
906 to evaluate SAMPL6 pK_a Challenge results, we developed a better way to capture microscopic pK_a predictions as presented in
907 an earlier paper [30]. This alternative reporting format consists of charge and relative free energy of microstates with respect to
908 a reference microstate and pH. This approach presents the most concise method of capturing all necessary information regard-
909 ing microscopic pK_a predictions and allows calculation of predicted microscopic pK_a s, microstate population with respect to pH,
910 macroscopic pK_a s, macroscopic population with respect to pH, and tautomer ratios. Still, there may be methods developed to
911 predict macroscopic pK_a s directly instead of computing it from microstate predictions that justifies allowing a macroscopic pK_a
912 reporting format. In future challenges, we recommend collecting pK_a predictions with two submission types: (1) macroscopic
913 pK_a values together with the charges of the macrostates and (2) microstates, their total charge, and relative free energies with
914 respect to a specified reference microstate and pH. This approach is being used in SAMPL7.

915 In SAMPL6 we provided an enumerated list of microstates and their assigned microstate IDs because we were worried about
916 parsing submitted microstates in SMILES from different sources correctly. There were two disadvantages to this approach. First,
917 this list of enumerated microstates was used as input by some participants which was not our intention. Second, the first it-
918 eration of enumerated microstates was not complete. We had to add new microstates and assign them microstate IDs for a
919 couple of rounds until reaching a complete list. In future challenges, a better way of handling the problem of capturing predicted
920 microstates would be asking participants to specify the predicted protonation states themselves and assigning identifiers after
921 the challenge deadline to aid comparative analysis. This would prevent the partial unblinding of protonation states and allow
922 the assessment of whether methods can predict all the relevant states independently, without relying on a provided list of mi-

923 crostates. Predicted states can be submitted as mol2 files that represents the microstate with explicit hydrogens. The organizers
924 must only provide the microstate that was selected as the reference state for the relative microstate free energy calculations.
925

926 In the SAMPL6 pK_a Challenge there was not a requirement that prediction sets should report predictions for all compounds.
927 Some participants reported predictions for only a subset of compounds which may have led these methods to look more ac-
928 curate than others, due to missing predictions. In the future, it will be better to allow submissions of only complete sets for a
929 better comparison of method performance.

930 A wide range of methods participated in the SAMPL6 pK_a Challenge from very fast QSPR methods to QM methods with a
931 high-level of theory and extensive exploration of conformational ensembles. In the future, it would be interesting to capture
932 computing costs in terms of average compute hours per molecule. This can provide guidance to future users of pK_a prediction
933 methods for selection of which method to use.

934 Future blind challenges can maximize learning opportunities by evaluating predictions of different physicochemical prop-
935 erties for the same molecules in consecutive challenges. In SAMPL6 we organized both pK_a and $\log P$ challenges. Unfortunately
936 only a subset of compounds in pK_a datasets were suitable for the potentiometric $\log P$ measurements. Still, comparing pre-
937 diction performance of common compounds in both challenges can lead to beneficial insights especially for physical modeling
938 techniques if there are common aspects that are beneficial or detrimental to prediction performance. For example, in SAMPL6
939 pK_a and $\log P$ Challenges COSMO-RS and EC-RISM solvation models achieved good performance. Having access to a variety
940 of physicochemical property measurements can also help identification of error sources. For example, dominant microstates
941 determined for pK_a challenge can provide information to check if correct tautomers are modeling in a $\log P$ or $\log D$ challenge.
942 pK_a prediction is a requirement for $\log D$ prediction and experimental pK_a values can help diagnosing the source of errors in
943 $\log D$ predictions better. The physical challenges in SAMPL7, which is currently running with a deadline of September 30th, 2020,
944 follow this principle and include both pK_a , $\log P$, and membrane permeability properties for a set of monoprotic compounds.
945 We hope that future pK_a challenges can focus on multiprotic drug-like compounds with microscopic pK_a measurements for an
946 in-depth analysis.

946 4 Conclusion

947 The first SAMPL6 pK_a Challenge focused on kinase inhibitor like molecules to assess the performance of pK_a predictions for
948 drug-like molecules. With wide participation we had an opportunity to prospectively evaluate pK_a predictions spanning vari-
949 ous empirical and QM based approaches. A small number of popular pK_a prediction methods that were missing from blind
950 submissions were added as reference calculations after the challenge deadline.

951 The experimental dataset consisted of spectrophotometric measurements of 24 molecules and some of which were mul-
952 tiprotic. There was also experimental data on the dominant microstate sequence of a subset of the challenge molecules, but
953 not direct microscopic pK_a measurements. We have performed a comparative analysis of methods represented in the blind
954 challenge in terms of both macroscopic and microscopic pK_a prediction performance avoiding any assumptions about the ex-
955 perimental pK_a s.

956 Here, we used Hungarian matching to assign predicted and experimental values for the calculation of accuracy and corre-
957 lation statistics, because the majority of the experimental data was macroscopic pK_a values. In addition to evaluating error in
958 predicted pK_a values, we also reported the macroscopic pK_a errors that were not captured by the match between experimental
959 and predicted pK_a values. These were extra or missing pK_a predictions which are important indicators that predictions are failing
960 to capture the correct ionization states.

961 We evaluated microscopic pK_a predictions utilizing the experimental dominant microstate sequence data of eight molecules.
962 This experimental data allowed us to use microstate-based matching for evaluating the accuracy of microscopic pK_a values
963 in a more realistic way. We have determined that QM and LFER predictions had lower accuracy in determining the dominant
964 tautomer of the charged microstates than the neutral states. For both macroscopic and microscopic pK_a predictions we have
965 determined methods that were consistently well-performing according to multiple statistical metrics. Focusing on the com-
966 parison of molecules instead of methods for macroscopic pK_a prediction accuracy indicated molecules with sulfur-containing
967 heterocycles, iodo, and bromo groups suffered from lower pK_a prediction accuracy.

968 The overall performance of pK_a predictions as captured in this challenge is concerning for the application of pK_a prediction
969 methods in computer-aided drug design. Many computational methods for predicting target affinities and physicochemical
970 properties rely on pK_a predictions for determining relevant protonation states and the free energy penalty of such states. 1 unit
971 of pK_a error is an optimistic estimate of current macroscopic pK_a predictions for drug-like molecules based on SAMPL6 Challenge

972 where errors in predicting the correct number of ionization states or determining the correct dominant microstate were also
973 common to many methods. In the absence of other sources of errors, we showed that 1 unit over- or underestimation of the
974 pK_a of a ligand can cause significant errors in the overall binding affinity calculation due to errors in multiple protonation state
975 correction factor.

976 The SAMPL6 GitHub Repository contains all information regarding the challenge structure, experimental data, blind prediction
977 submission sets, and evaluation of methods. The repository will be useful for future follow up analysis and the experimental
978 measurements can continue to serve as a benchmark dataset for testing methods.

979 In this article, we aimed to demonstrate not only the comparative analysis of the pK_a prediction performance of contemporary
980 methods for drug-like molecules, but also to propose a stringent pK_a prediction evaluation strategy that takes into account
981 differences in microscopic and macroscopic pK_a definitions. We hope that this study will guide and motivate further improvement
982 of pK_a prediction methods.

983 5 Code and data availability

- 984 • SAMPL6 pK_a challenge instructions, submissions, experimental data and analysis is available at
<https://github.com/samplchallenges/SAMPL6>

985 6 Overview of supplementary information

986 Contents of the Supplementary Information:

- 987 • TABLE S1: SMILES and InChI identifiers of SAMPL6 pK_a Challenge molecules.
- 988 • TABLE S2: Evaluation statistics calculated for all macroscopic pK_a prediction submissions based on Hungarian match for
989 24 molecules.
- 990 • TABLE S3: Evaluation statistics calculated for all microscopic pK_a prediction submissions based on Hungarian match for 8
991 molecules with NMR data.
- 992 • TABLE S4: Evaluation statistics calculated for all microscopic pK_a prediction submissions based on microstate match for 8
993 molecules with NMR data.
- 994 • FIGURE S1: Dominant microstates of 8 molecules were determined based on NMR measurements.
- 995 • FIGURE S2: MAE of macroscopic pK_a predictions of each molecule did not show any significant correlation with any molecular
996 descriptor.
- 997 • FIGURE S3: The value of macroscopic pK_a was not a factor affecting prediction error seen in SAMPL6 Challenge according
998 to the analysis with Hungarian matching.
- 999 • FIGURE S4: There was low agreement between experimental dominant microstate pairs and the predicted microstate pairs
1000 selected by Hungarian algorithm for microscopic pK_a predictions.

1001 Extra files included in *SAMPL6-supplementary-documents.tar.gz*:

- 1002 • SAMPL6-pKa-chemical-identifiers-table.csv
- 1003 • macroscopic-pKa-statistics-24mol-hungarian-match.csv
- 1004 • microscopic-pKa-statistics-8mol-hungarian-match-table.csv
- 1005 • microscopic-pKa-statistics-8mol-microstate-match-table.csv
- 1006 • experimental-microstates-of-8mol-based-on-NMR.csv
- 1007 • enumerate-microstates-with-Epik-and-OpenEye-QUACPAC.ipynb
- 1008 • molecule_ID_and_SMILES.csv

1009 7 Author Contributions

1010 Conceptualization, MI, JDC ; Methodology, MI, JDC, ASR ; Software, MI, AR, ASR ; Formal Analysis, MI, ASR ; Investigation, MI ; Re-
1011 sources, JDC, DLM; Data Curation, MI ; Writing-Original Draft, MI; Writing - Review and Editing, MI, JDC, ASR, AR, DLM; Visualization,
1012 MI, AR ; Supervision, JDC, DLM ; Project Administration, MI ; Funding Acquisition, JDC, DLM.

1013 8 Acknowledgments

1014 We would like to acknowledge the infrastructure and website support of Mike Chiu that allowed a seamless collection of chal-
1015 lenge submissions. Mike Chiu also provided assistance with constructing a submission validation script to ensure all submissions

1016 adhered to the machine-readable format. We are grateful to Kiril Lanevskij for suggesting the Hungarian algorithm for matching
1017 experimental and predicted pK_a values. We would like to thank Thomas Fox for providing MoKa reference calculations. We
1018 acknowledge Caitlin Bannan for guidance on defining a working microstate definition for the challenge and guidance for design-
1019 ing the challenge. We thank Brad Sherborne for his valuable insights at the conception of the pK_a challenge and connecting
1020 us with Timothy Rhodes and Dorothy Levorse who were able to provide resources and expertise for experimental measure-
1021 ments performed at MRL. We acknowledge Paul Czodrowski who provided feedback on multiple stages of this work: challenge
1022 construction, purchasable compound selection, and manuscript draft. MI, JDC, and DLM gratefully acknowledge support from
1023 NIH grant R01GM124270 supporting the SAMPL Blind Challenges. MI, ASR, AR, and JDC acknowledge support from the Sloan
1024 Kettering Institute. JDC acknowledges support from NIH grant P30CA008748 and NIH grant R01GM121505. DLM appreciates
1025 financial support from the National Institutes of Health (R01GM108889) and the National Science Foundation (CHE 1352608).
1026 MI acknowledges Doris J. Hutchinson Fellowship. MI, ASR, AR, and JDC are grateful to OpenEye Scientific for providing a free
1027 academic software license for use in this work. MI, ASR, AR, and JDC thank Janos Fejervari and ChemAxon team that gave us
1028 permission to include ChemAxon/Chemicalize pK_a predictions as a reference prediction in challenge analysis.

1029 **9 Disclaimers**

1030 The content is solely the responsibility of the authors and does not necessarily represent the official views of the National
1031 Institutes of Health.

1032 **10 Disclosures**

1033 JDC was a member of the Scientific Advisory Board for Schrödinger, LLC during part of this study, and is a current Scientific
1034 Advisory Board member for OpenEye Scientific and scientific advisor to Foresite Labs. DLM is a current member of the Scientific
1035 Advisory Board of OpenEye Scientific and an Open Science Fellow with Silicon Therapeutics.

1036 The Chodera laboratory receives or has received funding from multiple sources, including the National Institutes of Health,
1037 the National Science Foundation, the Parker Institute for Cancer Immunotherapy, Relay Therapeutics, Entasis Therapeutics, Vir
1038 Biotechnology, Silicon Therapeutics, EMD Serono (Merck KGaA), AstraZeneca, Vir Biotechnology, XtalPi, the Molecular Sciences
1039 Software Institute, the Starr Cancer Consortium, the Open Force Field Consortium, Cycle for Survival, a Louis V. Gerstner Young
1040 Investigator Award, The Einstein Foundation, and the Sloan Kettering Institute. A complete list of funding can be found at <http://choderolab.org/funding>.

1042 **References**

- 1043 [1] **Manallack DT**, Prankerd RJ, Yuriev E, Oprea TI, Chalmers DK. The Significance of Acid/Base Properties in Drug Discovery. *Chem Soc Rev.* 2013; 42(2):485–496. doi: [10.1039/C2CS35348B](https://doi.org/10.1039/C2CS35348B).
- 1044 [2] **Charifson PS**, Walters WP. Acidic and Basic Drugs in Medicinal Chemistry: A Perspective. *Journal of Medicinal Chemistry.* 2014 Dec; 57(23):9701–9717. doi: [10.1021/jm501000a](https://doi.org/10.1021/jm501000a).
- 1045 [3] **Manallack DT**, Prankerd RJ, Nassta GC, Ursu O, Oprea TI, Chalmers DK. A Chemogenomic Analysis of Ionization Constants-Implications for Drug Discovery. *ChemMedChem.* 2013 Feb; 8(2):242–255. doi: [10.1002/cmdc.201200507](https://doi.org/10.1002/cmdc.201200507).
- 1046 [4] **de Oliveira C**, Yu HS, Chen W, Abel R, Wang L. Rigorous Free Energy Perturbation Approach to Estimating Relative Binding Affinities between Ligands with Multiple Protonation and Tautomeric States. *Journal of Chemical Theory and Computation.* 2019 Jan; 15(1):424–435. doi: [10.1021/acs.jctc.8b00826](https://doi.org/10.1021/acs.jctc.8b00826).
- 1047 [5] **Darvey IG**. The Assignment of pK_a Values to Functional Groups in Amino Acids. *Biochemical Education.* 1995 Apr; 23(2):80–82. doi: [10.1016/0307-4412\(94\)00150-N](https://doi.org/10.1016/0307-4412(94)00150-N).
- 1048 [6] **Bodner GM**. Assigning the pK_a 's of Polyprotic Acids. *Journal of Chemical Education.* 1986 Mar; 63(3):246. doi: [10.1021/ed063p246](https://doi.org/10.1021/ed063p246).
- 1049 [7] **Murray R**. Microscopic Equilibria. *Analytical Chemistry.* 1995 Aug; p. 1.
- 1050 [8] **Işık M**, Levorse D, Rustenburg AS, Ndukwe IE, Wang H, Wang X, Reibarkh M, Martin GE, Makarov AA, Mobley DL, Rhodes T, Chodera JD. pK_a Measurements for the SAMPL6 Prediction Challenge for a Set of Kinase Inhibitor-like Fragments. *Journal of Computer-Aided Molecular Design.* 2018 Oct; 32(10):1117–1138. doi: [10.1007/s10822-018-0168-0](https://doi.org/10.1007/s10822-018-0168-0).
- 1051 [9] **Bochevarov AD**, Watson MA, Greenwood JR, Philipp DM. 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.* 2016 Dec; 12(12):6001–6019. doi: [10.1021/acs.jctc.6b00805](https://doi.org/10.1021/acs.jctc.6b00805).

- 1062 [10] Selwa E, Kenney IM, Beckstein O, Iorga BI. SAMPL6: Calculation of Macroscopic pKa Values from Ab Initio Quantum Mechanical Free
1063 Energies. *Journal of Computer-Aided Molecular Design*. 2018 Oct; 32(10):1203–1216. doi: [10.1007/s10822-018-0138-6](https://doi.org/10.1007/s10822-018-0138-6).
- 1064 [11] Pickard FC, König G, Tofoleanu F, Lee J, Simonett AC, Shao Y, Ponder JW, Brooks BR. Blind Prediction of Distribution in the SAMPL5
1065 Challenge with QM Based Protomer and pK a Corrections. *Journal of Computer-Aided Molecular Design*. 2016 Nov; 30(11):1087–1100. doi:
1066 [10.1007/s10822-016-9955-7](https://doi.org/10.1007/s10822-016-9955-7).
- 1067 [12] Bannan CC, Mobley DL, Skillman AG. SAMPL6 Challenge Results from \$\$pK_a\$\$ Predictions Based on a General Gaussian Process Model.
1068 *Journal of Computer-Aided Molecular Design*. 2018 Oct; 32(10):1165–1177. doi: [10.1007/s10822-018-0169-z](https://doi.org/10.1007/s10822-018-0169-z).
- 1069 [13] Işık M, Levorse D, Mobley DL, Rhodes T, Chodera JD. Octanol-Water Partition Coefficient Measurements for the SAMPL6 Blind Prediction
1070 Challenge. *Journal of Computer-Aided Molecular Design*. 2020 Apr; 34(4):405–420. doi: [10.1007/s10822-019-00271-3](https://doi.org/10.1007/s10822-019-00271-3).
- 1071 [14] Işık M, Bergazin TD, Fox T, Rizzi A, Chodera JD, Mobley DL. Assessing the Accuracy of Octanol-Water Partition Coefficient Predictions in the
1072 SAMPL6 Part II Log P Challenge. *Journal of Computer-Aided Molecular Design*. 2020 Apr; 34(4):335–370. doi: [10.1007/s10822-020-00295-0](https://doi.org/10.1007/s10822-020-00295-0).
- 1073 [15] Kogej T, Muresan S. Database Mining for pKa Prediction. *Current Drug Discovery Technologies*. 2005; 2(4):221–229. doi:
1074 [10.2174/157016305775202964](https://doi.org/10.2174/157016305775202964).
- 1075 [16] Perrin DD, Dempsey B, Serjeant EP. pKa Prediction for Organic Acids and Bases. 1 ed. London and New York: Chapman and Hall; 1981.
- 1076 [17] Hammett LP. Physical Organic Chemistry. New York: McGraw-Hill; 1940.
- 1077 [18] Taft RW, Lewis IC. Evaluation of Resonance Effects on Reactivity by Application of the Linear Inductive Energy Relationship. V. Concerning
1078 a σR Scale of Resonance Effects^{1,2}. *Journal of the American Chemical Society*. 1959; 81(20):5343–5352. doi: [10.1021/ja01529a025](https://doi.org/10.1021/ja01529a025).
- 1079 [19] Xing L, Glen RC, Clark RD. Predicting $p K_a$ by Molecular Tree Structured Fingerprints and PLS. *Journal of Chemical Information and
1080 Computer Sciences*. 2003 May; 43(3):870–879. doi: [10.1021/ci020386s](https://doi.org/10.1021/ci020386s).
- 1081 [20] Zhang J, Kleinöder T, Gasteiger J. Prediction of $p K_a$ Values for Aliphatic Carboxylic Acids and Alcohols with Empirical Atomic Charge
1082 Descriptors. *Journal of Chemical Information and Modeling*. 2006 Nov; 46(6):2256–2266. doi: [10.1021/ci060129d](https://doi.org/10.1021/ci060129d).
- 1083 [21] Cruciani G, Milletti F, Storchi L, Sforza G, Goracci L. *In Silico* $p K_a$ Prediction and ADME Profiling. *Chemistry & Biodiversity*. 2009 Nov;
1084 6(11):1812–1821. doi: [10.1002/cbdv.200900153](https://doi.org/10.1002/cbdv.200900153).
- 1085 [22] Milletti F, Storchi L, Sforza G, Cruciani G. New and Original $p K_a$ Prediction Method Using Grid Molecular Interaction Fields. *Journal of
1086 Chemical Information and Modeling*. 2007 Nov; 47(6):2172–2181. doi: [10.1021/ci700018y](https://doi.org/10.1021/ci700018y).
- 1087 [23] Fraczkiewicz R. In Silico Prediction of Ionization. In: *Reference Module in Chemistry, Molecular Sciences and Chemical Engineering* Elsevier;
1088 2013. doi: [10.1016/B978-0-12-409547-2.02610-X](https://doi.org/10.1016/B978-0-12-409547-2.02610-X).
- 1089 [24] Simulations Plus ADMET Predictor v8.5;. Simulations Plus, Lancaster, CA, 2018. <https://www.simulations-plus.com/software/admetpredictor/physicochemical-biopharmaceutical/>.
- 1090 [25] Radak BK, Chipot C, Suh D, Jo S, Jiang W, Phillips JC, Schulten K, Roux B. Constant-pH Molecular Dynamics Simulations for Large Biomolecular
1091 Systems. *Journal of Chemical Theory and Computation*. 2017 Dec; 13(12):5933–5944. doi: [10.1021/acs.jctc.7b00875](https://doi.org/10.1021/acs.jctc.7b00875).
- 1092 [26] Special Issue: SAMPL6 (Statistical Assessment of the Modeling of Proteins and Ligands); October 2018. Volume 32, Issue 10. *Journal of
1093 Computer-Aided Molecular Design*.
- 1094 [27] Shelley JC, Cholleti A, Frye LL, Greenwood JR, Timlin MR, Uchimaya M. Epik: A Software Program for pK a Prediction and Protonation State
1095 Generation for Drug-like Molecules. *Journal of Computer-Aided Molecular Design*. 2007 Dec; 21(12):681–691. doi: [10.1007/s10822-007-9133-z](https://doi.org/10.1007/s10822-007-9133-z).
- 1096 [28] QUACPAC Toolkit 2017.Feb.1;. OpenEye Scientific Software, Santa Fe, NM. <http://www.eyesopen.com>.
- 1097 [29] OEChem Toolkit 2017.Feb.1;. OpenEye Scientific Software, Santa Fe, NM. <http://www.eyesopen.com>.
- 1098 [30] Gunner MR, Murakami T, Rustenburg AS, Işık M, Chodera JD. Standard State Free Energies, Not pKas, Are Ideal for Describing Small
1099 Molecule Protonation and Tautomeric States. *Journal of Computer-Aided Molecular Design*. 2020 May; 34(5):561–573. doi: [10.1007/s10822-020-00280-7](https://doi.org/10.1007/s10822-020-00280-7).
- 1100 [31] Kuhn HW. The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly*. 1955 Mar; 2(1-2):83–97. doi:
1101 [10.1002/nav.3800020109](https://doi.org/10.1002/nav.3800020109).
- 1102 [32] Munkres J. Algorithms for the Assignment and Transportation Problems. *J SIAM*. 1957 Mar; 5(1):32–28.

- 1106 [33] SciPy v1.3.1, Linear Sum Assignment Documentation; Sep 27, 2019. The SciPy community. https://docs.scipy.org/doc/scipy-1.3.1/reference/generated/scipy.optimize.linear_sum_assignment.html.
- 1108 [34] OpenEye pKa Prospector;. OpenEye Scientific Software, Santa Fe, NM. Accessed on Jan 23, 2018. <https://www.eyesopen.com/pka-prospector>.
- 1109 [35] ACD/pKa GALAS (ACD/Percepta Kernel v1.6);. Advanced Chemistry Development, Inc., Toronto, ON, Canada, 2018. <https://www.acdlabs.com/products/percepta/predictors/pKa>.
- 1111 [36] ACD/pKa Classic (ACD/Percepta Kernel v1.6);. Advanced Chemistry Development, Inc., Toronto, ON, Canada, 2018. <https://www.acdlabs.com/products/percepta/predictors/pKa>.
- 1113 [37] Chemicalize v18.23 (ChemAxon MarvinSketch v18.23);. ChemAxon, Budapest, Hungary, 2018. <https://docs.chemaxon.com/display/docs/pKa+Plugin>.
- 1115 [38] MoKa;. Molecular Discovery, Hertfordshire, UK, 2018. <https://www.moldiscovery.com/software/moka>.
- 1116 [39] Zeng Q, Jones MR, Brooks BR. Absolute and Relative pKa Predictions via a DFT Approach Applied to the SAMPL6 Blind Challenge. *Journal of Computer-Aided Molecular Design*. 2018 Oct; 32(10):1179–1189. doi: [10.1007/s10822-018-0150-x](https://doi.org/10.1007/s10822-018-0150-x).
- 1118 [40] Bochevarov AD, Harder E, Hughes TF, Greenwood JR, Braden DA, Philipp DM, Rinaldo D, Halls MD, Zhang J, Friesner RA. Jaguar: A High-
1119 Performance Quantum Chemistry Software Program with Strengths in Life and Materials Sciences. *International Journal of Quantum
1120 Chemistry*. 2013 Sep; 113(18):2110–2142. doi: [10.1002/qua.24481](https://doi.org/10.1002/qua.24481).
- 1121 [41] Tielker N, Eberlein L, Güssregen S, Kast SM. The SAMPL6 Challenge on Predicting Aqueous pKa Values from EC-RISM Theory. *Journal of
1122 Computer-Aided Molecular Design*. 2018 Oct; 32(10):1151–1163. doi: [10.1007/s10822-018-0140-z](https://doi.org/10.1007/s10822-018-0140-z).
- 1123 [42] Klamt A, Eckert F, Diedenhofen M, Beck ME. First Principles Calculations of Aqueous pK_a Values for Organic and Inorganic Acids Using
1124 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:
1125 [10.1021/jp034680o](https://doi.org/10.1021/jp034680o).
- 1126 [43] Eckert F, Klamt A. Accurate Prediction of Basicity in Aqueous Solution with COSMO-RS. *Journal of Computational Chemistry*. 2006 Jan;
1127 27(1):11–19. doi: [10.1002/jcc.20309](https://doi.org/10.1002/jcc.20309).
- 1128 [44] Pracht P, Wilcken R, Udvarhelyi A, Rodde S, Grimme S. High Accuracy Quantum-Chemistry-Based Calculation and Blind Prediction of
1129 Macroscopic pKa Values in the Context of the SAMPL6 Challenge. *Journal of Computer-Aided Molecular Design*. 2018 Oct; 32(10):1139–
1130 1149. doi: [10.1007/s10822-018-0145-7](https://doi.org/10.1007/s10822-018-0145-7).
- 1131 [45] Prasad S, Huang J, Zeng Q, Brooks BR. An Explicit-Solvent Hybrid QM and MM Approach for Predicting pKa of Small Molecules in SAMPL6
1132 Challenge. *Journal of Computer-Aided Molecular Design*. 2018 Oct; 32(10):1191–1201. doi: [10.1007/s10822-018-0167-1](https://doi.org/10.1007/s10822-018-0167-1).
- 1133 [46] Robert Fraczkiewicz MW, SAMPL6 pKa Challenge: Predictions of ionization constants performed by the S+pKa method implemented in
1134 ADMET Predictor software; February 22, 2018. The Joint D3R/SAMPL Workshop 2018. <https://drugdesigndata.org/about/d3r-2018-workshop>.
- 1135 [47] Mansouri K, Cariello NF, Korotcov A, Tkachenko V, Grulke CM, Sprankle CS, Allen D, Casey WM, Kleinstreuer NC, Williams AJ. Open-
1136 Source QSAR Models for pKa Prediction Using Multiple Machine Learning Approaches. *Journal of Cheminformatics*. 2019 Dec; 11(1). doi:
1137 [10.1186/s13321-019-0384-1](https://doi.org/10.1186/s13321-019-0384-1).
- 1138 [48] Baltruschat M, Czodrowski P. Machine Learning Meets pKa [Version 2; Peer Review: 2 Approved]. *F1000Research*. 2020; 9 (Chem Inf
1139 Sci)(113). doi: [10.12688/f1000research.22090.2](https://doi.org/10.12688/f1000research.22090.2).
- 1140 [49] Hunt P, Hosseini-Gerami L, Chrien T, Plante J, Ponting DJ, Segall M. Predicting pK_a Using a Combination of Semi-Empirical Quan-
1141 tum Mechanics and Radial Basis Function Methods. *Journal of Chemical Information and Modeling*. 2020 Jun; 60(6):2989–2997. doi:
1142 [10.1021/acs.jcim.0c00105](https://doi.org/10.1021/acs.jcim.0c00105).
- 1143 [50] Zdrazil B, Guha R. The Rise and Fall of a Scaffold: A Trend Analysis of Scaffolds in the Medicinal Chemistry Literature. *Journal of Medicinal
1144 Chemistry*. 2018 Jun; 61(11):4688–4703. doi: [10.1021/acs.jmedchem.7b00954](https://doi.org/10.1021/acs.jmedchem.7b00954).
- 1145 [51] Ertl P, Altmann E, McKenna JM. The Most Common Functional Groups in Bioactive Molecules and How Their Popularity Has Evolved over
1146 Time. *Journal of Medicinal Chemistry*. 2020 Aug; 63(15):8408–8418. doi: [10.1021/acs.jmedchem.0c00754](https://doi.org/10.1021/acs.jmedchem.0c00754).
- 1147 [52] OEMolProp Toolkit 2017.Feb.1;. OpenEye Scientific Software, Santa Fe, NM. <http://www.eyesopen.com>.

Table S1. SMILES and InChI identifiers of SAMPL6 pK_a Challenge molecules. A CSV version of this table can be found in *SAMPL6-supplementary-documents.tar.gz*. SMILES were generated by OpenEye OEChem [29]

SAMPL6 Molecule ID	Isomeric SMILES	InChI
SM01	c1cc2c(cc1O)c3c(o2)C(=O)NCCC3	InChI=1S/C12H11NO3/c14-7-3-4-10-9(6-7)8-2-1-5-13-12(15)11(8)16-10/h3-4,6,14H,1-2,5H2,(H,13,15)
SM02	c1ccc2c(c1)c(ncn2)Nc3cccc(c3)C(F)(F)	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)
SM03	c1ccc(cc1)Cc2nnnc(s2)NC(=O)c3cccs3	InChI=1S/C14H11N3OS2/c18-13(11-7-4-8-19-11)15-14-17-16-12(20-14)9-10-5-2-1-3-6-10/h1-8H,9H2,(H,15,17,18)
SM04	c1ccc2c(c1)c(ncn2)NCc3ccc(cc3)Cl	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)
SM05	c1ccc(c(c1)NC(=O)c2ccc(o2)Cl)N3CCCCC3	InChI=1S/C16H17ClN2O2/c17-15-9-8-14(21-15)16(20)18-12-6-2-3-7-13(12)19-10-4-1-5-11-19/h2-3,6-9H,1,4-5,10-11H2,(H,18,20)
SM06	c1cc2cccnnc2c(c1)NC(=O)c3cc(cnc3)Br	InChI=1S/C15H10BrN3O/c16-12-7-11(8-17-9-12)15(20)19-13-5-1-3-10-4-2-6-18-14(10)13/h1-9H,(H,19,20)
SM07	c1ccc(cc1)CNc2c3cccc3ncn2	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	Cc1ccc2c(c1)c(c(c=O)[nH]2)CC(=O)O)c3cccc3	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	COc1cccc(c1)Nc2c3cccc3ncn2.Cl	InChI=1S/C15H13N3O.CIH/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
SM10	c1ccc(cc1)C(=O)NCC(=O)Nc2nc3cccc3s2	InChI=1S/C16H13N3O2S/c20-14(10-17-15(21)11-6-2-1-3-7-11)19-16-18-2-8-4-5-9-13(12)22-16/h1-9H,10H2,(H,17,21)(H,18,19,20)
SM11	c1ccc(cc1)n2c3c(cn2)c(ncn3)N	InChI=1S/C11H9N5/c12-10-9-6-15-16(11(9)14-7-13-10)8-4-2-1-3-5-8/h1-7H,(H,2,12,13,14)
SM12	c1ccc2c(c1)c(ncn2)Nc3cccc(c3)Cl.Cl	InChI=1S/C14H10ClN3.CIH/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	Cc1cccc(c1)Nc2c3cc(c(c3ncn2)OC)OC	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	c1ccc(cc1)n2ncn3c2ccc(c3)N	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	c1ccc2c(c1)ncn2c3ccc(cc3)O	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	c1cc(c(c(c1)Cl)C(=O)Nc2ccncc2)Cl	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)
SM17	c1ccc(cc1)CSc2nnc(o2)c3ccncc3	InChI=1S/C14H11N3OS/c1-2-4-11(5-3-1)10-19-14-17-16-13(18-14)12-6-8-15-9-7-12/h1-9H,10H2
SM18	c1ccc2c(c1)c(=O)[nH]c(n2)CCC(=O)Nc3ncc(s3)Cc4ccc(c(c4)F)F	InChI=1S/C21H16F2N4O2S/c22-15-6-5-12(10-16(15)23)9-13-11-24-21(30-13)27-19(28)8-7-18-25-17-4-2-1-3-14(17)20(29)26-18/h1-6,10-11H,7-9H2,(H,24,27,28)(H,25,26,29)
SM19	CCOc1ccc2c(c1)sc(n2)NC(=O)Cc3ccc(c(c3)Cl)Cl	InChI=1S/C17H14Cl2N2O2S/c1-2-23-11-4-6-14-15(9-11)24-17(20-14)21-6(22)8-10-3-5-12(18)13(9)7-10/h3-7,9H,2,8H2,1H3,(H,20,21,22)
SM20	c1cc(cc(c1)OCc2ccc(cc2Cl)Cl)/C=C/3\C(=O)NC(=O)S3	InChI=1S/C17H11Cl2NO3S/c18-12-5-4-11(14(19)8-12)9-23-13-3-1-2-10(6-13)7-15-16(21)20-17(22)24-15/h1-8H,9H2,(H,20,21,22)/b15-7+
SM21	c1cc(cc(c1)Br)Nc2c(cnc(n2)Nc3cccc(c3)Br)F	InChI=1S/C16H11Br2FN4/c17-10-3-1-5-12(7-10)21-15-14(19)9-20-16(23-15)22-13-6-2-4-11(18)8-13/h1-9H,(H,20,21,22,23)
SM22	c1cc2c(cc(c(c2nc1)O))l	InChI=1S/C9H5l2NO/c10-6-4-7(11)9(13)8-5(6)2-1-3-12-8/h1-4,13H
SM23	CCOC(=O)c1ccc(cc1)Nc2cc(cnc(n2)Nc3ccc(cc3)C(=O)OCC)C	InChI=1S/C23H24N4O4/c1-4-30-21(28)16-6-10-18(11-7-16)25-20-14-15(3)24-23(27-20)26-19-12-8-17(9-13-19)22(29)31-5-2/h6-14H,4-5H2,1-3H3,(H2,24,25,26,27)
SM24	COc1ccc(cc1)c2c3c(ncn3oc2c4ccc(cc4)OC)NCCO	InChI=1S/C22H21N3O4/c1-27-16-7-3-14(4-8-16)18-19-21(23-11-12-26)24-13-25-22(19)29-20(18)15-5-9-17(28-2)10-6-15/h3-10,13,26H,11-12H2,1-2H3,(H,23,24,25)

1148 11 Supplementary Information

Microstate ID of Deprotonated State (A)	Microstate ID of Protonated State (HA)	Molecule ID	pKa (exp)	pKa SEM (exp)	pKa ID	Microstate identification source
		SM07	6.08	0.01	SM07_pKa1	NMR measurement
		SM14	5.3	0.01	SM14_pKa2	NMR measurement
		SM14	2.58	0.01	SM14_pKa1	NMR measurement
		SM02	5.03	0.01	SM02_pKa1	Estimated based on SM07 NMR measurement
		SM04	6.02	0.01	SM04_pKa1	Estimated based on SM07 NMR measurement
		SM09	5.37	0.01	SM09_pKa1	Estimated based on SM07 NMR measurement
		SM12	5.28	0.01	SM12_pKa1	Estimated based on SM07 NMR measurement
		SM13	5.77	0.01	SM13_pKa1	Estimated based on SM07 NMR measurement
		SM15	8.94	0.01	SM15_pKa2	Estimated based on SM14 NMR measurement
		SM15	4.7	0.01	SM15_pKa1	Estimated based on SM14 NMR measurement

Figure S1. Dominant microstates of 8 molecules were determined based on NMR measurements. Dominant microstate sequence of 6 derivatives were determined taking SM07 and SM14 as reference. Matched experimental pK_a values were determined by spectrophotometric pK_a measurements [8]. A CSV version of this table can be found in SAMPL6-supplementary-documents.tar.gz.

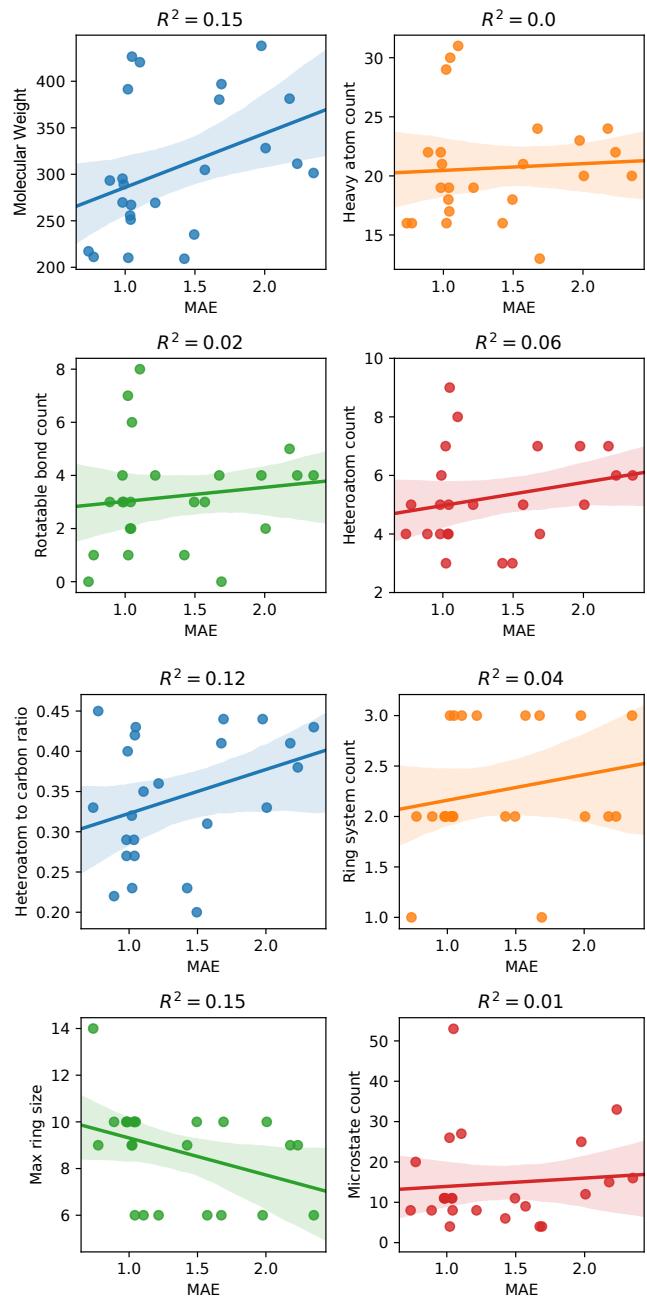


Figure S2. MAE of macroscopic pK_a predictions of each molecule did not show any significant correlation with any molecular descriptor.
 Plots show regression lines, 95% confidence intervals of the regression lines, and R^2 . The following molecular descriptors were calculated using OpenEye OEMolProp Toolkit [52]: molecular weight, non-terminal rotatable bond count, heteroatom to carbon ratio, maximum ring size, heavy atom count, heteroatom count, ring system count. Microstate count is based on the enumerated microstates for each compounds including additional microstates requested by participants.

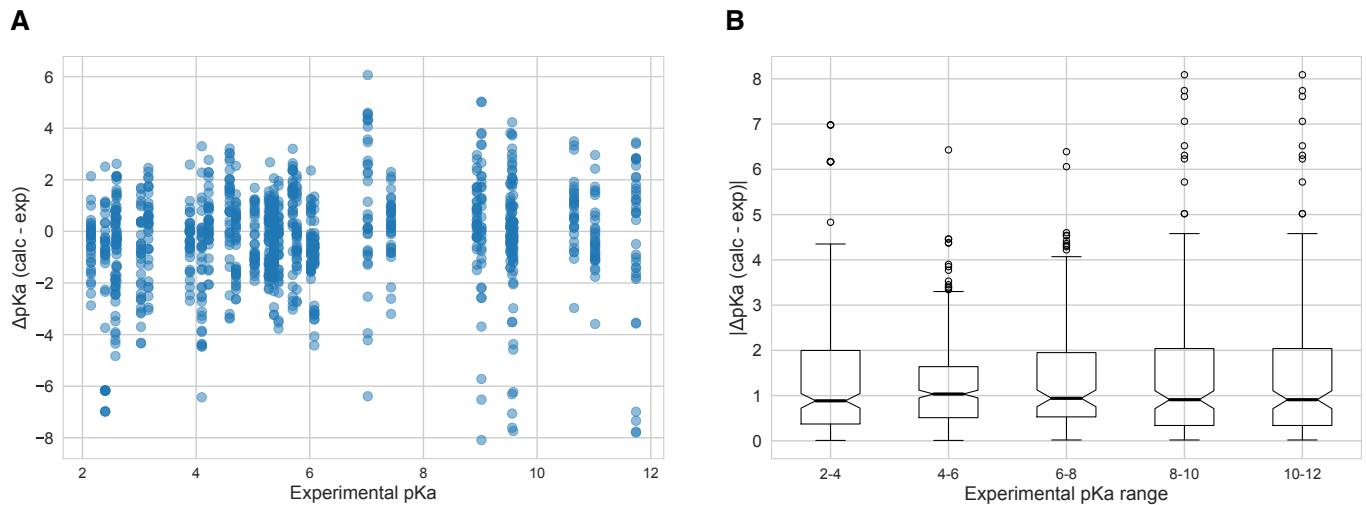


Figure S3. The value of macroscopic pK_a s was not a factor affecting prediction error seen in SAMPL6 Challenge according to the analysis with Hungarian matching. There was not clear trend between pK_a prediction error and the true pK_a error. Very high and very low pK_a values have similar inaccuracy compared to pK_a values close to 7. **A** Scatter plot of macroscopic pK_a prediction error calculated with Hungarian matching vs. experimental pK_a value **B** Box plot of absolute error of macroscopic pK_a predictions binned into 2 pK_a unit intervals of experimental pK_a .

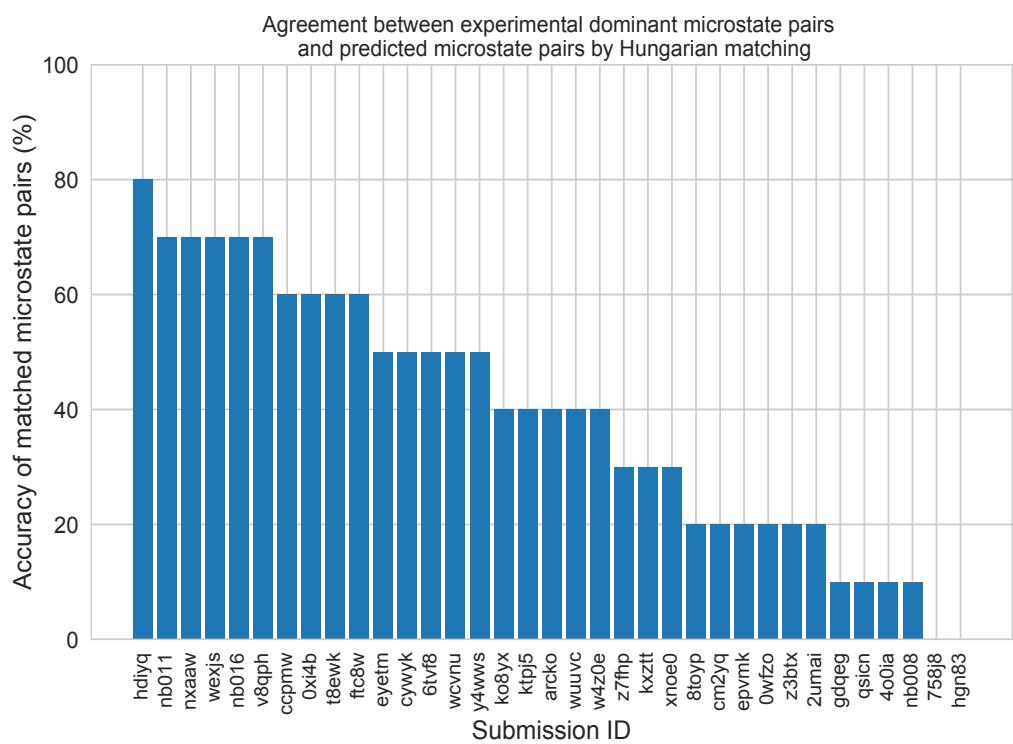


Figure S4. There was low agreement between experimental dominant microstate pairs and the predicted microstate pairs selected by Hungarian algorithm for microscopic pK_a predictions. This analysis could only be performed for 8 molecules with NMR data. Hungarian matching algorithm which matches predicted and experimental values considering only the closeness of the numerical value of pK_a and it often leads to predicted pK_a matches that described a different microstates pair than the experimentally observed dominant microstates..

Table S2. Evaluation statistics calculated for all macroscopic pK_a prediction submissions based on Hungarian match for 24 molecules. Methods are represented via their SAMPL6 submission IDs which can be cross-referenced with Table 1 for method details. There are eight 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), unmatched experimental pK_a s (number of missing pK_a predictions) and unmatched predicted pK_a s (number of extra pK_a predictions between 2 and 12. 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	Unmatched exp. pK_a s	Unmatched pred. pK_a s [2,12]
xvxzd	0.68 [0.54, 0.81]	0.58 [0.45, 0.71]	0.24 [-0.01, 0.45]	0.94 [0.88, 0.97]	0.92 [0.84, 1.02]	0.82 [0.68, 0.92]	2	4
gyuhx	0.73 [0.55, 0.91]	0.59 [0.44, 0.74]	0.03 [-0.23, 0.28]	0.93 [0.88, 0.96]	0.98 [0.90, 1.08]	0.88 [0.80, 0.94]	0	7
xmyhm	0.79 [0.52, 1.03]	0.56 [0.38, 0.77]	0.13 [-0.14, 0.41]	0.92 [0.85, 0.97]	0.96 [0.86, 1.08]	0.81 [0.68, 0.90]	0	3
nb017	0.94 [0.72, 1.16]	0.77 [0.58, 0.97]	-0.16 [-0.49, 0.16]	0.88 [0.81, 0.94]	0.94 [0.82, 1.08]	0.73 [0.60, 0.84]	0	6
nb007	0.95 [0.73, 1.15]	0.78 [0.60, 0.97]	0.05 [-0.29, 0.37]	0.88 [0.77, 0.95]	0.84 [0.77, 0.92]	0.79 [0.65, 0.89]	0	13
yqkga	1.01 [0.78, 1.23]	0.80 [0.59, 1.03]	-0.17 [-0.51, 0.19]	0.87 [0.78, 0.93]	0.93 [0.77, 1.08]	0.83 [0.72, 0.91]	0	1
nb010	1.03 [0.77, 1.26]	0.81 [0.61, 1.04]	0.24 [-0.11, 0.59]	0.87 [0.77, 0.94]	0.95 [0.83, 1.08]	0.80 [0.67, 0.90]	0	4
8xt50	1.07 [0.78, 1.36]	0.81 [0.58, 1.07]	-0.47 [-0.82, -0.14]	0.91 [0.84, 0.95]	1.08 [0.94, 1.22]	0.80 [0.68, 0.89]	0	0
nb013	1.10 [0.72, 1.47]	0.80 [0.56, 1.09]	-0.15 [-0.55, 0.22]	0.88 [0.78, 0.95]	1.09 [0.90, 1.25]	0.79 [0.64, 0.90]	0	6
nb015	1.27 [0.98, 1.56]	1.04 [0.80, 1.31]	0.13 [-0.32, 0.56]	0.87 [0.80, 0.93]	1.16 [0.94, 1.34]	0.78 [0.66, 0.86]	0	0
p0jba	1.31 [0.69, 1.73]	1.08 [0.43, 1.72]	-0.92 [-1.72, -0.11]	0.91 [0.51, 1.00]	1.18 [0.36, 1.72]	0.80 [0.00, 1.00]	0	0
37xm8	1.41 [0.93, 1.84]	1.01 [0.68, 1.38]	-0.18 [-0.69, 0.32]	0.83 [0.70, 0.93]	1.16 [0.98, 1.33]	0.70 [0.56, 0.83]	1	1
mkhqa	1.60 [1.13, 2.05]	1.24 [0.90, 1.62]	-0.32 [-0.89, 0.21]	0.80 [0.67, 0.91]	1.14 [0.98, 1.34]	0.64 [0.44, 0.79]	0	6
ttjd0	1.64 [1.20, 2.06]	1.30 [0.96, 1.67]	-0.12 [-0.70, 0.45]	0.81 [0.69, 0.91]	1.2 [1.03, 1.40]	0.65 [0.47, 0.80]	0	5
nb001	1.68 [1.05, 2.37]	1.21 [0.84, 1.68]	0.44 [-0.10, 1.03]	0.80 [0.70, 0.90]	1.16 [0.95, 1.42]	0.72 [0.55, 0.85]	0	7
nb002	1.70 [1.08, 2.38]	1.25 [0.89, 1.70]	0.51 [-0.04, 1.10]	0.80 [0.70, 0.90]	1.15 [0.95, 1.42]	0.72 [0.56, 0.84]	0	7
35bdm	1.72 [0.66, 2.34]	1.44 [0.62, 2.26]	-1.01 [-2.18, 0.13]	0.92 [0.46, 1.00]	1.45 [0.73, 2.15]	0.80 [0.00, 1.00]	0	0
ryzue	1.77 [1.42, 2.12]	1.50 [1.17, 1.84]	1.30 [0.86, 1.72]	0.91 [0.86, 0.95]	1.23 [1.06, 1.41]	0.82 [0.71, 0.91]	0	0
2ii2g	1.80 [1.31, 2.24]	1.39 [1.01, 1.82]	-0.74 [-1.29, -0.15]	0.79 [0.65, 0.89]	1.15 [0.96, 1.37]	0.68 [0.59, 0.82]	0	2
mpwiy	1.82 [1.39, 2.23]	1.48 [1.14, 1.88]	0.10 [-0.54, 0.73]	0.82 [0.70, 0.91]	1.29 [1.12, 1.51]	0.66 [0.49, 0.80]	0	5
5byn6	1.89 [1.50, 2.27]	1.59 [1.24, 1.97]	1.32 [0.84, 1.80]	0.91 [0.85, 0.95]	1.28 [1.10, 1.48]	0.83 [0.72, 0.92]	0	0
y75vj	1.90 [1.50, 2.26]	1.58 [1.21, 1.97]	1.04 [0.46, 1.60]	0.89 [0.79, 0.95]	1.34 [1.16, 1.53]	0.75 [0.57, 0.88]	1	0
w4iyd	1.93 [1.53, 2.28]	1.58 [1.20, 1.98]	1.26 [0.72, 1.76]	0.85 [0.74, 0.92]	1.21 [1.00, 1.40]	0.73 [0.57, 0.85]	0	1
np6b4	1.94 [1.21, 2.71]	1.44 [1.04, 1.94]	-0.47 [-1.08, 0.24]	0.71 [0.60, 0.87]	1.08 [0.81, 1.43]	0.75 [0.62, 0.86]	0	8
nb004	2.01 [1.38, 2.63]	1.57 [1.16, 2.04]	0.56 [-0.10, 1.27]	0.82 [0.72, 0.90]	1.35 [1.15, 1.60]	0.71 [0.54, 0.84]	0	5
nb003	2.01 [1.39, 2.64]	1.58 [1.18, 2.04]	0.52 [-0.14, 1.22]	0.82 [0.73, 0.91]	1.36 [1.16, 1.61]	0.71 [0.54, 0.84]	0	5
yc70m	2.03 [1.73, 2.33]	1.80 [1.48, 2.13]	-0.41 [-1.09, 0.31]	0.47 [0.28, 0.64]	0.56 [0.35, 0.83]	0.53 [0.35, 0.68]	0	27
hytjn	2.16 [1.24, 3.06]	1.39 [0.86, 2.04]	0.71 [0.03, 1.48]	0.45 [0.13, 0.78]	0.62 [0.26, 1.00]	0.47 [0.16, 0.73]	1	27
f0gew	2.18 [1.38, 2.95]	1.58 [1.09, 2.16]	-0.73 [-1.42, 0.04]	0.77 [0.67, 0.89]	1.29 [1.01, 1.63]	0.76 [0.63, 0.86]	0	0
q3pfp	2.19 [1.33, 3.09]	1.51 [0.99, 2.13]	0.59 [-0.10, 1.37]	0.44 [0.13, 0.77]	0.66 [0.27, 1.07]	0.50 [0.20, 0.75]	1	22
ds62k	2.22 [1.62, 2.81]	1.78 [1.34, 2.27]	0.78 [0.06, 1.52]	0.82 [0.70, 0.90]	1.41 [1.20, 1.63]	0.72 [0.55, 0.85]	0	4
xikp8	2.35 [1.94, 2.73]	2.06 [1.66, 2.47]	0.77 [-0.02, 1.58]	0.89 [0.80, 0.95]	1.59 [1.40, 1.81]	0.76 [0.59, 0.89]	1	0
nb005	2.38 [1.79, 2.95]	1.91 [1.44, 2.43]	0.31 [-0.49, 1.15]	0.84 [0.74, 0.91]	1.56 [1.34, 1.82]	0.71 [0.54, 0.83]	0	0
5nm4j	2.45 [1.42, 3.34]	1.58 [0.94, 2.34]	0.05 [-0.80, 1.07]	0.19 [0.00, 0.70]	0.40 [-0.06, 0.81]	0.34 [-0.04, 0.67]	4	1
ad5pu	2.54 [1.68, 3.30]	1.83 [1.24, 2.49]	-0.65 [-1.48, 0.25]	0.76 [0.64, 0.88]	1.43 [1.12, 1.78]	0.77 [0.63, 0.88]	0	0
pwn3m	2.60 [1.45, 3.53]	1.54 [0.83, 2.37]	0.79 [-0.06, 1.77]	0.21 [0.00, 0.63]	0.37 [0.01, 0.78]	0.34 [0.04, 0.63]	1	3
nb006	2.98 [2.37, 3.56]	2.53 [2.00, 3.10]	0.42 [-0.60, 1.47]	0.84 [0.74, 0.92]	1.78 [1.55, 2.06]	0.71 [0.54, 0.84]	0	0
0hxtm	3.26 [1.81, 4.39]	1.92 [1.03, 2.98]	1.38 [0.37, 2.56]	0.08 [0.00, 0.48]	0.28 [-0.17, 0.83]	0.29 [-0.04, 0.61]	3	7

Table S3. Evaluation statistics calculated for all microscopic pK_a prediction submissions based on Hungarian match for 8 molecules with NMR data. Methods are represented via their SAMPL6 submission IDs which can be cross-referenced with Table 1 for method details. There are eight 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), unmatched experimental pK_as (number of missing pK_a predictions) and unmatched predicted pK_as (number of extra pK_a predictions between 2 and 12). 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	Unmatched exp. pK _a s	Unmatched pred. pK _a s [2,12]
<i>nb011</i>	0.47 [0.30, 0.64]	0.33 [0.22, 0.46]	-0.02 [-0.18, 0.14]	0.97 [0.94, 0.99]	1.01 [0.97, 1.06]	0.90 [0.78, 0.96]	0	36
<i>hdlyq</i>	0.62 [0.47, 0.76]	0.47 [0.33, 0.62]	0.13 [-0.09, 0.34]	0.95 [0.92, 0.97]	0.34 [0.92, 1.09]	0.87 [0.79, 0.93]	0	16
<i>epvmk</i>	0.63 [0.43, 0.81]	0.47 [0.32, 0.63]	-0.02 [-0.25, 0.21]	0.95 [0.89, 0.98]	0.21 [0.91, 1.04]	0.81 [0.68, 0.91]	0	37
<i>xnoe0</i>	0.65 [0.47, 0.82]	0.50 [0.36, 0.66]	-0.1 [-0.32, 0.13]	0.95 [0.89, 0.98]	0.13 [0.92, 1.05]	0.82 [0.69, 0.91]	0	36
<i>gdqeg</i>	0.65 [0.41, 0.89]	0.43 [0.27, 0.62]	0.11 [-0.10, 0.35]	0.94 [0.88, 0.98]	0.35 [0.87, 1.02]	0.83 [0.67, 0.95]	0	53
<i>4o0ia</i>	0.66 [0.44, 0.86]	0.47 [0.31, 0.64]	0.00 [-0.22, 0.24]	0.94 [0.88, 0.98]	0.24 [0.87, 1.05]	0.85 [0.73, 0.94]	0	35
<i>nb008</i>	0.76 [0.48, 1.02]	0.52 [0.34, 0.73]	-0.08 [-0.37, 0.17]	0.93 [0.85, 0.98]	0.17 [0.79, 0.93]	0.84 [0.73, 0.92]	0	35
<i>ccpmw</i>	0.79 [0.62, 0.94]	0.62 [0.46, 0.80]	-0.17 [-0.44, 0.11]	0.92 [0.86, 0.96]	0.11 [0.82, 1.05]	0.80 [0.67, 0.89]	0	7
<i>0xi4b</i>	0.84 [0.58, 1.07]	0.61 [0.42, 0.83]	0.22 [-0.07, 0.51]	0.92 [0.84, 0.97]	0.51 [0.91, 1.09]	0.81 [0.65, 0.92]	0	32
<i>cwyk</i>	0.86 [0.60, 1.10]	0.62 [0.42, 0.84]	0.13 [-0.16, 0.44]	0.90 [0.82, 0.96]	0.44 [0.86, 1.08]	0.81 [0.64, 0.92]	0	35
<i>ftc8w</i>	0.86 [0.51, 1.17]	0.59 [0.39, 0.83]	0.10 [-0.19, 0.41]	0.90 [0.77, 0.97]	0.41 [0.84, 0.98]	0.75 [0.57, 0.88]	0	35
<i>nxaaw</i>	0.89 [0.56, 1.25]	0.61 [0.41, 0.87]	-0.02 [-0.35, 0.28]	0.89 [0.75, 0.97]	0.28 [0.85, 1.00]	0.79 [0.63, 0.91]	0	29
<i>nb016</i>	0.95 [0.71, 1.18]	0.77 [0.57, 0.98]	-0.23 [-0.56, 0.12]	0.89 [0.83, 0.95]	0.12 [0.82, 1.07]	0.75 [0.62, 0.85]	0	3
<i>kxzt</i>	0.96 [0.56, 1.33]	0.64 [0.41, 0.92]	0.00 [-0.32, 0.36]	0.90 [0.76, 0.97]	0.36 [0.96, 1.13]	0.79 [0.63, 0.91]	0	37
<i>eyetm</i>	0.98 [0.69, 1.27]	0.72 [0.50, 0.97]	-0.32 [-0.65, 0.00]	0.91 [0.86, 0.96]	0.00 [0.94, 1.22]	0.78 [0.64, 0.88]	0	7
<i>cm2yq</i>	0.99 [0.44, 1.54]	0.56 [0.31, 0.90]	0.10 [-0.21, 0.50]	0.91 [0.83, 0.98]	0.50 [0.96, 1.25]	0.89 [0.80, 0.96]	0	36
<i>2umai</i>	1.00 [0.46, 1.54]	0.57 [0.33, 0.91]	0.07 [-0.25, 0.46]	0.91 [0.82, 0.98]	0.46 [0.96, 1.26]	0.87 [0.76, 0.95]	0	36
<i>ko8yx</i>	1.01 [0.76, 1.25]	0.78 [0.56, 1.01]	0.35 [0.01, 0.67]	0.91 [0.82, 0.96]	0.67 [0.96, 1.19]	0.78 [0.64, 0.89]	0	26
<i>wuuvc</i>	1.02 [0.51, 1.53]	0.62 [0.38, 0.93]	0.19 [-0.13, 0.58]	0.88 [0.80, 0.96]	0.58 [0.85, 1.19]	0.90 [0.81, 0.96]	0	36
<i>ktpj5</i>	1.02 [0.51, 1.56]	0.61 [0.37, 0.95]	0.17 [-0.16, 0.57]	0.88 [0.80, 0.96]	0.57 [0.87, 1.22]	0.89 [0.80, 0.96]	0	36
<i>z7fhp</i>	1.02 [0.49, 1.55]	0.61 [0.36, 0.94]	0.08 [-0.24, 0.48]	0.90 [0.82, 0.97]	0.48 [0.97, 1.26]	0.88 [0.80, 0.95]	0	28
<i>arcko</i>	1.04 [0.73, 1.32]	0.77 [0.53, 1.02]	0.37 [0.05, 0.72]	0.89 [0.80, 0.94]	0.72 [0.90, 1.14]	0.78 [0.62, 0.90]	0	24
<i>y4wws</i>	1.04 [0.70, 1.33]	0.74 [0.49, 1.00]	-0.31 [-0.66, 0.05]	0.91 [0.85, 0.96]	0.05 [1.02, 1.26]	0.79 [0.68, 0.88]	0	30
<i>wcvnu</i>	1.11 [0.80, 1.39]	0.84 [0.59, 1.11]	0.28 [-0.10, 0.66]	0.89 [0.77, 0.95]	0.66 [0.98, 1.22]	0.73 [0.54, 0.88]	1	27
<i>8toyp</i>	1.13 [0.61, 1.65]	0.70 [0.42, 1.05]	0.13 [-0.25, 0.56]	0.88 [0.81, 0.96]	0.56 [0.98, 1.29]	0.83 [0.72, 0.92]	0	27
<i>qsicn</i>	1.17 [0.30, 1.65]	0.88 [0.23, 1.54]	-0.76 [-1.54, 0.01]	0.91 [0.46, 1.00]	0.01 [0.52, 1.59]	0.80 [0.00, 1.00]	0	2
<i>wexjs</i>	1.30 [0.95, 1.62]	0.98 [0.68, 1.29]	0.27 [-0.17, 0.74]	0.86 [0.74, 0.93]	0.74 [1.00, 1.29]	0.73 [0.55, 0.86]	0	25
<i>v8qph</i>	1.37 [0.92, 1.79]	0.98 [0.66, 1.34]	-0.15 [-0.64, 0.34]	0.84 [0.70, 0.93]	0.34 [0.97, 1.32]	0.70 [0.55, 0.82]	0	6
<i>w420e</i>	1.57 [1.18, 1.94]	1.23 [0.90, 1.58]	0.09 [-0.48, 0.62]	0.85 [0.76, 0.91]	0.62 [1.08, 1.46]	0.72 [0.60, 0.82]	0	19
<i>6tvf8</i>	1.88 [0.87, 2.85]	1.02 [0.54, 1.66]	0.45 [-0.14, 1.18]	0.51 [0.16, 0.87]	1.18 [0.26, 0.89]	0.61 [0.34, 0.82]	0	55
<i>0wfzo</i>	2.89 [1.73, 3.89]	1.88 [1.17, 2.68]	0.76 [-0.15, 1.77]	0.48 [0.21, 0.75]	1.77 [0.60, 1.37]	0.51 [0.30, 0.70]	0	4
<i>t8ewk</i>	3.30 [1.89, 4.39]	1.98 [1.06, 3.00]	1.32 [0.27, 2.49]	0.07 [0.00, 0.45]	2.49 [-0.17, 0.79]	0.28 [-0.03, 0.6]	0	6
<i>z3btx</i>	4.00 [2.30, 5.45]	2.49 [1.47, 3.65]	1.48 [0.26, 2.86]	0.29 [0.04, 0.60]	2.86 [0.31, 1.44]	0.43 [0.19, 0.63]	0	1
<i>758j8</i>	4.52 [2.64, 6.18]	2.95 [1.85, 4.25]	1.85 [0.48, 3.38]	0.24 [0.02, 0.58]	3.38 [0.20, 1.51]	0.34 [0.08, 0.57]	0	2
<i>hgn83</i>	6.38 [4.04, 8.47]	4.11 [2.52, 5.93]	2.13 [0.07, 4.28]	0.08 [0.00, 0.39]	4.28 [-0.18, 1.43]	0.32 [0.07, 0.56]	0	0

Table S4. Evaluation statistics calculated for all microscopic pK_a prediction submissions based on microstate pair match for 8 molecules with NMR data. Methods are represented via their SAMPL6 submission IDs which can be cross-referenced with Table 1 for method details. There are eight 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), unmatched experimental pK_a s (number of missing pK_a predictions) and unmatched predicted pK_a s (number of extra pK_a predictions between 2 and 12. 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	Unmatched exp. pK_a s	Unmatched pred. pK_a s [2,12]
nb016	0.52 [0.25, 0.71]	0.43 [0.23, 0.65]	-0.09 [-0.45, 0.30]	0.92 [0.05, 0.99]	0.99 [0.14, 1.16]	0.62 [-0.14, 1.00]	0	3
hdijq	0.68 [0.49, 0.83]	0.60 [0.39, 0.80]	0.38 [0.02, 0.70]	0.86 [0.47, 0.98]	0.91 [0.45, 1.26]	0.78 [0.4, 1.00]	0	16
nb011	0.72 [0.35, 1.07]	0.54 [0.28, 0.86]	0.45 [0.14, 0.83]	0.86 [0.18, 0.98]	0.93 [0.50, 1.21]	0.64 [0.26, 0.95]	0	36
ftc8w	0.75 [0.52, 0.96]	0.68 [0.50, 0.89]	-0.31 [-0.68, 0.16]	0.87 [0.02, 0.99]	1.12 [-0.11, 1.39]	0.56 [-0.10, 1.00]	0	35
6vf8	0.76 [0.55, 0.95]	0.68 [0.46, 0.90]	-0.63 [-0.89, -0.35]	0.92 [0.78, 0.99]	0.94 [0.69, 1.41]	0.87 [0.6, 1.00]	0	55
t8ewk	0.96 [0.65, 1.19]	0.81 [0.46, 1.13]	-0.77 [-1.12, -0.38]	0.80 [0.53, 0.96]	0.96 [0.76, 2.26]	0.78 [0.31, 1.00]	1	7
v8qph	0.99 [0.40, 1.52]	0.67 [0.29, 1.17]	-0.09 [-0.75, 0.45]	0.68 [0.11, 0.97]	0.96 [-1.26, 1.16]	0.38 [-0.3, 1.00]	0	6
ccpmw	1.07 [0.78, 1.27]	0.95 [0.60, 1.25]	-0.83 [-1.25, -0.37]	0.74 [0.43, 0.99]	0.95 [0.70, 2.32]	0.89 [0.52, 1.00]	1	8
Oxi4b	1.15 [0.75, 1.50]	0.98 [0.63, 1.36]	-0.30 [-0.94, 0.44]	0.77 [0.02, 0.98]	1.26 [0.09, 2.10]	0.51 [-0.14, 1.00]	0	33
cywyk	1.17 [0.88, 1.41]	1.06 [0.74, 1.35]	-0.47 [-1.09, 0.24]	0.73 [0.02, 0.98]	1.15 [-0.04, 2.00]	0.56 [-0.08, 1.00]	0	36
eyetm	1.17 [0.77, 1.52]	1.00 [0.61, 1.41]	-0.89 [-1.38, -0.38]	0.67 [0.30, 0.94]	0.93 [0.65, 2.59]	0.72 [0.29, 1.00]	1	8
nb008	1.26 [0.74, 1.71]	1.09 [0.63, 1.57]	0.47 [-0.40, 1.32]	0.79 [0.01, 0.99]	1.21 [-0.59, 1.85]	0.52 [-0.2, 1.00]	0	38
y4wws	1.41 [0.95, 1.80]	1.22 [0.78, 1.66]	-0.71 [-1.44, 0.06]	0.87 [0.05, 0.98]	1.55 [0.41, 2.02]	0.56 [-0.11, 1.00]	0	31
ktpj5	1.46 [0.83, 2.10]	1.15 [0.67, 1.77]	0.94 [0.29, 1.68]	0.77 [0.01, 0.98]	1.28 [-0.26, 1.60]	0.42 [-0.27, 0.95]	0	37
wuuvc	1.47 [0.84, 2.09]	1.18 [0.70, 1.77]	0.99 [0.36, 1.68]	0.78 [0.01, 0.98]	1.27 [-0.24, 1.58]	0.47 [-0.20, 1.00]	0	37
xnoe0	1.54 [1.09, 2.00]	1.39 [1.02, 1.83]	0.91 [0.11, 1.64]	0.82 [0.01, 0.98]	1.47 [-0.30, 1.79]	0.42 [-0.27, 0.95]	0	37
qsicn	1.58 [1.44, 1.70]	1.57 [1.44, 1.70]	-1.57 [-1.7, -1.44]	1.00 [0.00, 1.00]	1.06		0	2
epvmk	1.66 [1.20, 2.15]	1.50 [1.07, 1.96]	1.12 [0.31, 1.82]	0.82 [0.02, 0.98]	1.47 [-0.21, 1.8]	0.42 [-0.25, 0.95]	0	37
400ia	1.73 [1.33, 2.17]	1.62 [1.29, 2.02]	1.31 [0.53, 1.93]	0.87 [0.03, 0.99]	1.50 [0.07, 1.84]	0.56 [-0.07, 1.00]	0	36
ko8yx	1.75 [1.08, 2.45]	1.44 [0.87, 2.12]	1.38 [0.74, 2.10]	0.97 [0.88, 1.00]	1.66 [1.46, 2.28]	0.91 [0.69, 1.00]	0	27
Zumai	1.76 [1.21, 2.35]	1.54 [1.04, 2.11]	1.31 [0.55, 2.03]	0.82 [0.02, 0.98]	1.43 [-0.02, 1.77]	0.47 [-0.17, 0.95]	0	37
cm2yq	1.77 [1.22, 2.36]	1.55 [1.06, 2.12]	1.33 [0.57, 2.04]	0.82 [0.02, 0.98]	1.43 [-0.02, 1.76]	0.47 [-0.17, 0.95]	0	37
nxaaw	1.80 [0.84, 2.80]	1.34 [0.80, 2.18]	0.16 [-0.77, 1.41]	0.59 [0.02, 0.97]	1.37 [-0.08, 2.92]	0.6 [-0.05, 1.00]	0	30
wcvnu	1.90 [1.14, 2.64]	1.57 [0.97, 2.27]	1.44 [0.70, 2.24]	0.97 [0.91, 1.00]	1.78 [1.58, 2.48]	0.91 [0.69, 1.00]	0	27
kxzt	2.00 [1.13, 2.73]	1.64 [1.00, 2.39]	1.64 [1.00, 2.39]	0.83 [0.01, 0.98]	1.42 [-0.21, 1.99]	0.56 [-0.10, 1.00]	0	38
wexjs	2.05 [1.18, 2.93]	1.66 [1.01, 2.47]	1.48 [0.63, 2.39]	0.96 [0.55, 0.99]	1.87 [1.54, 2.29]	0.73 [0.20, 1.00]	0	26
z7fhp	2.14 [1.38, 2.87]	1.80 [1.12, 2.58]	1.28 [0.18, 2.34]	0.78 [0.02, 0.98]	1.71 [-0.41, 2.13]	0.42 [-0.25, 0.95]	0	30
gdqeg	2.38 [1.97, 2.71]	2.25 [1.74, 2.68]	-1.61 [-2.46, -0.37]	0.10 [0.00, 0.98]	0.31 [-0.60, 1.63]	0.29 [-0.45, 1.00]	0	53
8toyp	2.63 [1.89, 3.29]	2.34 [1.59, 3.07]	1.78 [0.47, 2.89]	0.82 [0.02, 0.98]	1.94 [-0.06, 2.39]	0.47 [-0.17, 0.95]	0	29
w4z0e	2.63 [1.81, 3.53]	2.34 [1.67, 3.18]	1.74 [0.46, 2.92]	0.98 [0.55, 1.00]	2.28 [1.52, 2.41]	0.73 [0.20, 1.00]	0	20
arcko	2.64 [1.23, 3.78]	2.08 [1.10, 3.24]	1.71 [0.44, 3.10]	0.57 [0.04, 0.95]	1.42 [0.56, 2.93]	0.56 [-0.06, 1.00]	0	28
0wfzo	18.72 [11.21, 25.03]	15.80 [9.9, 22.35]	15.09 [8.28, 22.12]	0.09 [0.01, 0.73]	2.35 [-10.18, 8.12]	0.02 [-0.65, 0.66]	0	12
z3btv	22.60 [15.03, 29.00]	19.70 [12.97, 26.69]	19.70 [12.97, 26.69]	0.09 [0.01, 0.72]	2.35 [-10.00, 8.28]	0.02 [-0.66, 0.66]	0	7
758j8	23.76 [16.33, 30.24]	21.00 [14.26, 28.00]	21.00 [14.26, 28.00]	0.09 [0.01, 0.71]	2.35 [-10.34, 8.12]	0.02 [-0.65, 0.65]	0	8
hgn83	27.91 [20.54, 34.52]	25.60 [18.9, 32.64]	25.60 [18.9, 32.64]	0.09 [0.01, 0.72]	2.35 [-10.21, 8.00]	0.02 [-0.65, 0.65]	0	5