Accuracy of macroscopic and microscopic pK_a predictions of small molecules evaluated by the SAMPL6 Blind Challenge

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

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

***For correspondence:**
mehtap.isik@choderalab.org (MI)

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 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 submissons, 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_as 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-mechanical and empirical approaches. These predictions included less than 8 extra/missing macroscopic pK_as for the set of 24 molecules. A large number of submissions had RMSE spanning 1–3 pK_a units. Molecules with sulfur-containing heterocycles, iodo, and bromo groups suffered from less accurate pK_a predictions on average considering all methods evaluated. For a subset of molecules, the available NMR-based dominant microstate sequence data was utilized to elucidate dominant tautomer prediction errors of microscopic pK_a predictions which was prominent for charged tautomers. The SAMPL6 pK_a Challenge demonstrated the need for improving pK_a prediction methods for drug-like molecules, especially for challenging moieties and multiprotic molecules. The inaccuracy of pK_a predictions as observed in this challenge can be detrimental to the performance of protein-ligand binding affinity predictions due to errors in

predicted dominant charge and tautomeric states and errors in the calculation of free energy correction for multiple protonation states of the ligand.

0.1 Keywords

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

0.2 Abbreviations

SAMPL Statistical Assessment of the Modeling of Proteins and Ligands
pK_a $-\log_{10}$ acid dissociation equilibrium constant
SEM Standard error of the mean
RMSE Root mean squared error
MAE Mean absolute error
τ Kendall's rank correlation coefficient (Tau)
R^2 Coefficient of determination (R-Squared)

1 Introduction

The acid dissociation constant (K_a) describes the protonation state equilibrium of a molecule given pH and pK_a is the negative logarithmic form of K_a. Predicting pK_a is a prerequisite for predicting many other properties of small molecules such as protein-ligand binding affinity, distribution coefficient (log *D*), membrane permeability, and solubility. Computer-aided drug design efforts include assessing the properties of virtual molecules to guide synthesis and prioritization decisions. In such cases, an experimental pK_a measurement is not possible. Therefore, accurate computational pK_a prediction methods are required.

Ionizable sites are found often in drug molecules and influence their pharmaceutical properties including target affinity, ADME/Tox, and formulation properties [1]. Drug molecules with titratable groups can exist in many different charge and protonation states based on the pH of the environment. Given that experimental data of protonation states and pK_a are often not available, we rely on predicted pK_a values to determine in which charge and protonation states the molecules exist and what are the relative populations of these states, so that we can assign the appropriate protonation state(s) in fixed-state calculations, or the appropriate solvent state weights/protonation penalty to calculations considering multiple states. The pH of the human gut ranges between 1-8 and 74% of approved drugs can change ionization states within this physiological pH range [2]. Because of this pK_a values of drug molecules provides essential information about their physicochemical and pharmaceutical properties. A wide distribution of acidic and basic pK_a values, ranging from 0 to 12, have been observed in approved drugs [1, 2].

Drug-like molecules present difficulties for pK_a prediction compared to simple monoprotic molecules. Drug-like molecules are frequently multiprotic, have large conjugated systems, heterocycles, tautomerization. Besides, the larger molecules with conformational flexibility can have intramolecular hydrogen bonding which shifts pK_a values. These shifts could be real or modeling artifacts due to collapsed conformations caused by deficiencies in solvation models. Yet predicting pK_as of drug-like molecules accurately is a prerequisite for computational drug discovery and design.

Small molecule pK_a predictions influence computational protein-ligand binding affinities in multiple ways. Errors in pK_a predictions can cause modeling the wrong charge, protonation, and tautomerization states which affect hydrogen bonding opportunities and charge distribution of the ligand. The prediction of the dominant protonation state and relative population of minor states in the aqueous medium is dictated by the pK_a values. The relative free energy of different protonation states in the aqueous state is a function of pK_a and pH, it contributes to the overall protein-ligand affinity in the form of a free energy penalty of reaching higher energy protonation states [3].

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

$$HA \rightleftharpoons A^- + H^+ \quad BH^+ \rightleftharpoons B + H^+ \tag{1}$$

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

$$pK_a = -\log_{10} K_a \tag{3}$$

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

The definition of pK_a diverges into two for multiprotic molecules: macroscopic pK_a and microscopic pK_a [4–6]. Macroscopic pK_a describes the equilibrium dissociation constant between different charged states of the molecule. Each charge state can be composed of multiple tautomers. Macroscopic pK_a is about the deprotonation of the molecule, not a particular titratable group. Microscopic pK_a describes the acid dissociation equilibrium between individual tautomeric states of different charges. We refer to the collection of all tautomeric states of different macroscopic states (charge states) as microscopic states. Microscopic pK_a value defined between two microstates captures the deprotonation of a single titrable group with a fixed background protonation state of other titratable groups. In molecules with multiple titratable groups, the protonation state of one group can affect the proton dissociation propensity of another functional group, therefore the same titratable group may have different microscopic pK_a values based on the protonation state of the rest of the molecule.

Different experimental methods capture different definitions of pK_as as explained in more detail in this prior publication [7]. Most common pK_a measurement techniques such as potentiometric and spectrophotometric methods measure macroscopic pK_as while NMR measurements can determine microscopic pK_as and microstate populations. Therefore, it is important to pay attention to the source and definition of pK_a values to interpret their meaning correctly. Computational methods can predict both microscopic and macroscopic pK_as. While microscopic pK_a predictions are more informative for determining relevant microstates/tautomers of a molecule and their relative free energies, computing predicted macroscopic pK_as is useful for direct comparison of methods to more common macroscopic experimental measurements. In this paper, we explore approaches to assess the performance of both macroscopic and microscopic pK_a predictions, taking advantage of available experimental data.

Microscopic pK_a predictions can be converted to macroscopic pK_a predictions either directly with the equation 5 [8] or through computing the macroscopic free energy of deprotonation between ionization states with charges N and N-1 via Boltzmann weighted sum of the relative free energy of microstates (G_i) as in equations 6 and 7 [9].

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

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

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

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 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 the ideal gas constant times the temperature.

1.1 Motivation for a blind pK_a challenge

SAMPL (Statistical Assessment of the Modeling of Proteins and Ligands) is a series of annual computational prediction challenges for the computational chemistry community. The goal of SAMPL is evaluate to the current performance of the models and to bring the attention of the quantitative biomolecular modeling field on major issues that limit the accuracy of protein-ligand binding models.

SAMPL Challenges that focus on different physical properties so far have assessed intermolecular binding models of various protein-ligand and host-guest systems, solvation models to predict hydration free energies and distribution coefficients. Potential benefits of these challenges are motivating improvement computational methods and revealing unexpected error contributors by focusing on interesting test systems. SAMPL Challenges have demonstrated the effects of force field accuracy, sampling issues, solvation modeling defects, and tautomer/protonation state predictions on protein-ligand binding predictions.

During the SAMPL5 log *D* Challenge, the performance of cyclohexane-water log *D* predictions were lower than expected and accuracy suffered when protonation states and tautomers were not taken into account [10, 11]. A common simplified approach to predicting the log *D* is to predict the log *P* of the neutral state, and use a pK_a prediction to correct for the free energy penalty of the neutral state. However, it does not matter how accurate the log *P* prediction is, if the protonation state is neglected, or the error in the predicted pK_a value is large. With the motivation of deconvoluting the different sources of error contributing to the large errors observed in the SAMPL5 log *D* Challenge, we organized separate pK_a and log *P* challenges in SAMPL6 [7, 12, 13]. For this iteration of the SAMPL challenge, we have taken one step back and isolated just the problem of predicting aqueous protonation states.

This is the first time a blind pK_a prediction challenge has been fielded as part of SAMPL. In this first iteration of the challenge, we aimed to assess the performance of current pK_a prediction methods for drug-like molecules, investigate potential causes of inaccurate pK_a estimates, and determine how much current level of expected accuracy might impact protein binding affinity predictions. In binding free energy predictions, any error in predicting the free energy of accessing a minor aqueous protonation state of ligand that contributes to the complex formation will directly add to the error in the predicted binding free energy. Similarly for log *D* predictions, inaccurate prediction aqueous protonation state that contributes to partitioning between phases or prediction of relative free energy of these states will be detrimental to the accuracy of transfer free energy predictions.

1.2 Approaches to predict small molecule pK_as

There is a large variety pK_a prediction methods developed until this day for aqueous pK_a prediction of small molecules. Broadly we can divide pK_a predictions as knowledge-based empirical methods and physical methods. Empirical method categories include Database Lookup (DL), Linear Free Energy Relationship (LFER), Quantitative Structure-Property Relationship (QSPR), and Machine Learning approaches. DL methods rely on the principle that structurally similar compounds have similar pK_a values and utilize an experimental database of complete structures or fragments. The pK_a values of the most similar database entries are reported as the predicted pK_a of the query molecule. In the QSPR approach, the pK_a values are predicted as a function of various quantitative molecular descriptors and the parameters of the function are trained on experimental datasets. A function in the form of multiple linear regression is common, although more complex forms can also be used such as the artificial neural networks in ML methods. The LFER approach is the oldest pK_a prediction strategy. They use Hammett-Taft type equations to predict pK_a based on classification of the molecule to a parent class (associated with a base pK_a value) and two parameters that describe how the base pK_a value must be modified based on its substituents. Physical modeling of pK_a predictions require Quantum Mechanics (QM) models. Molecular mechanics based pK_a prediction methods have not been found as a feasible approach for small molecule pK_a predictions as deprotonation is a covalent bond breaking event that can only be captured by QM. QM methods are often utilized together with linear empirical corrections (LEC) that are designed to rescale and unbias QM predictions for better accuracy.

2 Methods

2.1 Design and logistics of the SAMPL6 pK_a Challenge

The SAMPL6 pK_a Challenge was conducted as a blind prediction challenge focus on predicting aqueous pK_a value of 24 small molecules that resemble fragments of kinase inhibitors and contain heterocycles that are frequentyly found in FDA-approved kinase inhibitors. The compound selection process was described in depth in the prior publication reporting SAMPL6 pK_a Challenge experimental data collection [7]. The distribution of molecular weights, experimental pK_a values, number of rotatable bonds, and heteroatom to 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 ranged between 2-12 and roughly uniform. 2D representations of all compounds were provided in Fig. 5. Drug-like molecules are often larger and more complex than the ones used in this study, however, aimed for the

The dataset composition and details of the pK_a measurement technique, except the identity of the small molecules, were announced about a month before the challenge start time. 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 [7]. The instructions for participation and the identity of the challenge molecules were released at the challenge start date (October 25,

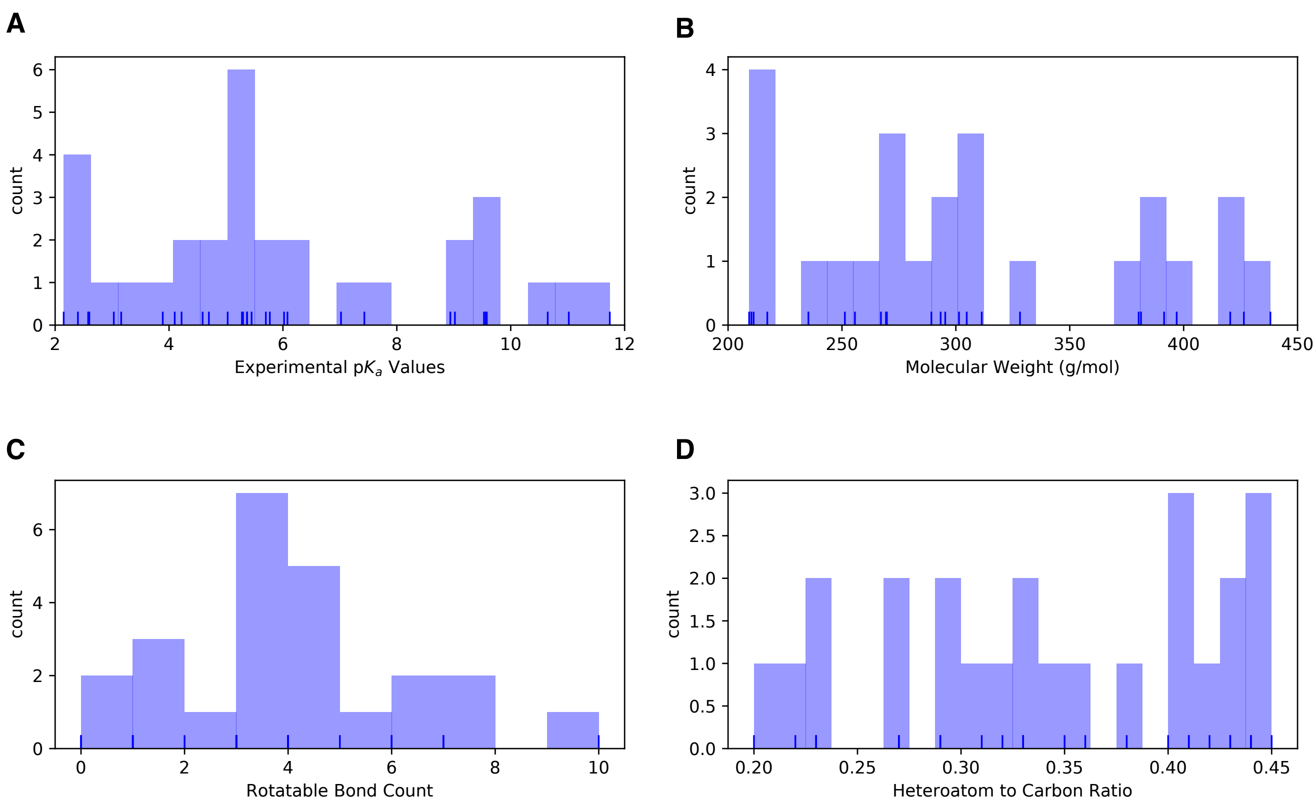


Figure 1. Distribution of molecular properties of 24 compounds in SAMPL6 pK_a Challenge. **A** Histogram of spectrophotometric pK_a measurements collected with Sirius T3 [7]. The overlayed carpet plot indicates the actual values. Five compounds have multiple measured pK_as in the range of 2-12. **B** Histogram of molecular weights of compounds in SAMPL6 set. Molecular weights were calculated by neglecting counter ions. **C** Histogram of the number of non-terminal rotatable bonds in each molecule. **D** The histogram of the ratio of heteroatom (non-carbon heavy atom) count to the number of carbon atoms.

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. The participants reported their results and their own evaluations in the special issue of the Journal of Computer-Aided Molecular Design [14].

In this first iteration of pK_a prediction challenge we were not sure what was the best way to capture all necessary information related to pK_a predictions. Our aim 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 microstate ID of protonated state, microstate ID of deprotonated state, microscopic pK_a, and microscopic pK_a SEM. The reason and method of microstate enumeration are discussed further in Section 2.2 "Enumeration of Microstates". The SEM captures the statistical uncertainty of the predicted 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 consecutive columns reporting 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, 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 have 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 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 with multiple methods as long as they create separate submissions files. Anonymous participation was allowed in the challenge, however, all participants opted to make their submissions public. All 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. There were many instances that 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 participant. Submission ID for both macroscopic (type III) and microscopic (type I) pK_a predictions of each method (when exists) 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 predictions to be reported together with the pair of deprotonated and protonated microstates that describes the transition. 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 have decided that the safest route was pre-enumerating all possible microstates of challenge compounds, assigning the microstates IDs to each in the form of SM##_micro###, and require participants to report microstate pairs using the provided microstates IDs.

We enumerated an initial list of microstates with Epik [15] and OpenEye QUACPAC [16] and took the union of results. Microstates with Epik were generated using Schrodinger Suite v2016-4, and 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, and carbonyl hybridization

False. Then we created a union of all enumerated states written as canonical isomeric SMILES. 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 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 documents. Because resonance and geometric isomerism should be ignored when matching predicted structures microstate IDs (except SM20 which should be modeled as E-isomer), we provided microstate ID tables with canonical SMILES and 2D-depictions.

Despite pooling together enumerated charge states and tautomers with Epik and OpenEye QUACPAC to our surprise the microstate lists were still incomplete. A better algorithm that can enumerate all possible microstates would be very beneficial. In SAMPL6 Challenge participants came up with new microstates that were not present in the initial list that we provided. Based on participant requests 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 the challenge participants.

A working pK_a microstate definition for this challenge was provided in challenge instructions for clarity. Physically meaningful microscopic pK_as 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 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 related to the same microstate. Pairs of resonance structures and geometric isomers (cis/trans, stereo) won't be considered as different microstates, as long as there is no change in the number of hydrogens bound to each heavy atom in these structures. Since we wanted participants to report only microscopic pK_as 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 have reported this dependency in their method descriptions explicitly, therefore we can not identify which submissions have used the enumerated microstate lists as input and which ones have followed the instructions.

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 only 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, while experimental data did not provide any information on which group(s) are being titrated, microscopic pK_a values, the identity of associated macrostates (which charge) or microstates (which tautomers). In this comparative performance evaluation of we let the experimental data lead the challenge analysis towards various evaluation routes. 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.

2.3.1 Matching algorithms for pairing predicted and experimental pK_as

Macroscopic pK_a predictions can be calculated from microscopic pK_as for direct comparison to experimental macroscopic pK_a values, although there is still a remaining issue. How to match predicted macroscopic pK_as to experimental macroscopic pK_as when there could multiple numbers of each reported for each molecule? Experimental data, in this case, 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.

For evaluating predictions taking the experimental data as reference Fraczkiewicz et al. delineated recommendations for fair comparative analysis of computational pK_a predictions [17]. In the absence of any experimental information that would aid the match, experimental and computational pK_as should be matched preserving the order of pK_a values and minimizing the sum of absolute errors.

We picked Hungarian matching algorithm [18, 19] to assign experimental and predicted macroscopic pK_as with squared error cost function as suggested by Kiril Lanevskij. The algorithm is available in SciPy package (*scipy.optimize.linear_sum_assignment*) [20]. This matching algorithm provides optimum global assignment that minimizes the linear sum of squared errors of all pairwise matches. The reason to select squared error cost function instead of the absolute error cost function is to avoid misordered matches, For instance, for a molecule with experimental pK_a values of 4 and 6, and predicted pK_as 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 9, 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 experimental data of the challenge did not contain information about what charge states the transitions were happening between and also because we have not collected the pair of macrostates associated with each pK_a predictions in submissions. There is no perfect solution to the numerical pK_a assignment problem, but we tried to determine the fairest way to penalize predictions based on their numerical deviation from the experimental values.

For the analysis of microscopic pK_a predictions we adopted a different matching approach. Only for the 8 molecules, we utilized the dominant microstate sequence inferred from NMR experiments to match computational predictions and experimental pK_as. We will refer to this assignment method as microstate matching, where experimental pK_a value is matched to the computational microscopic pK_a value which was reported for the dominant microstate pair observed for each transition. We have compared the results of Hungarian matching and microstate matching.

Inevitably the choice of matching algorithms to assign experimental and predicted values has an impact on the calculation of performance statistics. We believe the Hungarian algorithm for numerical matching and microstate-based were the best choices, providing the most unbiased matching without introducing assumptions outside of the experimental data.

2.3.2 Statistical metrics for submission performance

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

There are six error metrics reported for the numerical error of the pK_a values: the root-mean-squared error (RMSE), mean absolute error (MAE), mean error (ME), coefficient of determination (R^2), linear regression slope (m), and Kendall's Rank Correlation Coefficient (τ). Uncertainty in each performance statistic was calculated as 95% confidence intervals estimated by bootstrapping over predictions with 10000 bootstrap samples. Calculated errors statistics of all methods can be found in Table S2 for macroscopic pK_a predictions and Tables S4 and S4 for microscopic pK_a predictions.

In addition to the numerical error aspect of the pK_a values, we have also evaluated predictions in terms of their ability to capture the correct macrostates (ionization states) and microstates (tautomers of each ionization state) to the extent possible from the available experimental data. For macroscopic pK_as experiments did not provide any evidence of the identity of the ionization states. However, the number of ionization states indicates the number of macroscopic pK_as that exists between the experimental range of 2.0-12.0. For instance, SM14 has two experimental pK_as and therefore 3 different charge states were observed between the pH range of 2.0-12.0. If a prediction reported 4 macroscopic pK_as, it is clear that this method predicted an extra ionization state. With this perspective we reported the number of unmatched experimental pK_as (the number of missing pK_a predictions, i.e. missing ionization states) and the number of unmatched predicted pK_as (the number of extra pK_a predictions, i.e. extra ionization states) after Hungarian matching. The later count was restricted to only predictions with pK_a values between 2 and 12 because that was the range of the experimental method. Errors in extra or missing pK_a prediction errors highlight failure to predict the correct number of ionization states within a pH range.

For the evaluation of microscopic pK_a predictions, taking advantage of the available dominant microstate sequence data for a subset of 8 compounds, we calculated the dominant microstate prediction accuracy. Dominant microstate prediction accuracy is the ratio of correct dominant tautomer predictions for each charge state divided by, calculated over all ionization states of each molecule. In order to extract the sequence of dominant microstates from the microscopic pK_a predictions sets, we calculated the relative free energy of microstates selecting a neutral tautomer and pH 0 as reference following the Equation 8. Calculation of relative free energy of microstates was explained in more detail in a previous publication [21].

Relative free energy of state with respect to reference state B at pH 0.0 (arbitrary pH value selected as reference) can be calculated as follows:

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

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

We created a shortlist of top-performing methods for macroscopic and microscopic pK_a predictions. Top macroscopic pK_a predictions were selected based on the following criteria of consistent performance among different metrics: ranking in the top 10 consistently according to two error (RMSE, MAE) and two correlation metrics (R-Squared, and Kendall's Tau), and also having a combined count of less than 8 missing or extra macroscopic pK_as for the entire molecule set (a third of the number of compounds). These methods are presented in Table 2. A separate list of top-performing methods was constructed for microscopic pK_a with the following criteria: ranking in the top 10 methods when ranked by accuracy statistics (RMSE and MAE) and perfect dominant microstate prediction accuracy. These methods are presented in Table 3.

In addition to comparing the performance comparison of methods, we also wanted to compare pK_a prediction performance on the level of molecules to determine pK_as of which molecules in the challenge set were harder to predict considering all the methods in the challenge. For this purpose, we plotted prediction error distributions of each molecule considering all prediction methods. We also calculated MAE for each molecule over all prediction sets as well as for predictions from each method category separately.

2.4 Reference calculations

Including a null model as helpful in comparative performance analysis of predictive methods to establish what the performance statistics look like for a baseline method for the specific dataset. Null models or null predictions employ a simple prediction model which is not expected to be particularly successful, but it is useful for providing a simple point of comparison for more sophisticated methods. The expectation is for more sophisticated or costly prediction methods to outperform the predictions from a null model, otherwise the simpler null model would be preferable. In SAMPL6 pK_a Challenge there were two blind submissions that database lookup methods that were suitable to be considered as null predictions. These methods, with submission IDs *5nm4j* and *5nm4j* both used OpenEye pKa-Prospector database to find the most similar molecule to query molecule and report its pK_a as the predicted value. We acknowledge that database lookup methods with a rich experimental database present a quite challenging null model to beat, however, due to the accuracy level needed from pK_a predictions for computer-aided drug design we believe it is an appropriate performance baseline that physical and empirical pK_a prediction methods should strive to perform better than.

We have also included additional reference calculations in the comparative analysis to provide more perspective. The methods we chose to include as reference calculations were missing from the blind predictions sets although they are widely used methods by academia and industry. representing different methodological approaches: Schrodinger/Epik (*nb007, nb008, nb010*), Schrodinger/Jaguar (*nb011, nb013*), Chemaxon/Chemicalize (*nb015*), and Molecular Discovery/MoKa (*nb016, nb017*). Epik and Jaguar pK_a predictions were collected by Bas Rustenburg, Chemicalize predictions by Mehtap Isik, and MoKa predictions by Thomas Fox, after the challenge deadline avoiding any alterations to the respective standard procedures of the methods guided by the experimental data. Reference calculations were not formally blind, as experimental data of the challenge has been made publically available before their collection.

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

3 Results and Discussion

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

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

3.1 Analysis of macroscopic pK_a predictions

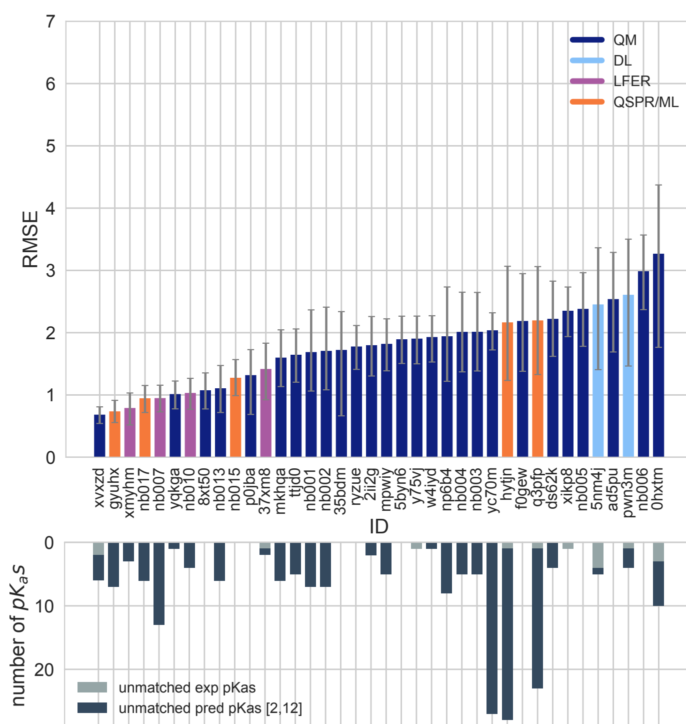


Figure 2. RMSE and unmatched pK_a counts vs. submission ID plots for macroscopic pK_a predictions based on Hungarian matching. Methods are indicated by submission IDs. RMSE is shown with error bars denoting 95% confidence intervals obtained by bootstrapping over challenge molecules. Submissions are colored 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_as (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

Table 1. Submission IDs, names, category, and type for all the pK_a prediction sets. Reference calculations are labeled as *nb###*. The method name column lists the names provided by each participant in the submission file. The "type" column indicates if 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 pK_a (Type I) Submission ID	Macroscopic pK_a (Type III) Submission ID	Submission Type	Ref.
DL	Substructure matches to experimental data in pKa OpenEye pKa Prospector Database v1.0		*5nm4j*	Null	[22]
DL	OpenEye pKa-Prospector 1.0.0.3 with Analog Search ion identification algorithm		*pwn3m*	Null	[22]
LFER	ACD/pKa GALAS (ACD/Percepta Kernel v1.6)	*v8qph*	*37xm8*	Blind	[23]
LFER	ACD/pKa Classic (ACD/Percepta Kernel, v1.6)		*xmyhm*	Blind	[24]
LFER	Epik Scan (Schrodinger v2017-4)		*nb007*	Reference	[15]
LFER	Epik Microscopic (Schrodinger v2017-4)	*nb008*	*nb010*	Reference	[15]
QSPR/ML	OpenEye Gaussian Process	*6tvf8*	*hytjn*	Blind	[11]
QSPR/ML	OpenEye Gaussian Process Resampled		*q3pfp*	Blind	[11]
QSPR/ML	S+pKa (ADMET Predictor v8.5, Simulations Plus)	*hdiyq*	*gyuhx*	Blind	[25]
QSPR/ML	Chemicalize v18.23 (ChemAxon MarvinSketch v18.23)		*nb015*	Reference	[26]
QSPR/ML	MoKa v3.1.3	*nb016*	*nb017*	Reference	[27, 28]
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	*ko8yx*	*ryzue*	Blind	[29]
QM	Direct scheme with single point correction: SMD/M06-2X//6-311++G(d,p)//M06-2X/6-31+G(d) for bases and SMD/M06-2X//6-311++G(d,p)//M06-2X/6-31G(d) for acids + thermal corrections	*w4z0e*	*xikp8*	Blind	[29]
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	*wcvnu*	*5byn6*	Blind	[29]
QM	Vertical scheme: thermodynamic cycle that uses only gas phase optimized structures to compute gas hase and solvation free energy. SMD/M06-2X/6-31+G(d) for bases and SMD/M06-2X/6-31G(d) for acids + Thermal corrections	*arcko*	*w4iyd*	Blind	[29]
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	*wexjs*	*y75vj*	Blind	[29]
QM + LEC	Jaguar (Schrodinger v2017-4)	*nb011*	*nb013*	Reference	[30]
QM + LEC	CPCM/B3LYP/6–311+G(d,p) and global fitting	*y4wws*	*35bdm*	Blind	[9]
QM + LEC	CPCM/B3LYP/6–311+G(d,p) and separate fitting for neutral to negative and for positive to neutral transformations	*qsicn*	*p0jba*	Blind	[9]
QM + LEC	EC-RISM/MP2/6-311+G(d,p)-P3NI-q-noThiols-2par	*kxztt*	*ds62k*	Blind	[31]
QM + LEC	EC-RISM/MP2/cc-pVTZ-P2-q-noThiols-2par	*ftc8w*	*2ii2g*	Blind	[31]
QM + LEC	EC-RISM/MP2/6-311+G(d,p)-P2-phi-all-2par	*ktpj5*	*nb001*	Blind*	[31]
QM + LEC	EC-RISM/MP2/6-311+G(d,p)-P2-phi-noThiols-2par	*wuuvc*	*nb002*	Blind*	[31]
QM + LEC	EC-RISM/MP2/6-311+G(d,p)-P3NI-phi-all-2par	*2umai*	*nb003*	Blind*	[31]
QM + LEC	EC-RISM/MP2/6-311+G(d,p)-P3NI-phi-noThiols-2par	*cm2yq*	*nb004*	Blind*	[31]
QM + LEC	EC-RISM/MP2/6-311+G(d,p)-P2-phi-all-1par	*z7fhp*	*nb005*	Blind*	[31]
QM + LEC	EC-RISM/MP2/6-311+G(d,p)-P3NI-phi-all-1par	*8toyp*	*nb006*	Blind*	[31]
QM + LEC	EC-RISM/MP2/cc-pVTZ-P2-phi-noThiols-2par	*epvmk*	*ttjd0*	Blind	[31]
QM + LEC	EC-RISM/MP2/cc-pVTZ-P2-phi-all-2par	*xnoe0*	*mkhqa*	Blind	[31]
QM + LEC	EC-RISM/MP2/cc-pVTZ-P3NI-phi-noThiols-2par	*4o0ia*	*mpwiy*	Blind	[31]
QM + LEC	EC-RISM/B3LYP/6-311+G(d,p)-P3NI-q-noThiols-2par	*nxaaw*	*ad5pu*	Blind	[31]
QM + LEC	EC-RISM/B3LYP/6-311+G(d,p)-P3NI-phi-noThiols-2par	*0xi4b*	*f0gew*	Blind	[31]
QM + LEC	EC-RISM/B3LYP/6-311+G(d,p)-P2-phi-noThiols-2par	*cywyk*	*np6b4*	Blind	[31]
QM + LEC	PCM/B3LYP/6-311+G(d,p)	*gdqeg*	*yc70m*	Blind	[31]
QM + LEC	COSMOtherm_FINE17 (COSMOtherm C30_1701, BP/TZVPD/FINE//BP/TZVP/COSMO)	*t8ewk*	*0hxtm*	Blind	[32, 33]
QM + LEC	DSD-BLYP-D3(BJ)/def2-TZVPD//PBEh-3c[DCOSMO-RS] + RRHO(GFN-xTB[GBSA]) + Gsolv(COSMO-RS[TZVPD]) and linear fit		*xvxzd*	Blind	[34]
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)	*eyetm*	*8xt50*	Blind	[34]
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)	*ccpmw*	*yqkga*	Blind	[34]
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	*0wfzo*		Blind	[35]
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	*z3btx*		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	*758j8*		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	*hgn83*		Blind	

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

lower than we have hoped for. 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(BJ)/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 Classsic), *nb007* (Schrodinger/Epik Scan)) and QSPR/ML categories (*gyuhx* (Simulations Plus), *nb017* (MoKa)). These five methods with RMSE less than 1 pK_a unit also are 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 uncertainty of the correlation statistics. 32 out of 38 methods have R higher than 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), *xvxzd* (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 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). 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 pK_a units after the challenge [29].

Null prediction methods based on database lookup (*5nm4j* and *pwn3m*) had similar performance, roughly RMSE of 2.5 pK_a units, MAE of 1.5 pK_a units, R^2 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.

Distribution of macroscopic pK_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 pK_a values.

There were four submissions of QM+LEC category that used COSMO-RS implicit solvation model. It was interesting that while three of these achieved the lowest RMSE among QM-based methods (*xvxzd*, *yqkga*, and *8xt50*) [34] and one of them showed the highest RMSE (*0hxtm* (COSMOtherm_FINE17)) in SAMPL6 Challenge macroscopic pK_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 oscilator (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 [34] *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 much 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 pK_a prediction results and RRHO correction provided a minor improvement [34]. 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 pK_a value, we also analyzed missing or extra pK_a predictions. Analysis of the pK_a values with accuracy- and correlation-based error metrics was only possible after the assignment of predicted macroscopic pK_as to experimental pK_as through the Hungarian matching, although, this approach masks pK_a prediction issues in the form of extra or missing macroscopic pK_a predictions. To capture this form of prediction errors we reported the number of unmatched experimental pK_as (missing pK_a predictions and the number of unmatched predicted pK_as (extra pK_a predictions) after Hungarian matching for each method. Both missing and extra pK_a prediction counts were only considered for the pH range of 2-12 which was the limits of experimental measurements. The lower subplot of Fig. 2 shows the total count of unmatched experimental or predicted pK_as 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 pK_a value accuracy and the number of pK_as can be viewed together. Presence of missing or extra macroscopic pK_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 challenge results, extra macroscopic pK_a predictions were found to be more common than missing pK_a predictions. In pK_a prediction evaluations usually accuracy of ionization states predicted within a pH range seen is neglected. When predictions are only evaluated for pK_a value accuracy with numerical matching algorithms more pK_a predictions are likely to lead to lower prediction errors. Therefore, it is not surprising that methods are biased to predict extra pK_a values. The SAMPL6 pK_a Challenge experimental data consists of 31 macroscopic pK_as in total, measured for 24 molecules (6 molecules in the set have multiple pK_as). Within the 10 methods with lowest RMSE only *xvxzd* method has an error of missing predicted pK_a (2 unmatched out of 31 experimental pK_as), and all other methods that rank top 10 according to RMSE have extra predicted pK_as ranging from 1 to 13. Two prediction sets without any extra pK_a predictions and low RMSE are *8xt50* (ReSCoSS conformations // DSD-BLYP-D3 reranking // COSMOtherm pKa) and *nb015* (ChemAxon/Chemicalize).

3.1.1 Consistently well-performing methods for macroscopic pK_a prediction

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

Consistently well performing methods for macroscopic pK_a prediction included methods from all categories. Two methods of the QM+LEC category were *xvxzd* (DSD-BLYP-D3(BJ)/def2-TZVPD//PBEh-3c[DCOSMO-RS] + RRHO(GFN-xTB[GBSA]) + Gsolv(COSMO-RS[TZVPD]) and linear fit) and *(*8xt50) (ReSCoSS conformations // DSD-BLYP-D3 reranking // COSMOtherm pKa) and both used COSMO-RS approach. Empirical pK_a predictions with top performance were both proprietary software. From QSPR and LFER categories, *gyuhx* (Simulation Plus) and *xmymhm* (ACD/pKa Classic) were the methods that made it to consistently well-performing methods list. Simulation Plus pK_a prediction method consisted of 10 artificial neural network ensembles trained on 16,000 compounds for 10 classes of ionizable atoms. Atom type and the local molecular environment was how the ionization class of each atom was determined [36]. ACD/pKa Classic which was trained on method 17,000 compounds uses Hammet-type equations and tries to capture effects related to tautomeric equilibria, covalent hydration, resonance effects, and α,β-unsaturated systems [24].

In Figure 4 prediction vs. experimental data correlation plots of macroscopic pK_a predictions with 4 consistently well-performing methods, a representative average method, and the null method(*5nm4j*). The representative method with average performance (*2ii2g* (EC-RISM/MP2/cc-pVTZ-P2-q-noThiols-2par)) was selected as the method with the highest RMSE below the median of all methods.

3.1.2 Which chemical properties are driving macroscopic pK_a prediction failures?

In addition to comparing the performance of methods that participated in the SAMPL6 Challenge, we also wanted to analyze macroscopic pK_a predictions from the perspective of challenge molecules and determine whether particular compounds suffer

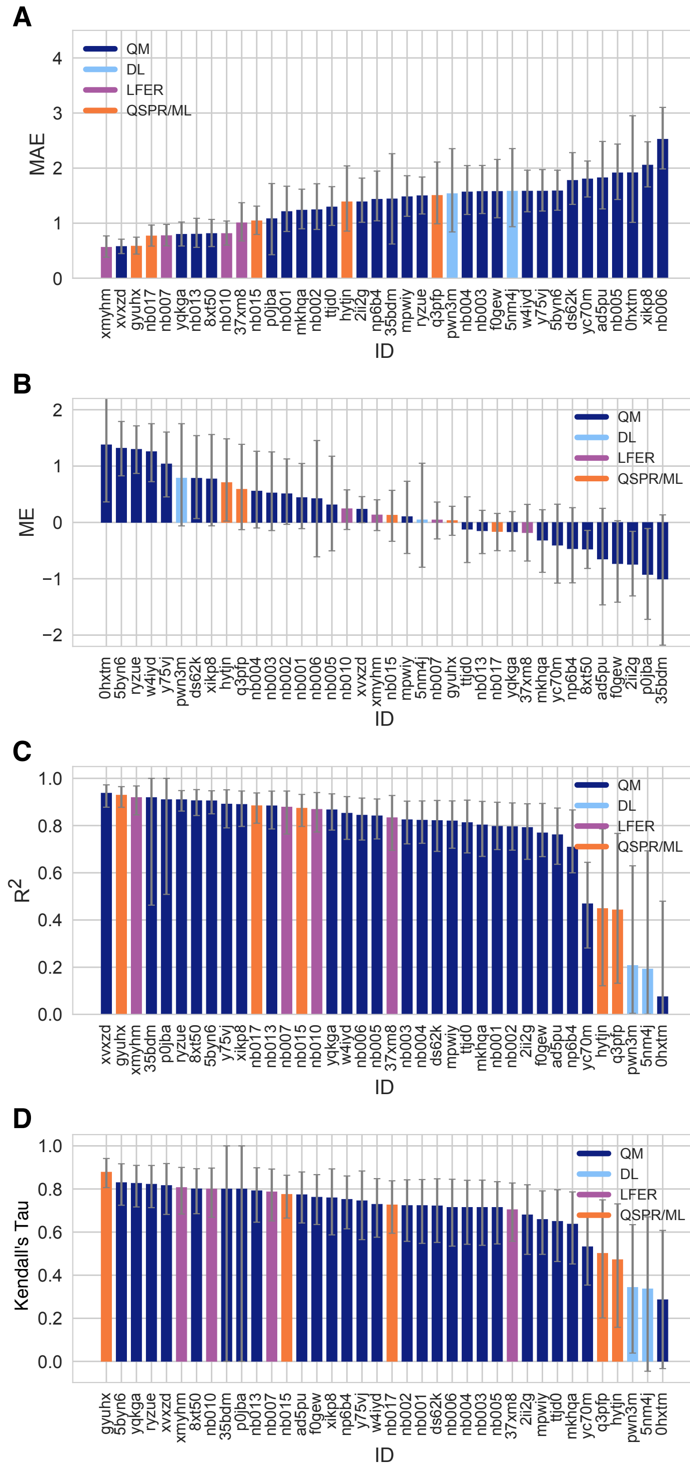


Figure 3. Additional performance statistics for macroscopic pK_a predictions based on Hungarian matching. Methods are indicated by submission IDs. Mean absolute error (MAE), mean error (ME), Pearson's R^2, and Kendall's Rank Correlation Coefficient Tau (τ) are shown, with error bars denoting 95% confidence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

---

from larger inaccuracy in pK_a predictions. The goal of this analysis is to provide insight on which molecular properties or moieties might be causing larger pK_a prediction errors. In Fig. 5 2D depictions of challenge molecules are presented with MAE calculated

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_as and less than 7 unmatched predicted pK_as 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]
xvxzd	Full quantum chemical calculation of free energies and fit to experimental pKa	0.68 [0.54, 0.81]	0.58 [0.45, 0.71]	0.94 [0.88, 0.97]	0.82 [0.68, 0.92]	2	4
gyuhx	S+pKa	0.73 [0.55, 0.91]	0.59 [0.44, 0.74]	0.93 [0.88, 0.96]	0.88 [0.8, 0.94]	0	7
xmyhm	ACD/pKa Classic	0.79 [0.52, 1.03]	0.56 [0.38, 0.77]	0.92 [0.85, 0.97]	0.81 [0.68, 0.9]	0	3
8xt50	ReSCoSS conformations // DSD-BLYP-D3 reranking // COSMOtherm 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

for their macroscopic pK_a predictions over all methods, based on Hungarian match. For multiprotic molecules, MAE was averaged over all the pK_as. For the analysis of pK_a prediction accuracy observed for each molecule, MAE is a more appropriate statistical value than RMSE for following global trends. This is because MAE value less sensitive to outliers than is RMSE.

A comparison of the prediction performance of individual molecules is shown in Fig. 6. In Fig. 6A MAE each molecule is shown considering all blind predictions and reference calculations. A cluster of molecules marked orange and red have higher than average MAE. Molecules marked red (SM06, SM21, and SM22) are the only compounds in the SAMPL6 dataset with bromo or 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 marked orange (SM03, SM10, SM18, SM19, and SM20) all have sulfur-containing heterocycles, and all molecules except SM18 of this group have MAE larger than 1.6 pK_a unit. SM18 despite containing thiazole group has a low MAE. SM18 is the only compound with three experimental pK_as and we suspect the presence of multiple experimental pK_as could have a masking effect on the errors captured by MAE with Hungarian matching due to more pairing choices.

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

SAMPL6 pK_a set consists of only 24 small molecules which limits our ability to do statistically confirm the determination of which chemical substructures cause greater errors in pK_a predictions. Still, the trends observed in this challenge point to molecules with iodo, bromo, and sulfur-containing heterocycles with larger prediction errors of macroscopic pK_a value. We hope that reporting this observation will lead to the improvement of methods for similar compounds with such moieties.

We have also looked for correlation with molecular descriptors for finding other potential explanations for why macroscopic pK_a predictions were larger in some molecules. While testing the correlation between errors and many molecular descriptors it is important to keep the possibility of spurious correlations in mind. We haven't observed any significant correlation between numerical pK_a predictions and the descriptors we have tested. First of all, higher number of experimental pK_as (Fig. 6A) did not seem to associate with lower pK_a prediction performance. But we need to keep in mind that there was a low representation of multiprotic compounds in the SAMPL6 set (5 molecules with 2 macroscopic pK_as and one molecule with 3 macroscopic pK_a). Other descriptors we checked for were amide group presence, molecular weight, heavy atom count, rotatable bond count, heteroatom count, heteroatom to carbon ratio, ring system count, maximum ring size, and the number of microstates (as enumerated for the challenge). Correlation plots and R^2 values can be seen in Fig. S2. We had suspected that pK_a prediction methods may be trained better for moderate values (4-10) than extreme values as molecules with extreme pK_as are less likely to change ionization states close to physiological pH. To test this we look at the distribution of absolute errors calculated for all molecules 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 macroscopic pK_as was not a factor affecting prediction error seen in SAMPL6 Challenge.

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

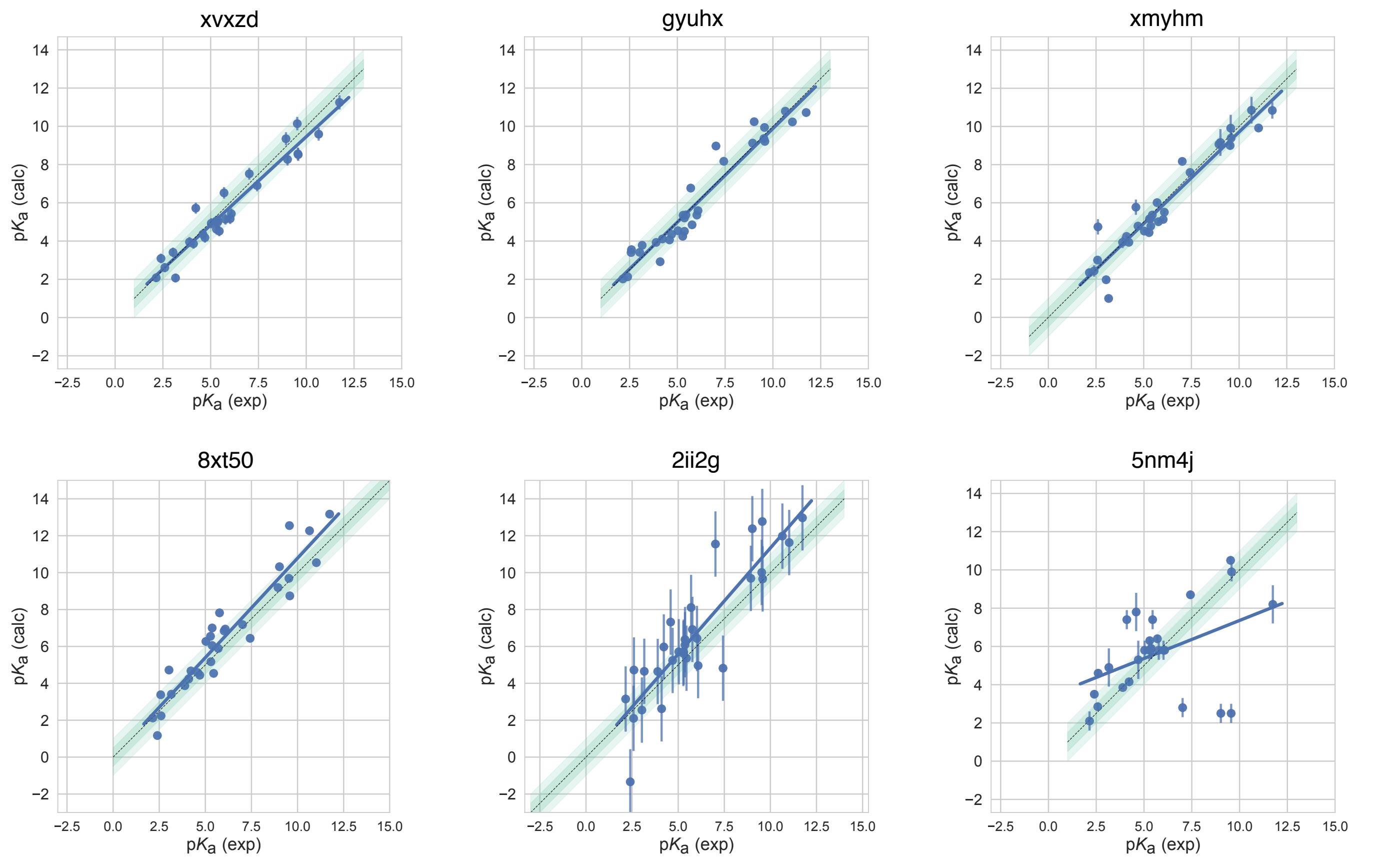


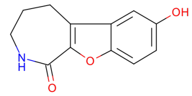
Figure 4. Predicted vs. experimental value correlation plots of 4 consistently well-performing methods, a representative method with average performance (*2ii2g*), and the null method (*5nm4j*). Dark and light green shaded areas indicate 0.5 and 1.0 units of error. Error bars indicate standard error of the mean of predicted and experimental values. Experimental pK_a SEM values are too small to be seen under the data points. EC-RISM/MP2/cc-pVTZ-P2-q-noThiols-2par method (*2ii2g*) was selected as the representative method with average performance because it is the method with the highest RMSE below the median.

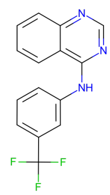
3.2 Analysis of microscopic pK_a predictions using microstates determined by NMR for 8 molecules

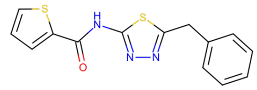
The common approach for analyzing microscopic pK_a prediction accuracy has been to compare it to experimental macroscopic pK_a data, assuming experimental pK_as describe titrations of distinguishable sides and, therefore, equal to microscopic pK_as. But this typical approach fails to evaluate the methods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

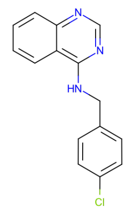
Analysis of microscopic pK_a predictions of the SAMPL6 Challenge was not straight-forward due to the lack of experimental data with microscopic detail. For 24 molecules macroscopic pK_as were determined with the spectrophotometric method. For 18 molecules single macroscopic titration was observed and for 6 molecules multiple experimental pK_as were reported. For 18 molecules with single experimental pK_a it is probable that the molecules are monoprotic and therefore macroscopic pK_a value is equal to the microscopic pK_a, but there is no direct experimental evidence to support that this is the case but only the support from prediction methods. There is always the possibility that the macroscopic pK_a observed is the result of two different titrations overlapping closely with respect to pH. We did not want to bias the blind challenge analysis with any prediction method. Therefore, we believe analyzing the microscopic pK_a predictions via Hungarian matching to experimental values with the assumption that the 18 molecules have a single titratable site is not the best approach. Instead, analysis at the level of macroscopic pK_as is much more appropriate when a numerical matching scheme is the only option to evaluate predictions using macroscopic experimental data.

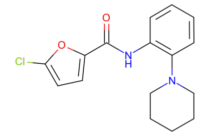
For a subset of the molecules in the dataset of 8 molecules, dominant microstates were inferred from NMR experiments. This dataset was extremely useful for guiding the assignment between experimental and predicted pK_a values based on microstates. In this section, we present the performance evaluations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

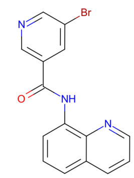

SM01
0.72 [0.51, 0.99]

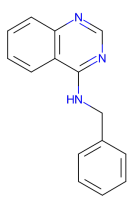

SM02
1.01 [0.88, 1.15]

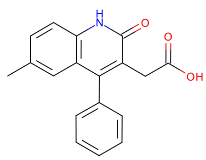

SM03
2.39 [1.78, 3.00]

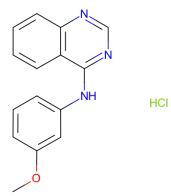

SM04
0.98 [0.81, 1.15]

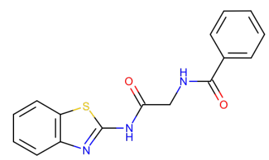

SM05
1.59 [1.24, 1.93]

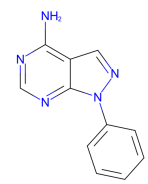

SM06
2.03 [1.65, 2.47]

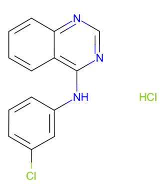

SM07
1.51 [1.11, 1.94]

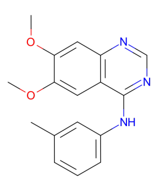

SM08
0.87 [0.62, 1.15]

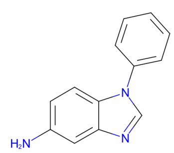

SM09
1.06 [0.87, 1.26]

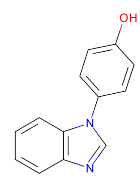

SM10
2.24 [1.60, 2.93]

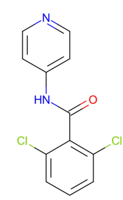

SM11
0.79 [0.54, 1.06]

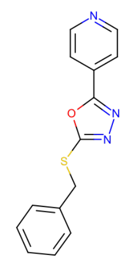

SM12
1.05 [0.85, 1.26]

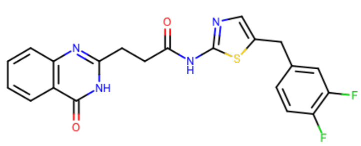

SM13
1.01 [0.76, 1.27]

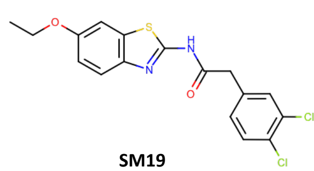

SM14
1.45 [1.19, 1.73]

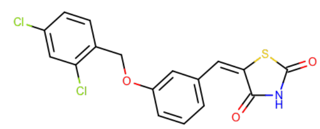

SM15
1.04 [0.86, 1.22]

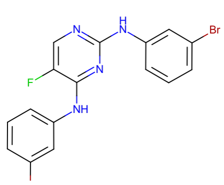

SM16
1.07 [0.88, 1.28]

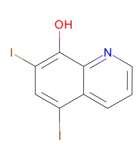

SM17
1.19 [0.92, 1.48]

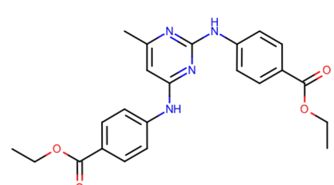

SM18
1.06 [0.84, 1.33]


SM19
2.23 [1.64, 2.89]


SM20
1.69 [1.44, 1.93]


SM21
2.00 [1.46, 2.56]


SM22
1.73 [1.29, 2.21]


SM23
1.12 [0.80, 1.46]

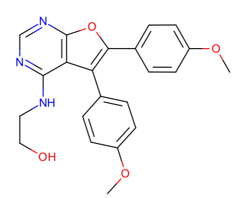

SM24
1.01 [0.77, 1.26]

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, which could be more than one per method for multiprotic molecules (SM06, SM14, SM15, SM16, SM18, SM22). Hungarian matching algorithm was employed for pairing experimental and predicted pK_a values. MAE values are reported with 95% confidence intervals.

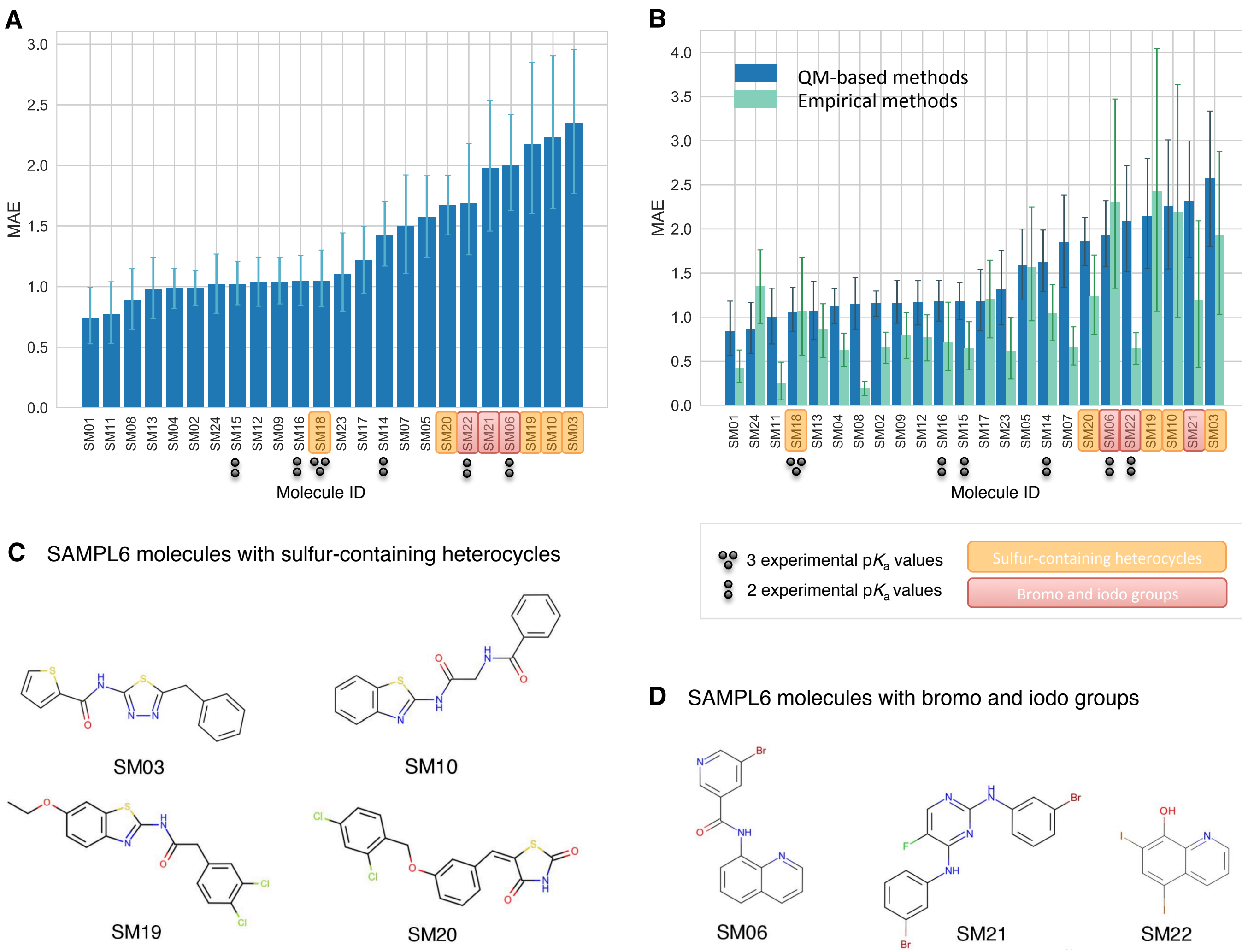


Figure 6. Average prediction accuracy calculated over all prediction methods was lower for molecules with sulfur-containing heterocycles, bromo, and iodo groups. (A) MAE calculated for each molecule as an average of all methods. **(B)** MAE of each molecule broken out by method category. QM-based methods (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.

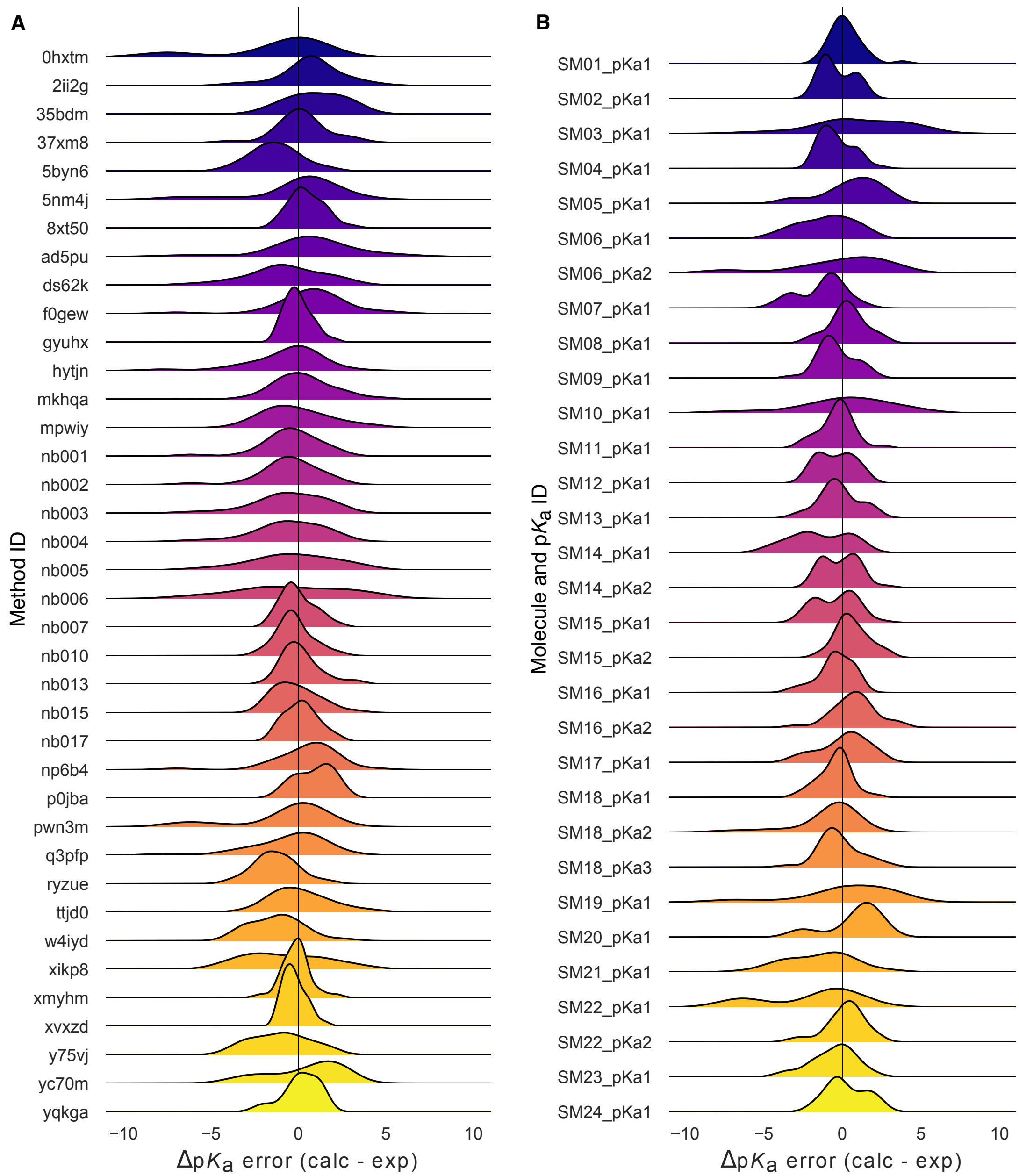


Figure 7. Macroscopic pK_a prediction error distribution plots show how prediction accuracy varies across methods and individual molecules. (A) pK_a prediction error distribution for each submission for all molecules according to Hungarian matching. **(B)** Error distribution for each SAMPL6 molecule 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.

with experimentally determined dominant microstates.

3.2.1 Microstate-based matching revealed errors masked by pK_a value-based matching between experimental and predicted pK_as

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

The number of unmatched predicted microscopic pK_as is shown in the lower bar plots of Fig. 8B and C, to emphasize the large number of microscopic pK_a predictions submitted by many methods. In the case of microscopic pK_a, the number of unmatched predictions does not indicate an error in the form of an extra predicted pK_a, because the spectrophotometric experiments do not capture all microscopic pK_as theoretically possible (transitions between all pairs of microstates that are 1 proton apart). pK_as of transitions to and from very high energy tautomers are very hard to measure by experimental methods, including the most sensitive methods like NMR. The reason we plotted them was more to demonstrate how the increased number of prediction value choices for Hungarian matching can lead to erroneously low RMSE values. We have also checked how often Hungarian matching led to the correct matches between predicted and experimental pK_a in terms of the microstate pairs, i.e. how often the microstate pair of the Hungarian match recapitulates the dominant microstate pair of the experiment. The overall accuracy of microstate pair matching was found to be low for SAMPL6 Challenge submission. Fig. S4 shows that for most methods the predicted microstate pair selected by Hungarian match did not match experimentally determined microstate pair. This means the lower RMSE results obtained from Hungarian matching are low for the wrong reason. Matching experimental and predicted values on the basis of microstate IDs do not suffer from this problem.

The disadvantage of the evaluation through the microstate-based matching approach is that the conclusions in this section are only about a subset of challenge compounds with limited diversity. This subset is composed of 6 molecules 4-aminoquinazoline and 2 molecules with benzimidazole scaffolds, and a total of 10 pK_a values. The sequences of dominant microstates for SM07 and SM14 were determined by NMR experiments directly [7] and dominant microstates of their derivatives were inferred by taking them as reference (Fig. 8). Although we believe that microstate-based evaluation is more informative, the lack of a large experimental dataset limits the conclusions to a very narrow chemical diversity. Microstate-based matching revealed errors masked by pK_a value-based matching between experimental and predicted pK_as.

3.2.2 Accuracy of pK_a predictions evaluated by microstate-based matching

Both accuracy and correlation based statistics were calculated for predicted microscopic pK_a values after microstate-based 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 statistics can be found in Table S4. Due to small number of data points in this set, correlation based statistics calculated shows large uncertainty and provide less utility for distinguishing better performing methods. Therefore we focused more on accuracy based metrics for the analysis of microscopic pK_as than correlation based metrics. In terms of accuracy of microscopic pK_a value, all three QSPR/ML based methods (*nb016* (MoKa),*hdiyq* (Simulations Plus), *6tvf8* (OE Gaussian Process)), three QM-based methods (*nb011* (Jaguar), *ftc8w* (EC-RISM/MP2/cc-pVTZ-P2-q-noThiols-2par), *t8ewk* (COSMOlogic_FINE17)), and one LFER method (*v8qph* (ACD/pKa GALAS)) achieved RMSE lower than 1 pK_a unit. The same 6 methods also have the lowest MAE.

3.2.3 Evaluation of dominant microstate prediction accuracy

For many computational chemistry approaches including structure-based modeling of protein-ligand interactions, predicting the ionization state and the exact position of protons is needed to establish what to include in the modeled system. This is why in addition to being able to predict pK_a values accurately, we need pK_a prediction methods to be able to capture microscopic protonation states accurately. Even when the predicted pK_a value is very accurate, the predicted protonation site can be

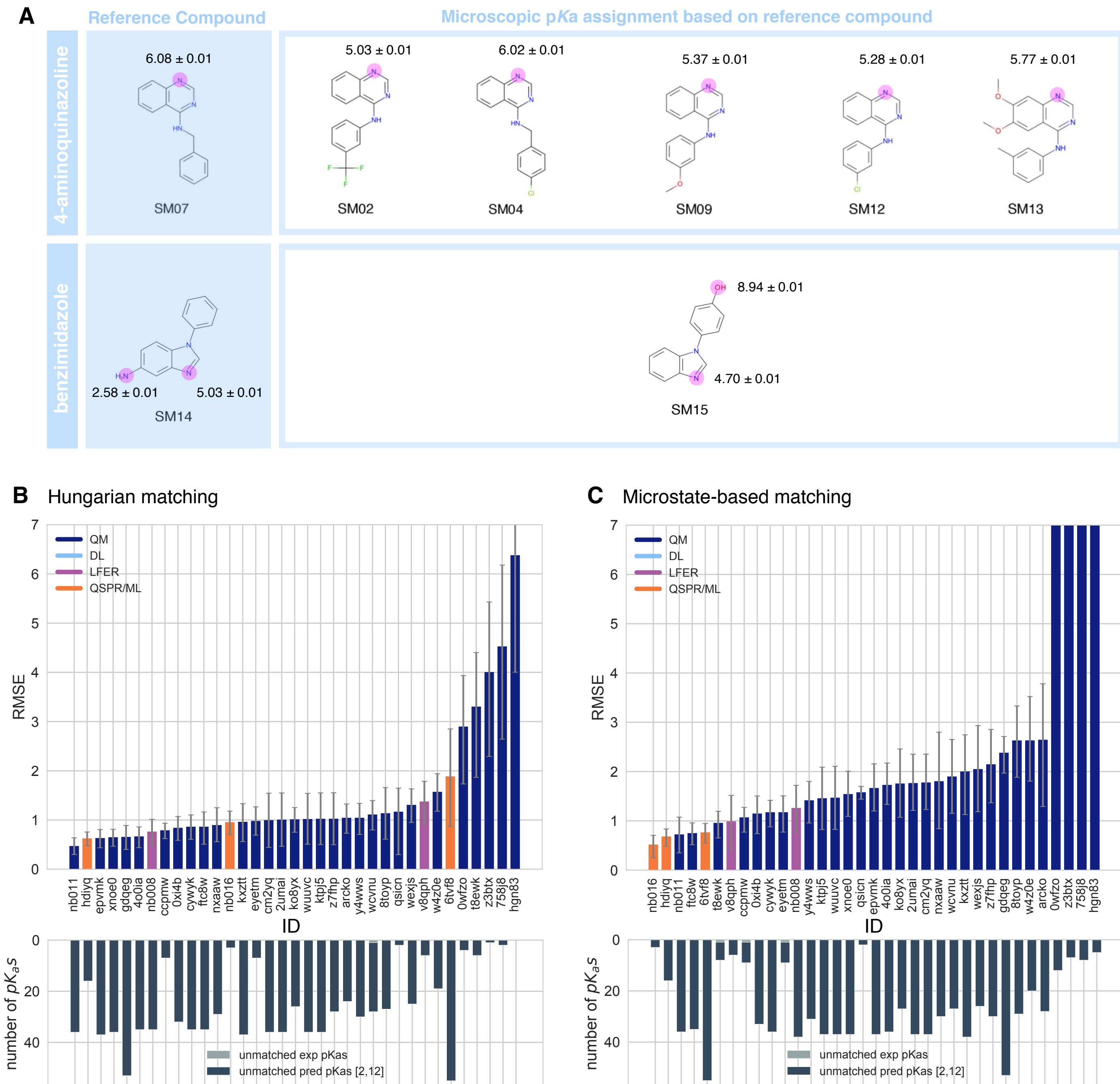


Figure 8. NMR determination of dominant microstates allowed in-depth evaluation of microscopic pK_a predictions of 8 compounds. **A** Dominant microstate sequence of two compounds (SM07 and SM14) were determined by NMR [7]. Based on these reference compounds dominant microstates of 6 other derivative 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_as. **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_as 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_as (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.

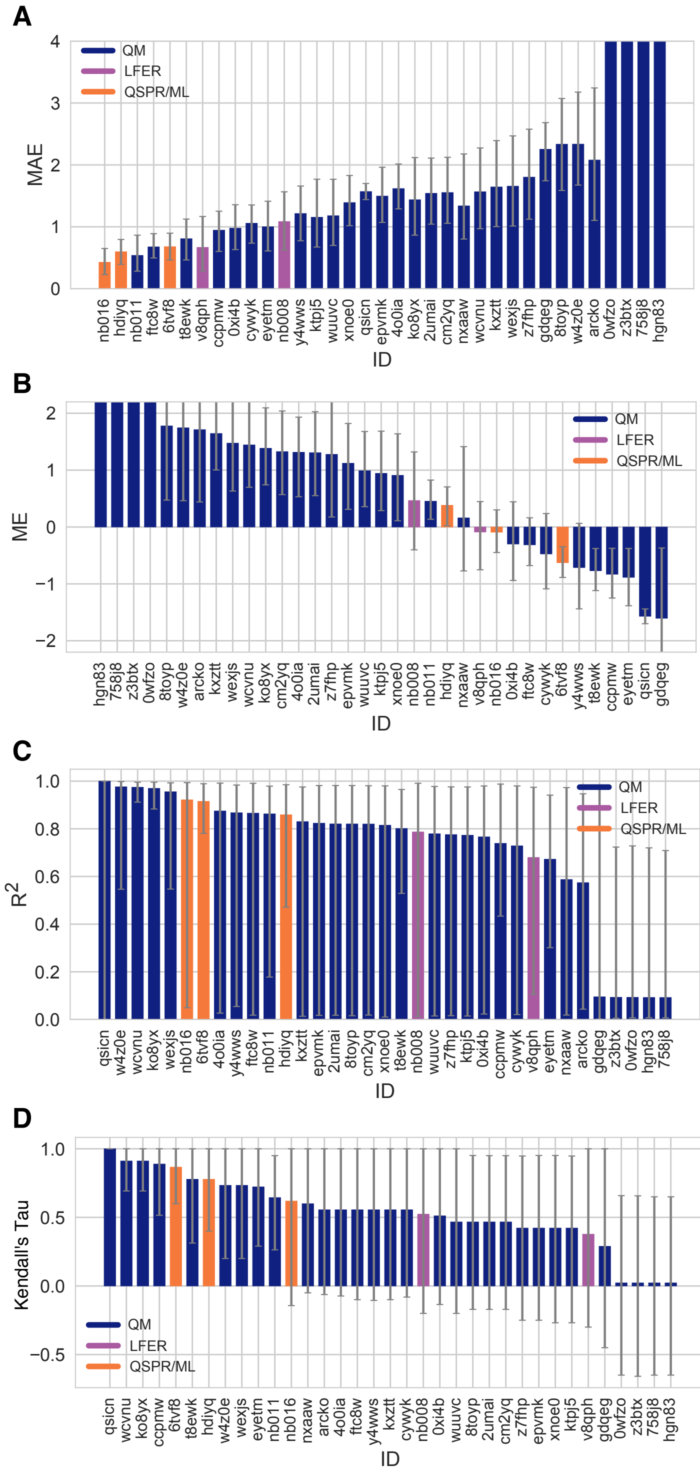


Figure 9. Additional performance statistics for microscopic pK_a predictions for 8 molecules with experimentally determined dominant microstates. Microstate-based matching was performed between experimental pK_a values and predicted microscopic pK_as. Mean absolute error (MAE), mean error (ME), Pearson's R^2, and Kendall's Rank Correlation Coefficient Tau (τ) are shown, with error bars denoting 95% confidence intervals obtained by bootstrapping over challenge molecules. 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, z3btx, 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.

wrong. Therefore, we assessed if methods participating in the SAMPL6 pK_a Challenge were predicting correctly the sequence of dominant microstates, i.e. dominant tautomers of each charge state observed between pH 2 and 12.

Dominant microstate prediction accuracy of microscopic pK_a prediction method are shown in Fig. 10. The dominant microstate sequence is essentially the sequence of states that are most visible experimentally, due to their higher fractional population and relative free energy within the tautomers of each charge. To extract the dominant tautomers predicted for the sequence of ionization states of each method, first, the relative free energy of microstates were calculated at reference pH 0 [21]. Then to determine the dominant microstate of each charge, we have selected the lowest energy tautomer for each ionization states of the charges -1, 0, 1, and 2 (the charge range captured by NMR) experiments. Than predicted and experimental dominant microstates were compared for each charge to calculate the fraction of correctly predicted dominant tautomers. This value is reported as the dominant microstate accuracy for all charges (Fig. 10A). Dominant microstate prediction errors were present the methods participating in the SAMPL6 pK_a Challenge. 10 QM and 3 QSPR/ML methods did not make any mistakes in dominant microstate predictions, although, they are expected to be making mistakes in the relative ratio of tautomers (free energy difference between microstates) as reflected by pK_a value errors. While all the participating QSPR/ML methods showed good performance in dominant microstate prediction, LFER and some QM methods made mistakes. Accuracy of the prediction of the neutral dominant tautomers was perfect for all methods, except *qsicn* (Fig. 10B). But errors in predicting the major tautomer of charge +1 were much more frequent. 22 out of 35 prediction sets made at least one error in prediction the lowest energy tautomer with +1 charge. We didn't include ionization states with charges -1 and +2 in this assessment because we had only one compound with these charges in the dataset. Never the less, dominant tautomer prediction errors seem to be a bigger problem for charged tautomers than the neutral tautomer.

Experimental data of the sequence of dominant microstates was only available for 8 compounds. Therefore conclusions the performance of methods in terms of dominant tautomer prediction are limited to this narrow chemical diversity (benzimidazole and 4-aminoquinazoline derivatives). We present this analysis as a prototype of how microscopic pK_a predictions should be evaluated. To reach broad conclusions about which methods are better for capturing dominant microstates and ratios of tautomers we hope that in the future more extensive evaluations can be done with larger experimental datasets following the strategy we are demonstrating here. Even if experimental microscopic pK_a measurement data is not available, experimental dominant tautomer determinations are still informative for assessing prediction methods. methods.

Focusing on dominant microstate sequence prediction accuracy from the perspective of molecules showed that major tautomer of SM14 cationic form was the most frequently mispredicted one. Fig. 10 shows the dominant microstate prediction accuracy calculated for individual molecules for charge states 0 and +1, averaged over all prediction methods. SM14, the molecule that exhibits the highest microstate prediction error, 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 predicting microstates accurately. Other molecules are monoprotic (4-aminoquinazolines) or their experimental pK_a values are very well separated (SM14, 4.2 pK_a units). It would be very interesting to expand this assessment to a larger variety of drug-like molecules to discover for which structures tautomer predictions are more accurate and for which structure computational predictions are not as reliable.

3.2.4 Consistently well-performing methods for microscopic pK_a predictions

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

Simulations Plus pK_a prediction method is the only method that appeared to be consistently well performing in both the assessment for macroscopic and microscopic pK_a prediction (*gyuhx* and *hdiyq*). However it is worth noting that two methods that were in consistently top-performing methods list for macroscopic pK_a predictions lacked equivalent submissions of their underlying microscopic pK_a predictions and therefore could not be evaluated at the microstate level. These methods were (ACD/Classic

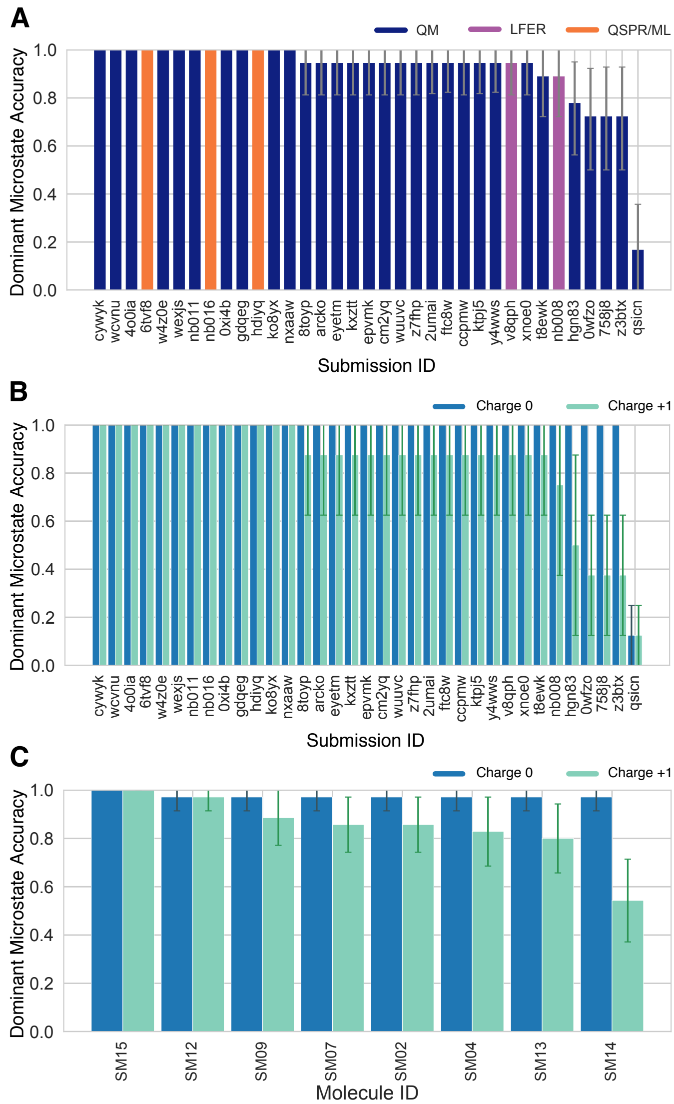


Figure 10. Some methods predicted the sequence of dominant tautomers inaccurately. Prediction accuracy of the dominant microstate of each charged state was calculated using the dominant microstate sequence determined by NMR for 8 molecules as reference. **(A)** Dominant microstate accuracy vs. submission 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 generating 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 showed 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.

pKa) and *xvxzd*(DSD-BLYP-D3(BJ)/def2-TZVPD//PBEh-3c[DCOSMO-RS] + RRHO(GFN-xTB[GBSA]) + Gsolv(COSMO-RS[TZVPD]) and linear fit).

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

Physical modeling methods for predicting protein-ligand binding affinities rely on pK_a predictions for modeling the protein and the ligand. As SAMPL6 pK_a Challenge only focused on small molecule pK_a prediction we will ignore the protonation state effects of the protein for now. Many affinity prediction methods such as docking, MM/PBSA, MM/GBSA, absolute or alchemical relative free energy calculation methods predict the affinity of a fixed protonation state of the ligand to a receptor. These models strictly depend on pK_a predictions for determining possible protonation states of the ligand in the aqueous environment and in a protein complex, as well as the free energy penalty to reach those states [3]. Accuracy of pK_a predictions can become a limitation for

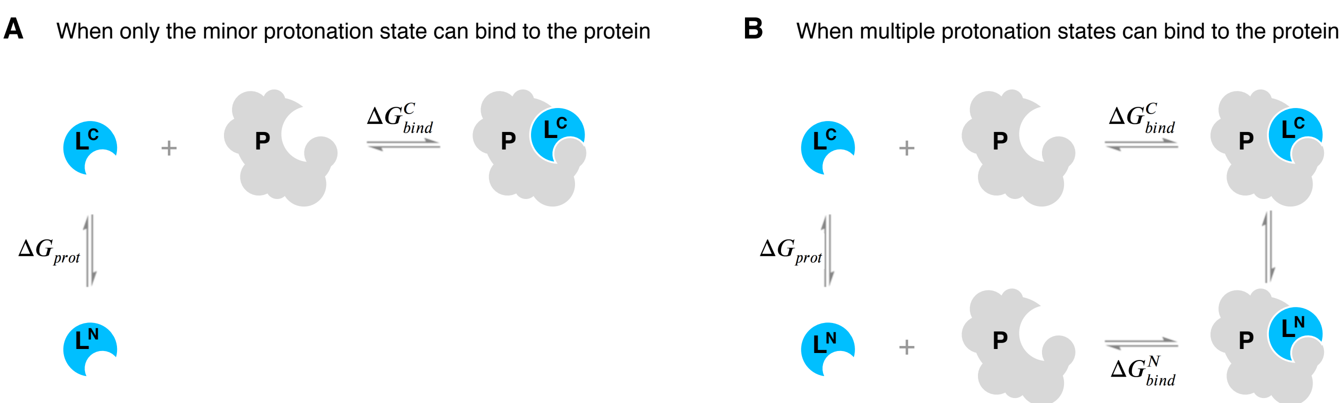
Table 3. Top-performing methods for microscopic pK_a predictions based on consistent ranking within the Top 10 according to various statistical metrics calculated for 8 molecule dataset. Performance statistics are provided as mean and 95% confidence intervals. Submissions that rank in the Top 10 according 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
hdiyq	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 performance of physical models that try to capture molecular association.

In terms of the ligand protonation states, there are two ways in which the pK_a prediction errors can influence the prediction accuracy for protein-ligand binding free energies as depicted in Fig. 11. The first scenario is when a ligand is present in aqueous solution in multiple protonation states (Fig. 11A). 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 (ΔG_{prot}). ΔG_{prot} is a function of pH and pK_a. A 1 unit of error in pK_a value would lead to 1.36 kcal/mol error in overall binding affinity if the protonation state with the minor population binds the protein. The equations in Fig. 11A show the calculation of overall affinity.

In addition to multiple protonation states being present in the aqueous environment, multiple charge states can contribute to the complex formation (Fig. 11B). Then, the overall free energy of binding needs to include 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 ($\Delta G^C_{bind} - \Delta G^N_{bind}$) as shown in Fig. 11B.



$$\Delta G_{bind} = \Delta G^C_{bind} + \Delta G_{prot}$$

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

$$\Delta G_{bind} = \Delta G^N_{bind} + \Delta G_{corr}$$

$$\Delta G_{bind} = \Delta G^N_{bind} - RT\ln\frac{1 + e^{-\frac{\Delta G^C_{bind} - \Delta G^N_{bind}}{RT}} 10^{pK_a - pH}}{1 + 10^{pK_a - pH}}$$

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 ($\Delta G^C_{bind} - \Delta G^N_{bind}$).

Using the equations in Fig. 11B we can model the true MPSC (ΔG_{corr}) value with respect to the difference between pH and the pK_a of the ligand, to see when this value has significant impact to overall binding free energy. In Fig. 12, true MPSC value that needs to be added to the ΔG^N_{bind} is shown for ligands with varying binding affinity difference between protonation states ($\Delta\Delta G = \Delta G^C_{bind} - \Delta G^N_{bind}$) and varying free energy of binding difference between the protonation states. Fig. 12A shows the simulation of a case where for a monoprotic base which has a charged state with a lower affinity than the neutral state. Solid lines show the true correction. In situations where pK_a is lower than pH, correction factor disappears as the ligand fully populates the neutral state ($\Delta G_{bind} = \Delta G^N_{bind}$). As the p$K_a$ value gets larger than the pH, the charged state is populated more and ΔG_{corr} value increases to approach significant $\Delta\Delta G$. What is interesting to note is the pH-pK_a range that ΔG_{corr} changes. It is often assumed that for a basic ligand if 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^C_{bind} only. 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 $\Delta\Delta G = 1 kcal/mol$ or 5 pH units for a system with $\Delta\Delta G = 4 kcal/mol$. 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 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 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 affinity differences between protonation states. For a system whose charged state has lower affinity than the neutral state ($\Delta\Delta G$ = 2 kcal/mol), the absolute error caused by underestimated pK_a by 1 unit only can be up to 0.9 kcal/mol. For a system whose charged state has even lower affinity than the neutral state ($\Delta\Delta G$ = 4 kcal/mol), the absolute error caused by underestimated pK_a by 1 unit only 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^N_{bind}, (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.

The same question about whether or not an MPSC calculated based on an inaccurate pK_a should be included in binding affinity predictions can be asked for different circumstances underestimated or overestimated pK_a values, 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 reality, we 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. But based on the analysis of extreme cases, with 1 unit of pK_a error including MPSC correction is more often than not helpful in improving binding affinity predictions. The detrimental effect of pK_a inaccuracy is still significant, however, 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. Achieving pK_a value prediction accuracy of 0.5 units would significantly help the binding affinity models to incorporate more accurate MPSC terms.

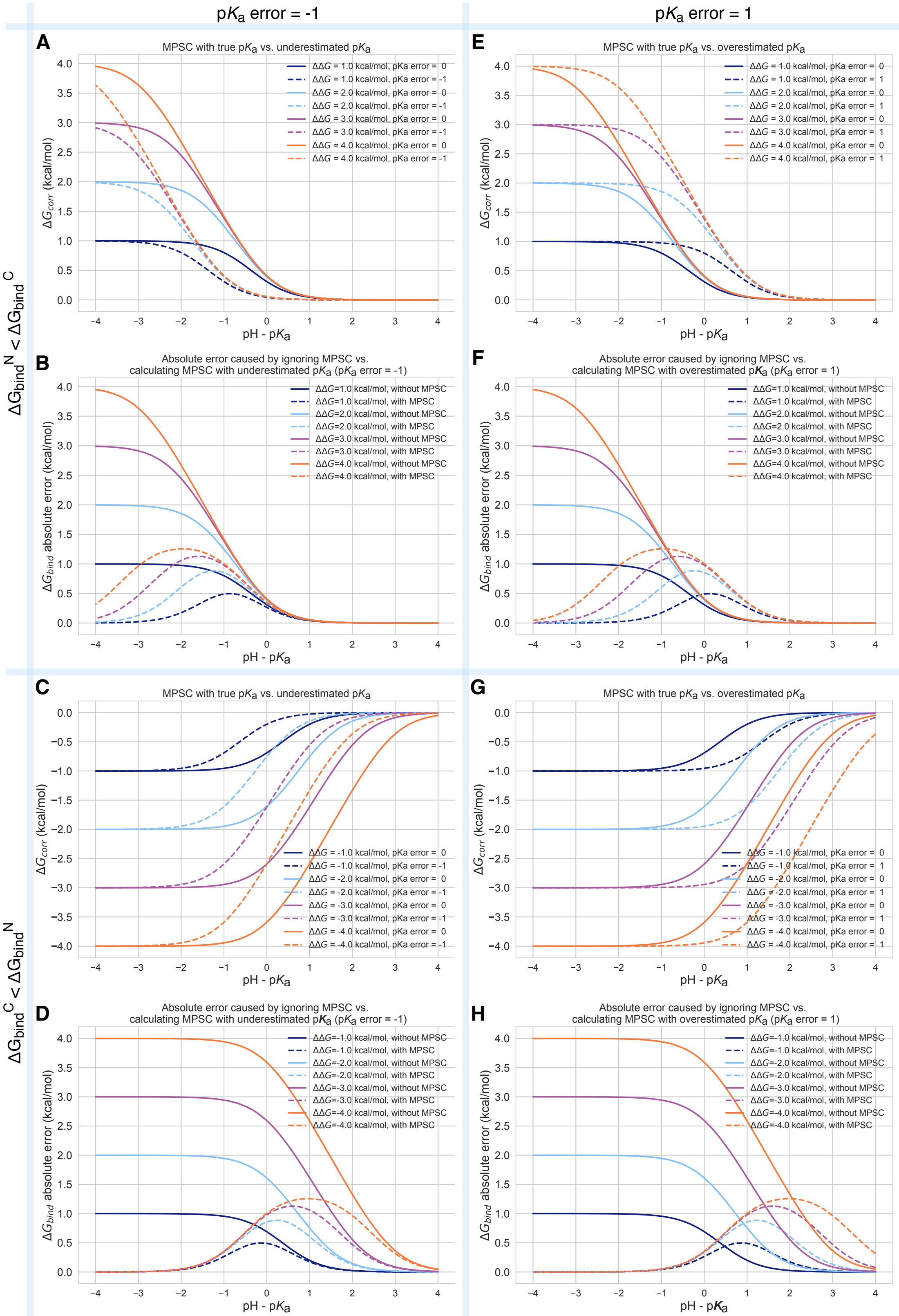


Figure 12. Inaccuracy of pK_a prediction ($\pm$ 1 unit) affects the the accuracy of MPSC and overall protein-ligand binding free energy calculation in varying amounts based on aqueous pK_a value and relative binding affinity of individual protonation states ($\Delta\Delta G = \Delta G^C_{bind} - \Delta G^N_{bind}$). 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 $\Delta\Delta G$.

3.4 Take-away lessons from SAMPL6 pK_a Challenge

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 to achieve prediction errors as low was 0.5 pK_a units and reliable predictions of charge and tautomer states. 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 are factors that complicate pK_a prediction of drug-like molecules. For macroscopic pK_a predictions have not yet reached experimental accuracy. Inter-method variability of macroscopic pK_a measurements can be around 0.5 pK_a units [17]. 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 algorithm (Hungarian matching) to pair predicted and experimental values to calculate performance statistics of macroscopic pK_a predictions. Accuracy, correlation, and extra/missing pK_a prediction counts were the main metrics for macroscopic pK_a evaluations. An RMSE range of 0.7 to 3.2 pK_a units. Only five methods achieved RMSE between 0.7-1 pK_a units, 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 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 included linear empirical corrections to rescale and unbias their pK_a predictions.

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

There were four submissions of QM-based methods that utilized COSMO-RS implicit solvation model. It was interesting that while three of these achieved the lowest RMSE among QM-based methods (*xvxzd*, *yqkga*, and *8xt50*) [34] and one of them showed the highest RMSE (*0hxtm* (COSMOtherm_FINE17)) in SAMPL6 Challenge macroscopic pK_a predictions. The comparison of these methods indicates that capturing conformational ensemble of microstates, high level QM calculations, and RRHO corrections were factors contributing to better macroscopic pK_a predictions. Linear empirical corrections applied QM calculations improved results, especially when the linear correction is calibrated for an experimental dataset using the same level of theory as the deprotonation free energy predictions (as in *xvxzd*). This challenge also points to the advantage of COSMO-RS solvation approach compared to other implicit solvent models.

Evaluation of macroscopic pK_a prediction accuracy of individual molecules on average considering all the predictions in SAMPL6 Challenge provided insight into which molecules posed greater difficulty for pK_a predictions. pK_a prediction errors were higher for compounds with sulfur-containing heterocycles, iodo, and bromo groups. This trend was also conserved when only QM-based methods were analyzed. SAMPL6 pK_a dataset consisted of only 24 small molecules which limited our ability to statistically confirm this conclusion, however, we believe it is worth reporting molecular features that coincided with larger errors even if we can not evaluate the driving reason for these failures.

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

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

Analysis of dominant microstate prediction accuracy of microscopic pK_a showed that some QM and LFER methods made mistakes in predicting the dominant tautomers of the ionization states seen experimentally. Dominant tautomer prediction seemed to be a more prominent problem for charged tautomers than the neutral tautomer. The easiest way to extract dominant microstate sequence from predictions is to calculate the relative free energy of microstates at any reference pH, and determining the lowest energy state in each ionization state. Errors in dominant microstate predictions were very rare for neutral tautomers but more frequent in cationic tautomers with +1 charge of the 8 molecule set. SM14 was the molecule with the lowest dominant microstate prediction accuracy, while dominant microstates predictions for SM15 were perfect for all molecules. SM14 and SM15 both have two experimental pK_as and benzimidazole scaffold. The difference between them is the distance between the experimental pK_a values which is smaller for SM14. These results make sense from the perspective of relative free energies of microstates. Closer pK_a values mean that the free energy difference between different microstates are smaller for SM14, and therefore any error in predicting the relative free energy of tautomers is more likely to cause reordering of relative populations of microstates and impact the accuracy of dominant microstate predictions. It would have been extremely informative to evaluate the tautomeric ratios and relative free energy predictions of microstates, however, experimental data was missing for this approach.

According to statistics calculated with microstate-based matching, we determined a shortlist of consistently well-performing methods for microscopic pK_a predictions of 8 molecule set. These methods that ranked in the top 10 according to RMSE, MEA, and had perfect dominant microstate prediction accuracy included three methods from QM+LEC category and three from QSPR/ML category. Simulations Plus pK_a prediction method was the only method that appeared to be consistently well-performing in both the assessment for macroscopic and microscopic pK_a prediction (*gyuhx* and *hdiyq*), although, due to the size of the experimental datasets evaluation of macroscopic pK_a prediction carried more weight in this performance assessment. Still microscopic pK_a evaluation can provide much more in-depth analysis and can be more informative about capturing reasons for failure.

The performance levels of microscopic and macroscopic pK_a prediction as seen in SAMPL6 pK_a Challenge assessment can be detrimental to the accuracy of protein-ligand affinity predictions and other pH-dependent physicochemical property predictions such as distribution coefficients, membrane permeability, and solubility. Protein-ligand binding affinity predictions rely on pK_a predictions in two ways: determination of relevant aqueous microstates and the free energy penalty to reach these states. Microscopic pK_a predictions with better accuracy are needed for accurate incorporation of multiple protonation state correction (MPSC) to overall binding affinity calculations. We simulated the effect of overestimating or underestimating pK_a of a ligand by one unit on overall binding affinity prediction for a ligand where both cation and neutral states contribute to binding affinity. pK_a prediction error of this magnitude (assuming dominant tautomers were predicted correctly) could cause up to 0.9 and 1.2 kcal/mol error in overall binding affinity when the binding affinity of protonation states are 2 or 4 kcal/mol different, respectively. For the case of 4 kcal/mol binding affinity difference between protonation states the 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 range of pH-pK_a value that MPSC needs to be incorporated in binding affinity predictions can be wider than the widely assumed range of 2 pH units, based on the affinity difference between protonation states. At the level of 1 unit pK_a error incorporating MSPC would improve binding affinity predictions more often than not. If microscopic pK_a could be predicted with 0.5 pK_a units of accuracy, MPSC calculations would be much more reliable.

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

We were disappointed to see that all of the top-performing empirical methods were developed as commercial software that require licenses to run, and there were not any open-source alternatives for empirical pK_a predictions. Since then two publications reported open-source machine learning-based pK_a prediction methods, however, one can only predict the most acidic or most basic macroscopic pK_a values of a molecule [37] and the second one is only trained for predicting pK_a values of monoprotic molecules [38]. Recently a pK_a prediction methodology was published that describes a mixed approach of semi-empirical QM calculations and machine learning that can predict macroscopic pK_as of both mono-and polyprotic species [39]. The authors reported RMSE of 0.85 for the retrospective analysis performed on the SAMPL6 dataset.

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

The first pK_a challenge of the SAMPL series was useful for understanding the current state of the field and led to many lessons. We believe the highest benefit can be achieved if further iteration so of small molecule pK_a prediction challenges can be organized, creating motivation for improving protonation state prediction methods for drug-like molecules. In future challenges, it is desirable to increase chemical diversity to cover more of common scaffolds [40] and functional groups [41] seen in drug-like molecules, and gradually increasing the complexity of molecules.

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

In this paper, we demonstrated how experimental microstate information can guide the analysis further than the typical pK_a evaluation approach that has been used so far. The traditional pK_a evaluation approach only focuses on the numerical error of the pK_a values and neglects the difference between macroscopic and microscopic pK_a definitions. This is mainly caused by the lack of pK_a datasets with microscopic detail. To improve pK_a and protonation state predictions of multiprotic molecules it is necessary to embrace the difference between macroscopic and microscopic pK_a definitions and select strategies for experimental data collection and prediction evaluation accordingly. In SAMPL6 Challenge the analysis was limited by the availability of experimental microscopic data as well. As usual macroscopic pK_a values were abundant (24 molecules) and limited data on microscopic states was available (8 molecules), although the later opened new avenues for evaluation. For future blind challenges for multiprotic compounds, striving to collect experimental datasets with microscopic pK_as would be very beneficial. Benchmark datasets of microscopic pK_as are currently missing. This limits the improvement of pK_a and tautomer prediction methods for multiprotic molecules. If the collection of experimental microscopic pK_as is not possible due to time and resource cost of such NMR experiments, at least supplementing the more automated macroscopic pK_a measurements with NMR-based determination of the dominant microstate sequence or tautomeric ratios of each ionization state can create very useful benchmark datasets. This supplementary information can allow microstate-based assignment between experimental and predicted pK_as and a more realistic assessment of method performance.

If the only available experimental data is in the form of macroscopic pK_a values, the best way to evaluate computational predictions is by calculating predicted macroscopic pK_a predictions. With the conversion of microscopic pK_a to macroscopic pK_as all the structural information about the titration site is lost and only remaining information is the total charge of macroscopic ionization states. Unfortunately, most macroscopic pK_a measurements including potentiometric and spectrophotometric methods do not capture the absolute charge of the macrostates. The spectrophotometric method does not measure charge at all. The potentiometric method can only capture the relative charge change between macrostates. Only pH-dependent solubility based pK_a estimations can differentiate the neutral and charged states from one another. So it is very common to have experimental datasets of macroscopic pK_a without any charge or protonation position information regarding the macrostates. This causes an issue of assigning predicted and experimental pK_a values before any error statistics can be calculated. As delineated by Fraczkiewicz et. al. the fairest and reasonable solution for pK_a matching problem involves an assignment algorithm that preserves the order of predicted and experimental microstates and uses the principle of smallest differences to pair values [17]. We recommend Hungarian matching with the squared error cost function. The algorithm is available in SciPy package (scipy.optimize.linear_sum_assignment) [20]. In addition to the analysis of numerical error statistics after Hungarian matching, at the very least number of missing and extra pK_a predictions must be reported based on unmatched pK_a values. Missing or extra pK_a predictions point to a problem with capturing the right number of ionization states within the pH interval of the experimental measurements. We have demonstrated that for microscopic pK_a predictions performance analysis based in Hungarian matching results in overly optimistic and misleading results, instead the employed microstate-based matching provided a more realistic assessment.

For capturing all the necessary information related to pK_a predictions we allowed three different submission types in SAMPL6:

(1) macroscopic pK_a values, (2) microscopic pK_a values and microstate pair identities, (3) fractional population of microstates with respect to pH. We realized later that collecting fractional populations of microstates was redundant since microscopic pK_a values and microstate pairs capture all the necessary information to construct fractional population vs. pH curves. Only microscopic and macroscopic pK_a values were used for the challenge analysis presented in this paper. While exploring ways to evaluate SAMPL6 pK_a Challenge results, we developed a better way to capture microscopic pK_a predictions as presented in an earlier paper [21]. This alternative reporting format consists of charge and relative free energy of microstates with respect to a reference microstate and pH predicted by pK_a predictions. This approach presents the most concise method of capturing all necessary information regarding microscopic pK_a predictions and allows calculation of predicted microscopic pK_as, microstate population with respect to pH, macroscopic pK_as, macroscopic population with respect to pH, and tautomer ratios. Still, there may be methods developed to predict macroscopic pK_as directly instead of computing it from microstate predictions that justifies allowing a macroscopic pK_a reporting format. In future challenges, we recommend collecting pK_a predictions with two submission types: (1) macroscopicpK_a values and (2) microstates, their total charge, and relative free energies with respect to a specified reference microstate and pH.

In SAMPL6 we provided an enumerated list of microstates and their assigned microstate IDs because we were worried about parsing submitted microstates in SMILES from different sources correctly. There were two disadvantages to this approach. First, this list of enumerated microstates was used as input by some participants which was not our intention. Second, the first iteration of enumerated microstates was not complete. We had to add new microstates and assign them microstate IDs for a couple of rounds until reaching a complete list. In future challenges, a better way of handling the problem of capturing predicted microstates would be asking participants to specify the predicted protonation states themselves and assigning identifiers after the challenge deadline to aid comparative analysis. This would prevent the partial unblinding of protonation states. Predicted states can be submitted as mol2 files that represents the microstate with explicit hydrogens. The organizers must only provide the microstate that was selected as the reference state for the relative microstate free energy calculations.

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

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

To maximize the lessons that can be learned from blind challenges we believe in the utility of evaluating predictions of different physicochemical properties for the same molecules in consecutive challenges. In SAMPL6 we organized both pK_a and log P challenges. Unfortunately only a subset of compounds in pK_a datasets were suitable for the potentiometric log P measurements. Still for the subset of compounds that were common in both challenges comparing prediction performance can lead to beneficial insights especially for physical modeling techniques if there are common aspects that are beneficial or detrimental to prediction performance. For example, in SAMPL6 pK_a and log P Challenges COSMO-RS and EC-RISM solvation models achieved good performance. Having a variety of experimental measurements of physicochemical properties can also help to identify sources of errors. For example, dominant microstates determined for pK_a challenge can provide information to check if correct tautomers are modeling in a log P or log D challenge. pK_a prediction is a requirement for log D prediction and experimental pK_a values can help diagnosing the source of errors in log D predictions better. The physical challenges in SAMPL7, which is currently running with a deadline of September 30th, 2020, follow this principle and include both pK_a, log P, and membrane permeability properties for a set of monoprotic compounds. We hope that future pK_a challenges can focus on multiprotic drug-like compounds with microscopic pK_a measurements for an in-depth analysis.

4 Conclusion

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

The experimental dataset consisted of spectrophotometric measurements of 24 molecules and some of which were multiprotic. There was also experimental data on the dominant microstate sequence of a subset of the challenge molecules, but not direct microscopic pK_a measurements. We have performed a comparative analysis of methods represented in the blind challenge in terms of both macroscopic and microscopic pK_a prediction performance avoiding any assumptions about the experimental pK_as.

As the majority of the experimental data was macroscopic pK_a values, we had to utilize Hungarian matching to assign predicted and experimental values before calculating accuracy and correlation statistics. In addition to evaluating error in predicted pK_a values, we also reported the macroscopic pK_a errors that were not captured by the match between experimental and predicted pK_a values. These were extra or missing pK_a predictions which are important indicators that predictions are failing to capture the correct ionization states.

We utilized the experimental dominant microstate sequence data of 8 molecules to evaluate microscopic pK_a predictions in more detail. This experimental data allowed us to use microstate-based matching for evaluating the accuracy of microscopic pK_a values in a more realistic way. We have determined that QM and LFER predictions had lower accuracy in determining the dominant tautomer of the charged microstates than the neutral states. For both macroscopic and microscopic pK_a predictions we have determined methods that were consistently well-performing according to multiple statistical metrics. Focusing on the comparison of molecules instead of methods for macroscopic pK_a prediction accuracy indicated molecules with sulfur-containing heterocycles, iodo, and bromo groups suffered from lower pK_a prediction accuracy.

The overall performance level observed for pK_a predictions in this challenge is concerning for the application of pK_a prediction methods in computer-aided drug design. Many methods for capturing target affinities and physicochemical properties rely on pK_a predictions for determining relevant protonation states and the free energy penalty of such states. 1 unit of p$K_a$error is an optimistic estimate of current macroscopic pK_a predictions for drug-like molecules based on SAMPL6 Challenge where errors in predicting the correct number of ionization states or determining the correct dominant microstate were also common to many methods. In the absence of other sources of errors, we showed that 1 unit over- or underestimation of the pK_a of a ligand can cause significant errors in the overall binding affinity calculation due to errors in multiple protonation state correction factor.

All information regarding the challenge structure, experimental data, blind prediction submission sets, and evaluation of methods are available in the SAMPL6 GitHub Repository for future follow up analysis and to serve as a benchmark dataset for testing methods.

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

5 Code and data availability

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

6 Overview of supplementary information

Contents of the Supplementary Information:

- TABLE S1: SMILES and InChI identifiers of SAMPL6 pK_a Challenge molecules.
- TABLE S2: Evaluation statistics calculated for all macroscopic pK_a prediction submissions based on Hungarian match for 24 molecules.
- TABLE S3: Evaluation statistics calculated for all microscopic pK_a prediction submissions based on Hungarian match for 8 molecules with NMR data.
- TABLE S4: Evaluation statistics calculated for all microscopic pK_a prediction submissions based on microstate match for 8 molecules with NMR data.
- FIGURE S1: Dominant microstates of 8 molecules were determined based on NMR measurements.
- FIGURE S2: MAE of macroscopic pK_a predictions of each molecule did not show any significant correlation with any molecular descriptor.
- FIGURE S3: The value of macroscopic pK_a was not a factor affecting prediction error seen in SAMPL6 Challenge according to the analysis with Hungarian matching.
- 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.

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

- SAMPL6-pKa-chemical-identifiers-table.csv
- macroscopic-pKa-statistics-24mol-hungarian-match.csv
- microscopic-pKa-statistics-8mol-hungarian-match-table.csv
- microscopic-pKa-statistics-8mol-microstate-match-table.csv
- experimental-microstates-of-8mol-based-on-NMR.csv
- enumerate-microstates-with-Epik-and-OpenEye-QUACPAC.ipynb
- molecule_ID_and_SMILES.csv

7 Author Contributions

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

8 Acknowledgments

We would like to acknowledge the infrastructure and website support of Mike Chui that allowed a seamless collection of challenge submissions. Mike Chui also provided assistance with constructing a submission validation script to ensure all submissions adhered to the machine-readable format. We are grateful to Kiril Lanevskij for suggesting the Hungarian algorithm for matching experimental and predicted pK_a values. We would like to thank Thomas Fox for providing MoKa reference calculations. We acknowledge Caitlin Bannan for guidance on defining a working microstate definition for the challenge and guidance for designing the challenge. We thank Brad Sherborne for his valuable insights at the conception of the pK_a challenge and connecting us with Timothy Rhodes and Dorothy Levorse who were able to provide resources and expertise for experimental measurements performed at MRL. We acknowledge Paul Czodrowski who provided feedback on multiple stages of this work: challenge construction, purchasable compound selection, and evaluation. MI, JDC, and DLM gratefully acknowledge support from NIH grant R01GM124270 supporting the SAMPL Blind Challenges. MI, ASR, AR, and JDC acknowledge support from the Sloan Kettering Institute. JDC acknowledges support from NIH grant P30 CA008748. DLM appreciates financial support from the National Institutes of Health (1R01GM108889-01) and the National Science Foundation (CHE 1352608). MI acknowledges Doris J. Hutchinson Fellowship. MI, ASR, AR, and JDC are grateful to OpenEye Scientific for providing a free academic software license for use in this work. MI, ASR, AR, and JDC thank Janos Fejervari and ChemAxon team that gave us permission to include ChemAxon/Chemicalize pK_a predictions as a reference prediction in challenge analysis.

9 Disclaimers

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

10 Disclosures

JDC was a member of the Scientific Advisory Board for Schrödinger, LLC during part of this study. JDC and DLM are current members of the Scientific Advisory Board of OpenEye Scientific Software, and DLM is an Open Science Fellow with Silicon Therapeutics. The Chodera laboratory receives or has received funding from multiple sources, including the National Institutes of Health, the National Science Foundation, the Parker Institute for Cancer Immunotherapy, Relay Therapeutics, Entasis Therapeutics, Vir Biotechnology, Silicon Therapeutics, EMD Serono (Merck KGaA), AstraZeneca, Vir Biotechnology, XtalPi, the Molecular Sciences Software Institute, the Starr Cancer Consortium, the Open Force Field Consortium, Cycle for Survival, a Louis V. Gerstner Young Investigator Award, The Einstein Foundation, and the Sloan Kettering Institute. A complete list of funding can be found at http://choderalab.org/funding.

References

[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.

[2] **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.

[3] **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.

[4] **Darvey IG**. The Assignment of pKa Values to Functional Groups in Amino Acids. Biochemical Education. 1995 Apr; 23(2):80–82. doi: 10.1016/0307-4412(94)00150-N.

[5] **Bodner GM**. Assigning the pKa's of Polyprotic Acids. Journal of Chemical Education. 1986 Mar; 63(3):246. doi: 10.1021/ed063p246.

[6] **Murray R**. Microscopic Equilibria. Analytical Chemistry,. 1995 Aug; p. 1.

[7] **Işık M**, Levorse D, Rustenburg AS, Ndukwe IE, Wang H, Wang X, Reibarkh M, Martin GE, Makarov AA, Mobley DL, Rhodes T, Chodera JD. pKa 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.

[8] **Bochevarov AD**, Watson MA, Greenwood JR, Philipp DM. Multiconformation, Density Functional Theory-Based p K_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.

[9] **Selwa E**, Kenney IM, Beckstein O, Iorga BI. SAMPL6: Calculation of Macroscopic pKa Values from Ab Initio Quantum Mechanical Free Energies. Journal of Computer-Aided Molecular Design. 2018 Oct; 32(10):1203–1216. doi: 10.1007/s10822-018-0138-6.

[10] **Pickard FC**, König G, Tofoleanu F, Lee J, Simmonett AC, Shao Y, Ponder JW, Brooks BR. Blind Prediction of Distribution in the SAMPL5 Challenge with QM Based Protomer and pK a Corrections. Journal of Computer-Aided Molecular Design. 2016 Nov; 30(11):1087–1100. doi: 10.1007/s10822-016-9955-7.

[11] **Bannan CC**, Mobley DL, Skillman AG. SAMPL6 Challenge Results from $$pK_a$$ Predictions Based on a General Gaussian Process Model. Journal of Computer-Aided Molecular Design. 2018 Oct; 32(10):1165–1177. doi: 10.1007/s10822-018-0169-z.

[12] **Işık M**, Levorse D, Mobley DL, Rhodes T, Chodera JD. Octanol–Water Partition Coefficient Measurements for the SAMPL6 Blind Prediction Challenge. Journal of Computer-Aided Molecular Design. 2020 Apr; 34(4):405–420. doi: 10.1007/s10822-019-00271-3.

[13] **Işık M**, Bergazin TD, Fox T, Rizzi A, Chodera JD, Mobley DL. Assessing the Accuracy of Octanol–Water Partition Coefficient Predictions in the 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.

[14] Special Issue: SAMPL6 (Statistical Assessment of the Modeling of Proteins and Ligands); October 2018. Volume 32, Issue 10. Journal of Computer-Aided Molecular Design.

[15] **Shelley JC**, Cholleti A, Frye LL, Greenwood JR, Timlin MR, Uchimaya M. Epik: A Software Program for pK a Prediction and Protonation State Generation for Drug-like Molecules. Journal of Computer-Aided Molecular Design. 2007 Dec; 21(12):681–691. doi: 10.1007/s10822-007-9133-z.

[16] QUACPAC Toolkit 2017.Feb.1;. OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com.

[17] **Fraczkiewicz R**. In Silico Prediction of Ionization. In: *Reference Module in Chemistry, Molecular Sciences and Chemical Engineering* Elsevier; 2013.doi: 10.1016/B978-0-12-409547-2.02610-X.

[18] **Kuhn HW**. The Hungarian Method for the Assignment Problem. Naval Research Logistics Quarterly. 1955 Mar; 2(1-2):83–97. doi: 10.1002/nav.3800020109.

[19] **Munkres J**. Algorithms for the Assignment and Transportation Problems. J SIAM. 1957 Mar; 5(1):32–28.

[20] 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.

[21] **Gunner MR**, Murakami T, Rustenburg AS, Işık M, Chodera JD. Standard State Free Energies, Not pKas, Are Ideal for Describing Small Molecule Protonation and Tautomeric States. Journal of Computer-Aided Molecular Design. 2020 May; 34(5):561–573. doi: 10.1007/s10822-020-00280-7.

[22] OpenEye pKa Prospector;. OpenEye Scientific Software, Santa Fe, NM. Accessed on Jan 23, 2018. https://www.eyesopen.com/pka-prospector.

[23] ACD/pKa GALAS (ACD/Percepta Kernel v1.6);. Advanced Chemistry Development, Inc., Toronto, ON, Canada, 2018. https://www.acdlabs.com/products/percepta/predictors/pKa/.

[24] ACD/pKa Classic (ACD/Percepta Kernel v1.6);. Advanced Chemistry Development, Inc., Toronto, ON, Canada, 2018. https://www.acdlabs.com/products/percepta/predictors/pKa/.

[25] Simulations Plus ADMET Predictor v8.5;. Simulations Plus, Lancaster, CA, 2018. https://www.simulations-plus.com/software/admetpredictor/physicochemical-biopharmaceutical/.

[26] Chemicalize v18.23 (ChemAxon MarvinSketch v18.23);. ChemAxon, Budapest, Hungary, 2018. https://docs.chemaxon.com/display/docs/pKa+Plugin.

[27] **Milletti F**, Storchi L, Sforna G, Cruciani G. New and Original p K_a Prediction Method Using Grid Molecular Interaction Fields. Journal of Chemical Information and Modeling. 2007 Nov; 47(6):2172–2181. doi: 10.1021/ci700018y.

[28] MoKa;. Molecular Discovery, Hertfordshire, UK, 2018. https://www.moldiscovery.com/software/moka/.

[29] **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.

[30] **Bochevarov AD**, Harder E, Hughes TF, Greenwood JR, Braden DA, Philipp DM, Rinaldo D, Halls MD, Zhang J, Friesner RA. Jaguar: A High-Performance Quantum Chemistry Software Program with Strengths in Life and Materials Sciences. International Journal of Quantum Chemistry. 2013 Sep; 113(18):2110–2142. doi: 10.1002/qua.24481.

[31] **Tielker N**, Eberlein L, Güssregen S, Kast SM. The SAMPL6 Challenge on Predicting Aqueous pKa Values from EC-RISM Theory. Journal of Computer-Aided Molecular Design. 2018 Oct; 32(10):1151–1163. doi: 10.1007/s10822-018-0140-z.

[32] **Klamt A**, Eckert F, Diedenhofen M, Beck ME. First Principles Calculations of Aqueous p K_a Values for Organic and Inorganic Acids Using COSMO-RS Reveal an Inconsistency in the Slope of the p K_a Scale. The Journal of Physical Chemistry A. 2003 Nov; 107(44):9380–9386. doi: 10.1021/jp034688o.

[33] **Eckert F**, Klamt A. Accurate Prediction of Basicity in Aqueous Solution with COSMO-RS. Journal of Computational Chemistry. 2006 Jan; 27(1):11–19. doi: 10.1002/jcc.20309.

[34] **Pracht P**, Wilcken R, Udvarhelyi A, Rodde S, Grimme S. High Accuracy Quantum-Chemistry-Based Calculation and Blind Prediction of Macroscopic pKa Values in the Context of the SAMPL6 Challenge. Journal of Computer-Aided Molecular Design. 2018 Oct; 32(10):1139–1149. doi: 10.1007/s10822-018-0145-7.

[35] **Prasad S**, Huang J, Zeng Q, Brooks BR. An Explicit-Solvent Hybrid QM and MM Approach for Predicting pKa of Small Molecules in SAMPL6 Challenge. Journal of Computer-Aided Molecular Design. 2018 Oct; 32(10):1191–1201. doi: 10.1007/s10822-018-0167-1.

[36] **Robert Fraczkiewicz MW**, SAMPL6 pKa Challenge: Predictions of ionization constants performed by the S+pKa method implemented in ADMET Predictor software; February 22, 2018. The Joint D3R/SAMPL Workshop 2018. https://drugdesigndata.org/about/d3r-2018-workshop.

[37] **Mansouri K**, Cariello NF, Korotcov A, Tkachenko V, Grulke CM, Sprankle CS, Allen D, Casey WM, Kleinstreuer NC, Williams AJ. Open-Source QSAR Models for pKa Prediction Using Multiple Machine Learning Approaches. Journal of Cheminformatics. 2019 Dec; 11(1). doi: 10.1186/s13321-019-0384-1.

[38] **Baltruschat M**, Czodrowski P. Machine Learning Meets pKa [Version 2; Peer Review: 2 Approved]. F1000Research. 2020; 9 (Chem Inf Sci)(113). doi: 10.12688/f1000research.22090.2.

[39] **Hunt P**, Hosseini-Gerami L, Chrien T, Plante J, Ponting DJ, Segall M. Predicting p K_a Using a Combination of Semi-Empirical Quantum Mechanics and Radial Basis Function Methods. Journal of Chemical Information and Modeling. 2020 Jun; 60(6):2989–2997. doi: 10.1021/acs.jcim.0c00105.

[40] **Zdrazil B**, Guha R. The Rise and Fall of a Scaffold: A Trend Analysis of Scaffolds in the Medicinal Chemistry Literature. Journal of Medicinal Chemistry. 2018 Jun; 61(11):4688–4703. doi: 10.1021/acs.jmedchem.7b00954.

[41] **Ertl P**, Altmann E, McKenna JM. The Most Common Functional Groups in Bioactive Molecules and How Their Popularity Has Evolved over Time. Journal of Medicinal Chemistry. 2020 Aug; 63(15):8408–8418. doi: 10.1021/acs.jmedchem.0c00754.

[42] 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*.

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)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)Cc2nnc(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	c1cc2cccnc2c(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)CNc2c3ccccc3ncn2	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)c3ccccc3	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)Nc2c3ccccc3ncn2.Cl	InChI=1S/C15H13N3O.ClH/c1-19-12-6-4-5-11(9-12)18-15-13-7-2-3-8-14(13)16-10-17-15;/h2-10H,1H3,(H,16,17,18);1H
SM10	c1ccc(cc1)C(=O)NCC(=O)Nc2nc3ccccc3s2	InChI=1S/C16H13N3O2S/c20-14(10-17-15(21)11-6-2-1-3-7-11)19-16-18-12-8-4-5-9-13(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,(H2,12,13,14)
SM12	c1ccc2c(c1)c(ncn2)Nc3cccc(c3)Cl.Cl	InChI=1S/C14H10ClN3.ClH/c15-10-4-3-5-11(8-10)18-14-12-6-1-2-7-13(12)16-9-17-14;/h1-9H,(H,16,17,18);1H
SM13	Cc1cccc(c1)Nc2c3cc(c(cc3ncn2)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)n2cnc3c2ccc(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)21-16(22)8-10-3-5-12(18)13(19)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,(H2,20,21,22,23)
SM22	c1cc2c(cc(c(c2nc1)O)I)I	InChI=1S/C9H5I2NO/c10-6-4-7(11)9(13)8-5(6)2-1-3-12-8/h1-4,13H
SM23	CCOC(=O)c1ccc(cc1)Nc2cc(nc(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(ncnc3oc2c4ccc(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)

11 Supplementary Information

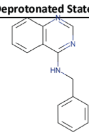
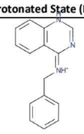
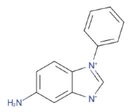
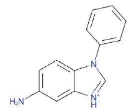
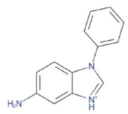
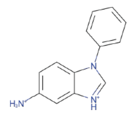
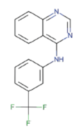
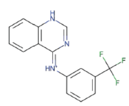
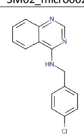
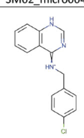
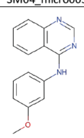
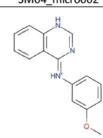
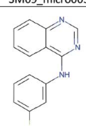
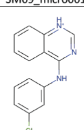
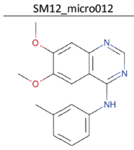
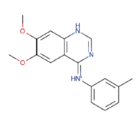
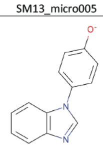
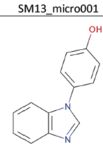
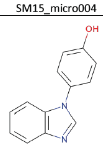
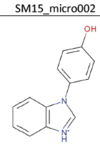
Microstate ID of Deprotonated State (A)	Microstate ID of Protonated State (HA)	Molecule ID	pKa (exp)	pKa SEM (exp)	pKa ID	Microstate identification source
 SM07_micro004	 SM07_micro006	SM07	6.08	0.01	SM07_pKa1	NMR measurement
 SM14_micro001	 SM14_micro002	SM14	5.3	0.01	SM14_pKa2	NMR measurement
 SM14_micro002	 SM14_micro003	SM14	2.58	0.01	SM14_pKa1	NMR measurement
 SM02_micro002	 SM02_micro004	SM02	5.03	0.01	SM02_pKa1	Estimated based on SM07 NMR measurement
 SM04_micro003	 SM04_micro002	SM04	6.02	0.01	SM04_pKa1	Estimated based on SM07 NMR measurement
 SM09_micro003	 SM09_micro001	SM09	5.37	0.01	SM09_pKa1	Estimated based on SM07 NMR measurement
 SM12_micro012	 SM12_micro006	SM12	5.28	0.01	SM12_pKa1	Estimated based on SM07 NMR measurement
 SM13_micro005	 SM13_micro001	SM13	5.77	0.01	SM13_pKa1	Estimated based on SM07 NMR measurement
 SM15_micro004	 SM15_micro002	SM15	8.94	0.01	SM15_pKa2	Estimated based on SM14 NMR measurement
 SM15_micro002	 SM15_micro003	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 [7]. A CSV version of this table can be found in *SAMPL6-supplementary-documents.tar.gz*.

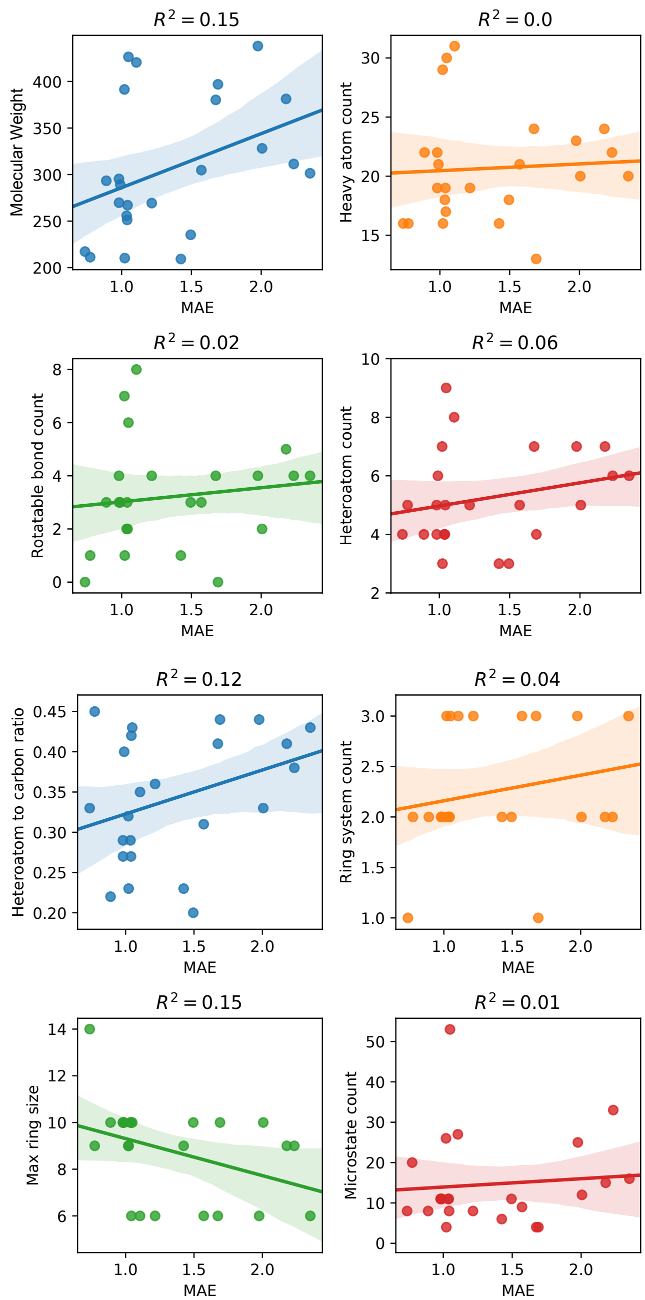


Figure S2. MAE of macroscopic pK_a predictions of each molecule did not show any significant correlation with any molecular descriptor. Plots show regression lines, 96% confidence intervals of the regression lines, and R_2. The following molecular descriptors were calculated using OpenEye OEMolProp Toolkit [42].

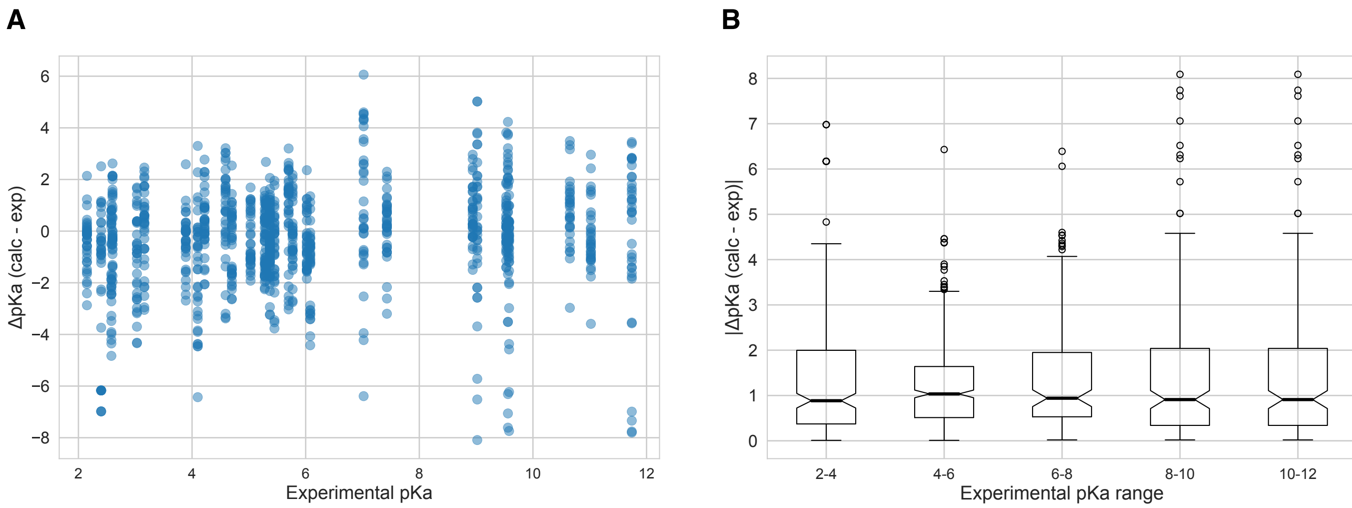


Figure S3. The value of macroscopic pK_as was not a factor affecting prediction error seen in SAMPL6 Challenge according to the analysis with Hungarian matching. There was not clear trend between pK_a prediction error and the true pK_a error. Very high and very low pK_a values have similar inaccuracy 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.

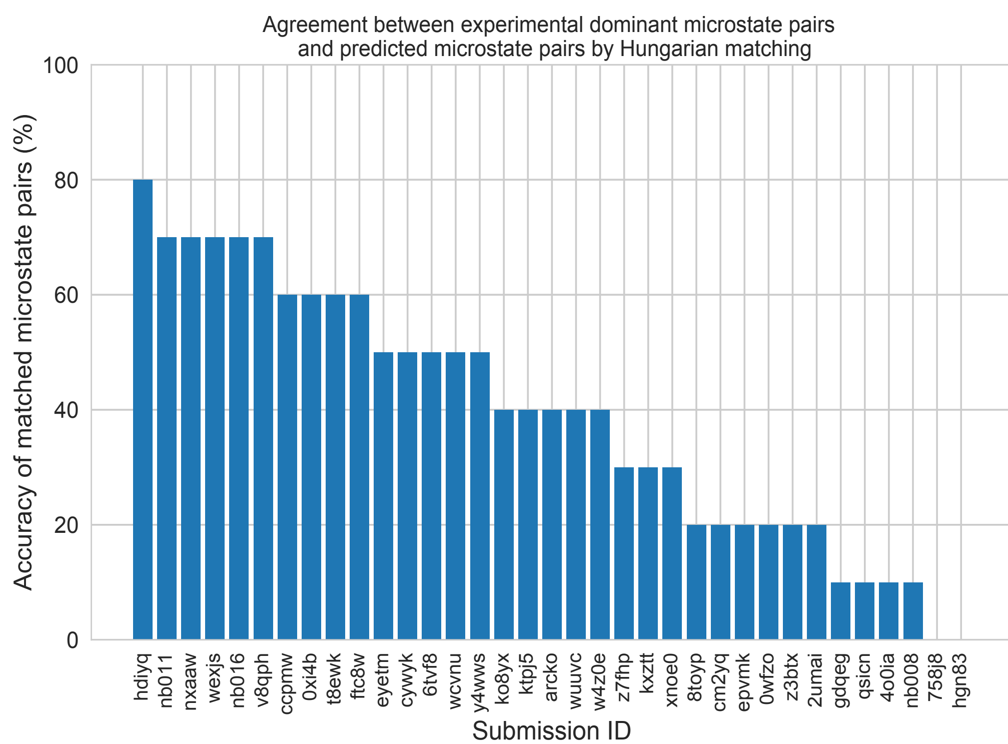


Figure S4. There was low agreement between experimental dominant microstate pairs and the predicted microstate pairs selected by Hungarian algorithm for microscopic pK_a predictions. This analysis could only be performed for 8 molecules with NMR data. Hungarian matching algorithm which matches predicted and experimental values considering only the closeness of the numerical value of pK_a and it often leds 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_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_as	Unmatched pred. pK_as [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, 1.4.0]	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_as	Unmatched pred. pK_as [2,12]
nb011	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
hdiyq	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
epvmk	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
xnoe0	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
gdqeg	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
4o0ia	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
nb008	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
ccpmw	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
0xi4b	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
cywyk	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
ftc8w	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
nxaaw	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
nb016	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
kxztt	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
eyetm	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
cm2yq	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
2umai	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
ko8yx	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
wuuvc	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
ktpj5	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
z7fhp	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
arcko	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
y4wws	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
wcvnu	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
8toyp	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
qsicn	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
wexjs	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
v8qph	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
w4z0e	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
6tvf8	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
0wfzo	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
t8ewk	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
z3btx	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
758j8	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
hgn83	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_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*.

Update this table with dominant microstate accuracy

Submission ID	RMSE	MAE	ME	R^2	m	Kendall's Tau	Unmatched exp. pK_as	Unmatched pred. pK_as [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
hdiyq	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
6tvf8	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
0xi4b	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
4o0ia	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
2umai	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
kxztt	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
z3btx	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