SAMPL6 pKa reference calculations using Epik and Jaguar

Ariën S. Rustenburg[1,2], **Mehtap Isik**[1], **Patrick B. Grinaway**[1,2], **Andrea Rizzi**[1,3], **Marilyn R Gunner**[5], **John D. Chodera**[1*]

[1]Computational and Systems Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY 10065; [2]Graduate Program in Physiology, Biophysics, and Systems Biology, Weill Cornell Medical College, New York, NY 10065; [3]Tri-Institutional Training Program in Computational Biology and Medicine, New York, NY 10065; [4]Schrödinger LLC, New York, NY 10036; [5]Department of Physics, City College of New York, New York, NY 10031

***For correspondence:**
john.chodera@choderalab.org (JDC)

NOTES: Objectives for this manuscript

- Provide a baseline for the expectation of performance of empirical models (Epik) and DFT models (Jaguar) on kinase inhibitor-like molecules and molecular fragments; draw comparisons to earlier benchmarks for these tools
- Describe basic concepts, along with strengths and weaknesses of each approach (Epik and Jaguar)
- Lessons learned for participating in the challenge
  - How should we predict macroscopic pKas from microscopic pKas? Explain possibilities and justify choice
    - Sequential titration (Epik)
    - <n protons>/charge vs pH inflection points: Matches electrochemical titration, but not UV-metric
    - Is there a way to predict what UV-metric would observe from populations vs pH?
- For microscopic pKa prediction: Epik may enumerate a different set of microstates than other tools (we proposed these as newly labeled states)
- Epik does not return state populations below a certain threshold, so we could not consider states proposed by SAMPL6 [Check this]
- For worst predicted compounds, provide an explanation as to why the methods may perform poorly (descriptors?) [may not have enough data points to draw conclusion]

Figures and Tables

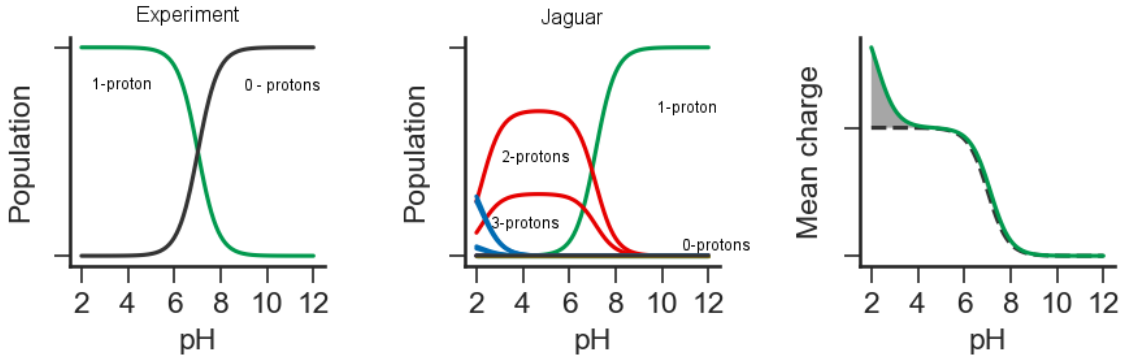


Figure 1. Comparing microscopic Jaguar pKa predictions to experiment . Jaguar was used to predict microscopic pKas (denoted as Type I predictions in SAMPL6), from which microstate populations (middle) and a charge titration curve (right) were derived. The example shows predictions for molecule SM03. The population of every microstate (between 0 to 1) is plotted in a color indicative of the number of protonations relative to the most deprotonated state observed. Black means deprotonated, green means 1 proton, red means 2 protons, and blue means 3 protons. In the charge-titration curve, the experimental prediction is shown as a dashed line. Since there is no microscopic information in the experimental details, the offset with the Jaguar charge curve is derived from aligning the two curves at integer charge offsets and finding the best fit. The mean charge is therefore only a relative number. Tick marks indicate integer charge values.

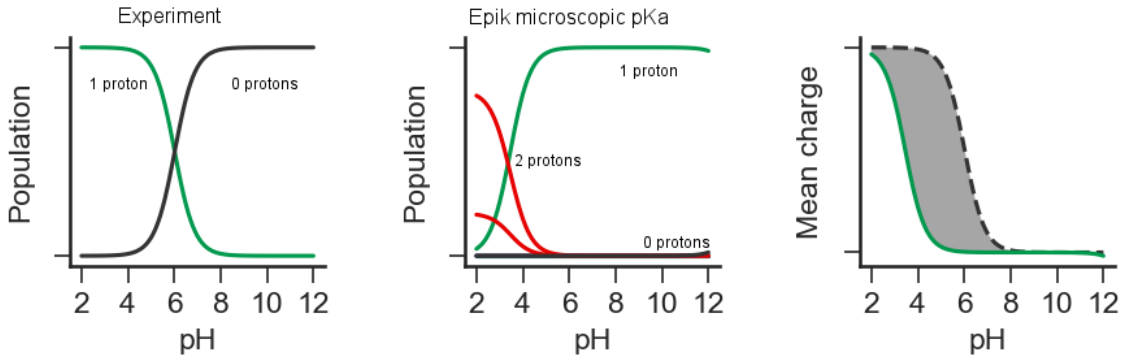


Figure 2. Comparing microscopic Epik pKa predictions to experiment . Epik was used to predict microscopic pKas (denoted as Type I predictions in SAMPL6), from which microstate populations (middle) and a charge titration curve (right) were derived. The example shows predictions for molecule SM04. The population of every microstate (between 0 to 1) is plotted in a color indicative of the number of protonations relative to the most deprotonated state observed. Black means deprotonated, green means 1 proton, red means 2 protons. In the charge-titration curve, the experimental prediction is shown as a dashed line. Since there is no microscopic information in the experimental details, the offset with the Epik charge curve is derived from aligning the two curves at integer charge offsets and finding the best fit. The mean charge is therefore only a relative number. Tick marks indicate integer charge values.

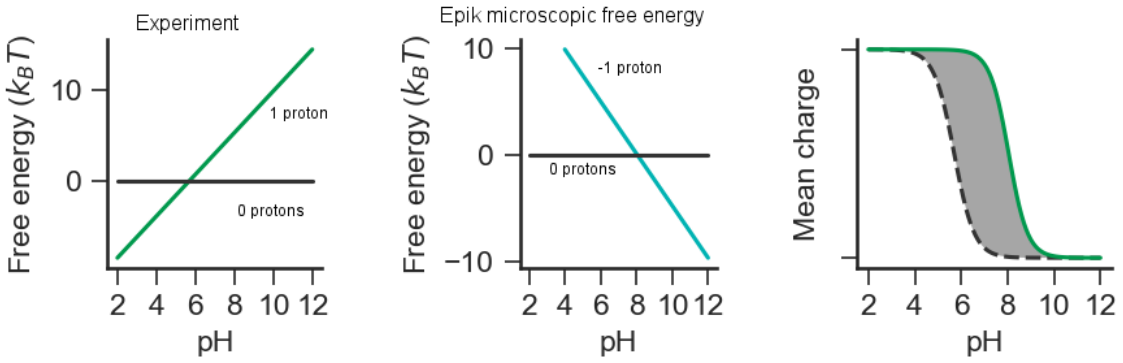


Figure 3. Comparing microscopic Epik free energies to experiment . Epik was used to predict microscopic free energies (denoted as Type II predictions in SAMPL6) (middle), from which a charge titration curve (right) was derived. The example shows predictions for molecule SM20. The population of every microstate (between 0 to 1) is plotted in a color indicative of the number of protonations relative to the most deprotonated state observed. Black means 0 protons, green means +1 proton, and cyan means -1 proton. In the charge-titration curve, the experimental prediction is shown as a dashed line. Since there is no microscopic information in the experimental details, the offset with the Epik charge curve is derived from aligning the two curves at integer charge offsets and finding the best fit. The mean charge is therefore only a relative number. Tick marks indicate integer charge values.

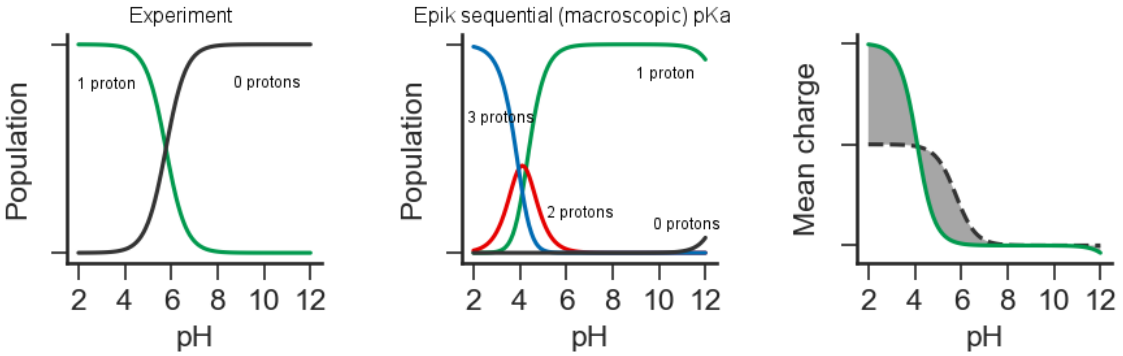


Figure 4. Comparing macroscopic Epik pKa values to experiment . Epik was used to predict macroscopic pKa values (denoted as Type III predictions in SAMPL6), from which macrostate populations (middle) and a charge titration curve (right) were derived. The example shows predictions for molecule SM13. The population of every microstate (between 0 to 1) is plotted in a color indicative of the number of protonations relative to the most deprotonated state observed. Black means deprotonated, green means 1 proton, red means 2 protons, and blue means 3 protons. In the charge-titration curve, the experimental prediction is shown as a dashed line. Since there is no microscopic information in the experimental details, the offset with the Epik charge curve is derived from aligning the two curves at integer charge offsets and finding the best fit. The mean charge is therefore only a relative number. Tick marks indicate integer charge values.

Table 1. Overall performance of each method using either the closest pKa matching, Hungarian pKa matching, and the area between the experimental and predicted macroscopic charge titration curve

	Closest			Hungarian			Titration curve
	RMSE	MAE	R^2	RMSE	MAE	R^2	Δ area
Epik sequential							
Epik micropka							
Jaguar micropKa							

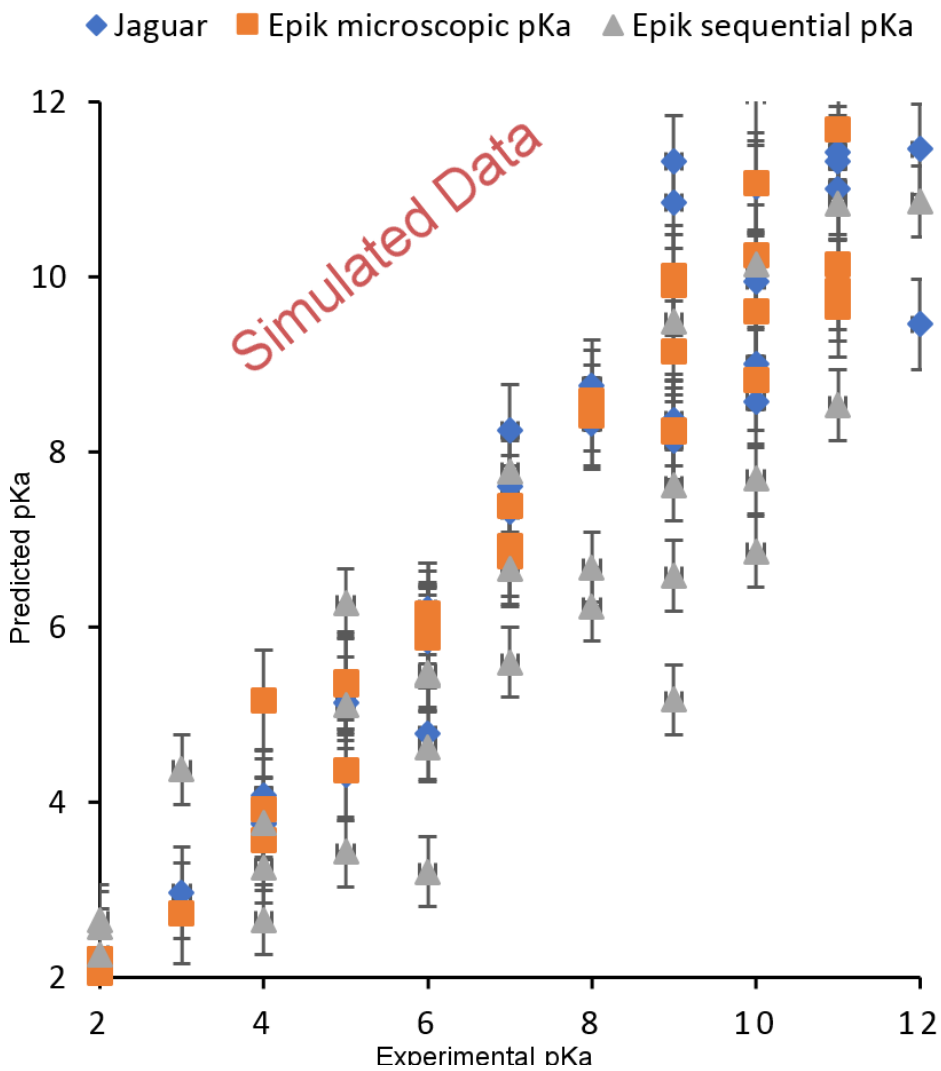


Figure 5. Comparison of the entire set of experimental pKa values to all predictions. Microscopic pKa values for Jaguar (diamonds) and Epik (squares), as well as macroscopic Epik pKa values (triangles) are compared to the experimental dataset. The pKa values were matched using algorithm 1. RMSE and r-squared values can be found in table 1. An alternative alignment of pKa values using **??** is available as a supplementary figure.

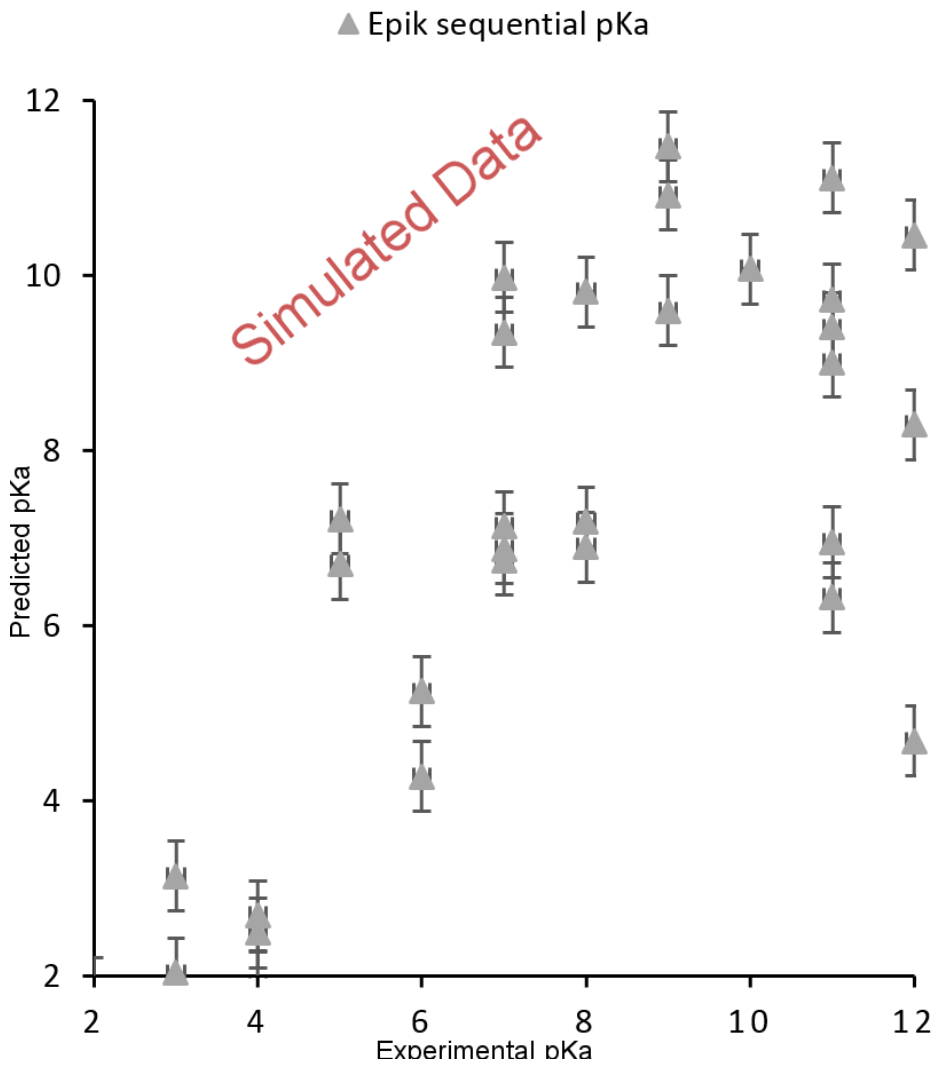


Figure 6. Epik sequential pKa values compared to experiment, by sequential alignment of pKa values.

Show RMSE in addition to (or instead of) r^2.

Comparison of the Epik sequential scan with entire experimental data set is shown, restricting to only sequentially mapping predicted pKa values to experiment.

Add columns for the other two algorithms

Table 2. Macroscopic pKas computed via Epik's sequential scan procedure. Experimental and computed pKa values matched using algorithm 1.

Molecule	Experiment ± SEM	Epik closest ± SEM
SM01	9.53±0.01	9±1
SM02	5.03±0.01	4±2
SM03	7.02±0.01	7±2
SM04	6.02±0.01	6±1
SM05	4.59±0.01	5±2
SM06	3.03±0.04	2±1
SM06	11.74±0.01	10±2
SM07	6.08±0.01	6±1
SM08	4.22±0.01	4±2
SM09	5.37±0.01	4±2
SM10	9.02±0.01	9±2
SM11	3.89±0.01	4±2
SM12	5.28±0.01	4±2
SM13	5.77±0.01	4±2
SM14	2.58±0.01	3±2
SM14	5.30±0.01	6±2
SM15	8.94±0.01	9±1
SM15	4.70±0.01	6±1
SM16	5.37±0.01	4.7±0.9
SM16	10.65±0.01	10±2
SM17	3.16±0.01	4.9±0.7
SM18	9.58±0.03	9±2
SM18	2.15±0.02	2.5±2
SM18	11.02±0.04	11±2
SM19	9.56±0.02	9±2
SM20	5.70±0.03	8±2
SM21	4.10±0.01	4±2
SM22	7.43±0.01	6.8±0.9
SM22	2.40±0.02	4±1
SM23	5.45±0.01	6.0±0.8
SM24	2.60±0.01	4±2

Add hungarian algorithm to SI.

Table 3. Microscopic pKa per molecule for each method as matched by algorithm 1, compared to experiment.

	Experimental pKa	Epik pKa	Epik pKa uncertainty	Epik error	Jaguar pKa	Jaguar error
SM01 pKa1						
SM01 pKa2						
SM02 pKa 1						
SM03 pKa 1						

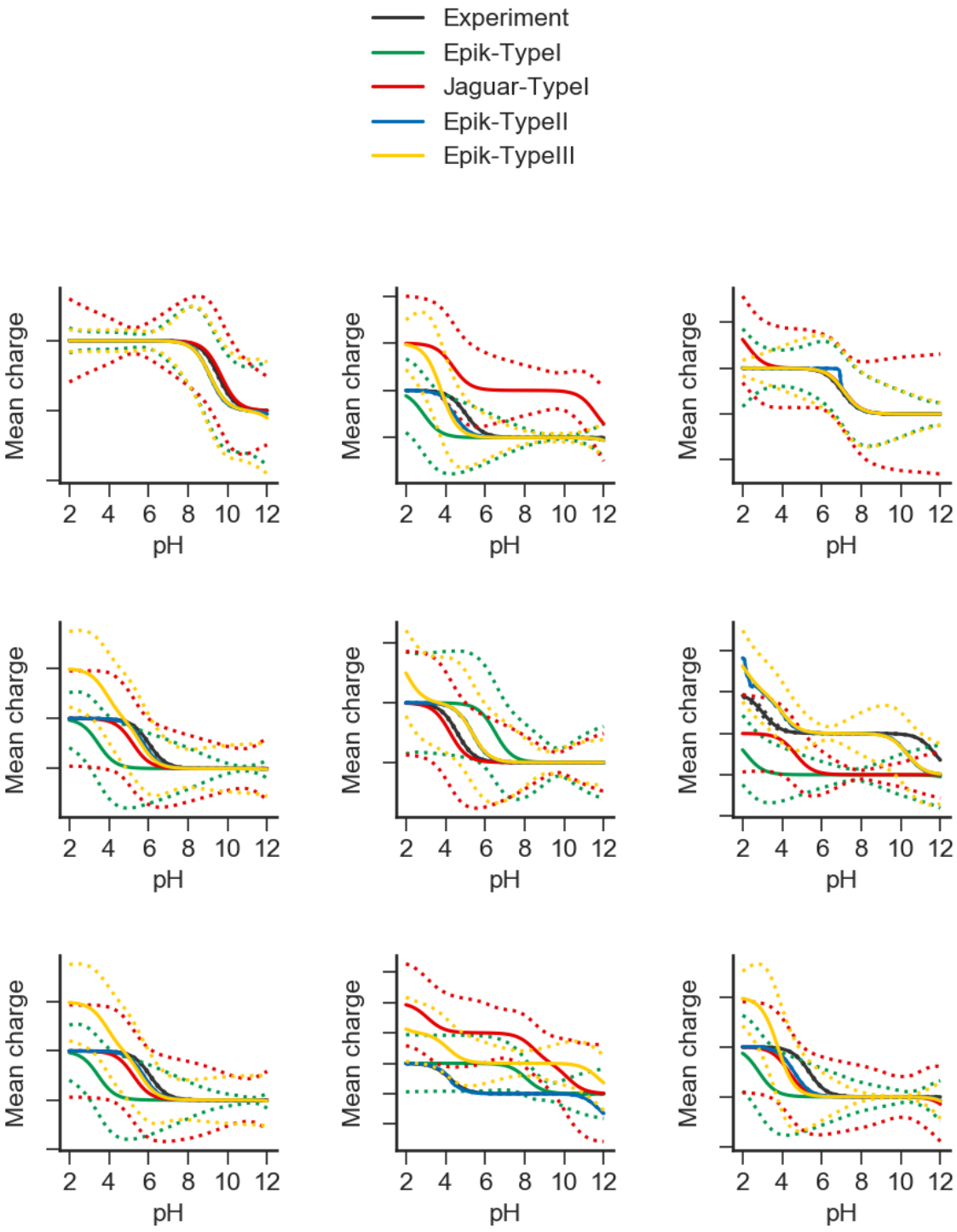


Figure 7. Bootstrap titration curves for each method compared to experiment for molecules SM01-SM09. The titration curve for each method is shown as a solid line, and 96 % confidence intervals from bootstrap have been shown as dotted lines. A numerical comparison to experiment is presented in table 4.

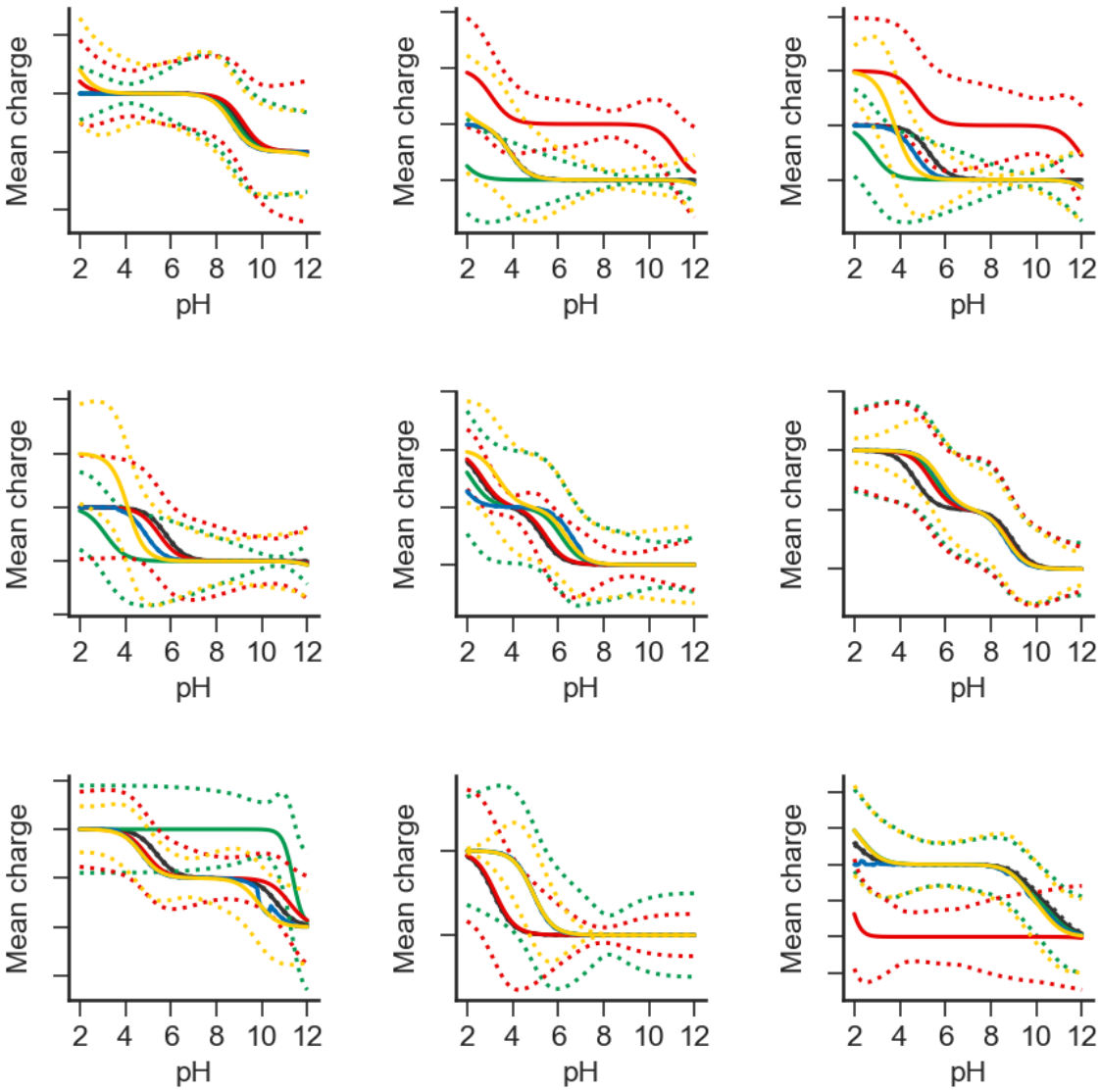


Figure 8. Bootstrap titration curves for each method compared to experiment for molecules SM10-SM18. The titration curve for each method is shown as a solid line, and 96 % confidence intervals from bootstrap have been shown as dotted lines. A numerical comparison to experiment is presented in table 4.

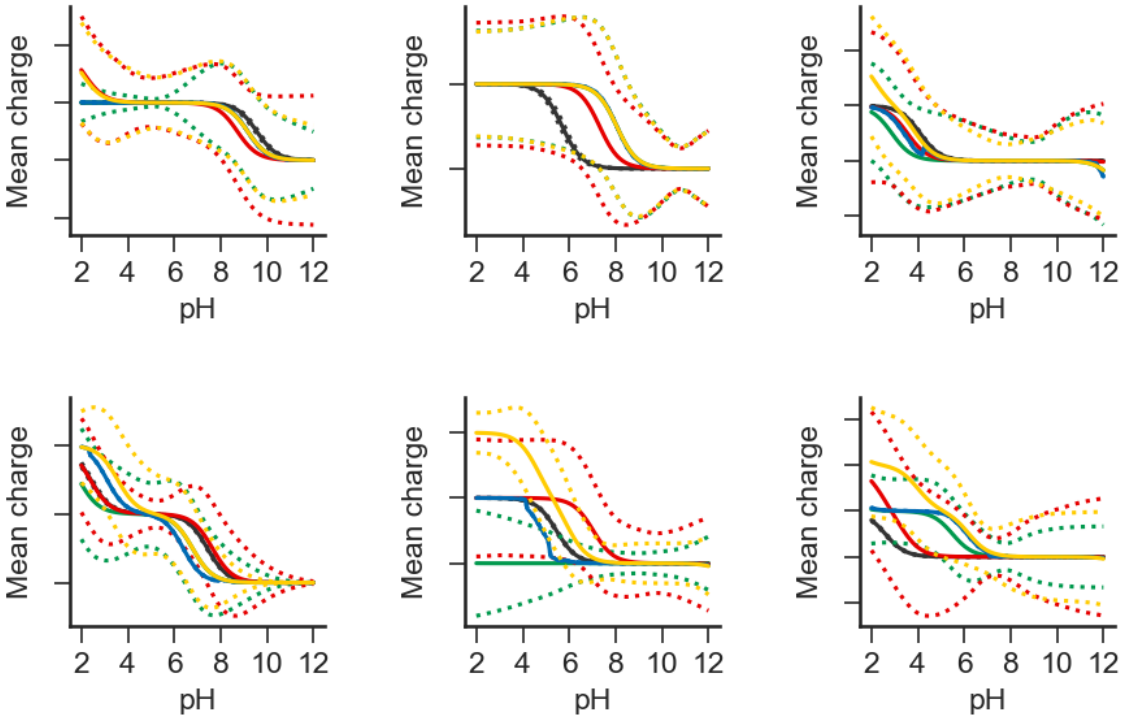


Figure 9. Bootstrap titration curves for each method compared to experiment for molecules SM19-SM24. The titration curve for each method is shown as a solid line, and 96 % confidence intervals from bootstrap have been shown as dotted lines. A numerical comparison to experiment is presented in table 4.

ASR:Need to add new data, and bootstrap error bars.

Table 4. Difference in area between derived macroscopic charge titration curves (Δ area) for each method. The macroscopic charge curve was constructed for each of the methods, and after alignment with the experimental curve by an integer offset, the area between the curves is reported.

Molecule	Epik microscopic pKa	Epik microscopic population	Epik scan	Jaguar microscopic pKa
SM01	0.43	0.44	0.46	0.15
SM02	2.12	0.59	1.99	1.21
SM03	0.10	0.29	0.09	0.57
SM04	2.55	0.26	2.20	0.81
SM05	1.93	0.74	1.08	0.51
SM06	2.60	2.52	2.62	2.09
SM07	2.61	0.32	2.25	0.84
SM08	1.27	0.57	0.63	5.82
SM09	2.53	0.78	2.32	1.09
SM10	0.17	0.34	0.62	0.28
SM11	1.84	0.06	0.14	1.68
SM12	2.48	0.71	2.26	0.88
SM13	2.64	0.88	2.80	0.30
SM14	1.07	1.74	1.75	0.25
SM15	1.07	1.32	1.31	0.85
SM16	4.36	1.10	1.44	0.97
SM17	1.71	1.71	1.71	0.07
SM18	0.74	1.02	1.00	NaN
SM19	0.38	0.39	0.73	1.22
SM20	2.34	2.34	2.34	1.60
SM21	1.26	0.70	0.55	1.68
SM22	0.94	1.62	1.64	0.23
SM23	0.38	0.68	3.17	NaN
SM24	0.61	3.46	5.62	NaN
Average	1.59	1.03	1.70	1.10

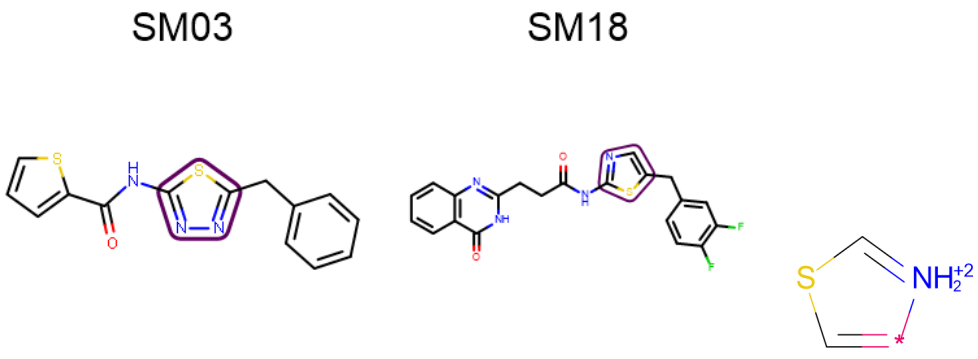


Figure 10. Problematic lewis structure generated by Epik. Epik generated protonation states with invalid Lewis structure for SM03 and SM18, with a thiazole/thiadiazole (right). One of these was also present in the SAMPL6 set of microstates, titles SM03_micro021. A valid Lewis structure could not be figured out for this protonation state.

Table 5. The pKa per molecule for each method, compared to experiment.

	Experimental pKa	Epik pKa	Epik pKa uncertainty	Epik error	Jaguar pKa	Jaguar error
SM01 pKa1						
SM01 pKa2						
SM02 pKa 1						
SM03 pKa 1						

See Images folder for titration curves and free energy images for each category. Will be added to supplementary info later.

Abstract The goal of the SAMPL6 pKa Challenge was to evaluate the performance of small molecule pKa prediction methods from challenge participants in a blinded fashion on a set of small, drug-like molecules resembling kinase inhibitor fragments. To provide a useful point of comparison for blind participant predictions that may use experimental methods still under development, we performed reference benchmark calculations using a popular empirical model (Epik) and quantum chemical approach (Jaguar) from Schrödinger. Epik predicts microstate populations and pKas using a Hammett-Taft type model, while Jaguar is a fast DFT quantum chemical method. In this work, we discuss how these reference calculations were performed, provide a broad assessment of the performance of the method, and highlight challenges and considerations in predicting pKas to benchmark against experiment macroscopic pKa measurements.

Introduction

Titratable sites are ubiquitous in druglike small molecules.

JDC: It would be very useful if we could cite some sort of reference discussing how important protonation states are in small molecule drugs or druglike molecules, but I'm failing to find useful references right now. Alternatively, we could pull pKas from DrugBank https://www.drugbank.ca/releases/latest or use Epik to make our own plots, like the kinase inhibitor pKa plot we had previously generated.

Large-scale computational surveys suggest that 60% of all protein-ligand complexes undergo a change in ionization state upon binding [1], either due to protonation state changes of the small molecule or the protein (where roughly a third of all protein residues are ionizable [2]). More generally, protonation state effects—in which the dominant protonation, charge, or tautomer state shifts upon binding, or a mixture of protonation states are significantly populated in complex or solution—has the potential to cause large modeling errors if these effects are neglected. In the SAMPL5 distribution coefficient (logD) challenge, for example, protonation state effects were determined to be a major contributor to loss in accuracy for the otherwise mundane task of predicting a transfer free energy between aqueous and cyclohexane phases [3].

To isolate the question of how well pKa effects could be modeled—and therefore how accurately the community could address these effects—the SAMPL6 challenge featured a blind pKa prediction component, as an intermediate step to logD predictions in which we provide participants with pKas and later have them predict both pKa and logD [**?**]. The SAMPL6 pKa challenge consisted of predicting macroscopic pKas measured by UV-metric titration for a set of small moelcules that resembled kinase inhibitors and their fragments [**?**]. As participants in the SAMPL6 pKa challenge were expected to utilize a wide variety of methods still under development, we endeavored to provide a useful baseline reference set of predictions using well-established, widely-deployed, commercially available methods. We selected both an empirical method (Epik [4]) and quantum chemical method (Jaguar [5]) from the Schrödinger Suite of computational chemistry software, version 20XX-X.

Reference calculations—which were not fully blinded—were performed in a manner that attempted to mimic standard use, using recommended settings for each program, without significantly modifying the input parameters. As the computation of UV-metric macroscopic pKas from the microscopic pKas predicted

by the tools is not necessarily completely straightforward, we considered several alternative possibilities, which we discuss in more detail. We provide an analysis and broad assessment of the performance of the two methods, and highlight challenges and considerations in predicting pKas to benchmark against experiment. All analysis tools used to perform this study are available via GitHub at https://github.com/choderalab/SAMPL6-Reference-pKa-Calculations

Methods

Epik

Summary of how Epik works

Could be filled in by John Shelley?

Jaguar

Summary of how Jaguar computes pKas

Could be filled in by Art Bochevarov?

Microstate predictions

Epik: Fast empirical pKa predictions

- were performed between pH 2-12
- Reported states with a minimum population of $e^{-10/RT}$
- States proposed by SAMPL6 but not predicted by Epik were not considered in analysis.

Jaguar: Ab initio quantum chemical pKa predictions

- Ran using default settings (5 conformations) for each of the pre-enumerated specified microstate pairs.
- Input structures were first minimized using MMFF.
- Additional minimization was performed for structures for which scf/geopt did not converge

Predicting macroscopic pKas

Epik and Jaguar predict microscopic pKas or microstate energies, which must be translated into macroscopic pKas for comparison to experiment. Several reasonable choices are possible for translating these microscopic properties into macroscopic pKas.

Sequential titration of dominant species

- Sequential titration avoids the need to compute the energies of *all* protonation states
- Epik can automatically perform a sequential scan (which can also be used in Jaguar with some automation)
- For this experiment, we started from the predicted highest-occupancy microstate at pH 7 and went in both directions, returning pKas between 2-12

Virtual electrochemical titration

- Average charge vs pH to match electrochemical titrations, look at the mean signed deviation (MSD) to see whether the charge curve behavior matches

Predicting UV-metric titration

- Actual experiments were UV-metric, where only UV-active transitions are observable
- There is potentially a route to predicting UV-active transitions via identifying of microstates that would have different UV spectra, perhaps using SMARTS matches or simple QM absorption spectra calculations?

Results

- How much conformation-dependence is there in Jaguar-derived pKas (or state energies)?
- Does sequential scan or mean molecular charge provide better agreement with experimental macroscopic pKas? Would it be worthwhile to develop predictive UV-metric models?
- Comparison of observed accuracies to previously reported/expected accuracies; expected accuracy on kinase inhibitors derived from this study
- How does Epik compare to Jaguar in terms of accuracy and computational cost?
- Discussion of outliers

Discussion

Overall performance

Some low probability structures produced by Epik included questionable protonation states of a heterycyclic moiety, present in SM03 and SM18 (Figure 10).

Overall performance of Epik and Jaguar based on various metrics (figures/tables)

Matching of experimental and calculated pKas

The experiments do not provide any microscopic information on what atom, or microstate a pKa belongs to. Therefore, it is necessary to perform a matching of pKa values between experiment and predicted pKa values. There are several ways one could go about this, each strategy can prioritize a different aspect to match. We consider three algorithms:

Closest pKa matching

Hungarian pKa matching

This algorithm, also known as linear sum assignment. It finds the combination of pKa that minimizes the overall cost by picking rows and columns in a matrix, also considering the cost of not mapping certain pKa values in the case of different numbers of predictions and experimental values.

Sequential pKa matching

For macroscopic only, because it doesn't make sense to sequentially align microscopic pKas to a set of macroscopic pKas algorithm 2

Macroscopic titration curves

- Compare macroscopic titration results

Population curves from pKa

- Description of the population curve generation from pKa

$$g_i(\mathrm{pH}) = \beta\left(n_i * \mathrm{pH} - \sum_j \mathrm{pKa}_j\right) \tag{1}$$

$$\pi_i(\mathrm{pH}) = \frac{e^{-g_i(\mathrm{pH})}}{\sum_i e^{-g_i(\mathrm{pH})}} \tag{2}$$

Virtual electrochemical titration

- Description of the calculation of the mean charge curve

$$\langle q_{\mathrm{total}} \rangle(\mathrm{pH}) = \sum_i q_i \times \pi_i(\mathrm{pH}) \tag{3}$$

Microscopic pKa values

- Compare microscopic pKa prediction results

Descriptor analysis

- which moieties are harder to predict?
- Do certain descriptors correlate with variance/absolute errors?
- Do number of rotatable bonds affect epik/jaguar results (conformations missing from prediction)?

Comparison with other methods

This subsection depends on how much information we already have in the overview paper

- Compared to other Hammet-Taft methods, does Epik stand out? [based on early overview results?]
- Compared to other QM methods, does Jaguar stand out? [based on early overview results?]

Suggestions for future challenges

- Future challenges could benefit from NMR experiments for microstates
- Potentiometric titrations could capture states that may have been left out
- Probabilistic models (i.e. bayesian hierarchical models) could be constructed for the analysis of experiments.

Conclusions

- Epik and Jaguar can be useful as baselines
- We can use various ways of judging each method.
- A conclusion about the accuracy
- Some take away points about difficulties about comparing microscopic to macroscopic
- How a future challenge could make this easier

Code and data availability

- Input files and analysis scripts are available at https://github.com/choderalab/SAMPL6-reference-pka-calculations

Author Contributions

(Follow the CRediT Taxonomy)

Acknowledgments

ASR, MI, AR, PBG, and JDC acknowledge support from the Sloan Kettering Institute. JDC acknowledges support from NIH grant P30 CA008748.

Disclosures

JDC is a member of the Scientific Advisory Board for Schrödinger, LLC.

References

[1] **Aguilar B**, Anandakrishnan R, Ruscio JZ, Onufriev AV. Statistics and Physical Origins of pK and Ionization State Changes upon Protein-Ligand Binding. Biophys J. 2010 Mar; 98(5):872–880. doi: 10.1016/j.bpj.2009.11.016.

[2] **Jordan IK**, Kondrashov FA, Adzhubei IA, Wolf YI, Koonin EV, Kondrashov AS, Sunyaev S. A Universal Trend of Amino Acid Gain and Loss in Protein Evolution. Nature. 2005 Feb; 433(7026):633–638. doi: 10.1038/nature03306.

[3] **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. J Comput Aided Mol Des. 2016 Nov; 30(11):1087–1100. doi: 10.1007/s10822-016-9955-7.

[4] **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. J Comput Aided Mol Des. 2007 Dec; 21(12):681–691. doi: 10.1007/s10822-007-9133-z.

[5] **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. Int J Quantum Chem. 2013 Sep; 113(18):2110–2142. doi: 10.1002/qua.24481.

[6] **McQuilton P**, St Pierre SE, Thurmond J, the FlyBase Consortium. FlyBase 101 – the basics of navigating FlyBase. Nucleic Acids Research. 2012; 40(D1):D706–D714. doi: 10.1093/nar/gkr1030.

[7] **Aivazian D**, Serrano RL, Pfeffer S. TIP47 is a key effector for Rab9 localization. The Journal of Cell Biology. 2006; 173(6):917–926. doi: 10.1083/jcb.200510010.

[8] **Bloss CS**, Wineinger NE, Peters M, Boeldt DL, Ariniello L, Kim JY, Sheard J, Komatireddy R, Barrett P, Topol EJ. A prospective randomized trial examining health care utilization in individuals using multiple smartphone-enabled biosensors. bioRxiv. 2016; doi: 10.1101/029983.

[9] **Brettar I**, Christen R, Höfle MG. Aquiflexum balticum gen. nov., sp. nov., a novel marine bacterium of the Cytophaga–Flavobacterium–Bacteroides group isolated from surface water of the central Baltic Sea. International Journal of Systematic and Evolutionary Microbiology. 2004; 54(6):2335–2341. doi: 10.1099/ijs.0.63255-0.

[10] **Brettar I**, Christen R, Höfle MG. Belliella baltica gen. nov., sp. nov., a novel marine bacterium of the Cytophaga–Flavobacterium–Bacteroides group isolated from surface water of the central Baltic Sea. International Journal of Systematic and Evolutionary Microbiology. 2004; 54(1):65–70.

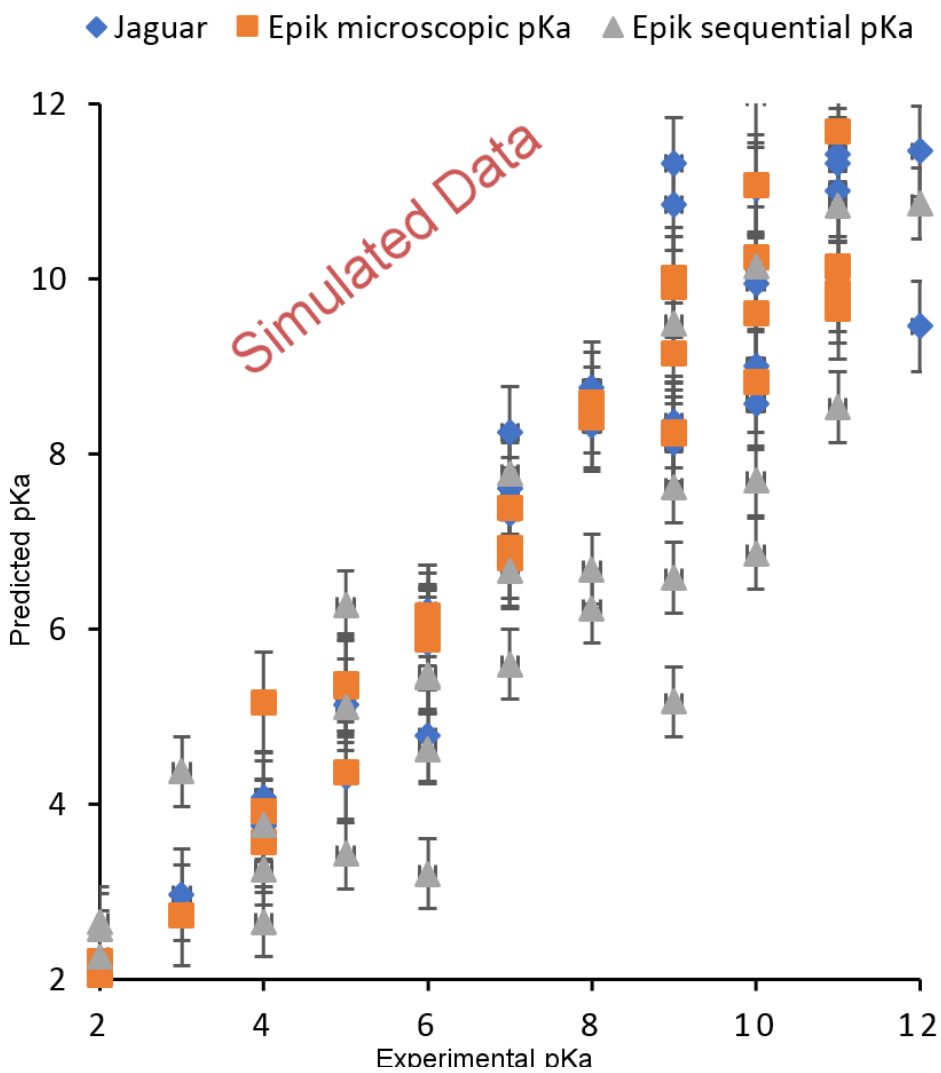
[11] **Van Kreel BK**. The Estimation of the Apparent Standard Free Energy Change ΔGopH of a Biochemical Reaction from the Standard Free Energy of Formation and Apparent Free Energy of Ionization of the Participating Molecules and Its Application to the Reactions of the Purine Metabolism. Biochem Mol Biol Educ. 1985; 13(3):125–130.

[12] **Król M**, Wrona M, Page CS, Bates PA. Macroscopic p K_a Calculations for Fluorescein and Its Derivatives. J Chem Theory Comput. 2006 Nov; 2(6):1520–1529. doi: 10.1021/ct600235y.

[13] **ten Brink T**, Exner TE. pKa Based Protonation States and Microspecies for Protein–ligand Docking. J Comput Aided Mol Des. 2010 Nov; 24(11):935–942. doi: 10.1007/s10822-010-9385-x.

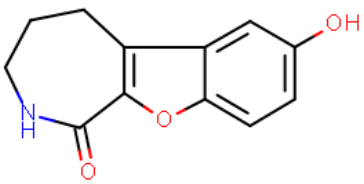
[14] **Manchester J**, Walkup G, Rivin O, You Z. Evaluation of p K_a Estimation Methods on 211 Druglike Compounds. J Chem Inf Model. 2010 Apr; 50(4):565–571. doi: 10.1021/ci100019p.

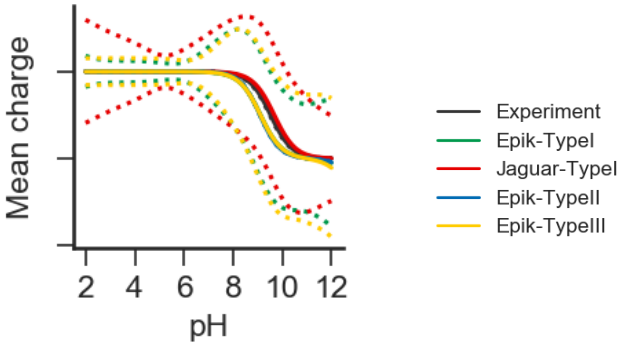
207 ## Supplementary Information



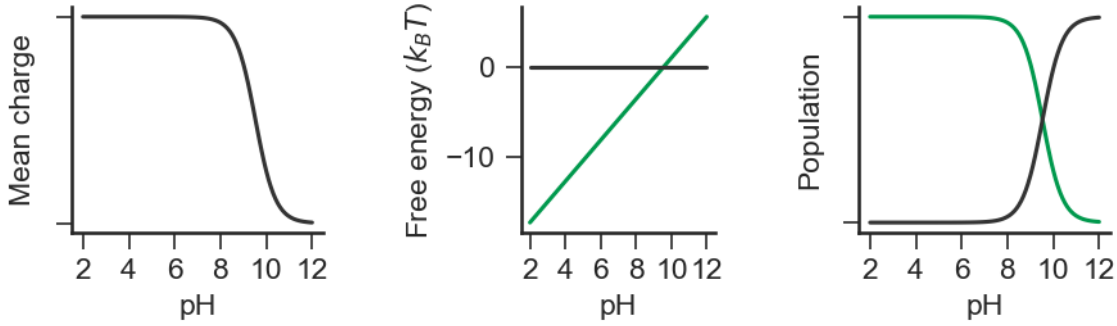
Appendix 0 Figure 11. Comparison of the entire set of experimental pKa values to all predictions using Hungarian pKa matching. Microscopic pKa values for Jaguar (diamonds) and Epik (squares), as well as macroscopic Epik pKa values (triangles) are compared to the experimental dataset. The pKa values were matched using algorithm 1. RMSE and r-squared values can be found in table 1. An alternative alignment of pKa values using **??** is available as a supplementary figure.

SM01

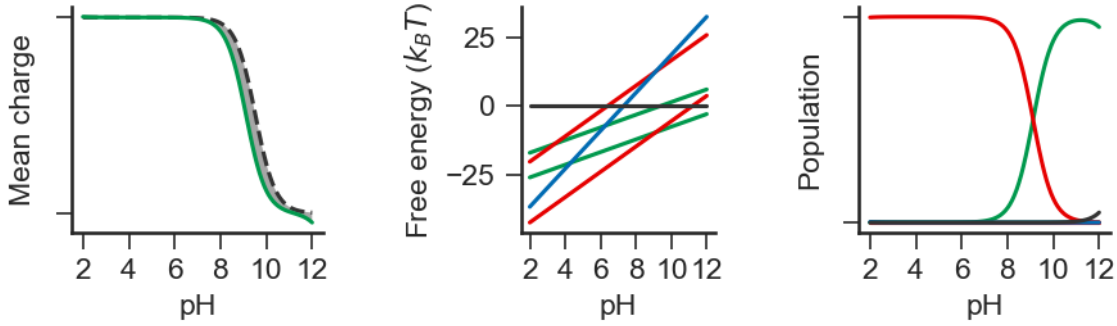




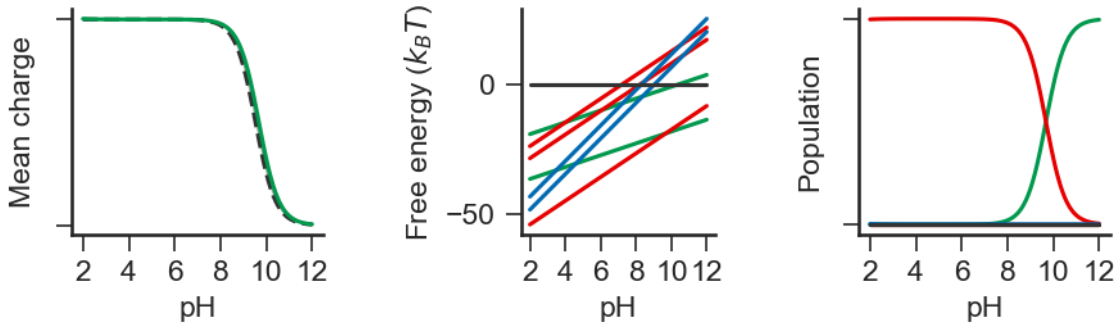
Experiment



Epik-TypeI



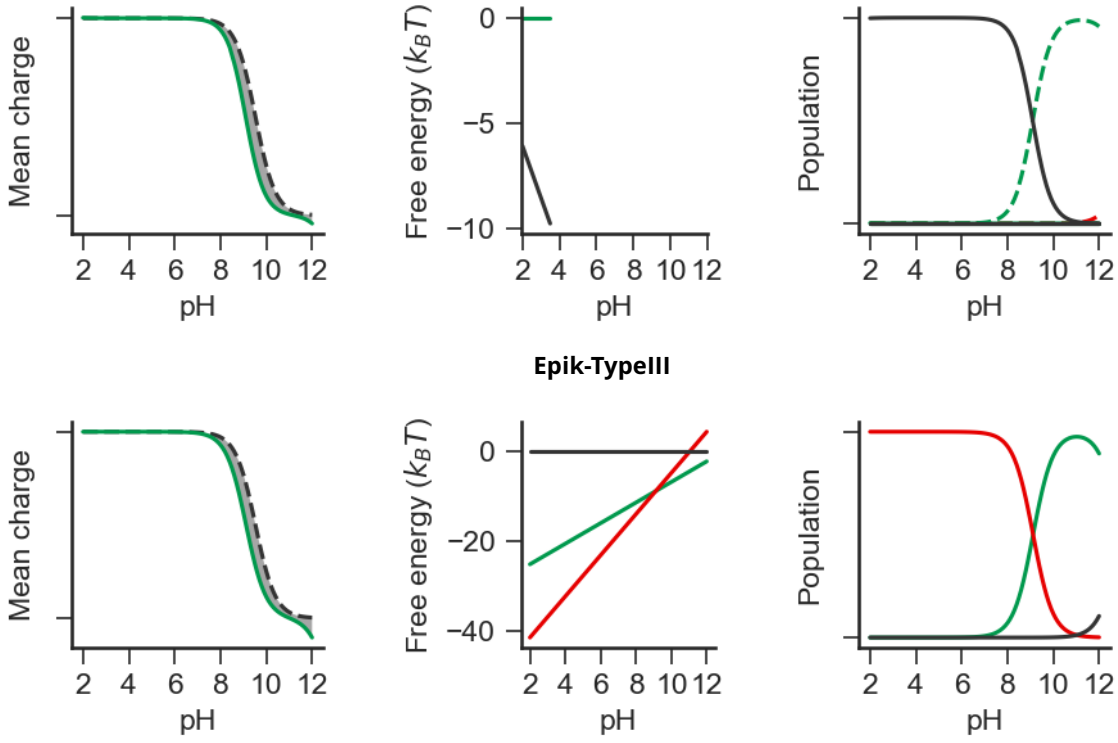
Jaguar-TypeI



Epik-TypeII

Result: Mapping of each experimental pKa i to predicted pKa j
C is constructed, where every row i is an experiment, and every column j a prediction;
C_{ij} = cost($\mathrm{pKa}_{\mathrm{exp},i}$, $\mathrm{pKa}_{\mathrm{pred},j}$);
while $C.size > 0$ **do**
  k, l = arg min(C_{ij});
  assign experimental value k to prediction l;
  remove row k, column l from C;
end
remaining values are unmatched;

Algorithm 1: This algorithm matches experiment with prediction based on how close each value is, one pKa value at a time. Unless the matrix C is square, some values will be unmatched. Those leftover pKas are returned at the end. It uses a cost function, such as root mean square deviation, to assess how close two values are.



Epik-TypeIII

- Titration curves and free energy plots for each compound, by each method
- The mean charge/deviation curves for each compound and each method
- .mae and .sdf files with results
- scripts and or jupyter notebooks for analysis
- csv version of tables

Result: Mapping of each experimental pKa i to predicted pKa j
I = sorted experimental pKas ;
J = sorted predicted pKas ;
length = max(I.size, J.size);
Append placeholders such that I.size = length ;
Append placeholders such that J.size = length ;
min = ∞;
solution = J rolled 0 times;
for *n in 0..length* **do**
  S = J rolled n times ;
  total = 0.0;
  for *m in 0..length* **do**
    if I_m *is placeholder* **then** unmatched experiment
      if J_m <= *7.0* **then**
        I_m = 0.0
      else
        I_m = 14.0
      end
    else if J_m *is placeholder* **then** unmatched prediction
      if I_m <= *7.0* **then**
        J_m = 0.0
      else
        J_m = 14.0
      end
    end
    total = total + cost(I_m, J_m);
  end
  if *total* < *min* **then** solution is better
    min = total;
    solution = S ;
  end
end

Algorithm 2: Sequential pKa mapping. It uses a cost function to measure cost, and it rolls (shifts by one, and reintroduces last element as first). Any unmatched pKas are represented by matching with a placeholder value. To calculate the cost, the placeholder is replaced by either 0, or 14, depending on whether the unmatched value is above or below 7.0.