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

track down type II artifacts, and dont use type I labels, use epik scan, sequential, et cetera

- 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]
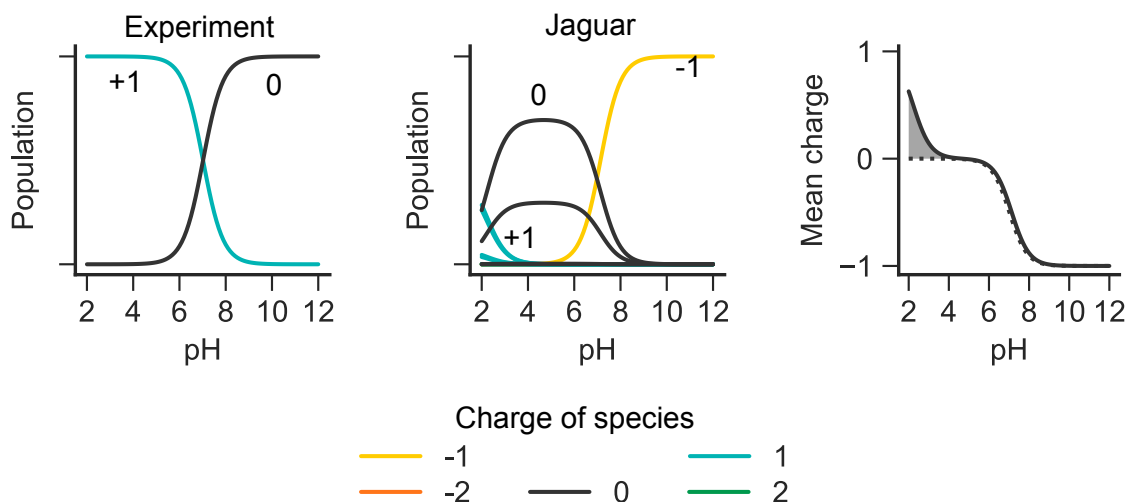
31 **Figures and Tables**

Figure 1. Comparing UV experiments for SM03 to microscopic pKa Jaguar predictions. This example shows how the UV experimental populations (left) for macrostates molecule SM03 as a function of the pH can be compared to Jaguar predictions. Jaguar was used to predict microscopic pKas (denoted as Type I predictions in SAMPL6), which can be used to derive relative microstate populations (middle). In the UV experiment, charge reflects the number of protonations, since absolute charge information is unknown. The charge combined with populations produces a macroscopic titration curve (right), with the prediction as a solid, and the experiment as a dashed line. The area between the curves (gray) is calculated after correcting for an integer offset in charges.
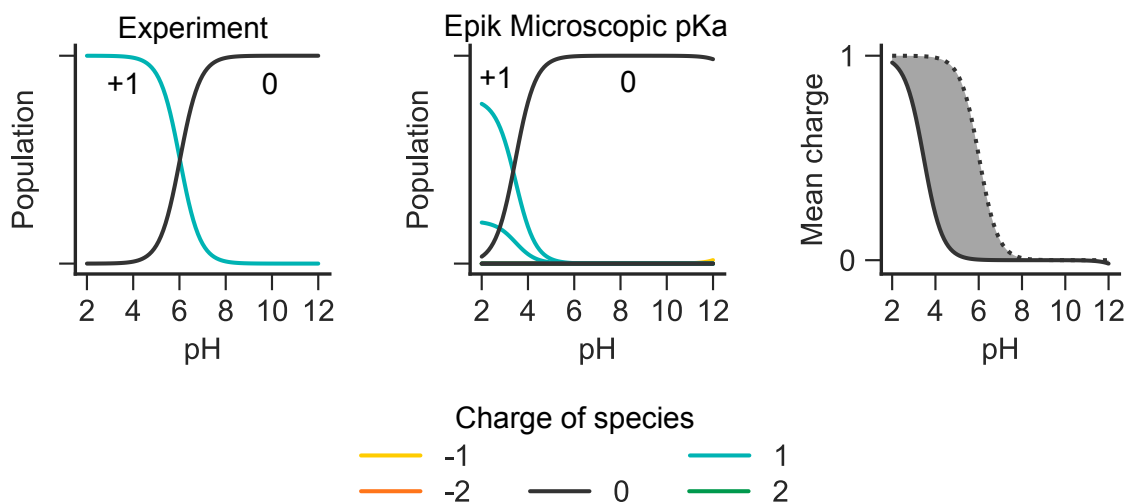
ASR: Add number of area to plot

Figure 2. Comparing UV experiments for SM04 to Epik microscopic pKa predictions .This example shows how the UV experimental populations (left) for macrostates molecule SM04 as a function of the pH can be used to compare to microscopic pKa Epik predictions. Epik was used to predict microscopic pKas (denoted as Type I predictions in SAMPL6), which can be used to derive relative microstate populations (middle). In the UV experiment, charge reflects the number of protonations, since absolute charge information is unknown. The charge combined with populations produces a macroscopic titration curve (right), with the prediction as a solid, and the experiment as a dashed line. The area between the curves (gray) is calculated after correcting for an integer offset in charges.
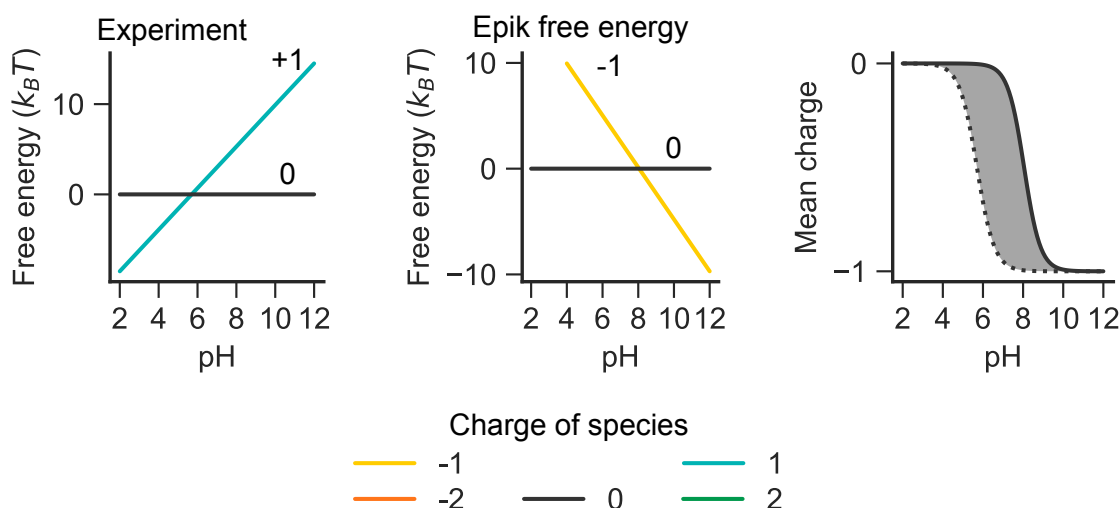
Figure 3. Comparing UV experiments for SM20 to Epik microscopic free energy predictions .This example shows how the UV experimental free energy (left) for macrostates molecule SM20 as a function of the pH can be compared to microscopic free energy Epik predictions. Epik was used to predict microscopic free energy (denoted as Type II predictions in SAMPL6) (middle). In the UV experiment, charge reflects the number of protonations, since absolute charge information is unknown. The charge combined with free energies produces a macroscopic titration curve (right), with the prediction as a solid, and the experiment as a dashed line. The area between the curves (gray) is calculated after correcting for an integer offset in charges.
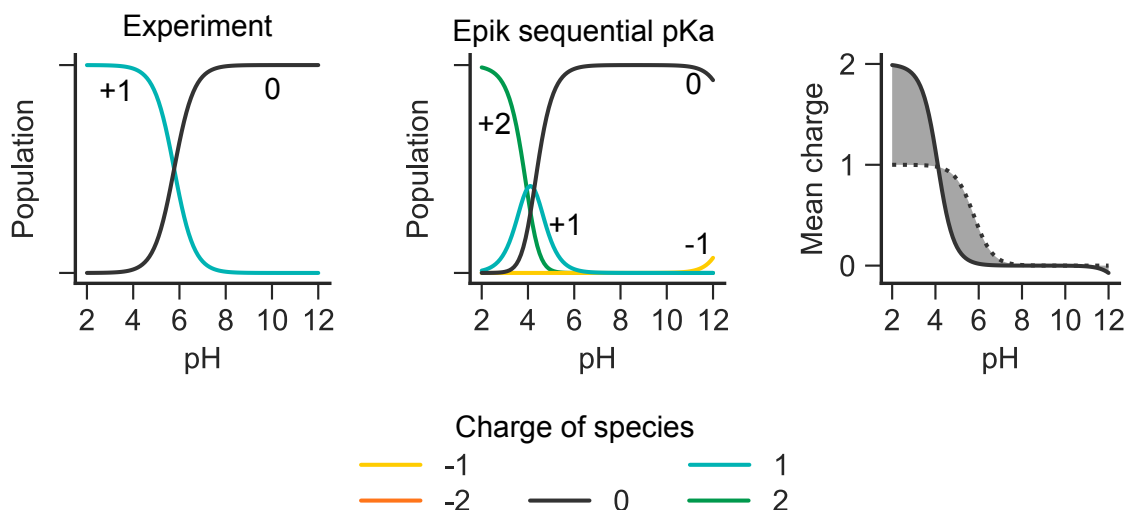
Figure 4. Comparing UV experiments for SM13 to Epik macroscopic pKa predictions .This example shows how the UV experimental populations (left) for macrostates molecule SM13 as a function of the pH can be compared to sequential pKa Epik predictions. Epik was used to predict macroscopic pKa values (denoted as Type III predictions in SAMPL6), which can be translated into relative macrostate populations (middle). In the UV experiment, charge reflects the number of protonations, since absolute charge information is unknown. The charge combined with populations produces a macroscopic titration curve (right), with the prediction as a solid, and the experiment as a dashed line. The area between the curves (gray) is calculated after correcting for an integer offset in charges.

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

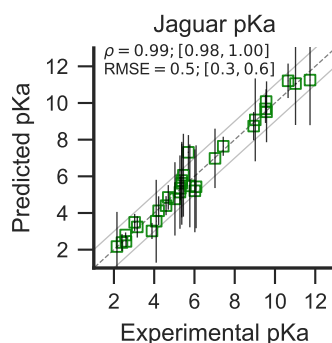|	Closest			Hungarian			Titration curve
	RMSE	MAE	R^2	RMSE	MAE	R^2	Δ area
Epik sequential							
Epik micropka							
Jaguar micropKa							

Figure 5. Comparison of microscopic Jaguar pKa values to experimental macroscopic pKa values. Microscopic pKa values for Jaguar were matched ot the experimental dataset using algorithm 1. An alternative alignment of pKa values using the Hungarian algorithm (**??**) is available as a supplementary figure.
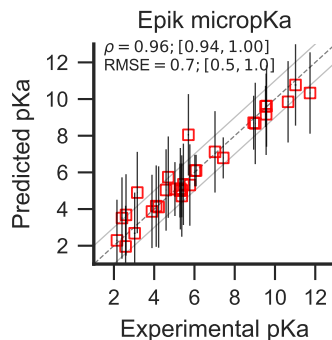

Figure 6. Comparison of microscopic Epik pKa values to experimental macroscopic pKa values. Microscopic pKa values for Epik were matched ot the experimental dataset using algorithm 1. An alternative alignment of pKa values using the Hungarian algorithm (**??**) is available as a supplementary figure.
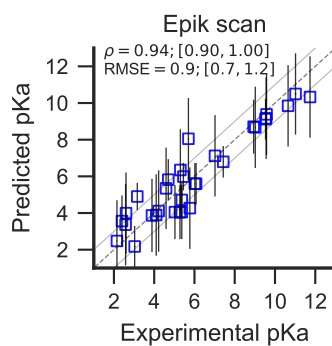
Figure 7. Comparison of macroscopic Epik pKa values to experimental macroscopic pKa values. Macroscopic pKa values for Jaguar were matched ot the experimental dataset using algorithm 1. An alternative alignment of pKa values using the Hungarian algorithm (**??**) is available as a supplementary figure.
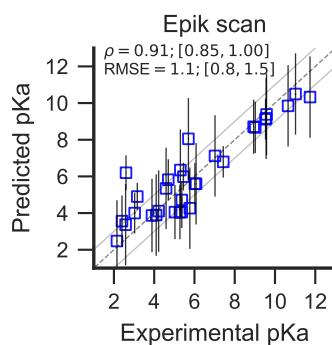

Figure 8. Epik macroscopic pKa values compared to macroscopic experimental pKa values, sequentially matching pKa values Comparison of the Epik sequential scan with entire experimental data set is shown, allowing only sequential pKa values to match experiment.

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

Add table for the 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						

ASR: Find a good place to put the legend, and regenerate figures with lines less thick, or more alpha, add molecules to these. Plot experiment as dashed with each curve.
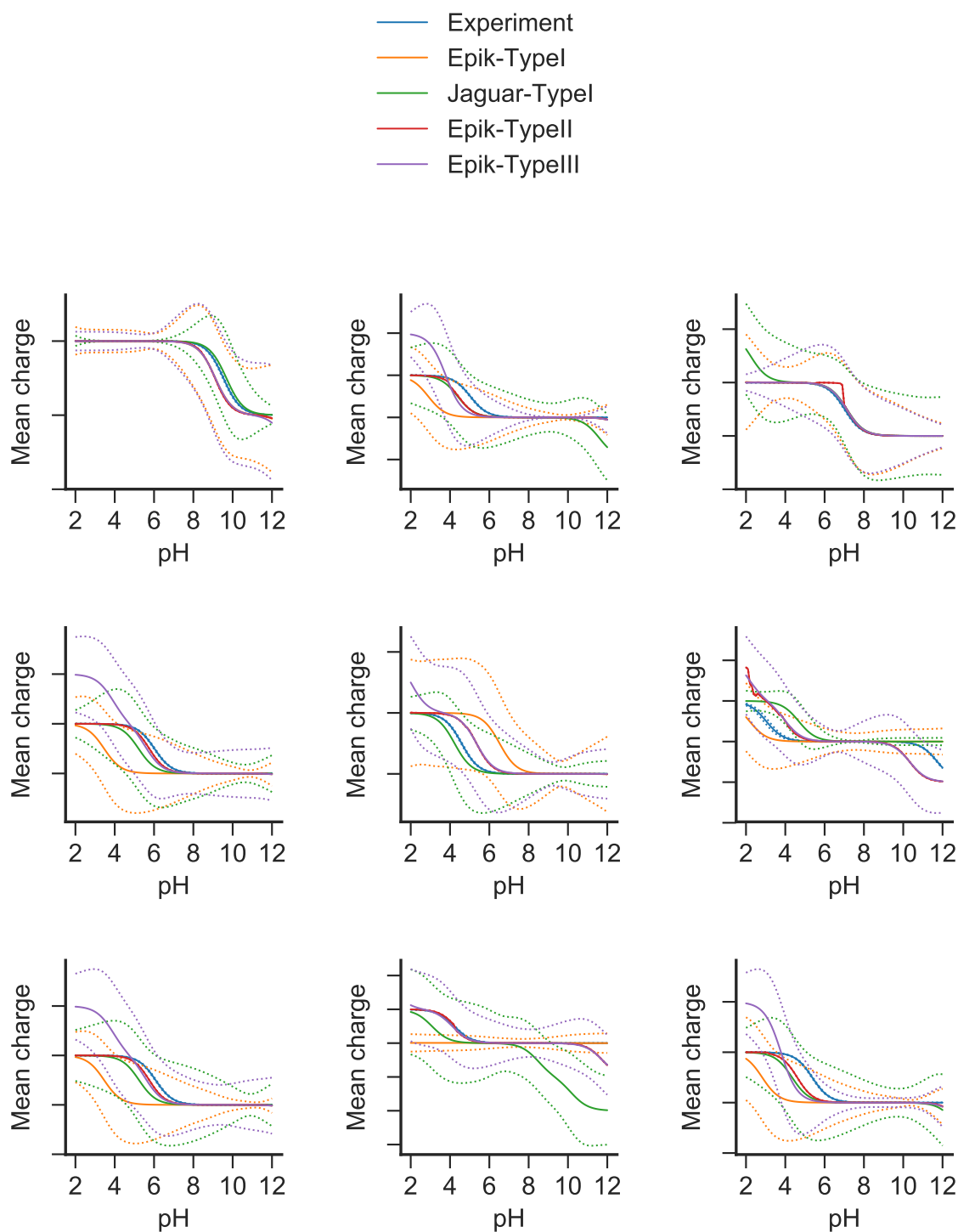

Figure 9. Bootstrap titration curves for each method compared to experiment for molecules SM01-SM09. The titration curve for each method is shown as a solid line, and 96 % confidence intervals from bootstrap have been shown as dotted lines. A numerical comparison to experiment is presented in table 4.
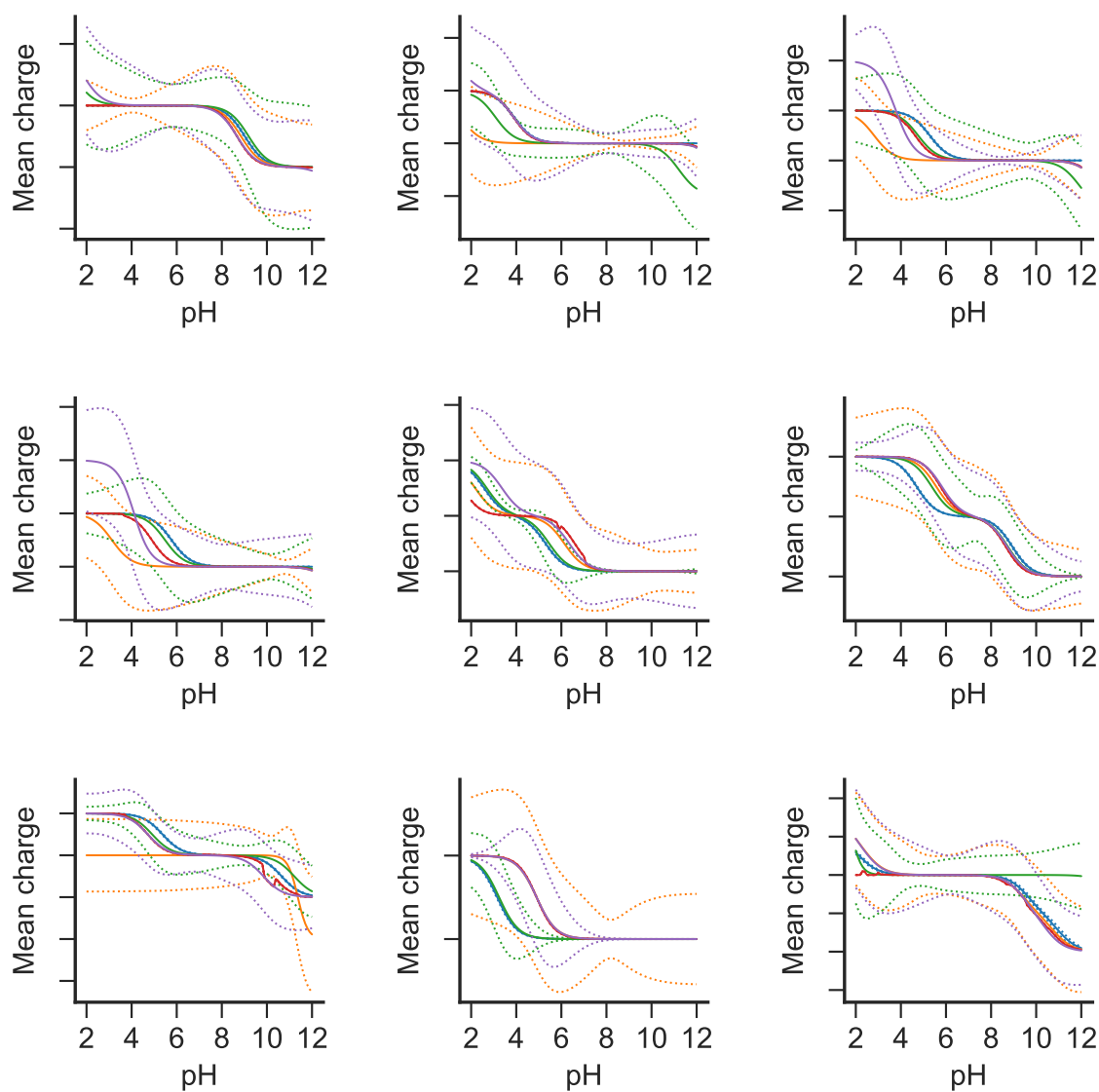
Figure 10. Bootstrap titration curves for each method compared to experiment for molecules SM10-SM18. The titration curve for each method is shown as a solid line, and 96 % confidence intervals from bootstrap have been shown as dotted lines. A numerical comparison to experiment is presented in table 4.

Figure 11. 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.
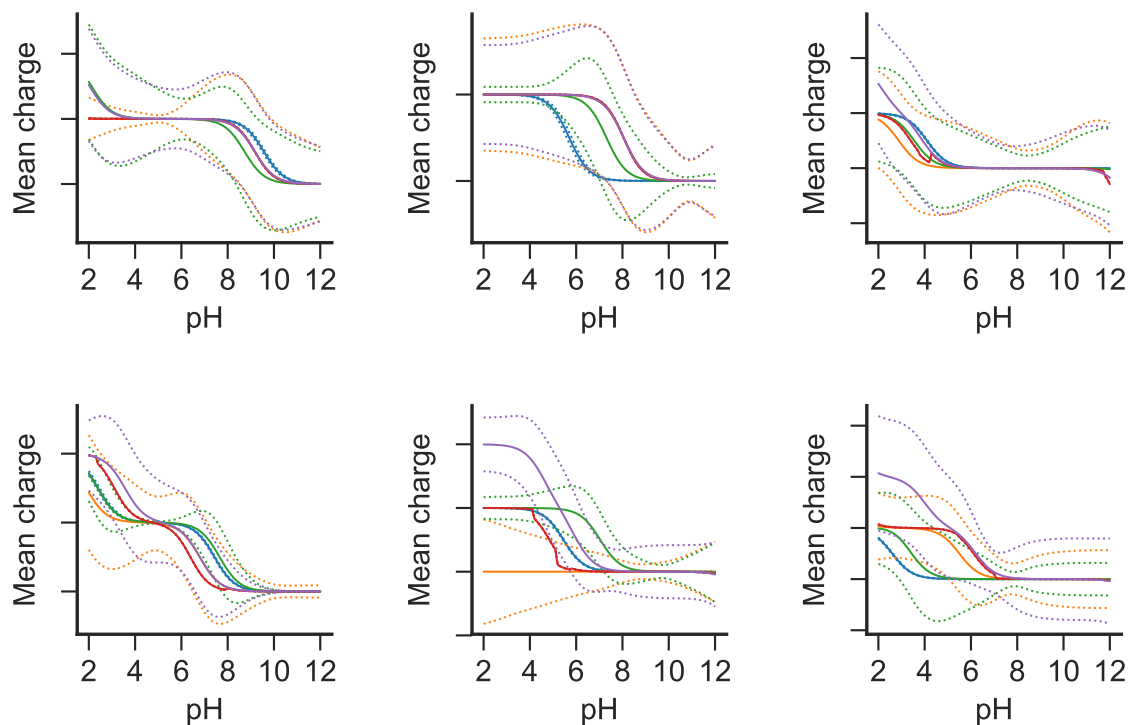
Table 4. Difference in area between derived macroscopic charge titration curves (∆ area) for each method. The macroscopic charge curve was constructed for each of the methods, and after alignment with the experimental curve by an integer offset, the area between the curves is reported.

| 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
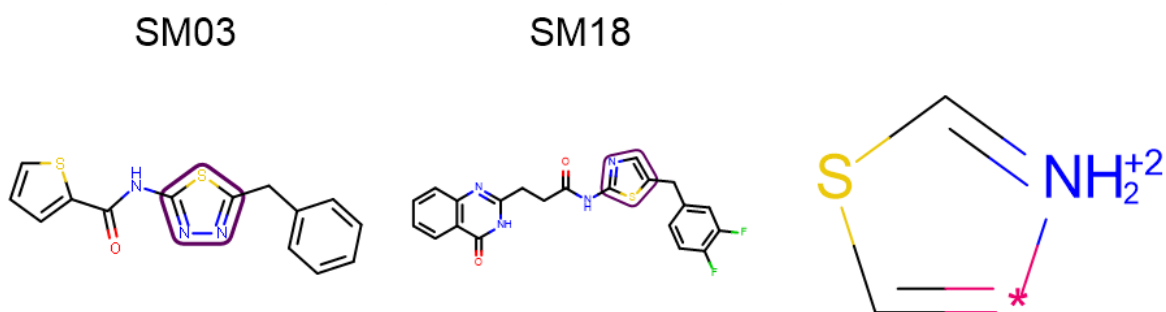
Figure 12. Problematic lewis structure generated by Epik. Epik generated protonation states with invalid Lewis structure for SM03 and SM18, with a pentavalent nitrogen present in the thiazole/thiadiazole (the protonation pattern displayed on the right). One of these was also present in the set provided as part of SAMPL6 microstates, namely SM03_micro021. A valid Lewis structure could not be figured out for this protonation state.

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

33

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

43

Introduction

45 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.

46

47 Large-scale computational surveys suggest that 60% of all protein-ligand complexes undergo a change
48 in ionization state upon binding [1], either due to protonation state changes of the small molecule or the
49 protein (where roughly a third of all protein residues are ionizable [2]). More generally, protonation state
50 effects—in which the dominant protonation, charge, or tautomer state shifts upon binding, or a mixture
51 of protonation states are significantly populated in complex or solution—has the potential to cause large
52 modeling errors if these effects are neglected. In the SAMPL5 distribution coefficient (logD) challenge, for
53 example, protonation state effects were determined to be a major contributor to loss in accuracy for the
54 otherwise mundane task of predicting a transfer free energy between aqueous and cyclohexane phases [3].
55 To isolate the question of how well pKa effects could be modeled—and therefore how accurately the
56 community could address these effects—the SAMPL6 challenge featured a blind pKa prediction component,
57 as an intermediate step to logD predictions in which we provide participants with pKas and later have
58 them predict both pKa and logD [**?**]. The SAMPL6 pKa challenge consisted of predicting macroscopic pKas
59 measured by UV-metric titration for a set of small moelcules that resembled kinase inhibitors and their
60 fragments [**?**]. As participants in the SAMPL6 pKa challenge were expected to utilize a wide variety of
61 methods still under development, we endeavored to provide a useful baseline reference set of predictions
62 using well-established, widely-deployed, commercially available methods. We selected both an empirical
63 method (Epik [4]) and quantum chemical method (Jaguar [5]) from the Schrödinger Suite of computational
64 chemistry software, version 20XX-X.
65 Reference calculations—which were not fully blinded—were performed in a manner that attempted to
66 mimic standard use, using recommended settings for each program, without significantly modifying the
67 input parameters. As the computation of UV-metric macroscopic pKas from the microscopic pKas predicted
68 by the tools is not necessarily completely straightforward, we considered several alternative possibilities,
69 which we discuss in more detail. We provide an analysis and broad assessment of the performance of
70 the two methods, and highlight challenges and considerations in predicting pKas to benchmark against
71 experiment. All analysis tools used to perform this study are available via GitHub at https://github.com/
72 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 12).

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

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

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

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

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 pK_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 pK_a Estimation Methods on 211 Druglike Compounds. J Chem Inf Model. 2010 Apr; 50(4):565–571. doi: 10.1021/ci100019p.
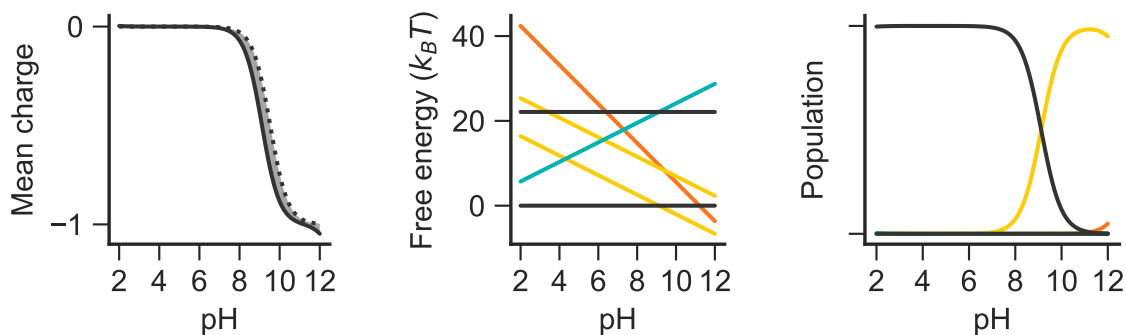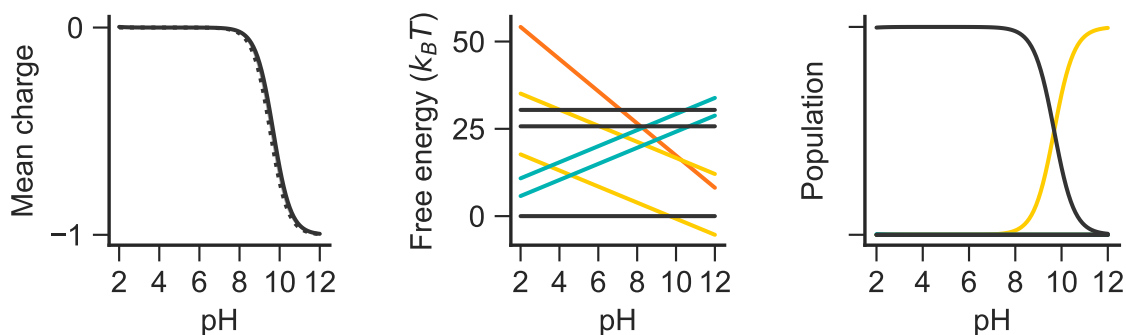
Supplementary Information

SM01

Experiment
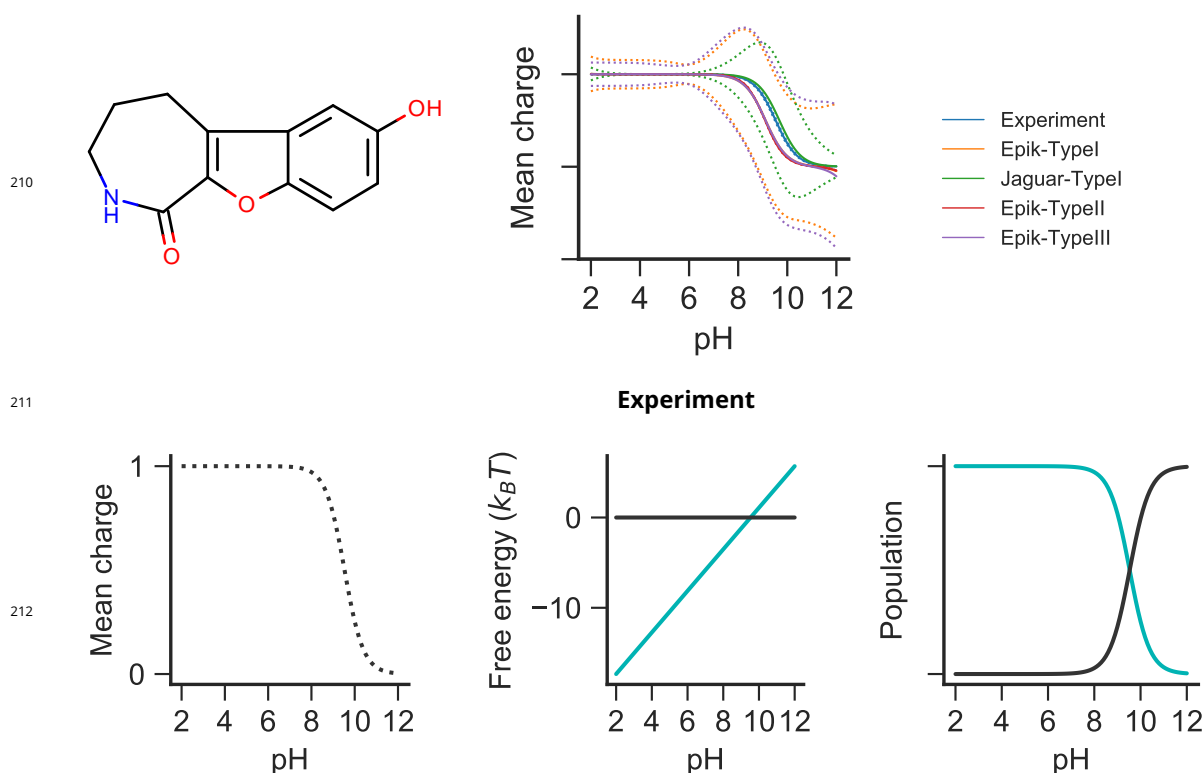
213

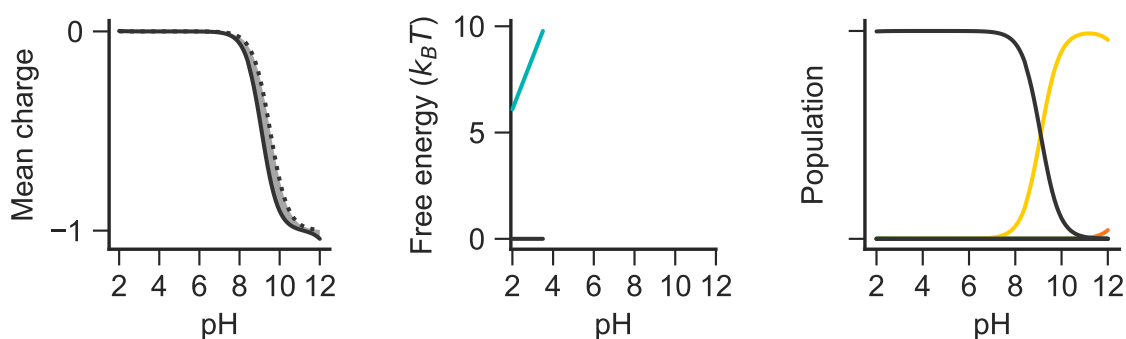**Epik-TypeI**

214

215

Jaguar-TypeI
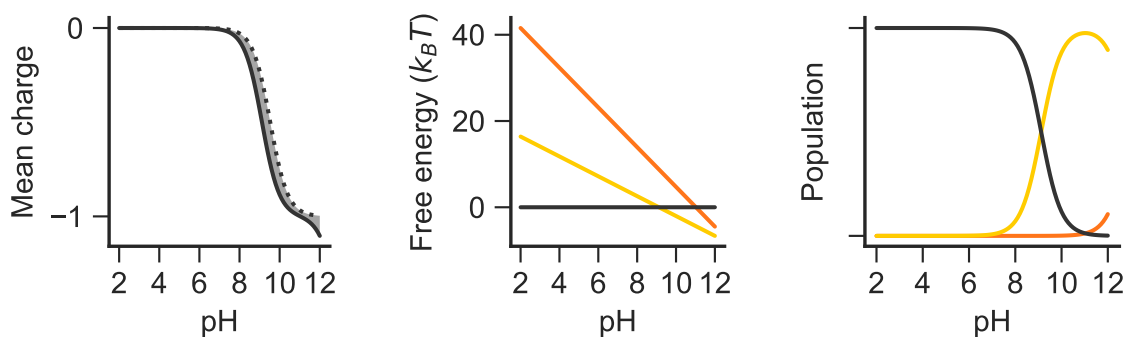
216

217

Epik-TypeII

218

219

Epik-TypeIII

220

221 • Titration curves and free energy plots for each compound, by each method
222 • The mean charge/deviation curves for each compound and each method

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(pKa$_{\text{exp},i}$, pKa$_{\text{pred},j}$);

while *C.size > 0* **do**

 $\quad$ k, l = arg min(C_{ij});

 $\quad$ assign experimental value k to prediction l;

 $\quad$ 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.

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**

 $\quad$ S = J rolled n times ;

 $\quad$ total = 0.0;

 $\quad$ **for** *m in 0..length* **do**

 $\quad\quad$ **if** *I_m is placeholder* **then** unmatched experiment

 $\quad\quad\quad$ **if** *J_m <= 7.0* **then**

 $\quad\quad\quad\quad$ $\mid$ I_m = 0.0

 $\quad\quad\quad$ **else**

 $\quad\quad\quad\quad$ $\mid$ I_m = 14.0

 $\quad\quad\quad$ **end**

 $\quad\quad$ **else if** *J_m is placeholder* **then** unmatched prediction

 $\quad\quad\quad$ **if** *I_m <= 7.0* **then**

 $\quad\quad\quad\quad$ $\mid$ J_m = 0.0

 $\quad\quad\quad$ **else**

 $\quad\quad\quad\quad$ $\mid$ J_m = 14.0

 $\quad\quad\quad$ **end**

 $\quad\quad$ **end**

 $\quad\quad$ total = total + cost(I_m, J_m);

 $\quad$ **end**

 $\quad$ **if** *total < min* **then** solution is better

 $\quad\quad$ min = total;

 $\quad\quad$ solution = S ;

 $\quad$ **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.

223   • .mae and .sdf files with results

224   • scripts and or jupyter notebooks for analysis

225   • csv version of tables