Benchmarking *in silico* Tools for Cysteine pK_a Prediction

Ernest Awoonor-Williams,[†*] Andrei A. Golosov,[†] and Viktor Hornak[†]

[†]Novartis Institutes for BioMedical Research, 181 Massachusetts Avenue, Cambridge, Massachusetts 02139, United States

Cysteine, Protein, pK_a prediction, thiol, thiolate, covalent drug discovery

ABSTRACT: Accurate estimation of the pK_a's of cysteine residues in proteins could inform targeted approaches in hit discovery. The pK_a of a targetable cysteine residue in a disease-related protein is an important physiochemical parameter in covalent drug discovery, as it influences the fraction of nucleophilic thiolate amenable to chemical protein modification. Traditional structure-based *in silico* tools are limited in their predictive accuracy of cysteine pK_a's relative to other titratable residues. Additionally, there are limited comprehensive benchmark assessments for cysteine pK_a predictive tools. This raises the need for extensive assessment and evaluation of methods for cysteine pK_a prediction. Here, we report the performance of several computational pK_a methods, including single structure and ensemble-based approaches, on a diverse test set of experimental cysteine pK_a's retrieved from the PKAD database. The dataset consisted of 16 wildtype and 10 mutant proteins with experimentally measured cysteine pK_a values. Our results highlight that these methods are varied in their overall predictive accuracies. Among the test set of wildtype proteins evaluated, the best method (MOE) yielded a mean absolute error of 2.3 pK units — highlighting the need for improvement of existing pK_a methods for accurate cysteine pK_a estimation. Given the limited accuracy of these methods, further development is needed before these approaches can be routinely employed to drive design decisions in early drug discovery efforts.

Methods for the accurate calculation of the pK_a of ionizable residues in proteins can enable targeted approaches in drug discovery. The pK_a of an ionizable residue provides insight into the protonation state of a residue at a specific pH and is an important physicochemical property in the experimental and computational analysis of a protein. Knowledge of the pK_a value of a titratable residue in a protein is extremely vital in understanding the pH-dependent properties governing the structure and dynamics of a protein,[1] — as well as in elucidating the catalytic mechanisms of enzymatic reactions.[2]

Cysteine (Cys) plays diverse functional roles in cell biology,[3] including regulatory and catalytic redox activities.[4] Cysteines are strong nucleophiles for binding metals and drugs[5]— and have been widely exploited in covalent drug discovery efforts.[6–16] The nucleophilicity of a Cys residue is dependent on the ionization state of the side chain, with the deprotonated thiolate form ($-S^-$) being more nucleophilic than its protonated thiol form (-SH). The reactivity and susceptibility of a Cys residue towards deprotonation and chemical protein modification is complex;[15,17–20] however, pK_a provides information about the relative stability of both the neutral and charged states. Cysteines with low pK_a's have readily accessible thiolates that are prone to covalent chemical modification by electrophilic inhibitors (**Figure 1**).

Several methods exist for determining the pK_a of ionizable residues in proteins;[21] however, *in silico* approaches are generally preferred to experiments — given the challenging and time-consuming nature of experiments. Computational prediction of protein pK_a's often rely on the three-dimensional structure of the protein, traditionally determined by X-ray crystallography or nuclear magnetic resonance (NMR). The general strategy for *in silico* pK_a prediction is to estimate a pK_a shift (i.e., ΔpK_a) from a reference or intrinsic residue pK_a in solvent. ΔpK_a arises from the differences in the electrostatic environment and the interactions experienced by the residue in solvent and in the full protein. Many *in silico* tools exist for protein pK_a prediction,[22–30] with a significant majority of methods based on continuum electrostatics approaches[23–26] and empirical methods.[28] Among these methods, the empirical PROPKA program[27,28] is arguably the most popular and widely used due to its speed, simplicity and availability. Recently, machine learning techniques based on deep-learning representation have been developed for protein residue pK_a predictions.[31–34]

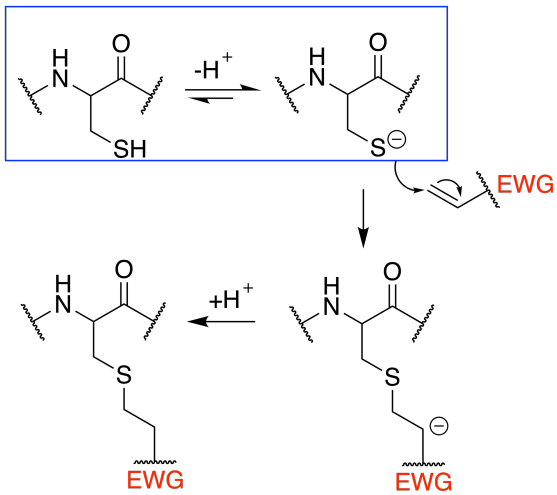


Figure 1. Mechanism of Michael addition showing the covalent modification of a cysteine residue by an electrophilic compound. *A low Cys pK_a means that a greater proportion of the thiolate anion is available to engage in chemical protein modification.*

Despite the plethora of predictive pK_a tools, significant and contrasting differences exist in their accuracy and overall predictive performance.[35–37] Notably, Cys pK_a prediction has proven challenging for these *in silico* methods[38] and very limited benchmark studies assessing their performance are present in the literature today.[38–40] Earlier effort by Awoonor-Williams and Rowley[38] evaluated four pK_a methods: continuum electrostatics-based methods (H++, MCCE), empirical PROPKA program, and explicit-solvent replica-exchange thermodynamic integration (RETI) algorithm implemented in GROMACS[41] using both CHARMM and Amber force fields, to predict 18 Cys pK_a's in a test set of 12 proteins. The explicit-solvent RETI approach with the CHARMM36 force field yielded the lowest root-mean-square error of 2.4 pK units from experiment, although this performance was comparable to the null model (RMSE = 2.7).[38] More recent work by the Shen group[40] have employed generalized Born-Neck2 continuous constant pH molecular dynamics (GB-Neck2 CpHMD) in the Amber MD suite to compute Cys pK_a's for a dataset of proteins mainly comprising the set evaluated in the Awoonor-Williams and Rowley benchmark study.[38] Their results suggest that GB-Neck2 CpHMD Cys pK_a predictions yielded RMSE of 1.2–1.3, surpassing traditional structure-based predictive pK_a methods.[38] However, the GB-Neck2 continuous CpHMD code is not freely distributed with the Amber MD package for use in our study.

Here, we revisit the evaluation of methods for predicting cysteine pK_a using a combination of freely accessible tools available in our setting to assess their predictive accuracies — prior to being employed to support medicinal chemistry projects. In our approach, we performed benchmark assessments of several different *in silico* tools to predict Cys pK_a's in proteins for which an experimental structure exists and pK_a has been determined. The experimental dataset was taken from the PKAD database,[42] and consisted of 16 wildtype (WT) and 10 mutant (MT) Cys pK_a's. We examined several methods, including industry-leading molecular modeling tools (MOE,[43] Maestro[44]), continuum electrostatics-based methods (H++,[22,23] PypKa[26]), empirical PROPKA[27,28] tool, deep-learning pKAI predictor,[34] and molecular dynamics-based sampling techniques using popular Amber and NAMD constant-pH MD codes.[45,46] We note that this is the largest test to date of cysteine pK_a prediction using a wide range of different recently-developed methods. Our aim for this study is to provide a comprehensive evaluation and assessment of these *in silico* tools for Cys pK_a prediction, to inform the broader scientific community about their predictive accuracies.

THEORY & METHODS

Data Set.

The structure files comprising the protein test set were downloaded from the Protein Data Bank (PDB).[47] Missing residues and loops within the protein model system were built using Prime[48] within Protein Preparation Wizard tool in Maestro. For protein systems with multiple chains, only chain A of the protein was considered. The Cys pK_a's considered in this work were for free cysteines and do not include cysteine residues involved in disulfide bonds or post-translational modifications. The pK_a's span a broad range of values from depressed to elevated pK_a's relative to the intrinsic solution Cys pK_a (8.6).[49] A cysteine pK_a test set comprising of 26 cysteine residues with experimentally determined pK_a's in wildtype and mutant proteins were obtained from the PKAD database:[42] 16 wild-type (WT) and 10 mutant (MT) proteins. In the mutant test set, protein structure files were not available, so single point mutations were introduced in the wildtype proteins prior to cysteine pK_a calculation. The pK_a's were determined using a wide range of experimental methods such as reaction kinetics, NMR, and spectrophotometric titration. **Tables 1** and **2** provide a summary of the test set of proteins studied in this work.

Table 1. Test Set of Wildtype Protein Cysteine pK_a's.

Protein	PDB ID	Cys ID	Exptl. pK_a
α-1-antitrypsin	1QLP	232	6.86 (0.05)[50]
AhpC	4MA9	46	5.94 (0.10)[51]
Cathepsin B	1THE	29	3.60 (0.04)[52]
DJ-1	1P5F	106	5.4 (0.1)[53]
HMCK	1I0E	283	5.6 (0.1)[54]
uMtCK	1QK1	278	5.6 (0.1)[54]
msrA	2L90	72	7.20 (0.12)[55]
O(6)-AGT	1EH6	145	5.3 (0.2)[56]
Papain	1PPN	25	3.32 (0.01)[57]
ppΩ	1PPO	25	2.88 (0.02)[57]
Thioredoxin	1ERT	32	6.3 (0.1)[58]
PTP1B	2HNP	215	5.57 (0.12)[59]
Ubc2	1JAS	88	10.2 (0.2)[60]
Ubc13	1JBB	87	11.1 (0.1)[60]
UbcH10	1I7K	102	10.9 (0.2)[60]
Yersinia PTP	1YPT	403	4.67 (0.15)[61]

Table 2. Test Set of Mutant Protein Cysteine pK_a's.

Protein	PDB ID	Cys ID	Exptl. pK_a
$ACBP^{E78C}$	1NTI	78	11.5 (0.1)[62]
$ACBP^{M46C}$	1NTI	46	8.2 (0.1)[62]
$ACBP^{S65C}$	1NTI	65	9.0 (0.1)[62]
$ACBP^{T17C}$	1NTI	17	9.8 (0.1)[62]
$ACBP^{V36C}$	1NTI	36	9.5 (0.1)[62]
$HMCK^{S285A}$	1I0E	283	6.7 (0.1)[54]
Mb^{A125C}	2MGE	125	8.43 (0.03)[63]
Mb^{G124C}	2MGE	124	6.53 (0.05)[63]
$msrA^{E115Q}$	2L90	72	8.2 (0.1)[55]
Yersinia PTP^{H402A}	1YPT	403	7.35 (0.04)[61]

Methods including traditional single-structure-based and ensemble-based sampling approaches were evaluated for their predictive accuracy in estimating experimental Cys pK_a. The pK_a methods examined include Poisson-Boltzmann continuum electrostatics-based approaches such as H++[22,23] and PypKa[26], empirical PROPKA program,[27,28] deep-learning pKAI+ predictor,[34] and constant-pH molecular dynamics simulations implemented in the Amber[64] and NAMD[65] codes. Additionally, pK_a algorithms implemented in industry leading molecular design and chemical simulation software suite (MOE[43] and Maestro[44])

were tested to access their predictive capabilities. For the ensemble-based pK_a calculations, both implicit and explicit solvent models were used for the calculations. In the following section, we provide a brief overview of the different pK_a methods used in our benchmark study. For more specific details about the input parameters used for the different methods, we refer readers to the Supporting Information.

Summary of Predictive pK_a Methods Used

H++ computes residue pK_a based on the established continuum electrostatics methodology by calculating the energetics of proton transfer of a titratable group.[22,23] The program uses atomic resolution structure as input and computes residue pK_a in addition to other molecular properties such as isoelectric point, titration curves, and protonation states. H++ is accessible via the url: http://newbiophysics.cs.vt.edu/H++/index.php

PROPKA computes residue pK_a based upon empirical relationships between factors influencing pK_a shifts and structures. More specifically, the model incorporates hydrogen bonding, desolvation, and charge-charge interaction effects into residue pK_a shifts to account for the environmental perturbation to the reference or intrinsic pK_a of a titratable group. More recent development of the model includes improved treatment of pK_a shifts in protein–ligand complexes via inductive intra- and inter-ligand coupling interactions.[27] In our study, PROPKA3[28] was used for cysteine residue pK_a prediction.

Chemical Computing Group (CCG) **MOE**[43] and Schrödinger **Maestro**[44] software suite which provide access to PROPKA program for residue pK_a prediction were also used to compute Cys pK_a's. For the Maestro software, residue pK_a is computed based upon PROPKA3 and was accessed through the Refine tab of the Protein Preparation Wizard. For the MOE pK_a application, Cys pK_a's were computed via the Protein Properties window after structure preparation and refinement. Also, ensemble pK_a calculations were also performed by sampling conformational and protonation states via LowModeMD[66] and Protonate3D[67] algorithms in MOE software (version 2022.02). Default setting pH range from 6.4–8.4 was used for the ensemble property pK_a calculations. Residue pK_a application in MOE program is based upon custom implementation of PROPKA2.[68]

PypKA is a tool to predict the pK_a values of titratable residues in proteins using Poisson-Boltzmann/Monte Carlo-based calculations.[26] The DelPhi[69] program numerically solves the Poisson-Boltzmann equation while the Monte-Carlo algorithm samples residue protonation/tautomeric states. In this work, cysteine pK_a's were calculated using the PypKA webserver which is accessible via the url: https://pypka.org/.

pKAI+ is a deep learning-based pK_a prediction tool trained on a database of residue pK_a's estimated from structures using the continuum electrostatics-based PypKa program.[34] The model was trained on a large database consisting of approximately 6 million pK_a values estimated from about 50,000 biomolecular structures. The pKAI+ model employed predicts experimental pK_a's of titratable residues from a single conformation or protein structure. The model is accessible from the GitHub repository: https://github.com/bayer-science-for-a-better-life/pKAI.

Constant pH Molecular Dynamics (CpHMD) is capable of sampling titratable residue protonation states in conjunction with conformational dynamics for accurate pK_a estimation. For this reason, CpHMD pK_a approaches are generally more computationally expensive than traditional single-structure-based pK_a methods. In this work, we employ the CpHMD methods implemented in the Amber and NAMD codes for Cys pK_a prediction. For the Amber approach, simulations were carried out using both the Generalized Born (GB) implicit solvent model[29] and explicit solvent[45] model via the pH-replica exchange MD (pH-REMD) algorithm using discrete protonation states. The protein was modelled using the Amber FF99SB[70] force field and simulations were performed in parallel mode by running pH-REMD. The pH-REMD runs consisted of 16 replicas and were run for either 2 ns or 5 ns for each pH-replica in implicit and explicit solvent, respectively. For the NAMD runs, we computed cysteine pK_a's in explicit solvent using the nonequilibrium molecular dynamics/Monte Carlo (neMD/MC) constant pH approach.[46] The CHARMMM36 protein force field was used for the simulations which were run in parallel at pH ranging 1–14 in steps of 1.0 pH unit. For each pH simulation window, we performed 10 ns sampling yielding a total sampling time of 140 ns per protein model system. More details about the protocol and the input parameters used in the atomistic CpHMD simulations can be found in the Supporting Information.

RESULTS AND DISCUSSION

We assess the accuracy of the different predictive pK_a methods employed in estimating the experimental Cys pK_a's and measure the correlation between the quantities. The different pK_a methods used were grouped into traditional single-structure-based and ensemble-based approaches. We refer to traditional single-structure-based approaches as methods that compute pK_a using a single protein structure conformation, whereas ensemble-based sampling methods couple the dynamic dependence of titratable residue pK_a/protonation state with conformational sampling. We discuss the overall performance of these methods for both the wildtype and mutant protein test sets.

Wildtype Protein Test Set.

The wildtype protein test set comprised of 16 proteins with experimental Cys pK_a's for which a PDB structure exists. Cysteine residue pK_a's were computed for the protein structure after system preparation, which includes filling in missing sidechain and loops within the protein model system.

Traditional Single-structure-based pK_a Methods.

Figure 2 shows a plot of the predicted Cys pK_a versus experiment for the wildtype protein test set using different single structure-based pK_a methods. The results show a significant variation in the predictive performance and accuracy of the different methods for Cys pK_a calculation (**Figure 2**). The average root-mean-square error (RMSE) and mean absolute error (MAE) of the pK_a predictions were 3.9 and 3.3, respectively. The results highlight intrinsic limitations and challenges in the predictive capabilities of these methods for accurate cysteine pK_a calculation, **Table 3**.

Among the different pK_a methods explored (**Figure 2**), results using the pK_a tool in the MOE program yielded the smallest deviation from experiment for the wildtype test set, **Table 3**. The MAE for the wildtype Cys pK_a predictions using the MOE pK_a tool was 2.3 pK units, which is ~1 pK unit better than the null model (**Figure 3**). All the other pK_a methods performed either similarly or worse than the null model for the wildtype test set. The null model assumes the reference pK_a of 8.6 for all cysteines predicted in the test set. The predictions from the PROPKA program gave the largest deviation from experiment (MAE = ~4.0). The poor agreement between the experimental and predicted Cys pK_a's for the empirical PROPKA program could most likely stem from weak parameterization of the method for cysteine residues due to a small amount of training data.[68] The difficulty of PROPKA in predicting experimental Cys pK_a's has been highlighted in previous studies.[38,40]

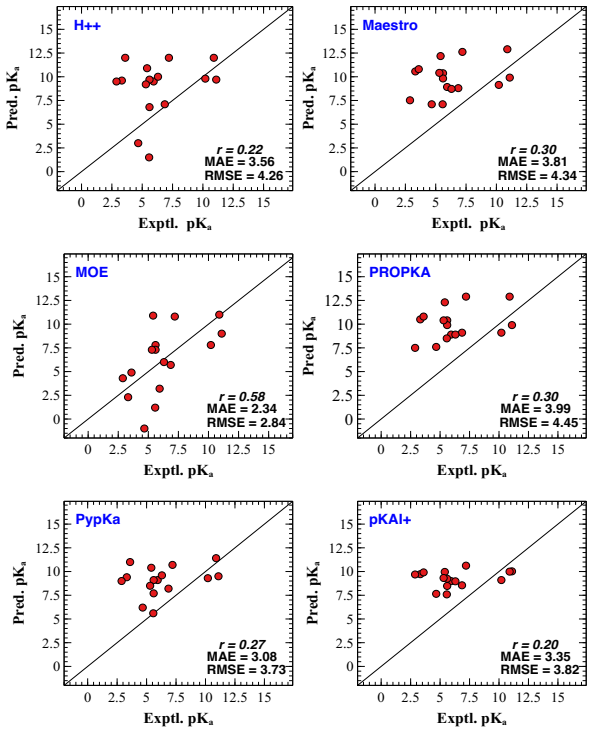


Figure 2. Predictive versus experimental Cys pK_a for wildtype protein test set using traditional single structure-based pK_a methods.

Table 3. Statistical significance in RMSE for the traditional pK_a methods used for the wildtype protein test set.

Method	σ	Range for $\sigma^2_{95\%}$
H++	4.26	$3.07 < \sigma < 6.27$
Maestro	4.34	$3.13 < \sigma < 6.40$
MOE	2.84	$2.05 < \sigma < 4.19$
PROPKA	4.45	$3.21 < \sigma < 6.56$
PypKa	3.73	$2.69 < \sigma < 5.49$
pKAI+	3.82	$2.75 < \sigma < 5.62$

Confidence limits in RMSE values (σ) were calculated using χ-squared function (Eqn. **74** of ref [71])[71] at a range of 95% for N=16.

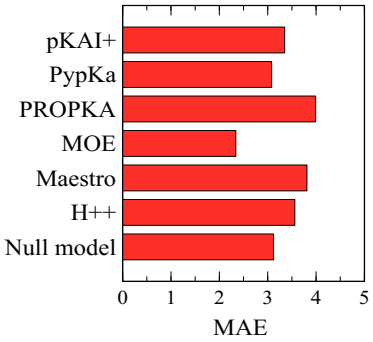


Figure 3. Mean absolute error (MAE) for Cys pK_a predictions using single structure-based pK_a methods on the wildtype protein set.

Although the MOE pK_a program showed the best predictive performance among the structure-based methods (**Figure 3**), there were a few outliers in the pK_a correlation plot (**Figure 2**). For example, cysteine pK_a's in the active site of protein tyrosine phosphatases: Cys-403 of yersinia PTP and Cys-215 of human PTP1B were significantly downshifted by 5.7 and 4.3 pK units, respectively, relative to experiment. In a similar vein, the active site Cys-106 and Cys-72 in the proteins DJ-1 and methionine sulfoxide reductase A (msrA) were overestimated by 5.5 and 3.6 pK units, respectively, relative to experiment. Both active-site cysteines have nearby charged glutamic acid residues in typical ionization fashion which destabilize the thiolate cysteine form (**Figure 4**), resulting in elevated cysteine pK_a predictions. The magnitude in predicted pK_a elevation for Cys-106 of DJ-1 relative to Cys-72 of msrA is potentially due to the closer proximity of the thiolate and carboxyl groups, **Figure 4**.

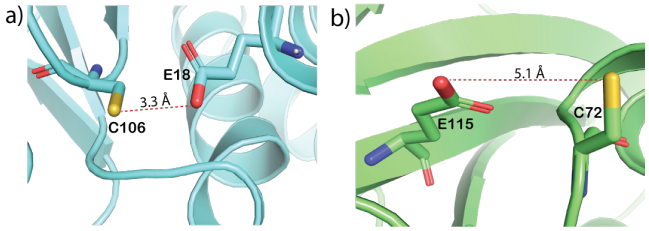


Figure 4. Representative configuration of the active site cysteine thiolates in the proteins (**a**) DJ-1 and (**b**) msrA. Both Cys-106 of DJ-1 and Cys-72 of msrA have a nearby Glu residue in the crystal structure. *Figure S2 shows the experimental XRD electron density of the Glu^{18}---Cys^{108} contact in DJ-1.*

The discrepancy between predicted and experimental Cys pK_a's is likely due to differences in protein crystal structure conformation relative to the biologically relevant state. A major limitation of traditional single structure-based pK_a methods is that they are unable to capture dynamic changes in the local environment of ionizable residues. In some cases, proteins may be trapped in nonrepresentative conformations[72] or largely coupled between conformational and protonation states.[73] It is also important to note the variability in pK_a measurements of titratable residues in proteins. The pK_a of Cys residue is strongly influenced by protein microenvironment[74] — nearby basic residues decrease Cys pK_a by stabilizing thiolate state.[75] We have examined the immediate environment of Cys residues to explore the presence of nearby basic residues (**Table S14** in SI). We hypothesize that for cases where there are no basic residues and there is a substantial drop in the reported Cys pK_a, such values are probably due to variability in experiments. For instance,

the very low experimental pK_a reported for some Cys residues in proteins (e.g., Cys-25 in ppΩ; PDB id: 1PPO) may be due to metal coordination that is present in pH titration experiments. In our pK_a calculations, such metal–Cys coordination was not considered in our model structures. These differences in structure could be a potential contributing factor to the discrepancy between predicted pK_a estimates and experiment.

Ensemble-based pK_a Methods

A primary source of inaccuracy for traditional single structure-based pK_a methods is that they only use a single protein conformation state to deduce titratable residue pK_a and protonation states. However, residue protonation state and conformation are strongly coupled to each other. Ensemble-based pK_a methods like constant-pH molecular dynamics are designed to sample multiple protein conformation and residue protonation states. To account for the coupling between protein conformational sampling and titratable residue protonation state changes, we employed constant-pH molecular dynamics simulations to calculate cysteine pK_a's. Constant-pH MD simulations can directly sample pH-induced conformational changes and its effect on titratable residue pK_a and protonation state. These methods have been shown to yield good estimates of experimental pK_a's for titratable residues, particularly Asp and Glu residues.[72,76,77]

To this effect, we applied the Amber and NAMD CpHMD codes to predict Cys pK_a's for the protein model systems studied. Both implicit and explicit solvent models were used in our simulations. We refer to the Amber CpHMD approach as Amber99SB/Amber, while the NAMD CpHMD is referred to as Charmm36/NAMD. In addition to the above CpHMD methods employed, we also performed ensemble pK_a calculations via pH-dependent protein conformation sampling using the MOE program by combining LowModeMD[66] and Protonate3D[67] algorithms. **Figure 5** depicts a summary of the predicted Cys pK_a results in comparison with experimental pK_a for the wildtype protein test set using the ensemble-based pK_a methods. **Table 4** reports the statistical significance in RMSE for the results.

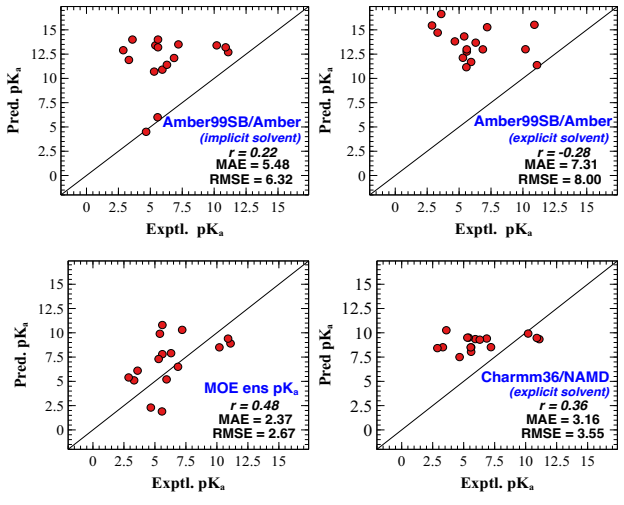


Figure 5. Predictive versus experimental Cys pK_a for a test set of wildtype proteins using ensemble-based pK_a sampling methods.

Table 4. Statistical significance in RMSE for the ensemble-based pK_a methods used for the wildtype protein test set.

Method	σ	Range for $\sigma^2_{95\%}$
Amber99SB/Amber (explicit)	8.00	$5.77 < \sigma < 11.8$
Amber99SB/Amber (implicit)	6.32	$4.56 < \sigma < 9.32$
Charmm36/NAMD (explicit)	3.55	$2.56 < \sigma < 5.23$
MOE ens pK_a	2.67	$1.93 < \sigma < 3.94$

Confidence limits in RMSE values (σ) were calculated using χ-squared function (Eqn. **74** of ref [71])[71] at a range of 95% for N=16.

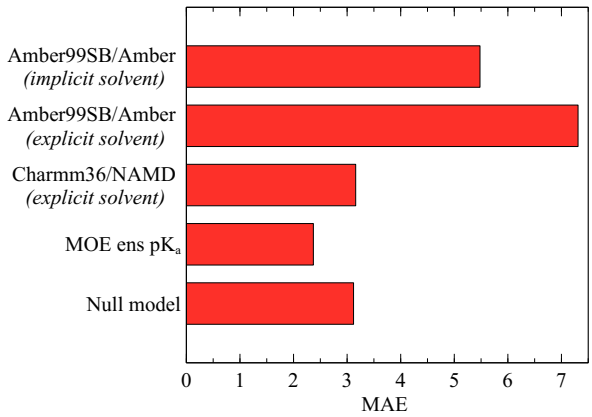


Figure 6. Mean absolute error (MAE) for test set of wildtype proteins using ensemble-based pK_a sampling methods.

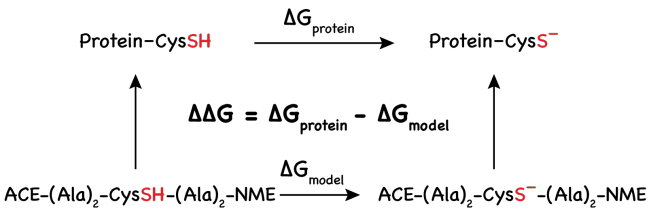
Analogous to the results obtained using the traditional structure-based pK_a approaches (**Figure 2**), there is a variation in the predictive accuracies of the different pK_a methods used (**Figure 5**). **Figure 6** shows the mean absolute error of the ensemble-based pK_a methods. Among these methods, the ensemble-average Cys pK_a results obtained using the MOE program yielded the smallest deviation from experiment (**Table 4**; **Figure 6**). The computed Cys pK_a's using the Amber99SB/Amber CpHMD code yielded the largest deviation from experiment (**Figure 6**). For both the implicit and explicit solvent models, predicted Cys pK_a's were severely overpredicted by the Amber99SB/Amber CpHMD code. This appeared to be more significant for simulations carried out in explicit solvent relative to the GB implicit solvent model, (**Figure 5**). To ensure that the large deviation in predictive pK_a results for the explicit solvent runs were not due to poor sampling, we extended the simulations by doubling the simulation time for each pH-replica window from 5 ns to 10 ns. We did not observe any significant improvement in the predictive Cys pK_a results for the extended runs (**Table S12** in SI), suggesting that the fundamental limitation in the accuracy of the method is not as a result of poor conformational sampling.

Although the mean absolute deviation is lower for the Charmm36/NAMD pK_a predictions relative to the Amber results (**Figure 6**), the predicted Cys pK_a values for the Charmm36/NAMD method have a narrow dynamic range. So, both methods are generally poor for accurate Cys pK_a prediction, with no predictive relative ranking among computed pK_a's. A plausible reason for the inaccuracy in the Amber99SB/Amber CpHMD pK_a predictions can be attributed to limitations in the force field model to describe cysteine thiol and

thiolate parameters distinctly.[78] In previous studies, we have observed that Cys pK_a calculations computed using all-atom RETI pK_a approach in GROMACS yielded slightly better performance when Charmm36 force field (RMSD= 2.40) was used relative to Amber99SB force field (RMSD= 3.20) for a test set oof 18 experimental cysteine pK_a's.[38] The Amber force field model uses the same Lennard-Jones (LJ) sigma and epsilon parameters for both thiol and thiolate forms of cysteine (see **Table S1** in SI) — a limitation which has been shown to impact protein pK_a calculations and the hydration structure of model thiolates in free energy calculations.[78] This raises the need for the development of improved force field parameters to enable accurate Cys pK_a calculations. Recent work by Roitberg, Estrin and coworkers[79] have developed novel set of LJ parameters for cysteine relevant species for use in classical Amber MD simulations. The parameters were derived to reproduce solute-water radial pair distribution functions *g(r)* (RDF) from *ab initio* molecular dynamics.[79]

To investigate the impact of the improved parameters on the predictive cysteine pK_a results, we performed Amber thermodynamic integration (TI) MD calculations for select few systems following the thermodynamic cycle depicted in **Scheme 1**, using both default and modified Amber parameters. The modified cysteine thiolate Amber parameters were derived from the recent work of Estrin and coworkers[79] (see **Table S2** in SI for details).

Scheme 1. Thermodynamic cycle used for the Amber-TI cysteine pK_a calculations. The model system used is alanine pentapeptide with capped termini. *$\Delta\Delta G$ refers to the relative free energy difference between the protein and the model peptide.*



For the select protein systems explored, our preliminary results from the Amber-TI MD calculations suggest a slight improvement in predictive performance when using the modified cysteine thiolate parameters relative to the default ones (**Table S3**). On average, we observed an improvement of 0.5 ΔpK_a units in the predicted pK_a error estimates when the modified parameters were used. However, the calculated Cys pK_a shifts are still elevated relative to experiment (**Table S3** in SI). The observed trend in elevation of the predicted Cys pK_a estimates for the Amber-TI calculations correlates with the previous results from the Amber constant-pH MD simulations. Although the quality and variety of the test set is quite limiting to draw any meaningful conclusions from these results, our findings suggest that more rigorous validation and parametrization efforts are needed for accurate Cys pK_a calculations. In particular, for use of the improved cysteine thiolate parameters in CpHMD simulations, the cysteine GB parameters would need to be updated. In addition, further improvement and a complete assessment of the parameters is needed, as they were developed to reproduce Cys sulfur and water oxygen interactions, and evaluated in a specific system of interest.[79] Overall, the analysis indicates that proper description and parameterization of cysteine thiol/thiolate parameters are required to achieve reliable results in MD-based simulations, including pK_a calculations — highlighting the need for improvement in force field parameters for accurate Cys pK_a prediction.

The best performing method among the ensemble-based pK_a sampling approaches employed for the wildtype test set was the MOE ensemble pK_a approach — which yielded an MAE of 2.4 pK units (**Figure 6**). This is on par with the MOE pK_a results obtained earlier where protein conformation sampling was not considered. We note that other factors beyond conformational changes in protein structure, including the variability in pH measurements and computational methodology may be contributing to the observed limited performance. In some cases, the deviation in the predicted Cys pK_a's from experiment slightly improved when pH-dependent conformational sampling was introduced, for example the active site Cys-403 of yersinia PTP. In this case, the predictive pK_a performance improved by 3.4 pK units upon sampling protein conformational and protonation states, although the predicted pK_a of Cys-403 was still downshifted by 2.4 pK units from experiment. This is better than the initial ~6 pK units downshift predicted relative to experiment for the single structure MOE pK_a approach. Visual inspection of representative configurations from both pK_a approaches suggest that rearrangement of protein side chains occur such that the active-site Cys-403 in yersinia PTP adopts a different orientation in both states (**Figure 7**).

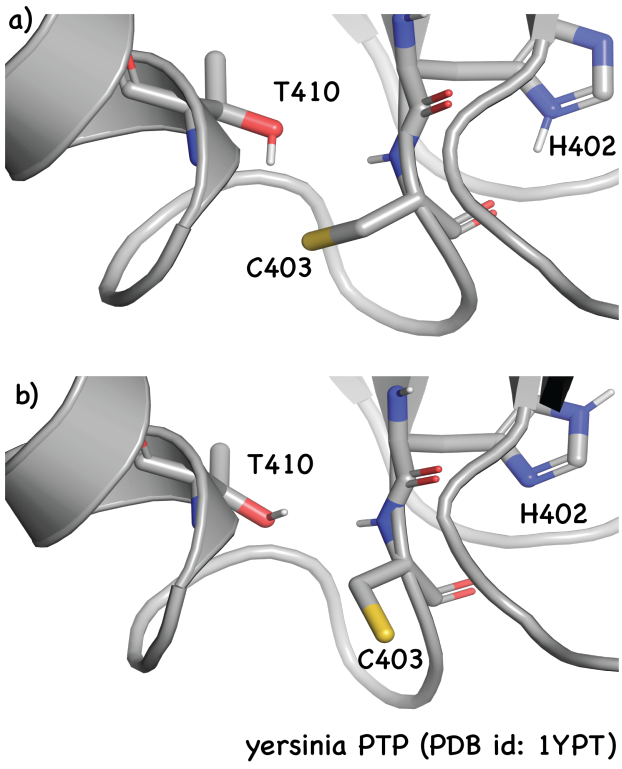


Figure 7. Representative configuration of the active-site Cys-403 in yersinia PTP from single structure **(a)** and ensemble-based (**b**) MOE pK_a calculations. The active site Cys-403 adopts a different position and orientation in both states, which leads to significantly different predicted cysteine pK_a's.

The representative position and orientation of Cys-403 of yersinia PTP in the ensemble pK_a approach is different from the single structure-based pK_a approach (**Figure 7**). Although adequate conformational and protonation state sampling is required for accurate pK_a prediction, the results highlight the importance of conformational sampling and structure dynamics effects in pK_a calculations, particularly for cysteines. This is particularly important for cysteines in catalytic environments (i.e., dyad and triad systems) where other residue protonation and rotameric/tautomeric states (e.g., histidine) can be largely coupled to one another and can influence pK_a.

Mutant Protein Test set

The mutant protein test set consisted of 10 proteins which have been listed in **Table 2**. For the mutant test set of proteins, crystal structure files were not available, so single point mutations were introduced in the wildtype proteins. The computationally mutated proteins were preprocessed using Protein Preparation Wizard in the Schrödinger Maestro program prior to cysteine pK_a calculation.

Single Structure and Ensemble-based pK_a Methods

Figures 8 and **9** depict the correlation plot between the experimental and predicted Cys pK_a's using the single structure-based and ensemble pK_a methods for the mutant protein test set. Both classes of pK_a methods appear to yield reasonable predictive performance (average MAE of ~2 pK_a shifts) in comparison with the wildtype Cys pK_a results, (**Figure 5**). Although the average performance of these methods seems encouraging for this test set, it is important to note that the experimental values have a narrow dynamic range for this set (6.5 –11.5), **Table 2**. The average ΔpK_a of the test set is ~1 pK_a unit, which lies within the distribution of the reference solution Cys pK_a. So, this data set is not representative and comprehensive enough to capture the true accuracy and predictive performance of the different methods. Thus, no meaningful conclusions about the predictive capabilities of these methods can be drawn from the results since they essentially predict the null cysteine pK_a. In other words, the performance of these methods is on par or worse than the null model — so they are not being predictive at all given the narrow dynamic range in residue pK_a's. The availability of more extensive cysteine pK_a datasets could help to better inform the predictive accuracies of these methods.

Although the average performance of these methods seems encouraging for this test set, it is important to note that the experimental values have a narrow dynamic range for this set (6.5 – 11.5), **Table 2**. The average ΔpK_a of the test set is ~1 pK_a unit, which lies within the distribution of the reference solution Cys pK_a. So, this data set is not representative and comprehensive enough to capture the true accuracy and predictive performance of the different methods. Thus, no meaningful conclusions about the predictive capabilities of these methods can be drawn from the results since they essentially predict the null cysteine pK_a. In other words, the performance of these methods is on par or worse than the null model — so they are not being predictive at all given the narrow dynamic range in residue pK_a's. The availability of more extensive cysteine pK_a datasets could help to better inform the predictive accuracies of these methods.

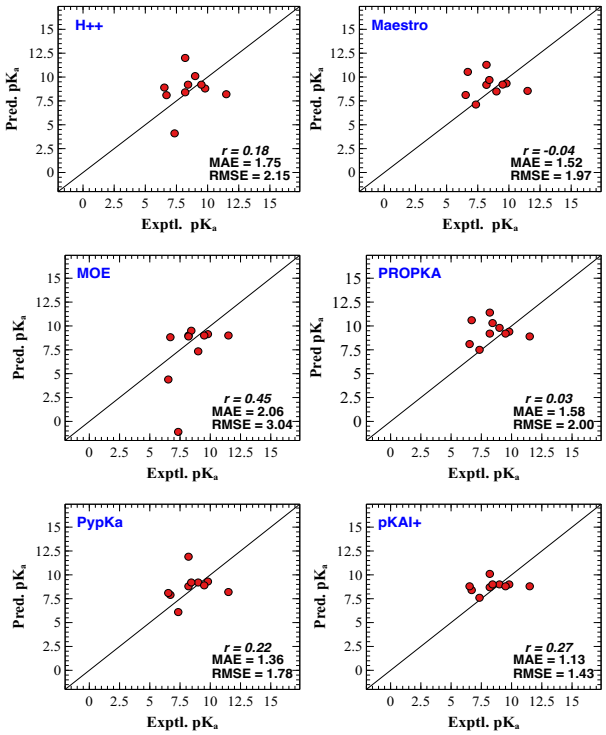


Figure 8. Cys pK_a results for the mutant protein test set using static-based pK_a methods.

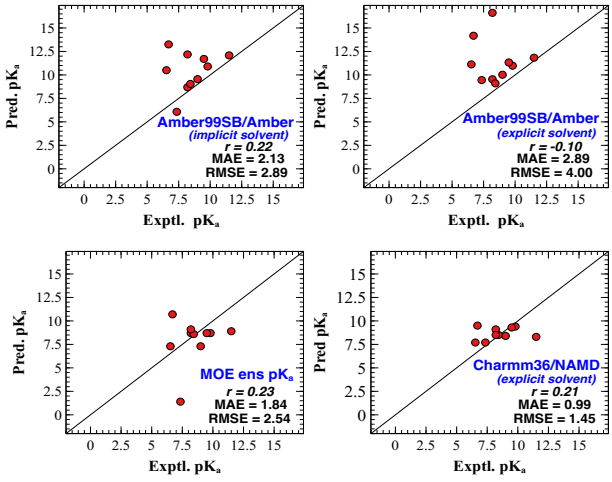
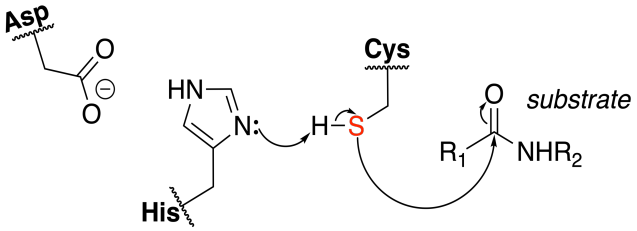


Figure 9. Cys pK_a results for mutant protein test set using different MD-based pK_a methods.

Limitations of *in silico* Methods for Cys pK_a Calculations

The results obtained in this study highlight intrinsic limitations in existing methods for accurate cysteine pK_a prediction. Issues stemming from multiple factors, including poor pK_a models to inadequate protein conformational space and protonation state sampling are plausible factors, to name a few. Not to mention, the variability in experimental pK_a measurements via different techniques and the lack of rigorous test sets for cysteine pK_a

validation. In addition, accurate description of cysteine thiol/thiolate parameters is lacking in conventional MM force fields and simulated protein structures may deviate from the biologically relevant or native state structure leading to discrepancies between predicted cysteine pK_a's and experiment. For catalytic-site cysteines which typically have depressed pK_a's, correct assignment of the protonation/rotameric/tautomeric states of neighboring residues (e.g., histidine) is crucial to achieve reliable pK_a results. This is because the protonation states of residues in these microenvironments are typically coupled to one another, which can have significant effect on the resultant pK_a. A classic example includes active-site cysteines (or serine residues) comprising catalytic dyad/triad systems in enzymes such as proteases, where neighboring acidic (e.g., Asp/Glu) and basic (i.e., His) residues polarize and activate the nucleophile for covalent catalysis, **Scheme 2**. This effect of induced polarization stabilizes the charged state of the nucleophile — lowering the resultant pK_a. Polarizable force fields[80,81] may be better suited for describing residue pK_a's in highly-charged catalytic enzyme sites which is a challenge for conventional force fields.



Scheme 2. Schematic representation of the classic catalytic triad mechanism showing the coordinated network of residues that lead to the polarization and activation of a cysteine nucleophile for covalent modification.

We note that we do not observe marked discrepancies between predicted and experimental pK_a's when these methods are used to estimate pK_a's for titratable residues other than cysteines (e.g., Asp, Glu, His, Lys, Tyr). To demonstrate this, we computed the pK_a's of Asp, Glu, His, and Lys residues in the classic hen egg white lysozyme (HEWL) system using all the methods discussed in this work. The results suggest that these methods are fairly accurate in predicting experimental pK_a's for residues in this system (average RMSE = 0.84 pK units), **Figure S1**. The contrast in the performance of these methods for cysteines relative to other titratable residues may reflect a poorer description in the underlying physics that is missing in these models or a greater complexity in the acid/base chemistry of cysteines.

Moving forward, the adoption and development of better pK_a models that capture the complex electrostatic microenvironment in proteins could prove useful in tackling the limitations in existing pK_a models for accurate cysteine pK_a prediction. In addition, more extensive sampling and coupling of the dynamic dependence in protein conformation and protonation states is another opportunity for improvement for these models. Lastly, the availability of large and comprehensive datasets of experimental cysteine pK_a's will enable rigorous validation of pK_a methods, including emerging machine learning-based pK_a predictors.

CONCLUSION

In summary, we have employed a broad range of pK_a tools to assess their predictive performance in accurately estimating cysteine pK_a's for a test set of proteins collected from the PKAD database. The protein test set consisted of 16 wildtype and 10 computationally mutated proteins with experimentally measured cysteine pK_a's. We examined traditional single-structure-based and ensemble-based pK_a approaches, including a deep learning-based pK_a prediction tool, pKAI+. Overall, the results highlight intrinsic limitations in the accuracy and predictive performance of *in silico* methods for cysteine pK_a calculation. For the wildtype test set of proteins, the performance of the best method (MOE) yielded a root-mean-squared error of 2.7 pK units. Although we observed a slightly better overall performance for the mutant test set, no meaningful conclusion could be drawn from the results given the narrow distribution of residue pK_a shift for this set (avg. $|\Delta pK_a|$ is ~1 unit from the reference solution Cys pK_a). The ensemble-based sampling and advanced CpHMD approaches did not significantly improve the accuracy of the Cys pK_a predictions for the test set evaluated. In particular, we found that the Amber CpHMD code using discrete protonation states greatly overestimated predicted Cys pK_a's — yielding the most significant deviation from experiment among the pK_a methods evaluated. We posit this is due to a poor description of the cysteine thiol/thiolate force field parameters in Amber MD force field, particularly LJ parameters. Improvement in the force field description of cysteine parameters could yield more accurate pK_a results for MD-based simulations. The continued development and rigorous evaluation of these methods on comprehensive datasets of experimental cysteine pK_a will go a long way to inform and improve their predictive capabilities. Progress in the calculation and prediction of cysteine pK_a's for drug discovery will require collaborative efforts from experimentalists and computational scientists,[37,82] — redefining the conceptual framework underpinning the complexity in acid-base chemistry of cysteines in biomolecules.

DATA AND SOFTWARE AVAILABILITY

Additional data and results for all the calculations performed are available in the Supporting Information. The PDB files used for the different cysteine pK_a calculations can be found at: https://github.com/awoonor/Cysteine_pKa_PDB_files.

The pK_a methods explored span a broad range of classes, which include continuum electrostatics-based methods to state-of-the-art enhanced sampling constant-pH MD approaches. The pK_a method include H++ (http://newbiophysics.cs.vt.edu/H++/), PROPKA (https://github.com/jensengroup/propka), PypKA (https://pypka.org/run-pypka/), and PKAI+ (https://github.com/bayer-science-for-a-better-life/pKAI). CCG **MOE** (https://www.chemcomp.com/) and Schrödinger **Maestro** (https://www.schrodinger.com/products/maestro) software suite were also used to compute cysteine residue pK_a's. Ensemble-based pK_a approaches were computed using popular Amber (https://ambermd.org/) and NAMD (http://www.ks.uiuc.edu/Research/namd/) software packages.

ASSOCIATED CONTENT

Supporting Information

Summary of the computed pK_a values of target Cys residues in the proteins studied, details of the pK_a calculations performed, including description of the constant pH molecular dynamics simulations and Amber-TI pK_a calculations and input parameters. (PDF).

AUTHOR INFORMATION

Corresponding Author

*** Ernest Awoonor-Williams** – Computer-Aided Drug Discovery, Global Discovery Chemistry, Novartis Institutes for BioMedical Research, 181 Massachusetts Avenue, Cambridge, MA 02139, United States.
Email: ernest.awoonor-williams[at]novartis.com

Author Contributions

E.A.-W conceptualized the research project and was the primary contributor to writing the manuscript. All authors were involved in reviewing the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

The authors thank Prof. Christopher Rowley, Dr. Callum Dickson, and Dr. Michael Schaefer for helpful discussions, including reviewing and providing valuable feedback to improve the manuscript. The authors also thank the reviewers for their constructive feedback which helped to improve the manuscript. E.A.-W thanks Dr. Sepehr Dehghani-Ghahnaviyeh for helpful discussions about Amber-TI GPU implementation. E.A.-W acknowledges the support of the Innovation Postdoctoral Fellowship Program at the Novartis Institutes for BioMedical Research.

ABBREVIATIONS

Asp, aspartate; Cys, cysteine; Glu, glutamate; His, histidine; Lys, lysine; Tyr, tyrosine; CpHMD, constant pH molecular dynamics; ens, ensemble; EWG, electron withdrawing group; GB, generalized Born; LJ, Lennard-Jones; MAE, mean absolute error; MD, molecular dynamics; neMD/MC, nonequilibrium Molecular Dynamics/Monte Carlo; MOE; Molecular Operating Environment; NMR, nuclear magnetic resonance; PDB; protein data bank; REMD, replica exchange molecular dynamics; RMSE, root-mean-square error; TI, thermodynamic integration; WT, wildtype.

REFERENCES

(1) Matthew, J. B.; Gurd, F. R.; Garcia-Moreno, B.; Flanagan, M. a; March, K. L.; Shire, S. J. PH-Dependent Processes in Proteins. *CRC Crit. Rev. Biochem.* **1985**, *18*, 91–197.

(2) Gutteridge, A.; Thornton, J. M. Understanding Nature's Catalytic Toolkit. *Trends Biochem. Sci.* **2005**, *30*, 622–629.

(3) Pace, N. J.; Weerapana, E. Diverse Functional Roles of Reactive Cysteines. *ACS Chem. Biol.* **2013**, *8*, 283–296.

(4) Giles, N. M.; Giles, G. I.; Jacob, C. Multiple Roles of Cysteine in Biocatalysis. *Biochem. Biophys. Res. Commun.* **2003**, *300*, 1–4.

(5) Giles, N. M.; Watts, A. B.; Giles, G. I.; Fry, F. H.; Littlechild, J. A.; Jacob, C. Metal and Redox Modulation of Cysteine Protein Function. *Chem. Biol.* **2003**, *10*, 677–693.

(6) Liu, Q.; Sabnis, Y.; Zhao, Z.; Zhang, T.; Buhrlage, S. J.; Jones, L. H.; Gray, N. S. Developing Irreversible Inhibitors of the Protein Kinase Cysteinome. *Chem. Biol.* **2013**, *20*, 146–159.

(7) Bauer, R. A. Covalent Inhibitors in Drug Discovery: From Accidental Discoveries to Avoided Liabilities and Designed Therapies. *Drug Discov. Today* **2015**, *20*, 1061–1073.

(8) Awoonor-Williams, E. Modelling Covalent Modification of Cysteine Residues in Proteins, Memorial University of Newfoundland, 2020.

(9) Baillie, T. A. Targeted Covalent Inhibitors for Drug Design. *Angew. Chemie - Int. Ed.* **2016**, *55*, 13408–13421.

(10) Visscher, M.; Arkin, M. R.; Dansen, T. B. Covalent Targeting of Acquired Cysteines in Cancer. *Curr. Opin. Chem. Biol.* **2016**, *30*, 61–67.

(11) Wu, S.; Luo Howard, H.; Wang, H.; Zhao, W.; Hu, Q.; Yang, Y. Cysteinome: The First Comprehensive Database for Proteins with Targetable Cysteine and Their Covalent Inhibitors. *Biochem. Biophys. Res. Commun.* **2016**, 6–11.

(12) Awoonor-Williams, E.; Walsh, A. G.; Rowley, C. N. Modeling Covalent-Modifier Drugs. *Biochim. Biophys. Acta - Proteins Proteomics* **2017**, *1865*, 1664–1675.

(13) Awoonor-Williams, E.; Rowley, C. N. How Reactive Are Druggable Cysteines in Protein Kinases? *J. Chem. Inf. Model.* **2018**, *58*, 1935–1946.

(14) Gehringer, M.; Laufer, S. A. Emerging and Re-Emerging Warheads for Targeted Covalent Inhibitors: Applications in Medicinal Chemistry and Chemical Biology. *J. Med. Chem.* **2019**, *62*, 5673–5724.

(15) Awoonor-Williams, E.; Isley, W. C.; Dale, S. G.; Johnson, E. R.; Yu, H.; Becke, A. D.; Roux, B.; Rowley, C. N. Quantum Chemical Methods for Modeling Covalent Modification of Biological Thiols. *J. Comput. Chem.* **2020**, *41*, 427–438.

(16) Awoonor-Williams, E.; Rowley, C. N. Modeling the Binding and Conformational Energetics of a Targeted Covalent Inhibitor to Bruton's Tyrosine Kinase. *J. Chem. Inf. Model.* **2021**, *61*, 5234–5242.

(17) Ferrer-Sueta, G.; Manta, B.; Botti, H.; Radi, R.; Trujillo, M.; Denicola, A. Factors Affecting Protein Thiol Reactivity and Specificity in Peroxide Reduction. *Chem. Res. Toxicol.* **2011**, *24*, 434–450.

(18) Awoonor-Williams, E.; Kennedy, J.; Rowley, C. N. Measuring and Predicting Warhead and Residue Reactivity. In *Annual Reports in Medicinal Chemistry*; Ward, R. A., Grimster, N. P., Eds.; Elsevier Inc., 2021; pp 203–227.

(19) Roseli, R. B.; Keto, A. B.; Krenske, E. H. Mechanistic Aspects of Thiol Additions to Michael Acceptors: Insights from Computations. *WIREs Comput. Mol. Sci.* **2022**.

(20) Watt, S. K. I.; Charlebois, J. G.; Rowley, C. N.; Keillor, J. W. A Mechanistic Study of Thiol Addition to N -Phenylacrylamide. *Org. Biomol. Chem.* **2022**.

(21) Reijenga, J.; van Hoof, A.; van Loon, A.; Teunissen, B. Development of Methods for the Determination of PKa Values. *Anal. Chem. Insights* **2013**, *8*, 53–71.

(22) Gordon, J. C.; Myers, J. B.; Folta, T.; Shoja, V.; Heath, L. S.; Onufriev, A. H++: A Server for Estimating PKas and Adding Missing Hydrogens to Macromolecules. *Nucleic Acids Res.* **2005**, *33*, 368–371.

(23) Anandakrishnan, R.; Aguilar, B.; Onufriev, A. V. H++ 3.0: Automating PK Prediction and the Preparation of Biomolecular Structures for Atomistic Molecular Modeling and Simulations. *Nucleic Acids Res.* **2012**, *40*, 537–541.

(24) Song, Y.; Mao, J.; Gunner, M. R. MCCE2: Improving Protein PKa Calculations with Extensive Side Chain Rotamer Sampling. *J. Comput. Chem.* **2009**, *30*, 2231–2247.

(25) Wang, L.; Zhang, M.; Alexov, E. DelPhiPKa Web Server: Predicting p K a of Proteins, RNAs and DNAs. *Bioinformatics* **2016**, *32*, 614–615.

(26) Reis, P. B. P. S.; Vila-Viçosa, D.; Rocchia, W.; Machuqueiro, M. PypKa: A Flexible Python Module for Poisson–Boltzmann-Based p K a Calculations. *J. Chem. Inf. Model.* **2020**, *60*, 4442–4448.

(27) Sondergard, C. R.; Olsson, M. H. M.; Rostkowski, M.; Jensen, J. H. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of PKa Values. *J. Chem. Theory Comput.* **2011**, *7*, 2284–2295.

(28) Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical PKa Predictions. *J. Chem. Theory Comput.*

2011, *7*, 525–537.
(29) Mongan, J.; Case, D. A.; McCammon, J. A. Constant PH Molecular Dynamics in Generalized Born Implicit Solvent. *J. Comput. Chem.* **2004**, *25*, 2038–2048.
(30) Riccardi, D.; Schaefer, P.; Cui, Q. P K a Calculations in Solution and Proteins with QM/MM Free Energy Perturbation Simulations: A Quantitative Test of QM/MM Protocols. *J. Phys. Chem. B* **2005**, *109*, 17715–17733.
(31) Cai, Z.; Luo, F.; Wang, Y.; Li, E.; Huang, Y. Protein p K a Prediction with Machine Learning. *ACS Omega* **2021**, *6*, 34823–34831.
(32) Gokcan, H.; Isayev, O. Prediction of Protein p K a with Representation Learning. *Chem. Sci.* **2022**, *13*, 2462–2474.
(33) Chen, A. Y.; Lee, J.; Damjanovic, A.; Brooks, B. R. Protein p K a Prediction by Tree-Based Machine Learning. *J. Chem. Theory Comput.* **2022**, *18*, 2673–2686.
(34) Reis, P. B. P. S.; Bertolini, M.; Montanari, F.; Rocchia, W.; Machuqueiro, M.; Clevert, D.-A. A Fast and Interpretable Deep Learning Approach for Accurate Electrostatics-Driven p K a Predictions in Proteins. *J. Chem. Theory Comput.* **2022**.
(35) Stanton, C. L.; Houk, K. N. Benchmarking p K Prediction Methods for Residues in Proteins Benchmarking p K a Prediction Methods for Residues in Proteins. **2008**, *4*, 951–966.
(36) Lee, A. C.; Crippen, G. M. Predicting PKa. *J. Chem. Inf. Model.* **2009**, *49*, 2013–2033.
(37) Alexov, E.; Mehler, E. L.; Baker, N.; Baptista, A. M.; Huang, Y.; Milletti, F.; Nielsen, J. E.; Farrell, D.; Carstensen, T.; Olsson, M. H. M.; Shen, J. K.; Warwicker, J.; Williams, S.; Word, J. M. Progress in the Prediction of PKa Values in Proteins. *Proteins* **2011**, *79*, 3260–3275.
(38) Awoonor-Williams, E.; Rowley, C. N. Evaluation of Methods for the Calculation of the PKa of Cysteine Residues in Proteins. *J. Chem. Theory Comput.* **2016**, *12*, 4662–4673.
(39) Pahari, S.; Sun, L.; Basu, S.; Alexov, E. DelPhiPKa: Including Salt in the Calculations and Enabling Polar Residues to Titrate. *Proteins Struct. Funct. Bioinforma.* **2018**, *86*, 1277–1283.
(40) Harris, R. C.; Liu, R.; Shen, J. Predicting Reactive Cysteines with Implicit-Solvent-Based Continuous Constant PH Molecular Dynamics in Amber. *J. Chem. Theory Comput.* **2020**, *16*, 3689–3698.
(41) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1–2*, 19–25.
(42) Pahari, S.; Sun, L.; Alexov, E. PKAD: A Database of Experimentally Measured PKa Values of Ionizable Groups in Proteins. *Database* **2019**, *2019*.
(43) Molecular Operating Environment (MOE), 2022.02 Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910. Montreal, QC H3A 2R7, Canada, 2022.
(44) Schrödinger Release 2022-3: Maestro, Schrödinger, LLC. New York, NY, 2022.
(45) Swails, J. M.; York, D. M.; Roitberg, A. E. Constant PH Replica Exchange Molecular Dynamics in Explicit Solvent Using Discrete Protonation States: Implementation, Testing, and Validation. *J. Chem. Theory Comput.* **2014**, *10*, 1341–1352.
(46) Radak, B. K.; Chipot, C.; Suh, D.; Jo, S.; Jiang, W.; Phillips, J. C.; Schulten, K.; Roux, B. Constant-PH Molecular Dynamics Simulations for Large Biomolecular Systems. *J. Chem. Theory Comput.* **2017**, *13*, 5933–5944.
(47) Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
(48) Schrödinger Release 2022-3: Prime, Schrödinger, LLC. New York, NY, 2022.
(49) Thurlkill, R. L.; Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. PK Values of the Ionizable Groups of Proteins. *Protein Sci.* **2006**, *15*, 1214–1218.
(50) Griffiths, S. W.; King, J.; Cooney, C. L. The Reactivity and Oxidation Pathway of Cysteine 232 in Recombinant Human ??1-Antitrypsin. *J. Biol. Chem.* **2002**, *277*, 25486–25492.
(51) Nelson, K. J.; Parsonage, D.; Hall, A.; Karplus, P. A.; Poole, L. B.; Nelson, K. J.; Parsonage, D.; Hall, A.; Karplus, P. A.; Poole, L. B. Cysteine p K Values for the Bacterial Peroxiredoxin AhpC Cysteine p K a Values for the Bacterial Peroxiredoxin AhpC †,‡. **2008**, *47*, 12860–12868.
(52) Hasnain, S.; Hirama, T.; Tam, A.; Mort, J. S. Characterization of Recombinant Rat Cathepsin B and Nonglycosylated Mutants Expressed in Yeast. New Insights into the PH Dependence of Cathepsin B-Catalyzed Hydrolyses. *J. Biol. Chem.* **1992**, *267*, 4713–4721.
(53) Witt, A. C.; Lakshminarasimhan, M.; Remington, B. C.; Hasim, S.; Pozharski, E.; Wilson, M. A. Cysteine PKa Depression by a Protonated Glutamic Acid in Human DJ-1. *Biochemistry* **2008**, *47*, 7430–7440.
(54) Wang, P.-F.; McLeish, M. J.; Kneen, M. M.; Lee, G.; Kenyon, G. L. An Unusually Low p K a for Cys282 in the Active Site of Human Muscle Creatine Kinase †. *Biochemistry* **2001**, *40*, 11698–11705.
(55) Lim, J. C.; Gruschus, J. M.; Kim, G.; Berlett, B. S.; Tjandra, N.; Levine, R. L. A Low PK Cysteine at the Active Site of Mouse Methionine Sulfoxide Reductase A. *J. Biol. Chem.* **2012**, *287*, 25596–25601.
(56) Guengerich, F. P.; Fang, Q.; Liu, L.; Hachey, D. L.; Pegg, A. E. O 6 -Alkylguanine-DNA Alkyltransferase : Low p K a and High Reactivity Of. **2003**, *1*, 10965–10970.
(57) Pinitglang, S.; Watts, A. B.; Patel, M.; Reid, J. D.; Noble, M. A.; Gul, S.; Bokth, A.; Naeem, A.; Patel, H.; Thomas, E. W.; Sreedharan, S. K.; Verma, C.; Brocklehurst, K. A Classical Enzyme Active Center Motif Lacks Catalytic Competence until Modulated Electrostatically. *Biochemistry* **1997**, *36*, 9968–9982.
(58) Forman-Kay, J. D.; Clore, G. M.; Gronenborn, A. M. Relationship between Electrostatics and Redox Function in Human Thioredoxin: Characterization of PH Titration Shifts Using Two-Dimensional Homo- and Heteronuclear NMR. *Biochemistry* **1992**, *31*, 3442–3452.
(59) Lohse, D. L.; Denu, J. M.; Santoro, N.; Dixon, J. E. Roles of Aspartic Acid-181 and Serine-222 in Intermediate Formation and Hydrolysis of the Mammalian Protein-Tyrosine-Phosphatase PTP1. *Biochemistry* **1997**, *36*, 4568–4575.
(60) Tolbert, B. S.; Tajc, S. G.; Webb, H.; Snyder, J.; Nielsen, J. E.; Miller, B. L.; Basavappa, R. The Active Site Cysteine of Ubiquitin-Conjugating Enzymes Has a Significantly Elevated PKa: Functional Implications. *Biochemistry* **2005**, *44*, 16385–16391.
(61) Zhang, Z.-Y. Y.; Dixon, J. E. Active Site Labeling of the Yersinia Protein Tyrosine Phosphatase: The Determination of the PKa of the Active Site Cysteine and the Function of the Conserved Histidine 402. *Biochemistry* **1993**, *32*, 9340–9345.
(62) Jensen, K. S.; Pedersen, J. T.; Winther, J. R.; Teilum, K. The p K a Value and Accessibility of Cysteine Residues Are Key Determinants for Protein Substrate Discrimination by Glutaredoxin. *Biochemistry* **2014**, *53*, 2533–2540.
(63) Quillin, M. L.; Arduini, R. M.; Olson, J. S.; Phillips, G. N. High-Resolution Crystal Structures of Distal Histidine Mutants of Sperm Whale Myoglobin. *Journal of Molecular Biology*. 1993, pp 140–155.
(64) Case, D. A.; Belfon, K.; Ben-Shalom, I. Y.; Brozell, S. R.; Cerutti, D. S.; T.E. Cheatham, I.; Cruzeiro, V. W. D.; Darden, T. A.; Duke, R. E.; Giambasu, G.; G.; Gilson, M. K.; Gohlke, H.; Goetz A. W.; Harris, R.; Izadi, S.; Izmailov, S. A.; Kasavajhala, K.; Kovalenko, A.; Krasny, R.; Kurtzman, T.; Lee, T. S.; LeGrand, S.; Li, P.; Lin, C.; Liu, J.; Luchko, T.; Luo, R.; Man, V.; Merz, K. M.; Miao, O.; Mikhailovskii, O.; Monard, G.; Nguyen, H.; Onufriev, A.; Pan, F.; Pantano, S.; Qi, R.; Roe, D. R.; Roitberg, A.; Sagui, C.; Schott-Verdugo, S.; Shen, J.; Simmerling, C. L.; Skrynnikov, N. R.; Smith, J.; Swails, J.; Walker, R. C.; Wang, J.; Wilson, L.; Wolf, R. M.; Wu, X.; Xiong, Y.; Xue, Y.; York, D. M.; Kollman, P. A. (2020) Amber 2020, University of California, San Francisco.
(65) Phillips, J. C.; Hardy, D. J.; Maia, J. D. C.; Stone, J. E.; Ribeiro, J. V.; Bernardi, R. C.; Buch, R.; Fiorin, G.; Hénin, J.; Jiang, W.; McGreevy, R.; Melo, M. C. R.; Radak, B. K.; Skeel, R. D.; Singharoy, A.; Wang, Y.; Roux, B.; Aksimentiev, A.; Luthey-Schulten, Z.; Kalé, L. V.; Schulten, K.; Chipot, C.; Tajkhorshid, E. Scalable Molecular Dynamics on CPU and GPU Architectures with NAMD. *J. Chem. Phys.* **2020**, *153*.
(66) Labute, P. LowModeMD—Implicit Low-Mode Velocity Filtering Applied to Conformational Search of Macrocycles and Protein Loops. *J. Chem. Inf. Model.* **2010**, *50*, 792–800.
(67) Labute, P. Protonate3D: Assignment of Ionization States and

Hydrogen Coordinates to Macromolecular Structures. *Proteins Struct. Funct. Bioinforma.* **2009**, *75*, 187–205.
(68) Li, H.; Robertson, A. D.; Jensen, J. H. Very Fast Empirical Prediction and Rationalization of Protein PKa Values. *Proteins* **2005**, *61*, 704–721.
(69) Rocchia, W.; Alexov, E.; Honig, B. Extending the Applicability of the Nonlinear Poisson−Boltzmann Equation: Multiple Dielectric Constants and Multivalent Ions. *J. Phys. Chem. B* **2001**, *105*, 6507–6514.
(70) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins Struct. Funct. Bioinforma.* **2006**, *65*, 712–725.
(71) Nicholls, A. Confidence Limits, Error Bars and Method Comparison in Molecular Modeling. Part 1: The Calculation of Confidence Intervals. *J. Comput. Aided. Mol. Des.* **2014**, *28*, 887–918.
(72) Sequeira, J. G. N.; Rodrigues, F. E. P.; Silva, T. G. D.; Reis, P. B. P. S.; Machuqueiro, M. Extending the Stochastic Titration CpHMD to CHARMM36m. *J. Phys. Chem. B* **2022**, *126*, 7870–7882.
(73) Sarkar, A.; Roitberg, A. E. PH-Dependent Conformational Changes Lead to a Highly Shifted p K a for a Buried Glutamic Acid Mutant of SNase. *J. Phys. Chem. B* **2020**, *124*, 11072–11080.
(74) Roos, G.; Foloppe, N.; Messens, J. Understanding the PK(a) of Redox Cysteines: The Key Role of Hydrogen Bonding. *Antioxid. Redox Signal.* **2013**, *18*, 94–127.
(75) Poole, L. B. The Basics of Thiols and Cysteines in Redox Biology and Chemistry. *Free Radic. Biol. Med.* **2015**, *80*, 148–157.
(76) Wallace, J. A.; Shen, J. K. *Predicting PKa Values with Continuous Constant PH Molecular Dynamics.*, 1st ed.; Elsevier Inc., 2009; Vol. 466.
(77) Wallace, J. A.; Wang, Y.; Shi, C.; Pastoor, K. J.; Nguyen, B.-L.; Xia, K.; Shen, J. K. Toward Accurate Prediction of PKa Values for Internal Protein Residues: The Importance of Conformational Relaxation and Desolvation Energy. *Proteins Struct. Funct. Bioinforma.* **2011**, *79*, 3364–3373.
(78) Awoonor-Williams, E.; Rowley, C. N. The Hydration Structure of Methylthiolate from QM/MM Molecular Dynamics. *J. Chem. Phys.* **2018**, *149*, 045103.
(79) Pedron, F. N.; Messias, A.; Zeida, A.; Roitberg, A. E.; Estrin, D. A. Novel Lennard-Jones Parameters for Cysteine and Selenocysteine in the AMBER Force Field. *J. Chem. Inf. Model.* **2023**, *63*, 595–604.
(80) Jing, Z.; Liu, C.; Cheng, S. Y.; Qi, R.; Walker, B. D.; Piquemal, J.-P.; Ren, P. Polarizable Force Fields for Biomolecular Simulations: Recent Advances and Applications. *Annu. Rev. Biophys.* **2019**, *48*, 371–394.
(81) Baker, C. M. Polarizable Force Fields for Molecular Dynamics Simulations of Biomolecules. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2015**, *5*, 241–254.
(82) Nielsen, J. E.; Gunner, M. R.; García-Moreno E., B. The p K a Cooperative: A Collaborative Effort to Advance Structure-Based Calculations of p K a Values and Electrostatic Effects in Proteins. *Proteins Struct. Funct. Bioinforma.* **2011**, *79*, 3249–3259.