PKAD-R: curated, redesigned and expanded database of experimental pKa values in proteins

Ada Y. Chen[1], Shailesh K. Panday[2], Kaoru Ri[3], Emil Alexov[2], Bernard R. Brooks[1], Ana Damjanovic[4,5,1]*

[1]*Laboratory of Computational Biology, National Heart, Lung and Blood Institute, NIH, Bethesda, MD 20892*

[2]*Department of Physics and Astronomy, Clemson University, Clemson, SC 29634*

[3]*Wilmer Eye Institute, Johns Hopkins University School of Medicine, Baltimore, MD 21287*

[4]*Department of Biophysics, Johns Hopkins University, Baltimore, MD 21218*

[5]*Department of Physics and Astronomy, Johns Hopkins University, Baltimore, MD 21218*

** Email: ad@jhu.edu.*

Abstract

Understanding pKa values in ionizable protein residues is critical for understanding fundamental protein properties, such as structure, function and interactions. We present a new version of PKAD, named PKAD-R, which is a curated database of experimentally determined protein pKa values. The database builds upon its predecessors, PKAD and PKAD-2, with significant updates and improvements through: (1) **careful data curation** to remove incorrect entries and consolidate redundant entries by offering alternative structures and pKa values for each unique residue (2) **database redesign**, to enhance its usability by adding additional information such as protein and species names, detailed notes, as well as sequence identity (3) **database expansion** through identification of 214 new (128 non-redundant) pKa entries from the literature. The database currently contains 877 unique pKa entries for wild type structures and 147 for mutant structures, however, we aim to keep updating the database with new entries. The PKAD-R database is available as a stand-alone downloadable file as well as web servers. The database is designed to provide both a set of pKa entries for unique residues suitable for machine learning applications, as well as modularity by providing alternative pKa values and structures, allowing the user to decide which entries to include.

Keywords: Machine learning, ionizable residue, pKa prediction

1. Introduction

Protonation states of ionizable residues in proteins are critical for determining protein's properties such as folding, aggregation, function, and interactions with ligands and other biomolecules.[1–7] Ionizable residues, capable of gaining or losing protons, participate in essential biological processes.[8–12] The pKa values of these residues are therefore fundamental, as they dictate protonation state under specific pH conditions. For example, in enzymes, the pKa of active site residues can predict catalytic behavior, as only certain protonation states support substrate binding or catalysis.[13–15] In addition, ionizable residues contribute to the stability of a protein's structure through electrostatic interactions, which are pH-dependent.[1,16–18] In drug design, understanding pKa values helps optimize ligand binding by predicting the protonation states of binding site residues.[9,11,19] Overall, accurate knowledge of pKa values is essential for understanding and predicting the behavior of proteins.

Building upon the importance of ionizable residues in protein function, pH-dependent biological processes are ubiquitous. Many such processes rely on the protonation-deprotonation equilibria of protein ionizable residues, influencing cellular transport, enzymatic catalysis, signal transduction, structural changes, and more. In cellular transport, for example, proton pumps V-ATPase rely on cycles of protonation and deprotonation of lipid-exposed glutamic acid residues to facilitate proton transport.[20–22] In voltage-gated sodium channels, protonation and deprotonation of ionizable residues in the selectivity filter can modulate the channel's conductance and selectivity.[23–25] Acid-sensing ion channels exhibit proton-dependent gating mechanisms involving the protonation/deprotonation of their "acidic pocket".[26] Similarly, the NhaA Na+/H+ antiporter employs a "pH sensor", a cluster of amino acyl side chains on the cytoplasmic side, to achieve pH activation.[27,28] In enzymatic catalysis, a classic example is the lysozyme: a pair of Asp-Glu residues in its active site acts as a general acid catalyst which is essential for glycoside hydrolysis.[29] In signal transduction, human pH-sensing G protein-coupled receptors (GPCRs), which transduce pH signals into physiological effects, utilize protonation networks to govern their activation, during which the protonation of a buried acidic residue is thought to drive this process.[30] Changes in protonation states can be coupled to structural changes in proteins such as nitrophorin 4 and mutants of staphylococcal nuclease (SNase) with buried ionizable residues.[31–34] These examples underscore the central role of protonation states in regulating pH-dependent processes and highlight the value of pKa databases for modeling and understanding these phenomena with precision.

Among the methods for determining pKa values of protein residues, NMR spectroscopy stands out as the gold standard due to its accuracy and residue-specific resolution with a typical error of 0.1 pKa unit.[35–38] By detecting chemical shift changes in response to pH, NMR provides precise pKa values, accounting for the local microenvironment and electrostatic effects.[37,39–41] The two most common NMR methods are 1H and 13C NMR. 1H NMR is faster and more convenient due to its high sensitivity and natural abundance,[42] though signal overlap and solvent exchange can limit its accuracy in larger proteins.[37] In contrast, 13C NMR offers higher accuracy and resolution for crowded or buried residues but is slower and more resource-intensive due to lower sensitivity and the need for isotopic labeling.[43,44] Despite its strengths, NMR is resource-intensive, requiring isotopic labeling, extensive sample preparation, and long acquisition times, and is limited to small or moderately sized proteins.[35,37,45] Emerging advancements, such as solid-state NMR, are expanding its applicability to fibrillar and less-soluble proteins, further solidifying its role in challenging systems.[46] In contrast, UV spectrophotometry and fluorescence methods offer faster alternatives but lack the resolution and sensitivity of NMR; while UV spectrophotometry is effective for proteins with UV-active residues, it struggles with spectral overlap, and fluorescence relies on intrinsic or labeled fluorophores, limiting its generalizability.[39,47] Methods like reaction kinetics and thermodynamic stability analyses are indirect, providing broader functional insights but sacrificing accuracy at the residue level.[39] Overall, while other methods may excel in speed or specialized contexts, NMR remains the most robust and comprehensive tool for residue-specific pKa determination in proteins.

Computational pKa calculation methods fall into two main categories: macroscopic and microscopic methods. Macroscopic methods, like continuum electrostatics (CE),[48–52] calculate electrostatic potentials using models such as the Poisson-Boltzmann equation or generalized Born model, but often rely on a single dielectric constant, which may not accurately capture protein flexibility, especially for buried residues. Using a Gaussian-based smooth dielectric function improves CE's accuracy and resulted in highly accurate pKa's predictions via DelPhiPKa.[51,53] Empirical methods like PROPKA[54] are fast and reasonably accurate but are primarily parameterized on surface-proximal residues, which may lead to reduced accuracy for buried residues.[8] Microscopic methods, such as by molecular dynamics (MD) and QM/MM simulations,[55,56] provide detailed modeling but are computationally intensive. One type of the MD-based method is using free energy calculations which typically involves calculating the free energy difference between protonated and deprotonated states. Common approaches include thermodynamic integration,[57] free energy perturbation,[58] and the Bennett acceptance ratio (BAR) method.[59] Another type of MD-based method is constant-pH approach which can simulate protonation state changes at constant pH, accommodating conformational changes. Many MD programs implement constant-pH methods, such as the λ-dynamics method and enveloping distribution sampling (EDS) in CHARMM,[60–62] the hybrid explicit/implicit solvent constant-pH method in Amber,[63,64] the hybrid nonequilibrium MD/Monte Carlo (neMD/MC) constant-pH in NAMD,[65] and the λ-dynamics in GROMACS.[66,67] Enhanced sampling techniques, such as pH replica exchange, are often integrated with constant-pH simulations to improve sampling and accelerate convergence.[68,69] Despite their accuracy, constant-pH methods inherit the computational demands of classical MD, requiring long simulation times to achieve convergence.

Recently, machine learning methods,[70–75] including deep learning and tree-based models, have been explored for protein pKa prediction as well, offering new avenues for handling complex data more effectively. While some ML methods utilized pKa data generated computationally,[70–72] all machine learning methods trained on experimental dataset utilized the PKAD database.[73–75] Nonetheless, utilization of this dataset requires additional curation/structure removal. Here, we present a curated dataset, PKAD-R, with all inconsistencies and errors removed, and handling of redundant entries (we list these as alternative entries when multiple pKa measurements or mutant pKa values are available for the same residue), as well as sequence identity calculation for all entries. In such a way, for machine learning application a user can choose a threshold at which they want to keep pKa entries (30% similar or 90% similar). We chose those two values based on previous ML studies.[70,72] Removing redundant entries in machine learning is necessary for removing overlap between training and test data (needed for fair evaluation of the error) and for generalization of the ML model to non-homologous proteins. Our database offers meticulously curated data which allows direct suitability in machine learning applications and serves as a valuable dataset for evaluating computational pKa calculation methods. Finally, our dataset is also a repository for pKa values, and we will keep updating it with new entries.

2. Database Development

Starting point for PKAD-R database was the PKAD-2 database[38] which was extensively redesigned to improve its functionality. We have performed the following changes to the original database:

2.1. *Database curation*

(1) Identification and removal of erroneous entries. These included: duplicates, entries based on calculated rather than experimentally determined pKa values, entries where the pKa was absent in the original literature, and cases where the pKa could not be assigned to a specific residue.

(2) Correction of incorrect structures. During the update, 187 entries were identified in which the mutations in the PDB structures did not match those in the structures used for their pKa measurements. For 46 of them, we located the PDB with correct mutations and updated these entries to use the correct PDB, ensuring consistency. We excluded the remaining 141 entries and did not use the old PDBs as approximate structures, as these entries either have mutations at the target ionizable site, mutations involving other ionizable residues, or mutations near the target site. These entries are available in a separate list, as described below.

(3) pKa values without PDBs provided. Additionally, we compiled a separate list of 213 entries with known pKa values but without suitable PDB structures. These unpaired entries are provided for future use, allowing users to pair them with structures they deem appropriate, such as those generated by AlphaFold. Any updates to these entries will be incorporated into subsequent database versions.

(4) Additional small corrections. We have also performed several other cleaning tasks for this database, including correcting any residue number offsets, fixing incorrect or missing chain IDs, and updating PDB entries whenever a newer version is available (e.g., 1LZ3 has been superseded by 135L, and 4GDH by 4QYT). Additionally, for one protein, we updated the PDB from theoretical model (1ILB) to experimental structure (1FOV).

(5) Selection of representative entries. In some cases, multiple pKa values or PDB structures were reported for the same ionizable residue. To address this issue, we identified unique ionizable residues for each protein and species. For each unique residue, we selected or created a single representative entry, referred to as the “Main” entry. This entry contains the most appropriate pKa value and PDB structure, based on information from the original literature. For residues with multiple pKa values reported, the most recommended entry is labeled “Main” in the “pKa classification” column, while other alternative pKa values are provided in separate rows labeled as “Alt. pKa” in that column with detailed explanations provided in the “Notes” column. When an "Alt. pKa" entry corresponds to a measurement in a mutated version of the protein or a different state of the protein, it is labeled

as "Alt. pKa (mutant)" or "Alt. pKa (state)", respectively. “Main” entries for a unique protein are associated with a single primary PDB structure, with additional suitable PDBs provided in a separate "Alternative PDBs" column. "Alt. pKa (mutant)" and "Alt. pKa (state)" entries are paired with different PDBs that correspond to the specific mutant or protein state, respectively. Below we describe the general strategy how we selected the “Main” entry for each unique ionizable residue.

The pKa and PDB selection for the “Main” entry were manually performed on a case-by-case basis, examining individual residues rather than the entire protein to ensure a thorough clean-up while preserving as many entries as possible without unnecessary exclusions. First, this selection of process prioritized pKa values measured using NMR techniques, which are generally more reliable than other methods, with NMR being the first choice. Among NMR techniques, ^{13}C NMR was preferred over ^{1}H NMR. Values obtained under moderate experimental conditions, typically around 298 K and 0.1 M salt concentration, were prioritized. In some cases, pKa values were averaged to produce a single and reasonable representative value, such as for Bovine Ribonuclease A, where measurements at 30 mM and 1 M salt concentrations were averaged. When searching for and selecting PDB structures, we ensure that mutations in the PDB align with those used in the pKa measurements. If multiple PDB structures are available, preference is given to structures determined by X-ray crystallography over those obtained by solution NMR, as it generally provides higher atomic-level resolution and a clearer, more consistent baseline for pKa calculation. The structure with the highest resolution is then selected. Selected PDBs were carefully cross-checked to align as closely as possible with the experimental conditions for pKa measurements, such as protein species and the presence of cofactors used in the measurements. For example, in the case of calbindin, the pKa values for the Ca^{2+}-loaded form were paired with the PDB structure for the same form (4ICB) to ensure that the cofactors (Ca^{2+} ions) are present in both the pKa measurements and the PDB structure; other pKa values for the apo-form were listed as alternative entries, labeled as "Alt. pKa (state)". Users should refer to the "Notes" column for the most accurate selection criteria for each residue, as these may be updated over time. While this paper outlines our initial methodology, we reserve the right to refine the criteria to improve the database's reliability.

The pKa values for mutants are retained either as "Main" or alternative entries, depending on whether the mutation occurs directly at the site where the pKa was measured. If the mutated residue itself is the ionizable site being measured, it is treated as a distinct residue (compared to the wildtype counterpart). In such cases, the pKa of the mutated residue is designated as a "Main" entry rather than an alternative entry for the corresponding wildtype residue. For example, in SNase, pKa measurements for residue 23 in the mutants V23D, V23E, and V23K are all designated as "Main" entries, each paired with mutant-specific PDB structures. These measurements involve different residue types, using them together does not introduce redundancy, so they are kept as separate "Main" entries. However, if a pKa value is measured for a mutated protein but the mutation does not involve the target ionizable residue, the pKa is listed as an alternative entry, labeled “Alt. pKa (mutant).” We classify them this way because, while the mutation may alter the environment of the target residue, it may remain highly similar to the wildtype and could introduce redundancy if all were treated as "Main". For example, again in SNase, three pKa measurements are available for Glu-43: one in the wildtype protein and two in the mutants L38K and L38E. Since the mutations do not occur at the target residue, the wildtype entry is labeled as "Main", while the mutant entries are classified as "Alt. pKa (mutant)." Despite being labeled as alternatives, each mutant entry is still paired with a PDB structure with the correct mutation, enabling users to study the impact of mutations on the target residue within an appropriate structural context.

Below, we present an example of how we handle proteins with two states, using human hemoglobin, where pKa values are available for both oxy- and deoxyhemoglobin. Given the high similarity between the two states, we designate deoxyhemoglobin’s pKa entries as "Main" and oxyhemoglobin’s as “Alt. pKa (state)”, indicating they are alternatives measured in different protein states. This choice of "Main" state is arbitrary, and users can select the other state if preferred. For each state, we selected the best-matched PDB: PDB 4HHB for deoxyhemoglobin, which includes the heme cofactor but no oxygen molecules, and PDB 1HHO for oxyhemoglobin, which contains both heme and oxygen molecules. This careful pKa classification and PDB selection ensures minimal redundancy while maintaining data richness and accuracy.

In previous versions of PKAD, a single pKa value was sometimes paired with multiple similar PDBs, resulting in the same pKa value appearing in multiple entries. In this version, each pKa value appears only once, paired with a single primary PDB. However, if pKa values reported by different papers happen to be identical, all are retained as separate entries, but only one entry is designated as "Main," with the others labeled as alternatives, enabling users to access the different experimental details associated with each. Prior to the “Main” selection clean-up, the database contained 1,667 entries. After keeping only one PDB for each unique pKa value, 1,024 entries remain, which are the entries we present in the PKAD-R database. Users may opt to use only the “Main” entries, offering 778 non-redundant entries for unique residues. These entries are particularly well-suited for tasks like machine learning training and testing, where a high degree of data non-redundancy is essential.

2.2. *Database redesign*

To provide friendliness of usage, we have redesigned the database, by adding several additional columns. Table 1 shows explanation of all columns. Below we introduce new columns.

- **“Protein Name” and “Species”** columns provide protein names and species for each entry, enabling quick and efficient searches. Users can quickly determine whether a pKa value from the literature is already included in the database by searching for the protein name and species.
- **“pKa Classification”** column categorizes each entry as “Main,” “Alt. pKa,” “Alt. pKa (mutant),” or “Alt. pKa (state).” Entries labeled as “Main” represent the most recommended pKa value for a given residue, paired with the most appropriate PDB structure. The label “Alt. pKa” indicates another pKa measurement for the same residue in the same protein as the “Main” entry. “Alt. pKa (mutant)” refers to a pKa measured for the same residue in a mutated version of the protein. “Alt. pKa (state)” denotes a pKa measured in a different state of the protein, such as deoxyhemoglobin versus oxyhemoglobin. For more details on selection of “Alt. pKa (mutant)” and “Alt. pKa (state)” see the description in the “Database Curation” section.
- **“Alternative PDBs”** column lists additional PDB structures that are also available for use and similar to the primary PDB structure listed in the "PDB" column.
- **"Sequence Identity > 30%" and "Sequence Identity > 90%"** columns list chains in the format PDB-ID.Chain-ID (e.g., "1EX3.A") included in this database that have sequence identities greater than 30% and 90%, respectively, compared to the sequence of the current entry's chain. We include a tag, 'mutation_on_site,' for chains with a mutation directly on the measured ionizable site, such as some of the SNase variants, serving as a warning to alert users that these entries should not be discarded based solely on sequence identity. Sequence identity is calculated using the PairwiseAligner class from the Bio.Align module within the Biopython library.[76]
- **“ResID in PDB” and “ResID in pKa paper”** columns: "ResID in PDB" lists the residue ID as it appears in the corresponding PDB structure, while "ResID in pKa paper" indicates the residue ID referenced in the original pKa publication when it differs from the PDB.
- **“Notes”** column provides, for each residue, details about the selection of the most appropriate pKa value and PDB structure for the "Main" entry, as well as any additional relevant information such as the specific state of the protein and important cofactors.
- “**Warning**” column labels entries under specific conditions: 1) when the pKa is a range or an approximation (labeled as “pKa: range or ~”); 2) when the residue is the C-terminus or N-terminus (labeled as “C/N-term”); 3) when the residue does not exist in the PDB structure but is present in the protein (labeled as “ResID NOT exist”), likely due to its high flexibility and disorder, which makes accurate structural definition difficult; 4) when a mutated structure is used for a wildtype protein, but the mutation is distant from the targeted residue, allowing the structure to be approximately treated as wildtype (labeled as 'approx. WT'). This column helps users quickly filter out entries that may not meet their needs, such as those unsuitable for direct use in machine learning studies.

Table 1. Columns of the database.

Column Name	**Meaning**

Protein Name	Name of the protein in which the pKa was measured.
Species	Species of the protein in which the pKa was measured.
pKa Classification	This column indicates the classification of pKa for each entry. Options are “Main”, “Alt. pKa”, “Alt. pKa (mutant)” and “Alt. pKa (state)”. See Database Development section for details.
PDB	PDB ID of the recommended primary protein structure.
Chain	Chain ID of the target residue.
ResID in PDB	Residue ID of the target residue in the PDB file. In cases where this ID is different from the ID in the pKa paper, the column “ResID in pKa paper” provides the residue ID referenced there.
ResID in pKa paper	Residue ID referenced in the original publication of the pKa measurement, when this ID is different from the “ResID in PDB”.
ResName	Residue name of the target residue.
Expt. pKa	Experimental pKa value.
Mut. Pos.	Mutation position, if applicable. If this column is empty, the entry corresponds to a wildtype protein.
Alternative PDBs	PDB IDs for alternative protein structures.
Notes	Notes of criteria used for determining the “Main” entry and any other information the authors deem relevant.
Sequence Identity > 30%	PDB IDs with sequence identities greater than 30%, compared to the current entry's PDB
Sequence Identity > 90%	PDB IDs with sequence identities greater than 90%, compared to the current entry's PDB
Warning	This column indicates specific conditions that users should be aware of when using the data. See Database Development section for details.
Expt. Uncertainty	Experimental uncertainty.
Expt. Temp.	Experimental temperature.
Expt. pH	Experimental pH value.
Expt. Salt Concentration	Experimental salt concentration.
Expt. Method	Experimental method for measuring the pKa values.
Reference	The link to the original publication for the pKa measurement.

2.3. *New entries*

We identified 214 new pKa entries from the literature. Among these, 88 entries were directly associated with PDB structures in the original papers, while 126 were not. For the latter group, suitable PDB structures were identified for 77 entries, resulting in 165 new entries with paired pKa-PDB information. After integrating these entries with PKAD-2 and completing the clean-up/curation process (explained above), 128 new entries were retained, while redundant ones were removed. Notably, among these 128 entries, 42 correspond to cysteine (Cys), significantly increasing the number of non-redundant Cys entries from 23 in PKAD-2 to 65, increased by 183%. This substantial increase in Cys pKa data largely enhances the dataset’s utility for Cys pKa calculations.

Additionally, we include pKa values for nine SNase variants with Asp mutations that were previously unpublished,[77] from personal communication with Dr. Bertrand García-Moreno. The pKa values of these mutated residues, introduced into the interior of SNase, are challenging to predict due to their large deviation from standard pKa values in water, making them a valuable addition to the PKAD-R database.

2.4. *Database repository*

In this work, we present the new curated and redesigned version of the PKAD database, named PKAD-R. Here, "R" stands for "repository", as we provide instructions for user contributions and submission of new pKa values not present in PKAD-R. The new pKa values that are submitted will be curated and included in future updates. All data and statistics discussed in this paper pertain to the first version of PKAD-R.

The PKAD-R web servers facilitate data filtering and sorting based on various fields, e.g.
"Protein Name", "Species", "PDB", "Chain", "ResID in PDB", "ResName", "Expt. pKa", "Mut. Pos.", "pKa Classification", "Alternative PDBs", "Sequence Identity > 30%", "Sequence Identity > 90%", "ResID in pKa paper", "Warning", "Expt. Uncertainty", "Expt. Temp.", "Expt. pH", "Expt. Salt Concentration", "Expt. Method" and "Reference" in the dataset. The servers are available via http://compbio.clemson.edu/PKAD-R/ and https://sites.krieger.jhu.edu/damjanovic-lab/pkad-r/. The former web server is developed using php v8.3.6 for back-end and the JavaScript for the front-end. In addition, for each pKa entry a visualization of the residue over the protein structure is provided to aid in analyzing the spatial placement of the target residue. Reference paper is also cross-referenced to allow the user to directly reach the appropriate publication that originally reported it.

3. Results and Discussions

Table 2 summarizes the key statistics of the pKa database PKAD-R, which comprises 1,024 non-redundant entries. Among these, 778 are "Main" entries (explained in the Database Development section) corresponding to unique residues, while the remaining 246 entries represent alternative pKa values. These alternative entries provide additional pKa values for the same residue as their corresponding "Main" entry, which may include pKa measurements from different studies, pKa values obtained in mutated versions of the protein, or pKa measurements taken in different states of the protein. Among "Main" entries, 578 entries (74.3%) have no similar chains in this database with sequence identity greater than 90%, and 391 entries (50.3%) have no similar chains with sequence identity greater than 30%.

Glutamic acid (Glu) and aspartic acid (Asp) are the most represented residues, with 290 and 259 entries, respectively. Their average pKa values are 4.38 and 3.93, spanning broad ranges (Asp: 0.50 to 9.90, Glu: 2.10 to 10.50), reflecting the diverse environments in which these residues are found. Histidine (His) and lysine (Lys) are moderately represented in the database, with 211 and 122 entries, respectively. His has an average pKa of 6.55, spanning a range of 2.50–9.19, while Lys has an average pKa of 10.10, with values ranging from 5.30 to 12.12. Tyrosine (Tyr) and cysteine (Cys) are less common in the dataset, with 41 and 65 entries, respectively. The average pKa value of Cys deviates notably from its pKa value of model compound in water. Specifically, the average pKa of Cys is 6.24, significantly lower than its model compound pKa of 8.55. This deviation may suggest that Cys residues in this dataset are often located in less typical protein environments. In fact, Cys residues are more frequently located in protein interiors, as indicated by their average %SASA value of 7%. This value is significantly lower than those of other residue types, i.e., Asp (37%), Glu (38%), His (36%), Lys (48%), and Tyr (20%). In contrast, the average pKa values for Asp (3.93), Glu (4.38), His (6.55), Lys (10.10), and Tyr (10.01) closely aligned with their model compound pKa values (Asp: 3.67, Glu: 4.25, His: 6.54, Lys: 10.40, Tyr: 9.84), possibly reflecting a more comprehensive sampling that captures a diverse range of protein environments.

The database spans a wide pKa range (0.50–12.12) across all residue types, with residues such as Asp and Glu exhibiting particularly broad ranges. This provides a robust foundation for analyzing pKa variability in diverse biological and chemical contexts.

Table 2. Database key statistics.

Residue Name	No. of Entries	No. of "Main"	No. of "Main" Wildtype	No. of "Main" Mutant	Avg. pKa*	Lowest pKa*	Highest pKa*

Asp	259	220	198	22	3.93	0.50	9.90
Glu	290	223	199	24	4.38	2.10	10.50
His	211	124	120	4	6.55	2.50	9.19
Lys	122	94	79	15	10.10	5.30	12.12
Tyr	41	39	38	1	10.01	6.28	11.70
Cys	65	44	40	4	6.24	2.88	11.10
N-term	17	15	15	0	7.67	6.91	9.14
C-term	19	19	19	0	3.16	2.40	4.03
All	1024	778	708	70	5.68	0.50	12.12

* Considering all entries including both “Main” and alternative pKa values.

Table 3 presents a detailed comparison of pKa statistics between wildtype and mutant residues in the whole dataset, including both “Main” and alternative entries. Of the 1,024 total entries, 86% (877 entries) are wildtype, while 14% (147 entries) are mutant. Cys has the largest proportion of mutant entries, with 31% of its total entries being mutants, followed by Lys and Glu, which both have 18% mutant entries.

Both Asp and Glu exhibit a significant increase in pKa in the mutant proteins: Asp’s average pKa rises by 1.94, and Glu’s increases by 0.96. In contrast, Lys shows a substantial decrease in average pKa in mutant proteins, with a reduction of 2.43. It is likely because a large number of entries come from mutants of staphylococcal nuclease which are often buried in protein cores. This particular environment tends to shift pKa values towards the neutral form, which in the case of Asp and Glu favors an increase in the pKa shift, and in the case of Lys favors a decrease in the pKa shift. However, His and Cys display small differences in average pKa between wildtype and mutant proteins, with His changing from 6.60 to 6.09 and Cys shifting from 6.02 to 6.66. None of the staphylococcal nuclease variants contain substitutions of His and Cys. Tyr also shows a large decrease of 3.92 in average pKa; however, the small number of Tyr entries limits the statistical significance of this observation.

Asp and Glu residues dominate both the wildtype and mutant datasets, together accounting for 54% of the entries. In the wildtype dataset, Asp and Glu make up 53% (465 entries out of 877), while in the mutant dataset, they comprise 58% (84 entries out of 144).

Regarding pKa variability, Asp shows the largest pKa range in both wildtype and mutant datasets, with a range of 0.50 to 9.90 in the wildtype and 2.10 to 9.70 in the mutant. Glu and Cys have the second-largest ranges: Glu spans 2.10 to 10.50 in the wildtype and 2.90 to 9.40 in the mutant, while Cys ranges from 2.88 to 11.10 in the wildtype and 3.75 to 10.90 in the mutant.

Table 3. Statistics of all pKa entries by wildtype and mutant residues.

Residue Name	**Wildtype**				**Mutant**			
	No.	Avg. pKa	Lowest pKa	Highest pKa	No.	Avg. pKa	Lowest pKa	Highest pKa
Asp	227	3.67	0.50	9.90	32	5.61	2.10	9.70
Glu	238	4.21	2.10	10.50	52	5.17	2.90	9.40

His	193	6.60	2.50	9.19	18	6.09	4.03	8.90
Lys	100	10.56	8.40	12.12	22	8.13	5.30	10.90
Tyr	38	10.54	9.14	11.70	3	6.62	6.28	6.86
Cys	45	6.02	2.88	11.10	20	6.66	3.75	10.90
N-term	17	7.67	6.91	9.14	0	\	\	\
C-term	19	3.16	2.40	4.03	0	\	\	\
All	877	5.61	0.50	12.12	147	6.05	2.10	10.90

Figure 1 illustrates the distribution of pKa values for six amino acid residues (Asp, His, Cys, Glu, Lys, and Tyr) in the database, comparing the overall dataset (labeled as "All" in the figure legends) with residues that were experimentally mutated (labeled as "Mutant"). The pKa distributions for most residue types are approximately Gaussian, with notable rightward tails for Asp and Glu and a leftward tail for Lys. For Asp and Glu, pKa values predominantly cluster around 4, consistent with their expected acidic behavior, though mutant entries reveal a secondary cluster near 8. His displays a distribution centered between 6 and 7, reflecting its role in enzymatic active sites; mutant entries follow this trend. Lys exhibits a peak around 11, highlighting its basic nature, with the mutant distribution shifted leftward, centering around 8. Cys shows a broader, more dispersed distribution centering near 6, with mutant entries following a similar but sparser pattern. Tyr has a peak near 10 in the wildtype dataset, while the mutant data exhibits a distinct shift with only three entries around 6.5.

While His and Cys mutant distributions closely match their respective overall distributions, Asp, Glu, and Lys display distinct differences between the overall and mutant distributions. For Asp and Glu, the mutant distributions exhibit a clear bimodal pattern: one group aligns with the mean of the overall distribution (~4), while the other group is shifted significantly to higher pKa values (~8). This bimodality suggests distinct subpopulations of mutants with different ionization properties. Lys, on the other hand, shows a continuous mutant distribution that is shifted leftward, centering around 8. The observed correlation between mutations and pKa shifts in Asp, Glu, and Lys may stem from experimental design choices. Experimentalists sometimes intentionally select internal sites for mutation, as seen in studies involving Staphylococcal nuclease.[41,78–84] These sites are typically buried within the protein, and buried sites often exhibit shifts in the pKa values. This suggests that the large pKa shifts observed in these mutants may not necessarily reflect a general property of all mutations but rather a consequence of targeted experimental design. If mutation sites were selected randomly across a protein, the mutant distribution might not exhibit the same degree of pKa shifts.

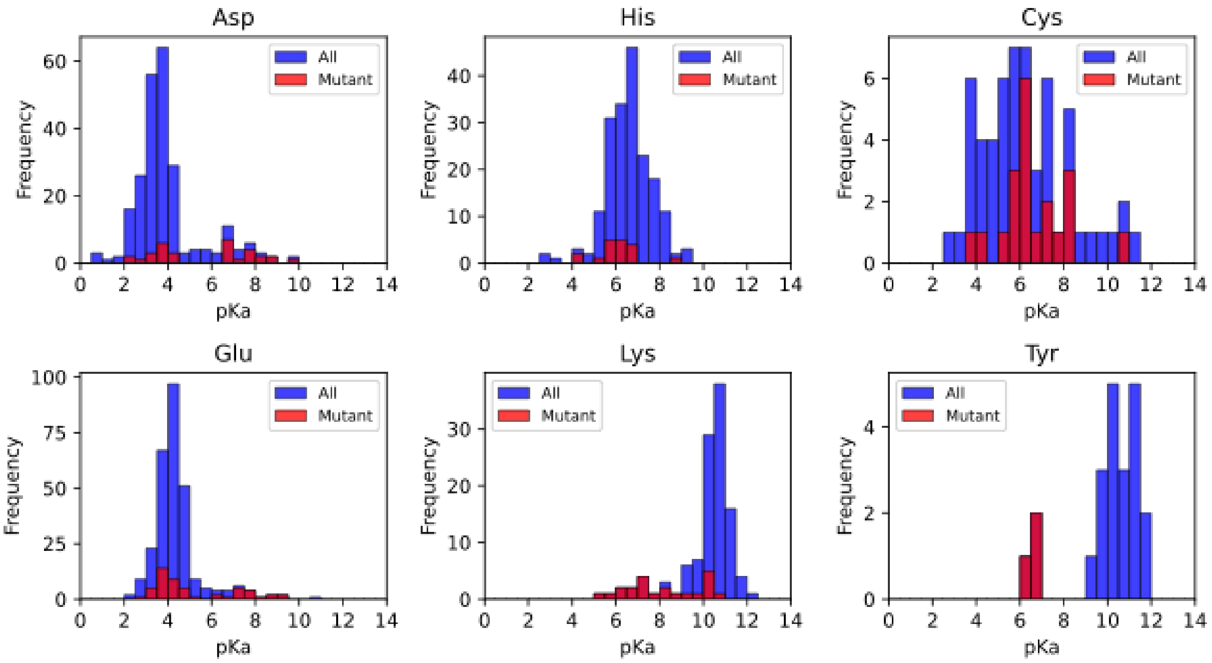

Figure 1. Distribution of pKa values for six ionizable residue types (Asp, His, Cys, Glu, Lys, and Tyr) in the database. Blue bars represent the overall dataset ("All"), while red bars represent the mutant subset ("Mutant").

Figure 2 presents the side chain percentage solvent-accessible surface area (%SASA), which is the SASA normalized by the side chain SASA of a single amino acid in water, in relation to the pKa values for each residue type. Panel A displays individual scatter plots for Asp, His, Cys, Glu, Lys, and Tyr, showing the relationship between pKa values and %SASA for both wildtype and mutant residues. Five residue types (Asp, His, Glu, Lys, and Tyr) exhibit a wide range of %SASA values, indicating varying degrees of solvent accessibility, from fully buried to highly exposed residues. In contrast, Cys residues show very low solvent accessibility, with most data points having %SASA values below 0.2.

For Asp, Glu, and Lys, the mutant data points can be divided into two distinct groups: one group lies within the major distribution of the entire dataset, while the other exhibits a highly shifted pKa and a very low %SASA. This shift in pKa is strongly correlated with low %SASA values, consistent with the idea that many of these mutants are engineered to reside deep within the protein interior. For Lys, all data points with highly shifted pKa values are mutants, while for Asp and Glu, a few wildtype residues also show high pKa shifts and low %SASA values. This suggests that such pKa shifts can occur naturally in some residues, rather than solely as a result of artificial mutations. His and Cys residues, in contrast, exhibit similar patterns for both wildtype and mutant residues on the 2D plane of %SASA and pKa. For Tyr, wildtype and mutant data show two distinct clusters, with three mutant entries located at low %SASA and pKa values around 6.5, while wildtype data points span a range of %SASA but do not exhibit any correlation between %SASA and pKa.

Panel B further investigates the relationship between %SASA and pKa shifts by plotting the ΔpKa (the difference between the experimental pKa and the model compound pKa) against %SASA for all residues combined. The scatter plot reveals a broad distribution of ΔpKa values across the entire %SASA spectrum. Residues with high %SASA (above 0.5) tend to show smaller ΔpKa values, while those with lower %SASA exhibit a wider range of ΔpKa shifts. Importantly, the plot demonstrates that significant pKa shifts are always associated with low solvent accessibility, although residues with low solvent accessibility can also exhibit low ΔpKa values, as seen in the data points in the lower center region. This trend supports the hypothesis that solvent-exposed residues experience less perturbation in their pKa values compared to buried residues. Overall, these findings suggest that the burial of side chains within the protein core leads to more pronounced pKa shifts.

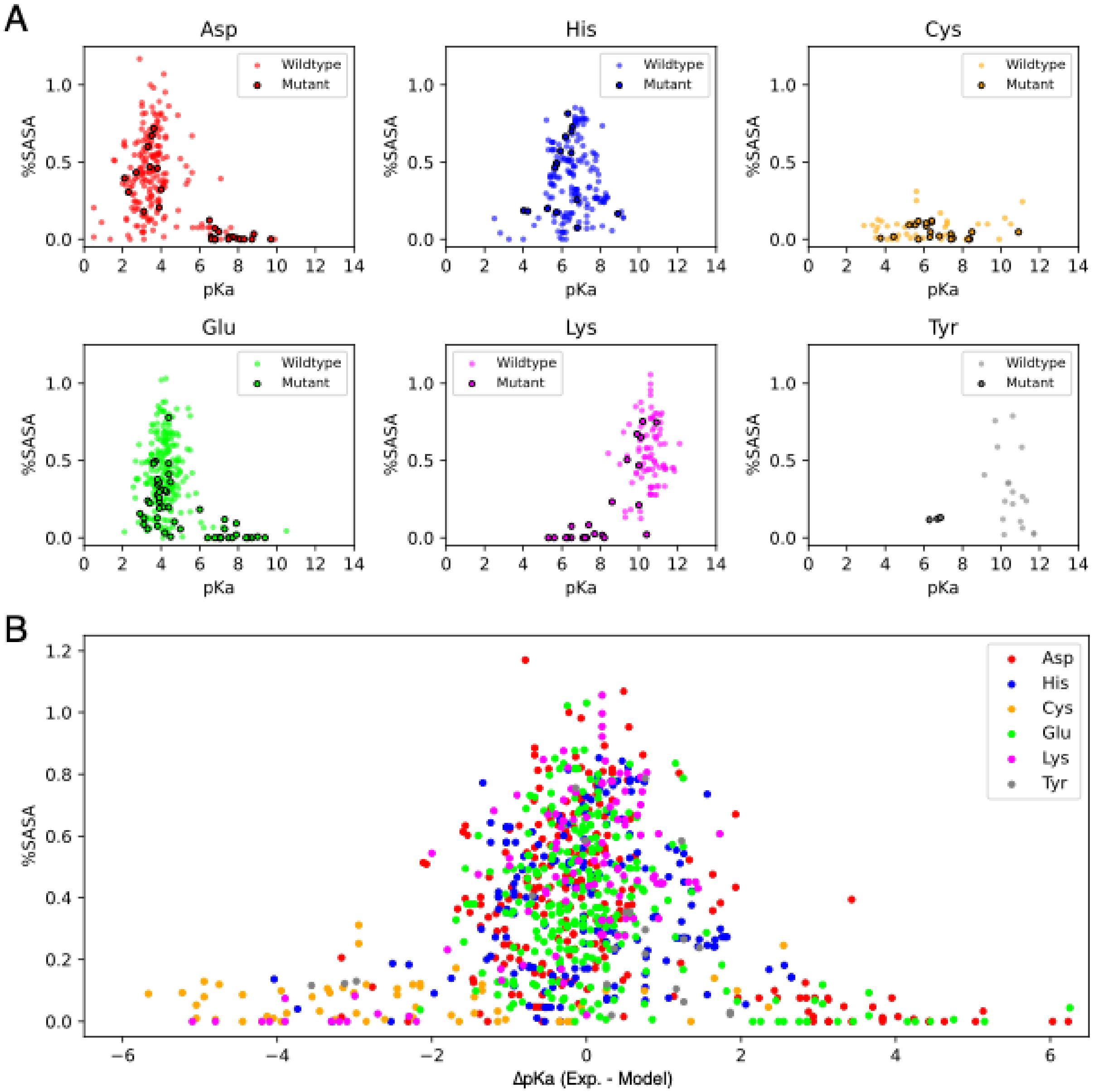

Figure 2. Side chain %SASA and its relationship with pKa. (A) %SASA against pKa values for six ionizable residue types (Asp, His, Cys, Glu, Lys, and Tyr); (B) %SASA against ΔpKa.

4. Conclusion

This study presents PKAD-R, an updated version of PKAD-2 database, with improvements in non-redundancy, usability, and data quality. A total of 128 new pKa entries paired with PDB structures were also added, increasing particularly the dataset for cysteine residues by 183%. Rigorous manual curation addressed inconsistencies between experimental pKa data and corresponding PDB structures. Each residue was carefully reviewed, with a case-by-case examination of the original literature to evaluate and select the most representative pKa value and PDB structure as the “Main” entry, prioritizing NMR-measured pKa values and well-aligned PDB structures. Where available, alternative pKa values and PDB structures are provided, along with comprehensive notes that explain the selection criteria and offer guidance for users in choosing the most appropriate entries. This redundancy reduction process takes into account factors such as measurement techniques, experimental conditions, and the structural alignment of PDBs with pKa experiments. By systematically addressing redundancy and thoroughly documenting our decisions, our database provides users with transparent, curated, and highly informative data. The resulting database offers 1,024 unique pKa entries for 778 unique residues, making it a valuable dataset for machine learning and other applications. The database further offers filtering of proteins by sequence identity, offering users the ability to further remove similar proteins for machine learning purposes. The PKAD-R database is downloadable as a stand-alone file, and it is accessible as web servers via http://compbio.clemson.edu/PKAD-R/ and https://sites.krieger.jhu.edu/damjanovic-lab/pkad-r/, where users can perform selections and create subsets of entries.

Conflict of Interest

The authors declare that they have no conflict of interest.

Funding Information

The work was supported by the Intramural Research Program of the NIH, NHLBI (ZIA HL001051-27 to B.R.B.). The work of S.P. and E.A. was supported by NIH grant R35GM151964. A.D. acknowledges NHLBI grant 75N92020P00042.

Acknowledgements

We sincerely thank Dr. Bertrand García-Moreno for sharing SNase data. Calculations were performed on the LoBoS supercomputing cluster in NIH, NHLBI.

Author Contributions

A.Y.C. curated the data, conducted the analyses, prepared the figures and tables, and drafted and revised the manuscript. A.D. did the initial round of data curation, reviewed database entries and revised the manuscript. S.P. developed the database web server. S.P., E.A., and B.R.B. reviewed and edited the manuscript. K.R. contributed to data collection. All authors contributed ideas and participated in discussions of the work.

Data Availability

Data in this work is accessible via http://compbio.clemson.edu/PKAD-R/ and https://sites.krieger.jhu.edu/damjanovic-lab/pkad-r/.

Statement of Usage of Artificial Intelligence

We utilized Artificial Intelligence (ChatGPT) to assist in polishing the English writing of this paper. All content is entirely original and developed by the authors.

References

(1) Zhou, H.-X.; Pang, X. Electrostatic Interactions in Protein Structure, Folding, Binding, and Condensation. *Chem. Rev.* 2018, *118* (4), 1691–1741. https://doi.org/10.1021/acs.chemrev.7b00305.

(2) Matthew, J. B.; Gurd, F. R. N.; Garcia-Moreno, B. E.; Flanagan, M. A.; March, K. L.; Shire, S. J. PH-Dependent Processes in Protein. *Crit. Rev. Biochem.* 1985, *18* (2), 91–197. https://doi.org/10.3109/10409238509085133.

(3) Silverstein, T. P. The Proton in Biochemistry: Impacts on Bioenergetics, Biophysical Chemistry, and Bioorganic Chemistry. *Front. Mol. Biosci.* 2021, *8*. https://doi.org/10.3389/fmolb.2021.764099.

(4) Schönichen, A.; Webb, B. A.; Jacobson, M. P.; Barber, D. L. Considering Protonation as a Posttranslational Modification Regulating Protein Structure and Function. *Annu. Rev. Biophys.* 2013, *42* (Volume 42, 2013), 289–314. https://doi.org/https://doi.org/10.1146/annurev-biophys-050511-102349.

(5) Raab, S. A.; El-Baba, T. J.; Laganowsky, A.; Russell, D. H.; Valentine, S. J.; Clemmer, D. E. Protons Are Fast and Smart; Proteins Are Slow and Dumb: On the Relationship of Electrospray Ionization Charge States and Conformations. *J. Am. Soc. Mass Spectrom.* 2021, *32* (7), 1553–1561. https://doi.org/10.1021/jasms.1c00100.

(6) Onufriev, A. V; Alexov, E. Protonation and PK Changes in Protein–Ligand Binding. *Q. Rev. Biophys.* 2013, *46* (2), 181–209. https://doi.org/DOI: 10.1017/S0033583513000024.

(7) Longo, G. S.; Pérez-Chávez, N. A.; Szleifer, I. How Protonation Modulates the Interaction between Proteins and PH-Responsive Hydrogel Films. *Curr. Opin. Colloid Interface Sci.* 2019, *41*, 27–39. https://doi.org/10.1016/J.COCIS.2018.11.009.

(8) Alexov, E.; Mehler, E. L.; Baker, N.; M. Baptista, A.; Huang, Y.; Milletti, F.; Erik Nielsen, J.; 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 Struct. Funct. Bioinforma.* 2011, *79* (12), 3260–3275. https://doi.org/https://doi.org/10.1002/prot.23189.

(9) Sussman, F.; Villaverde, M. C.; Dominguez, J. L.; Danielson, U. H. On the Active Site Protonation State in Aspartic Proteases: Implications for Drug Design. *Current Pharmaceutical Design*. 2013, pp 4257–4275. https://doi.org/http://dx.doi.org/10.2174/1381612811319230009.

(10) Petukh, M.; Stefl, S.; Alexov, E. The Role of Protonation States in Ligand-Receptor Recognition and Binding. *Current Pharmaceutical Design*. 2013, pp 4182–4190. https://doi.org/http://dx.doi.org/10.2174/1381612811319230004.

(11) Poian, A. T. Da; Carneiro, F. A.; Stauffer, F. Viral Inactivation Based on Inhibition of Membrane Fusion: Understanding the Role of Histidine Protonation to Develop New Viral Vaccines. *Protein & Peptide Letters*. 2009, pp 779–785. https://doi.org/http://dx.doi.org/10.2174/092986609788681823.

(12) CHESLER, M. Regulation and Modulation of PH in the Brain. *Physiol. Rev.* 2003, *83* (4), 1183–1221. https://doi.org/10.1152/physrev.00010.2003.

(13) Warshel, A.; Sharma, P. K.; Kato, M.; Xiang, Y.; Liu, H.; Olsson, M. H. M. Electrostatic Basis for Enzyme Catalysis. *Chem. Rev.* 2006, *106* (8), 3210–3235. https://doi.org/10.1021/cr0503106.

(14) Harris, T. K.; Turner, G. J. Structural Basis of Perturbed PKa Values of Catalytic Groups in Enzyme Active Sites. *IUBMB Life* 2002, *53* (2), 85–98. https://doi.org/https://doi.org/10.1080/15216540211468.

(15) Chang, C. A.; Huang, Y. M.; Mueller, L. J.; You, W. Investigation of Structural Dynamics of Enzymes and Protonation States of Substrates Using Computational Tools. *Catalysts*. 2016. https://doi.org/10.3390/catal6060082.

(16) Yang, A. S.; Honig, B. On the PH Dependence of Protein Stability. *J. Mol. Biol.* 1993, *231* (2), 459–474. https://doi.org/10.1006/JMBI.1993.1294.

(17) Pace, C. N.; Grimsley, G. R.; Scholtz, J. M. Protein Ionizable Groups: P<em>K</Em> Values and Their Contribution to Protein Stability and Solubility *. *J. Biol. Chem.* 2009, *284* (20), 13285–13289. https://doi.org/10.1074/jbc.R800080200.

(18) Talley, K.; Alexov, E. On the PH-Optimum of Activity and Stability of Proteins. *Proteins Struct. Funct. Bioinforma.* 2010, *78* (12), 2699–2706. https://doi.org/https://doi.org/10.1002/prot.22786.

(19) Fischer, A.; Smieško, M.; Sellner, M.; Lill, M. A. Decision Making in Structure-Based Drug Discovery: Visual Inspection of Docking Results. *J. Med. Chem.* 2021, *64* (5), 2489–2500. https://doi.org/10.1021/acs.jmedchem.0c02227.

(20) Bello, S. A.; Yu, S.; Wang, C.; Adam, J. M.; Li, J. Review: Deep Learning on 3D Point Clouds. *Remote Sens.* 2020, *12* (11). https://doi.org/10.3390/rs12111729.

(21) Abbas, Y. M.; Wu, D.; Bueler, S. A.; Robinson, C. V; Rubinstein, J. L. Structure of V-ATPase from the Mammalian Brain. *Science (80-.).* 2020, *367* (6483), 1240–1246. https://doi.org/10.1126/science.aaz2924.

(22) Roh, S.-H.; Stam, N. J.; Hryc, C. F.; Couoh-Cardel, S.; Pintilie, G.; Chiu, W.; Wilkens, S. The 3.5-&#xc5; CryoEM Structure of Nanodisc-Reconstituted Yeast Vacuolar ATPase V_o Proton Channel. *Mol. Cell* 2018, *69* (6), 993-1004.e3. https://doi.org/10.1016/j.molcel.2018.02.006.

(23) Damjanovic, A.; Chen, A. Y.; Rosenberg, R. L.; Roe, D. R.; Wu, X.; Brooks, B. R. Protonation State of the Selectivity Filter of Bacterial Voltage-Gated Sodium Channels Is Modulated by Ions. *Proteins Struct. Funct. Bioinforma.* 2020, *88* (3), 527–539. https://doi.org/https://doi.org/10.1002/prot.25831.

(24) Chen, A. Y.; Brooks, B. R.; Damjanovic, A. Determinants of Conductance of a Bacterial Voltage-Gated Sodium Channel. *Biophys. J.* 2021, *120* (15), 3050–3069. https://doi.org/https://doi.org/10.1016/j.bpj.2021.06.013.

(25) Chen, A. Y.; Brooks, B. R.; Damjanovic, A. Ion Channel Selectivity through Ion-Modulated Changes of Selectivity Filter PKa Values. *Proc. Natl. Acad. Sci.* 2023, *120* (26), e2220343120. https://doi.org/10.1073/pnas.2220343120.

(26) Yoder, N.; Yoshioka, C.; Gouaux, E. Gating Mechanisms of Acid-Sensing Ion Channels. *Nature* 2018, *555* (7696), 397–401. https://doi.org/10.1038/nature25782.

(27) Padan, E. The Enlightening Encounter between Structure and Function in the NhaA Na$^+$&#x2013;H$^+$ Antiporter. *Trends Biochem. Sci.* 2008, *33* (9), 435–443. https://doi.org/10.1016/j.tibs.2008.06.007.

(28) Kozachkov, L.; Padan, E. Conformational Changes in NhaA Na+/H+ Antiporter. *Mol. Membr. Biol.* 2013, *30* (1), 90–100. https://doi.org/10.3109/09687688.2012.693209.

(29) Phillips, D. C. THE HEN EGG-WHITE LYSOZYME MOLECULE. *Proc. Natl. Acad. Sci.* 1967, *57* (3), 483–495. https://doi.org/10.1073/pnas.57.3.483.

(30) Howard, M. K.; Hoppe, N.; Huang, X.-P.; Macdonald, C. B.; Mehrota, E.; Rockefeller Grimes, P.; Zahm, A.; Trinidad, D. D.; English, J.; Coyote-Maestas, W.; Manglik, A. Molecular Basis of Proton-Sensing by G Protein-Coupled Receptors. *bioRxiv* 2024, 2024.04.17.590000. https://doi.org/10.1101/2024.04.17.590000.

(31) Di Russo, N. V; Martí, M. A.; Roitberg, A. E. Underlying Thermodynamics of PH-Dependent Allostery. *J. Phys. Chem. B* 2014, *118* (45), 12818–12826. https://doi.org/10.1021/jp507971v.

(32) Damjanović, A.; García-Moreno, B.; Lattman, E. E.; García, A. E. Molecular Dynamics Study of Water Penetration in Staphylococcal Nuclease. *Proteins Struct. Funct. Bioinforma.* 2005, *60* (3), 433–449. https://doi.org/https://doi.org/10.1002/prot.20486.

(33) Damjanović, A.; Brooks, B. R.; García-Moreno E., B. Conformational Relaxation and Water Penetration Coupled to Ionization of Internal Groups in Proteins. *J. Phys. Chem. A* 2011, *115* (16), 4042–4053. https://doi.org/10.1021/jp110373f.

(34) Damjanović, A.; Wu, X.; García-Moreno E., B.; Brooks, B. R. Backbone Relaxation Coupled to the Ionization of Internal Groups in Proteins: A Self-Guided Langevin Dynamics Study. *Biophys. J.* 2008, *95* (9), 4091–4101. https://doi.org/https://doi.org/10.1529/biophysj.108.130906.

(35) Hass, M. A. S.; Mulder, F. A. A. Contemporary NMR Studies of Protein Electrostatics. *Annu. Rev. Biophys.* 2015, *44* (Volume 44, 2015), 53–75. https://doi.org/https://doi.org/10.1146/annurev-biophys-083012-130351.

(36) André, I.; Linse, S.; Mulder, F. A. A. Residue-Specific PKa Determination of Lysine and Arginine Side Chains by Indirect 15N and 13C NMR Spectroscopy: Application to Apo Calmodulin. *J. Am. Chem. Soc.* 2007, *129* (51), 15805–15813. https://doi.org/10.1021/ja0721824.

(37) Webb, H.; Tynan-Connolly, B. M.; Lee, G. M.; Farrell, D.; O'Meara, F.; Søndergaard, C. R.; Teilum, K.; Hewage, C.; McIntosh, L. P.; Nielsen, J. E. Remeasuring HEWL PKa Values by NMR Spectroscopy: Methods, Analysis, Accuracy, and Implications for Theoretical PKa Calculations. *Proteins Struct. Funct. Bioinforma.* 2011, *79* (3), 685–702. https://doi.org/https://doi.org/10.1002/prot.22886.

(38) Ancona, N.; Bastola, A.; Alexov, E. PKAD-2: New Entries and Expansion of Functionalities of the Database of Experimentally Measured PKa's of Proteins. *J. Comput. Biophys. Chem.* 2023, *22* (05), 515–524. https://doi.org/10.1142/S2737416523500230.

(39) Reijenga, J.; van Hoof, A.; van Loon, A.; Teunissen, B. Development of Methods for the Determination of PKa Values. *Anal. Chem. Insights* 2013, *8*, ACI.S12304. https://doi.org/10.4137/ACI.S12304.

(40) Wallerstein, J.; Weininger, U.; Khan, M. A. I.; Linse, S.; Akke, M. Site-Specific Protonation Kinetics of Acidic Side Chains in Proteins Determined by PH-Dependent Carboxyl 13C NMR Relaxation. *J. Am. Chem. Soc.* 2015, *137* (8), 3093–3101. https://doi.org/10.1021/ja513205s.

(41) Fitch, C. A.; Karp, D. A.; Lee, K. K.; Stites, W. E.; Lattman, E. E.; García-Moreno, E. B. Experimental PK$_a$ Values of Buried Residues: Analysis with Continuum Methods and Role of Water Penetration. *Biophys. J.* 2002, *82* (6), 3289–3304. https://doi.org/10.1016/S0006-3495(02)75670-1.

(42) Kohda, D.; Sawada, T.; Inagaki, F. Characterization of PH Titration Shifts for All the Nonlabile Proton Resonances in a Protein by Two-Dimensional NMR: The Case of Mouse Epidermal Growth Factor. *Biochemistry* 1991, *30* (20), 4896–4900. https://doi.org/10.1021/bi00234a009.

(43) Kesvatera, T.; Jönsson, B.; Thulin, E.; Linse, S. Measurement and Modelling of Sequence-Specific PKaValues of Lysine Residues in Calbindin D9k. *J. Mol. Biol.* 1996, *259* (4), 828–839. https://doi.org/10.1006/JMBI.1996.0361.

(44) Zhang, M.; Thulin, E.; Vogel, H. J. Reductive Methylation AndpKa Determination of the Lysine Side Chains in Calbindin D9k. *J. Protein Chem.* 1994, *13* (6), 527–535. https://doi.org/10.1007/BF01901534.

(45) Popov, K.; Rönkkömäki, H.; Lajunen, L. H. J. Guidelines for NMR Measurements for Determination of High and Low PKa Values (IUPAC Technical Report). *Pure Appl. Chem.* 2006, *78* (3), 663–675. https://doi.org/doi:10.1351/pac200678030663.

(46) Frericks Schmidt, H. L.; Shah, G. J.; Sperling, L. J.; Rienstra, C. M. NMR Determination of Protein PKa Values in the Solid State. *J. Phys. Chem. Lett.* 2010, *1* (10), 1623–1628. https://doi.org/10.1021/jz1004413.

(47) Chan, C.-H.; Wilbanks, C. C.; Makhatadze, G. I.; Wong, K.-B. Electrostatic Contribution of Surface Charge Residues to the Stability of a Thermophilic Protein: Benchmarking Experimental and Predicted PKa Values. *PLoS One* 2012, *7* (1), e30296-.

(48) 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* (W1), W537–W541. https://doi.org/10.1093/nar/gks375.

(49) Georgescu, R. E.; Alexov, E. G.; Gunner, M. R. Combining Conformational Flexibility and Continuum Electrostatics for Calculating PKas in Proteins. *Biophys. J.* 2002, *83* (4), 1731–1748. https://doi.org/https://doi.org/10.1016/S0006-3495(02)73940-4.

(50) Reis, P. B. P. S.; Vila-Viçosa, D.; Rocchia, W.; Machuqueiro, M. PypKa: A Flexible Python Module for Poisson–Boltzmann-Based PKa Calculations. *J. Chem. Inf. Model.* 2020, *60* (10), 4442–4448. https://doi.org/10.1021/acs.jcim.0c00718.

(51) Wang, L.; Zhang, M.; Alexov, E. DelPhiPKa Web Server: Predicting PKa of Proteins, RNAs and DNAs. *Bioinformatics* 2016, *32* (4), 614–615. https://doi.org/10.1093/bioinformatics/btv607.

(52) Soler, M. A.; Ozkilinc, O.; Hunashal, Y.; Giannozzi, P.; Esposito, G.; Fogolari, F. Molecular Electrostatics and PKa Shifts Calculations with the Generalized Born Model. A Tutorial through Examples with Bluues2. *Comput. Phys. Commun.* 2023, *287*, 108716. https://doi.org/10.1016/J.CPC.2023.108716.

(53) 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* (12), 1277–1283. https://doi.org/https://doi.org/10.1002/prot.25608.

(54) 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* (2), 525–537. https://doi.org/10.1021/ct100578z.

(55) Jensen, J. H.; Li, H.; Robertson, A. D.; Molina, P. A. Prediction and Rationalization of Protein PKa Values Using QM and QM/MM Methods. *J. Phys. Chem. A* 2005, *109* (30), 6634–6643. https://doi.org/10.1021/jp051922x.

(56) Riccardi, D.; Schaefer, P.; Yang; Yu, H.; Ghosh, N.; Prat-Resina, X.; König, P.; Li, G.; Xu, D.; Guo, H.; Elstner, M.; Cui, Q. Development of Effective Quantum Mechanical/Molecular Mechanical (QM/MM) Methods for Complex Biological Processes. *J. Phys. Chem. B* 2006, *110* (13), 6458–6469. https://doi.org/10.1021/jp056361o.

(57) Simonson, T.; Carlsson, J.; Case, D. A. Proton Binding to Proteins: PKa Calculations with Explicit and Implicit Solvent Models. *J. Am. Chem. Soc.* 2004, *126* (13), 4167–4180. https://doi.org/10.1021/ja039788m.

(58) Jorgensen, W. L.; Thomas, L. L. Perspective on Free-Energy Perturbation Calculations for Chemical Equilibria. *J. Chem. Theory Comput.* 2008, *4* (6), 869–876. https://doi.org/10.1021/ct800011m.

(59) Bennett, C. H. Efficient Estimation of Free Energy Differences from Monte Carlo Data. *J. Comput. Phys.* 1976, *22* (2), 245–268. https://doi.org/https://doi.org/10.1016/0021-9991(76)90078-4.

(60) Khandogin, J.; Brooks, C. L. Constant PH Molecular Dynamics with Proton Tautomerism. *Biophys. J.* 2005, *89* (1), 141–157. https://doi.org/https://doi.org/10.1529/biophysj.105.061341.

(61) Lee, J.; Miller, B. T.; Damjanović, A.; Brooks, B. R. Constant PH Molecular Dynamics in Explicit Solvent with Enveloping Distribution Sampling and Hamiltonian Exchange. *J. Chem. Theory Comput.* 2014, *10* (7), 2738–2750. https://doi.org/10.1021/ct500175m.

(62) Lee, J.; Miller, B. T.; Damjanović, A.; Brooks, B. R. Enhancing Constant-PH Simulation in Explicit Solvent with a Two-Dimensional Replica Exchange Method. *J. Chem. Theory Comput.* 2015, *11* (6), 2560–2574. https://doi.org/10.1021/ct501101f.

(63) Mongan, J.; Case, D. A.; McCammon, J. A. Constant PH Molecular Dynamics in Generalized Born Implicit Solvent. *J. Comput. Chem.* 2004, *25* (16), 2038–2048. https://doi.org/10.1002/jcc.20139.

(64) 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* (3), 1341–1352. https://doi.org/10.1021/ct401042b.

(65) 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* (12), 5933–5944. https://doi.org/10.1021/acs.jctc.7b00875.

(66) Kong, X.; Brooks, C. L. Λ-dynamics: A New Approach to Free Energy Calculations. *J. Chem. Phys.* 1996, *105* (6), 2414–2423. https://doi.org/10.1063/1.472109.

(67) Aho, N.; Buslaev, P.; Jansen, A.; Bauer, P.; Groenhof, G.; Hess, B. Scalable Constant PH Molecular Dynamics in GROMACS. *J. Chem. Theory Comput.* 2022, *18* (10), 6148–6160. https://doi.org/10.1021/acs.jctc.2c00516.

(68) Damjanovic, A.; Miller, B. T.; Okur, A.; Brooks, B. R. Reservoir PH Replica Exchange. *J. Chem. Phys.* 2018, *149* (7), 72321. https://doi.org/10.1063/1.5027413.

(69) Itoh, S. G.; Damjanović, A.; Brooks, B. R. PH Replica-Exchange Method Based on Discrete Protonation States. *Proteins Struct. Funct. Bioinforma.* 2011, *79* (12), 3420–3436. https://doi.org/10.1002/prot.23176.

(70) Cai, Z.; Luo, F.; Wang, Y.; Li, E.; Huang, Y. Protein PKa Prediction with Machine Learning. *ACS Omega* 2021, *6* (50), 34823–34831. https://doi.org/10.1021/acsomega.1c05440.

(71) Cai, Z.; Peng, H.; Sun, S.; He, J.; Luo, F.; Huang, Y. DeepKa Web Server: High-Throughput Protein PKa Prediction. *J. Chem. Inf. Model.* 2024, *64* (8), 2933–2940. https://doi.org/10.1021/acs.jcim.3c02013.

(72) 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 PKa Predictions in Proteins. *J. Chem. Theory Comput.* 2022, *18* (8), 5068–5078. https://doi.org/10.1021/acs.jctc.2c00308.

(73) Gokcan, H.; Isayev, O. Prediction of Protein PKa with Representation Learning. *Chem. Sci.* 2022, *13* (8), 2462–2474. https://doi.org/10.1039/D1SC05610G.

(74) Chen, A. Y.; Lee, J.; Damjanovic, A.; Brooks, B. R. Protein PKa Prediction by Tree-Based Machine Learning. *J. Chem. Theory Comput.* 2022, *18* (4), 2673–2686. https://doi.org/10.1021/acs.jctc.1c01257.

(75) Liu, S.; Yang, Q.; Zhang, L.; Luo, S. Accurate Protein PKa Prediction with Physical Organic Chemistry Guided 3D Protein Representation. *J. Chem. Inf. Model.* 2024, *64* (11), 4410–4418. https://doi.org/10.1021/acs.jcim.4c00354.

(76) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. L. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* 2009, *25* (11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163.

(77) Cannon, B. R. Thermodynamic Consequences of Substitutions of Internal Positions in Proteins with Polar and Ionizable Residues, Johns Hopkins University, 2008.

(78) Karp, D. A.; Gittis, A. G.; Stahley, M. R.; Fitch, C. A.; Stites, W. E.; García-Moreno E., B. High Apparent Dielectric Constant Inside a Protein Reflects Structural Reorganization Coupled to the Ionization of an Internal Asp. *Biophys. J.* 2007, *92* (6), 2041–2053. https://doi.org/10.1529/biophysj.106.090266.

(79) Isom, D. G.; Castañeda, C. A.; Cannon, B. R.; Velu, P. D.; Garc\'\ia-Moreno E., B. Charges in the Hydrophobic Interior of Proteins. *Proc. Natl. Acad. Sci.* 2010, *107* (37), 16096–16100. https://doi.org/10.1073/pnas.1004213107.

(80) Isom, D. G.; Castañeda, C. A.; Cannon, B. R.; Garc\'\ia-Moreno E., B. Large Shifts in PKa Values of Lysine Residues Buried inside a Protein. *Proc. Natl. Acad. Sci.* 2011, *108* (13), 5260–5265. https://doi.org/10.1073/pnas.1010750108.

(81) Harms, M. J.; Castañeda, C. A.; Schlessman, J. L.; Sue, G. R.; Isom, D. G.; Cannon, B. R.; García-Moreno E., B. The PKa Values of Acidic and Basic Residues Buried at the Same Internal Location in a Protein Are Governed by Different Factors. *J. Mol. Biol.* 2009, *389* (1), 34–47. https://doi.org/https://doi.org/10.1016/j.jmb.2009.03.039.

(82) Castañeda, C. A.; Fitch, C. A.; Majumdar, A.; Khangulov, V.; Schlessman, J. L.; García-Moreno, B. E. Molecular Determinants of the PKa Values of Asp and Glu Residues in Staphylococcal Nuclease. *Proteins Struct. Funct. Bioinforma.* 2009, *77* (3), 570–588. https://doi.org/https://doi.org/10.1002/prot.22470.

(83) Baran, K. L.; Chimenti, M. S.; Schlessman, J. L.; Fitch, C. A.; Herbst, K. J.; Garcia-Moreno, B. E. Electrostatic Effects in a Network of Polar and Ionizable Groups in Staphylococcal Nuclease. *J. Mol. Biol.* 2008, *379* (5), 1045–1062. https://doi.org/https://doi.org/10.1016/j.jmb.2008.04.021.

(84) Arthur, E. J.; Yesselman, J. D.; Brooks III, C. L. Predicting Extreme PKa Shifts in Staphylococcal Nuclease Mutants with Constant PH Molecular Dynamics. *Proteins Struct. Funct. Bioinforma.* 2011, *79* (12), 3276–3286. https://doi.org/https://doi.org/10.1002/prot.23195.