JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

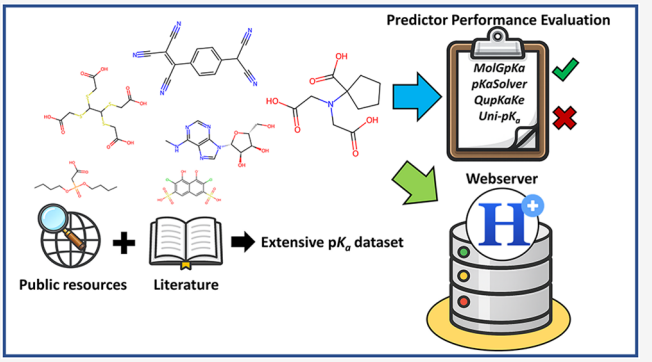
pubs.acs.org/jcim Article

Benchmarking pK_a Prediction Algorithms against an Extensive, Public Data Set

Levente Sipos-Szabó, Dávid Bajusz,* György T. Balogh, and György M. Keserű

Cite This: J. Chem. Inf. Model. 2026, 66, 4607−4619 Read Online

ABSTRACT: Accurate prediction of proton dissociation constants (pK_a) is essential for downstream drug discovery and molecular modeling workflows. While several proprietary pK_a prediction tools have been established as popular choices in the field, open-source solutions provide viable alternatives for large-scale computational workflows. In particular, machine learning approaches have recently emerged as a promising orthogonal route to traditional empirical methods. However, many of these algorithms were benchmarked on small, disjoint data sets, with inconsistencies in pK_a data interpretation, particularly for polyprotic molecules. To address these challenges, we assembled a comprehensive data set of over 90,000 experimental aqueous pK_a values spanning over 31,000 unique molecules from scattered online resources, with each entry annotated for charge state transitions and microspecies distributions. This data set, made accessible through the pKahub online database, represents one of the largest publicly available collections of annotated pK_a data to date. We used this resource to benchmark seven pK_a prediction methods, including three commercial tools (ACD/Labs, Chemaxon, and Epik) and four open-source machine learning models (MolGpKa, pKaSolver, QupKake, and Uni-pK_a).



INTRODUCTION

Many molecules tend to undergo spontaneous ionization in aqueous (or more generally, protic) media, mainly due to association or dissociation of mobile protons.[1] According to the Brønsted−Lowry acid−base theory, this behavior can be quantified by the reaction equilibrium constant of proton dissociation or association corrected by the concentration of the solvent. In practical scenarios, the negative logarithm is used for equilibrium constants; these are called pK_a and pK_b for acidic and basic reactions (proton dissociation and association), respectively. Ionization can have a drastic effect on physicochemical properties, such as solvation, lipophilicity, membrane permeability, and supramolecular interactions, all of them being crucially important aspects of drug efficacy and safety.[2,3] Enumerating ionized states for small molecules is especially important for structure-based molecular modeling and virtual screening workflows. The protonation state of the ligand greatly affects the secondary interaction types and strengths with protein binding sites, such as electrostatic interactions, hydrogen bonding, and van der Waals interactions.[4]

Since large-scale proton dissociation constant determination experiments are infeasible for many, there are multiple methods for the computational prediction of pK_a values. (Crucially, these methods can also process virtual molecules not yet synthesized, where experimentation is, by definition, not available.) These can be theoretical approaches utilizing high-level energy calculations,[5,6] such as density functional theory (DFT) along with thermodynamic cycles and advanced solvation models like COSMO[7] or EC-RISM.[8,9] Trading precision for higher throughput, empirical methods are based on QSPR approaches, using large, experimentally determined pK_a data sets to train prediction models on chemical features to determine the pK_a value of the query compound or functional group.[10,11] For common use cases, empirical methods are far more utilized due to their computation time advantage, as ab initio or DFT-based methods can require a day's worth of computation on computer clusters for a single molecule, while empirical methods usually take a few seconds or even less to evaluate a compound on a desktop computer.[6] A perhaps less discussed advantage of empirical methods is that while they have a more limited applicability domain in accordance with their training data set, they usually outperform theoretical approaches in benchmarking studies.[12] Therefore, empirical methods have been the established choices for general use cases such as ADMET property prediction or structure preparation for molecular modeling, with the most widely used alternatives being

Received: January 13, 2026
Revised: March 26, 2026
Accepted: March 27, 2026
Published: April 6, 2026

https://doi.org/10.1021/acs.jcim.6c00107
J. Chem. Inf. Model. 2026, 66, 4607−4619

proprietary software, like Percepta from Advanced Chemistry Development, Inc. (ACD/Labs), and the property calculator modules of Chemaxon or Schrödinger Epik.[13] This is especially the case for methodologies relying on linear free energy relationships (LFER), which were one of the leading paradigms in empirical pK_a prediction before the advent of machine learning approaches.[14,15]

In contrast, freely available and open-source applications for general use have been scarcer. This is especially the case for the full enumeration of possible protonation states for a molecule, which is required by advanced property predictions and downstream molecular modeling tasks such as virtual screening campaigns. To the best of our knowledge, the only mature tool for this task is Dimorphite-DL, which uses heuristic rules for enumerating the ionization states of predefined centers, according to a given pH.[16] A freely available tool for the quantitative evaluation of ionized microspecies is still a great need. This is also the case for many commercial products, as they usually provide only microspecies distributions through graphical user interfaces, limiting the use cases for enumerating large compound libraries.

Recent years have seen the large-scale adaptation of machine learning methods.[17−19] In machine learning, the model adapts to the chemical input data, which circumvents the need for the laborious, hand-crafted equations needed for previous methods such as LFER, making academic and open-source development more feasible. In the past few years, several new predictor models based on deep neural networks were released and many of these models are freely available and show competitive performance with commercial methods.[17,20,21]

The cornerstone of every machine learning method is the quality and amount of data used to train it.[22] Compared with the training requirements of advanced deep learning methods, the available amount of experimental pK_a data is rather low. To handle this, almost every recently published method applies a pretraining-fine-tuning paradigm: in the pretraining phase, a large amount of calculated pK_a values are utilized from the ChEMBL database,[23] which are then followed by fine-tuning on a smaller set of experimental data. For experimental pK_a data, the largest public data set is the iBonD databank,[24] which contains over 40,000 pK_a values to date. Interestingly, there is no restricting license on these data, but the database can only be parsed in a purely visual manner, making it difficult to adapt for programmatic use cases. Large portions of this data set were manually collected and released as training sets in recently published predictors, making them more easily accessible.[25] Other predictors used orthogonal training sets like the pK_a data set released with the Datawarrior[26] software, or its modified versions.[27] This means that most method developers utilize only a subset of the publicly available data, for example, the large and reliable IUPAC digitized K_a dataset is rarely used for training or benchmarking these new methods.[28]

The increasing impetus in method development has also highlighted some serious misconceptions concerning pK_a data interpretation in the scientific community, particularly with polyprotic small molecules.[29] The main source of confusion seems to be the oversimplification of the more complex ensemble of the ionized microspecies of molecules with more than one ionization center. A common mistake arises from the conceptually tempting notion of assigning a single pK_a value to the whole molecule, or to a single functional group. This practice works for simple acids and bases but leads to inconsistencies when dealing with more complex molecules. This is also reinforced by the fact that almost every available experimental pK_a data point describes a macroscopic average of a given set of microspecies. For a comprehensive discussion of this topic, we would like to point the reader to the article of Fraczkiewicz.[30] These problems also leaked into frequently used data sources such as the ChEMBL database and affect the quality of many predictors using these data, as pointed out by Zheng et al.[29] To this end, it is recommended to clearly annotate the exact charge state transition (e.g., "0 to −1"), along with published pK_a values, an information that is lacking in public data sources. To date, there is also no widely accessible pK_a data repository that contains detailed microspecies data along with pK_a values. To address the above-mentioned issues, we assembled the largest data set of experimental aqueous pK_a values from the literature and various public sources to date, containing over 90,000 data points across more than 31,000 unique molecules. Each pK_a value was annotated with a charge state transition and relevant microspecies distributions. We make this data set accessible through a Web server called pKaHub (http://pkahub.ttk.hu). Here, we use this extensive data set to benchmark and compare three commercial and four open-source machine learning-based pK_a prediction methods, namely, ACD/Labs, Chemaxon, Epik,[13] MolGpKa,[17] pKaSolver,[31] QupKake,[20] and Uni-pK_a.[21]

METHODS

pK_a Terminology

The term pK_a denotes the negative base-10 logarithm of the equilibrium constant for a proton dissociation reaction expressed relative to the activity of the solvent. In the simplest case, it describes deprotonation of a conjugate acid at a single acidic functional group. More generally, for molecules with multiple ionizable sites, a deprotonation step occurs between two microspecies that differ in the presence of a single proton; the equilibrium constant for such a site-specific transition is described by a microscopic pK_a (micro-pK_a). Most experimental approaches for determining proton-transfer equilibria rely on titration-based measurements, which typically report transitions between overall (net) charge states of the molecule. These observed transitions therefore correspond to averages over multiple underlying microscopic protonation/deprotonation events rather than a single site-specific process and are commonly referred to as macroscopic pK_a (macro-pK_a) values.

Predictors

We selected four open-source, deep learning-based pK_a predictors published in recent years as the basis of our evaluation. These methods have been reported to perform comparably to commercial tools on small benchmark data sets (e.g., SAMPL), while remaining readily accessible through publicly available code and pretrained weights. Together, they also reflect the recent evolution of deep-learning architectures for pK_a prediction. MolGpKa was among the earliest graph neural network (GNN)-based predictors for small organic molecules. It was trained primarily on calculated pK_a values (ACD/Labs) available through the ChEMBL v25 data set. pKaSolver was likewise trained on ChEMBL-derived pK_a values calculated with Epik, but it additionally introduces a fine-tuning stage by using experimentally measured pK_a data to improve predictive accuracy. pKaSolver uses Dimorphite-DL to enumerate one dominant microspecies for the different charge states of the query molecule and performs sequential ionization prediction by predicting the micro-pK_a between the enumerated forms; accordingly, we treat it as a microscopic pK_a predictor. For the Epik-derived version in the original work, the model weights could not be shared due to licensing restrictions. The open version (called pKaSolver-lite in the original article) of the model was trained using only the experimental data. As our goal was to track the performance of freely available methods, we did not attempt to recreate the proprietary variant. QupKake similarly includes an initial training phase on calculated pK_a values made accessible through ChEMBL, although in this case, the calculations were performed with

https://doi.org/10.1021/acs.jcim.6c00107
J. Chem. Inf. Model. 2026, 66, 4607−4619

Marvin (Chemaxon), rather than ACD/Labs, followed by experimental-data fine-tuning. A distinctive feature of QupKake is its use of quantum-chemical descriptors computed with the semiempirical GFN2-xTB method.[32] Its protonation-site identification is also atypical: whereas most methods rely on predefined protonation templates, QupKake employs a separate neural network to detect ionizable sites. This model was trained using ionization-site scanning calculations for ChEMBL molecules generated with the CREST semiempirical quantum chemistry workflow.[33] Finally, Uni-pK_a follows a different strategy. For a given query molecule, it enumerates possible microspecies using a template-based set of ionization rules to identify candidate sites, generating states through sequential protonation and deprotonation. For each microspecies, it predicts a dimensionless free energy from which macroscopic and microscopic pK_a values can be derived via Boltzmann averaging. Uni-pK_a also uses ChEMBL-based training and experimental-data fine-tuning, primarily drawing experimental labels from the DataWarrior and Novartis data sets. We evaluated Uni-pK_a in two configurations. One with the simplified template with around 30 site definitions, as recommended by the authors for general use, and the other with the full template compiled by the authors, which includes many ionization sites that are rarely relevant for typical drug-like molecules. Although the weights we used differ from those reported in the original publication, this implementation remains the most accessible in terms of practical utility. The pK_a predictors used in this study are summarized in Table 1.

Table 1. Open-Source and Commercial pK_a Predictors Included in the Benchmarks

name	proprietary or open-source	predicted pK_a type (with batch calculations)	predictor architecture
MolGpKa[17]	open-source	micro	graph neural network
pKaSolver[31]	open-source	micro with sequential ionization	graph neural network
QupKake[20]	open-source	micro	graph neural network
Uni-pK_a with simple template[21]	open-source	micro and macro	SE(3)-transformer-based foundational model
Uni-pK_a with full template[21]	open-source	micro and macro	SE(3)-transformer-based foundational model
ACD/Labs Classic[34]	proprietary	macro	LFER
ACD/Labs GALAS[34]	proprietary	macro	undisclosed
Chemaxon[35]	proprietary	micro and macro	undisclosed
Epik[13,36]	proprietary	macro	graph neural network

In the following sections, we generally refer to MolGpKa, pKaSolver, QupKake, and Uni-pK_a as open-source, and Chemaxon, Epik, and ACD/Labs as commercial or proprietary predictors. MolGpKa, pKaSolver, and QupKake were implemented using the code and weights distributed on GitHub, with minimal modifications in some cases to carry out batch calculations. QupKake was used as the command-line utility. Uni-pK_a was used according to the code on Bohrium with the uploaded weights (www.bohrium.com/notebooks/38543442597). The Uni-pK_a code for microstate enumeration was changed to enumerate states in the −6 to +6 charge window. Epik was used in the Schrodinger 2025-2 package (epikx), and input was processed using Ligprep before predictions with enumerated charge states set to −6 to +6. The same Epik workflow was used for the creation of the database and for performance evaluations. Chemaxon pK_a calculation was carried out using the cxcalc command-line utility with the "large model" selected.

Data Set Collection and Preparation

We collected various pK_a data compilations from the literature and public data resources. Many literature data sets already had citations in other works and somewhat-established names, being mostly referred to by the name of the first author of the original publication. Here, we kept this convention for unnamed, lesser-known data sets. As an exception, the name "AvLiLuMoVe" refers to the filename used in the original publication's Supporting Information, where it is labeled the "Literature" data set (we avoided this label here, as it could be misleading in our context). In the text, we refer to these as "source data sets".

The literature compilations included the Settimo, Hunt, Jensen, Caine, Manchester, Organic Oxygen Acids, and Nitrogen Bases data sets, which were acquired from the Supporting Information of these articles. Chemical structures and pK_a values from the Caine and Manchester data sets were digitized by us using the OSRA optical recognition software.[37] The Baltruschat ChEMBL data consisted of experimental proton dissociation constant data mined by Baltruschat et al. from ChEMBL. The AvLiLuMoVe and Novartis data sets were also acquired from this source. The Datawarrior data set was downloaded from the databanks of the Datawarrior software. The IUPAC digitized pK_a dataset is a combined effort to digitalize the pK_a data from various chemistry handbooks. The AttenGpKa training set is a large compilation of data which was reportedly acquired via the manual mining of the iBonD database.[24] The OCHEM data was acquired from the Online Chemical Modeling Environment website, a public data repository. The QSARToolbox data was downloaded from the "pKa" and "pKa OASIS" database end points of the QSARToolbox software. The SAMPL data sets were acquired from the associated GitHub repositories of the challenges. The collected data sets are summarized in Table 2.

Table 2. Summary of the Collected Source Datasets before and after Data set Processing, with the Number of Contained pK_a and Other Types of Dissociation Constant Values and the Number of Unique Molecules (Unique at the Dataset Level)

	before processing		after processing	
source data set name	proton dissociation constants	unique molecules	pK_a entries	unique molecules
Settimo[38]	622	426	622	426
Hunt[39]	2435	2242	2430	2237
Jensen[40]	53	48	53	48
Caine[41]	78	71	78	71
Manchester[42]	142	85	142	85
Organic Oxygen Acids and Nitrogen Bases[43]	1143	1123	1143	1123
Baltruschat ChEMBL[44]	8106	6303	7752	6093
AvLiLuMoVe[44]	123	123	123	123
Novartis[44]	280	280	280	280
Datawarrior[26]	7913	6375	7495	6050
IUPAC digitized pK_a dataset[28]	24,211	10,615	18,513	8180
AttenGpKa training set[25]	26,522	15,712	13,556	10,384
OCHEM[45]	21,687	11,855	20,708	11,112
QSARToolbox[46]	28,926	13,390	24,609	12,341
SAMPL6[47]	31	24	31	24
SAMPL7[48]	20	20	20	20
SAMPL8[49]	24	21	24	21
euroSAMPL1[12]	35	35	35	35
combined	122,351	38,701	97,614	31,250

Individual source data sets were processed separately. For the processing, we included only pK_a or pK_b type of proton dissociation or association constants and converted all pK_b values to pK_a by subtracting them from 14.0 (pK_w, from the self-dissociation constant of water). We excluded entries where the chemical structure could not be processed by RDKit, except in a few cases when we manually repaired structures with a problematic SMILES representation regarding the aromaticity of indole rings.

https://doi.org/10.1021/acs.jcim.6c00107
J. Chem. Inf. Model. 2026, 66, 4607−4619

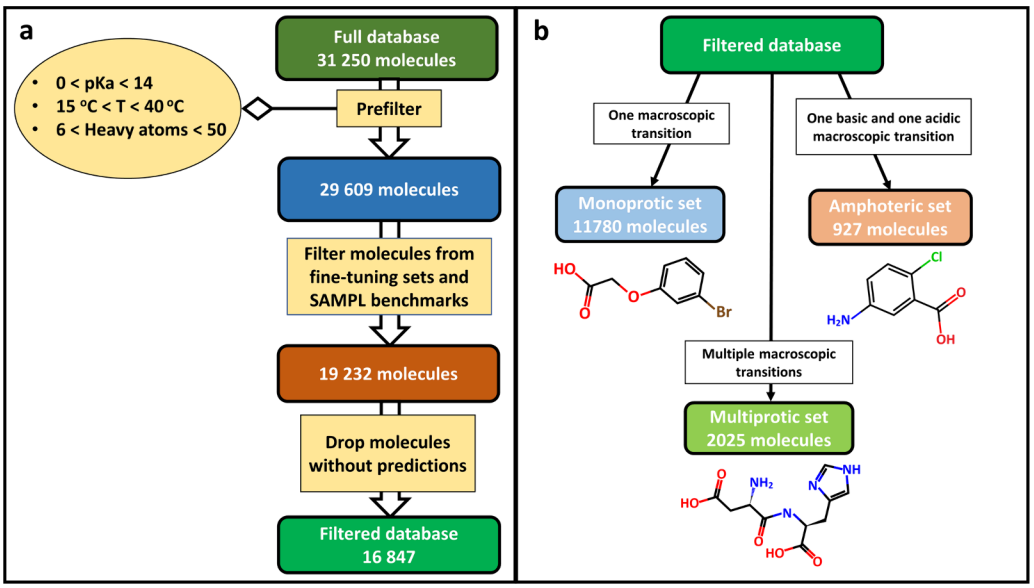


Figure 1. (a) Filtering steps applied to the full database. (b) Three benchmark data sets for predictor validation were compiled from the filtered and curated database.

Entries were also excluded if there was an implication that the measurement was not carried out in water. This included entries from the AttenGpKa training set and the IUPAC digitized pK_a data set, where we excluded all entries where the cosolvent content was greater than 5% or when the cosolvent content was not quantitatively specified. Entries in the IUPAC digitized data set with nonconventional experimental settings, such as covalent hydration, excited states, or extreme pressure, were also filtered out, as well as entries where the remark specified a lack of information on the experimental conditions.

Charge state transition assignment was carried out for every experimental pK_a value by matching them to the calculated macroscopic pK_a values by Epik. By manual examination of the Jensen, Manchester, and all SAMPL data sets, we could assume that in cases of multiple, different pK_a entries for the same molecule, the entries refer to different charge state transitions. For these values, we used an order-preserving matching algorithm (Figure 2), which matches the order of the experimental and predicted pK_a values when ordering them in an ascending (or descending) order. The AvLiLuMoVe and Novartis data sets contained only one pK_a value for a single molecule: in this case, the charge state transition with the lowest difference to the predicted value was chosen. For the Settimo, Hunt, Caine, Organic Oxygen Acids and Nitrogen Bases, AttenGpKa, OCHEM, QSARToolbox, Baltruschat CheMBL, and DataWarrior data sets, it was not certain whether different values belong to different transitions for the same molecule. In this case, a given experimental value was matched to the predicted value that was closest to it. For the IUPAC digitized pK_a dataset, there is a "pKa type" label associated with every entry. This describes the type of ionization constant the numeric value refers to, such as "pKb" or "pKa", and in some cases, also describes whether the constant refers to a different dissociation step (e.g., "pKa1" and "pKa2") for polyprotic molecules. This proved to be consistent among a large section of the data set. However, in a few hundred cases, we found the notations to be somewhat inconsistent. Therefore, for every unique molecule in the data set, we grouped the pK_a values by their types and checked (i) if there were any two values within one group with an absolute difference greater than 1.0, or (ii) if there were any pair of values between two groups with an absolute difference smaller than 0.5. When the groupings passed the check, we assigned the experimental pK_a values with order-preserving matching.

We excluded all data where the structure could not be processed by Epik or where we could not get any predictions. This usually involved cases with an erroneous structure or structures with more exotic ionization centers, such as aliphatic or aromatic C−H deprotonation, except for 8 values from the Novartis data set and two values from the Hunt data set which were processed manually. To filter out possibly misannotated or erroneous experimental values from further processing, we excluded all molecules and associated pK_a entries in a given data set, where there was at least one matched experimental vs predicted pK_a value pair with a greater difference than 4.

We also extracted temperature and ionic strength values for each data point where possible, although temperature values were only specified for some of the IUPAC and OCHEM entries. Ionic strength values were only provided by the IUPAC digitized pK_a dataset for a subset of entries.

Averaging pK_a Values for Charge State Transitions

We define an aggregated set of experimental pK_a values for every unique molecule in the full database by averaging every experimental pK_a value assigned to the same charge state transition of that molecule (from now on, we refer to these values as the experimental pK_a values of that molecule for that transition). Recognizing that several pK_a values across different source data sets could be originating from the same measurement, we applied a cross-data set deduplication procedure before this averaging step. For each molecule, pK_a values are grouped by source data set and charge state transition. If the same numeric pK_a value appears for the same transition in multiple source data sets, it is likely a duplicate entry propagated across source data sets, so it is counted only as many times as the minimum occurrence count in any of those source data sets. (If a value appears in only one source data set, its count is kept as-is.) The resulting counts determine how many times each pK_a value enters the final list and contributes to the aggregated pK_a for the given charge state transition during averaging.

Benchmark Data Sets for Predictor Evaluation

We compiled three benchmark data sets from the combined database of 31,250 molecules to test the open-source predictors against the commercial ones by applying a sequence of filtering steps and grouping the remaining molecules. The process for creating the three data sets is shown in Figure 1.

We included only molecules with all of their pK_a values between 0 and 14. We filtered out every molecule that has a heavy atom count lower than 6 or greater than 50. We also filtered every pK_a entry which had an assigned temperature value outside of the 15−40 °C interval. This reduced the number of molecules to 29,609. From this, we filtered out every molecule which was also present in the fine-tuning data of any of the open-source predictors or was in any of the popular benchmark sets, such as the SAMPL sets or the Novartis set. (The latter was also excluded because it was also a part of the Uni-pK_a fine-tuning set for the used weights.) This filtering left 19,232 molecules in the set.

https://doi.org/10.1021/acs.jcim.6c00107
J. Chem. Inf. Model. 2026, 66, 4607−4619

Stereochemistry was omitted when these comparisons were made with the fine-tune set, as some predictors do not consider stereocenters in their predictions and in some cases, we found the handling of stereochemistry information inconsistent in these data sets. To test the predictors on equal footing using the same set of molecules, we filtered out any molecule for which at least one of the relevant predictors did not yield a valid prediction. A prediction was considered sufficient if the number of predicted pK_a values was at least the number of experimental pK_a values available for that molecule. We defined three groups of molecules. Molecules with a single experimental pK_a and an assigned charge-state transition of 0 to −1 (acid) or 1 to 0 (base) were labeled as monoprotic. Molecules with exactly two experimental pK_a values and assigned charge-state transitions of 0 to −1 and 1 to 0 formed the amphoteric group. Molecules with more than one experimental pK_a value that were not labeled as amphoteric were labeled as polyprotic. For the monoprotic and amphoteric groups, we excluded any molecule with an insufficient prediction from any predictors. For the polyprotic group, only macroscopic predictors were considered. After filtering, the final data set comprised 16,847 molecules. The filtered molecules were compiled into the monoprotic, amphoteric, and polyprotic data sets according to their labels. We did not include polyprotic molecules in the benchmarking data sets, which had only one assigned transition where both involved states had a nonzero net charge, as it implies that the molecule has experimentally undetected charge state transitions.

Cheminformatics Tools

In the majority of cases, RDKit (version 2024.03.5) was used for the processing of chemical data. For molecular similarity calculations, Morgan fingerprints of radius 4 with 2048 bits were generated, and similarity was calculated using the Tanimoto index.[50] In the evaluations, two molecules were considered equivalent if they had the same InChI strings. Data set processing and the evaluation of predictors were performed using in-house Python scripts. Similarity between two sets of molecules was calculated by averaging the similarity values of all of the molecules in the smaller set with their most similar counterpart from the partner set.

One-Way ANOVA

To establish statistical significance between the distributions of absolute prediction errors among the evaluated micro- and macro-pK_a predictors, one-way analysis of variance (ANOVA) was performed using the scipy (version 1.16.3) Python package. Absolute errors between experimental and predicted pK_a values were aggregated across all molecular benchmark data sets (monoprotic, amphoteric, and polyprotic benchmark sets) for each predictor, yielding one independent sample per predictor. The null hypothesis that all predictors share a common population mean absolute error was tested at a significance level of $\alpha = 0.05$. Upon rejection of the null hypothesis, pairwise post hoc comparisons were conducted to identify which specific predictor pairs differed significantly. Five established parametric post hoc tests were applied: Scheffé's test, Tamhane's T2 test, the pairwise Student's *t*-test with Bonferroni-Holm correction, Tukey's range test, and Tukey's honestly significant difference (HSD) test. All post hoc procedures were implemented using the scikit-posthocs library (version 0.12.0) and applied at the same $\alpha = 0.05$ threshold. To obtain a single, robust significance determination for each predictor pair, the binary outcomes of the five post hoc tests were aggregated by majority vote: a pair was deemed statistically significantly different if at least three of the five tests independently reported $p < 0.05$ for that pair.

Web Server

The pKaHub web server was implemented using the Django (version 5.2) framework with the SQLite database engine. The microspecies distribution for a given charge state of a molecule is also provided by using the microstate enumerator of Uni-pK_a. Here, we primarily used simple templates for generating microspecies; in cases when the enumeration with simple templates was insufficient to provide any microspecies for the given charge state, full template enumeration was used. Microspecies in a given charge state were filtered by discarding every structure with a state population of less than 5% according to the predicted free energy by Uni-pK_a.

RESULTS AND DISCUSSION

Assignment of pK_a Values to Calculated Charge State Transitions

A typical pK_a determination involves titrating the solution of the molecule across a defined pH range while continuously monitoring changes in a suitable physicochemical signal, most commonly via potentiometric or spectrophotometric measurements. In ideal cases, the observed signal can be described as a population-weighted sum of contributions from the relevant protonation states, provided that the individual states exhibit sufficiently distinct responses. (This assumption does not always hold, e.g., in spectrophotometric titrations of molecules lacking an adequate chromophore, different protonation states may produce indistinguishable spectra.) By fitting the pH dependence of this signal with an appropriate equilibrium model, one can extract the equilibrium constants (pK_a values) describing the system's pH-dependent state transitions. In this setting, the measurement yields only a sequence of apparent transition values consistent with the measured pH range but does not directly identify the specific protonation states involved in each transition.

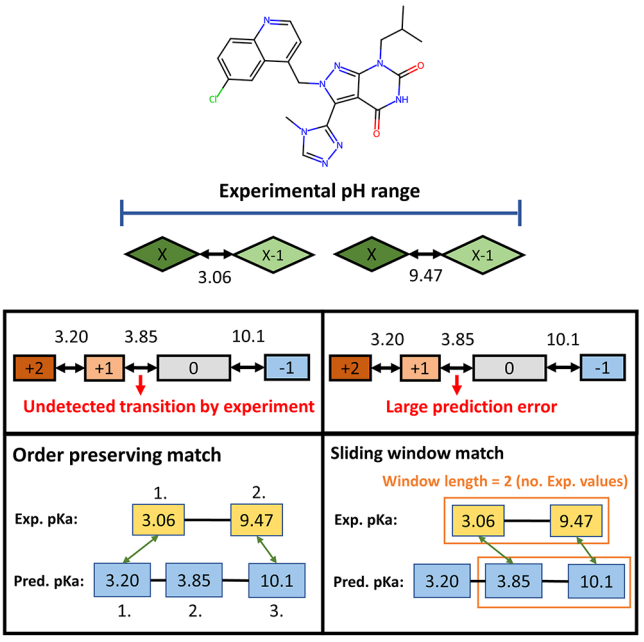


Figure 2. Assignment of experimental pK_a values to macroscopic charge state transitions through matching the numeric values of the experimental and calculated pK_as. Boxes refer to relevant (nonexotic) macroscopic charge states within a given pH range, and bidirectional arrows refer to macroscopic charge state transitions, with the assigned numbers referring to macro-pK_a values. When using order-preserving matching of the pK_a values, we assume that there could be "gaps" in the detected experimental state transitions. If we accept that the predictor could have large errors and the experiment registers all relevant macroscopic transitions, we can use a window search technique by enforcing a one-to-one correspondence in a consecutive order between the experimental and predicted values.

Assigning specific microscopic states to experimentally derived macroscopic constants is generally not straightforward, especially by experiment. A practical route to a more detailed description of the underlying transitions is, therefore, to align the experimental values with computationally predicted dissociation constants. In structure-based prediction frameworks, each

https://doi.org/10.1021/acs.jcim.6c00107
J. Chem. Inf. Model. 2026, 66, 4607−4619

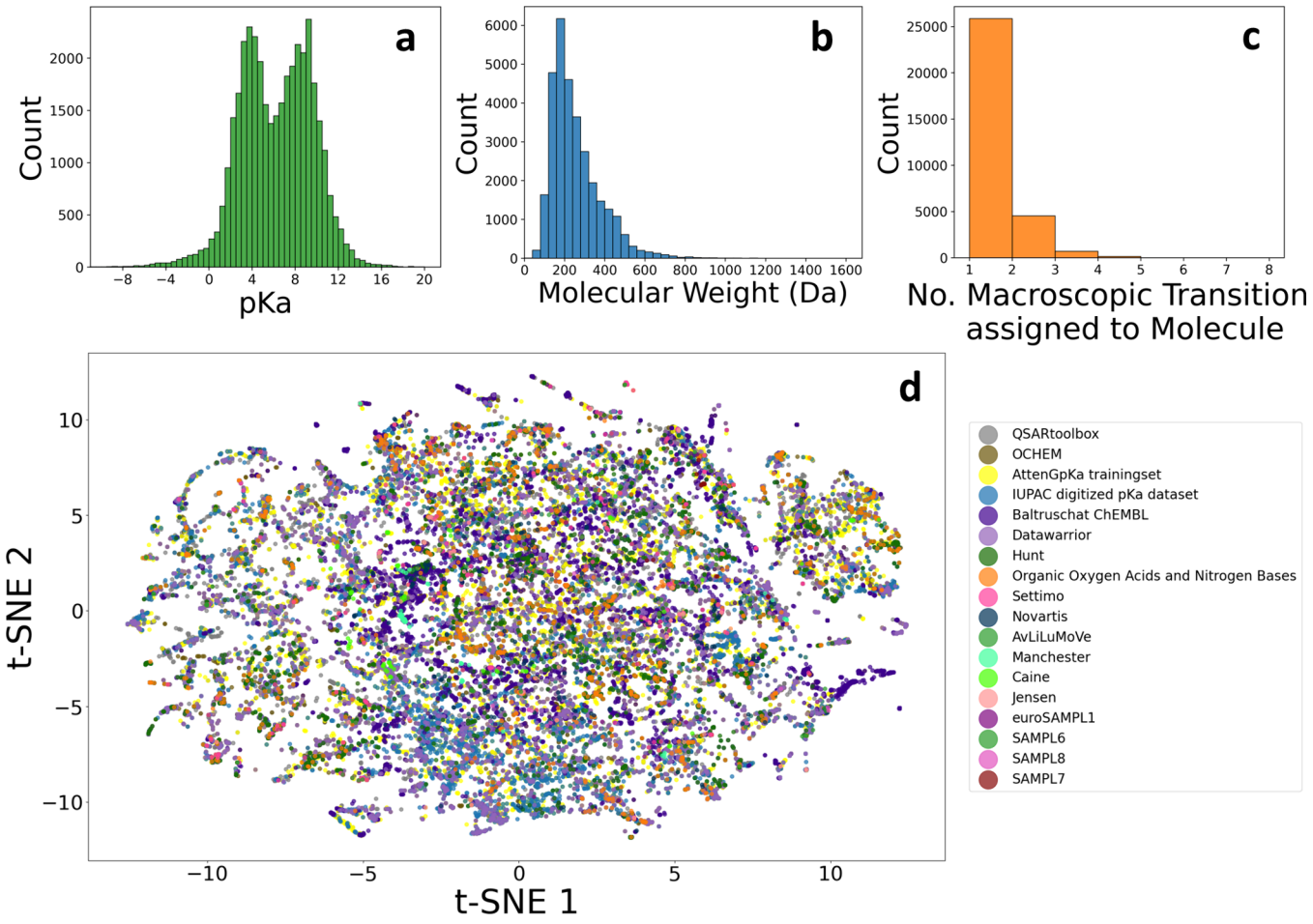


Figure 3. Distribution of (a) pK_a values in the combined database. (Averaged per charge state transition for a molecule, pK_a values with experimental temperatures outside the 15−40 °C temperature window were excluded from the average, numerically equivalent pK_a values from different source data sets were excluded.) (b) Molecular weights. (c) Number of charge state transitions assigned to a molecule. (d) t-SNE plot representing the chemical space of the source data sets.

predicted pK_a is associated with a defined protonation or microstate transition, providing the state-level annotation that is typically missing from experiment.

Formally, this amounts to matching two ordered sequences: an experimental set of length N and a predicted set of length M, where the goal is to assign each experimental value to one of the predicted values. We generally expect $M \geq N$; if fewer transitions are predicted than observed, then the predictor has likely failed to enumerate the relevant ionization states. To obtain a fair and chemically consistent alignment, it is recommended that the matching algorithm should satisfy two principles.[51,52] First, it should be order-preserving: if experimental values e_i and e_j (with $i < j$) are matched to predicted values p_k and p_l, then the indices must satisfy $k < l$, preventing "crossing" assignments and preserving the monotonic progression of transitions with pH. Second, among all order-preserving matchings, the method should minimize the sum of absolute deviations between the matched values, $\sum_1^N |e_n - p_m|$. These principles ensure numerically precise correspondences while respecting the relative magnitudes of the predicted transitions.

Notably, these principles do not require the matched predicted values to form a consecutive block within the predicted sequence. Allowing gaps allows the possibility of experimentally undetected transitions, leaving some predicted pK_a values unassigned. This choice is deliberately agnostic with respect to experimental sensitivity and imposes minimal constraints on the alignment problem. However, it implicitly assumes that experimentally undetected transitions are more common than large predictor errors. If, instead, one assumes that the experiment captures all transitions within the relevant pH window, then the matched predicted values should be consecutive; in that regime, the assignment reduces to a sliding-window alignment of the N experimental values against all length-N consecutive segments of the M-predicted values. We illustrate the two kinds of assignment paradigms for an example molecule in the data set on Figure 2.

The problem becomes more challenging in large aggregated data sets, where metadata rarely indicate whether two reported pK_a values originate from the same experiment, or whether they refer to distinct experimental state transitions; in many cases, a pK_a value only has a single assigned chemical structure without sufficient labels. In this case, a maximally permissive strategy is to assign each experimental value to its closest predicted value. Under a stricter "complete observation" assumption, one would instead minimize the total assignment error subject to the consecutiveness of the selected predicted values, potentially allowing multiple experimental measurements to map to the same predicted transition.

As our compiled database spans a wide range of experimental methodologies and, possibly, a wide range of conditions, it is

https://doi.org/10.1021/acs.jcim.6c00107
J. Chem. Inf. Model. 2026, 66, 4607−4619

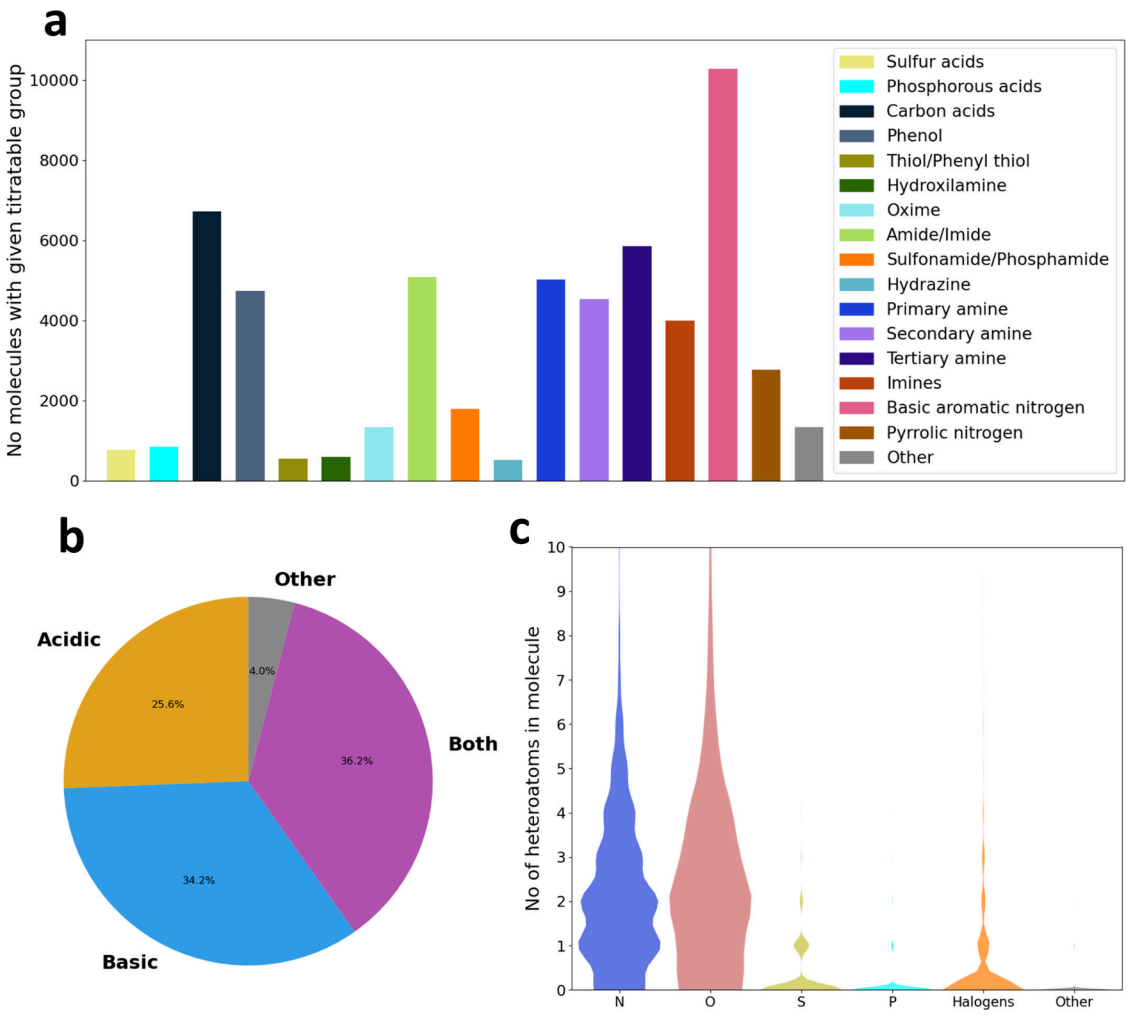


Figure 4. Distributions of (a) the number of molecules in the database that contain at least one titratable group in the given categories; (b) molecules with only acidic, basic, or both kinds of titratable groups ("Other" denotes molecules with only nonconventional titratable groups such as C−H acids). (c) Violin plot showing the distribution of the number of heteroatoms per molecule in the database.

Table 3. Summaries of the Monoprotic, Amphoteric, and Polyprotic Benchmark Datasets with Self-Similarity Values and Similarity Values for the Combined Fine-Tuning Dataset for All of the Methods

evaluation set	number of pK_a values	number of molecules	self-similarity of the data set	combined fine-tune set similarity
monoprotic set	11,780	11,780	0.557	0.380
amphoteric set	1854	927	0.514	0.390
polyprotic set	4669	2025	0.553	0.380

difficult to quantify the relative frequency of experimentally undetected transitions compared to the intrinsic prediction error of Epik, our chosen reference predictor method. Therefore, we apply the more agnostic order-preserving matching method, as this enforces the least amount of constraint and provides the least bias on our end.

Combined Database Composition

The combined database consisted of 97,614 experimental pK_a values from 31,250 molecules after processing. The pK_a values that were assigned to the same charge state transition of the same molecule were averaged: we assigned these averaged pK_a values as the given molecule's pK_a value for that given charge state transition (values with experimental temperatures outside of the 15−40 °C temperature window were excluded from the average). To filter duplicate data points between different data sets, we excluded any numerically equivalent pK_a values from different data sets within the same charge state transition, as described in Methods. With this method, we arrived at 37,654 distinct pK_a values for the molecules in the combined data set (including pK_a values with excluded temperature ranges, which extends this set with only 44 new pK_a averages). The distributions of averaged pK_a values, average number of experimental pK_a's assigned to a charge state transition, molecular weights, and charge state transitions, as well as the similarity of the different subdata sets (t-SNE plot) are shown in Figure 3.

The combined pK_a database aggregates values from diverse literature sources, yet experimental method annotations are frequently absent or nonstandardized. Consequently, deviations observed among replicate pK_a values for the same molecule and charge-state transition may reflect a mixture of true measurement variability and intermethod differences. To assess this, we quantified two complementary dispersion measures for transitions with at least two experimental entries: (i) the standard deviation and (ii) the largest pairwise difference in pK_a

https://doi.org/10.1021/acs.jcim.6c00107
J. Chem. Inf. Model. 2026, 66, 4607−4619

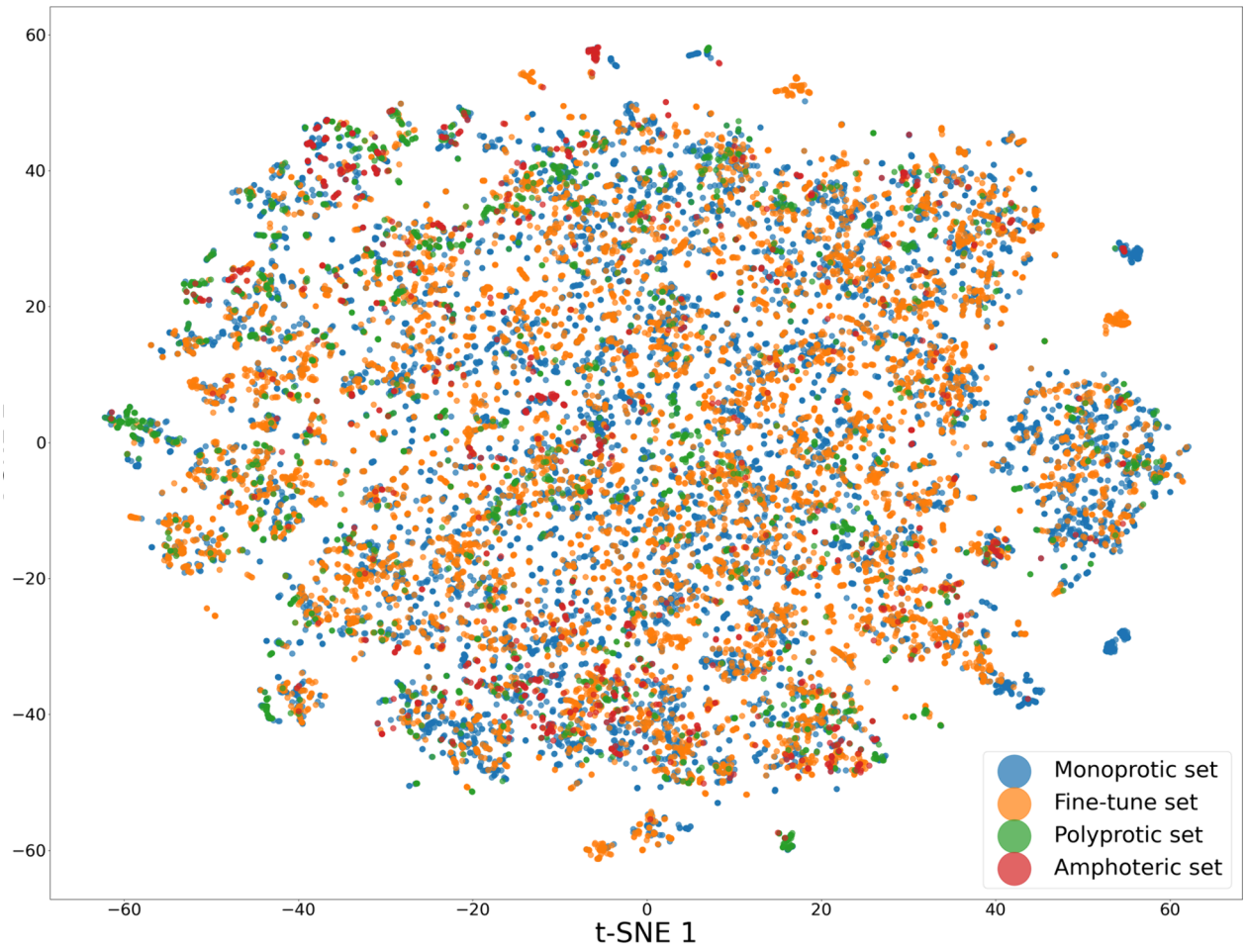


Figure 5. t-SNE plot of the monoprotic, amphoteric, and polyprotic data sets along with the combined fine-tuning data set used by the open-source predictors.

values. The range metric is particularly informative because it directly captures worst-case disagreement and remains interpretable, even for small replicate counts. Finally, to contextualize these results, we repeated the same analyses on the IUPAC digitized pK_a dataset, which includes explicit experimental method labels, allowing us to compare overall dispersion with method-stratified dispersion and thereby estimate how much of the observed spread is attributable to method heterogeneity versus within-method reproducibility. The distributions are found in Supplementary Figure S2. The mean standard deviation and largest pairwise difference are 0.20 and 0.38, respectively, with larger values quickly trailing off. These values are consistent with reported interexperiment deviations in other benchmarking studies,[38,44] and are reasonably robust within a range that is sufficiently narrower than the deviation of predicted values vs experimental values.

To assess how the different source data sets overlap in their molecule content, we checked the number of unique molecules that only appear in the given data set and calculated the fraction of these molecules compared to the size of the source data set. The full report on this can be found in Supplementary Table S1. The most unique source data sets (by fraction of data set-specific molecules) proved to be the SAMPL, Novartis, Caine, and Manchester data sets. This is consistent with the notion that SAMPL sets and the Novartis set are used as benchmarking sets, and we can see that the digitization of the Caine and Manchester data set provided many unique molecules, even if the overall size of the data sets is small. The least unique source data sets were the Datawarrior, Organic Oxygen Acids and Nitrogen Bases, and OCHEM data sets. Interestingly, the SAMPL7 data set also proved to contain quite few unique molecules (15%), considering that this data set was used as the base of a competition between different pK_a predictors.

We provide a characterization of the distribution of acidic and basic structural moieties based on SMARTS pattern matching and the elemental composition of the molecules in the database in Figure 4. The database has a high prevalence of amines, basic aromatic nitrogens, and carboxylic acids and contains more molecules with only basic titratable groups than molecules with only acidic groups, but molecules with both basic and acidic groups are the most numerous (numerical data are provided in Table S2 by specific titratable groups—some of these are shown under umbrella terms like "carbon acids" in Figure 4). These trends reflect the overall composition of chemical spaces that are investigated in drug discovery.

Evaluation of Commercial and Open-Source Predictors

For benchmarking recently developed prediction methods, the SAMPL and Novartis data sets have become widely used validation sets. The SAMPL data sets are particularly suitable because they were designed and carefully curated specifically to evaluate predictor performance. The Novartis data set originates from the work of Baltruschat et al.,[44] who obtained the measurements through an industry collaboration with Novartis.

https://doi.org/10.1021/acs.jcim.6c00107
J. Chem. Inf. Model. 2026, 66, 4607−4619

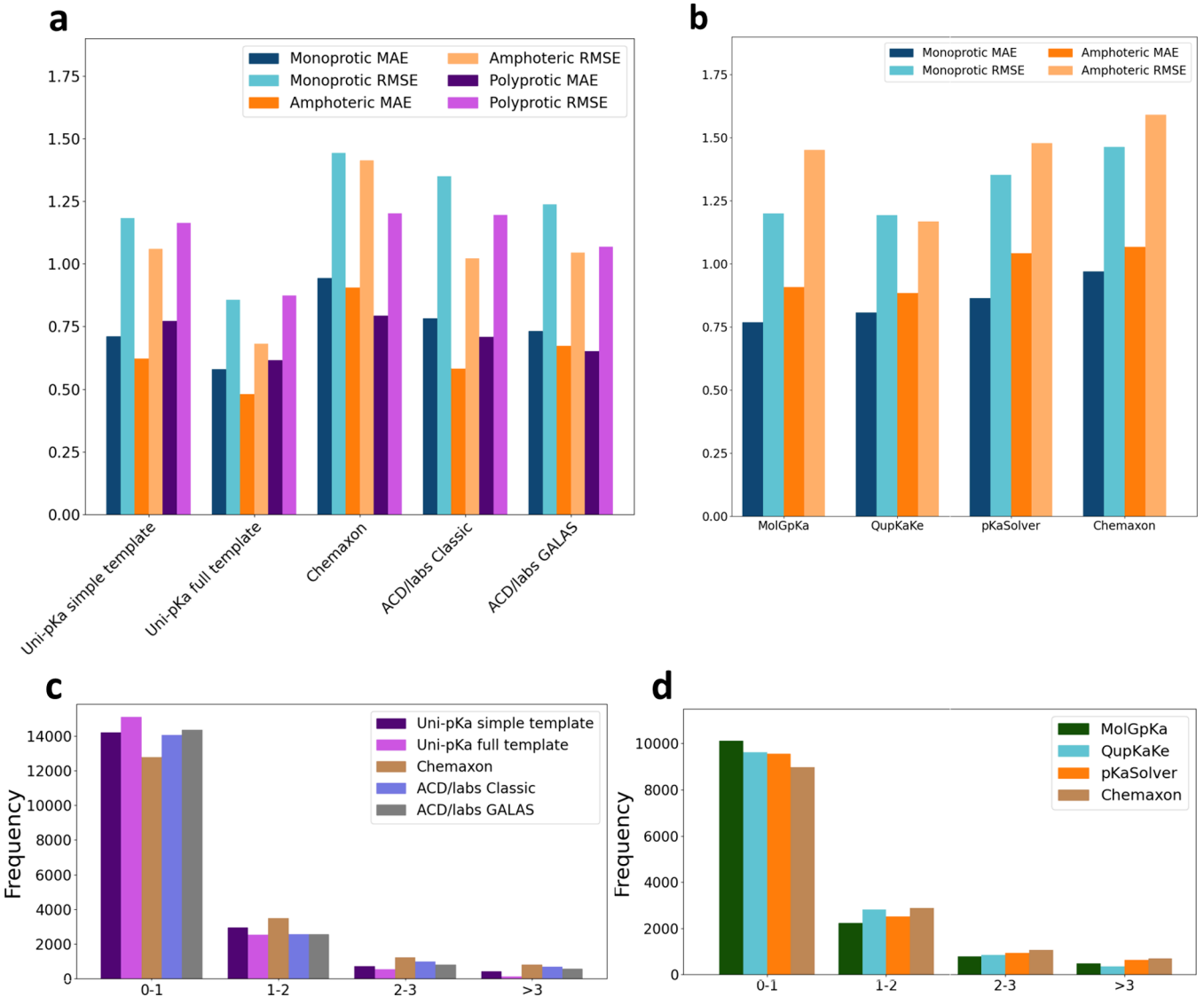


Figure 6. Mean absolute error (MAE) and root mean squared error (RMSE) distributions for the macro pK_a predictors on the monoprotic, amphoteric, and polyprotic benchmark sets (a) and for the micro pK_a predictors on the monoprotic and amphoteric benchmark sets (b). Distribution of the number of absolute prediction errors in the given bins (left closed, right open intervals) for the macro pK_a predictors (c) and the micro pK_a predictors (d).

We note that this data set is frequently misattributed to Liao et al.,[54] likely because the original article itself appears to include an incorrect citation to the mentioned paper, instead of Novartis when citing this data set. Important limitations of these benchmarks are their size (few hundred molecules) and their composition, which tends to avoid more complex, polyprotic molecules. As noted previously, this can restrict conclusions about performance on chemically challenging systems.[53] To evaluate open-source predictors against commercial reference methods under increasingly demanding conditions, we therefore compiled three benchmark data sets of increasing difficulty: a monoprotic data set containing molecules with a single reported charge-state transition, an amphoteric data set containing molecules with two transitions between the exact charge states of −1, 0, and +1, and a polyprotic data set consisting exclusively of molecules with a more complex set of charge state transitions. It is generally inadvisable to evaluate proprietary predictors on literature-derived data sets, because potential overlap between a model's training data and the validation set cannot be assessed and may bias the results. Accordingly, our goal is not to compare commercial products against one another, but to benchmark open-source methods against commercial references across multiple data sets, comprising thousands of pK_a values. This approach also reflects the fact that the training set composition and potential data leakage can be controlled only for the open-source predictors.

Information on the monoprotic, amphoteric, and polyprotic benchmark data sets is found in Table 3, along with similarity values to the combined set of molecules used for fine-tuning in any of the open-source models. Comparison of the three benchmark data sets to the combined fine-tuning data set on a structural basis reveals that they provide equivalent and balanced coverage of roughly the same chemical space (Figure 5). The self-similarity values of the benchmark data sets (defined as the average value of each molecule's highest similarity to any other molecule in the same data set) and fine-tuning set similarity values are comparable to the values we got for the SAMPL data sets reported in Supplementary Table S3. Distribution of pK_a

https://doi.org/10.1021/acs.jcim.6c00107
J. Chem. Inf. Model. 2026, 66, 4607−4619

Molecule ID: mol47
SMILES: O=C(O)c1cn(C2CC2)c2cc(N3CCNCC3)c(F)cc2c1=O
InChI: InChI=1S/C17H18FN3O3/c18-13-7-11-14(8-15(13)20-5-3-19-4-6-20)21(10-1-2-10)9-12(16(11)22)17(23)24/h7-10,19H,1-6H2,(H,23,24)

Experimental Macro pKa Values

Download: TSV JSON SMILES SDF

Macro pKa value	Dataset	Assigned charge state transition
6.16	Settimo	1 » 0
6.25	Settimo	1 » 0
8.62	Settimo	0 » -1

Download Microspecies: MICRO SMILES MICRO SDF

Charge States and Microspecies Visualization

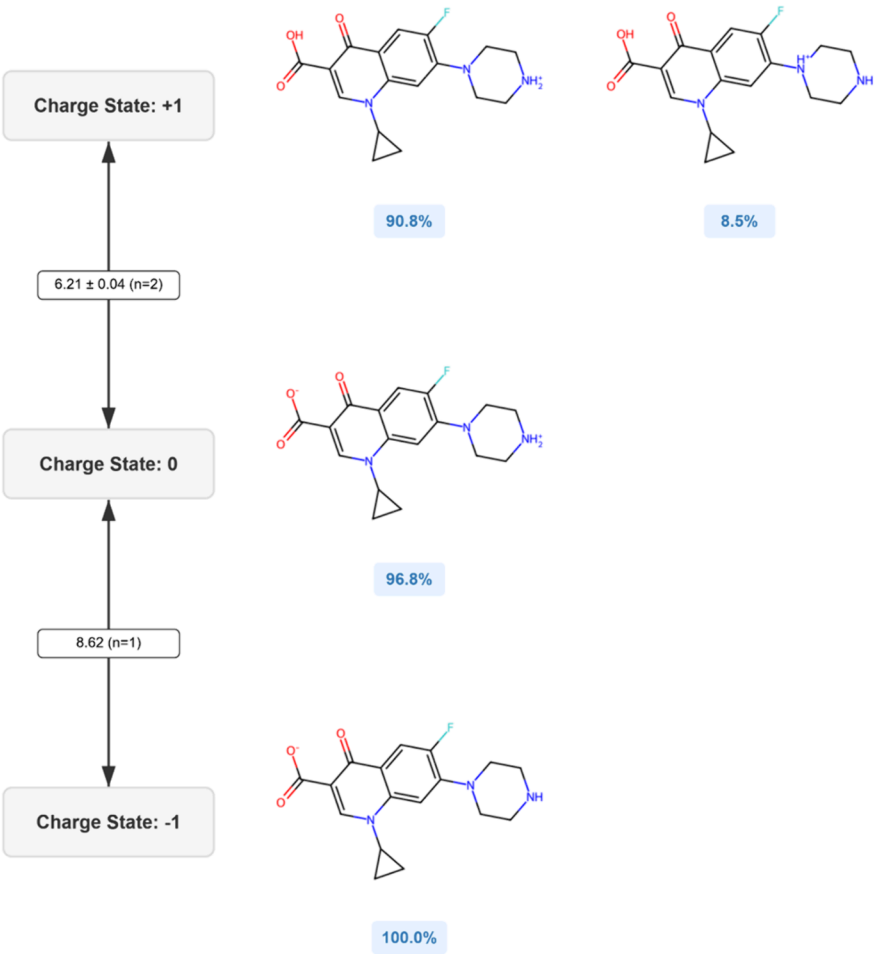


Figure 7. For a given molecule, the user can examine and download the microspecies distribution along with the dissociation constants annotated to macroscopic charge state transitions.

values, molecular weights, and others is provided in Supplementary Figure S1.

The micro-pK_a predictors were tested against the monoprotic and the amphoteric benchmark data sets. For the monoprotic set, as there was only one pK_a value to assign per molecule, we matched the predicted value, which was the closest to the experimental value. For the amphoteric sets, we used a simple linear assignment algorithm to assign those two predictions which provided the lowest sum of absolute deviations. Graphical illustrations of the results along with the distribution of the absolute errors are provided in Figure 6, and the results are also presented in tabular format in Supplementary Table S4.

The macro-pK_a predictors were tested against all three benchmark data sets, including the polyprotic set. The matching

https://doi.org/10.1021/acs.jcim.6c00107
J. Chem. Inf. Model. 2026, 66, 4607−4619

algorithm was the order-preserving, error-sum-minimizing matching that was also used for charge state transition assignment during database processing. Results are presented in Figure 6, and are available in tabular format in Supplementary Table S5. (The performance of Epik could be positively biased, as this algorithm was used in the preparation of the whole database: we therefore leave it out of the head-to-head comparison here but include its performance metrics for reference in Supplementary Table S6.)

To establish statistical significance between the distribution of prediction errors of different macro and micro-pK_a predictors, we performed one-way ANOVA, followed by a majority vote of five independent post hoc statistical tests (Supplementary Figure S3). At a significance level of $p = 0.05$, the distribution of the prediction errors of most micro-pK_a predictors (all except QupKake and MolGpKa) and also most macro-pK_a predictors (except Uni-pK_a simple vs ACD/Labs Classic and GALAS) are significantly different.

We can see that open-source predictors have comparable and sometimes significantly better performance than some of the commercial ones by the MAE and RMSE metrics. In general, macro-pK_a predictors usually outperform micro-pK_a predictors: this can be due to the more consistent pK_a description of these models with the nature of experimental results. It is also interesting that QupKake significantly outperforms the Chemaxon predictions, as this model was pretrained on Chemaxon-calculated pK_a values for molecules in ChEMBL. The simple template enumeration of Uni-pK_a returns comparable (and, in two cases, significantly better) performances compared to commercial predictors. Altogether, in terms of numerical accuracy, open-source methods, and especially Uni-pK_a, have caught up with commercial products. Reassuringly, the mean absolute errors rarely exceed one pK_a unit for any of the predictors on any of the data sets. By contrast, commercial providers have a clear edge in exception rates (percentage of compounds that cannot be processed): over the 31,250 compounds of the full database, typically less than 1% threw exceptions on commercial software, while all open-source predictors have shown >1% exception rates, going as high as 6.6% (Uni-pK_a with the simple template) and 7.4% (pKaSolver).

For drug design purposes, prediction errors within 1 pK_a unit are typically acceptable, although they can be more problematic as the pK_a value becomes closer to the physiological pH of 7.4. Here, a larger prediction error could translate to a mischaracterization of the most relevant microspecies (a standard step in ligand preparation), which could propagate to downstream molecular modeling tasks such as solubility or permeability prediction, or docking/virtual screening (via a failed interpretation of protein−ligand interactions). For the latter, typical examples are GPCR ligands, where a protonated amine is usually required to establish a key salt bridge with the $D^{3.32}$ residue, with the pK_a values of some typical groups being close to 7.4, e.g. 7.9 for the piperazine group of cariprazine, a D2/D3 receptor agonist.[55,56] (In a broader sense, the specific protonation states at physiologically relevant conditions affect several aspects of drug action, like efficacy, pharmacokinetics, all the way to in vivo pharmacology, and safety.) Larger errors for micro-pK_a predictors can mostly be attributed to molecules with multiple titratable groups (causing problems due to the incomplete description of the ionization processes by micro-pK_a predictors). For macro-pK_a predictors, large errors can be attributed to exotic structures such as strongly lipophilic molecules, macrocycles, conjugated systems, and unusual ionization centers. Some examples of strong outliers (absolute error >3 for at least two predictors) for the macro-pK_a predictors are provided in Supplementary Figure S4.

pKaHub: Online Database for the Combined, Annotated Data Set

We uploaded the combined database to a user-friendly web server, accessible at http://pkahub.ttk.hu, with all source code shared on GitHub (https://github.com/keserulab/pkahub). The server implements suitable Django data models for macroscopic charge states and transitions, as well as molecules and microspecies, and stores them along with the curated experimental data, which can be accessed through a graphical interface (as illustrated in Figure 7) or downloaded in tabular, JSON, SMILES, or SDF formats.

CONCLUSION

We compiled a large collection of experimental pK_a values from the literature and various public resources spanning over 31,000 molecules. The data were carefully curated and filtered: most importantly, this entailed the assignment of experimental pK_a values to macroscopic charge state transitions through a nonrestrictive, order-preserving matching algorithm. This large corpus of data allowed the thorough testing of four commercial and four open-source predictor models against three data sets with increasingly complex molecules, each with over a thousand data points that were not overlapping with the training data of any of the predictors. The evaluation proved that open-source pK_a predictors are viable alternatives to commercial ones in terms of the numeric accuracy of the results.

The full curated database is made available for public use on the pKaHub web server (http://pkahub.ttk.hu). The database is annotated according to the modern microequilibria theory of ionization states.

ASSOCIATED CONTENT

Data Availability Statement

Molecular structures and experimental pK_a data, collected from a wide selection of public and literature sources and annotated with macroscopic charge states as described in this paper, are made available for viewing one-by-one or downloading in bulk in multiple formats at the pKaHub web server (http://pkahub.ttk.hu), developed as part of this work. The source code for the web server is provided via GitHub (https://github.com/keserulab/pkahub).

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.6c00107.

Molecule counts over source data sets (Table S1), titratable group definitions and occurrences (Table S2), similarity values (Table S3), property distributions (Figure S1), distribution of experimental errors (Figure S2), performance metrics (Tables S4−S6), ANOVA results (Figure S3), and prediction outliers (Figure S4) (PDF)

AUTHOR INFORMATION

Corresponding Author

Dávid Bajusz − *Medicinal Chemistry Research Group and Drug Innovation Centre, HUN-REN Research Centre for Natural Sciences, 1117 Budapest, Hungary;* orcid.org/0000-0003-4277-9481; Email: bajusz.david@ttk.hu

https://doi.org/10.1021/acs.jcim.6c00107
J. Chem. Inf. Model. 2026, 66, 4607−4619

Authors

Levente Sipos-Szabó − *Medicinal Chemistry Research Group and Drug Innovation Centre, HUN-REN Research Centre for Natural Sciences, 1117 Budapest, Hungary; Department of Organic Chemistry and Technology, Faculty of Chemical Technology and Biotechnology, Budapest University of Technology and Economics, H-1111 Budapest, Hungary;* orcid.org/0009-0002-1210-7940

György T. Balogh − *Department of Pharmaceutical Chemistry, Semmelweis University, 1092 Budapest, Hungary; Center for Pharmacology and Drug Research & Development, Semmelweis University, 1085 Budapest, Hungary; Department of Chemical and Environmental Process Engineering, Faculty of Chemical Technology and Biotechnology, Budapest University of Technology and Economics, H-1111 Budapest, Hungary;* orcid.org/0000-0001-8273-1760

György M. Keserű − *Medicinal Chemistry Research Group and Drug Innovation Centre, HUN-REN Research Centre for Natural Sciences, 1117 Budapest, Hungary; Department of Organic Chemistry and Technology, Faculty of Chemical Technology and Biotechnology, Budapest University of Technology and Economics, H-1111 Budapest, Hungary;* orcid.org/0000-0003-1039-7809

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.6c00107

Author Contributions

L.S.-S. and G.T.B. have performed computations. L.S.-S. and D.B. have collected data sets and developed the online database. D.B. and G.M.K. have conceptualized and directed the study. All authors have participated in writing the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Levente M. Mihalovits and Adrián Papp for their help with some of the calculations. This work was supported by the Servier-Beregi Scholarship. This work was supported by the National Research Development and Innovation Office of Hungary [contracts FK146063 to D.B., NKKP 152458 to G.T.B., NKKP 153477, 152137, and PharmaLab (RRF-2.3.1-21-2022-00015) to G.M.K.]. The work of D.B. was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. The authors acknowledge the HPC time and support of the Governmental Information Technology Development Agency, Hungary.

■ REFERENCES

(1) Manallack, D. T. The pK_a Distribution of Drugs: Application to Drug Discovery. *Perspect. Med. Chem.* **2007**, *1*, 1177391X0700100003.

(2) Harris, J.; Chipot, C.; Roux, B. How Is Membrane Permeation of Small Ionizable Molecules Affected by Protonation Kinetics? *J. Phys. Chem. B* **2024**, *128* (3), 795−811.

(3) Zhang, Y.-Q.; Guo, R.-R.; Chen, Y.-H.; Li, T.-C.; Du, W.-Z.; Xiang, R.-W.; Guan, J.-B.; Li, Y.-P.; Huang, Y.-Y.; Yu, Z.-Q.; Cai, Y.; Zhang, P.; Ling, G.-X. Ionizable Drug Delivery Systems for Efficient and Selective Gene Therapy. *Mil. Med. Res.* **2023**, *10* (1), 9.

(4) Isin, B.; Estiu, G.; Wiest, O.; Oltvai, Z. N. Identifying Ligand Binding Conformations of the B2-Adrenergic Receptor by Using Its Agonists as Computational Probes. *PLoS One* **2012**, *7* (12), No. e50186.

(5) Bochevarov, A. D.; Watson, M. A.; Greenwood, J. R.; Philipp, D. M. Multiconformation, Density Functional Theory-Based pKa Prediction in Application to Large, Flexible Organic Molecules with Diverse Functional Groups. *J. Chem. Theory Comput.* **2016**, *12* (12), 6001−6019.

(6) Pracht, P.; Wilcken, R.; Udvarhelyi, A.; Rodde, S.; Grimme, S. High Accuracy Quantum-Chemistry-Based Calculation and Blind Prediction of Macroscopic pKa Values in the Context of the SAMPL6 Challenge. *J. Comput. Aided Mol. Des.* **2018**, *32* (10), 1139−1149.

(7) Warnau, J.; Wichmann, K.; Reinisch, J. COSMO-RS Predictions of logP in the SAMPL7 Blind Challenge. *J. Comput. Aided Mol. Des.* **2021**, *35* (7), 813−818.

(8) Ganyecz, Á.; Kállay, M. Implementation and Optimization of the Embedded Cluster Reference Interaction Site Model with Atomic Charges. *J. Phys. Chem. A* **2022**, *126* (15), 2417−2429.

(9) Tielker, N.; Eberlein, L.; Güssregen, S.; Kast, S. M. The SAMPL6 Challenge on Predicting Aqueous pKa Values from EC-RISM Theory. *J. Comput. Aided Mol. Des.* **2018**, *32* (10), 1151−1163.

(10) Fraczkiewicz, R.; Lobell, M.; Göller, A. H.; Krenz, U.; Schoenneis, R.; Clark, R. D.; Hillisch, A. Best of Both Worlds: Combining Pharma Data and State of the Art Modeling Technology To Improve in Silico pKa Prediction. *J. Chem. Inf. Model.* **2015**, *55* (2), 389−397.

(11) Lee, A. C.; Yu, J.; Crippen, G. M. pKa Prediction of Monoprotic Small Molecules the SMARTS Way. *J. Chem. Inf. Model.* **2008**, *48* (10), 2042−2053.

(12) Tielker, N.; Lim, M.; Kibies, P.; Gretz, J.; Hein-Janke, B.; Chodun, C.; Mata, R. A.; Czodrowski, P.; Kast, S. M. The euroSAMPL1 pKa Blind Prediction and Reproducible Research Data Management Challenge. *Phys. Chem. Chem. Phys.* **2025**, *27* (36), 18855−18869.

(13) Johnston, R. C.; Yao, K.; Kaplan, Z.; Chelliah, M.; Leswing, K.; Seekins, S.; Watts, S.; Calkins, D.; Chief Elk, J.; Jerome, S. V.; Repasky, M. P.; Shelley, J. C. Epik: pKa and Protonation State Prediction through Machine Learning. *J. Chem. Theory Comput.* **2023**, *19* (8), 2380−2388.

(14) Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. Epik: A Software Program for pKaprediction and Protonation State Generation for Drug-like Molecules. *J. Comput. Aided Mol. Des.* **2007**, *21* (12), 681−691.

(15) Baikété, J.; Malloum, A.; Conradie, J. pKa Prediction for Small Molecules: An Overview of Experimental, Quantum, and Machine Learning-Based Approaches. *J. Comput. Aided Mol. Des.* **2026**, *40* (1), 5.

(16) Ropp, P. J.; Kaminsky, J. C.; Yablonski, S.; Durrant, J. D. Dimorphite-DL: An Open-Source Program for Enumerating the Ionization States of Drug-like Small Molecules. *J. Cheminformatics* **2019**, *11* (1), 14.

(17) Pan, X.; Wang, H.; Li, C.; Zhang, J. Z. H.; Ji, C. MolGpka: A Web Server for Small Molecule pKa Prediction Using a Graph-Convolutional Neural Network. *J. Chem. Inf. Model.* **2021**, *61* (7), 3159−3165.

(18) Yang, Y.; Yao, K.; Repasky, M. P.; Leswing, K.; Abel, R.; Shoichet, B. K.; Jerome, S. V. Efficient Exploration of Chemical Space with Docking and Deep Learning. *J. Chem. Theory Comput.* **2021**, *17* (11), 7106−7119.

(19) Rezaee, M.; Ekrami, S.; Hashemianzadeh, S. M. Comparing ANI-2x, ANI-1ccx Neural Networks, Force Field, and DFT Methods for Predicting Conformational Potential Energy of Organic Molecules. *Sci. Rep.* **2024**, *14* (1), 11791.

(20) Abarbanel, O. D.; Hutchison, G. R. QupKake: Integrating Machine Learning and Quantum Chemistry for Micro-pKa Predictions. *J. Chem. Theory Comput.* **2024**, *20* (15), 6946−6956.

(21) Luo, W.; Zhou, G.; Zhu, Z.; Yuan, Y.; Ke, G.; Wei, Z.; Gao, Z.; Zheng, H. Bridging Machine Learning and Thermodynamics for Accurate pKa Prediction. *JACS Au* **2024**, *4* (9), 3451−3465.

(22) Artrith, N.; Butler, K. T.; Coudert, F.-X.; Han, S.; Isayev, O.; Jain, A.; Walsh, A. Best Practices in Machine Learning for Chemistry. *Nat. Chem.* **2021**, *13* (6), 505−508.

(23) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40* (D1), D1100−D1107.

(24) *iBonD*. http://ibond.nankai.edu.cn/ (accessed Jan 06, 2026).

(25) An, H.; Liu, X.; Cai, W.; Shao, X. AttenGpKa: A Universal Predictor of Solvation Acidity Using Graph Neural Network and Molecular Topology. *J. Chem. Inf. Model.* **2024**, *64* (14), 5480−5491.
(26) Sander, T.; Freyss, J.; von Korff, M.; Rufener, C. DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis. *J. Chem. Inf. Model.* **2015**, *55* (2), 460−473.
(27) Mansouri, K.; Cariello, N. F.; Korotcov, A.; Tkachenko, V.; Grulke, C. M.; Sprankle, C. S.; Allen, D.; Casey, W. M.; Kleinstreuer, N. C.; Williams, A. J. Open-Source QSAR Models for pKa Prediction Using Multiple Machine Learning Approaches. *J. Cheminformatics* **2019**, *11* (1), 60.
(28) Zheng, J. *IUPAC/Dissociation-Constants: V1.0*, 2022. .
(29) Zheng, J. W.; Leito, I.; Green, W. H. Widespread Misinterpretation of pKa Terminology for Zwitterionic Compounds and Its Consequences. *J. Chem. Inf. Model.* **2024**, *64* (23), 8838−8847.
(30) Fraczkiewicz, R. What, This "Base" Is Not a Base? Common Misconceptions about Aqueous Ionization That May Hinder Drug Discovery and Development. *J. Med. Chem.* **2025**, *68* (20), 21003−21011.
(31) Mayr, F.; Wieder, M.; Wieder, O.; Langer, T. Improving Small Molecule pKa Prediction Using Transfer Learning With Graph Neural Networks. *Front. Chem.* **2022**, *10*, No. 866585.
(32) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15* (3), 1652−1671.
(33) Pracht, P.; Grimme, S.; Bannwarth, C.; Bohle, F.; Ehlert, S.; Feldmann, G.; Gorges, J.; Müller, M.; Neudecker, T.; Plett, C.; Spicher, S.; Steinbach, P.; Wesołowski, P. A.; Zeller, F. CREST—A Program for the Exploration of Low-Energy Molecular Chemical Space. *J. Chem. Phys.* **2024**, *160* (11), 114110.
(34) *Acid Dissociation Constant Calculator|pKa Prediction Software*; ACD/Labs. https://www.acdlabs.com/products/percepta-platform/physchem-suite/pka/ (accessed Jan 08, 2026).
(35) *Chemaxon*. https://chemaxon.com (accessed Jan 08, 2026).
(36) *Schrödinger - Physics-based Software Platform for Molecular Discovery & Design*; Schrödinger. https://www.schrodinger.com/ (accessed Jan 09, 2026).
(37) Filippov, I. V.; Nicklaus, M. C. Optical Structure Recognition Software To Recover Chemical Information: OSRA, An Open Source Solution. *J. Chem. Inf. Model.* **2009**, *49* (3), 740−743.
(38) Settimo, L.; Bellman, K.; Knegtel, R. M. A. Comparison of the Accuracy of Experimental and Predicted pKa Values of Basic and Acidic Compounds. *Pharm. Res.* **2014**, *31* (4), 1082−1095.
(39) Hunt, P.; Hosseini-Gerami, L.; Chrien, T.; Plante, J.; Ponting, D. J.; Segall, M. Predicting pKa Using a Combination of Semi-Empirical Quantum Mechanics and Radial Basis Function Methods. *J. Chem. Inf. Model.* **2020**, *60* (6), 2989−2997.
(40) Jensen, J. H.; Swain, C. J.; Olsen, L. Prediction of pKa Values for Druglike Molecules Using Semiempirical Quantum Chemical Methods. *J. Phys. Chem. A* **2017**, *121* (3), 699−707.
(41) Caine, B. A.; Bronzato, M.; Fraser, T.; Kidley, N.; Dardonville, C.; Popelier, P. L. A. Aqueous pKa Prediction for Tautomerizable Compounds Using Equilibrium Bond Lengths. *Commun. Chem.* **2020**, *3* (1), 21.
(42) Manchester, J.; Walkup, G.; Rivin, O.; You, Z. Evaluation of pKa Estimation Methods on 211 Druglike Compounds. *J. Chem. Inf. Model.* **2010**, *50* (4), 565−571.
(43) Yu, H.; Kühne, R.; Ebert, R.-U.; Schüürmann, G. Comparative Analysis of QSAR Models for Predicting pKa of Organic Oxygen Acids and Nitrogen Bases from Molecular Structure. *J. Chem. Inf. Model.* **2010**, *50* (11), 1949−1960.
(44) Baltruschat, M.; Czodrowski, P. Machine Learning Meets pK_a. *F1000Research* **2020**, *9*, 113.
(45) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V.; Tanchuk, V. Y.; Todeschini, R.; Varnek, A.; Marcou, G.; Ertl, P.; Potemkin, V.; Grishina, M.; Gasteiger, J.; Schwab, C.; Baskin, I. I.; Palyulin, V. A.; Radchenko, E. V.; Welsh, W. J.; Kholodovych, V.; Chekmarev, D.; Cherkasov, A.; Aires-de-Sousa, J.; Zhang, Q.-Y.; Bender, A.; Nigsch, F.; Patiny, L.; Williams, A.; Tkachenko, V.; Tetko, I. V. Online Chemical Modeling Environment (OCHEM): Web Platform for Data Storage, Model Development and Publishing of Chemical Information. *J. Comput. Aided Mol. Des.* **2011**, *25* (6), 533−554.
(46) *QSAR Toolbox*. https://qsartoolbox.org/ (accessed Jan 06, 2026).
(47) Işık, M.; Levorse, D.; Rustenburg, A. S.; Ndukwe, I. E.; Wang, H.; Wang, X.; Reibarkh, M.; Martin, G. E.; Makarov, A. A.; Mobley, D. L.; Rhodes, T.; Chodera, J. D. pKa Measurements for the SAMPL6 Prediction Challenge for a Set of Kinase Inhibitor-like Fragments. *J. Comput. Aided Mol. Des.* **2018**, *32* (10), 1117−1138.
(48) Bergazin, T. D.; Tielker, N.; Zhang, Y.; Mao, J.; Gunner, M. R.; Francisco, K.; Ballatore, C.; Kast, S. M.; Mobley, D. L. Evaluation of Log P, pKa, and Log D Predictions from the SAMPL7 Blind Challenge. *J. Comput. Aided Mol. Des.* **2021**, *35* (7), 771−802.
(49) Bahr, M. N.; Nandkeolyar, A.; Kenna, J. K.; Nevins, N.; Da Vià, L.; Işık, M.; Chodera, J. D.; Mobley, D. L. Automated High Throughput pKa and Distribution Coefficient Measurements of Pharmaceutical Compounds for the SAMPL8 Blind Prediction Challenge. *J. Comput. Aided Mol. Des.* **2021**, *35* (11), 1141−1155.
(50) Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J. Cheminformatics* **2015**, *7* (1), 20.
(51) Işık, M.; Rustenburg, A. S.; Rizzi, A.; Gunner, M. R.; Mobley, D. L.; Chodera, J. D. Overview of the SAMPL6 pKa Challenge: Evaluating Small Molecule Microscopic and Macroscopic pKa Predictions. *J. Comput. Aided Mol. Des.* **2021**, *35* (2), 131−166.
(52) Fraczkiewicz, R. In Silico Prediction of Ionization. *In Comprehensive Medicinal Chemistry II* **2007**, *5*, 603−626.
(53) Fraczkiewicz, R. Thermodynamics-Informed Neural Networks and Extensive Data Sets: Key Factors to Accurate Blind Predictions of Apparent pKa Values in the euroSAMPL Challenge. *Phys. Chem. Chem. Phys.* **2025**, *27* (16), 8039−8042.
(54) Liao, C.; Nicklaus, M. C. Comparison of Nine Programs Predicting pKa Values of Pharmaceutical Substances. *J. Chem. Inf. Model.* **2009**, *49* (12), 2801−2812.
(55) Chen, Z.; Fan, L.; Wang, H.; Yu, J.; Lu, D.; Qi, J.; Nie, F.; Luo, Z.; Liu, Z.; Cheng, J.; Wang, S. Structure-Based Design of a Novel Third-Generation Antipsychotic Drug Lead with Potential Antidepressant Properties. *Nat. Neurosci.* **2022**, *25* (1), 39−49.
(56) Ágai-Csongor, É.; Domány, G.; Nógrádi, K.; Galambos, J.; Vágó, I.; Keserű, G. M.; Greiner, I.; Laszlovszky, I.; Gere, A.; Schmidt, É.; Kiss, B.; Vastag, M.; Tihanyi, K.; Sághy, K.; Laszy, J.; Gyertyán, I.; Zájer-Balázs, M.; Gémesi, L.; Kapás, M.; Szombathelyi, Z. Discovery of Cariprazine (RGH-188): A Novel Antipsychotic Acting on Dopamine D3/D2 Receptors. *Bioorg. Med. Chem. Lett.* **2012**, *22* (10), 3437−3440.

https://doi.org/10.1021/acs.jcim.6c00107
J. Chem. Inf. Model. 2026, 66, 4607−4619