Widespread misinterpretation of pK_a terminology and its consequences

Jonathan W. Zheng,[†] Ivo Leito,[‡] and William H. Green[*,†]

[†]*Massachusetts Institute of Technology, Cambridge, MA, 02139, U.S.A.*

[‡]*University of Tartu, Ravila 14A, Tartu 50411, Estonia*

E-mail: whgreen@mit.edu

Abstract

The acid dissociation constant (pK_a), which quantifies the propensity for a solute to donate a proton to its solvent, is crucial for drug design and synthesis, environmental fate studies, chemical manufacturing, and many other fields. Unfortunately, the terminology used for describing acid-base phenomena is inconsistent, causing large potential for misinterpretation. In this work, we examine a systematic confusion underlying the definition of "acidic" and "basic" pK_a values for zwitterionic compounds. Due to this confusion, some pK_a data is misrepresented in data repositories, including the widely-used and highly trusted ChEMBL Database. Such datasets are widely used to supply training data for pK_a prediction models, and hence, confusion and errors in the data makes model performance worse. Herein, we discuss the intricacies of this issue. We make suggestions for describing acid-base phenomena, training pK_a prediction models, and stewarding pK_a datasets, given the high potential for confusion and potentially high impact of accurately describing acid-base phenomena.

Introduction

The acid dissociation constant, or pK_a, significantly affects the behavior of compounds and is thus very important in pharmacokinetics, environmental chemistry, chemical manufacturing, and numerous other applications.[1–3] It is used extensively in drug design to calculate ionizability, solubility, and partitioning between aqueous and organic phases (e.g. lipophilicity), which are key heuristics in ADME screening.[4,5] The pK_a is defined as the equilibrium constant of the hydrogen dissociation reaction for acid AH in arbitrary solvent,

$$\mathrm{AH} + \mathrm{S} \rightleftharpoons \mathrm{SH^+} + \mathrm{A^-} \tag{1}$$

where AH refers to a Brønsted acid, S is the solvent, SH^+ is the protonated form of the solvent, and A^- is the conjugate base. The equilibrium constant corresponding to this reaction is

$$K_a = \frac{a_{\mathrm{A^-}} \cdot a_{\mathrm{SH^+}}}{a_{\mathrm{AH}}} \tag{2}$$

and the pK_a is defined as

$$pK_a = -\log_{10}(K_a) \tag{3}$$

Although pK_a values are termed "constants", they correspond to a given temperature, ionic strength, and solvent.[6] In this work, we will assume that all values are measured at 25 °C, extrapolated to 0 molar ionic strength, in water.

Under these thermodynamic definitions, pK_a strictly refers to the strength of the *acidic* dissociation of the acid AH. It is widely accepted and understood that a dissociation involving proton *loss* should be referred to as a pK_a of the compound.

In contrast, terminology for describing proton *gain* (i.e., basicity) is significantly less clear. It is common for chemists to describe the "pK_a of a base" to refer to its *basicity*, when

in reality they are speaking of the base's conjugate acid. This term is also sometimes called a *basic* $\mathrm{p}K_\mathrm{a}$, and sometimes referred as $\mathrm{p}K_\mathrm{aH}$ or $\mathrm{p}K_\mathrm{BH+}$, corresponding to the $\mathrm{p}K_\mathrm{a}$ of the conjugate base of A:

$$\mathrm{AH^+} \rightleftharpoons \mathrm{A} + \mathrm{H^+} \tag{4}$$

Yet another term is $\mathrm{p}K_\mathrm{b}$, which represents an activity scale in hydroxide (or deprotonated form of solvent, in non-aqueous systems). This is generally less popular because it requires another step of converting activity scales before it can be directly compared with $\mathrm{p}K_\mathrm{a}$ values. Furthermore, its practical usage requires that the deprotonated solvent is stable, and that the autoprotolysis constant of the solvent is known, which for many solvents is not fulfilled.

Despite this potential for confusion, the literature has not yet converged on terminology that universally describes the acid dissociation constant of conjugate acids. Each of these terms is commonly encountered in the literature. This in itself already can be problematic.

Another topic that must be clarified is the difference between "macroscopic" versus "microscopic" $\mathrm{p}K_\mathrm{a}$ values. A *macroscopic* $\mathrm{p}K_\mathrm{a}$ corresponds to the observed equilibrium between different *protonation states*, each state corresponding to an ensemble of microstates, whereas a *microscopic* $\mathrm{p}K_\mathrm{a}$ corresponds to the equilibrium between two specific isomers at different charge states.[7,8] For monoprotic compounds with only one dominant tautomer at each charge state, these two values are the same. For polyprotic compounds that have few dissociation sites, with values far different from each other, and also only one dominant tautomer at each charge state, both types of $\mathrm{p}K_\mathrm{a}$ are also approximately the same. However, in all other cases, these values are different.

Whereas most experimental efforts (and thereby most experimental data) are focused on macroscopic $\mathrm{p}K_\mathrm{a}$ values, theoretical efforts are often interested in predicting microscopic ones. Owing to the ease of synthesizing, testing, and analyzing, and simulating compounds with few acidity centers, most effort has also focused on simpler compounds, e.g., monoprotic

or with few acidity centers. However, in pharmaceutical discovery and design, which is among the largest use cases for pK_a values (both experimental and predicted), many (if not most) molecules of interest are substantially more complex. Drug molecules tend to have multiple dissociation sites, form zwitterions in water, and sometimes tautomerize. Failing to consider the difference between "macroscopic" and "microscopic" pK_a can lead to markedly incorrect predictions.

But there is an even more insidious terminological issue that has led to a systematic misinterpretation of pK_a data. When reporting the pK_a of any compound, three fundamental pieces of information are crucial:

1. What is the pK_a value?
2. What is the species that loses the proton, and what is the species that remains after proton loss?
3. If the above is not available, then what is the overall, macroscopic charge transition that the pK_a value corresponds to, and what is the identity of the uncharged species?

We have observed significant confusion in the literature regarding points 2 and 3, especially for amino acids and other ampholytes that form zwitterions in solution. This confusion arises from an unfortunate development of terminology that will be discussed in this work. This terminology has appeared in databases including ChEMBL, a widely-trusted repository for biochemical data. This situation has caused confusion in the development of pK_a models trained on such data, contaminating model performance and thereby affecting point 1. In addition, due to the limited assessment of model performance on relevant benchmarks, such contamination has largely evaded discussion in the literature.

The acid dissociation constant is used in drug design as a metric for ADME properties, blood-brain barrier penetration,[9] protein binding,[10] solubility,[11] purification processes,[12] and others; hence, these errors in training have the potential to affect medicine and thereby the health and welfare of people. It also affects kinetic simulations, whether from being used to

compute ratios of rate coefficients from the equilibrium constant, or to compute solvation free energies of ionic reactants[13,14] which are then combined with calculations of transition state thermodynamics to compute liquid-phase rate coefficients.[15] In the following sections, we will explain the origin and depth of the problem, examine potential downstream effects on predicting ADME properties, and discuss our perspective on how to resolve this issue for future research efforts.

The acidity label issue

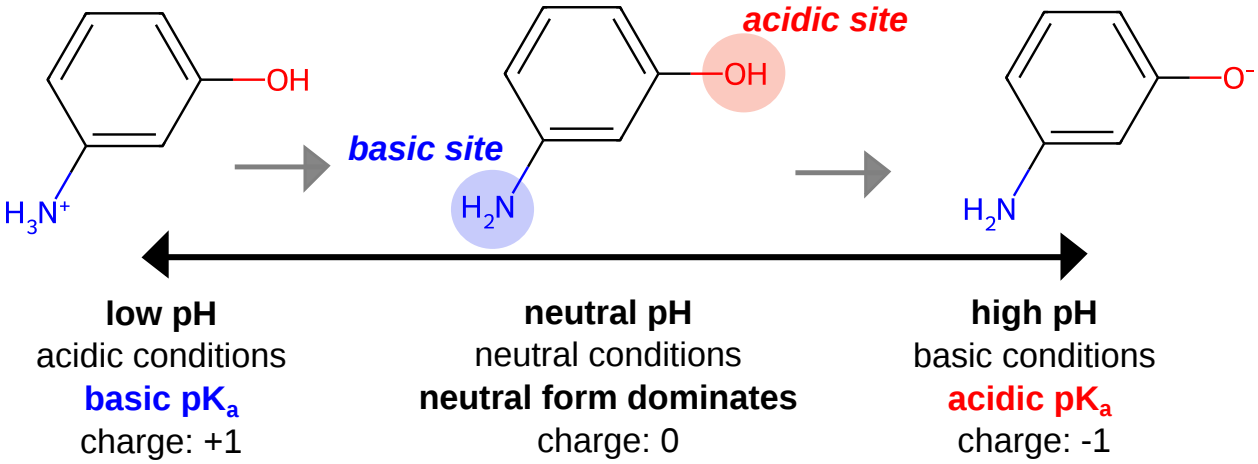

Figure 1: pH scale for m-aminophenol.

An ampholyte is a molecule that can either act as an acid or base. For a given microstate that has one acidic and one basic site, the lower pK_a corresponds always to proton gain and the higher value to a proton loss. At a low pH, the activity of protons is high; therefore, the chemical equilibrium favors the reaction of $RH + H^+$ to form RH_2^+, and the predominant charge state will be +1. In this sense, RH acts a Brønsted-Lowry base. Similarly, at a high pH, the activity of protons is low, so the reaction forming protons (i.e. the acid dissociation of RH) is favored.

The key issue is as follows: For most molecules, the low and high macroscopic p$K_a$values are termed "basic" and "acidic" pK_a values, respectively. However, for compounds that tautomerize to zwitterions (such as amino acids), the order is flipped, and the pK_a labels

are generally "acidic" and "basic", respectively.

This is at first quite counterintuitive. From a *macroscopic* pK_a perspective, the low pK_a corresponds to proton gain and the high pK_a corresponds to proton loss. These are usual definitions of Brønsted acidity. For example, Figure 1 depicts the amphiprotic nature of m-aminophenol and the major microspecies at different pH regimes, as well as the corresponding pK_a types, corresponding to the usual ordering of basicity and acidity.[16]

But for zwitterion-forming compounds, the situation is different. Let us examine glycine (Figure 2). The macroscopic pK_a values of glycine are commonly reported to be approximately 2.4 and 9.8.

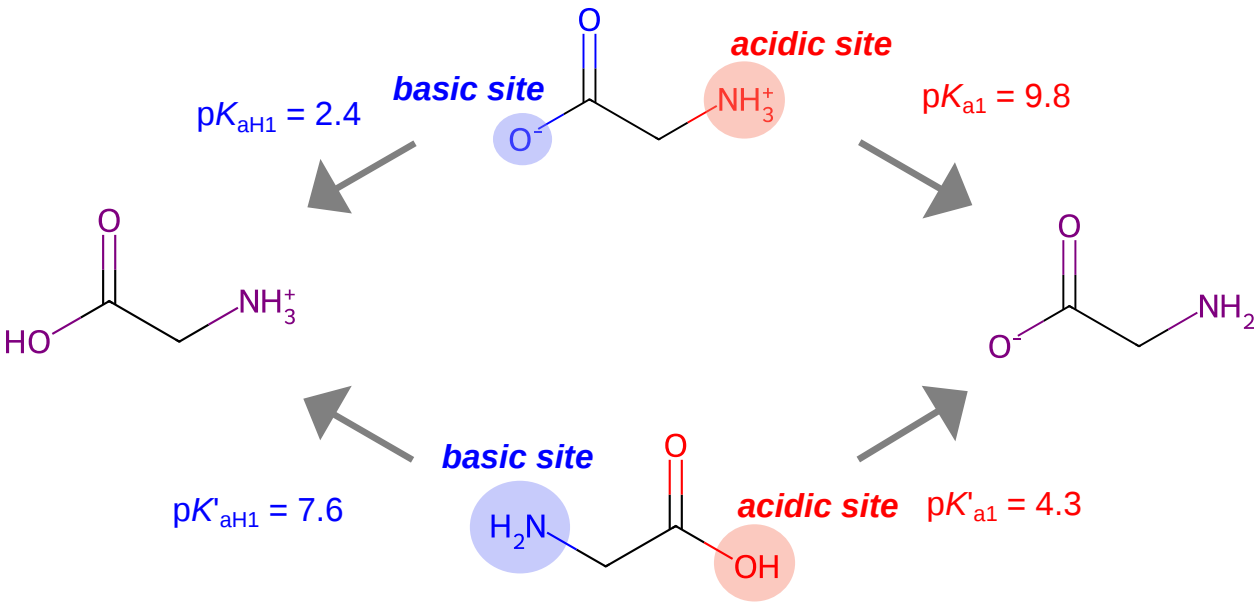

Figure 2: Microscopic pK_a values of glycine in both the uncharged and zwitterionic tautomers.

In contrast to compounds with only an alcohol and amine site such as m-aminophenol, glycine in aqueous solution exists predominantly as a zwitterion. In fact, its tautomerization is dictated by its uncharged form's microscopic pK_a values: pK_a(-OH) = 4.3 (acidic) and pK_{aH}(-amino) = 7.6 (basic), respectively. From a thermodynamic perspective, it is *because* the pK_a of the -COOH group in the non-zwitterionic neutral form is lower than that of the -NH_3^+ group in the zwitterionic neutral form that the zwitterion can exist; and the relative

population of each tautomer is dictated by the ratio of the corresponding equilibria.[17] In water, the zwitterionic form of glycine predominates over the non-zwitterionic neutral form at all pH, representing more than 99.9% of microspecies for the neutral protonation state. The zwitterionic tautomer has stronger acidic and basic sites, with microscopic pK_a values of pK_{aH} = 2.4 (basic, proton gain) and pK_a = 9.8 (acidic, proton loss).

Hence, the numerical values observed in the macroscopic pK_a (2.4 and 9.8) correspond to those of the dominant zwitterion microstate. But confusingly, the macroscopic pK_a at 2.4 is called *acidic* and the pK_a at 9.8 is called *basic*, because such is the case for the uncharged microstate, even though in this case the zwitterion dominates the microstate population. (We note that it is not always true that the zwitterion tautomer dominates - but it is often a significant microstate.)

Hence, we arrive at this unfortunate terminology, in which the macroscopic acidity label represents its *uncharged* tautomer's microscopic pK_a, but the value represents the zwitterion. The set of information is inconsistent with either microstate, and is inverted in how "acidity" and "basicity" are defined for other amphiprotic molecules like alkanolamines. An example of this convention can be seen as early as 1929, shortly after the zwitterion was introduced as a concept in the early 1920s.[18–20] Later, in 1984, Harris and Serjeant suggest avoiding the "basic" and "acidic" pK_a labels unless one is sure about the presence of zwitterions, instead naming the lower value "proton gain" and the higher value "proton loss".[16]

This terminology of "basic" and "acidic" persists today, with many sources, including PubChem, ChEMBL, and others labeling that the numerically lower pK_a is acidic for many amino acids. Exceptions infrequently occur in the literature - several publications describe the carbonyl group's low pK_a as basic and the aminium group as acidic.[21–23] In the following sections, we examine how this confusing terminology has led to problems in cheminformatics, which has impacted the performance of predictive models.

Inconsistency in the ChEMBL Database

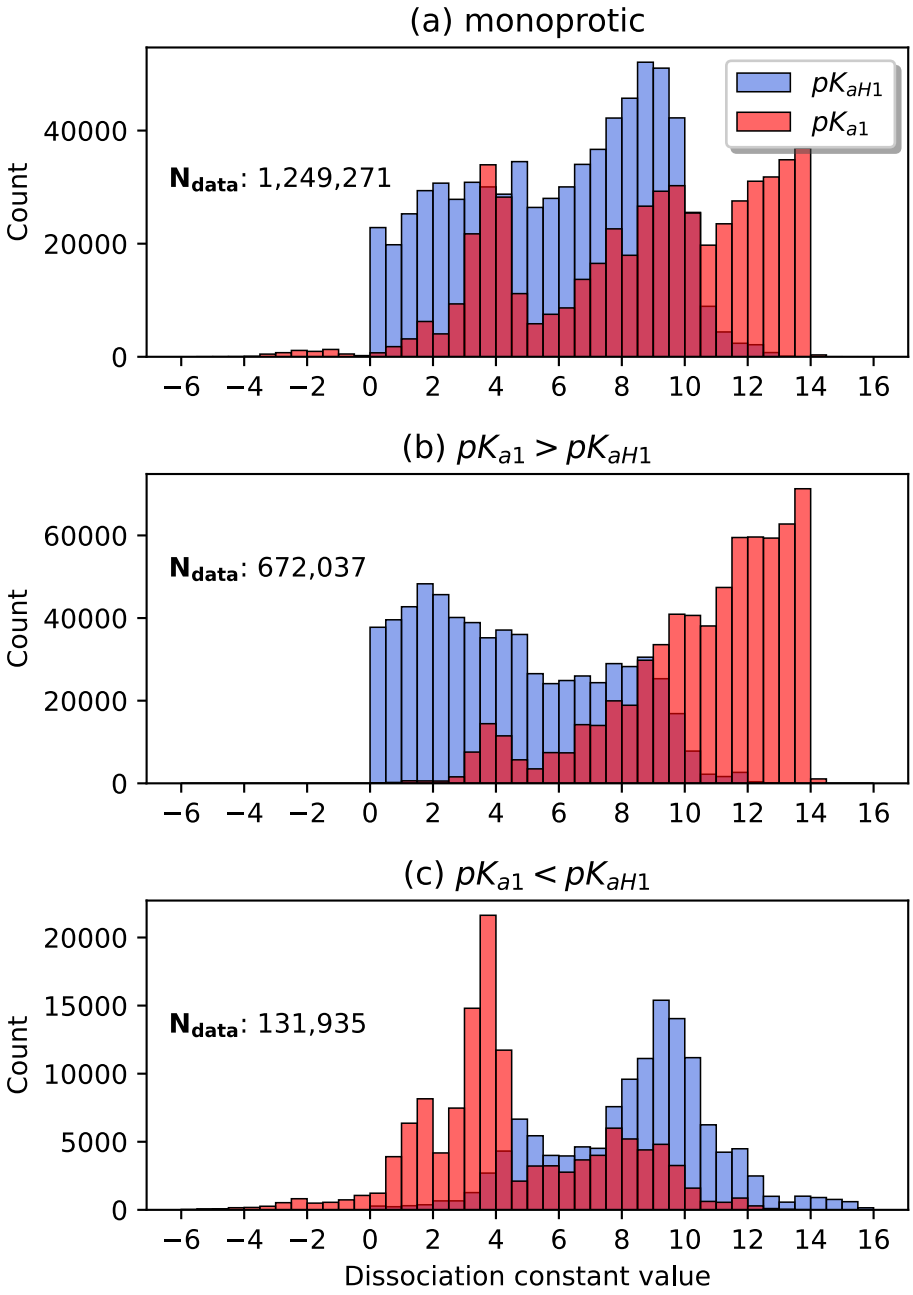

Figure 3: Histograms comparing the distribution of pK_{a} values for polyprotic ChEMBL compounds; a significant portion have a listed acidic pK_{a} less than the basic pK_{a}.

ChEMBL is a manually curated database that includes 2.4+ million distinct compounds as of version 34.[7,8] The database includes calculations for the "most acidic" and "most basic"

sites for each compound. Owing to the low availability of high-fidelity experimental pK_a data, the ChEMBL database has recently been used as a pretraining dataset for many recent data-driven pK_a models. The pK_a values in ChEMBL are from ChemAxon's Protonation pK_a calculator, and are consistent with predictions in the software's macroscopic static pK_a mode. The "static" mode indicates that charged forms of input molecules are converted to neutral forms before calculation.[24] We emphasize these are *macroscopic*, despite many recent efforts to treat them as *microscopic*.

Unfortunately, another issue manifests in the ChEMBL data due to the terminology. ChEMBL reports only the numerically lowest acidic pK_a and highest basic pK_a predicted for each compound. For non-zwitterion forming compounds, this would correspond to what might be best understood as the "most acidic" and "most basic" pK_a values. But for these flipped-label compounds, this policy instead samples the *least* basic and *least* acidic macroscopic values within the examined pK_a range. For models that sample the ChEMBL dataset, it is commonly assumed that these values correspond to the +1 and -1 charge states, even though this is not the case in these circumstances.[25–29] This can lead to corruption of the training data and poor inference performance for such compounds.

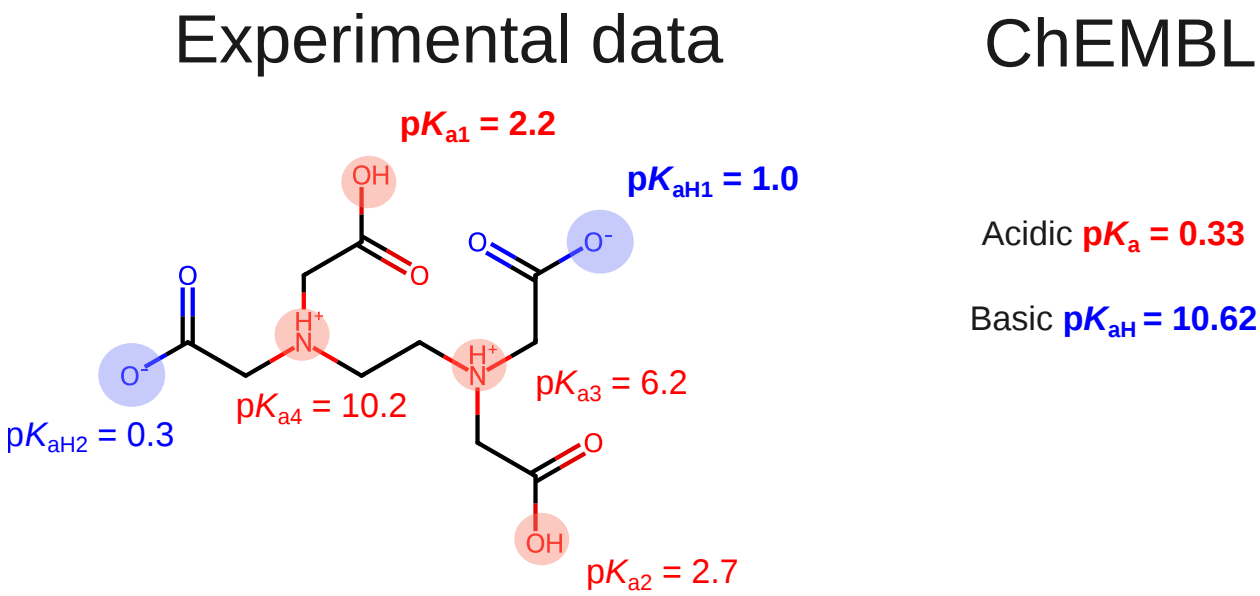

Figure 4: Dissociations for one of the zwitterionic forms of EDTA. Because the acidity labels are flipped and then the most extreme values are taken, the ChEMBL data is reported the pK_a values corresponding to dissociations for the +2 and -3 charge states, rather than the +1 and +0.

To illustrate this, let us examine ethylenediaminetetraacetic acid (EDTA), a compound commonly used in applications including manufacturing,[30] environmental remediation,[31] dentistry,[32,33] medicine,[34] and others. The uncharged tautomer of EDTA has four carboxylic acid groups and two tertiary amine groups. In water, as shown in Figure 4, EDTA can tautomerize. It has six pK_a values: 0.3, 1.0, 2.2, 2.7, 6.2, and 10.2. Four correspond to carboxylic acid groups, and two to the aminium groups.[35] Two of the four carboxylic acid pK_a values correspond to proton gain. The most basic and most acidic pK_a values relative to the neutral compound (i.e., those corresponding to charge state +1 to 0 and then from +0 to -1) are 1.0 and 2.2, respectively. But ChEMBL reports 10.62 and 0.33, which instead represent charge transitions from -3 to -4 and from +2 to +1, respectively. Because ChEMBL's overall ranking of acidities and basicities has flipped, it is reporting the *least* acidic and *least* basic pK_a values among the heteroatom functional sites.

As of version 34, in ChEMBL, the acidic pK_a values are lower than the basic pK_a values for 131,935 entries (see Figure 3). In total, ChEMBL includes 2,053,423 compounds with

calculated pK_a values, among which 803,972 are polyprotic (i.e., an acidic and basic pK_a are reported). Hence, about 5% of all acidic/basic entries, or 16% of all polyprotic compounds, have potential for error. These include values for medical compounds such as levothyroxine and ampicillin, biochemicals such as phosphocreatine and ATP, dye molecules such as Methyl Red, and practically all amino acids found in living organisms. As these compounds are important in biochemical applications, there is significant potential for errors in pK_a data and modeling to affect human life. We therefore suggest that these calculations are carefully curated by ChEMBL. The solution need not be complicated: if the acidic pK_a values are lower than the basic ones, then choose the *highest* acidic and *lowest* basic values.

The largest currently existing set of pK_a data with full metadata (including temperatures, methods, original references, and critical evlauation) is the IUPAC Digitized Aqueous pK_a Dataset,[36] which is a digitized version of several compilations of pK_a data.[37–39] Most of the values in this dataset are macroscopic. We found a set of compounds in ChEMBL that have acidity values lower than their basic values; from these, we downsampled to those that also have at least an acidic or basic pK_a value in the IUPAC dataset. This resulted in a set of 171 proton gain and 195 proton loss pK_a values. The parity is shown between the ChEMBL dataset and the IUPAC data (Figure 5), standardized such that the values align to the same change in protonation state. Although a large portion of the ChEMBL data match closely to the experimental data, many deviate significantly. Some of the low deviations are attributed to calculation error. The high deviations often correspond to those cases such as EDTA, in which the "most basic" and "most acidic" pK_a values are incongruent with the proton gain/loss terminology. Proton gain values in ChEMBL are systematically too low, and proton loss values too high, which agrees with this hypothesis that there are issues with the pK_a value sampling.

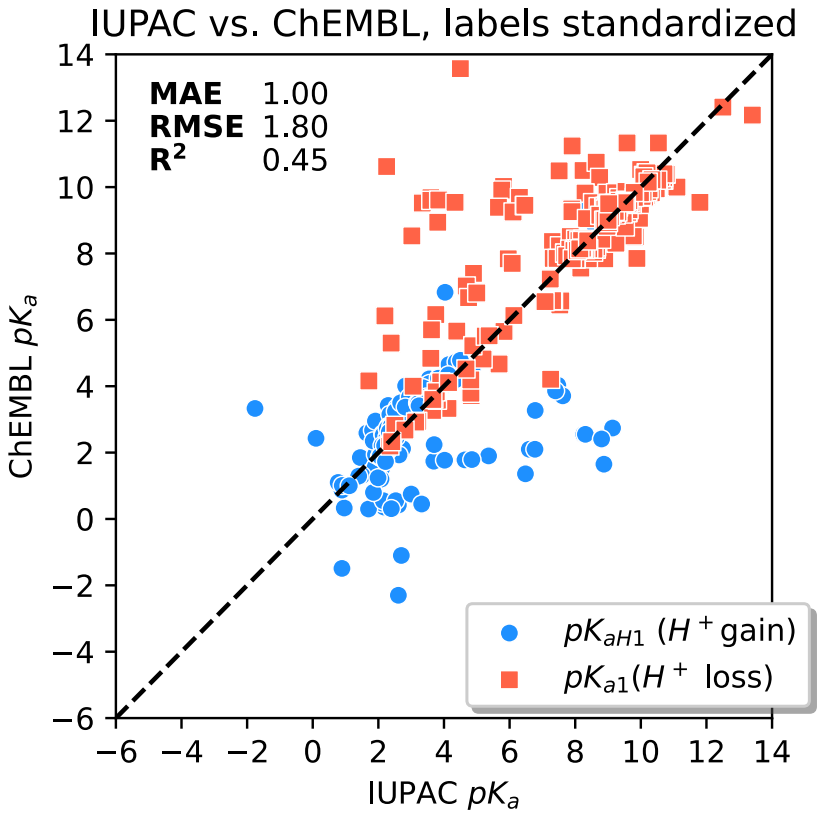

Figure 5: Parity between the IUPAC data against the ChEMBL synthetic data.

Several recently published models use ChEMBL data for pre-training and full training, respectively, each with varying degrees of errors. These data are thereby misinterpreted, at which point the terminology goes from confusing to *deleterious*, causing model misprediction.

Model error due to ChEMBL data

In recent years, numerous models have used ChEMBL Database pK_a calculations during training.

MolGpKa is a pK_a predictor published in 2021.[26] The model was trained on values from the ChEMBL database and is publicly available as a web-server with a conveniently-accessible API. As of June 2024, the web server site has over 500,000 page views. Although trained on macroscopic pK_a values, it attempts to compute microscopic pK_a values, and

hence follows the same acidity ordering as seen in ChEMBL. The model converts zwitterions into their uncharged tautomers, and predicts that glycine's uncharged tautomer has an acidic pK_a of 2.1 and a basic pK_a of 9.6. It has the same issues with inconsistency as discussed previously in the ChEMBL data, and is hence not discussed further here.

Several models have involved pretraining on ChEMBL values and then fine-tuning on experimental data. MF-SuP-pK_a[28] and pkasolver,[29] both released in 2022, are two such recent examples. Unfortunately, because the models are not publicly available, we could not assess their performance on amino acids.

Uni-pK_a, published in 2024, also leverages the ChEMBL dataset to pretrain a module that leverages 3D information and computed free energies of individual microstates to calculate the overall macroscopic pK_a. The model accounts for tautomerism, capturing the microscopic pK_a of both the uncharged and zwitterionic tautomers. To our knowledge, this is the only recently-released ML model that correctly distinguishes between those microstates. While showing excellent performance, the weights of the model in the original manuscript are not publicly available, and we do not examine it further here.[25]

QupKake is a recently-published machine learning model for the prediction of microscopic p$K_a$values.[27] When assessed against the SAMPL6-8, Novartis, and literature datasets, the model was shown to exhibit state-of-the-art performance, with test RMSEs between 0.5 to 0.8 pH units. The model is pre-trained on data from the ChEMBL database, whose most acidic and most basic protonation sites are determined using a surrogate model trained on CREST, which uses semi-empirical quantum mechanical methods to estimate the relative stability of protomers.[40] In the training process, zwitterions were not considered. Due to the fine tuning, predictions from QupKake are not always similar to the calculations in ChEMBL.

Because QupKake is publicly available, is considered state-of-the-art in several metrics, and involves fine-tuning (which yields a more complicated relationship with the ChEMBL data), we examine this model more closely here. Although QupKake attempts to predict

microscopic pK_a information, it is pre-trained on macroscopic pK_a calculations. We hence believe it more accurate to compare QupKake predictions to macroscopic pK_a values in the IUPAC digitized dataset.

As a simple test case, let us compare the model prediction for glycine to the data. Using the uncharged tautomer as input to the model, we might expect the model to perform well; it is the simplest amino acid, and contains two of the most common acidity centers across all acids and bases. However, the "acidic" and "basic" pK_a values (interpreted by QupKake as proton loss and gain) predicted are 2.4 and 7.8. The different interpretations of pK_a labeling make direct comparison difficult. For the zwitterion microscopic pK_a (which largely matches the values of the macroscopic pK_a), the ChEMBL values are 9.2 and 2.3, and the experimental values are 9.4 for proton loss and 2.4 for proton gain. For the uncharged microstate, the experimental values are instead 4.3 and 7.6. The model predictions of 2.4 and 7.8 are therefore inconsistent with both microstates: for the uncharged state, the lower pK_a is nearly 2 pH units in error, whereas for the zwitterion state, the higher pK_a is almost 2 pH units in error (and the acidity ordering is flipped).

To better assess the accuracy of the model, we further examined a larger set of potentially zwitterionic compounds. To do this, we took the subset of ChEMBL calculations whose reported acidic pK_a was lower than the basic one, and then found the intersection of that set with the IUPAC dataset. To further simplify the comparison, we downsampled to compounds with just one acidic center and one basic acidity center. This resulted in a subset of 52 proton gain and 69 proton loss pK_a values. Then, we compared both the ChEMBL calculation and QupKake prediction against the IUPAC data.

We ran the calculations with the following assumptions:

1. We used uncharged tautomers, as the QupKake model was not trained on zwitterionic species and therefore predicts values close to 7 when presented with zwitterions.

2. We assume that the "basic" pK_a in ChEMBL and QupKake correspond to proton *loss*, and likewise for the "acidic" pK_a, reflecting the confusing nature of the acidity labels.

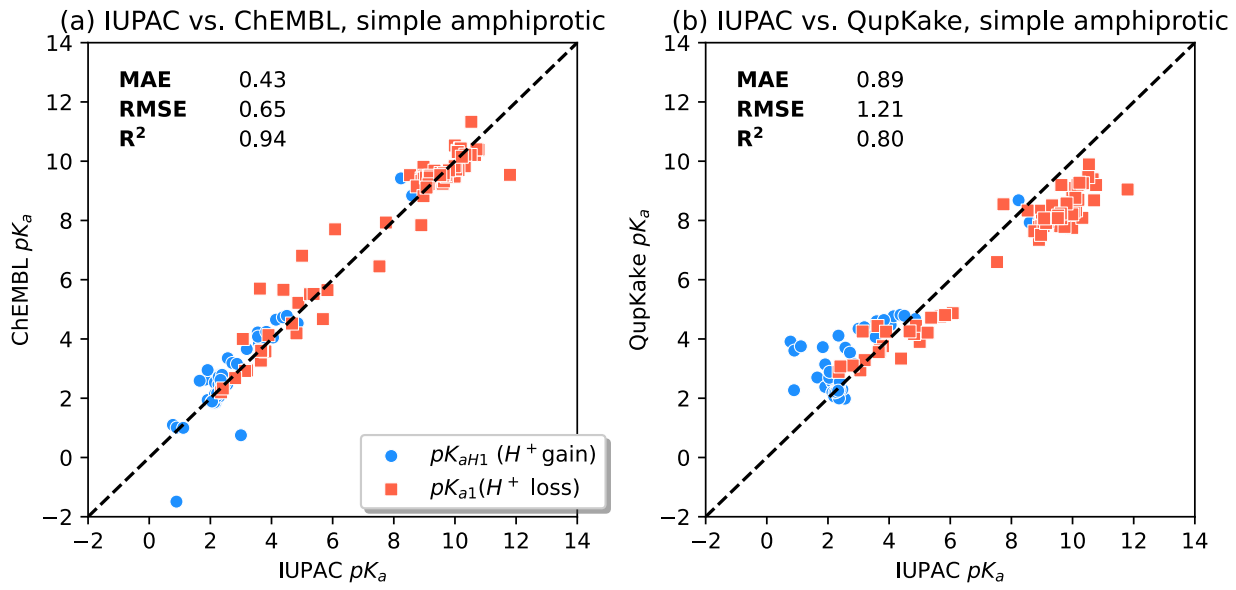

Figure 6: Parity plot comparing the downsampled, simple test set of compounds from IUPAC experimental data vs. (a) ChEMBL calculated values, and (b) QupKake machine learning model predictions.

Figure 6 shows the comparison of both ChEMBL and QupKake predictions against the experimental data. The ChEMBL calculations very accurately recreate the experimental data. For QupKake, the RMSE shown is 1.2 pH units; while this is still decent, this is greater than the test RMSEs of 0.5 to 0.8 pH units reported by QupKake for the SAMPL and Novartis test sets, and greater than the deviation of experimental data compared to ChEMBL. If the QupKake model were able to learn perfectly from the ChEMBL data, then it should predict closer to the ChEMBL calculation errors, i.e., closer to 0.6 RMSE. The parity plots also show that such errors are systematic, with many proton gain values too high and proton loss values too low by about 1-2 pK_a units. We believe that this lower accuracy is for the most part due to inconsistency of microstate information as provided to the model during training, and that predictions for such compounds can be learned just as they can for other simple polyprotic species like aminophenols. The misinterpretation of pK_a labels in datasets has the potential to impair model predictions, and should be rectified in both the dataset *and* modeling sides.

This issue has not attracted significant discussion because amino acids and other zwitterionic compounds are not frequently seen and analyzed in benchmarking sets. We also emphasize that this process involved prior knowledge of potential labeling issues - as we ascribed the values predicted of the *uncharged* microstate to the acidity behavior of the *zwitterionic* microstate.

Implications of erroneous pK_a

Acid dissociation constants are widely used to predict biochemical properties. They are also used to predict thermodynamic properties such as the free energy of ions in solution, which can be used to estimate solubilities and rate coefficients. Here, we assess the sensitivity of these properties to changes in pK_a.

If we use pK_a predictions for amino acids assuming they follow the same acidity / basicity labels as simple ampholytes such as m-aminophenol, then we will have a flipped understanding of the pK_a label. Let us call this the "naive" case, in which we consistently obtain pK_{a1} values that are about 5 pH units too low and pK_{aH1} values 5 pH units too high.

In the optimistic scenario that we are correctly mapping the acidity labels, poor performance is still occasionally observed. We compute pK_{aH1} = 2.3 and pK_{a1} = 7.8 using QupKake, compared to the microscopic values of 2.4 and 9.8, respectively for the zwitterion, and 4.3 and 7.6 for the uncharged tautomer. This corresponds to errors of 0.1 and 2.0 pH units for the former and 2.0 and 0.2 pH units for the latter microstates.

In the remainder of this section, we will use QupKake's predicted values for glycine to assess the significance of pK_a errors in both of these "naive" and "corrected" usage. We emphasize, again, that QupKake is not necessarily the only model that has such errors - and indeed, it exhibits excellent performance in other metrics - and is examined here only because it is open-source, state-of-the-art, and utilizes the ChEMBL data in a complex manner.

Solubility

The pK_a is used directly to calculate solubility for ionizable compounds and partitioning (e.g. in lipophilicity). A common way is to use it in the Henderson-Hasselbalch expression, through which we obtain the following expressions for solubility for an ampholyte with one acidic and one basic pK_a:

$$\frac{S(pH)}{S_0} = 10^{\mathrm{pK_{aH1}-pH}} + 10^{-\mathrm{pK_{a1}+pH}} + 1 \tag{5}$$

where $S(pH)$ is the solubility of the drug at a given pH and S_0 is the solubility of the un-ionized drug (i.e. at the isoelectric point).[11,41,42] The pK_a and pK_{aH1} used are the experimentally observed macroscopic values of solubility.[43]

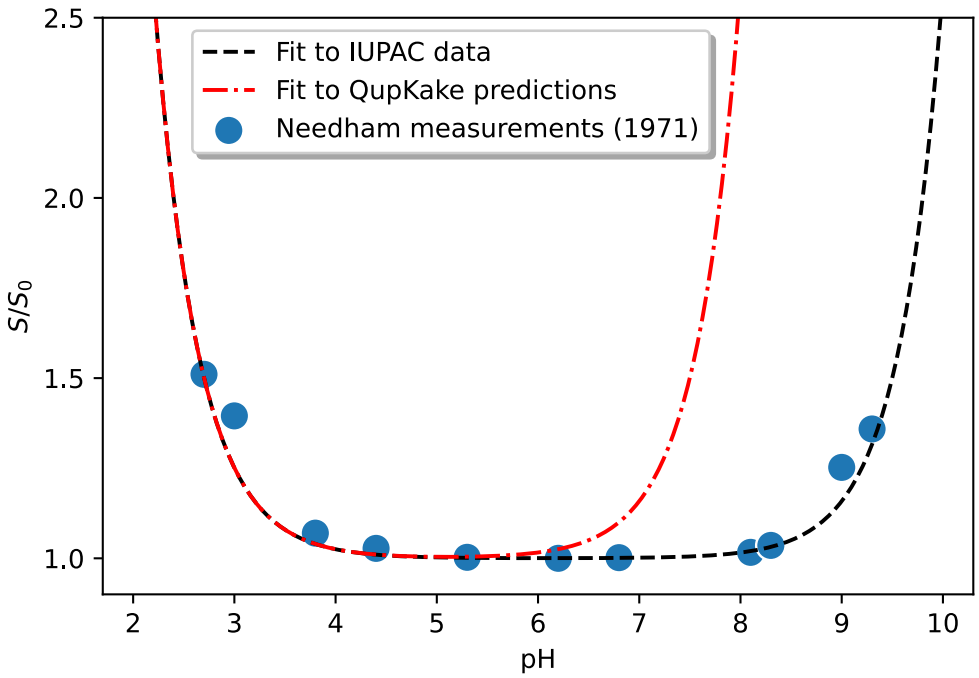

Figure 7: Solubility predictions using the Henderson-Hasselbalch prediction with experimental data versus QupKake predictions.

Figure 7 shows that the Henderson-Hasselbalch relationship accurately depicts the relationship between solubility and pH if parameterized with the IUPAC data. In the naive

case that pK_a data is used with flipped labels, the relationship predicts complete dissolution across all pH (for instance, at pH = 5, the predicted S/S_0 is 1000; its corresponding S-pH curve is not shown in Figure 7 due to the difference in magnitudes). If using QupKake, one must choose to use the uncharged microstate (trained on *macroscopic* pK_a data), which already introduces a modeling error. Using the predicted proton loss pK_a of 7.8, one sees that the solubility predictions are significantly shifted. At the pH of blood (roughly 7.4) the effect of ionization would lead to a less than 1% increase in solubility, whereas using the ML prediction would arrive at an estimate of 40% increased solubility. Such errors therefore have potential to affect drug interactions in the human body.

Chemical kinetics and thermodynamics

The Gibbs free energy difference for an acid dissociation can be related from the equilibrium constant. For equation 1, for instance, the free energy of reaction is:

$$\Delta G_{\text{rxn}} = \ln 10 \cdot \text{RT} \cdot \text{p}K_a(\text{AH}) \tag{6}$$

where R is the molar gas constant, T is the temperature, and ΔG_{rxn} is the Gibbs free energy of reaction for an acid dissociation. At 25 °C, the term $\ln 10 \cdot \text{RT}$ equals 1.36 kcal mol^{-1}. ΔG_{rxn} is connected to kinetics; for one, it can be used to compute the equilibrium coefficient, allowing the reverse reaction's rate to be determined if the forward rate is known. It can be used via a thermodynamic cycle to compute the free energy of an ionic reactant in solution,[13,14] which can then be used (along with information about the transition state) in kinetic simulations.[44]

If naively using the pK_a values for glycine, mistakenly flipping their order, this would lead to errors of 5 pH units which propagates to an error of nearly 7 kcal mol^{-1}. This level of inaccuracy is unacceptable in chemical kinetics - if used in either of the two kinetic applications described above, this would correspond to a factor of ≈136,500 error in computed room-temperature rate coefficients. The errors would be even higher (closer to 6 pH units,

or 8 kcal mol^{-1}) if using the data from ChEMBL or models such as MolGpKa.

Using the computed values with the corrected acidity order would still result in errors of 2 pH units for one of the pK_a values for either the uncharged or zwitterion tautomers, which corresponds to 2.7 kcal mol^{-1}, or a factor of about 100 error in rate coefficients. Generally, thermodynamic errors of <0.5 kcal mol^{-1}, corresponding to rate errors of about 2, are considered to be within chemical accuracy.

Recommendations for future researchers

We are excited by recently-developed models such as QupKake and Uni-pK_a which attempt to utilize microstate information in pK_a prediction. We believe that including 3D conformations and energies computed using QM will lead to more accurate and less ambiguous future models. However, future model developers must take utmost care to distinguish exactly what property their model is predicting - macroscopic or microscopic - as well as the corresponding nature of the training data.

When using ChEMBL versions ≤34, we encourage researchers to identify wherever the acidic pK_a value is less than the basic one, and consider correcting or discarding this subset of data. Furthermore, zwitterionic tautomers should not be excluded during training by default.

This issue is far more fundamental than just pK_a and amino acids; it speaks to the difficulty in representing and standardizing the tautomeric information of a molecule. Although convenient to relate just a single tautomer of a 2D molecular graph to a macroscopic property, it is increasingly apparent that *ensembles* of tautomers and conformations are critical to accurate pK_a prediction for polyprotic and zwitterionic compounds, and perhaps other properties as well. Some standard open-source cheminformatics packages include methods for identifying the possibility for an ampholyte to form a zwitterion *a priori*; for example, Dimorphite-DL,[45] which enumerates the microstates of a compound at a given pH, and Open Babel.[46] However, there is not yet an open-source method for accurately predicting

the relative population of zwitterionic and uncharged microstates based on molecular graphs. This issue of *macroscopic* and *microscopic* pK_a increasingly becomes important if multiple ionizable groups are present. Efforts in microscopic pK_a prediction should include quantum-chemical computation of free energies for all relevant tautomers at given charge states, as is being explored in recent machine-learning models.

We suggest the following guidelines to avoid future confusion:

- If known, pK_a data should be labeled as macroscopic or microscopic.
- pK_a should *only* refer to proton loss; i.e., if speaking of "the pK_a of AH", this should always refer to AH transforming to A^-, and not into AH_2^+.
- If microscopic, the data point should indicate a description of the exact microstate and ionization center.
- For a macroscopic pK_a, if a chemical identifier for the compound at that charge state is known, then it should be provided. If not, then an identifier for the neutral form, along with the charge state transition or numerical order of the dissociation with respect to the neutral charge state, should be provided. For instance, it is far preferable to describe the dissociation reaction HCO_3^- forming CO_3^{2-} as the bicarbonate anion's pK_a, but it could also be described as carbonic acid's pK_{a2}.
- If no information about the charge state is known because the molecule is amphiprotic, then the pK_a may simply be arranged in ascending order along with an identifier for its neutral form. However, such cases should include some form of annotation to avoid the impression that all such values are acidic with respect to the +0 charge state.
- The usage of "acidic" and "basic" pK_a is best avoided in dataset labels.

Conclusion

Considerable progress has been made in recent years toward prediction of pK_a for monoprotic and some polyprotic species. However, inconsistent and confusing terminology serves as an obstacle for pK_a prediction as pertains to true microscopic pK_a prediction, especially for drug compounds which tend to have many acidity centers and/or form zwitterions. The situation is even worse for pK_a data in non-aqueous solvents, for which different acidity scales frequently appear, and solvent effects such as homoconjugation and ion pairing are treated inconsistently. Such systematic errors in data can exceed 3 pH units.[47,48]

We have used glycine as a comparison case to show that errors in pK_a prediction of amino acids can lead to large errors in predicted solubility and kinetics. Many drug molecules are simple amino acids similar to glycine, such as aminocaproic acid (a clotting promoter), pregabalin (a treatment for seizures and nerve pain), and methyldopa (a drug for hypertension). Such errors will be higher and even more difficult to correct in more complex drug molecules. The potential for impact on life is substantial, and these errors should be corrected in existing databases to reduce the risk of impact in drug development.

Much effort in recent years has also focused on developing, tweaking, and experimenting with model architectures. We strongly encourage future researchers to also focus on examining datasets, understanding the data provenance, limitations, and potential issues, before using the data for model development. Errors in data can ossify, leading to misconceptions that could perpetuate for years to come. Improving data will naturally lead to better models. They will also lead to better benchmarks and allow for a more holistic understanding of model performance. It is our perspective that the low quantity of high-quality, consistent data is currently the primary issue precluding the development of AI in chemistry, and we hence encourage further research in this domain.

Data and Software Availability

The full set of ChEMBL data is openly available on the ChEMBL website at: `https://www.ebi.ac.uk/chembl/downloads/`. The experimental pK_a data were obtained from the IUPAC Aqueous pK_a dataset, openly available on Zenodo at: `https://zenodo.org/doi/10.5281/zenodo.7236452`. The QupKake model used in this work was obtained from the QupKake GitHub repository at: `https://github.com/Shualdon/QupKake`. The pK_a and solubility data used in this study are available in the Supporting Information.

Acknowledgement

J. W. Z. and W. H. G. acknowledge the Machine Learning for Pharmaceutical Discovery and Synthesis Consortium (MLPDS) for funding. The work of I. L. was supported by the Estonian Research Council grant (PRG960).

Supporting Information Available

Supporting information available:

- .csv files containing the data and calculations used in this study

References

(1) Manallack, D. T.; Prankerd, R. J.; Yuriev, E.; Oprea, T. I.; Chalmers, D. K. The significance of acid/base properties in drug discovery. *Chemical Society Reviews* **2013**, *42*, 485–496.

(2) Mamy, L.; Patureau, D.; Barriuso, E.; Bedos, C.; Bessac, F.; Louchart, X.; Martin-Laurent, F.; Miege, C.; Benoit, P. Prediction of the fate of organic compounds in the

environment from their molecular properties: A review. *Critical Reviews in Environmental Science and Technology* **2015**, *45*, 1277–1377.

(3) Bell, R. P. *The proton in chemistry*; Springer Science & Business Media, 2013.

(4) Young, R. J.; Leeson, P. D. Mapping the efficiency and physicochemical trajectories of successful optimizations. *Journal of Medicinal Chemistry* **2018**, *61*, 6421–6467.

(5) Shultz, M. D. Two decades under the influence of the rule of five and the changing properties of approved oral drugs: miniperspective. *Journal of Medicinal Chemistry* **2018**, *62*, 1701–1714.

(6) Reijenga, J.; Van Hoof, A.; Van Loon, A.; Teunissen, B. Development of methods for the determination of pKa values. *Analytical chemistry insights* **2013**, *8*, ACI–S12304.

(7) Davies, M.; Nowotka, M.; Papadatos, G.; Dedman, N.; Gaulton, A.; Atkinson, F.; Bellis, L.; Overington, J. P. ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic acids research* **2015**, *43*, W612–W620.

(8) Zdrazil, B.; Felix, E.; Hunter, F.; Manners, E. J.; Blackshaw, J.; Corbett, S.; de Veij, M.; Ioannidis, H.; Lopez, D. M.; Mosquera, J. F.; others The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic acids research* **2024**, *52*, D1180–D1192.

(9) Zhang, H. A QSAR study of the brain/blood partition coefficients on the basis of pKa values. *QSAR & Combinatorial Science* **2006**, *25*, 15–24.

(10) Bas, D. C.; Rogers, D. M.; Jensen, J. H. Very fast prediction and rationalization of pKa values for protein–ligand complexes. *Proteins: Structure, Function, and Bioinformatics* **2008**, *73*, 765–783.

(11) Avdeef, A. Solubility of sparingly-soluble ionizable drugs. *Advanced drug delivery reviews* **2007**, *59*, 568–590.

(12) Caliaro, G. A.; Herbots, C. A. Determination of pKa values of basic new drug substances by CE. *Journal of pharmaceutical and biomedical analysis* **2001**, *26*, 427–434.

(13) Zheng, J. W.; Green, W. H. Experimental Compilation and Computation of Hydration Free Energies for Ionic Solutes. *The Journal of Physical Chemistry A* **2023**, *127*, 10268–10281.

(14) Kröger, L. C.; Müller, S.; Smirnova, I.; Leonhard, K. Prediction of solvation free energies of ionic solutes in neutral solvents. *The Journal of Physical Chemistry A* **2020**, *124*, 4171–4181.

(15) Chung, Y.; Green, W. H. Computing Kinetic Solvent Effects and Liquid Phase Rate Constants Using Quantum Chemistry and COSMO-RS Methods. *The Journal of Physical Chemistry A* **2023**, *127*, 5637–5651.

(16) Albert, A.; Serjeant, E.; Albert, A.; Serjeant, E. Zwitterions (dipolar ions). *The Determination of Ionization Constants: A Laboratory Manual* **1984**, 126–134.

(17) Scholz, F.; Kahlert, H. Acid–base equilibria of amino acids: microscopic and macroscopic acidity constants. *ChemTexts* **2018**, *4*, 1–9.

(18) Harris, L. J. The combination of proteins, amino-acids, &c., with acids and alkalis. Part II.—Titration curves of amino-acids in presence of formaldehyde. *Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character* **1929**, *104*, 412–439.

(19) Harris, L. J.; Birch, T. W. Zwitterions: Proof of the zwitterion constitution of the amino-acid molecule. II. Amino-acids, polypeptides, etc., and proteins as zwitterions, with instances of non-zwitterion ampholytes. *Biochemical Journal* **1930**, *24*, 1080.

(20) Bjerrum, N. Dissoziationskonstanten von mehrbasischen Säuren und ihre Anwendung

zur Berechnung molekularer Dimensionen. *Zeitschrift für physikalische Chemie* **1923**, *106*, 219–242.

(21) Locke, M. J.; Hunter, R. L.; McIver Jr, R. T. Experimental determination of the acidity and basicity of glycine in the gas phase. *Journal of the American Chemical Society* **1979**, *101*, 272–273.

(22) Locke, M. J.; McIver Jr, R. T. Effect of solvation on the acid/base properties of glycine. *Journal of the American Chemical Society* **1983**, *105*, 4226–4232.

(23) Laughlin, R. G. Fundamentals of the zwitterionic hydrophilic group. *Langmuir* **1991**, *7*, 842–847.

(24) ChemAxon Docs: Red and blue representation of pKa values. `https://docs.chemaxon.com/display/docs/calculators_red-and-blue-representation-of-pka-values.md`, Accessed: 7-31-2024.

(25) Luo, W.; Zhou, G.; Zhu, Z.; Yuan, Y.; Ke, G.; Wei, Z.; Gao, Z.; Zheng, H. Bridging Machine Learning and Thermodynamics for Accurate p K a Prediction. *JACS Au* **2024**,

(26) Pan, X.; Wang, H.; Li, C.; Zhang, J. Z.; Ji, C. MolGpka: A Web Server for Small Molecule pKa Prediction Using a Graph-Convolutional Neural Network. *Journal of Chemical Information and Modeling* **2021**, *61*, 3159–3165.

(27) Abarbanel, O. D.; Hutchison, G. R. QupKake: Integrating Machine Learning and Quantum Chemistry for Micro-pKa Predictions. *Journal of Chemical Theory and Computation* **2024**,

(28) Wu, J.; Wan, Y.; Wu, Z.; Zhang, S.; Cao, D.; Hsieh, C.-Y.; Hou, T. MF-SuP-pKa: Multi-fidelity modeling with subgraph pooling mechanism for pKa prediction. *Acta Pharmaceutica Sinica B* **2022**,

(29) Mayr, F.; Wieder, M.; Wieder, O.; Langer, T. Improving small molecule pKa prediction using transfer learning with graph neural networks. *Frontiers in Chemistry* **2022**, *10*, 866585.

(30) Marín, D.; Vega, M.; Lebrero, R.; Muñoz, R. Optimization of a chemical scrubbing process based on a Fe-EDTA-carbonate based solvent for the simultaneous removal of CO2 and H2S from biogas. *Journal of Water Process Engineering* **2020**, *37*, 101476.

(31) Grčman, H.; Velikonja-Bolta, Š.; Vodnik, D.; Kos, B.; Leštan, D. EDTA enhanced heavy metal phytoextraction: metal accumulation, leaching and toxicity. *Plant and soil* **2001**, *235*, 105–114.

(32) Serper, A.; Çalt, S. The demineralizing effects of EDTA at different concentrations and pH. *Journal of Endodontics* **2002**, *28*, 501–502.

(33) Calt, S.; Serper, A. Time-dependent effects of EDTA on dentin structures. *Journal of endodontics* **2002**, *28*, 17–19.

(34) Banfi, G.; Salvagno, G. L.; Lippi, G. The role of ethylenediamine tetraacetic acid (EDTA) as in vitro anticoagulant for diagnostic purposes. **2007**,

(35) Belle-Oudry, D. Quantitative analysis of sulfate in water by indirect EDTA titration. *Journal of chemical education* **2008**, *85*, 1269.

(36) Zheng, J.; Lafontant-Joseph, O. IUPAC/Dissociation-Constants. 2024; `https://doi.org/10.5281/zenodo.7236452`.

(37) Perrin, D. D. *Dissociation Constants of Organic Bases in Aqueous Solutions*; Franklin Book Company, 1965; Vol. 1.

(38) Perrin, D. D. *Dissociation Constants of Organic Bases in Aqueous Solutions: Supplement*; Franklin Book Company, 1972; Vol. 1.

(39) Serjeant, E. P.; Dempsey, B. Ionisation constants of organic acids in aqueous solution. *IUPAC chemical data series* **1979**, *23*, 160–190.

(40) Pracht, P.; Bohle, F.; Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Physical Chemistry Chemical Physics* **2020**, *22*, 7169–7192.

(41) Shin, H. K.; Kang, Y.-M.; No, K. T.; others Predicting ADME properties of chemicals. *Handbook of computational chemistry* **2017**, *59*, 2265–2301.

(42) Hansen, N. T.; Kouskoumvekaki, I.; Jørgensen, F. S.; Brunak, S.; Jonsdottir, S. O. Prediction of pH-dependent aqueous solubility of druglike molecules. *Journal of chemical information and modeling* **2006**, *46*, 2601–2609.

(43) Needham Jr, T.; Paruta, A.; Gerraughty, R. Solubility of amino acids in pure solvent systems. *Journal of Pharmaceutical Sciences* **1971**, *60*, 565–567.

(44) Chung, Y.; Green, W. H. Machine learning from quantum chemistry to predict experimental solvent effects on reaction rates. *Chemical Science* **2024**,

(45) 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. *Journal of Cheminformatics* **2019**, *11*, 1–8.

(46) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *Journal of cheminformatics* **2011**, *3*, 1–14.

(47) Kütt, A.; Selberg, S.; Kaljurand, I.; Tshepelevitsh, S.; Heering, A.; Darnell, A.; Kaupmees, K.; Piirsalu, M.; Leito, I. pKa values in organic chemistry–Making maximum use of the available data. *Tetrahedron letters* **2018**, *59*, 3738–3748.

(48) Zheng, J. W.; Al Ibrahim, E.; Green, W. H. *Journal of Computational Chemistry*, In press, forthcoming.

TOC Graphic