Supporting Information for: Digitized dataset of aqueous acid dissociation constants

Jonathan W. Zheng, Olivier Lafontant-Joseph, and William H. Green*

Massachusetts Institute of Technology, Department of Chemical Engineering

E-mail: whgreen@mit.edu

Additional reference works

We additionally procured permission to digitize the information in the following three reference books:

1. **Kortum**: Dissociation Constants of Organic Acids in Aqueous Solution; G. Kortum, W. Vogel and K. Andrussow; Butterworths (1961)

2. **Perrin Inorganic**: Dissociation Constants of Inorganic Acids and Bases in Aqueous Solution; D. D. Perrin; Butterworths (1969)

3. **Izutsu**: Acid-Base Dissociation Constants in Dipolar Aprotic Solvents; Izutsu, K; Blackwell (1990)

Data from these books are excluded from this dataset, as the cleanup and curation are still in preliminary stages. However, we anticipate that portions of some of these datasets will soon become publicly available.

Temperature and pressure variation examples

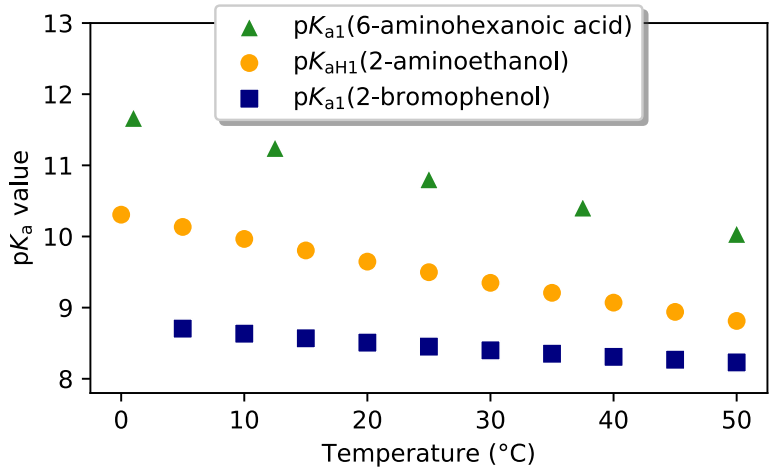


Figure S1: Example of temperature profiles in dataset for several different types of compounds between 0 and 50 °C.

Figure S1 shows some examples of temperature dependence between 0 to 50 °C at 1 bar pressure. Whereas the temperature dependence is weak for 2-bromophenol, it has a much stronger effect (around 1 pK_a unit deviation) for 6-aminohexanoic acid. For a discussion of temperature dependence, see the references cited here.[1–3]

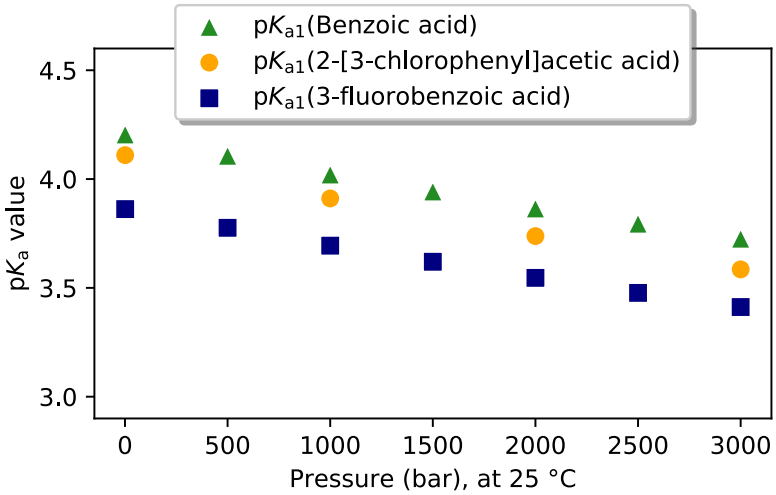


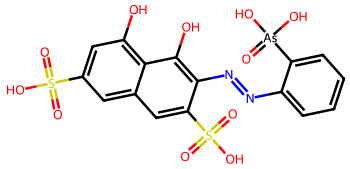
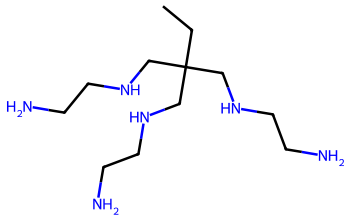
Figure S2: Example of pressure profiles in dataset for several different types of compounds between 1 and 3000 bar.

Figure S2 shows some examples of pressure dependence between 1 to 3000 bar at 25 °C.

Pressure has a moderate effect, with pK_a in these cases changing by about 0.5 pK_a units. For a discussion of pressure dependence, see the references cited here.[3,4]

Most numerous acid/base sites

Table S1: Species with most measured pK_a and pK_{aH} values.

Number of sites	Molecule
6 pK_a	
6 pK_{aH}	

Examples of species with data for the most acidic and basic sites (6 pK_a and 6 pK_{aH}, respectively) are shown in Table S1. The first compound, Arsenazo I, includes two sulfonic acid sites, two hydroxyl sites attached to the naphthalene moiety, and an arsono group. The second compound includes three primary and three secondary amine groups.

Data leakage details

ChEMBL

About half of the SAMPL 6-8 pK_a challenge molecules (30 of the 67 total) are present in the ChEMBL version 35 (CHEMBL35) pK_a database, nearly all of which are direct InChI matches (in total spanning 39 matches in the ChEMBL database; the higher number is because some are salts or similar isomers). Of these 30, 23 are from SAMPL6 (out of a total of 24 total compounds in SAMPL6), 1 is from SAMPL7 (out of a total of 22), and 6 are from SAMPL8 (out of a total of 21). The ChEMBL35 dataset also includes 269 of the 280 species that are present in the Novartis dataset - comprising 95% of the dataset. Of these, 110 entries are for acids (out of the 112 acids in the Novartis test set), and 159 are for bases (out of the 168 bases in the Novartis test set).

Many recently-developed machine learning models are trained on ChEMBL but do not report pruning out molecules that are present in both training and test sets, as should have been done. There is therefore a high potential for data leakage in the existing literature, especially considering that *nearly the entire* Novartis dataset is found within the ChEMBL database. This data leakage would make resulting models appear better at predicting pK_a than they really are.

Additionally, some of the datapoints in ChEMBL and throughout the literature are serious errors, due to the presence of zwitterionic protomers labeled in a confusing way which leads to inversion of the meaning of "acidic" and "basic" labels. In many cases ChEMBL actually reports the "least basic" and "least acidic" values instead, leading sometimes to errors of more than 9 pK_a units.[5] One possible solution is for users to filter out any pK_a values from ChEMBL whose most "acidic" pK_a value is lower than its most "basic" value. This was done in this work, although unfortunately it also has the side effect that zwitterionic compounds will be underrepresented during training.

iBonD

The commonly-used[6–9] iBonD data also has data leakage. Using the version of the database released by An et al.,[7] four entries corresponding to three distinct species in the SAMPL8 challenge are also present in iBonD. For the Novartis set, two bases and seven acids are also present in iBonD. iBonD spans numerous solvents, but these overlapping species all only have values in water.

DataWarrior

The commonly-used DataWarrior dataset[10–14] also has four values that overlap with the SAMPL test sets, corresponding to three species (noted previously by Wu et al.[15]). Three of the values are in SAMPL6 and one in SAMPL8. These 4 datapoints should be removed before testing with SAMPL to avoid data leakage contaminating the test. By nature of its construction, the DataWarrior set does not leak with any Novartis data.

Modeling details

Model training details

The IUPAC models shown in the maintext were trained using 80% train / 10% validation / 10% test splits. The 10 ChEMBL models were trained with 90% train / 10% validation splits. All models were trained using the D-MPNN architecture implemented in Chemprop v2.2, with additional features to represent the temperature and pK_a type. Chemprop is available at https://github.com/chemprop/chemprop and can be used either through the command-line interface or Python.

For training the models, the IUPAC data was processed such that only values between temperatures of 20 to 30 °C were included. If multiple temperatures were present in this range, the pK_a and temperature values were averaged and assigned to a single datapoint.

The models are trained explicitly on *macroscopic* pK_a prediction, taking as input the SMILES string, temperature, and dissociation type. The pK_a type was encoded as an integer feature: -1 for pK_{a1} and 1 for pK_{aH1}. The temperature was encoded as a float. This results in an (N, 2) array where N represents the number of entries, wherein the first column is the pK_a type and the second column is the temperature.

For finetuning: during pretraining, the FFNN and D-MPNN encoder were allowed to update their weights. For finetuning, the encoder weights were frozen, allowing just the FFNN weights to update.

In all analyses, we exclude the "p$K_a > 12$" values from the SAMPL7 data.

The hyperparameters used for the non-ChEMBL models are below.

Table S2: Hyperparameters used in GNN models. All other hyperparameters not shown here are the default values in Chemprop v2.2.

Hyperparameter	**Value**
Depth	2
FFNN number of layers	1
Dropout	0.4
MPNN hidden dimension	800
FFNN hidden dimension	1500

Data cleaning and augmentation for model training

- InChI strings were sanitized by standardizing different protonation states, different salt forms, and different tautomeric representations to the same string.

- Entries with non-numeric or missing values were dropped. Only first dissociations, for pK_{aH} and pK_a, were kept. From the IUPAC data, only "Reliable" and "Approximate" values were kept (which have cited pK_a uncertainties less than 0.04 pK_a units).

For the ChEMBL data, we removed any compounds with acidic pK_a values lower than pK_{aH} values, based on findings from our previous work that many such values are in error.[5]

For all datasets - IUPAC and ChEMBL - we removed species found in the SAMPL or Novartis test sets, based on matches to sanitized InChI strings. The ChEMBL dataset was also pruned to remove any IUPAC dataset compounds.

References

(1) Leung, C. S.; Grunwald, E. Temperature dependence of. DELTA.-Cp. deg. for the self-ionization of water and for the acid dissociation of acetic acid and benzoic acid in water. *The Journal of Physical Chemistry* **1970**, *74*, 687–696.

(2) 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.

(3) Samuelsen, L.; Holm, R.; Lathuile, A.; Schönbeck, C. Buffer solutions in drug formulation and processing: How pKa values depend on temperature, pressure and ionic strength. *International journal of pharmaceutics* **2019**, *560*, 357–364.

(4) Neuman Jr, R. C.; Kauzmann, W.; Zipp, A. Pressure dependence of weak acid ionization in aqueous buffers. *The Journal of Physical Chemistry* **1973**, *77*, 2687–2691.

(5) Zheng, J. W.; Leito, I.; Green, W. H. Widespread Misinterpretation of pKa Terminology for Zwitterionic Compounds and Its Consequences. *Journal of Chemical Information and Modeling* **2024**, *64*, 8838–8847.

(6) Liu, S.; Yang, Q.; Zhang, L.; Luo, S. Highly Precise Prediction of Micro-and Supra-pKa Based on 3D Descriptors Integrating Non-Covalent Interactions. *Angewandte Chemie* **2025**, *137*, e202424069.

(7) An, H.; Liu, X.; Cai, W.; Shao, X. AttenGpKa: a universal predictor of solvation acidity using graph neural network and molecular topology. *Journal of Chemical Information and Modeling* **2024**, *64*, 5480–5491.

(8) Yang, Q.; Li, Y.; Yang, J. D.; Liu, Y.; Zhang, L.; Luo, S.; Cheng, J. P. Holistic Prediction of the pKa in Diverse Solvents Based on a Machine-Learning Approach. *Angewandte Chemie - International Edition* **2020**, *59*, 19282–19291.

(9) Nevolianis, T.; Zheng, J.; Müller, S.; Baumann, M.; Tshepelevitsh, S.; Kaljurand, I.; Leito, I.; Smirnova, I.; Green, W.; Leonhard, K. Solvation free energies of anions: from new reference data to predictive models. *ChemRxiv* **2025**,

(10) Wagen, C. Physics-Informed Machine Learning Enables Rapid Macroscopic pKa Prediction. *arXiv* **2025**,

(11) Sander, T.; Freyss, J.; Von Korff, M.; Rufener, C. DataWarrior: An open-source program for chemistry aware data visualization and analysis. *Journal of Chemical Information and Modeling* **2015**, *55*, 460–473.

(12) 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. *Journal of Cheminformatics* **2019**, *11*, 1–20.

(13) Abarbanel, O.; Hutchison, G. QupKake: Integrating Machine Learning and Quantum Chemistry for micro-pKa Predictions. **2023**,

(14) Yang, C.; Gong, C.; Zhang, Z.; Fang, J.; Li, W.; Liu, G.; Tang, Y. In Silico Prediction of pKa Values Using Explainable Deep Learning Methods. *Journal of Pharmaceutical Analysis* **2024**, 101174.

(15) 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**,