Predicting Antioxidant Activity

CHEG/CISC/ECEG/MSEG 848
Spring 2025

Prof. Arthi Jayaraman

Industry:

Arkema

Industry mentors:
Dr. Brian Koo, Dr. Christian Becker, and Dr. Katie Daisey

Scott Coia
Master's Student, Chemical Engineering
University of Delaware

Aakanksh Chittiprolu
Master's Student, Data Science Program
University of Delaware

Tejas Pawar
Master's Student, Data Science Program
University of Delaware

May 15th, 2025

Abstract

The primary objective of this project is to develop a machine learning-based predictive workflow to estimate the ability of antioxidants to scavenge hydroxyl ($\cdot$OH) radicals generated via Fenton and Fenton-like reactions. Specifically, the goal is to model the rate constant, k, which represents the rate of hydroxyl radical scavenging.

Our approach involved compiling a dataset of 563 reactants with rate constants obtained from NIST[1] and molecular structures encoded in the form of Simplified Molecular Input Line Entry Specification (SMILES) strings. We computed 1621 molecular descriptors using Mordred. We strategically filtered out descriptors that consisted of too many similar values and were too highly correlated. We then ran the model for regression based models (Linear, Lasso, and Ridge) and tree-based models (Random Forest, XGBoost) and we tried using Principal Component Analysis (PCA) and not using PCA for all models. Our best model for the entire dataset used a version of Random Forest where only the top 50 features are inputted. The key components of this model for its performance are that the top 128 possible protonation states were considered, features that contained over 80% constant values were removed, and for each pair of features that were correlated above .999, one of the features were dropped. Our best model to predict the entire dataset had an R^2 score of .91 and a root mean squared error (RMSE) of .27. We were able to achieve even better results by filtering the dataset into a smaller subset of just aliphatic alcohols, where our best model had an R^2 score of .97 and an RMSE of .05.

The results demonstrate the feasibility of using data-driven models for antioxidant screening and suggest the potential for generalization to new molecules. This workflow offers a foundation for integrating computational screening into chemical rate constant data.

Introduction

Arkema is producer of hydrogen peroxide (H_2O_2). Utilized in many applications, including water treatment, aseptic food packaging, chemical processing, cleaning, and electronics. H_2O_2 is a strong inorganic oxidant, and is "green" because it has H_2O and O_2 as by-products. However, the Fenton reaction, consisting of $Fe^{2+} + H_2O_2 \rightarrow Fe^{3+} + \cdot OH + OH^-$ and $Fe^{3+} + H_2O_2 \rightarrow Fe^{2+} + \cdot OOH + H^+$ as well as Fenton-like reactions involving Cu lead to rapid degradation of H_2O_2. As shown in Figure A1 in the appendix, just 500ppm of Fe and 500ppm of Cu at ambient temperature will cause all the H_2O_2 to degrade within a few minutes. It is known that the presence of free radicals causes the propagation of Fenton and Fenton-like reactions. Therefore, if there are antioxidants present in the solution that could react to $\cdot OH$ radicals and form much larger, essentially nonreactive radicals, that would significantly slow down the rate of H_2O_2 degradation. Arkema is specifically interested in studying antioxidant scavenging ability of $\cdot OH$ radicals generated via the Fenton Reactions, and for their applications specifically for low pH environments. There are four main mechanisms through which antioxidants can be scavenged.[2] The most prominent is hydrogen abstraction, which is when a radical reacts to an antioxidant to form a larger radical and water (e.g $\cdot OH + CH_3OH \rightarrow H_2O + \cdot CH_2OH$). Another major mechanism that can occur is addition, which is when the hydroxyl radical adds to an unsaturated compound, aliphatic or aromatic, to form a free radical product (e.g $\cdot OH + C_6H_6 \rightarrow \cdot OHC_6H_6$). There is also electron transfer, where the hydroxyl radical is oxidized to form the hydroxide ion (e.g $\cdot OH + Fe(CN)_6{}^{4-} \rightarrow OH^- + \cdot Fe(CN)_6{}^{3-}$). There is also the possibility of radical-radical reactions such as $\cdot OH + \cdot OH \rightarrow H_2O$, however this project will not be concerned with such reactions as there is not much data available and the available data is for small radicals that could further catalyze the Fenton reaction.[1] To study these reactions, kinetic studies can be done, however they can be expensive and time-consuming[2]. In contrast, we utilized antioxidant reactivity data to create a structure-based machine learning (ML) model to predict hydrogen peroxide radical scavenging rates from molecular parameters.

To test and train such a structure-based ML model, we compiled a dataset of 563 data points for the rate constant of antioxidants, each representing a unique combination of a molecule name, reaction mechanism, and pH. We then converted the molecular names to SMILES (Simplified Molecular Input Line Entry System). Using the python library Mordred, we were able to extract 1621 descriptors, each of which describe a chemical and/or topological property of the molecule. Mordred was chosen due to its computational speed and accuracy compared to alternative descriptor calculators and it being open source.[3] After this we performed feature engineering to get rid of highly correlated and irrelevant features. Then we were able to put the selected features into machine learning models in order to predict antioxidant activity.

Project Goal

The goal of our project is to identify and build a data-driven model to quantify and/or rank existing and new antioxidants in terms of their activity against the hydrogen peroxide free hydroxyl radical ($\cdot OH$) (or more generally $\cdot OR$) generated via the Fenton and Fenton-like reactions in low pH

environments. To achieve the goal, we divided the development into three tasks. Task 1: Compile all available literature data on antioxidant reactivity. Task 2: Create a structure-based ML model to predict the hydrogen peroxide radical scavenging rate of various antioxidants. Task 3: Gain nontrivial insights into what aspects of the chemistry most affect reaction rates.

Computational Approach

Data Collection

Data was collected from Dorfman & Adams at NIST[1], where they obtained the rate constants from various literature sources. They normalized the rate constants were normalized to reference reactions, so the provided rate constants provided are not what were in the original citations. However, the normalized data is more useful than the original data because the normalized data compares different molecules without regard to experimental conditions. Figure A2 shows a sample of what the data looks like. The data was exclusively at or within a few degrees of room temperature, therefore temperature is not a factor we are considering in this project. For a given combination of molecule name, pH, and reaction mechanism, only one data point was taken, and when multiple values were present they were averaged, and outliers were removed. Data that was incomplete or had a range of pH values $>\pm 1$ were removed. The curated dataset consists of 563 unique entries, each comprising a reactant name, its Canonical SMILES representation, the pH at which the reaction occurs, the associated reaction mechanism type, and the experimentally determined rate constant. A summary of the distribution of pH values, k values, reaction mechanisms, and listed chemical classes shown in the dataset are shown in Figures A3, A4, A5, and A6, resepctively.

SMILES Retrieval

The molecule name had to be converted into a machine-readable format, and SMILES was chosen as it is used in many libraries. We used the python library RDKit to automatically get the SMILES from the molecule name for most of the molecules. However, for some entries in the dataset RDKit did not automatically generate SMILES representations, so Pubchem[4] and Chemical Structure Search[5] were used to generate SMILES. SMILES were then converted to canonical SMILES to ensure every molecule has a unique SMILES representation.

Our project is concerned with the effects of pH on antioxidant activity, so the Dimorphite-DL python library was used to generate SMILES of different protonation states of molecules that exist in different pH values.[6] In reality, only one protonation state will dominate in most cases for a molecule in a given pH solution, however there is uncertainty in both the pH measurements ($\pm$ .5) of our data and the pKa values estimated by Dimorphie-DL, so expanding the dataset to include multiple possible protonation states can lead to a better performing model. We used a pKa precision factor of 1, meaning that we consider estimated pKa values to have an uncertainty of ± 1 standard deviation. Dimorphite-DL returns the top x most likely present protonation states, where you can choose x.

Descriptor-Based Feature Extraction

To extract chemically meaningful features from the SMILES representations, we used the descriptor library Mordred, which generated 1621 descriptors per molecule, and provided a rich and diverse set of features capturing topological, electronic, and steric properties. We attempted to use RDKit descriptors, atomic level encodings, character level encodings, and molecular transfer embeddings in addition to using Mordred descriptors, but none of those performed as well as solely using Mordred, as adding additional descriptors likely led to overfitting and redundancy.

Feature Engineering and Dimensionality Reduction

For each subset of the data, we attempted to find the optimal features to remove. We thought to remove highly correlated features and features that contain mostly the same values, however what is considered "highly correlated" and "mostly the same" values is subjective. For each subset of the data, we first removed features with a correlation >.95. We then attempted to see if we removed features that contained over a threshold percentage of the same values, and we varied that threshold from 60% to 100% with intervals of 5%. In parallel, we ran the model to test the optimal threshold for dropping correlated features that are correlated above a certain threshold. We tested thresholds of 1, .999, .99, .98, .95, .90, .85, .80, .75, .70, .65, .60, .55, and .50. Once we found an optimal value of features with similar values to remove and the optimal amount of correlated features to remove, we then attempted to find the optimal amount of protonation states to use. We tested not accounting for protonation states at all, and using the top 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 16, 32, 64, and 128 protonation states. Varying the optimal amount of correlated features to remove, similar features to remove, and number of top protonation states to use greatly affected model performance.

Model Development and Evaluation

The rate constant was transformed using a base 10 logarithmic scale to normalize its distribution. The final dataset included approximately 200 features per sample and was used to train regression models, including linear regression, lasso and ridge regression, random forest regression, and XGBoost regression. Zhang et al.[7] had success in predicting the rate constants of alkanes using XGBoost. We tested the models using both PCA and not using PCA.

Each model was trained on an 80% training split and evaluated on a 20% test split using standard metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 score. Hyperparameter tuning was performed using GridSearchCV to optimize performance.

Feature importance scores were derived from the Random Forest model to identify the most influential descriptors. Additionally, we built and evaluated models under specific chemical constraints—such as pH ≤ 3 and aliphatic alcohol only subsets to understand how these factors influence predictive accuracy.

Results

For the entire dataset, our base model involved removing features with correlation above .95, not utilizing protonation states, and not removing features that contained mostly similar values, and got an R^2 of .66. We then found that it is optimal to remove features that contain over 80% constant values where we got an R^2 of .73. In parallel, we found that it is optimal to one of the features from a pair of features that are correlated above .999, and got an R^2 of .76. Combining the improvements of removing features that contain over 80% constant values and features that are correlated above .999, we got an R^2 of .82. We found that 128 was optimal value for the number of top protonation states to use, leading us to the best result of an R^2 of .91 and RMSE of .27.

Among the trained models, for most of the subsets of the data, tree-based regressors outperformed linear models, except for the smallest subsets of data such as the aliphatic alcohols (N = 73) and carboxylic acids (N = 100). For the dataset as a whole, Random Forest and XGBoost models showed superior predictive power with lower error metrics and higher R^2 scores. Figure 1 shows the results of the best combination for the overall data, and PCA was not used. RFTop, which is a specifc type of Random Forest where the top 50 random forest features are inputted into the model, resulted in the best performance where the R^2 is equal to .91 and the RMSE is equal to .27.

	Model	MAE	RMSE	R² Score
0	Linear Regression	12360.412480	182508.200012	-4.071238e+10
1	Lasso Regression	0.371459	0.508251	6.842678e-01
2	Ridge Regression	0.234013	0.379773	8.237174e-01
3	Random Forest	0.122394	0.282609	9.023809e-01
4	XGBoost	0.119396	0.288518	8.982565e-01
5	Hyperparameter Tuned RF	0.130541	0.281266	9.033070e-01
6	RF Top	0.116985	0.269433	9.112718e-01

Figure 1: Results of the overall dataset by type of ML model. RFTop resulted in the best performance, with a R^2 of .91 and a RMSE of .27.

One concern by using so many protonation states could be that our dataset then gets skewed towards including more biomolecules which are the only molecules in our dataset which could have more than 8 protonation states. However this is not the case, when we compare the results of the molecules that are not biomolecules in the model with the top 10 protonation states vs the top 128 protonation states, the MAE decreases from .28 to .24. We are interested in knowing which molecules the model works for and which molecules the model does not work for. To assess the precision of the model in individual predictions, we compared the predicted values $\log_{10}(k)$ against the experimental values, as shown in Figure 2. The data points are mostly closely clustered around the red identity line, indicating a strong correlation between the predicted and actual rate constants throughout the full range of reactivity. This agreement reinforces that the model is not overfitting too much, and generalizes well across a variety of reaction and molecule types. However, there are some chemical classes that perform significantly better or worse than others, as shown in Figure 3.

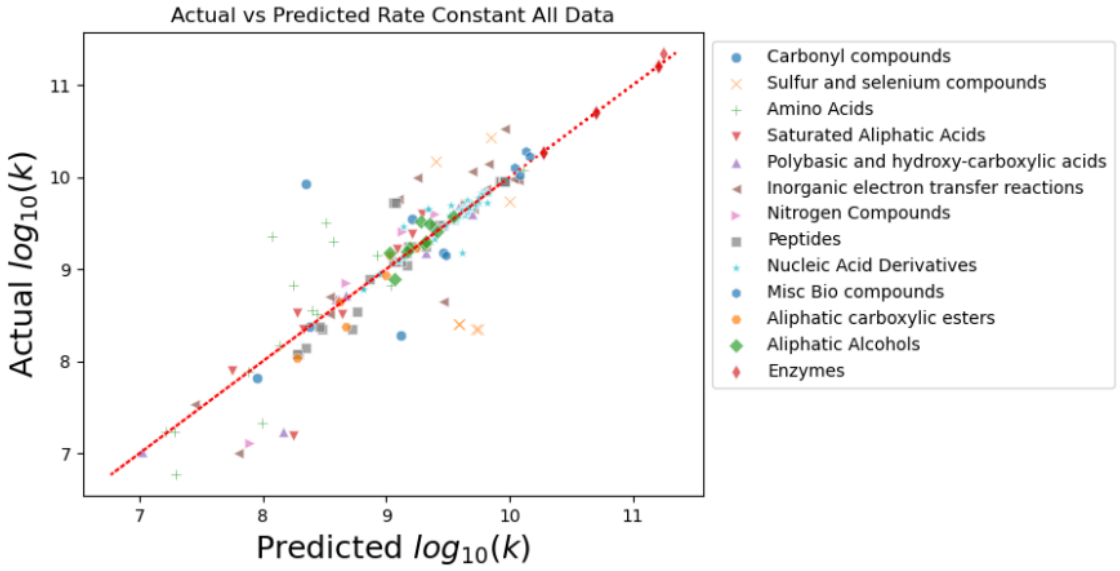

Figure 2: Predicted vs actual rate constants for the testing data of the entire dataset $R^2 = .91$

Molecule Type	Average Error	Number of Molecules in the Testing Dataset
Sulfur and selenium compounds	0.76	9
Carbonyl compounds	0.53	6
Nitrogen Compounds	0.36	4
Inorganic electron transfer reactions	0.34	14
Amino Acids	0.30	19
Saturated Aliphatic Acids	0.28	8
Unspecified Molecule Type	0.14	18
Misc Bio compounds	0.13	5
Polybasic and hydroxy-carboxylic acids	0.13	12
Peptides	0.11	30
Aliphatic carboxylic esters	0.11	7
Aliphatic Alcohols	0.08	10
Nucleic Acid Derivatives	0.01	86
Enzymes	0.001	76

Figure 3: MAE for testing data of the entire dataset by reaction type

Since the overall model applies to various different molecule types, we decided to test the results if we limited the data to just a subset of the data, to see if we can get improved results if we know the molecule type or reaction mechanism we want our antioxidant to be. We tried various subsets of data, and all of our attempts are shown in Figure 4. For each subset, we had to find the optimal amount of correlated features to remove, similar features to remove, and number of top protonation states to use from scratch. We first performed one layer of filtering, which was limiting the data that was believed from NIST to undergo hydrogen abstraction, addition, and pH ≤ 3. For the addition reactions, we then further limited it to Nucleic Acids which was the largest molecule type

within addition. For the addition data as a whole, it was optimal to not use any protonation states, but for the nucleic acid data it was optimal to use the top 128 protonation states. This is why there is more data for the best model for the nucleic acids despite nucleic acids being a subset of the addition reaction data. For the hydrogen abstraction, we further limited the dataset to pH $\leq$ 3, aliphatic alcohols, and carboxylic acids. Limiting the dataset and building models exclusively on the limited dataset gives us two takeaways. One, we can get better R^2 for subsets of the data than the entire dataset. More importantly, we get improved results for building a model with just the subset than the subset would perform in the model of the entire dataset. For example, Arkema is interested in hydrogen abstraction at pH $\leq$ 3 and that subset has a MAE of .177 in the model for the entire dataset, however for the model of the subset, it has a MAE of .143, which is a significant improvement. This result illustrates that a subset of the data can perform better in a reduced model than in the model for the entire dataset.

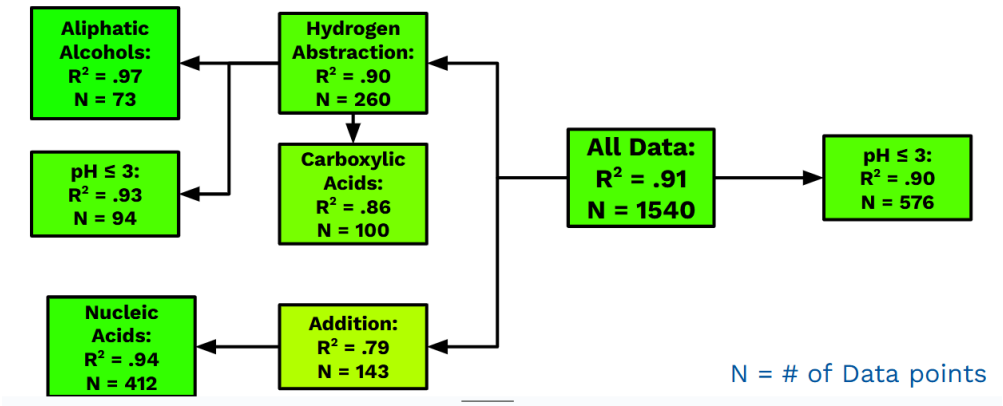

Figure 4: Summary of results from different subsets of the data

Conclusion

We have developed a structure-based ML model using Random Forest that given a SMILES string, will be able to predict the rate constant of an antioxidant scavenging the free hydroxyl radical ($\cdot$OH) generated via Fenton and Fenton-like reactions, with reasonable accuracy (R^2 = .91). The model works for various reaction types and model accuracy can improve if the scope of chemical classes are narrowed. The model accuracy is greatly affected by proper selection of the amount of correlated features to remove, similar features to remove, and number of top protonation states to use. In the future, a proper design of experiments approach could be implemented to choose an even better combination of those hyperparameters. Some limitations of our model are that our data are sparse, dated, and have significant uncertainty. The model can potentially be improved if Density Functional Theory (DFT) calculations are performed on each of the molecules, and the results are incorporated in the model. The most significant limitation of this project is that the data only applies to hydroxyl radical scavenging, when the Fenton reaction also produces $\cdot$OOH and $\cdot O_2^-$, which we did not consider in this project. Overall, this approach can potentially applied to other attempts at using structure-based ML models to predict chemical rate constants.

Acknowledgments

We greatly appreciate the guidance and support from our industry mentors: Dr. Brian Koo, Dr. Christian Becker, & Dr. Katie Daisey on behalf of Arkema S.A. We also deeply grateful for Professor Arthi Jayaraman at the University of Delaware for offering a course (CHEG/CISC/ECEG/MSEG 848) that involves industry collaborations, and for providing guidance and support throughout the semester.

References

1. Dorfman, L. M.; Adams, G. R. Reactivity of the Hydroxyl Radical in Aqueous Solutions. 1973. https://doi.org/10.6028/nbs.nsrds.46.

2. Koo B.; Becker C. Predicting Antioxidant Activities. Presented at the University of Delaware, Newark, DE, February 5, 2025.

3. Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. Journal of Cheminformatics 2018, 10 (1). DOI:10.1186/s13321-018-0258-y.

4. PubChem. https://pubchem.ncbi.nlm.nih.gov/ (accessed 2025-02-17).

5. Chemical Structure Search. https://www.fishersci.com/us/en/search/chemical/substructure.html (accessed 2025-02-17).

6. 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). DOI:10.1186/s13321-019-0336-9.

7. Zhang, Y.; Yu, J.; Song, H.; Yang, M. Structure-Based Reaction Descriptors for Predicting Rate Constants by Machine Learning: Application to Hydrogen Abstraction from Alkanes by CH3/h/O Radicals. Journal of Chemical Information and Modeling 2023, 63 (16), 5097–5106. DOI:10.1021/acs.jcim.3c00892.

8. Descriptor List — mordred 1.2.1a1 documentation. Github.io. https://mordred-descriptor.github.io/documentation/master/descriptors.html (accessed 2025-04-17).

A Appendix A: Additional Figures

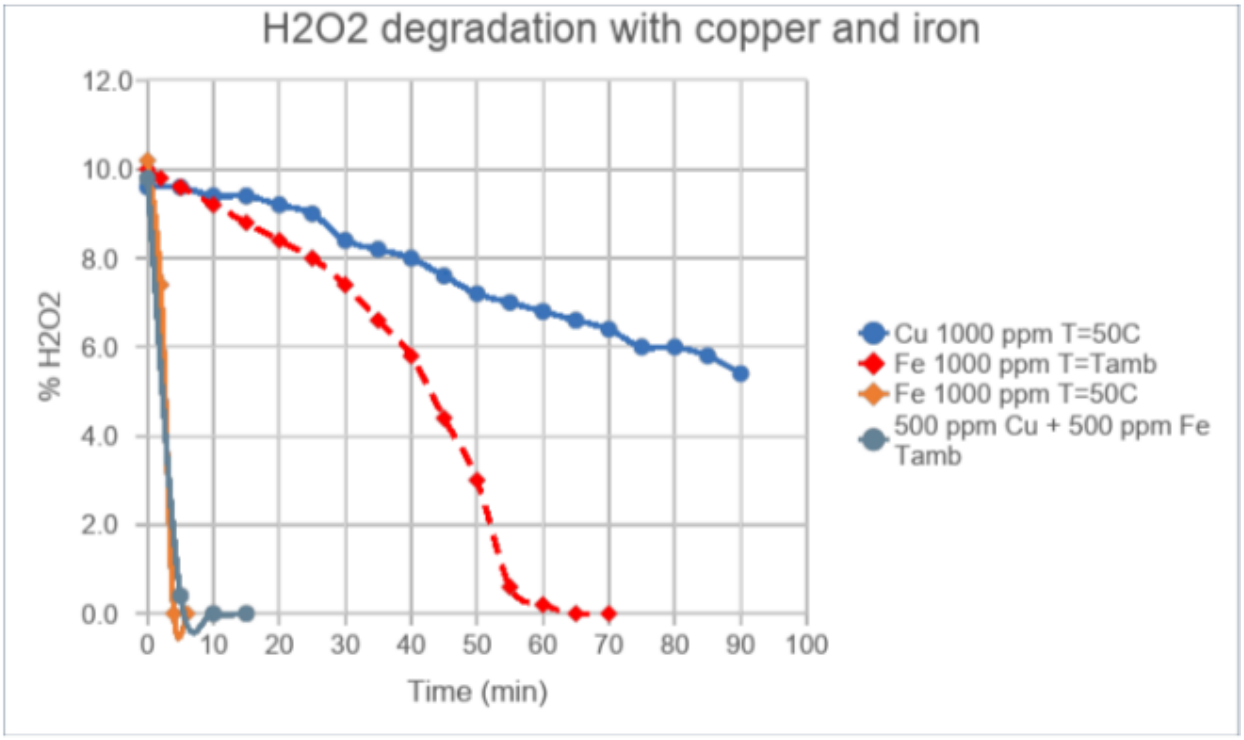

Figure A1: From Arkema on 2/5/25: H_2O_2 degradation over time in the presence of Copper and Iron.[1]

TABLE 2. *Addition reactions*

Reactant	$k(M^{-1}\ s^{-1})$	T(°C)	pH	Ref.	Method	Comments
acetanilide	5.0×10^9	25	9	[7]	S irrad. with PNDA	Rel. to $k_{OH+PNDA}=1.25\times10^{10}$
acetophenone	†*$(6.5\pm0.7)\times10^9$	RT	7	[8]	PR prod. form.	Absolute. N_2O added. H-corr.
	4.8×10^9	25	9	[7]	S irrad. with PNDA	Rel. to $k_{OH+PNDA}=1.25\times10^{10}$

Figure A2: Snippet of NIST data.[2] * = absolute values independent of reference rate constant; †= uncertainty $\leq$ 30%.

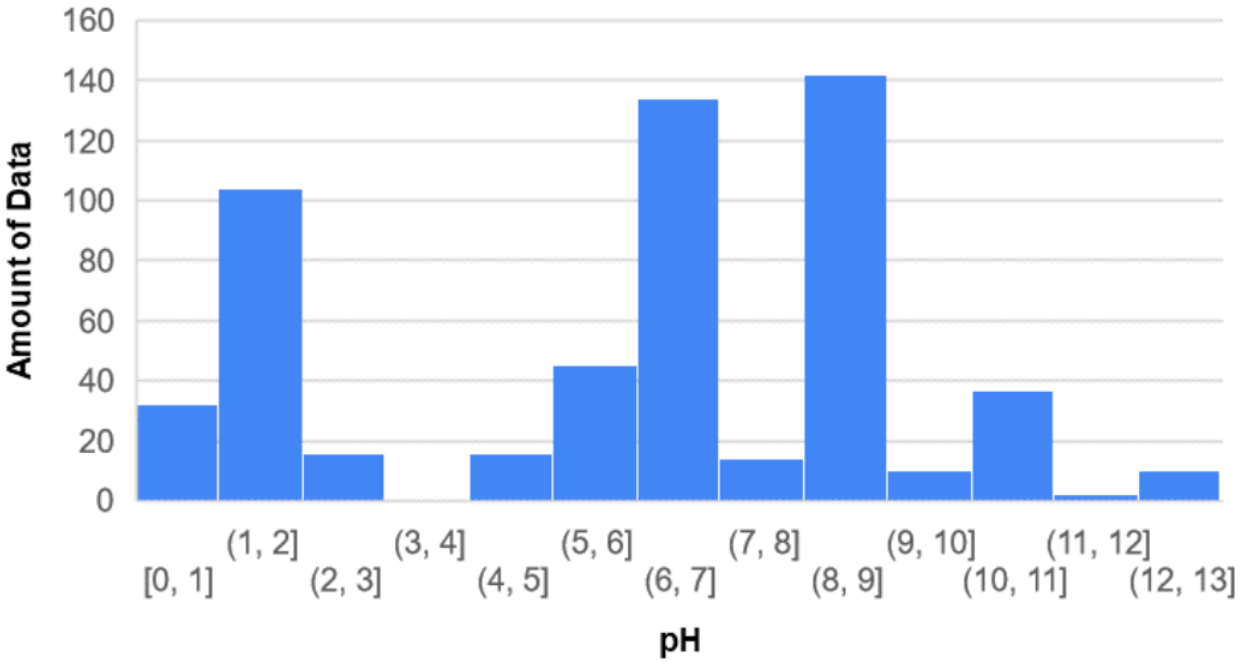

Figure A3: pH value distribution of our data

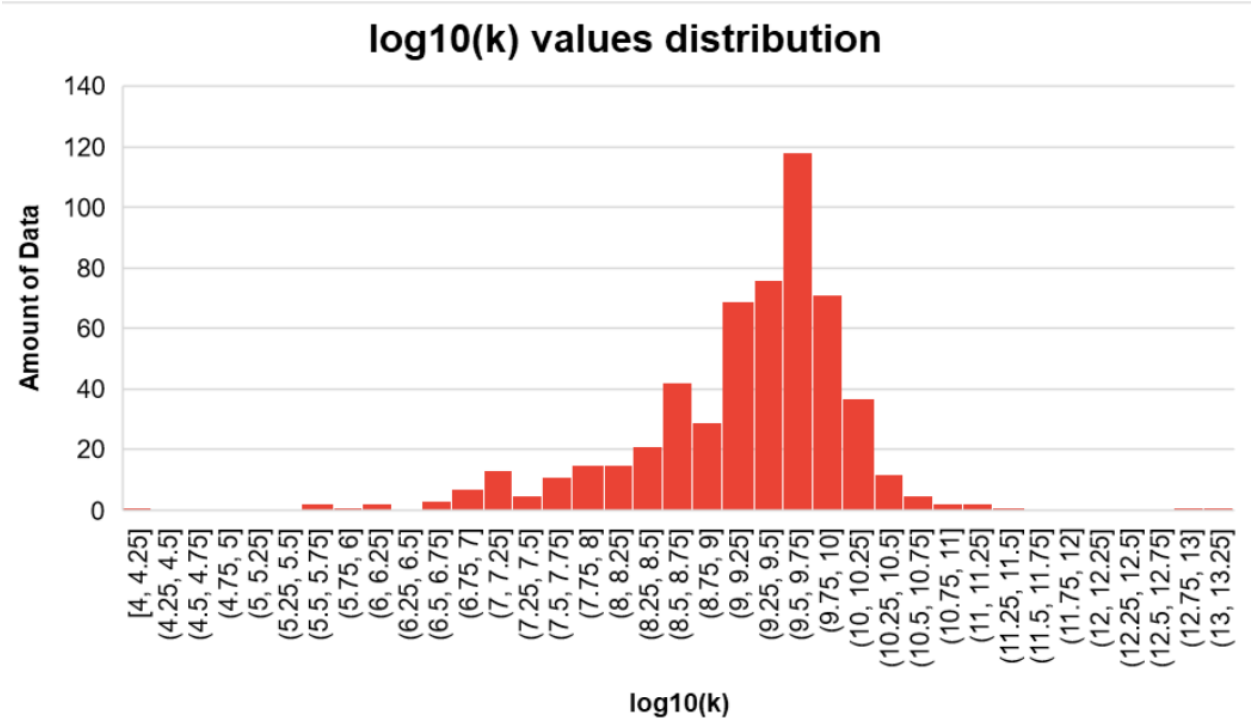

Figure A4: $\log_{10} k$ distribution of our data

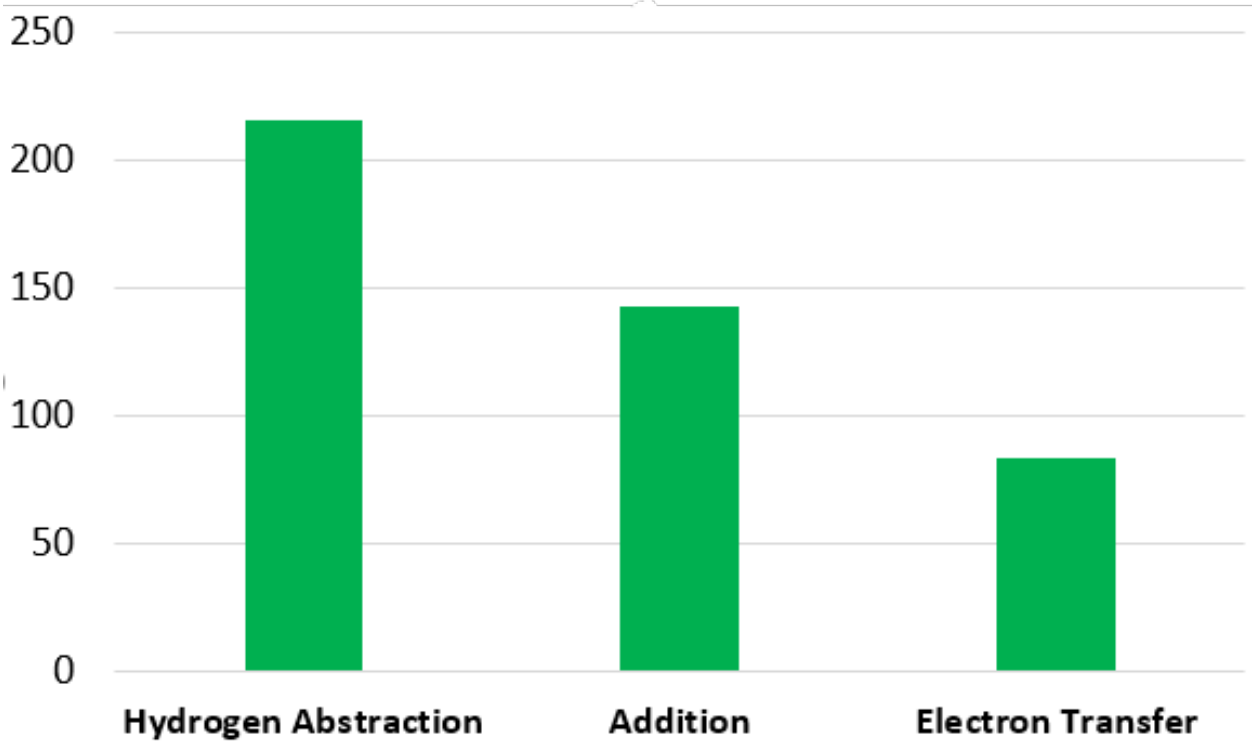

Figure A5: Amount of data by confirmed mechanism. Note that not all molecules in the dataset have a confirmed mechanism.

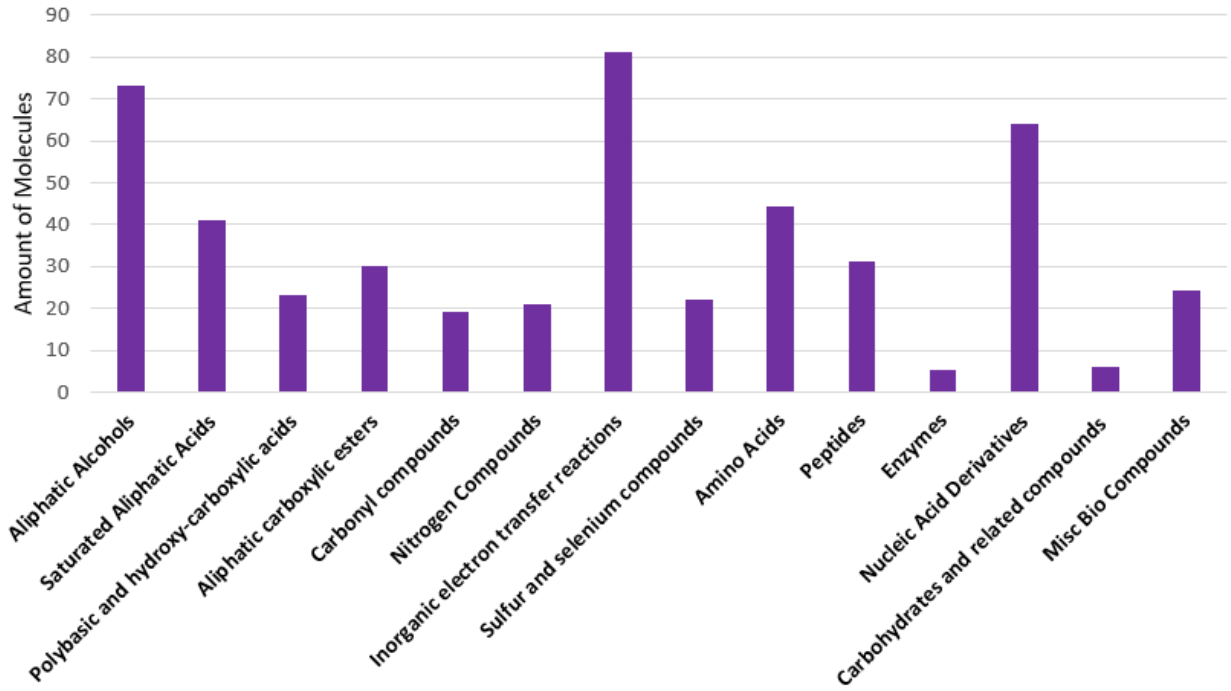

Figure A6: Data by molecule type (as sorted by NIST)

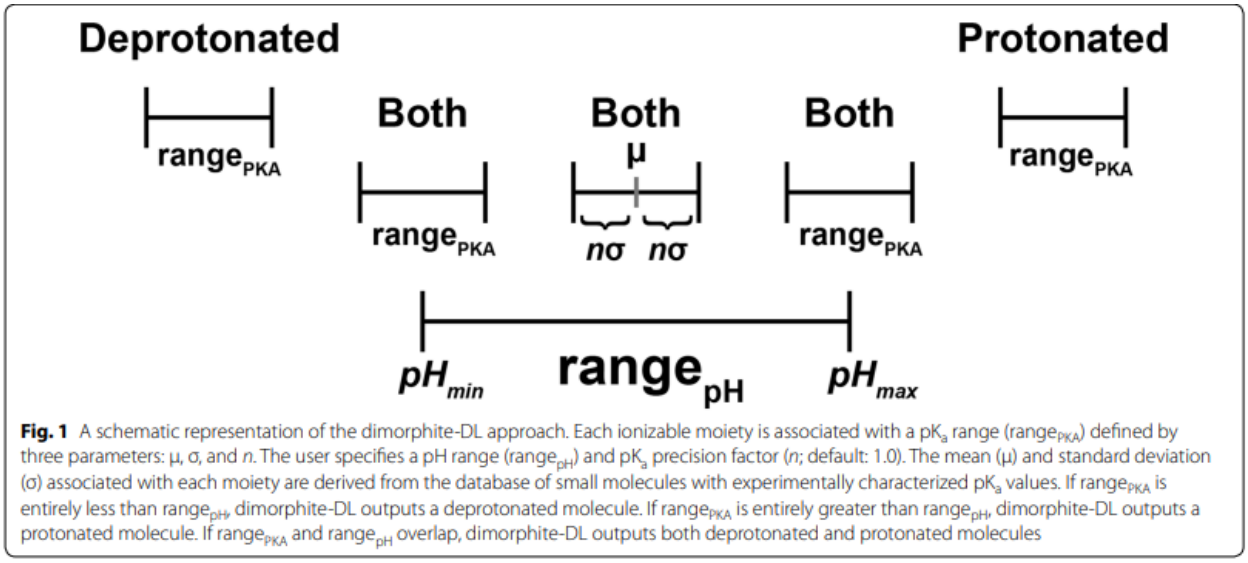

Fig. 1 A schematic representation of the dimorphite-DL approach. Each ionizable moiety is associated with a pK_a range ($range_{PKA}$) defined by three parameters: μ, σ, and *n*. The user specifies a pH range ($range_{pH}$) and pK_a precision factor (*n*; default: 1.0). The mean (μ) and standard deviation (σ) associated with each moiety are derived from the database of small molecules with experimentally characterized pK_a values. If $range_{PKA}$ is entirely less than $range_{pH}$, dimorphite-DL outputs a deprotonated molecule. If $range_{PKA}$ is entirely greater than $range_{pH}$, dimorphite-DL outputs a protonated molecule. If $range_{PKA}$ and $range_{pH}$ overlap, dimorphite-DL outputs both deprotonated and protonated molecules

Figure A7: Schematic of Dimorphite-DL library for protonation state prediction.[6]