HEPOM: Using Graph Neural Networks for the accelerated predictions of Hydrolysis Free Energies in different pH conditions.

Rishabh D. Guha,[†,‡,△] Santiago Vargas,[¶,△] Evan Walter Clark Spotte-Smith,[†,§] Alexander Rizzolo Epstein,[‖] Maxwell Venetos,[†,§] Ryan Kingsbury,[⊥] Mingjian Wen,[#] Samuel M. Blau,[@] and Kristin A. Persson[*,§,†]

†*Materials Science Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA, 94720 USA*

‡*Present Address: Schrödinger Inc. 1540 Broadway, New York, NY 10024, USA*

¶*Chemical Sciences Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA, 94720 USA*

§*Department of Materials Science and Engineering, University of California, Berkeley, 210 Hearst Memorial Mining Building, Berkeley, CA, 94720 USA*

‖ *Yusuf Hamied Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, UK*

⊥*Department of Civil and Environmental Engineering, Princeton University*

#*Department of Chemical and Biomolecular Engineering, University of Houston, Houston, Texas 77204, United States*

@*Energy Storage and Distributed Resources, Lawrence Berkeley National Laboratory*

△*These authors contributed equally to this work*

E-mail: kapersson@lbl.gov

https://doi.org/10.26434/chemrxiv-2024-2v1nx **ORCID:** https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Abstract

Hydrolysis is a fundamental family of chemical reactions where water facilitates the cleavage of bonds. The process is ubiquitous in biological and chemical systems, owing to water's remarkable versatility as a solvent. However, accurately predicting the feasibility of hydrolysis through computational techniques is a difficult task, as subtle changes in reactant structure like heteroatom substitutions or neighboring functional groups can influence the reaction outcome. Furthermore, hydrolysis is sensitive to the pH of the aqueous medium, and the same reaction can have different reaction properties at different pH conditions. In this work, we have combined reaction templates and high-throughput *ab initio* calculations to construct a diverse dataset of hydrolysis free energies. The developed framework automatically identifies reaction centers, generates hydrolysis products, and utilizes a trained Graph Neural Network(GNN) model to predict ΔG values for all potential hydrolysis reactions in a given molecule. The long-term goal of the work is to develop a data-driven, computational tool for high-throughput screening of pH-specific hydrolytic stability and the rapid prediction of reaction products, which can then be applied in a wide array of applications including chemical recycling of polymers and ion-conducting membranes for clean energy generation and storage.

1. Introduction

Water is one of the most essential molecules in chemistry, and yet, its unique properties make it notoriously difficult to characterize.[1,2] The significant electronegativity differences between its oxygen and hydrogen atoms gives water a highly polar character that leads to its recognition as the "universal solvent".[3,4] Hydrolysis, or any reaction where water acts as both a reactant and the solvent medium,[5,6] are a prevalent class of reactions across chemistry. Hydrolytic reactions are fundamental in biological[7,8] and synthetic chemistry[9,10] and play a critical role in various essential scientific processes and significant technological applications.

https://doi.org/10.26434/chemrxiv-2024-2v1nx **ORCID:** https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

These range from processes, such as human digestion,[8,11] where enzymes facilitate the hydrolytic breakdown of complex macronutrients into simpler compounds, to the degradation of hazardous pollutants[12] and alternative plastic chemistries.[13]

At the molecular level, hydrolysis begins when a water molecule attacks specific sites on the reactant, initiating a sequence of bond cleavages and formations that lead to new product(s). The mechanism and the associated rate of this reaction is closely tied to the pH of the aqueous medium.[14,15] The availability of protons (H^+) or hydroxide (OH^-) ions catalyzes the formation of charged species, which have markedly different reactivities compared to their neutral counterparts.[13,16] These ionized reactants can exhibit enhanced solubility[17,18] by forming stronger hydrogen bonds with the solvent. Additionally, water can act as catalyst, facilitating ion transfer through the solvent and creating alternate reaction pathways with lower energy barriers.[19,20] As a result, acid/base-catalyzed hydrolysis of the same reactant can have significantly different reaction rates compared to its neutral form, adding complexity to the study of these reactions.

Given activation barriers ($\Delta G^{\ddagger}$), the experimental rate of a hydrolysis reaction can be directly correlated via the Eyring equation.[16,21,22] This involves determining computationally-intensive and difficult to find transition states for each individual reaction along the reaction coordinate of the potential energy surface (PES).[16,23,24] In contrast, within a specific reaction family, the Bell-Evans-Polanyi principle (BEP)[25] can offer a qualitative linear correlation between the thermodynamic Gibbs Free Energy change (ΔG_r) and the kinetic parameter $\Delta G^{\ddagger}$.[26–28] In cases where it holds, BEP allows us to leverage the thermochemistry of products and reactants (ΔG_r) to approximate trends in the kinetic rates of the reaction. This opens the avenue for the development of a computational screening tool that can calculate the respective ΔG_r's of all potential hydrolysis pathways and screen molecules for a specific hydrolysis-related application. Despite this, quantifying this thermochemical quantity (ΔG_r), with high accuracy still requires DFT calculations with large basis sets and refined hybrid functionals at both reaction endpoints.[29,30] Depending on the size of molecules, these

https://doi.org/10.26434/chemrxiv-2024-2v1nx **ORCID:** https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

calculations can take anywhere from several hours to days, particularly when employing implicit solvent models[31] to approximate the contributions from the reaction environment.

Since computational cost is a severe bottleneck for any form of high-throughput screening, deep learning approaches have emerged as promising alternatives in the past decade, especially for tasks that involve the establishment of structure-to-property relationships.[32,33] Recently, graph convolutions, which iteratively update node and edge features based on connectivity and local environment, have proven to be extremely effective in learning molecular[34,35] and reaction representations.[36,37] Despite these methodological advances, the largest roadblock to the development of an accurate model is typically the procurement of diverse, representative data. For instance, the model developed by Grambow et al.[33] was facilitated by a dataset of 12,000 gas-phase reactions[38] sampled from a subset of molecules in the GDB-17 dataset.[39] The bond dissociation energy (BDE) prediction framework developed by Wen et al.[40] was trained on a dataset of over 60,000 homolytic and heterolytic bond dissociation reactions.[41] In the realm of hydrolysis, no such comprehensive dataset currently exists.

In this work, we have attempted to address these shortcomings by first developing a predictive framework based on reaction templates for different functional groups that can automatically generate hydrolysis products for multiple pathways in any molecule. This framework was then applied on a subset of the `QM9`[42] and the `Alchemy`[43] databases to generate a database of over 65,000 hydrolysis reactions in an implicit aqueous solvation environment. For a given reactant molecule in the `QM9` subset of the data, we have also generated corresponding hydroxylated and protonated states of the reactant molecule to approximate the effects of extreme pH on the ΔG_r of hydrolysis. In addition, the neutral fold of the dataset was developed with reactants from the `QM9` database and later augmented with the inclusion of larger reactant molecules from the Alchemy[43] dataset. Combined, we provide a new dataset that encompasses thermodynamic properties at different protonation states along with an exploratory set of neutral-pH reactions for analysis and model development.

We then proceeded to use this comprehensive dataset to train a GNN model, which

https://doi.org/10.26434/chemrxiv-2024-2v1nx **ORCID:** https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

serves as a **H**ydrolysis **E**nergy **P**redictor for **O**rganic **M**olecules (*HEPOM*). The model leverages the difference features of the atom (node), bond (edge) and global features between the products and the reactants to directly predict the DFT-calculated ΔG_r . The global reaction atom mapping allows the model to simultaneously track multiple bond dissociations and formations. For the neutral dataset, the model achieved a low mean absolute error (MAE) of 1.73 kcal/mol on a diverse holdout set of hydrolysis reactions and it was also successful in outperforming a diverse set of benchmark models on the smaller and more complex protonated and hydroxylated datasets.

2. Methods

2.1 Reaction Generation

As mentioned, we segmented the construction of our dataset into four main parts: three derived from the QM9 dataset (representing neutral, protonated, and hydroxylated reactions) and another, neutral reaction set from the Alchemy dataset. Hydrolyzable molecules in `QM9` were screened using `RDKit`[44] substructure matching for 20 standard, hydrolyzable functional groups (Figure S3(b)). These templates were adapted from the work by Tebes-Stevens et al.[45] and integrated into an automated framework to predict reaction products. For instance, in *Figure 1*, if an ester functional group is detected in a molecule, the reaction template would yield a carboxylic acid and an alcohol as the respective hydrolysis products. Bond 'a' in the reactant and bond 'b' in the water molecule is deleted with the `RemoveBond` functionality in `RDKit`. Then, `AddBond` is used to create bonds 'c' and 'd' between atoms R1-W2 and R2-W3 respectively, to yield a carboxylic acid and an alcohol as the respective products. Similar reaction templates were implemented for all functional groups. Nitriles are treated differently: the reaction template yields amides and these can be further hydrolyzed into an amine and a carboxylic acid (*Figure S1*). As a result, the intermediate products of nitrile reactions serve as reactants in additional hydrolysis reactions, thereby augmenting

https://doi.org/10.26434/chemrxiv-2024-2v1nx **ORCID:** https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

the dataset. We generated a total of **16,264** hydrolysis reactions from the `QM9` dataset. An initial model was trained on **15,264** of these neutral reactions, while **1,000** reactions were kept in an unseen holdout test set. The performance of this model is discussed in Section 3.2 *(vide infra)*.

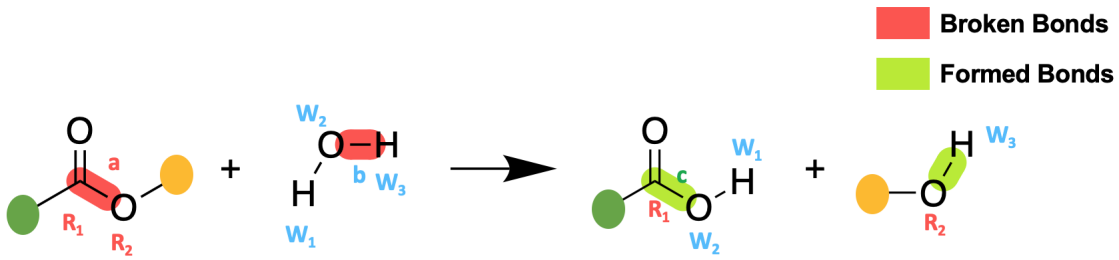


Figure 1: The set of bond cleavage and formations necessary to generate hydrolysis products for a representative ester molecule.

The broader goal of this work is to develop a framework capable of enumerating potential hydrolysis pathways for a wide range of molecules and predicting the thermodynamic free energies of these pathways with high accuracy. With this in mind, we screened molecules from the `Alchemy` dataset[43] to generate reactions representing larger, more complex reactions. The `Alchemy` dataset includes molecules with up to 14 heavy atoms, though we only considered molecules with more than 10 heavy atoms. Neutral reactions generated from the `Alchemy` dataset were added to the original dataset, resulting in a combined dataset of **41,006** reactions. Of these, **2,800** reactions were filtered out and added to the previous `QM9` test set, creating an unseen holdout test set of **3,800** reactions (10% of the training set size). In the following sections of the manuscript, we refer to the original dataset as the `QM9` dataset and the expanded dataset as the `QM9+Alchemy` dataset.

While the data generated above are useful and novel, they are examples of hydrolysis in a neutral reaction medium. However, hydrolysis is often catalyzed in an acidic or basic reaction medium. For example, the hydrolysis rate of amides in a neutral medium is negligible, even after heating, but amidic hydrolysis proceeds at a moderate rate in an acidic or basic medium,[46,47] forming carboxylic acid and an amine. Consequently, we explored whether this

https://doi.org/10.26434/chemrxiv-2024-2v1nx **ORCID:** https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

framework could extend to broader reaction conditions, potentially serving as a screening tool for identifying molecules amenable to pH-specific hydrolysis. It is important to clarify that under these reaction conditions, hydrolysis is initiated by the protonation or hydroxylation of the reacting functional group,[20,48] and the overall reaction rate is heavily influenced by the pKa values of the functional groups.[49,50] Generating a diverse dataset in high throughput for acid- or base-catalyzed hydrolysis while accounting for pKa was intractable with our current data generation scheme. Therefore, in this work, we focused our efforts on developing a unified model that can predict the differences in hydrolysis reaction-free energies for the same functional group under different pH conditions. The datasets generated for neutral or basic pH assume that the reaction medium is at an extreme pH, i.e., if a functional group can protonate or hydroxylate, it will. An alternative approach could involve applying a separate ML model which predicts the pKa and identifies the most probable site for protonation or hydroxylation at a specific pH,[51–53] but this is beyond the scope of the present study.

We separated extreme pH hydrolysis reactions into two reactions schemes. For an acidic medium, the reacting functional group was assumed to be protonated at the most electron-rich atom site (e.g., the carbonyl oxygen in an ester or amide, or the nitrogen atom in a nitrile). Similarly, for a basic pH, the relevant atom site in the functional group moiety was hydroxylated. The acidic pH reaction was then executed between the protonated reactant and two water molecules to maintain reaction stoichiometry. A representative example elucidating the differences in the hydrolysis reaction in acidic and neutral pH for a hydrolyzing carbamate molecule is demonstrated in *Figure S2* (a) and (b) of the SI. The extra water molecule on the reactant side absorbs the proton to generate hydronium as one of the reaction products. This approach circumvents the erroneous DFT-calculated energies of an isolated proton in an implicit solvent medium.[54] In the case of basic pH, the hydroxylated reactant decomposes into the reaction products and a hydroxide ion. For these two datasets, we focused on the QM9 molecules to limit the scope of computations, yielding a protonated dataset of **11,323** reactions and a hydroxylated dataset of **16,732** reactions. Holdout test

https://doi.org/10.26434/chemrxiv-2024-2v1nx **ORCID:** https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

sets consistent with the neutral QM9 dataset were also extracted before model training. Since the protonated reactants have a +1 charge and the hydroxylated reactants a -1 charge, we will refer to these datasets as $\texttt{QM9}^{+}$ and $\texttt{QM9}^{-}$, respectively, in the subsequent sections.

2.2 Density-Functional Theory

QChem (version 5 or 6)[55] was used to perform all the DFT calculations necessary to generate the dataset. A specialized frequency-flattening optimization (FFOpt) workflow, originally developed by Spotte-Smith et al.[41] and currently implemented in atomate[56] was used to optimize the reactant and product structures to a true minima and also obtain thermochemical quantities from the vibrational frequencies. The workflow iteratively performs successive geometry optimizations and frequency calculations until there are either none or a single negligible negative frequency ($<$15 cm^{-1}). This approach ensures that the optimized structure is a true local minimum of the PES and not a saddle point. Moreover, the workflow parses the necessary enthalpy and entropy terms from the QChem frequency output document for the free energy calculations. For all the DFT calculations, we used the range-separated meta-GGA hybrid functional, ωB97M-V,[57] which employs the VV10 dispersion correction,[58] to improve the non-covalent interactions. The def2-SVPD basis set[59] was employed for the FFOpt workflow and the solvation effects were implicitly accounted for with the water SMD solvent model.[17] The electronic energies of the optimized structures were refined with single-point calculations using a larger def2-QZVPPD basis set.[59]

2.3 Model Architecture

The GNN model(Figure 2) is based on the previous BonDNet architecture.[40] Here we briefly discuss that architecture before highlighting our key departures from this model, and how these differences are key to working on each datasets presented.

The original algorithm uses gated graph convolutional (GatedGC) layers to propagate initial node features within the graphs of individual species on both sides of a reaction.

https://doi.org/10.26434/chemrxiv-2024-2v1nx ORCID: https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

While GatedGC layers were used widely for structure-to-property models in chemistry and materials science,[60,61] BonDNet improved on these previous implementations by integrating update and message-passing equations between global nodes and atom/bond type nodes; this allows for the treatment of species of different charges and provides a framework to include molecular-level features. Similar to other graph neural networks, more distant graph relationships are treated by iteratively stacking several (typically 2-4 layers) GatedGC layers. With updated species graphs, reaction graphs are built to hold reaction feature differences - atom and bond nodes are mapped to each other on both sides of a reaction and their features are subtracted between corresponding atoms/bonds. Broken bonds are represented by zero vectors in this scheme. Here BonDNet implemented a custom set2set[62] global pooling feature to map reaction graphs to fixed-sized vectors. These vectors are passed to fully-connected layers for property prediction.

Our implementation extends global pooling by integrating a diverse set of global pooling functions, including `set2set`,[62] `WeightedMeanPooling`, `Self-attention pooling`,[63] and `Mean Pooling`. This diverse set of global pooling functions was intended to provide a more comprehensive set of architectures across different dataset sizes, as previous benchmarks showed `set2set` layers did not always outperform simpler `MeanPooling` approaches.[64]

In this implementation, the reaction mapping is altered from the original BonDNet to a global reaction graph is constructed between the union set of bonds in products and reactants. Originally, BonDNet used the product graph as a scaffold and subtracted reactant features from corresponding nodes in the product graph. This limited the model to only being applicable for $A \rightarrow B$ and $A \rightarrow B + C$ type reactions with a single bond dissociation. The previous framework could not interpret a hydrolysis reaction that involves at least two elementary bond dissociation and formation reactions. Our algorithm builds a global reaction graph by taking the union set of atoms and bonds in products and reactants and uses this to build a graph structure with bonds from each side of reaction. This change allows for an arbitrary number of bond changes, simultaneous breaking and forming, to be treated by

https://doi.org/10.26434/chemrxiv-2024-2v1nx **ORCID:** https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

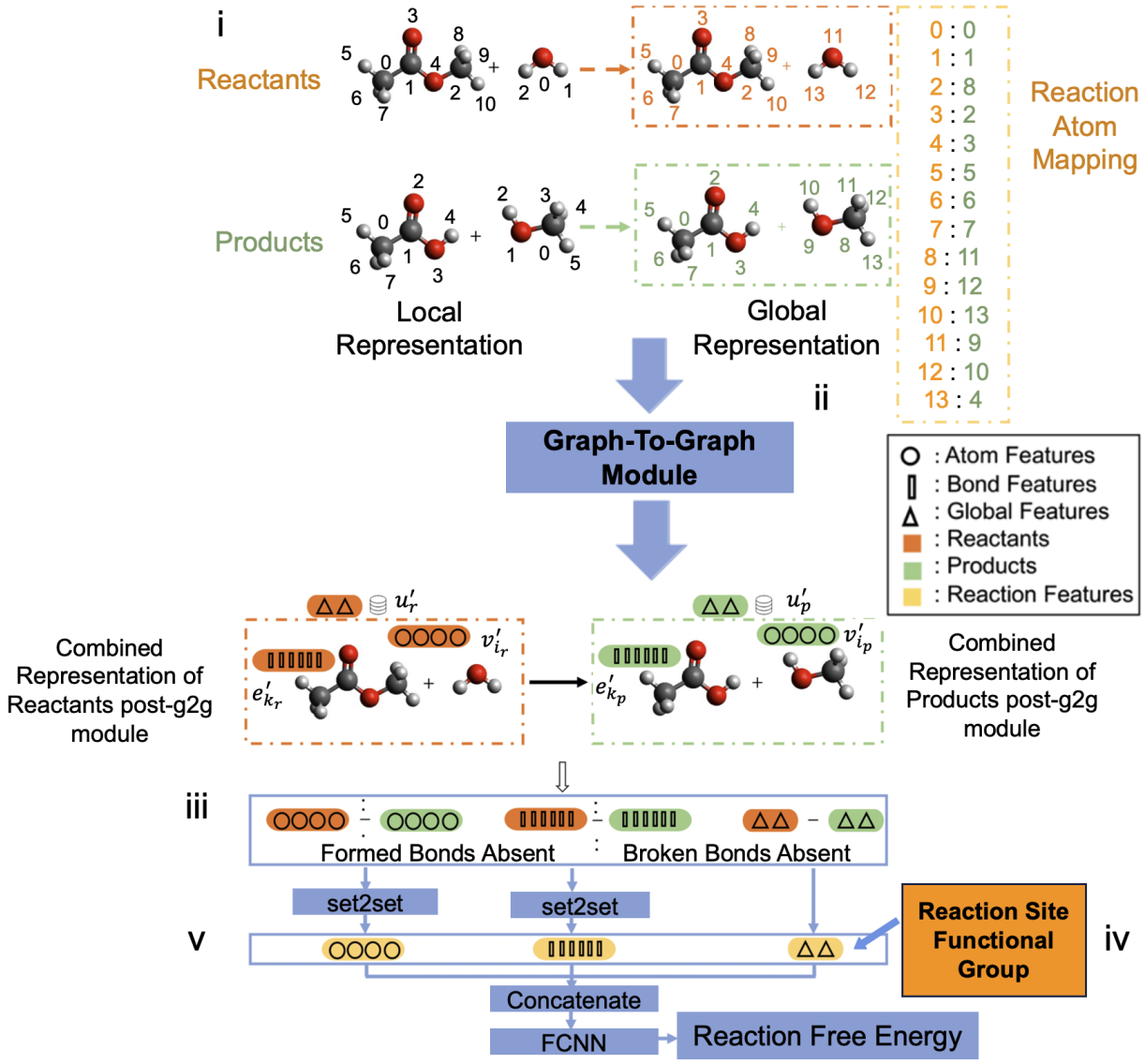


Figure 2: GNN Architecture: The user inputs atom-mapped sets of reactants and products (i) which undergo message-passing and update steps (ii). Using the user-specified mappings, these updated features are mapped to a global reaction graph (iii) where functional groups are the reaction site is added as a global feature (iv). Embeddings of bond and atom features plus global features directly serve as the fixed-size vector used in a conventional dense neural network for property prediction.

https://doi.org/10.26434/chemrxiv-2024-2v1nx **ORCID:** https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

the model (Figure 2). In addition, we are able to generalize our model to any number of species on either side of the reaction - a feature critical for hydrolysis where no reaction can fit BonDNet's original implementation.

For the task of hydrolysis, where we have a consistent reaction framework, we incorporated a one-hot encoding of functional group identity[45] into the global feature nodes. This encoding provides a simple, yet effective, descriptor that captures the reaction site of hydrolysis reactions alongside the more distant features generated by stacked message-passing layers. This is a particularly attractive feature as sequential stacking of message-passing layers rapidly increases compute time and can lead to problems such as oversmoothing.[65,66] We also implemented a host of computational features such as multi-GPU compatibility, a pytorch-lightning implementation, and added support for preprocesssing datasets.

3. Results and Discussion

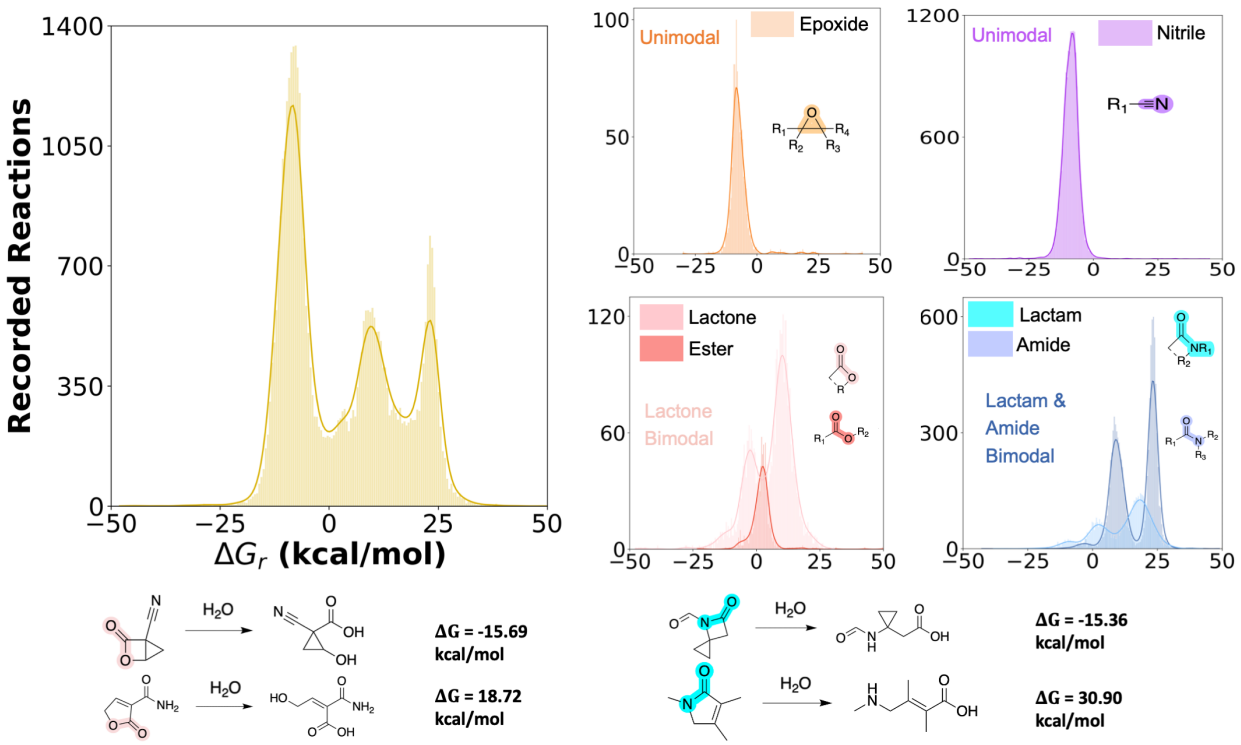
3.1 Dataset Overview

As detailed in *Section 2.1*, our hydrolysis database, in its current form, comprises a total of **68,761** reactions, making it the largest molecular database for hydrolysis reactions. Among these, the `QM9+Alchemy` dataset contains **41,006** reactions with reactant molecules in their neutral state, while the remaining reactions are approximately evenly split between the $\texttt{QM9}^{+}$ (protonated) and $\texttt{QM9}^{-}$ datasets, representing acidic and basic pH conditions, respectively.

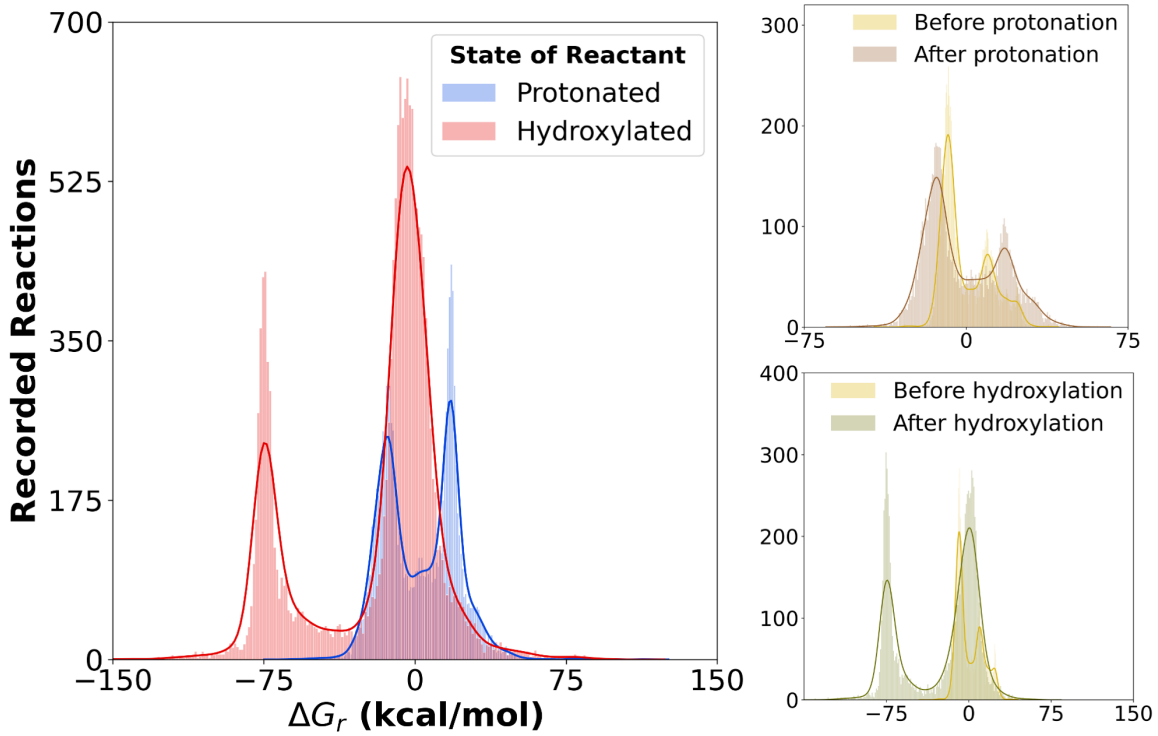
The `QM9+Alchemy` dataset contains reactants with up to 12 heavy atoms. The distribution of reactants based on the number of heavy elements is illustrated in Figure S3(a). For the charged subsets, ($\texttt{QM9}^{+}$ and $\texttt{QM9}^{-}$), the reactants are restricted to a maximum of 9 heavy atoms.

The number of hydrolyzed products varies depending on the reacting functional group, with reactions yielding one, two, and, in some cases (e.g., urea and carbamates), three products. Figure S3(b) visualizes the distribution of reactions based on the number of

https://doi.org/10.26434/chemrxiv-2024-2v1nx **ORCID:** https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0



(a)



(b)

Figure 3: (a) Distribution of hydrolysis ΔG_r for the `QM9 + Alchemy` dataset. (b) Distribution of hydrolysis ΔG_r for the $\texttt{QM9}^{+}$ and $\texttt{QM9}^{-}$ datasets. The shift in the distribution becomes clear when they are overlaid on the neutral ΔG_r for a subset of common reactants in the three datasets

https://doi.org/10.26434/chemrxiv-2024-2v1nx **ORCID:** https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

products generated, and Figure S3(c) illustrates the distribution across different hydrolyzed functional groups.

The hydrolysis ΔG_r distribution for the `QM9 + Alchemy` dataset is presented in Figure 3(a), where three peaks are observed: two distinct peaks in the endergonic region ($\Delta G_r > 0$) and one larger peak in the exergonic regime ($\Delta G_r < 0$). Interestingly, the ΔG_r distribution in Figure 3(a) is almost perfectly balanced with **20,547** reactions (50.11%) of the neutral reactions falling within the endergonic regime.

Further analysis across different functional groups reveals some interesting insights. Distributions of epoxides, nitriles, and esters exhibit unimodal energy distributions, while cyclic esters and cyclic amides (e.g., lactones and lactams) are bimodal. Sampling random lactone and lactam reactions from the endergonic and exergonic regimes indicates that cyclic structures with strained rings have more favorable thermodynamic hydrolysis pathways, whereas stable five-membered rings are more resistant to hydrolysis.[67,68] The hydrolysis of amides shows a distinctly bimodal nature, with both peaks centered in the endergonic regime, consistent with the established trend of the thermodynamic infeasibility of amide hydrolysis in a neutral reaction medium.[69]

The energy distribution for the protonated (`QM9`$^+$) and hydroxylated (`QM9`$^-$) datasets is shown in Figure 3(b). It is evident that the ΔG_r distribution for hydroxylated reactants shifts strongly toward the exergonic regime with greater than 70% of the reactions with a thermodynamically exergonic hydrolytic pathway. The shift in the protonated dataset is more subtle; however, when comparing the corresponding slices of the same reactions in the neutral and protonated states (Figure 3(b)), it is clear that the distribution broadens after protonation. Schematic S4 provides examples of hydrolysis for the same reactant in neutral, protonated, and hydroxylated states, illustrating the wide range of DFT-calculated ΔG_r values depending on the reactant's state.

The following section will discuss our model's performance on these different datasets and how it compares to existing benchmarks.

https://doi.org/10.26434/chemrxiv-2024-2v1nx **ORCID:** https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

3.2 Model Performance - Neutral Dataset

In the initial round of model training, we utilized a dataset of hydrolysis reactions generated exclusively from reactants extracted from the QM9 database. Despite a modest training set of **15,264** reactions, the model performed well on the holdout test set of **1,000** reactions across ten different functional groups. The Mean Absolute Error (MAE) was **2.44 kcal/mol**. Further details on the model's performance with this smaller dataset are provided in Section S4 of the SI.

Compared to other studies using molecular GNNs for property prediction,[35,40,70] this training dataset is relatively small. However, to the best of our knowledge, there are no publicly available datasets specifically for hydrolysis reactions. To evaluate the impact of additional training data, we curated **24,742** more reactions from the Alchemy database, focusing on molecules with 10, 11, or 12 heavy atoms. This expansion also increased the variety of hydrolyzing functional groups from 10 to 13. The expanded dataset was randomly split into training and testing sets at roughly a 9:1 ratio, resulting in a holdout test set of **3,800** reactions. As shown in *Figure 4(a)*, the model generalized effectively on this test set, with the MAE improving to **1.73 kcal/mol**. The parity plot comparing model predictions with DFT labels for the test set, shown in *Figure 4(c)*, demonstrates a high coefficient of determination (R^2) of **0.96**. The distribution of deviations between model predictions and DFT labels, illustrated in *Figure 4(b)*, indicates that errors are closely centered around a mean of zero kcal/mol. A detailed breakdown of these errors is provided in Table 1, showing that only 53 out of 3,800 reactions had prediction errors exceeding 10 kcal/mol, thus suggesting the model is suitable for screening purposes in this regime. Section S5 of the SI includes examples of five such outlier reactions, which often feature unusual structures with multiple strained rings, potentially contributing to the larger prediction errors. Another important aspect of evaluating the model's applicability is its capability to classify the overall thermodynamic feasibility of hydrolysis reactions as either exergonic or endergonic, based on the DFT labels. The model correctly classifies 97.1% of the reactions in the test

https://doi.org/10.26434/chemrxiv-2024-2v1nx **ORCID:** https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

set, demonstrating its strong predictive power in distinguishing between these two thermodynamic outcomes. Among the 117 misclassified reactions, a significant proportion (72) had DFT-calculated ΔG_r values close to zero, highlighting the inherent difficulty of correctly classifying these reactions.

Table 2 summarizes MAE statistics by functional group, revealing that no specific functional group performs poorly. Notably, functional groups like lactones and lactams, which exhibit a broader range of ΔG_r as shown in *Figure 3(a)*, tend to have higher MAEs. The higher MAE for imides may be attributed to their lower representation in the database (165 out of 41,006 reactions). However, interestingly, some functional groups with lower representation, such as aliphatic fluoride (191) and cyclic carbonate (116), show lower MAEs compared to the model average.

To assess our model's performance relative to other reaction-property prediction algorithms, we benchmarked it against several models. As detailed in Section 2.3, our model is highly generalizable and capable of handling reactions with varying numbers of bond changes—a feature not commonly found in reaction-property algorithms. This limitation narrowed the range of models suitable for benchmarking. We tested a simple reactant-only graph neural network with both atom and bond features, incorporating standard chemoinformatic features such as bond degree, element identity, atomic weight, ring inclusion, and hybridization, as well as global features such as the number of atoms and bonds, molecular weight, and one-hot encoding for the hydrolyzing functional group and charge. Additionally, we evaluated an XGBoost model with Morgan Fingerprints and Chemprop,[70] another modern algorithm. Both the XGBoost and Chemprop models were tuned using Bayesian optimization before final testing. The performance of these models is summarized in Table 3. Although Chemprop performs competitively (MAE: 2.25 kcal/mol vs. 1.73 kcal/mol), our model outperforms all benchmarked models in terms of performance metrics on the holdout test set. The performance of individual benchmark models is shown in *Figure 5*.

https://doi.org/10.26434/chemrxiv-2024-2v1nx **ORCID:** https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

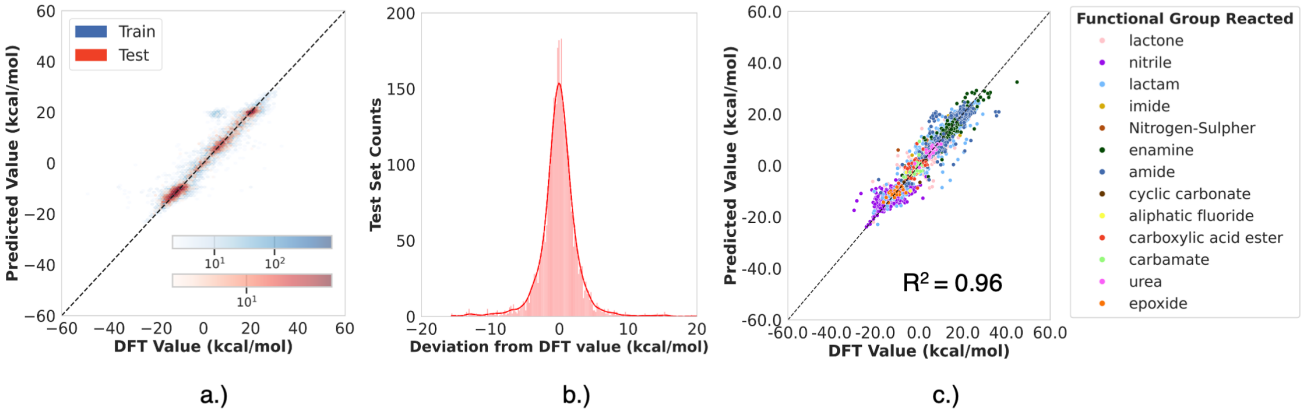


Figure 4: Performance of HEPOM on the `QM9 + Alchemy` dataset. (a) ΔG_r predicted by HEPOM versus DFT reference labels for the train and test sets; (b) histogram of the prediction errors; (c) parity plot for the holdout test set segregated on the basis of hydrolyzed functional group.

Table 1: Error distribution for the `QM9+Alchemy` holdout test set

Absolute Error (kcal/mol)	**Counts**
< 2	2674
> 2 and < 5	893
> 5 and < 10	180
> 10	53

3.3 Model Embeddings - Neutral Dataset

Visualizing the feature space provides insight into the underlying patterns the model learns during training. To analyze the learned representations for the trained model, we extracted high-dimensional difference feature vectors for each test set reaction before they are implemented into the fully-connected layer for prediction. These vectors were then reduced to a two-dimensional (2D) space using the Uniform Manifold Approximation and Projection (UMAP) method.[71] The evolution of these 2D embeddings at different epochs during training is shown in Figure S6 of the SI.

Initially, the embeddings are loosely clustered based on the functional groups of the hydrolyzing reactants, as expected. However, as training progresses, these clusters become tighter and more defined, reflecting not only functional groups but also other underlying

https://doi.org/10.26434/chemrxiv-2024-2v1nx **ORCID:** https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

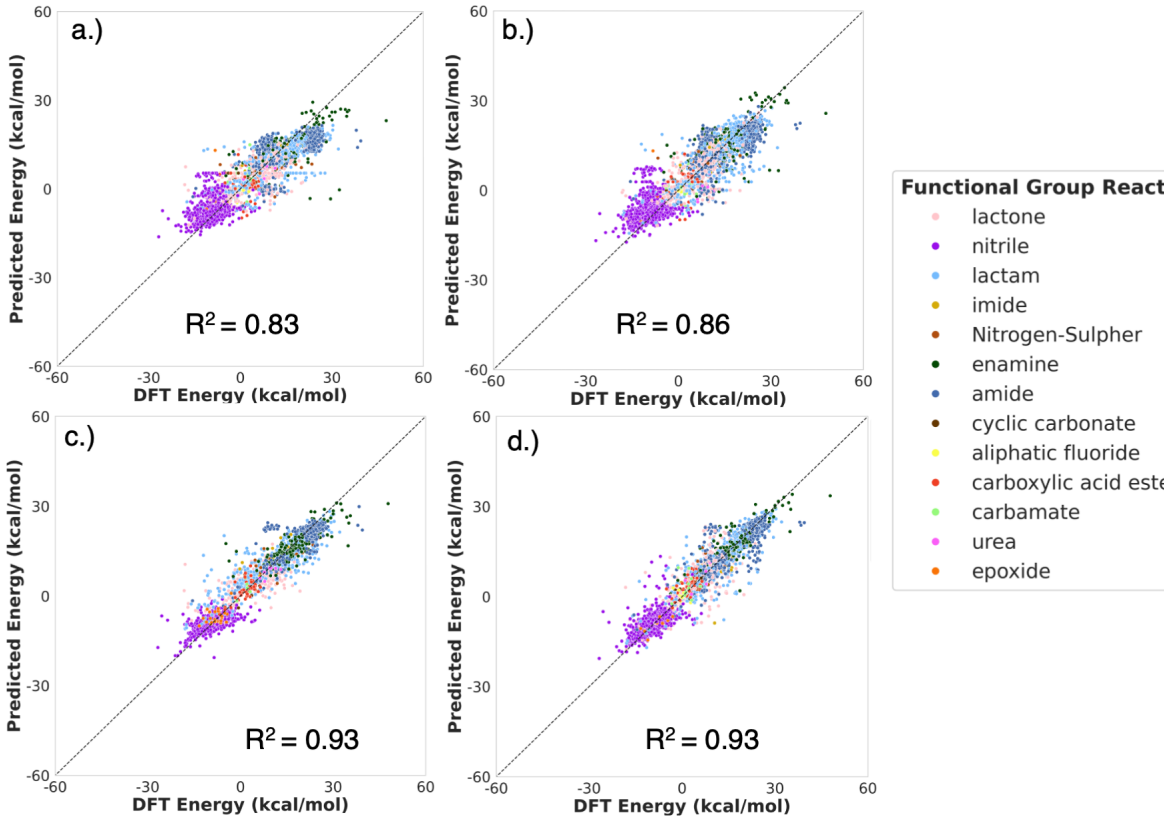


Figure 5: Parity plots for the performance of benchmark models on the holdout test of the `QM9 + Alchemy` dataset. (a) Reactant only GNN - node features; (b) Reactant only GNN - node + edge features; (c) XGBoost+Morgan fingerprints; (d) Chemprop[70]

https://doi.org/10.26434/chemrxiv-2024-2v1nx **ORCID:** https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Table 2: MAE Statistics Based on Functional Group Hydrolyzed

Functional Group	**MAE (kcal/mol)**
Lactone	2.198
Nitrile	1.433
Lactam	2.408
Imide	2.498
Nitrogen-Sulphur cleavage	2.176
Enamine	2.098
Amide	1.724
Cyclic Carbonate	1.252
Aliphatic Fluoride	1.022
Carboxylic Acid Ester	1.436
Carbamate	1.591
Urea	1.091
Epoxide	1.571
Overall Average	**1.731**

chemical similarities not explicitly known to the model. Figure 3 illustrates the final 2D representations of the feature vectors for the test set, each tagged with its respective hydrolyzing functional group. In addition to clustering by functional groups, a clear distinction emerges between the embeddings of uni-product reactions and those of bi-product and tri-product reactions. Uni-product reactions predominantly cluster on one side of the feature vector space, while reactions yielding more than one product aggregate oppositely. For uni-product reactions, the model forms a distinct cluster for cyclic functional groups (lactams, lactones, and imides). This suggests that the model identifies additional common features beyond functional group similarity, such as ring-opening during hydrolysis.

3.4 Model Performance: QM9$^+$ and QM9$^-$ datasets

In Section 2.1, we discussed that in most practical scenarios, hydrolysis occurs in a reaction medium where acidic or basic pH expedites the reaction. To extend the model's applicabil-

https://doi.org/10.26434/chemrxiv-2024-2v1nx **ORCID:** https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Table 3: Performance comparison against benchmark models for `QM9+Alchemy` dataset. We also include a benchmark to trivially guessing the mean of the training set

Model	**Test MAE (kcal/mol)**	**Test RMSE (kcal/mol)**
Mean	12.745	14.670
Reactant GNN(atom)	4.008	5.429
Reactant GNN(atom+bond)	3.445	4.875
XGB + Morgan	2.448	3.705
Chemprop	2.257	3.528
HEPOM	**1.731**	**2.674**

ity, we generated the $\texttt{QM9}^{+}$ and $\texttt{QM9}^{-}$ datasets, which include protonated and hydroxylated reactants, respectively, to simulate extreme *pH* conditions. In the current work, the protonation/hydroxylation was limited to only the `QM9` molecules. Therefore these charged datasets are considerably smaller than the neutral `QM9 + Alchemy` dataset. The trained models' performance on the holdout test sets is shown in *Figure 7* and summarized in Table 4.

As expected, the model's performance on these datasets deteriorates, evidenced by the lower coefficient of determination (R^2) and higher Mean Absolute Errors (MAEs) for both the $\texttt{QM9}^{+}$ and $\texttt{QM9}^{-}$ test sets. This performance decrease is particularly pronounced for the hydroxylated model, which shows a relatively high MAE of 6.607 kcal/mol. However, it is important to contextualize this result by noting that the range of ΔG_r values in this data set is radically different, roughly spanning between -150 and 100 kcal / mol, compared to the `QM9 + Alchemy` data sets (-40 to 40 kcal/mol). Notably, the MAE value also corresponds to a strong R^2 of 0.93.

Given these differences, a fair comparison of model performance should be made against relevant benchmarks rather than the neutral dataset. For these two datasets, we conducted hyperparameter optimization using XGBoost and Chemprop models. As shown in Table 4, our model outperforms the benchmarks on most metrics across this smaller and more complex dataset. The exception is the mean absolute error (MAE) on the $\texttt{QM9}^{-}$ dataset, where

https://doi.org/10.26434/chemrxiv-2024-2v1nx **ORCID:** https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

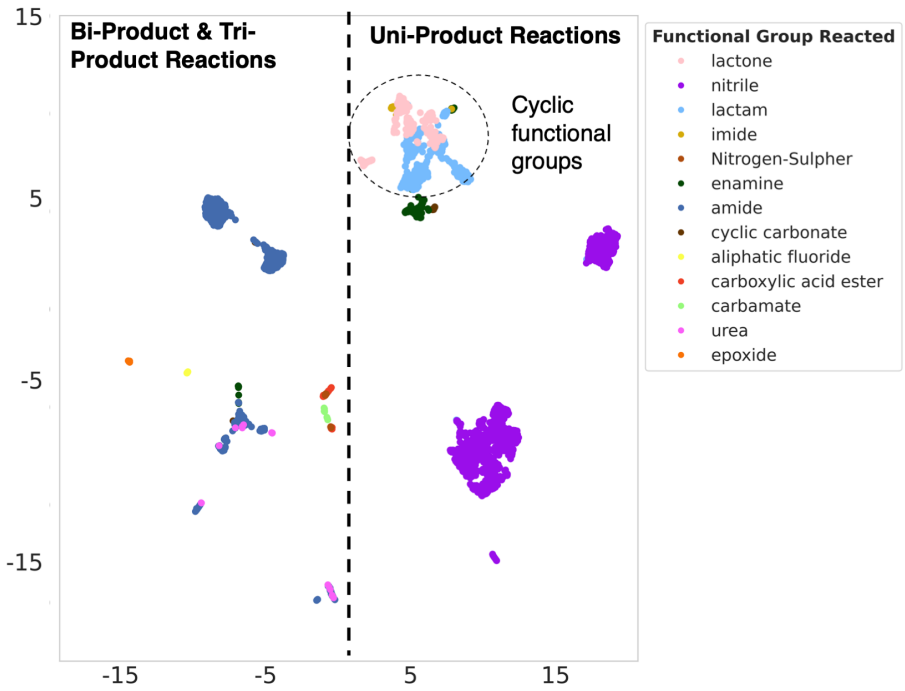


Figure 6: UMAP Embeddings of the high-dimensional feature vectors representing the hydrolysis reactions into a two-dimensional space

Chemprop slightly outperforms HEPOM. Interestingly, despite the higher MAE for HEPOM on the QM9$^-$ test set, it demonstrates greater robustness to outlier predictions, achieving a lower root mean squared error (RMSE) compared to Chemprop. Furthermore, HEPOM significantly outperforms Chemprop in correctly classifying the thermodynamic feasibility (endergonic vs. exergonic) of reactions in the QM9$^-$ dataset, achieving a classification accuracy of 95.5% compared to Chemprop's 84.8%. XGBoost exhibits the lowest classification accuracy of 76.3% on the same test set. Additional details, including parity plots for the benchmarks and MAE statistics by functional group for these holdout test sets, are provided in Section S7 of the SI.

3.5 Combined Model Training

For the neutral dataset, we observed a significant improvement in model performance after incorporating additional data from the Alchemy dataset. Given this result, it can be

https://doi.org/10.26434/chemrxiv-2024-2v1nx ORCID: https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

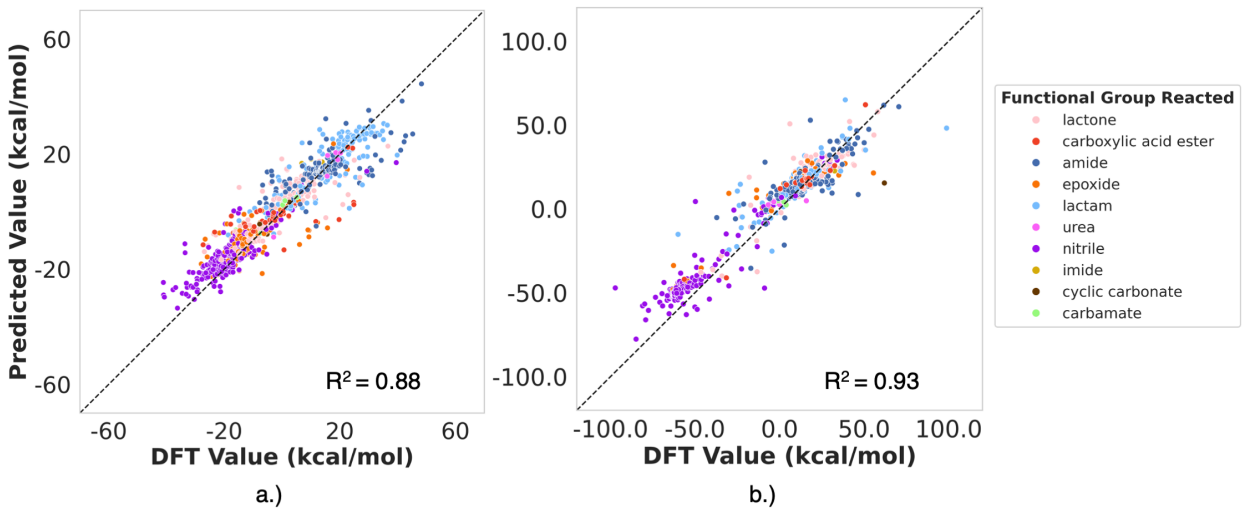


Figure 7: Test Set Performance on models trained with the a.) QM9$^+$ protonated and b.) QM9$^-$ hydroxylated datasets.

Table 4: Performance comparison against benchmark models for QM9$^+$ and QM9$^-$ datasets, best model is bolded.

Model	QM9$^+$ **Dataset**		QM9$^-$ **Dataset**	
	Test MAE (kcal/mol)	**Test RMSE (kcal/mol)**	**Test MAE (kcal/mol)**	**Test RMSE (kcal/mol)**
Mean	15.234	17.381	34.831	36.919
XGB + Morgan	5.394	8.195	8.687	14.375
Chemprop	6.275	8.864	**5.373**	9.682
HEPOM	**4.282**	**6.213**	6.607	**9.326**

expected that the performance of the protonated (QM9$^+$) and hydroxylated (QM9$^-$) models would also benefit from more data. However, obtaining this would require another round of computationally intensive data curation. Instead, we chose to augment the datasets by combining all three datasets. We found that adding the neutral data significantly improved the performance of both charged models, as shown in Figure 8.

Importantly, this improvement was not limited to our model but was also observed in the benchmark models, particularly Chemprop. This suggests that the higher MAEs of the QM9$^+$ and QM9$^-$ models may be simply due to insufficient data. However, as seen in Table 5,

https://doi.org/10.26434/chemrxiv-2024-2v1nx **ORCID:** https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

the most pronounced improvement was in our HEPOM model, which allowed it to surpass the other benchmarks, including in the previously higher MAE for the QM9$^-$ test set (Table 4).

The parity plots in Figure 8(a) and the MAE split based on the state of the reactant, compiled in Table 6, show that the improved model performance for the charged datasets comes at the cost of a slightly higher MAE for the neutral test set. The corresponding parity plots for the benchmark models and the MAE split based on functional groups are included in Section S8 of the SI.

Table 5: Performance comparison against benchmark models the combined dataset, best model is bolded.

Model	**Test MAE (kcal/mol)**	**Test RMSE (kcal/mol)**
XGB + Morgan	9.998	14.756
Chemprop	4.920	8.562
HEPOM	**3.054**	**4.281**

Table 6: Our Model MAE stratified on the combined test set based on the reactant state.

State of Reactant	**Mean Absolute Error (kcal/mol)**
Hydroxylated	4.057
Neutral	2.805
Protonated	2.838

4. Conclusion

In this work, we combined reaction templates and high-throughput DFT calculations to generate a large and diverse dataset of ΔG_r values of hydrolytic pathways for molecules screened from two popular molecular databases (QM9 and Alchemy). We then utilized this

https://doi.org/10.26434/chemrxiv-2024-2v1nx **ORCID:** https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

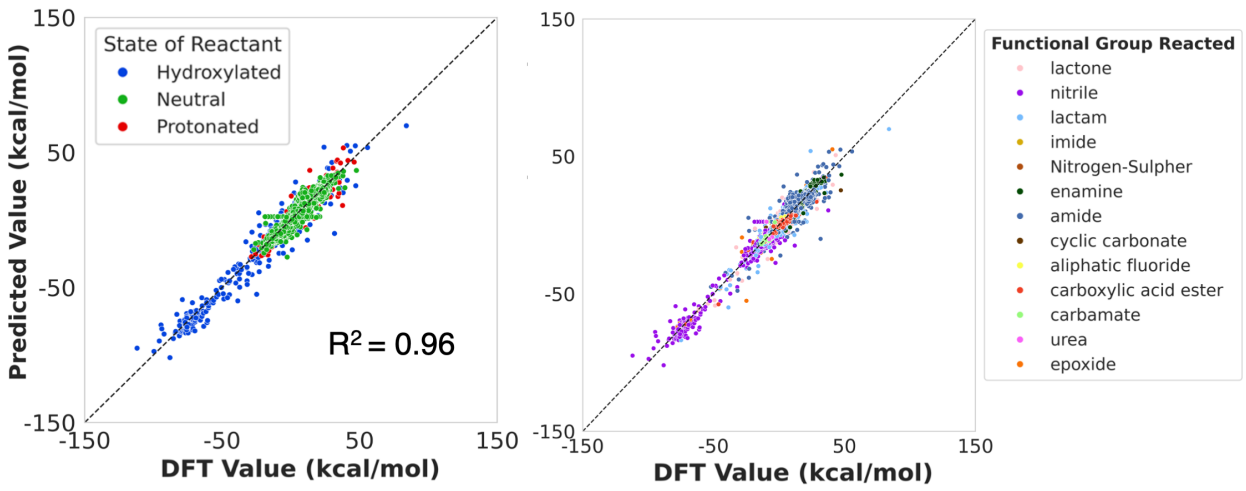


Figure 8: Test Set Performance on models trained with the combined dataset.

dataset to train a custom message-passing GNN on the difference features of the products and reactants, resulting in a model, capable of predicting the thermodynamic feasibility (ΔG_r) of hydrolysis reactions. The model demonstrates remarkable accuracy on the neutral dataset of hydrolysis reactions and outperforms benchmark models on smaller, more complex datasets involving charged reactants, simulating extreme pH conditions. In addition, by combining all three of our datasets, we find that our model is able to reasonably predict across all three classes at once.

We believe this model to be valuable for high-throughput screening of molecules and automated chemical synthesis in various domains, including drug development, environmental chemistry, and chemical deconstruction. The comprehensive dataset developed in this work also serves as a critical resource for training other machine learning models. In terms of the model, although this study focuses on hydrolysis, the model can be easily fine-tuned for any arbitrary reaction datasets with available reactant and product molecule graphs.

Training and holdout test sets for all models are publicly accessible via Figshare, and detailed information about the reactant and product molecules for the `QM9` database is available

https://doi.org/10.26434/chemrxiv-2024-2v1nx **ORCID:** https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

via the MPCules[72] interface, with future plans to integrate the `Alchemy` dataset reactants and products as well. The code for training the model can be accessed at the `GitHub` repository.

Acknowledgments

This work was intellectually led by the Moore Foundation Grant, which is funded by the Gordon and Betty Moore Foundation, under Grant no. 10454. Additional support was provided by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Materials Sciences and Engineering Division under contract no. DE-AC02-05- CH11231, Unlocking Chemical Circularity in Recycling by Controlling Polymer Reactivity across Scales program CUP-LBL- Helms. Data for this study were produced using computational resources provided by the Eagle and Swift high-performance computing (HPC) systems at the National Renewable Energy Laboratory and the Savio HPC cluster at University of California, Berkeley. The GNN models were trained on the Eagle HPC. This research was supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Department of Energy Computational Science Graduate Fellowship under Award Number DE-SC0020347.

References

(1) Breynaert, E.; Houlleberghs, M.; Radhakrishnan, S.; Grübel, G.; Taulelle, F.; Martens, J. A. Water as a tuneable solvent: a perspective. *Chemical Society Reviews* **2020**, *49*, 2557–2569.

(2) Idolor, O.; Guha, R. D.; Berkowitz, K.; Grace, L. An experimental study of the dynamic molecular state of transient moisture in damaged polymer composites. *Polymer Composites* **2021**, *42*, 3391–3403.

https://doi.org/10.26434/chemrxiv-2024-2v1nx **ORCID:** https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(3) Franks, F. *Water in Crystalline Hydrates Aqueous Solutions of Simple Nonelectrolytes*; Springer, 1973; pp 1–54.

(4) Pohorille, A.; Pratt, L. R. Is water the universal solvent for life? *Origins of Life and Evolution of Biospheres* **2012**, *42*, 405–409.

(5) Butler, R. N.; Coyne, A. G. Water: Nature's Reaction Enforcer Comparative Effects for Organic Synthesis "In-Water" and "On-Water". *Chemical reviews* **2010**, *110*, 6302–6337.

(6) Gorb, L.; Asensio, A.; Tuñón, I.; Ruiz-López, M. F. The mechanism of formamide hydrolysis in water from ab initio calculations and simulations. *Chemistry–A European Journal* **2005**, *11*, 6743–6753.

(7) Arumugam, P.; Gruber, S.; Tanaka, K.; Haering, C. H.; Mechtler, K.; Nasmyth, K. ATP Hydrolysis Is Required for Cohesin's Association with Chromosomes. *Current Biology* **2003**, *13*, 1941–1953.

(8) Bashkin, J. K. Hydrolysis of phosphates, esters and related substrates by models of biological catalysts. *Current Opinion in Chemical Biology* **1999**, *3*, 752–758.

(9) Blazek, J.; Gilbert, E. P. Effect of Enzymatic Hydrolysis on Native Starch Granule Structure. *Biomacromolecules* **2010**, *11*, 3275–3289, Publisher: American Chemical Society.

(10) Helms, B. A. Polydiketoenamines for a Circular Plastics Economy. *Accounts of Chemical Research* **2022**, *55*, 2753–2765.

(11) Le Feunteun, S.; Verkempinck, S.; Floury, J.; Janssen, A.; Kondjoyan, A.; Marze, S.; Mirade, P.-S.; Pluschke, A.; Sicard, J.; Van Aken, G., et al. Mathematical modelling of food hydrolysis during in vitro digestion: From single nutrient to complex foods

https://doi.org/10.26434/chemrxiv-2024-2v1nx **ORCID:** https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

in static and dynamic conditions. *Trends in Food Science & Technology* **2021**, *116*, 870–883.

(12) Meng, X.; Guo, Y.; Wang, Y.; Fan, S.; Wang, K.; Han, W. A systematic review of photolysis and hydrolysis degradation modes, degradation mechanisms, and identification methods of pesticides. *Journal of Chemistry* **2022**, *2022*, 9552466.

(13) Demarteau, J.; Epstein, A. R.; Christensen, P. R.; Abubekerov, M.; Wang, H.; Teat, S. J.; Seguin, T. J.; Chan, C. W.; Scown, C. D.; Russell, T. P.; Keasling, J. D.; Persson, K. A.; Helms, B. A. Circularity in mixed-plastic chemical recycling enabled by variable rates of polydiketoenamine hydrolysis. *Science Advances* **2022**, *8*, eabp8823.

(14) Olsson, E.; Menzel, C.; Johansson, C.; Andersson, R.; Koch, K.; Järnström, L. The effect of pH on hydrolysis, cross-linking and barrier properties of starch barriers containing citric acid. *Carbohydrate Polymers* **2013**, *98*, 1505–1513.

(15) Mitchell, S. M.; Ullman, J. L.; Teel, A. L.; Watts, R. J. pH and temperature effects on the hydrolysis of three β-lactam antibiotics: Ampicillin, cefalotin and cefoxitin. *Science of the total environment* **2014**, *466*, 547–555.

(16) Epstein, A. R.; Demarteau, J.; Helms, B. A.; Persson, K. A. Variable Amine Spacing Determines Depolymerization Rate in Polydiketoenamines. *Journal of the American Chemical Society* **2023**, *145*, 8082–8089.

(17) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *The Journal of Physical Chemistry B* **2009**, *113*, 6378–6396, Publisher: American Chemical Society.

(18) Bergström, C. A.; Luthman, K.; Artursson, P. Accuracy of calculated pH-dependent aqueous drug solubility. *European journal of pharmaceutical sciences* **2004**, *22*, 387–398.

https://doi.org/10.26434/chemrxiv-2024-2v1nx **ORCID:** https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(19) Girisuta, B.; Janssen, L. P. B. M.; Heeres, H. J. Kinetic Study on the Acid-Catalyzed Hydrolysis of Cellulose to Levulinic Acid. *Industrial & Engineering Chemistry Research* **2007**, *46*, 1696–1708, Publisher: American Chemical Society.

(20) Carlson, D. L.; Than, K. D.; Roberts, A. L. Acid- and Base-Catalyzed Hydrolysis of Chloroacetamide Herbicides. *Journal of Agricultural and Food Chemistry* **2006**, *54*, 4740–4750, Publisher: American Chemical Society.

(21) Jorner, K.; Brinck, T.; Norrby, P.-O.; Buttar, D. Machine learning meets mechanistic modelling for accurate prediction of experimental activation energies. *Chemical Science* **2021**, *12*, 1163–1175.

(22) Hirao, H.; Que Jr, L.; Nam, W.; Shaik, S. A Two-State Reactivity Rationale for Counterintuitive Axial Ligand Effects on the C H Activation Reactivity of Nonheme FeIV O Oxidants. *Chemistry–A European Journal* **2008**, *14*, 1740–1756.

(23) Epstein, A. R.; Spotte-Smith, E. W. C.; Venetos, M. C.; Andriuc, O.; Persson, K. A. Assessing the Accuracy of Density Functional Approximations for Predicting Hydrolysis Reaction Kinetics. *Journal of Chemical Theory and Computation* **2023**, *19*, 3159–3171.

(24) Malick, D. K.; Petersson, G. A.; Montgomery, J. A. Transition states for chemical reactions I. Geometry and classical barrier height. *The Journal of Chemical Physics* **1998**, *108*, 5704–5713.

(25) Evans, M. G.; Polanyi, M. Further considerations on the thermodynamics of chemical equilibria and reaction rates. *Transactions of the Faraday Society* **1936**, *32*, 1333.

(26) Stuyver, T.; Coley, C. W. Machine Learning-Guided Computational Screening of New Candidate Reactions with High Bioorthogonal Click Potential. *Chemistry (Weinheim an Der Bergstrasse, Germany)* **2023**, *29*, e202300387.

https://doi.org/10.26434/chemrxiv-2024-2v1nx **ORCID:** https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(27) Zhou, S.; Nguyen, B. T.; Richard, J. P.; Kluger, R.; Gao, J. Origin of Free Energy Barriers of Decarboxylation and the Reverse Process of CO2 Capture in Dimethylformamide and in Water. *Journal of the American Chemical Society* **2021**, *143*, 137–141, Publisher: American Chemical Society.

(28) Lawson, K. E.; Dekle, J. K.; Adamczyk, A. J. Towards pharmaceutical protein stabilization: DFT and statistical learning studies on non-enzymatic peptide hydrolysis degradation mechanisms. *Computational and Theoretical Chemistry* **2022**, *1218*, 113938.

(29) Mardirossian, N.; Head-Gordon, M. Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. *Molecular Physics* **2017**, *115*, 2315–2372.

(30) Ribeiro, A. J. M.; Ramos, M. J.; Fernandes, P. A. Benchmarking of DFT Functionals for the Hydrolysis of Phosphodiester Bonds. *Journal of Chemical Theory and Computation* **2010**, *6*, 2281–2292, Publisher: American Chemical Society.

(31) Cramer, C. J.; Truhlar, D. G. A universal approach to solvation modeling. *Accounts of chemical research* **2008**, *41*, 760–768.

(32) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. *arXiv:1704.01212 [cs]* **2017**, arXiv: 1704.01212.

(33) Grambow, C. A.; Pattanaik, L.; Green, W. H. Deep Learning of Activation Energies. *The Journal of Physical Chemistry Letters* **2020**, *11*, 2992–2997.

(34) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design* **2016**, *30*, 595–608.

(35) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.;

https://doi.org/10.26434/chemrxiv-2024-2v1nx **ORCID:** https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *Journal of Chemical Information and Modeling* **2019**, *59*, 3370–3388.

(36) Heid, E.; Green, W. H. Machine Learning of Reaction Properties via Learned Representations of the Condensed Graph of Reaction. *Journal of Chemical Information and Modeling* **2022**, *62*, 2101–2110.

(37) Wen, M.; Blau, S. M.; Xie, X.; Dwaraknath, S.; Persson, K. A. Improving machine learning performance on small chemical reaction data with unsupervised contrastive pretraining. *Chemical Science* **2022**, *13*, 1446–1458.

(38) Grambow, C. A.; Pattanaik, L.; Green, W. H. Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry. *Scientific Data* **2020**, *7*, 137.

(39) Ruddigkeit, L.; Van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *Journal of Chemical Information and Modeling* **2012**, *52*, 2864–2875.

(40) Wen, M.; Blau, S. M.; Spotte-Smith, E. W. C.; Dwaraknath, S.; Persson, K. A. BonDNet: a graph neural network for the prediction of bond dissociation energies for charged molecules. *Chemical Science* **2021**, *12*, 1858–1868.

(41) Spotte-Smith, E. W. C.; Blau, S. M.; Xie, X.; Patel, H. D.; Wen, M.; Wood, B.; Dwaraknath, S.; Persson, K. A. Quantum chemical calculations of lithium-ion battery electrolyte and interphase species. *Scientific Data* **2021**, *8*, 203.

(42) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data* **2014**, *1*, 140022.

(43) Chen, G.; Chen, P.; Hsieh, C.-Y.; Lee, C.-K.; Liao, B.; Liao, R.; Liu, W.; Qiu, J.;

https://doi.org/10.26434/chemrxiv-2024-2v1nx **ORCID:** https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Sun, Q.; Tang, J., et al. Alchemy: A quantum chemistry dataset for benchmarking ai models. *arXiv preprint arXiv:1906.09427* **2019**,

(44) Landrum, G. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling.

(45) Tebes-Stevens, C.; Patel, J. M.; Jones, W. J.; Weber, E. J. Prediction of hydrolysis products of organic chemicals under environmental pH conditions. *Environmental science & technology* **2017**, *51*, 5008–5016.

(46) Schowen, R. L.; Jayaraman, H.; Kershner, L. Catalytic Efficiencies in Amide Hydrolysis. The Two-Step Mechanism1, 2. *Journal of the American Chemical Society* **1966**, *88*, 3373–3375.

(47) Zahn, D. On the role of water in amide hydrolysis. *European Journal of Organic Chemistry* **2004**, *2004*, 4020–4023.

(48) Jencks, W. P.; Carriuolo, J. General base catalysis of ester Hydrolysis1. *Journal of the American Chemical Society* **1961**, *83*, 1743–1750.

(49) Rupp, M.; Korner, R.; V Tetko, I. Predicting the pKa of small molecules. *Combinatorial chemistry & high throughput screening* **2011**, *14*, 307–327.

(50) Baba, T.; Matsui, T.; Kamiya, K.; Nakano, M.; Shigeta, Y. A density functional study on the pKa of small polyprotic molecules. *International Journal of Quantum Chemistry* **2014**, *114*, 1128–1134.

(51) 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.

https://doi.org/10.26434/chemrxiv-2024-2v1nx **ORCID:** https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(52) Wu, J.; Kang, Y.; Pan, P.; Hou, T. Machine learning methods for pKa prediction of small molecules: Advances and challenges. *Drug Discovery Today* **2022**, *27*, 103372.

(53) 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* **2020**, *132*, 19444–19453.

(54) Dutra, F. R.; Silva, C. d. S.; Custodio, R. On the Accuracy of the Direct Method to Calculate pKa from Electronic Structure Calculations. *The Journal of Physical Chemistry A* **2021**, *125*, 65–73, Publisher: American Chemical Society.

(55) Epifanovsky, E.; Gilbert, A. T. B.; Feng, X.; Lee, J.; Mao, Y.; Mardirossian, N.; Pokhilko, P.; White, A. F.; Coons, M. P.; Dempwolff, A. L.; Gan, Z.; Hait, D.; Horn, P. R.; Jacobson, L. D.; Kaliman, I.; Kussmann, J.; Lange, A. W.; Lao, K. U.; Levine, D. S.; Liu, J.; McKenzie, S. C.; Morrison, A. F.; Nanda, K. D.; Plasser, F.; Rehn, D. R.; Vidal, M. L.; You, Z.-Q.; Zhu, Y.; Alam, B.; Albrecht, B. J.; Aldossary, A.; Alguire, E.; Andersen, J. H.; Athavale, V.; Barton, D.; Begam, K.; Behn, A.; Bellonzi, N.; Bernard, Y. A.; Berquist, E. J.; Burton, H. G. A.; Carreras, A.; Carter-Fenk, K.; Chakraborty, R.; Chien, A. D.; Closser, K. D.; Cofer-Shabica, V.; Dasgupta, S.; de Wergifosse, M.; Deng, J.; Diedenhofen, M.; Do, H.; Ehlert, S.; Fang, P.-T.; Fatehi, S.; Feng, Q.; Friedhoff, T.; Gayvert, J.; Ge, Q.; Gidofalvi, G.; Goldey, M.; Gomes, J.; González-Espinoza, C. E.; Gulania, S.; Gunina, A. O.; Hanson-Heine, M. W. D.; Harbach, P. H. P.; Hauser, A.; Herbst, M. F.; Hernández Vera, M.; Hodecker, M.; Holden, Z. C.; Houck, S.; Huang, X.; Hui, K.; Huynh, B. C.; Ivanov, M.; Jász, ; Ji, H.; Jiang, H.; Kaduk, B.; Kähler, S.; Khistyaev, K.; Kim, J.; Kis, G.; Klunzinger, P.; Koczor-Benda, Z.; Koh, J. H.; Kosenkov, D.; Koulias, L.; Kowalczyk, T.; Krauter, C. M.; Kue, K.; Kunitsa, A.; Kus, T.; Ladjánszki, I.; Landau, A.; Lawler, K. V.; Lefrancois, D.; Lehtola, S.; Li, R. R.; Li, Y.-P.; Liang, J.; Liebenthal, M.; Lin, H.-H.; Lin, Y.-S.; Liu, F.; Liu, K.-Y.; Loipersberger, M.; Luenser, A.;

https://doi.org/10.26434/chemrxiv-2024-2v1nx **ORCID:** https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Manjanath, A.; Manohar, P.; Mansoor, E.; Manzer, S. F.; Mao, S.-P.; Marenich, A. V.; Markovich, T.; Mason, S.; Maurer, S. A.; McLaughlin, P. F.; Menger, M. F. S. J.; Mewes, J.-M.; Mewes, S. A.; Morgante, P.; Mullinax, J. W.; Oosterbaan, K. J.; Paran, G.; Paul, A. C.; Paul, S. K.; Pavošević, F.; Pei, Z.; Prager, S.; Proynov, E. I.; Rák, ; Ramos-Cordoba, E.; Rana, B.; Rask, A. E.; Rettig, A.; Richard, R. M.; Rob, F.; Rossomme, E.; Scheele, T.; Scheurer, M.; Schneider, M.; Sergueev, N.; Sharada, S. M.; Skomorowski, W.; Small, D. W.; Stein, C. J.; Su, Y.-C.; Sundstrom, E. J.; Tao, Z.; Thirman, J.; Tornai, G. J.; Tsuchimochi, T.; Tubman, N. M.; Veccham, S. P.; Vydrov, O.; Wenzel, J.; Witte, J.; Yamada, A.; Yao, K.; Yeganeh, S.; Yost, S. R.; Zech, A.; Zhang, I. Y.; Zhang, X.; Zhang, Y.; Zuev, D.; Aspuru-Guzik, A.; Bell, A. T.; Besley, N. A.; Bravaya, K. B.; Brooks, B. R.; Casanova, D.; Chai, J.-D.; Coriani, S.; Cramer, C. J.; Cserey, G.; DePrince, A. E., III; DiStasio, R. A., Jr.; Dreuw, A.; Dunietz, B. D.; Furlani, T. R.; Goddard, W. A., III; Hammes-Schiffer, S.; Head-Gordon, T.; Hehre, W. J.; Hsu, C.-P.; Jagau, T.-C.; Jung, Y.; Klamt, A.; Kong, J.; Lambrecht, D. S.; Liang, W.; Mayhall, N. J.; McCurdy, C. W.; Neaton, J. B.; Ochsenfeld, C.; Parkhill, J. A.; Peverati, R.; Rassolov, V. A.; Shao, Y.; Slipchenko, L. V.; Stauch, T.; Steele, R. P.; Subotnik, J. E.; Thom, A. J. W.; Tkatchenko, A.; Truhlar, D. G.; Van Voorhis, T.; Wesolowski, T. A.; Whaley, K. B.; Woodcock, H. L., III; Zimmerman, P. M.; Faraji, S.; Gill, P. M. W.; Head-Gordon, M.; Herbert, J. M.; Krylov, A. I. Software for the frontiers of quantum chemistry: An overview of developments in the Q-Chem 5 package. *The Journal of Chemical Physics* **2021**, *155*, 084801.

(56) Mathew, K.; Montoya, J. H.; Faghaninia, A.; Dwarakanath, S.; Aykol, M.; Tang, H.; Chu, I.-h.; Smidt, T.; Bocklund, B.; Horton, M.; Dagdelen, J.; Wood, B.; Liu, Z.-K.; Neaton, J.; Ong, S. P.; Persson, K.; Jain, A. Atomate: A high-level interface to generate, execute, and analyze computational materials science workflows. *Computational Materials Science* **2017**, *139*, 140–152.

https://doi.org/10.26434/chemrxiv-2024-2v1nx **ORCID:** https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(57) Mardirossian, N.; Head-Gordon, M. B97M-V: A combinatorially optimized, range-separated hybrid, meta-GGA density functional with VV10 nonlocal correlation. *The Journal of Chemical Physics* **2016**, *144*, 214110.

(58) Vydrov, O. A.; Van Voorhis, T. Nonlocal van der Waals density functional: The simpler the better. *The Journal of Chemical Physics* **2010**, *133*, 244103.

(59) Hellweg, A.; Rappoport, D. Development of new auxiliary basis functions of the Karlsruhe segmented contracted basis sets including diffuse basis functions (def2-SVPD, def2-TZVPPD, and def2-QVPPD) for RI-MP2 and RI-CC calculations. *Physical Chemistry Chemical Physics* **2014**, *17*, 1010–1017, Publisher: The Royal Society of Chemistry.

(60) Bresson, X.; Laurent, T. Residual Gated Graph ConvNets. 2018; `http://arxiv.org/abs/1711.07553`, arXiv:1711.07553 [cs, stat].

(61) Dwivedi, V. P.; Joshi, C. K.; Luu, A. T.; Laurent, T.; Bengio, Y.; Bresson, X. Benchmarking Graph Neural Networks. 2022; `http://arxiv.org/abs/2003.00982`, arXiv:2003.00982 [cs, stat].

(62) Vinyals, O.; Bengio, S.; Kudlur, M. Order Matters: Sequence to sequence for sets. 2016; `http://arxiv.org/abs/1511.06391`, arXiv:1511.06391 [cs, stat].

(63) Lee, J.; Lee, I.; Kang, J. Self-attention graph pooling. International conference on machine learning. 2019; pp 3734–3743.

(64) Schweidtmann, A. M.; Rittig, J. G.; Weber, J. M.; Grohe, M.; Dahmen, M.; Leonhard, K.; Mitsos, A. Physical pooling functions in graph neural networks for molecular property prediction. *Computers & Chemical Engineering* **2023**, *172*, 108202.

(65) Zhou, K.; Dong, Y.; Wang, K.; Lee, W. S.; Hooi, B.; Xu, H.; Feng, J. Understanding

https://doi.org/10.26434/chemrxiv-2024-2v1nx **ORCID:** https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

and Resolving Performance Degradation in Graph Convolutional Networks. 2021; http://arxiv.org/abs/2006.07107, arXiv:2006.07107 [cs, stat].

(66) Rusch, T. K.; Bronstein, M. M.; Mishra, S. A Survey on Oversmoothing in Graph Neural Networks. 2023; http://arxiv.org/abs/2303.10993, arXiv:2303.10993 [cs].

(67) Hall Jr, H.; Brandt, M.; Mason, R. Hydrolysis rates and mechanisms of cyclic monomers. *Journal of the American Chemical Society* **1958**, *80*, 6420–6427.

(68) Yang, J.-C.; Gorenstein, D. G. Contribution of ring strain and the stereoelectronic effect to the hydrolysis of cyclic five-membered ring phosphorus esters. *Tetrahedron* **1987**, *43*, 479–486.

(69) Krug, J.; Popelier, P.; Bader, R. Theoretical study of neutral and of acid and base-promoted hydrolysis of formamide. *The Journal of Physical Chemistry* **1992**, *96*, 7604–7616.

(70) Heid, E.; Greenman, K. P.; Chung, Y.; Li, S.-C.; Graff, D. E.; Vermeire, F. H.; Wu, H.; Green, W. H.; McGill, C. J. *Chemprop: Machine Learning Package for Chemical Property Prediction*; preprint, 2023.

(71) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2020; http://arxiv.org/abs/1802.03426, arXiv:1802.03426 [cs, stat].

(72) Spotte-Smith, E. W. C.; Cohen, O. A.; Blau, S. M.; Munro, J. M.; Yang, R.; Guha, R. D.; Patel, H. D.; Vijay, S.; Huck, P.; Kingsbury, R.; Horton, M. K.; Persson, K. A. A database of molecular properties integrated in the Materials Project. *Digital Discovery* **2023**, Publisher: RSC.

https://doi.org/10.26434/chemrxiv-2024-2v1nx **ORCID:** https://orcid.org/0000-0003-1977-6039 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0