HEPOM: A Predictive Framework for Accelerated Hydrolysis Energy Predictions of Organic Molecules

Rishabh D. Guha [*]
Lawrence Berkeley National Laboratory

Santiago Vargas [*]
University of California, Los Angeles

Evan Walter Clark Spotte-Smith
University of California, Berkeley
Lawrence Berkeley National Laboratory

Alexander R. Epstein
University of California, Berkeley

Maxwell C. Venetos
University of California, Berkeley
Lawrence Berkeley National Laboratory

Mingjian Wen
University of Houston

Ryan S. Kingsbury
Princeton University

Samuel M. Blau
Lawrence Berkeley National Laboratory

Kristin A. Persson [†]
University of California, Berkeley
Lawrence Berkeley National Laboratory

Abstract

Hydrolysis is a fundamental chemical reaction where water facilitates the cleavage of bonds in a reactant molecule. 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. Subsequently, we use a Graph Neural Network (GNN) to predict the free energy changes (ΔG) for all hydrolytic pathways within a subset of the QM9 molecular dataset. The framework automatically identifies reaction centers, generates hydrolysis products, and utilizes a trained 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.

[*] Equal Contribution
[†] email: kapersson@lbl.gov

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

1 Introduction

Water is arguably the most widely known compound, and yet, its deceptively simple structure fails to suggest the complex relationships it forms with itself and with other compounds in reactions. In the case of hydrolysis, which is ubiquitous in both biological(1) and synthetic chemistry(2; 3), water doubles as a reactant and solvent medium in the reaction. At the molecular level, hydrolysis is initiated by the attack of a water, hydronium, or hydroxide molecule at specific sites in the reactant, triggering a sequence of bond cleavage and formations, leading to the formation of new product(s). The thermodynamic feasibility of this reaction is fundamentally tied to the pH of the aqueous reaction medium (4; 5). The availability of protons (H^{+}) or hydroxide (OH^{-}) ions, generates charged species with different reactivities than the neutral molecule. Consequently, acid or base-catalyzed hydrolysis (6; 7) of the same reactant can have prominently different reaction rates than its neutral counterparts and further complicates the study of these prominent reactions.

The Eyring equation provides a means to quantify experimental reaction rates by evaluating activation barriers ($\Delta G^{\ddagger}$) through computational methods (8; 9). However, this approach demands computationally intensive transition state calculations for each reaction along the complex potential energy surface (PES) (10; 11). In contrast, within a specific reaction family, the Bell-Evans-Polanyi principle (12) can offer a qualitative linear correlation between the thermodynamic Gibbs Free Energy change (ΔG_r) and the kinetic parameter $\Delta G^{\ddagger}$ (13; 14; 15). Nevertheless, quantifying this thermochemical quantity (ΔG_r) with high accuracy still requires DFT calculations with large basis sets and refined hybrid functionals for both reaction endpoints (16; 17). Depending on the size of the molecules, these calculations can take anywhere from several hours to days, particularly when employing implicit solvent models 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 (18; 19). Recently, graph convolutions, which can iteratively update node and edge features based on connectivity and local environment, have proven to be extremely effective in learning molecular (20; 21) and reaction representations (22; 23). 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. (19) was facilitated by a dataset of 12,000 gas-phase reactions (24) sampled from a subset of molecules in the GDB-17 dataset (25). The bond dissociation energy (BDE) prediction framework developed by Wen et al. (26) was trained on a dataset of over 60,000 homolytic and heterolytic bond dissociation reactions (27). In the realm of hydrolysis, no such comprehensive dataset currently exists.

In this work, we first developed a predictive framework based on reaction templates for different functional groups which can automatically generate products for multiple hydrolysis pathways in any molecule. This framework was then applied to a subset of the QM9 database (28) to generate a database of over 25,000 hydrolysis reactions in an implicit aqueous solvation environment. For a subset of the database, both the neutral and protonated states of the reactant molecule were considered to approximate hydrolysis in neutral and highly acidic pH conditions. Finally, we propose a GNN model that utilizes the difference features of the atom, bond, and global features between the products and the reactants to predict the DFT-calculated ΔG_r. The utilization of the global reaction atom mapping enables the model to track multiple elementary bond dissociation and formations, resulting in a mean absolute error (MAE) of 2.25 kcal mol^{-1} across a diverse holdout test set.

2 Methods

2.1 Reaction Generation

The hydrolyzable molecules in the QM9 database were screened through `RDKit` (29) substructure matching of 20 prototypically hydrolyzable functional groups. We then adapted hydrolysis reaction templates for the aforementioned groups from previous work by Tebes-Stevens et al. (30) into an automated framework for determining reaction products. For instance, as shown in Schematic S1, if an ester functional group was detected in a molecule, the reaction template used would yield a carboxylic acid and an alcohol as the respective hydrolysis products. Similar reaction templates were implemented for all functional groups. As seen in Schematic S2, the reaction template for nitriles yielded amides which can be further hydrolyzed into an amine and a carboxylic acid. Therefore, the products of the nitrile reactions were redirected as reactants for separate hydrolysis reactions to augment the dataset.

Hydrolysis reactions in neutral and strongly acidic pH were differentiated through two separate reaction schemes. For neutral pH, we assumed separate hydrolysis reactions between each detected functional group and one molecule of water. For an acidic medium, the reacting functional group was assumed to be protonated at the most basic atom site in the functional group moiety. The acidic pH reaction was then executed between the protonated reactant and two molecules of water to maintain reaction stoichiometry. A representative example of these two reaction conditions for a hydrolyzing carbamate molecule has been demonstrated in schematics S3 (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 was done to circumvent the erroneous DFT calculated energies of an isolated proton in an implicit solvent medium (31).

2.2 Density-Functional Theory

QChem (version 5 or 6) (32) 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. (27) and currently implemented in `atomate` (33) 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 (34), which employs the vv10 dispersion correction (35), to improve the non-covalent interactions. The def2-SVPD basis set (36) was employed for the `FFOpt` workflow and the solvation effects were implicitly accounted for with the water SMD solvent model (37). The electronic energies of the optimized structures were refined with single-point calculations using a larger def2-QZVPPD basis set (36).

2.3 Model Architecture

The GNN model, visually depicted in S4, is heavily based on the previously published BonDNet architecture (26). This algorithm uses gated graph convolutional (GatedGC) layers to propagate starting node features within the graphs of individual species on both sides of a reaction. While GatedGC layers have been used widely for structure-to-property models in chemistry and materials science (38; 39), 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. In order to propagate more distant graph relationships, several (typically 2-4 layers) GatedGC layers were stacked. With updated species' graphs, we constructed a reaction graph to hold reaction feature differences. Atom and bond nodes were mapped to each other on both sides of a reaction and features were subtracted from their corresponding node with zero-padding added to represent broken bonds. From here, a set2set (40) layer was applied to bond and atom node types in the reaction difference graph to obtain a vectorized representation of the reaction that is passed through a multilayer perceptron (MLP) for property prediction.

In this implementation, the reaction mapping is altered from the original BonDNet as 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 then subtracted reactant features from corresponding nodes in this scaffold. 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. In the presented model, we shift the atom-mapping to a prior task where atoms and bonds are labeled according to their mappings. This reduces the overhead of the model where it no longer has to determine mappings on-the-fly. More importantly, this change allows for an arbitrary number of bond changes to be treated by the model (both breaking and forming sequences in concert)(Fig. S4). With this, we extend the applicability of the original BonDNet immensely, not just for this task but to other, more complex chemical reactions.

We also attempted to leverage the consistent reaction framework of hydrolysis by incorporating a one-hot encoding of functional group identity 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 (41; 42). While this modification does not improve performance in the context of the neutral training/holdout sets, testing on the protonated and hydroxylated datasets remain.

3 Results and discussion

3.1 Dataset Overview

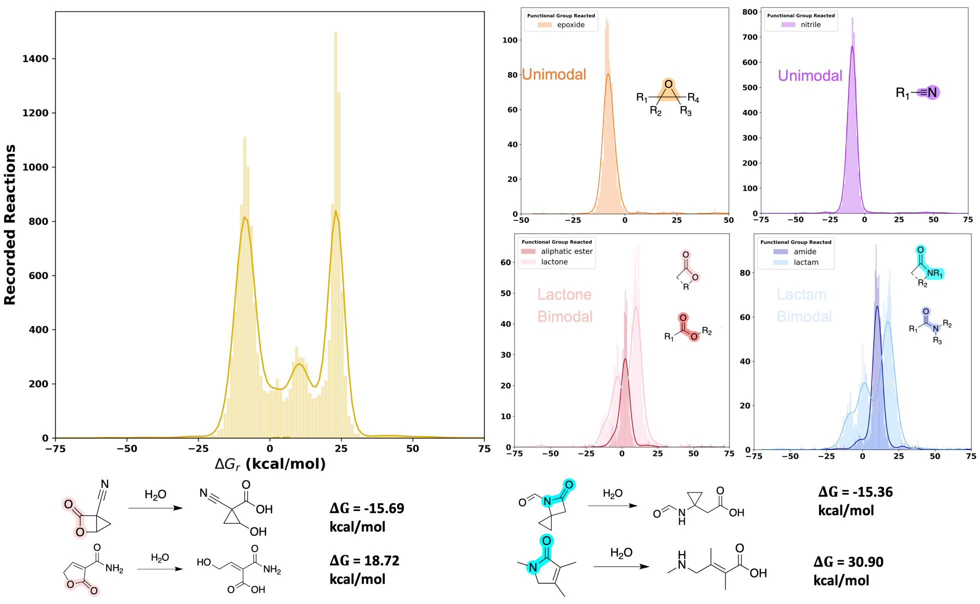

Figure 1: Distribution of ΔG_r for the compiled hydrolysis reactions.

In its current state, the dataset comprises a total of 25,599 reactions. Among these, 16,264 reactions correspond to reactants with a net zero charge, representing neutral pH conditions. The remaining reactions were generated from a subset of reactants from the neutral dataset. The hydrolyzable functional groups of these reactants were protonated at the relevant atom site to get positively charged reactants representing highly acidic pH conditions. The number of hydrolyzed products varies depending on the specific reacting functional group, with reactions yielding 1, 2, and in some instances (e.g., urea and carbamates), 3 products. The distribution of reactions based on the number of products generated is visualized in Figure S5(a) of the SI and the distribution across different hydrolyzed functional groups is also included in Figure S5(b). The ΔG_r distribution for the neutral dataset is presented in Figure 1. Here, we observe a bimodal nature, characterized by two distinct peaks in the endergonic and exergonic regimes. Approximately 54% (8837) of the neutral reactions fall within the endergonic regime. Further analysis across different functional groups reveals some interesting insights. Functional groups such as epoxides, nitriles, esters, and amides exhibit a unimodal energy distribution. Conversely, cyclic esters and cyclic amides, such as lactones and lactams, significantly contribute to the bimodal nature of the dataset. When we sample random lactone and lactam reactions from the endergonic and exergonic regimes, it becomes clear that cyclic structures with a strained ring structure have a more favorable thermodynamic hydrolysis pathway while stable 5-membered rings are more resistant to hydrolysis. The energy distribution for the protonated dataset and its differences when compared to the neutral, is included in Figure S6 of the

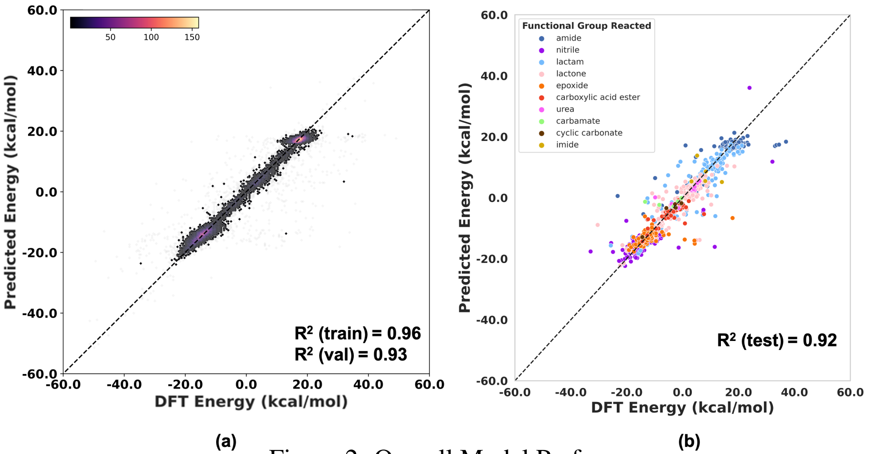

Figure 2: Overall Model Performance.

SI. However, for the scope of this work, our discussion regarding model performance is limited to the neutral dataset shown in Figure 1.

3.2 Overall Model Performance

To evaluate the model's performance and generalizability, we tested it on an independent holdout test set (Figure 2(b)) of hydrolysis reactions generated from QM9 molecules. This holdout set is comprised of 1000 reactions spanning diverse hydrolyzable functional groups and ΔG_r values ranging between -40 kcal/mol to 40 kcal/mol. Overall the predictions align accurately on the parity plot with an impressive coefficient of determination (R^2) and Mean Absolute Error (MAE) for both the validation and test sets. The model performance on the test set demonstrates generalizability, achieving an R^2 of 0.92 and a MAE of 2.25 kcal/mol vs. DFT-calculated values (Figure 2(a)). The classification accuracy for the model correctly classifying reactions endergonic vs. exergonic was also 95.3% in the test set.

Furthermore, to assess the model performance vs. other reaction-property prediction algorithms, we benchmarked our implementation to a host of other models. As discussed in Section 2.3, our model is highly generalizable and able to ingest reactions involving an arbitrary number of bond changes - a feature not common among reaction property algorithms. This, however, limited the range of models that could be selected for benchmarking. Nonetheless, we tested a simple reactant-only graph neural network with atom features and with both atom and bond features included. These features included a range of standard cheminformatic features such as bond degree, element identity, atomic weight, ring inclusion, hybridization, etc, coupled with global features like the total number of atoms and bonds in the reactant, molecular weight, and a one-hot encoding for the hydrolyzing functional group. An XGBoost model coupled with Morgan Fingerprints was also tested. Finally, Chemprop (43) was used as a more modern algorithm also based on graph neural networks and arbitrary bond changes. The XGBoost and Chemprop models were first tuned via a Bayesian optimization hyperparameter tuning scheme prior to final testing. We summarized the models' performance in Table 1 where our model outperforms all benchmarks on the holdout test set. We note that training performance for all the benchmarked models was close to the best-performing model, but their ability to generalize on the test set is limited.

3.3 Model Embeddings

To investigate the model-learned representations of the hydrolysis reactions, we reduced the high-dimensional difference feature vectors for each hydrolysis reaction into a two-dimensional (2D) space using the uniform manifold approximation and projection (UMAP) method (44). Figure 3 displays the 2D representations of the feature vectors for the test set, each tagged with its respective hydrolyzing functional group. A few interesting insights emerge from the visual patterns of the embeddings. As expected, the feature vectors for the hydrolysis reactions of similar functional groups cluster together. Specifically, in the case of lactones and lactams, we observe two distinctly separated clusters. In

Table 1: Our model performance vs. other comparable models and baselines

Model	Test MAE (kcal/mol)
Mean	12.3
Reactant GNN(atom)	3.54
Reactant GNN(atom+bond)	3.45
Chemprop	4.14
XGB + Morgan	3.23
Our Model w/ Funct. Group	**2.7**
Our Model	**2.25**

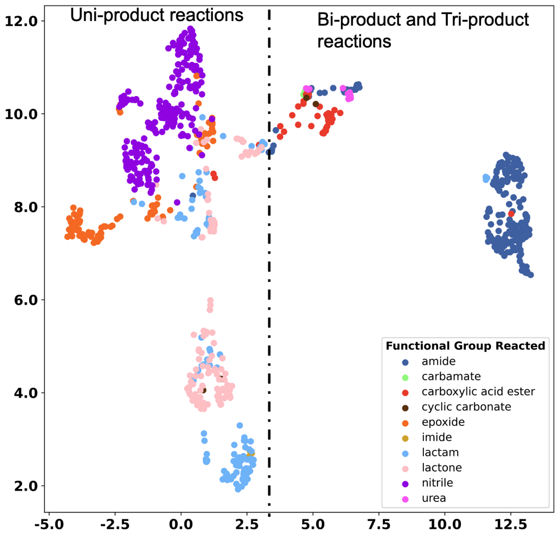

Figure 3: UMAP embedding of the reaction features

Figure S8 of the SI, we have separately visualized the two-dimensional UMAP embeddings of the exergonic and endergonic reactions for the lactams and lactones where the cluster on the top left is dominated by the exergonic reactions while the bottom left section broadly corresponds to the endergonic hydrolysis of these two functional groups. This implies that the model also learns to distinguish separate sub-classes for the same functional group. Furthermore, the uni-product reactions are all clustered to the left of the feature vector space while the reactions which yield more than one product aggregate on the right of the dataset.

4 Conclusion

Utilizing a combination of reaction templates, high-throughput DFT calculations, and graph neural networks, we have developed a predictive model capable of assessing the thermodynamic feasibility of hydrolysis reactions. Our current focus is on expanding the model's predictive capabilities to encompass acidic and basic pH conditions, which could prove invaluable in high-throughput screening of molecules and automated chemical synthesis for pH-dependent applications. The training and holdout test sets are publicly accessible through figshare and granular information regarding the individual reactant and product molecules is also available in the newly developed MPCules (45)interface. The code for training the model can be accessed at `https://github.com/HEPOM/HEPOM`.

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. 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 and the UCLA Hofmann2 compute cluster. 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] P. Arumugam, S. Gruber, K. Tanaka, C. H. Haering, K. Mechtler, and K. Nasmyth, "ATP Hydrolysis Is Required for Cohesin's Association with Chromosomes," *Current Biology*, vol. 13, pp. 1941–1953, Nov. 2003.

[2] J. Blazek and E. P. Gilbert, "Effect of Enzymatic Hydrolysis on Native Starch Granule Structure," *Biomacromolecules*, vol. 11, pp. 3275–3289, Dec. 2010. Publisher: American Chemical Society.

[3] B. A. Helms, "Polydiketoenamines for a Circular Plastics Economy," *Accounts of Chemical Research*, vol. 55, pp. 2753–2765, Oct. 2022.

[4] E. Olsson, C. Menzel, C. Johansson, R. Andersson, K. Koch, and L. Järnström, "The effect of pH on hydrolysis, cross-linking and barrier properties of starch barriers containing citric acid," *Carbohydrate Polymers*, vol. 98, pp. 1505–1513, Nov. 2013.

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

[6] B. Girisuta, L. P. B. M. Janssen, and H. J. Heeres, "Kinetic Study on the Acid-Catalyzed Hydrolysis of Cellulose to Levulinic Acid," *Industrial & Engineering Chemistry Research*, vol. 46, pp. 1696–1708, Mar. 2007. Publisher: American Chemical Society.

[7] D. L. Carlson, K. D. Than, and A. L. Roberts, "Acid- and Base-Catalyzed Hydrolysis of Chloroacetamide Herbicides," *Journal of Agricultural and Food Chemistry*, vol. 54, pp. 4740–4750, June 2006. Publisher: American Chemical Society.

[8] A. R. Epstein, J. Demarteau, B. A. Helms, and K. A. Persson, "Variable Amine Spacing Determines Depolymerization Rate in Polydiketoenamines," *Journal of the American Chemical Society*, vol. 145, pp. 8082–8089, Apr. 2023.

[9] D. K. Malick, G. A. Petersson, and J. A. Montgomery, "Transition states for chemical reactions I. Geometry and classical barrier height," *The Journal of Chemical Physics*, vol. 108, pp. 5704–5713, Apr. 1998.

[10] I. M. Alecu and D. G. Truhlar, "Computational Study of the Reactions of Methanol with the Hydroperoxyl and Methyl Radicals. 1. Accurate Thermochemistry and Barrier Heights," *The Journal of Physical Chemistry A*, vol. 115, pp. 2811–2829, Apr. 2011. Publisher: American Chemical Society.

[11] J. Aguilera-Iparraguirre, H. J. Curran, W. Klopper, and J. M. Simmie, "Accurate Benchmark Calculation of the Reaction Barrier Height for Hydrogen Abstraction by the Hydroperoxyl Radical from Methane. Implications for CnH2n+2 where n = $2 \rightarrow 4$," *The Journal of Physical Chemistry A*, vol. 112, pp. 7047–7054, July 2008. Publisher: American Chemical Society.

[12] M. G. Evans and M. Polanyi, "Further considerations on the thermodynamics of chemical equilibria and reaction rates," *Transactions of the Faraday Society*, vol. 32, p. 1333, 1936.

[13] T. Stuyver and C. W. Coley, "Machine Learning-Guided Computational Screening of New Candidate Reactions with High Bioorthogonal Click Potential," *Chemistry (Weinheim an Der Bergstrasse, Germany)*, vol. 29, p. e202300387, May 2023.

[14] S. Zhou, B. T. Nguyen, J. P. Richard, R. Kluger, and J. Gao, "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*, vol. 143, pp. 137–141, Jan. 2021. Publisher: American Chemical Society.

[15] K. E. Lawson, J. K. Dekle, and A. J. Adamczyk, "Towards pharmaceutical protein stabilization: DFT and statistical learning studies on non-enzymatic peptide hydrolysis degradation mechanisms," *Computational and Theoretical Chemistry*, vol. 1218, p. 113938, Dec. 2022.

[16] N. Mardirossian and M. Head-Gordon, "Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals," *Molecular Physics*, vol. 115, pp. 2315–2372, Oct. 2017.

[17] A. J. M. Ribeiro, M. J. Ramos, and P. A. Fernandes, "Benchmarking of DFT Functionals for the Hydrolysis of Phosphodiester Bonds," *Journal of Chemical Theory and Computation*, vol. 6, pp. 2281–2292, Aug. 2010. Publisher: American Chemical Society.

[18] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural Message Passing for Quantum Chemistry," *arXiv:1704.01212 [cs]*, June 2017. arXiv: 1704.01212.

[19] C. A. Grambow, L. Pattanaik, and W. H. Green, "Deep Learning of Activation Energies," *The Journal of Physical Chemistry Letters*, vol. 11, pp. 2992–2997, Apr. 2020.

[20] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, and P. Riley, "Molecular graph convolutions: moving beyond fingerprints," *Journal of Computer-Aided Molecular Design*, vol. 30, pp. 595–608, Aug. 2016.

[21] K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen, and R. Barzilay, "Analyzing Learned Molecular Representations for Property Prediction," *Journal of Chemical Information and Modeling*, vol. 59, pp. 3370–3388, Aug. 2019.

[22] E. Heid and W. H. Green, "Machine Learning of Reaction Properties via Learned Representations of the Condensed Graph of Reaction," *Journal of Chemical Information and Modeling*, vol. 62, pp. 2101–2110, May 2022.

[23] M. Wen, S. M. Blau, X. Xie, S. Dwaraknath, and K. A. Persson, "Improving machine learning performance on small chemical reaction data with unsupervised contrastive pretraining," *Chemical Science*, vol. 13, no. 5, pp. 1446–1458, 2022.

[24] C. A. Grambow, L. Pattanaik, and W. H. Green, "Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry," *Scientific Data*, vol. 7, p. 137, Dec. 2020.

[25] L. Ruddigkeit, R. Van Deursen, L. C. Blum, and J.-L. Reymond, "Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17," *Journal of Chemical Information and Modeling*, vol. 52, pp. 2864–2875, Nov. 2012.

[26] M. Wen, S. M. Blau, E. W. C. Spotte-Smith, S. Dwaraknath, and K. A. Persson, "BonDNet: a graph neural network for the prediction of bond dissociation energies for charged molecules," *Chemical Science*, vol. 12, no. 5, pp. 1858–1868, 2021.

[27] E. W. C. Spotte-Smith, S. M. Blau, X. Xie, H. D. Patel, M. Wen, B. Wood, S. Dwaraknath, and K. A. Persson, "Quantum chemical calculations of lithium-ion battery electrolyte and interphase species," *Scientific Data*, vol. 8, p. 203, Dec. 2021.

[28] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, "Quantum chemistry structures and properties of 134 kilo molecules," *Scientific Data*, vol. 1, p. 140022, 2014.

[29] G. Landrum, “RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling,”

[30] C. Tebes-Stevens, J. M. Patel, W. J. Jones, and E. J. Weber, “Prediction of hydrolysis products of organic chemicals under environmental pH conditions,” *Environmental science & technology*, vol. 51, pp. 5008–5016, May 2017.

[31] F. R. Dutra, C. d. S. Silva, and R. Custodio, “On the Accuracy of the Direct Method to Calculate pKa from Electronic Structure Calculations,” *The Journal of Physical Chemistry A*, vol. 125, pp. 65–73, Jan. 2021. Publisher: American Chemical Society.

[32] E. Epifanovsky, A. T. B. Gilbert, X. Feng, J. Lee, Y. Mao, N. Mardirossian, P. Pokhilko, A. F. White, M. P. Coons, A. L. Dempwolff, Z. Gan, D. Hait, P. R. Horn, L. D. Jacobson, I. Kaliman, J. Kussmann, A. W. Lange, K. U. Lao, D. S. Levine, J. Liu, S. C. McKenzie, A. F. Morrison, K. D. Nanda, F. Plasser, D. R. Rehn, M. L. Vidal, Z.-Q. You, Y. Zhu, B. Alam, B. J. Albrecht, A. Aldossary, E. Alguire, J. H. Andersen, V. Athavale, D. Barton, K. Begam, A. Behn, N. Bellonzi, Y. A. Bernard, E. J. Berquist, H. G. A. Burton, A. Carreras, K. Carter-Fenk, R. Chakraborty, A. D. Chien, K. D. Closser, V. Cofer-Shabica, S. Dasgupta, M. de Wergifosse, J. Deng, M. Diedenhofen, H. Do, S. Ehlert, P.-T. Fang, S. Fatehi, Q. Feng, T. Friedhoff, J. Gayvert, Q. Ge, G. Gidofalvi, M. Goldey, J. Gomes, C. E. González-Espinoza, S. Gulania, A. O. Gunina, M. W. D. Hanson-Heine, P. H. P. Harbach, A. Hauser, M. F. Herbst, M. Hernández Vera, M. Hodecker, Z. C. Holden, S. Houck, X. Huang, K. Hui, B. C. Huynh, M. Ivanov, Jász, H. Ji, H. Jiang, B. Kaduk, S. Kähler, K. Khistyaev, J. Kim, G. Kis, P. Klunzinger, Z. Koczor-Benda, J. H. Koh, D. Kosenkov, L. Koulias, T. Kowalczyk, C. M. Krauter, K. Kue, A. Kunitsa, T. Kus, I. Ladjánszki, A. Landau, K. V. Lawler, D. Lefrancois, S. Lehtola, R. R. Li, Y.-P. Li, J. Liang, M. Liebenthal, H.-H. Lin, Y.-S. Lin, F. Liu, K.-Y. Liu, M. Loipersberger, A. Luenser, A. Manjanath, P. Manohar, E. Mansoor, S. F. Manzer, S.-P. Mao, A. V. Marenich, T. Markovich, S. Mason, S. A. Maurer, P. F. McLaughlin, M. F. S. J. Menger, J.-M. Mewes, S. A. Mewes, P. Morgante, J. W. Mullinax, K. J. Oosterbaan, G. Paran, A. C. Paul, S. K. Paul, F. Pavošević, Z. Pei, S. Prager, E. I. Proynov, Rák, E. Ramos-Cordoba, B. Rana, A. E. Rask, A. Rettig, R. M. Richard, F. Rob, E. Rossomme, T. Scheele, M. Scheurer, M. Schneider, N. Sergueev, S. M. Sharada, W. Skomorowski, D. W. Small, C. J. Stein, Y.-C. Su, E. J. Sundstrom, Z. Tao, J. Thirman, G. J. Tornai, T. Tsuchimochi, N. M. Tubman, S. P. Veccham, O. Vydrov, J. Wenzel, J. Witte, A. Yamada, K. Yao, S. Yeganeh, S. R. Yost, A. Zech, I. Y. Zhang, X. Zhang, Y. Zhang, D. Zuev, A. Aspuru-Guzik, A. T. Bell, N. A. Besley, K. B. Bravaya, B. R. Brooks, D. Casanova, J.-D. Chai, S. Coriani, C. J. Cramer, G. Cserey, A. E. DePrince, III, R. A. DiStasio, Jr., A. Dreuw, B. D. Dunietz, T. R. Furlani, W. A. Goddard, III, S. Hammes-Schiffer, T. Head-Gordon, W. J. Hehre, C.-P. Hsu, T.-C. Jagau, Y. Jung, A. Klamt, J. Kong, D. S. Lambrecht, W. Liang, N. J. Mayhall, C. W. McCurdy, J. B. Neaton, C. Ochsenfeld, J. A. Parkhill, R. Peverati, V. A. Rassolov, Y. Shao, L. V. Slipchenko, T. Stauch, R. P. Steele, J. E. Subotnik, A. J. W. Thom, A. Tkatchenko, D. G. Truhlar, T. Van Voorhis, T. A. Wesolowski, K. B. Whaley, H. L. Woodcock, III, P. M. Zimmerman, S. Faraji, P. M. W. Gill, M. Head-Gordon, J. M. Herbert, and A. I. Krylov, “Software for the frontiers of quantum chemistry: An overview of developments in the Q-Chem 5 package,” *The Journal of Chemical Physics*, vol. 155, p. 084801, Aug. 2021.

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

[34] N. Mardirossian and M. Head-Gordon, “ B97M-V: A combinatorially optimized, range-separated hybrid, meta-GGA density functional with VV10 nonlocal correlation,” *The Journal of Chemical Physics*, vol. 144, p. 214110, June 2016.

[35] O. A. Vydrov and T. Van Voorhis, “Nonlocal van der Waals density functional: The simpler the better,” *The Journal of Chemical Physics*, vol. 133, p. 244103, Dec. 2010.

[36] A. Hellweg and D. Rappoport, “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*, vol. 17, pp. 1010–1017, Dec. 2014. Publisher: The Royal Society of Chemistry.

[37] A. V. Marenich, C. J. Cramer, and D. G. Truhlar, "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*, vol. 113, pp. 6378–6396, May 2009. Publisher: American Chemical Society.

[38] X. Bresson and T. Laurent, "Residual Gated Graph ConvNets," Apr. 2018. arXiv:1711.07553 [cs, stat].

[39] V. P. Dwivedi, C. K. Joshi, A. T. Luu, T. Laurent, Y. Bengio, and X. Bresson, "Benchmarking Graph Neural Networks," Dec. 2022. arXiv:2003.00982 [cs, stat].

[40] O. Vinyals, S. Bengio, and M. Kudlur, "Order Matters: Sequence to sequence for sets," Feb. 2016. arXiv:1511.06391 [cs, stat].

[41] K. Zhou, Y. Dong, K. Wang, W. S. Lee, B. Hooi, H. Xu, and J. Feng, "Understanding and Resolving Performance Degradation in Graph Convolutional Networks," Sept. 2021. arXiv:2006.07107 [cs, stat].

[42] T. K. Rusch, M. M. Bronstein, and S. Mishra, "A Survey on Oversmoothing in Graph Neural Networks," Mar. 2023. arXiv:2303.10993 [cs].

[43] E. Heid, K. P. Greenman, Y. Chung, S.-C. Li, D. E. Graff, F. H. Vermeire, H. Wu, W. H. Green, and C. J. McGill, "Chemprop: Machine Learning Package for Chemical Property Prediction," preprint, Chemistry, July 2023.

[44] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," Sept. 2020. arXiv:1802.03426 [cs, stat].

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