HEPOM: A predictive framework for accelerated Hydrolysis Energy Predictions of Organic Molecules

Rishabh D. Guha * Lawrence Berkeley National Laboratory

Evan Walter Clark Spotte-Smith University of California, Berkeley Lawrence Berkeley National Laboratory **Santiago Vargas** * University of California, Los Angeles

Alexander R. Epstein University of California, Berkeley

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

1 Introduction

Water is arguably the most widely known compound, and yet, its deceptively simple structure fails to underscore 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, 2] and synthetic chemistry[3, 4], water doubles as a reactant and solvent medium in which the reaction proceeds. At the molecular level, hydrolysis is initiated by the attack of a water, hydronium, or hydroxide molecule on specific sites in the reactant, triggering a sequence of bond cleavage and formations, which in turn, leads to the formation of new product(s). The thermodynamic feasibility of this reaction is fundamentally tied to the pH of the aqueous reaction medium [5, 6]. 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 [7, 8] of the same reactant can have prominently different reaction rates than its neutral counterparts.

The Eyring equation provides a means to quantify experimental reaction rates by evaluating activation barriers (ΔG^{\ddagger}) through computational methods [9, 10]. However, this approach demands computationally intensive transition state (TS) calculations for each reaction along the complex potential energy surface (PES) [11, 12]. In contrast, within a specific reaction family, the Bell-Evans-Polanyi principle [13] can offer a qualitative linear correlation between the thermodynamic Gibbs Free Energy change (ΔG_r) and the kinetic parameter ΔG^{\ddagger} [14, 15, 16]. 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 [17, 18]. 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 stepped up as promising alternatives in the past decade, especially for tasks that involve the establishment of structure–property relationships [19, 20, 21]. 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 [22, 23] and reaction representations [24, 25]. Despite these methodological advances, the largest roadblock to the development of an accurate model is typically the procurement of diverse, problem-representative data. For instance, the model developed by Grambow et al. [21] was facilitated by a dataset of 12,000 gas-phase reactions [26] sampled from a subset of molecules in the GDB-17 dataset [27]. The bond dissociation energy (BDE) prediction framework developed by Wen et al. [28] was trained on a dataset of over 60,000 homolytic and heterolytic bond dissociation reactions [29]. In the realm of hydrolysis, no such comprehensive dataset currently exists.

In this work, we first develop 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 [30] 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 [31] 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. [32] into an automated framework for determining reaction products. For instance, as shown in Schematic S1 of the SI, 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 of the SI, 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. 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 [33].

2.2 Density-Functional Theory

The electronic structure code QChem (version 5 or 6) [34] was used to perform all the DFT calculations necessary to generate the dataset. A specialized frequency-flattening optimization (FF0pt) workflow, originally developed by Spotte-Smith et al. [29] and currently implemented in atomate [35] 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 can parse the enthalpy and entropy from the QChem frequency output document, necessary for the free energy calculations. The range-separated meta-GGA hybrid functional, ω B97M-V [36], which employs the vv10 dispersion correction [37] for improving the non-covalent interactions was used for all the calculations. In case of the geometry optimizations and frequency calculations (FFOpt workflow), the def2-SVPD basis set [38] was used with the water SMD solvent model [39] to implicitly account for solvation effects. The electronic energies of the optimized structures were refined with single-point calculations using a larger def2-QZVPPD basis set [38].

2.3 Model Architecture

The GNN model, visually depicted in Figure S4 of the SI, is heavily based on the previously published BonDNet architecture [28]. 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-property models in chemistry and materials science [40, 41], 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 molecularlevel 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. In this implementation, the reaction mapping is altered from the original BonDNet as a reaction graph is constructed between the union set of bonds in products and reactants. The small change allows us to treat reactions where bonds both break and form as well as reactions where several bonds change (Fig. S4). From here, a set2set [42] 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 order to leverage the consistent reaction framework of hydrolysis, we made one more fundamental change to BonDNet: we incorporated 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. Additionally, an expanded global feature serves as a "memory bank" to allow the rapid propagation of messages between nodes further away from each other. This is a particularly attractive feature as sequential stacking of message-passing layers, alternative to better propagate further

relationships in the graph, rapidly increases compute time and can also lead to problems such as oversmoothing [43, 44]. With this modification, we limit the number of graph message-passing layers to 1-2 layers in our model architecture, and thus, drastically reduce the number of parameters.

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 hydrolysis pathway while stable 5-membered rings are much 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 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 robustness, 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 (y=x) with a high coefficient of determination (R^2) for both the validation and test sets. The model performance on the test set demonstrates impressive generalizability, achieving an R^2 of 0.92 and a Mean Absolute Error (MAE) of 2.25 kcal/mol compared to the DFT-calculated values (Figure 2(a)). The classification accuracy for the model correctly classifying reactions as endergonic vs. exergonic was also 95.3

3.3 Model Embeddings

To investigate the model-learned representations of the hydrolysis reactions, we reduced the highdimensional difference feature vectors for each hydrolysis reaction into a two-dimensional (2D) space using the uniform manifold approximation and projection (UMAP) method [45]. 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 separated clusters and we hypothesize that these are the distinct endothermic and exothermic reactions. This implies that the model also learns to distinguish separate subclasses 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.



Figure 3: UMAP embedding of the reaction features

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 dataset will be publicly accessible and the code can be accessed at https://github.com/santi921/bondnet.

References

- Prakash Arumugam, Stephan Gruber, Koichi Tanaka, Christian H. Haering, Karl Mechtler, and Kim Nasmyth. ATP Hydrolysis Is Required for Cohesin's Association with Chromosomes. *Current Biology*, 13(22):1941–1953, November 2003.
- [2] David Y. Hui. Molecular biology of enzymes involved with cholesterol ester hydrolysis in mammalian tissues. *Biochimica et Biophysica Acta (BBA) - Lipids and Lipid Metabolism*, 1303(1):1–14, September 1996.
- [3] Jaroslav Blazek and Elliot Paul Gilbert. Effect of Enzymatic Hydrolysis on Native Starch Granule Structure. *Biomacromolecules*, 11(12):3275–3289, December 2010. Publisher: American Chemical Society.
- [4] Brett A. Helms. Polydiketoenamines for a Circular Plastics Economy. Accounts of Chemical Research, 55(19):2753–2765, October 2022.
- [5] Erik Olsson, Carolin Menzel, Caisa Johansson, Roger Andersson, Kristine Koch, and Lars Järnström. The effect of pH on hydrolysis, cross-linking and barrier properties of starch barriers containing citric acid. *Carbohydrate Polymers*, 98(2):1505–1513, November 2013.
- [6] Jeremy Demarteau, Alexander R. Epstein, Peter R. Christensen, Mark Abubekerov, Hai Wang, Simon J. Teat, Trevor J. Seguin, Christopher W. Chan, Corinne D. Scown, Thomas P. Russell, Jay D. Keasling, Kristin A. Persson, and Brett A. Helms. Circularity in mixed-plastic chemical recycling enabled by variable rates of polydiketoenamine hydrolysis. *Science Advances*, 8(29):eabp8823, July 2022.
- [7] 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*, 46(6):1696–1708, March 2007. Publisher: American Chemical Society.
- [8] Daniel L. Carlson, Khoi D. Than, and A. Lynn Roberts. Acid- and Base-Catalyzed Hydrolysis of Chloroacetamide Herbicides. *Journal of Agricultural and Food Chemistry*, 54(13):4740–4750, June 2006. Publisher: American Chemical Society.
- [9] Alexander R. Epstein, Jeremy Demarteau, Brett A. Helms, and Kristin A. Persson. Variable Amine Spacing Determines Depolymerization Rate in Polydiketoenamines. *Journal of the American Chemical Society*, 145(14):8082–8089, April 2023.
- [10] David K. Malick, G. A. Petersson, and John A. Montgomery. Transition states for chemical reactions I. Geometry and classical barrier height. *The Journal of Chemical Physics*, 108(14):5704–5713, April 1998.
- [11] I. M. Alecu and Donald 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*, 115(13):2811–2829, April 2011. Publisher: American Chemical Society.
- [12] Jorge Aguilera-Iparraguirre, Henry J. Curran, Wim Klopper, and John 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*, 112(30):7047–7054, July 2008. Publisher: American Chemical Society.
- [13] M. G. Evans and M. Polanyi. Further considerations on the thermodynamics of chemical equilibria and reaction rates. *Transactions of the Faraday Society*, 32:1333, 1936.
- [14] Thijs Stuyver and Connor W. Coley. Machine Learning-Guided Computational Screening of New Candidate Reactions with High Bioorthogonal Click Potential. *Chemistry (Weinheim an Der Bergstrasse, Germany)*, 29(28):e202300387, May 2023.
- [15] Shaoyuan Zhou, Bach T. Nguyen, John P. Richard, Ronald Kluger, and Jiali 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*, 143(1):137–141, January 2021. Publisher: American Chemical Society.

- [16] Katherine E. Lawson, Joseph K. Dekle, and Andrew J. Adamczyk. Towards pharmaceutical protein stabilization: DFT and statistical learning studies on non-enzymatic peptide hydrolysis degradation mechanisms. *Computational and Theoretical Chemistry*, 1218:113938, December 2022.
- [17] Narbe Mardirossian and Martin Head-Gordon. Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. *Molecular Physics*, 115(19):2315–2372, October 2017.
- [18] António J. M. Ribeiro, Maria J. Ramos, and Pedro A. Fernandes. Benchmarking of DFT Functionals for the Hydrolysis of Phosphodiester Bonds. *Journal of Chemical Theory and Computation*, 6(8):2281–2292, August 2010. Publisher: American Chemical Society.
- [19] Mingjian Wen, Matthew K. Horton, Jason M. Munro, Patrick Huck, and Kristin A. Persson. A universal equivariant graph neural network for the elasticity tensors of any crystal system, July 2023. arXiv:2307.15242 [cond-mat].
- [20] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural Message Passing for Quantum Chemistry. arXiv:1704.01212 [cs], June 2017. arXiv: 1704.01212.
- [21] Colin A. Grambow, Lagnajit Pattanaik, and William H. Green. Deep Learning of Activation Energies. *The Journal of Physical Chemistry Letters*, 11(8):2992–2997, April 2020.
- [22] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design*, 30(8):595–608, August 2016.
- [23] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing Learned Molecular Representations for Property Prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, August 2019.
- [24] Esther Heid and William H. Green. Machine Learning of Reaction Properties via Learned Representations of the Condensed Graph of Reaction. *Journal of Chemical Information and Modeling*, 62(9):2101–2110, May 2022.
- [25] Mingjian Wen, Samuel M Blau, Xiaowei Xie, Shyam Dwaraknath, and Kristin A Persson. Improving machine learning performance on small chemical reaction data with unsupervised contrastive pretraining. *Chemical Science*, 13:1446–1458, 2022.
- [26] Colin A. Grambow, Lagnajit Pattanaik, and William H. Green. Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry. *Scientific Data*, 7(1):137, December 2020.
- [27] Lars Ruddigkeit, Ruud Van Deursen, Lorenz C. Blum, and Jean-Louis Reymond. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *Journal of Chemical Information and Modeling*, 52(11):2864–2875, November 2012.
- [28] Mingjian Wen, Samuel M. Blau, Evan Walter Clark Spotte-Smith, Shyam Dwaraknath, and Kristin A. Persson. BonDNet: a graph neural network for the prediction of bond dissociation energies for charged molecules. *Chemical Science*, 12(5):1858–1868, 2021.
- [29] Evan Walter Clark Spotte-Smith, Samuel M. Blau, Xiaowei Xie, Hetal D. Patel, Mingjian Wen, Brandon Wood, Shyam Dwaraknath, and Kristin Aslaug Persson. Quantum chemical calculations of lithium-ion battery electrolyte and interphase species. *Scientific Data*, 8(1):203, December 2021.
- [30] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1:140022, 2014.

- [31] Greg Landrum. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling.
- [32] Caroline Tebes-Stevens, Jay M. Patel, W. Jack Jones, and Eric J. Weber. Prediction of hydrolysis products of organic chemicals under environmental pH conditions. *Environmental science & technology*, 51(9):5008–5016, May 2017.
- [33] Felipe Ribeiro Dutra, Cleuton de Souza Silva, and Rogério Custodio. On the Accuracy of the Direct Method to Calculate pKa from Electronic Structure Calculations. *The Journal of Physical Chemistry A*, 125(1):65–73, January 2021. Publisher: American Chemical Society.
- [34] Evgeny Epifanovsky, Andrew T. B. Gilbert, Xintian Feng, Joonho Lee, Yuezhi Mao, Narbe Mardirossian, Pavel Pokhilko, Alec F. White, Marc P. Coons, Adrian L. Dempwolff, Zhengting Gan, Diptarka Hait, Paul R. Horn, Leif D. Jacobson, Ilya Kaliman, Jörg Kussmann, Adrian W. Lange, Ka Un Lao, Daniel S. Levine, Jie Liu, Simon C. McKenzie, Adrian F. Morrison, Kaushik D. Nanda, Felix Plasser, Dirk R. Rehn, Marta L. Vidal, Zhi-Qiang You, Ying Zhu, Bushra Alam, Benjamin J. Albrecht, Abdulrahman Aldossary, Ethan Alguire, Josefine H. Andersen, Vishikh Athavale, Dennis Barton, Khadiza Begam, Andrew Behn, Nicole Bellonzi, Yves A. Bernard, Eric J. Berquist, Hugh G. A. Burton, Abel Carreras, Kevin Carter-Fenk, Romit Chakraborty, Alan D. Chien, Kristina D. Closser, Vale Cofer-Shabica, Saswata Dasgupta, Marc de Wergifosse, Jia Deng, Michael Diedenhofen, Hainam Do, Sebastian Ehlert, Po-Tung Fang, Shervin Fatehi, Qingguo Feng, Triet Friedhoff, James Gayvert, Qinghui Ge, Gergely Gidofalvi, Matthew Goldey, Joe Gomes, Cristina E. González-Espinoza, Sahil Gulania, Anastasia O. Gunina, Magnus W. D. Hanson-Heine, Phillip H. P. Harbach, Andreas Hauser, Michael F. Herbst, Mario Hernández Vera, Manuel Hodecker, Zachary C. Holden, Shannon Houck, Xunkun Huang, Kerwin Hui, Bang C. Huynh, Maxim Ivanov, Ádám Jász, Hyunjun Ji, Hanjie Jiang, Benjamin Kaduk, Sven Kähler, Kirill Khistyaev, Jaehoon Kim, Gergely Kis, Phil Klunzinger, Zsuzsanna Koczor-Benda, Joong Hoon Koh, Dimitri Kosenkov, Laura Koulias, Tim Kowalczyk, Caroline M. Krauter, Karl Kue, Alexander Kunitsa, Thomas Kus, István Ladjánszki, Arie Landau, Keith V. Lawler, Daniel Lefrancois, Susi Lehtola, Run R. Li, Yi-Pei Li, Jiashu Liang, Marcus Liebenthal, Hung-Hsuan Lin, You-Sheng Lin, Fenglai Liu, Kuan-Yu Liu, Matthias Loipersberger, Arne Luenser, Aaditya Manjanath, Prashant Manohar, Erum Mansoor, Sam F. Manzer, Shan-Ping Mao, Aleksandr V. Marenich, Thomas Markovich, Stephen Mason, Simon A. Maurer, Peter F. McLaughlin, Maximilian F. S. J. Menger, Jan-Michael Mewes, Stefanie A. Mewes, Pierpaolo Morgante, J. Wayne Mullinax, Katherine J. Oosterbaan, Garrette Paran, Alexander C. Paul, Suranjan K. Paul, Fabijan Pavošević, Zheng Pei, Stefan Prager, Emil I. Provnov, Ádám Rák, Eloy Ramos-Cordoba, Bhaskar Rana, Alan E. Rask, Adam Rettig, Ryan M. Richard, Fazle Rob, Elliot Rossomme, Tarek Scheele, Maximilian Scheurer, Matthias Schneider, Nickolai Sergueev, Shaama M. Sharada, Wojciech Skomorowski, David W. Small, Christopher J. Stein, Yu-Chuan Su, Eric J. Sundstrom, Zhen Tao, Jonathan Thirman, Gábor J. Tornai, Takashi Tsuchimochi, Norm M. Tubman, Srimukh Prasad Veccham, Oleg Vydrov, Jan Wenzel, Jon Witte, Atsushi Yamada, Kun Yao, Sina Yeganeh, Shane R. Yost, Alexander Zech, Igor Ying Zhang, Xing Zhang, Yu Zhang, Dmitry Zuev, Alán Aspuru-Guzik, Alexis T. Bell, Nicholas A. Besley, Ksenia B. Bravaya, Bernard R. Brooks, David Casanova, Jeng-Da Chai, Sonia Coriani, Christopher J. Cramer, György Cserey, A. Eugene DePrince, III, Robert A. DiStasio, Jr., Andreas Dreuw, Barry D. Dunietz, Thomas R. Furlani, William A. Goddard, III, Sharon Hammes-Schiffer, Teresa Head-Gordon, Warren J. Hehre, Chao-Ping Hsu, Thomas-C. Jagau, Yousung Jung, Andreas Klamt, Jing Kong, Daniel S. Lambrecht, WanZhen Liang, Nicholas J. Mayhall, C. William McCurdy, Jeffrey B. Neaton, Christian Ochsenfeld, John A. Parkhill, Roberto Peverati, Vitaly A. Rassolov, Yihan Shao, Lyudmila V. Slipchenko, Tim Stauch, Ryan P. Steele, Joseph E. Subotnik, Alex J. W. Thom, Alexandre Tkatchenko, Donald G. Truhlar, Troy Van Voorhis, Tomasz A. Wesolowski, K. Birgitta Whaley, H. Lee Woodcock, III, Paul M. Zimmerman, Shirin Faraji, Peter M. W. Gill, Martin Head-Gordon, John M. Herbert, and Anna I. Krylov. Software for the frontiers of quantum chemistry: An overview of developments in the Q-Chem 5 package. The Journal of Chemical Physics, 155(8):084801, August 2021.
- [35] Kiran Mathew, Joseph H. Montoya, Alireza Faghaninia, Shyam Dwarakanath, Muratahan Aykol, Hanmei Tang, Iek-heng Chu, Tess Smidt, Brandon Bocklund, Matthew Horton, John Dagdelen, Brandon Wood, Zi-Kui Liu, Jeffrey Neaton, Shyue Ping Ong, Kristin Persson, and Anubhav Jain. Atomate: A high-level interface to generate, execute, and analyze computational materials science workflows. *Computational Materials Science*, 139:140–152, November 2017.

- [36] Narbe Mardirossian and Martin Head-Gordon. ω B97M-V: A combinatorially optimized, range-separated hybrid, meta-GGA density functional with VV10 nonlocal correlation. *The Journal of Chemical Physics*, 144(21):214110, June 2016.
- [37] Oleg A. Vydrov and Troy Van Voorhis. Nonlocal van der Waals density functional: The simpler the better. *The Journal of Chemical Physics*, 133(24):244103, December 2010.
- [38] Arnim Hellweg and Dmitrij 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*, 17(2):1010–1017, December 2014. Publisher: The Royal Society of Chemistry.
- [39] Aleksandr V. Marenich, Christopher J. Cramer, and Donald 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*, 113(18):6378–6396, May 2009. Publisher: American Chemical Society.
- [40] Xavier Bresson and Thomas Laurent. Residual Gated Graph ConvNets, April 2018. arXiv:1711.07553 [cs, stat].
- [41] Vijay Prakash Dwivedi, Chaitanya K. Joshi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Benchmarking Graph Neural Networks, December 2022. arXiv:2003.00982 [cs, stat].
- [42] Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. Order Matters: Sequence to sequence for sets, February 2016. arXiv:1511.06391 [cs, stat].
- [43] Kuangqi Zhou, Yanfei Dong, Kaixin Wang, Wee Sun Lee, Bryan Hooi, Huan Xu, and Jiashi Feng. Understanding and Resolving Performance Degradation in Graph Convolutional Networks, September 2021. arXiv:2006.07107 [cs, stat].
- [44] T. Konstantin Rusch, Michael M. Bronstein, and Siddhartha Mishra. A Survey on Oversmoothing in Graph Neural Networks, March 2023. arXiv:2303.10993 [cs].
- [45] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018.