Enabling Accurate and High-Throughput Kinetics Predictions via Message Passing Neural Networks

by

Kevin A. Spiekermann

B.S. Chemical Engineering
University of California, San Diego (2018)

M.S. Chemical Engineering Practice
Massachusetts Institute of Technology (2021)

Submitted to the Department of Chemical Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Chemical Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2023

Authored by . . . .
Department of Chemical Engineering
August 10, 2023

Certified by . . . .
William H. Green
Hoyt C. Hottel Professor in Chemical Engineering
Thesis Supervisor

Accepted by . . . .
Hadley D. Sikes
Esther and Harold E. Edgerton Professor in Chemical Engineering
Graduate Officer

Enabling Accurate and High-Throughput Kinetics Predictions via Message Passing Neural Networks

by

Kevin A. Spiekermann

Submitted to the Department of Chemical Engineering
on August 10, 2023, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Chemical Engineering

Abstract

Quantitative estimates for kinetic properties, namely reaction barrier heights and reaction energies, are essential for developing kinetic mechanisms, predicting reaction outcomes, and optimizing chemical processes. While ab initio methods, such as quantum chemistry, can be incredibly useful for providing accurate kinetic data, their high computational cost severely limits their utility for large-scale applications. High-quality experimental data is often even more rare, and such approaches are less amenable to exploring the vastness of chemical space due to monetary cost, time, and safety considerations. Modern machine learning (ML) techniques offer a promising option since they can quickly provide estimates to narrow the search space for more expensive ab initio or experimental methods. Unfortunately, the paucity of reliable quantitative chemical reaction data to train such models has presented a major hindrance for these data-driven approaches.

Here, this thesis focuses on the intersection of ML and quantum chemistry with the goal of enabling automatic high-fidelity predictions of kinetic parameters. The novel contributions can be grouped into three main categories:

1. Large-scale dataset generation, with an emphasis on high-quality methods and reaction diversity. Although much of the presented work studies reactions in the gas phase, this thesis also contributes a large dataset calculated in many popular solvents.

2. Train various ML models to quickly predict accurate kinetic parameters, which avoids the challenging task of finding transition state structures. Importantly, these models operate on simple input representations and hence are ideal for automated, high-throughput applications.

3. Provide best-practice guidelines and an open-source software package to improve the status quo of ML for chemistry research.

The contributions from this thesis, and from similar work, will be essential for modern high-throughput workflows and the future of automated predictive chemistry.

Thesis Supervisor: William H. Green
Title: Hoyt C. Hottel Professor in Chemical Engineering

Acknowledgments

There are numerous people who have contributed to my education and success at MIT. Firstly, I would like to express my gratitude to Professor William Green for serving as my graduate advisor. Bill has been both a great mentor and leader. His guidance, support, and detailed feedback have been instrumental throughout my research journey and have shaped my intellectual growth. Bill's deep knowledge of kinetics is impressive and proved invaluable for numerous projects, though his intuition to quickly identify issues in a kinetic model and subsequently offer helpful suggestions is arguably even more impressive. Thank you for consistently showing positive outlook and supporting me as I pursued my own research ideas.

I would also like to thank Professor Kris Prather. I had actually started my Ph.D. program by joining her research group. Kris is also a brilliant advisor and an incredibly supportive mentor. Although I realized I would rather explore research questions computationally and thus switched into Bill's group early on, I'm still grateful for the time I spent in her lab. Switching labs is a stressful process, so I'm very thankful for both Kris and Bill for making this a smooth transition.

I would like to thank my thesis committee members. Their suggestions and guidance helped improve the quality of my thesis. Professor Connor Coley consistently provided thoughtful feedback during each of our committee meetings and for all of the corresponding written progress reports. Your advice, analysis, and questions have been very valuable. Professor Ahmed Ghoniem similarly provided useful guidance. Professor Alon Dana was a former post-doc in our group, and I am grateful to have had the opportunity to work closely with and learn from him. Alon has insightful knowledge of kinetics and software engineering. He was eager to learn new topics and consistently demonstrated a passion to teach and mentor others.

I also want to thank Gwen Wilcox for quickly solving countless non-research related issues. Her logistical prowess and impressive organization provided valuable support to ensure that we are all working productively. Gwen has received numerous department awards at MIT for consistently bringing a positive energy and genuinely caring about everyone.

I have worked with many collaborators throughout my Ph.D, both within the Green Group and externally. In particular, I am very grateful to have closely worked with Lagnajit (Lucky) Pattanaik on multiple projects. He took the time to mentor me, answer a myriad of questions about machine learning, provide valuable feedback, and encourage me to think and grow as a researcher. I also had useful brainstorming sessions with Chas McGill, Esther Heid, Oscar Wu, and Kevin Greenman related to machine learning. Mark Payne, Matt

Johnson, Mengjie (Max) Liu, Hao-Wei Pang, and Xiaorui Dong helped me learn about our group's Reaction Mechanism Generator software. Thank you to Duminda Ranasinghe, Xiaorui Dong, Shih-Cheng Li, and Angiras Menon for answering several questions about quantum chemistry. I learned a lot about software engineering from Jackson Burns, who was an exceptional collaborator as we combined our prior work and ideas to develop an open-source Python package. Although I didn't directly collaborate on a project with Ryan Hawtof, Sayandeep Biswas, Jonathan Zheng, Rob Jones, or Kariana Moreno, we still had many valuable and fun discussions. To all Green Group members, thank you for creating such a positive community. It has been a great pleasure to interact with all of you; from the 'dles of the day to our gym sessions and delicious fun food outings, you made my Ph.D. journey much more enjoyable.

I also want to thank Xiao Liu from Professor Martin Head-Gordon's group U.C. Berkeley for collaborating as we benchmarked different quantum methods. It was a pleasure to work with our many collaborators at BASF–Maike Bergeler, Mark Pfeifle, Frederik Sandfort, and Oliver Welz.

I want to thank all members of my first-year cohort for making this experience worthwhile. My special thanks goes to Marc Berliner, Alexi Khechfe, Kaylee McCormack, Bert Neyhouse, and Mo Alkhadra and many others from the ChemE department with whom I shared fun memories. I also thank Aditya Limaye for being a great first-year roommate. I want to thank both Practice School teams for being amazing teammates and friends. Likewise, I enjoyed meeting people on the Sidney Pacific Student Government and at various events. I'd like to give a shout-out to Team Manatee for some rigorous swim workouts that helped keep me healthy as well to many other friends outside of MIT.

Importantly, I want to thank my mom and dad. None of this work would have been possible without your unconditional love and support. Last but not least, I want to express my love and appreciation for my partner, Courtney Cochrane. Thank you for supporting me through all the ups and downs. We've made so many great memories together, and I'll cherish them all.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Accurately predicting the time evolution of reacting chemical systems, as well as the yields of various products and side products, has been one of the main goals of physical chemistry for more than 135 years [1, 2]. Detailed kinetic mechanisms aim to achieve this goal by enumerating all relevant elementary reactions to make quantitative predictions for diverse chemical systems. A crucial step in developing a kinetic mechanism is to accurately compute the thermochemical and kinetic parameters of the relevant species and reactions. Historically, the field has relied on regressing these parameters from a relatively narrow set of experimental measurements. Not only are experiments often time-consuming, expensive, and have several safety considerations, the fitted parameters from this approach generally results in poor extrapolation when trying to model new systems. Due to the impressive accuracy of quantum mechanical (QM) methods and growth of compute power, it has become possible to accurately compute rate coefficients for individual reactions using first principles methods like transition state theory (TST) [3, 4], thus allowing the field of chemical kinetics to begin transitioning from a post-dictive to predictive modeling approach that can better extrapolate to new reaction conditions [5]. Indeed, computational approaches have enabled successful generation of complex mechanisms whose predictions can give good agreement with experimental measurements [6–15].

Although this progress is exciting, the QM workflow involves many steps for both the stable species and transition state (TS), including conformer generation, geometry optimization, frequency calculation, and refining the energy via accurate single-point calculation. Furthermore, identifying a suitable TS geometry that can successfully optimize to a structure that passes an intrinsic reaction coordinate verification is a non-trivial task. Given that each of these steps is computationally expensive, this workflow is poorly suited to provide kinetic estimates for the tens of thousands of reactions that must often be considered during automatic mechanism generation [16]. In reality, it is common for only a small fraction of the numerous rate coefficients in these mechanisms to be computed accurately, which can make the predictions somewhat erratic. The vision of reliable predictive chemical kinetics can only be realized if accurate rate coefficients can be calculated much more rapidly than is possible today. Hence, it is prudent to explore how data-driven methods can accelerate portions of this workflow to enable efficient identification of relevant low-barrier reactions. Ideally, the predictions are already within an acceptable error tolerance, but, depending on the application, it may be worth spending the compute time to calculate this much smaller subset of reactions with very high accuracy. The goal of this thesis is to accelerate high-throughput predictive chemistry via machine learning by providing accurate estimation of parameters relevant in automated mechanism generation.

1.1 Automated Mechanism Generation for Quantitative Predictive Chemistry

Detailed chemical reaction mechanisms help explain physical phenomena driven by chemical kinetics and are recognized as a necessary tool for chemical selection and process optimization. A reaction mechanism is a list of species and reactions, ideally elementary, along with their thermochemical and kinetic parameters respectively; transport parameters may also be necessary depending on the system and modeling approach. Together, these define a system of ordinary differential equations that can be numerically integrated to simulate the time evolution of the system. A schematic of this workflow is shown in Figure 1.1. This framework is powerful as it allows researchers to explore these systems computationally, which should be cheaper, faster, and safer in comparison to running experiments. With this increased understanding, kineticists can then offer guidance regarding how to manipulate these dynamics to obtain a desired outcome.

There are two very important questions that must be answered in order to successfully construct a kinetic model. First, how do we know which species and reactions are relevant to the system of interest? Once those have been defined, how do we know what thermochemical and kinetic values to use, which ultimately become coefficients in the system of equations? This thesis will primarily focus on the second question. However, these questions are related so it is informative to provide a brief overview for each.

Regarding the first task, mechanism development was historically a manual process that required people with an expert understanding of the niche system to tediously enumerate all

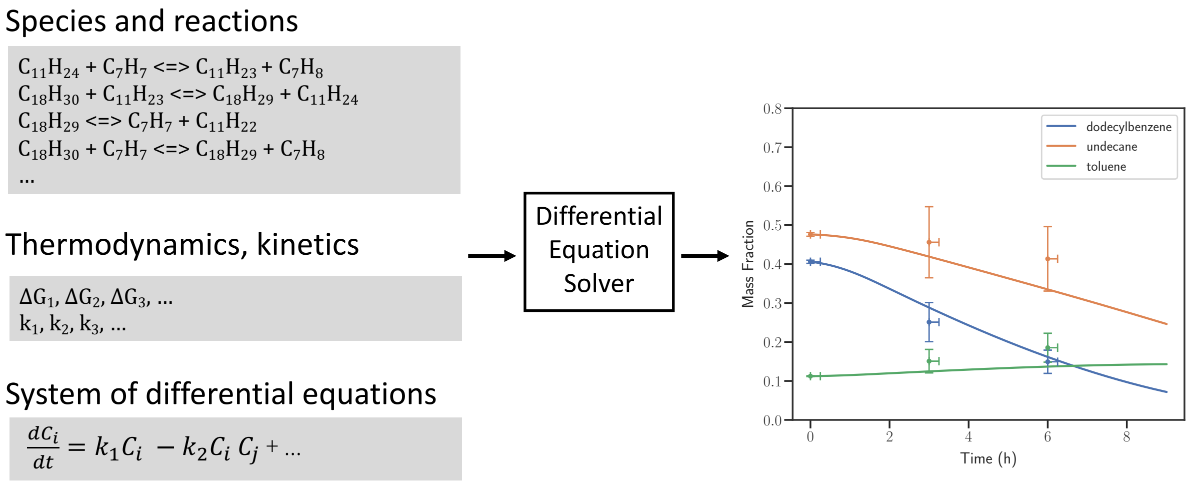

Species and reactions
$C_{11}H_{24}$ + C_7H_7 <=> $C_{11}H_{23}$ + C_7H_8
$C_{18}H_{30}$ + $C_{11}H_{23}$ <=> $C_{18}H_{29}$ + $C_{11}H_{24}$
$C_{18}H_{29}$ <=> C_7H_7 + $C_{11}H_{22}$
$C_{18}H_{30}$ + C_7H_7 <=> $C_{18}H_{29}$ + C_7H_8
...
Thermodynamics, kinetics
$\Delta G_1, \Delta G_2, \Delta G_3$, ...
k_1, k_2, k_3, ...
System of differential equations
$\frac{dC_i}{dt} = k_1 C_i - k_2 C_i C_j + \dots$
Differential Equation Solver
Mass Fraction
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Time (h)
0
2
4
6
8
dodecylbenzene
undecane
toluene

Figure 1.1. Schematic showing that integrating the differential equations that comprise a chemical kinetic model yields the time evolution of the system. The plot on the right is taken from ref. [14].

of the relevant species and reactions. This approach is not only error-prone, but it also does not scale well with larger systems or more complex chemistry. In recent decades, the development of computational methods for building reaction mechanisms has helped transition mechanism development to an automated procedure that leverages theoretical insight and experimental data. Several software packages exist for generating reaction mechanisms automatically, with examples including NETGEN [17, 18], Reaction [19], EXGAS [20], Genesys [21], and the Reaction Mechanism Generator (RMG) [22, 23]. Since RMG has become the most widely-used software, and it is also developed by our research group, portions of this thesis are dedicated to improving RMG.

RMG is an open-source software that constructs kinetic models composed of elementary chemical reaction steps using domain knowledge of how molecules react. A schematic overview of RMG's mechanism generation workflow is shown in Figure 1.2. After identifying a system of interest, RMG operates by taking the initial conditions (species concentrations, temperature, and pressure) as input. From here, RMG relies on two main components to generate the mechanism. The first component is RMG's database, which stores thermodynamic and kinetic values. If the database happens to have the value for the exact species or reaction that is under consideration for the mechanism, then RMG will directly query that value. However, more commonly, RMG uses simple data-driven estimators to quickly estimate the values, which answers the second broad question raised earlier. For example, to estimate thermodynamics, it uses Benson group additivity [24–26], which is a linear model for adding enthalpies of formation of predefined groups to obtain the enthalpy of the molecule. To estimate the kinetics, it uses decision trees [27]. These simple estimators are convenient since they can quickly provide predictions for tens of thousands of reactions that must often be considered during automatic mechanism generation [16]. The second major component of RMG is the flux based expansion algorithm [28]. At each iteration of mechanism generation, RMG simulates the set of species and reactions it has already chosen to include in the model, referred to as the “core”. During this simulation, it calculates fluxes through a set of reactions that are under consideration to enter the core, referred to as the “edge”. If the flux toward a species in the edge is high enough, it is added to the core. This iterative expansion will continue until the flux of all edge species is below the user defined tolerance. Poor thermochemistry and kinetic estimation can cause model generation to miss relevant chemical pathways, highlighting the connection between the two key questions raised earlier as well as the crucial role of RMG database in providing good estimates. When all relevant edge species have been added to the core, mechanism generation is finished, and the resulting kinetic model can be simulated (i.e., integrated with respect to time) with any popular software, such as Chemkin [29], Cantera [30], or RMS [31].

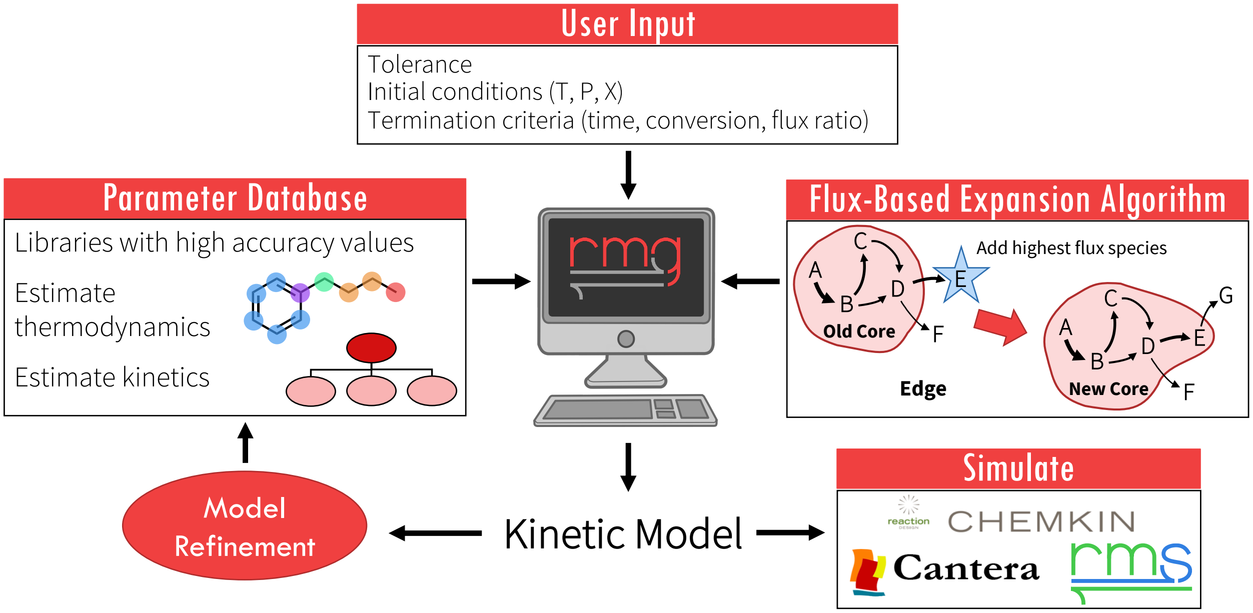


Figure 1.2. Schematic showing the high-level workflow used by RMG to create detailed kinetic mechanisms that are used to simulate real chemical systems. After receiving input from a user, RMG applies reaction templates to exhaustively enumerate a list of possible reactions, and the corresponding thermochemical and kinetic parameters are estimated via RMG’s database. The flux-based expansion algorithm is used to prioritize important (i.e., high-flux) reactions and iteratively expand the mechanism until it satisfies the user-defined termination criteria. Some parameters in the kinetic model must often be refined using the workflow shown in Figure 1.3. The final kinetic model can be simulated to yield the time evolution of the chemical system.

In theory, the resulting kinetic mechanism is complete and closely models the chemistry for the chosen system. In reality, the accuracy of the predicted concentration profiles is rarely satisfactory on the first try. While validating the kinetic model, it is common to identify several parameters which were poorly estimated during mechanism generation and thus must be refined. Model refinement involves updating RMG's database, running RMG again with these better values to obtain a new kinetic model, and likely iterating through this process a few times. Although identifying key parameters for refinement via sensitivity analysis is an automated process and many steps for refining thermochemical parameters have been automated as well [32], obtaining new kinetic parameters involves a more complicated and tedious workflow.

Historically, these parameters were regressed from experimental measurements. Today, advances in compute power have increased the popularity of QM calculations. Figure 1.3 outlines the steps necessary to calculate updated kinetic parameters for a single reaction. The workflow starts by embedding 3D structures–along with several conformers–for the reactant(s) and product(s) via distance geometry methods [33] that are then cheaply minimized with molecular mechanics force fields [34]. The lowest energy conformers given by molecular mechanics are further optimized using quantum chemical methods. Next, these reactant and product geometries are used to generate an initial guess for the transition state (TS) structure, which is then optimized and verified with an intrinsic reaction coordinate (IRC) calculation to confirm that the TS connects the corresponding reactant(s) and product(s) [35]. The energies of the stable species and TS are re-computed at a highly accurate level of theory (e.g., coupled-cluster which is commonly considered the gold standard in quantum chemistry [36, 37]) with zero-point energy corrections to provide an accurate barrier height and reaction energy. If the structures of interest are flexible with many rotatable bonds, conformational effects must also be considered. This is commonly done by approximating each hindered internal rotor as independent [38–40], but this approximation is often not accurate enough and some treatment of the coupling between different rotors is then required [41]. Methods that utilize multiple conformational minima for the reactant(s) and TS and combine conformational effects with low frequency anharmonic torsional modes–such as the multistructure methods developed by Truhlar and coworkers–have become popular alternatives, but identifying enough relevant conformers can be quite challenging [42–45]. Finally, the workflow culminates in calculating the partition functions, which canonical transition state theory (TST) uses to estimate the high-pressure limit rate coefficient $k_\infty(T)$; the effects of symmetry and reaction path degeneracy, tunneling, and other corrections should be included as well [4, 46–48]. Although Figure 1.3, and most work in this thesis, focuses on the gas-phase, if the reaction is carried out in solution, the free energies of each structure must be corrected with an additional solvation free energy term [49], which is commonly computed with either implicit [50], explicit [51], or hybrid [52, 53] solvation models; additional details are discussed in Chapter 6.

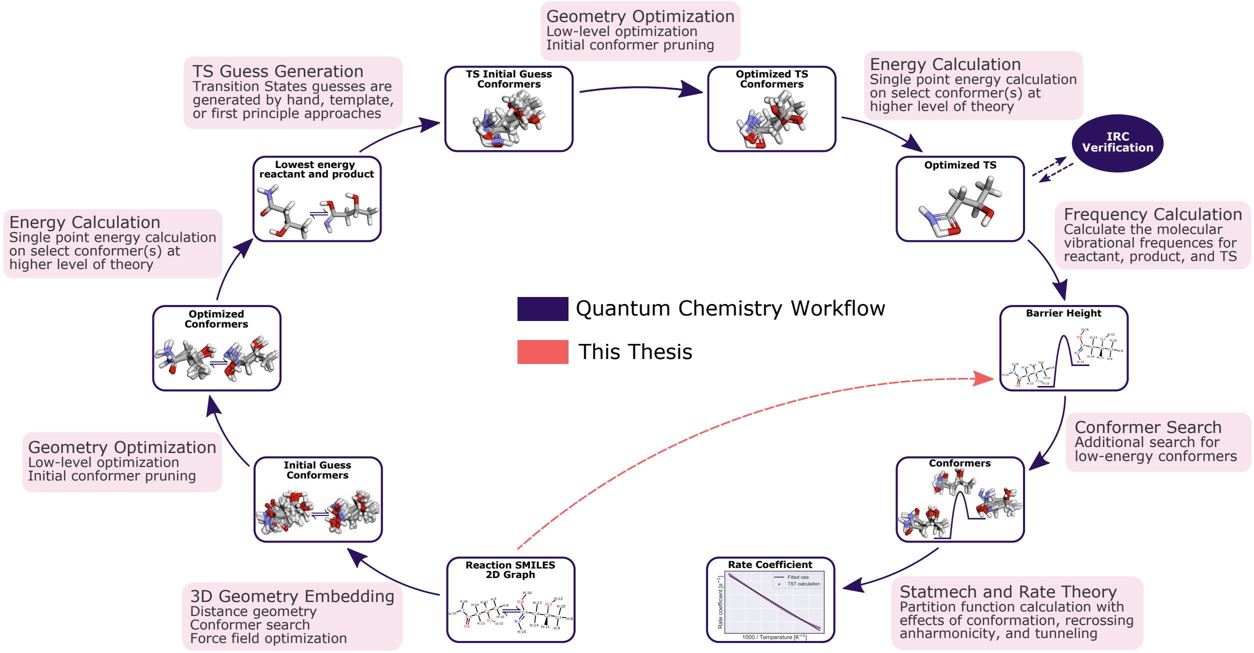

Geometry Optimization
Low-level optimization
Initial conformer pruning
TS Guess Generation
Transition States guesses are generated by hand, template, or first principle approaches
TS Initial Guess Conformers
Optimized TS Conformers
Energy Calculation
Single point energy calculation on select conformer(s) at higher level of theory
IRC Verification
Optimized TS
Lowest energy reactant and product
Energy Calculation
Single point energy calculation on select conformer(s) at higher level of theory
Frequency Calculation
Calculate the molecular vibrational frequences for reactant, product, and TS
Quantum Chemistry Workflow
This Thesis
Optimized Conformers
Barrier Height
Geometry Optimization
Low-level optimization
Initial conformer pruning
Initial Guess Conformers
Conformer Search
Additional search for low-energy conformers
Conformers
Reaction SMILES 2D Graph
Rate Coefficient
Fitted rate
TST calculation
1000 / Temperature [K⁻¹]
3D Geometry Embedding
Distance geometry
Conformer search
Force field optimization
Statmech and Rate Theory
Partition function calculation with effects of conformation, recrossing anharmonicity, and tunneling

Figure 1.3. The conventional workflow for combining QM with TST to predict a high-pressure limit rate coefficient for a single reaction is shown in the outer loop with dark purple arrows. Since this workflow is computationally expensive, this thesis focuses on quickly providing accurate estimates for barrier heights and reaction energies by using models that operate on simple input representations, making them ideal for automated, high-throughput application. This schematic is modified from ref. [54].

Despite the tremendous progress the field has made towards a predictive modeling approach, the process to obtain improved kinetic parameters for even a single reaction is unfortunately quite involved and computationally expensive, which continues to present many challenges towards automation and high-throughput. For example, geometry optimization and single-point energy refinement are essential steps in this workflow, yet they scale anywhere from $\mathcal{O}(N^4)$ for density functional methods with exchange correlation like B3LYP [55–58] to $\mathcal{O}(N^7)$ for highly accurate CCSD(T) methods [59, 60], such that N is the number of basis functions. Considering multiple conformers is crucial because using high-energy conformers to calculate partition functions can lead to inaccurate rates. However, this results in many more intensive calculations that each scale poorly with system size. The search for a 3D transition state (TS) structure is often time-consuming as well since automated saddle point finders frequently have high failure rates so many geometries must often be provided as initial guesses for the optimization to converge to a valid TS [61, 62].

Finally, this computational expense becomes especially daunting when thinking about how tremendously large chemical space is and how many possible reactions might need to be characterized. Even when considering just organic species comprised of the elements H, C, N, O, S, and halogens, there exist at least 166 billion organic molecules with up to 17 heavy atoms [63]. For context, this size range is still very small as many modern medicines and materials contain much larger molecules. The scope of possible unimolecular reactions from this limited set is substantially larger since each molecule contains multiple reaction sites and can participate in multiple reaction templates that can create radical species not present in the original set. The number of possible reactions becomes immensely larger when considering the combinatorial explosion that results from multi-molecule reactions (e.g., 166×10^9 choose 2 is 1.37×10^{22}). Given that the workflow from Figure 1.3 requires millions of CPU hours to generate high-quality kinetics datasets with just 6,000 to 12,000 reactions, such as those presented in Chapter 6 and Chapter 3 respectively, exploring this entire space via first-principles approaches is essentially impossible. It is imperative to develop tools that can make predictions for a much larger magnitude of reactions in order to achieve the grand vision of quantitative predictive chemistry.

Here, the goal of this thesis is to use machine learning (ML) as a tool to directly predict barrier heights and reaction energies as it is convenient for users to obtain both reaction properties from a single model. The main aspiration is to create a symbiotic relationship in which modern ML models can quickly screen through a multitude of reactions so that compute hours from the expensive QM workflow are only used when necessary. Important properties of these models include avoiding the challenging task of finding TS structures and operating on simple input representations that do not require expensive QM calculations and conformer searches. Such properties make these models ideal for automated high-throughput screening of large regions of reaction space. Example applications include estimating barrier heights during the automated generation of kinetic models, identifying relevant low-barrier reactions produced from an automated enumeration, and offering substantial speedup when refining existing kinetic models. These models are also relevant to inverse molecular design i.e., the task of designing molecules with specific properties. Once a target molecule has been identified, the synthesis route will likely involve many competing reactions, and these models can predict the relevant kinetic properties. Given that all known chemistry is dwarfed by the vastness of chemical space, most of which still remains unexplored [63–65], these models will satisfy an essential need by providing easy access to accurate kinetic predictions that will encourage further progress towards quantitative predictive chemistry.

1.2 Thesis Overview

This thesis has made several novel contributions towards the field of chemical kinetics. After Chapter 2 provides some basic background information regarding the methods and tools used in this thesis, the following chapters detail the original research in detail. Since quantitative chemical reaction data are often difficult to find, and high-quality datasets are especially rare, this thesis creates two kinetic datasets. The values for species thermochemistry and reaction properties from these datasets are extremely valuable by themselves for developing detailed kinetic mechanisms and predicting reaction outcomes. They also present several opportunities to benchmark the accuracy of other quantum chemical methods and evaluate various machine learning approaches to accelerate and automate steps in the traditional quantum chemistry workflow to calculate kinetic parameters. This thesis presents results for each of these tasks. Finally, this thesis provides guidelines for best practices in machine learning as well as an open-source software package, both of which can help improve the status quo of machine learning for chemistry. A summary of each chapter is given below.

Chapter 3 addresses the current paucity of quantitative chemical reaction data by creating the largest open-source high-quality kinetics database with atom-mapped reactions and transition state geometries. The CCSD(T)-F12a/cc-pVDZ-F12//ωB97X-D3/def2-TZVP level of theory is used to obtain highly accurate single-point energy values for nearly 22,000 unique stable species and transition states. This work reports the results from these quantum chemistry calculations and extracts the barrier heights and reaction enthalpies to create a kinetics dataset of nearly 12,000 elementary gas-phase organic reactions, whose species contain the elements H, C, N, and O. This work also reports accurate transition state theory rate coefficients $k_\infty(T)$ between 300 K and 2000 K and the corresponding Arrhenius parameters for a subset of rigid reactions. This dataset provides a foundation for this thesis, as well as for the broader kinetics community, since it provides opportunities to evaluate the best ways to represent reactions and predict kinetic parameters. This data should accelerate development of automated and reliable methods for quantitative reaction prediction.

Chapter 4 offers several significant improvements to the Reaction Mechanism Generator (RMG) database. The RMG database consists of curated datasets and estimators to quickly predict the tens of thousands of parameters necessary for automatically constructing a wide variety of chemical kinetic mechanisms. Here, I integrate values from Chapter 3 into this open-source database. Contributions include frequency scaling factors, atom energy corrections, bond additivity corrections, and a new thermodynamics library, all of which will improve species thermochemistry. Several high-quality training reactions are added as well, which result in significant improvements to many kinetic estimators within the RMG database. Altogether, these updates should improve RMG's reliability when automatically generating kinetic mechanisms.

Chapter 5 trains a directed message passing neural network to predict the high-quality barrier heights and reaction enthalpies calculated at CCSD(T)-F12a/cc-pVDZ-F12//ωB97X-D3/def2-TZVP from Chapter 3. Utilizing both the forward and reverse reactions gives over 20,000 diverse gas-phase reactions as training data. Our model uses 75% fewer parameters compared to previously published models trained on the same data, an improved reaction representation, and proper data splits to accurately estimate performance on unseen reactions. Using information from only the reactant and product, our model quickly predicts barrier heights with a testing MAE of 2.6 kcal mol^{-1} relative to the coupled-cluster data, which is comparable in accuracy to a good DFT calculation. Consistent with several previous studies, our results show that learned representations from graphs outperform traditional ML approaches (e.g., support vector regression, random forest, extreme gradient boosting, and multi-layer perceptron). Further, our results show that future modeling efforts to estimate reaction properties would significantly benefit from fine-tuning calibration using a transfer learning technique. This model is anticipated to offer substantial time savings and improve quantitative kinetic predictions for small molecule gas-phase chemistry.

Chapter 6 is conceptually similar to Chapter 5, but focuses on two important contributions for the field of solvated kinetics. First, a novel dataset of nearly 6,000 elementary radical reactions is created using the M06-2X/def2-QZVP//B3LYP-D3(BJ)/def2-TZVP level of theory in the gas phase. A conformer search is done for each species using TPSS/def2-TZVP in the COSMO phase to account for generic solvent effects. Solvation effects are treated using the BP-TZVP procedure from COSMO-RS, giving access to the Gibbs free energies of activation and of reaction for these radical reactions in 40 popular solvents. These balanced reactions involve the elements H, C, N, O, and S, contain up to 19 heavy atoms, and have canonical atom-mapped SMILES. All transition states are verified by an intrinsic reaction coordinate calculation. The second contribution is training a deep graph network to quickly estimate the Gibbs free energy of activation and of reaction in both gas and solution phase using only the SMILES of the reactant, product, and solvent. To properly measure performance, results are reported for both interpolative and extrapolative data splits (i.e., random and K-Means clustering of reactant substructures respectively). Performance is also compared to several baseline models. Accounting for both the forward and reverse reactions, as well as data augmentation, leads to approximately half a million entries to train the model, which achieves a mean absolute error of 3.00 kcal mol^{-1} for the Gibbs free energy of activation in solution in the test set when using extrapolation-based splits. Our results corroborate existing literature that shows learned representations from graphs typically outperform traditional ML approaches within the field of chemical kinetics. This model should accelerate predictions for high-throughput screening to quickly identify relevant reactions in solution, and the new dataset will serve as a benchmark for future studies.

Chapter 7 continues to address the data scarcity within chemical kinetics by enumerating 750 million canonical atom-mapped reaction SMILES. These reactions involve the elements H, C, N, and O and contain molecules with up to 11 heavy (non-hydrogen) atoms that participate in 22 popular reaction templates from the Reaction Mechanism Generator, thus this dataset is named RMG-DB-11. This dataset is used to pre-train both a deep graph network and a language model that the kinetics community can fine-tune on various downstream tasks focused on gas-phase kinetics. I also present a new dataset of barrier heights for 1,100 radical reactions calculated at CCSD(T)-F12a/cc-pVDZ-F12//ωB97X-D3/def2-TZVP, which is used to fine-tune the models. I also fine-tune using the gas-phase dataset of barrier heights for 6,000 reactions from Chapter 6 calculated at M06-2X/def2-QZVP//B3LYP-D3(BJ)/def2-TZVP. The specific results in this chapter indicate that graph networks outperform for predicting barrier heights. However, identifying learned reaction representations that generalize well remains a grand challenge for computational kinetics so future work should continue to explore this further. RMG-DB-11 should push the field to create foundation models that ultimately accelerate and improve kinetic predictions for small molecule chemistry.

While the previous chapters focused on an in-depth analysis of one specific dataset or a specific goal of predicting barrier heights, the next two chapters transition to discuss more fundamental and widespread issues within the field of ML for chemistry. I then directly provide guidelines and a software package that can help improve the status quo. Chapter 8 offers a broad perspective on what constitutes best practices for creating useful ML models. The discussion focuses within the field of chemical kinetics, but the principals are generalizable to other fields as well. Overall, I present three aspects that are essential for researchers to consider when constructing ML models: (1) Are the model's prediction targets and associated errors sufficient for practical applications? (2) Does the model prioritize user-friendly inputs so it is practical for others to integrate into prediction workflows? (3) Does the analysis report performance on both interpolative and more challenging extrapolative data splits so users have a realistic idea of the likely errors in the model's predictions? Altogether, these criteria should improve the quality of published models in the literature and encourage more rigorous analysis that enables better scientific progress.

Chapter 9 expands on the third main idea from Chapter 8, namely using well-thought-out data splits to properly analyze model performance. Critical to the use of ML is the process of splitting datasets into training, validation, and testing subsets that are used to develop and evaluate models. Common practice in the literature is to assign these subsets randomly, partly because this approach is computationally fast and efficient and also because this is often the only option conveniently provided in popular packages like `sklearn`. However, random splitting only measures a model's capacity to interpolate. Testing errors reported from models trained on random data splits may be overly optimistic if given new data that is dissimilar to the scope of the training set; thus, there is a growing need to easily measure performance for extrapolation tasks. To address this issue, a new open-source Python package called `astartes` is created, which implements many similarity- and distance-based algorithms to partition data for either interpolation tasks or more challenging extrapolation tasks. The examples in this chapter focus on use-cases within cheminformatics. However, our package operates on arbitrary vector inputs; its principals are generalizable to other ML domains as well, and it can be easily integrated into existing workflows.

Finally, Chapter 10 summarizes the key contributions of the thesis and discusses several recommendations for future work related to the field of automated chemical kinetics and machine learning applied to chemistry.

1.3 References

(1) Van't Hoff, J. H., *Etudes De Dynamique Chimique*; Muller: 1884; Vol. 1.

(2) Arrhenius, S. Über Die Reaktionsgeschwindigkeit Bei Der Inversion Von Rohrzucker Durch SÄUren. *Zeitschrift für physikalische Chemie* **1889**, *4*, a translation of the four pages in this paper that deal with temperature dependence is included in Back, M. H., and Laidler, K. J., "Selected Readings in Chemical Kinetics," Pergamon Press, Oxford, 1967, pp. 31-35., 226–248.

(3) Truhlar, D. G.; Garrett, B. C.; Klippenstein, S. J. Current Status of Transition-State Theory. *J. Phys. Chem.* **1996**, *100*, 12771–12800.

(4) Klippenstein, S. J. From Theoretical Reaction Dynamics to Chemical Modeling of Combustion. *Proc. Combust. Inst.* **2017**, *36*, 77–111.

(5) Green, W. H. Moving from Postdictive to Predictive Kinetics in Reaction Engineering. *AIChE Journal* **2020**, *66*, e17059.

(6) Harper, M. R.; Van Geem, K. M.; Pyl, S. P.; Marin, G. B.; Green, W. H. Comprehensive Reaction Mechanism for n-Butanol Pyrolysis and Combustion. *Combust. Flame* **2011**, *158*, 16–41.

(7) Pio, G.; Dong, X.; Salzano, E.; Green, W. H. Automatically Generated Model for Light Alkene Combustion. *Combust. Flame* **2022**, *241*, 112080.

(8) Magoon, G. R.; Aguilera-Iparraguirre, J.; Green, W. H.; Lutz, J. J.; Piecuch, P.; Wong, H.-W.; Oluwole, O. O. Detailed Chemical Kinetic Modeling of JP-10 (exo-tetrahydrodicyclopentadiene) High-temperature Oxidation: Exploring the Role of Biradical Species in Initial Decomposition Steps. *Int. J. Chem. Kinet.* **2012**, *44*, 179–193.

(9) Allen, J. W.; Scheer, A. M.; Gao, C. W.; Merchant, S. S.; Vasu, S. S.; Welz, O.; Savee, J. D.; Osborn, D. L.; Lee, C.; Vranckx, S., et al. A Coordinated Investigation of the Combustion Chemistry of Diisopropyl Ketone, a Prototype for Biofuels Produced by Endophytic Fungi. *Combust. Flame* **2014**, *161*, 711–724.

(10) Grinberg Dana, A.; Gudiyella, S.; Green, W. H.; Shanbhogue, S. J.; Michaels, D.; Chakroun, N. W.; Ghoniem, A. In *55th AIAA Aerospace Sciences Meeting*, 2017, p 0835.

(11) Zhang, P.; Yee, N. W.; Filip, S. V.; Hetrick, C. E.; Yang, B.; Green, W. H. Modeling Study of the Anti-Knock Tendency of Substituted Phenols as Additives: An Application of the Reaction Mechanism Generator (RMG). *Phys. Chem. Chem. Phys.* **2018**, *20*, 10637–10649.

(12) Class, Caleb A. and Vasiliou, AnGayle K. and Kida, Yuko and Timko, Michael T. and Green, William H. Detailed Kinetic Model for Hexyl Sulfide Pyrolysis and its Desulfurization by Supercritical Water. *Phys. Chem. Chem. Phys.* **2019**, *21*, 10311–10324.

(13) Johnson, M. S.; Nimlos, M. R.; Ninnemann, E.; Laich, A.; Fioroni, G. M.; Kang, D.; Bu, L.; Ranasinghe, D.; Khanniche, S.; Goldsborough, S. S., et al. Oxidation and Pyrolysis of Methyl Propyl Ether. *Int. J. Chem. Kinet.* **2021**, *53*, 915–938.

(14) Payne, A. M.*; Spiekermann, K. A.*; Green, W. H. Detailed Reaction Mechanism for 350–400°C Pyrolysis of an Alkane, Aromatic, and Long-Chain Alkylaromatic Mixture. *Energy & Fuels* **2022**, *36*, 1635–1646.

(15) Vermeire, F. H.; Aravindakshan, S. U.; Jocher, A.; Liu, M.; Chu, T.-C.; Hawtof, R. E.; Van de Vijver, R.; Prendergast, M. B.; Van Geem, K. M.; Green, W. H. Detailed Kinetic Modeling for the Pyrolysis of a Jet A Surrogate. *Energy Fuels* **2022**, *36*, 1304–1315.

(16) Yuan, W.; Li, Y.; Qi, F. Challenges and Perspectives of Combustion Chemistry Research. *Sci. China Chem.* **2017**, *60*, 1391–1401.

(17) Broadbelt, L. J.; Stark, S. M.; Klein, M. T. Computer Generated Pyrolysis Modeling: On-the-fly Generation of Species, Reactions, and Rates. *Ind. Eng. Chem. Res.* **1994**, *33*, 790–799.

(18) Broadbelt, L. J.; Stark, S. M.; Klein, M. T. Computer Generated Reaction Modelling: Decomposition and Encoding Algorithms for Determining Species Uniqueness. *Comput. Chem. Eng.* **1996**, *20*, 113–129.

(19) Blurock, E. S. Reaction: System for Modeling Chemical Reactions. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 607–616.

(20) Warth, V.; Battin-Leclerc, F.; Fournet, R.; Glaude, P.-A.; Côme, G.-M.; Scacchi, G. Computer Based Generation of Reaction Mechanisms for Gas-Phase Oxidation. *Comput. & Chem.* **2000**, *24*, 541–560.

(21) Vandewiele, N. M.; Van Geem, K. M.; Reyniers, M.-F.; Marin, G. B. Genesys: Kinetic Model Construction using Chemo-informatics. *Chem. Eng. J.* **2012**, *207*, 526–538.

(22) Gao, C. W.; Allen, J. W.; Green, W. H.; West, R. H. Reaction Mechanism Generator: Automatic Construction of Chemical Kinetic Mechanisms. *Comput. Phys. Commun.* **2016**, *203*, 212–225.

(23) Liu, M.; Grinberg Dana, A.; Johnson, M. S.; Goldman, M. J.; Jocher, A.; Payne, A. M.; Grambow, C. A.; Han, K.; Yee, N. W.; Mazeau, E. J., et al. Reaction Mechanism Generator v3.0: Advances in Automatic Mechanism Generation. *J. Chem. Inf. Model.* **2021**, *61*, 2686–2696.

(24) Benson, S. W.; Buss, J. H. Additivity Rules for the Estimation of Molecular Properties. Thermodynamic Properties. *J. Chem. Phys.* **1958**, *29*, 546–572.

(25) Benson, S. W., *Thermochemical Kinetics*; Wiley: 1976.

(26) Cohen, N.; Benson, S. W. Estimation of Heats of Formation of Organic Compounds by Additivity Methods. *Chem. Rev.* **1993**, *93*, 2419–2438.

(27) Johnson, M. S.; Green, W. H. A Machine Learning Based Approach to Reaction Rate Estimation. *ChemRxiv* **2022**, DOI: 10.26434/chemrxiv-2022-c98gc.

(28) Susnow, R. G.; Dean, A. M.; Green, W. H.; Peczak, P.; Broadbelt, L. J. Rate-Based Construction of Kinetic Models for Complex Systems. *J. Phys. Chem. A* **1997**, *101*, 3731–3740.

(29) CHEMKIN-PRO, R. 15112, Reaction Design. *Inc., San Diego, CA* **2011**.

(30) Goodwin, D. G.; Moffat, H. K.; Schoegl, I.; Speth, R. L.; Weber, B. W. Cantera: An Object-oriented Software Toolkit for Chemical Kinetics, Thermodynamics, and Transport Processes, 2022.

(31) Johnson, M. S.; Pang, H.-W.; Payne, A. M.; Green, W. H. ReactionMechanismSimulator.jl: A Modern Approach to Chemical Kinetic Mechanism Simulation and Analysis. *ChemRxiv* **2023**, DOI: 10.26434/chemrxiv-2023-tj34t.

(32) Grinberg Dana, A.; Ranasinghe, D.; Wu, H.; Grambow, C.; Dong, X.; Johnson, M.; Goldman, M.; Liu, M.; Green, W. ARC - Automated Rate Calculator, version 1.1.0, 2019.

(33) Crippen, G. M.; Havel, T. F., et al. Distance Geometry and Molecular Conformation, vol. 74, 1988.

(34) Halgren, T. A. MMFF VI. MMFF94s Option for Energy Minimization Studies. *J. Comp. Chem.* **1999**, *20*, 720–729.

(35) Fukui, K. The Path of Chemical Reactions-The IRC Approach. *Acc. Chem. Res.* **1981**, *14*, 363–368.

(36) Tajti, A.; Szalay, P. G.; Császár, A. G.; Kállay, M.; Gauss, J.; Valeev, E. F.; Flowers, B. A.; Vázquez, J.; Stanton, J. F. HEAT: High Accuracy Extrapolated Ab initio Thermochemistry. *J. Chem. Phys.* **2004**, *121*, 11599–11613.

(37) Karton, A.; Rabinovich, E.; Martin, J. M. L.; Ruscic, B. W4 Theory for Computational Thermochemistry: In Pursuit of Confident Sub-kJ/mol Predictions. *J. Chem. Phys.* **2006**, *125*, 144108.

(38) Pitzer, K. S.; Gwinn, W. D. Energy Levels and Thermodynamic Functions for Molecules with Internal Rotation I. Rigid Frame with Attached Tops. *J. Chem. Phys.* **1942**, *10*, 428–440.

(39) Truhlar, D. G. A Simple Approximation for the Vibrational Partition Function of a Hindered Internal Rotation. *J. Comput. Chem.* **1991**, *12*, 266–270.

(40) Pfaendtner, J.; Yu, X.; Broadbelt, L. J. The 1-D Hindered Rotor Approximation. *Theor. Chem. Acc.* **2007**, *118*, 881–898.

(41) Sharma, S.; Raman, S.; Green, W. H. Intramolecular Hydrogen Migration in Alkylperoxy and Hydroperoxyalkylperoxy Radicals: Accurate Treatment of Hindered Rotors. *J. Phys. Chem. A* **2010**, *114*, 5689–5701.

(42) Zheng, J.; Yu, T.; Papajak, E.; Alecu, I. M.; Mielke, S. L.; Truhlar, D. G. Practical Methods for Including Torsional Anharmonicity in Thermochemical Calculations on Complex Molecules: The Internal-Coordinate Multi-Structural Approximation. *Phys. Chem. Chem. Phys.* **2011**, *13*, 10885–10907.

(43) Yu, T.; Zheng, J.; Truhlar, D. G. Multi-Structural Variational Transition State Theory. Kinetics of the 1,4-Hydrogen Shift Isomerization of the Pentyl Radical with Torsional Anharmonicity. *Chem. Sci.* **2011**, *2*, 2199–2213.

(44) Zheng, J.; Truhlar, D. G. Quantum Thermochemistry: Multistructural Method with Torsional Anharmonicity Based on a Coupled Torsional Potential. *J. Chem. Theory Comput.* **2013**, *9*, 1356–1367.

(45) Zheng, J.; Meana-Pañeda, R.; Truhlar, D. G. MSTor Version 2013: A New Version of the Computer Code for the Multi-Structural Torsional Anharmonicity, Now With a Coupled Torsional Potential. *Comput. Phys. Commun.* **2013**, *184*, 2032–2033.

(46) Gonzalez-Lavado, E.; Corchado, J. C.; Suleimanov, Y. V.; Green, W. H.; Espinosa-Garcia, J. Theoretical Kinetics Study of the O(^{3}P)+ CH_4/CD_4 Hydrogen Abstraction Reaction: The Role of Anharmonicity, Recrossing Effects, and Quantum Mechanical Tunneling. *J. Phys. Chem. A* **2014**, *118*, 3243–3252.

(47) Zheng, J.; Bao, J. L.; Meana-Pañeda, R.; Zhang, S.; Lynch, B. J.; Corchado, J. C.; Chuang, Y.-Y.; Fast, P. L.; Hu, W.-P.; Liu, Y.-P.; Lynch, G. C.; Nguyen, K. A.; Jackels, C. F.; Fernandez Ramos, A.; Ellingson, B. A.; Melissas, V. S.; Villà, J.; Rossi, I.; Coitiño, E. L.; Pu, J.; Albu, T. V.; Ratkiewicz, A.; Steckler, R.; Garrett, B. C.; Isaacson, A. D.; Truhlar, D. G. *Polyrate*-version 2017-C, University of Minnesota: Minneapolis, 2017.

(48) Goldman, M. J.; Ono, S.; Green, W. H. Correct Symmetry Treatment for X+ X Reactions Prevents Large Errors in Predicted Isotope Enrichment. *J. Phys. Chem. A* **2019**, *123*, 2320–2324.

(49) Hellweg, A.; Eckert, F. Brick by Brick Computation of the Gibbs Free Energy of Reaction in Solution using Quantum Chemistry and COSMO-RS. *AIChE J.* **2017**, *63*, 3944–3954.

(50) Klamt, A. The COSMO and COSMO-RS Solvation Models. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2018**, *8*, e1338.

(51) Boereboom, J. M.; Fleurat-Lessard, P.; Bulo, R. E. Explicit Solvation Matters: Performance of QM/MM Solvation Models in Nucleophilic Addition. *J. Chem. Theory Comput.* **2018**, *14*, 1841–1852.

(52) Klamt, A.; Schüürmann, G. COSMO: A New Approach to Dielectric Screening in Solvents with Explicit Expressions for the Screening Energy and its Gradient. *Chem. J. Soc. Perkin Trans. 2* **1993**, 799–805.

(53) Klamt, A.; Jonas, V.; Bürger, T.; Lohrenz, J. C. W. Refinement and Parametrization of COSMO-RS. *J. Phys. Chem. A* **1998**, *102*, 5074–5085.

(54) Spiekermann, K. A.; Pattanaik, L.; Green, W. H. Fast Predictions of Reaction Barrier Heights: Toward Coupled-Cluster Accuracy. *J. Phys. Chem. A* **2022**, *126*, 3976–3986.

(55) Becke, A. D. Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys* **1993**, *98*, 5648–5652.

(56) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B* **1988**, *37*, 785–789.

(57) Kim, K.; Jordan, K. D. Comparison of Density Functional and MP2 Calculations on the Water Monomer and Dimer. *J. Phys. Chem.* **1994**, *98*, 10089–10094.

(58) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra using Density Functional Force Fields. *J. Phys. Chem.* **1994**, *98*, 11623–11627.

(59) Adler, T. B.; Knizia, G.; Werner, H.-J. A Simple and Efficient CCSD(T)-F12 Approximation. *J. Chem. Phys.* **2007**, *127*, 221106.

(60) Feller, D.; Peterson, K. A.; Dixon, D. A. A Survey of Factors Contributing to Accurate Theoretical Predictions of Atomization Energies and Molecular Structures. *J. Chem. Phys.* **2008**, *129*, 204105.

(61) Grambow, C. A.; Jamal, A.; Li, Y.-P.; Green, W. H.; Zador, J.; Suleimanov, Y. V. Unimolecular Reaction Pathways of a γ-Ketohydroperoxide from Combined Application of Automated Reaction Discovery Methods. *J. Am. Chem. Soc.* **2018**, *140*, 1035–1048.

(62) Maeda, S.; Harabuchi, Y. On Benchmarking of Automated Methods for Performing Exhaustive Reaction Path Search. *J. Chem. Theory Comput.* **2019**, *15*, 2111–2115.

(63) 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. *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875.

(64) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the Size of Drug-like Chemical Space Based on GDB-17 Data. *J. Comput. Aided Mol. Des.* **2013**, *27*, 675–679.

(65) Hoffmann, T.; Gastreich, M. The Next Level in Chemical Space Navigation: Going Far Beyond Enumerable Compound Libraries. *Drug Discovery Today* **2019**, *24*, 1148–1156.

Chapter 2

Background

The following sections provide a brief overview of the methods relevant to this thesis. The first section discusses some basics of computational chemistry, which is necessary to explain and predict chemical properties and reactivity. The next section provides a brief introduction to machine learning, specifically graph neural networks, which are the primary model architecture used throughout this thesis to accelerate portions of the workflow from Figure 1.3.

2.1 Computational Chemistry

Although experimental techniques are crucial to obtaining reliable thermodynamic and kinetic data, these approaches can be expensive, time-consuming, and have several safety considerations. Furthermore, studying reactive intermediates experimentally can be prohibitively difficult. With these limitations in mind, first-principle approaches can be another useful method of obtaining thermodynamic and kinetic data that is needed to construct kinetic models and understand chemical systems. This thesis uses quantum chemistry to calculate the geometry, vibrational frequency, and energy of chemical species. Only a brief overview of these techniques is presented here. Additional detail regarding thermodynamics and statistical mechanics can be found in several textbooks [1–8].

2.1.1 Quantum Mechanics

The primary goal of quantum chemistry is to describe the molecular properties of a chemical system. This involves solving the time-independent Schrödinger equation:

$$\hat{H}\Psi(\vec{r},\vec{R}) = E\Psi(\vec{r},\vec{R}) \tag{2.1}$$

such that $\hat{H}$ is the Hamiltonian operator, which is often separated into terms for the kinetic and potential energy for the individual nuclei and electrons, along with terms for their interactions. Ψ is wavefunction of the system, $\vec{r}$ and $\vec{R}$ are the positions of electrons and nuclei respectively, and E is the total energy of the system, which is also an eigenvalue. The physical interpretation of the wavefunction comes from the quantity $|\Psi|^2$, which relates to the probability of finding a particle in a given volume.

Equation (2.1) can only be solved analytically for very simple systems (e.g., those with only 1 electron). Analytical solutions for multiple electron systems are not possible, so obtaining information for larger systems requires approximations as well as numerical methods. The first assumption is the Born-Oppenheimer approximation, which allows us to treat the nuclear and electronic wavefunctions independently. The rational for this is that nuclei are much heavier than electrons, so nuclei effectively appear as stationary particles relative to the much faster moving electrons. Many computational quantum chemistry softwares utilize this approximation to solve the electronic Schrödinger Equation, with the nuclei-nuclei potential energy term added in at the very end.

Several methods have been used to obtain approximate numerical solutions to the Schrödinger Equation. One approach is to use wavefunction-based methods, such as Hartree-Fock (HF) [9, 10], which constructs an initial trial wavefunction using Slater determinants to satisfy the antisymmetric constraint of electrons and uses a variational approach to solve Equation (2.1). This invokes the independent particle model, in which the electrons are modelled as a combination of the various one-electron wavefunctions. Rigorously, the motion of each individual electron is correlated to all others. However, a key approximation in HF is the mean field approximation, which instead allows each electron to behave as if it sees an average potential of all the other electrons and omits correlation between the electrons. Still, the mean field approximation can cause substantial errors, so several post-HF methods have been developed to incorporate correlation effects. Examples include Møller-Plesset (MP) perturbation theory [11] and coupled-cluster (CC) theory [12]. Although CC is commonly considered the gold standard for quantum chemistry methods due its close agreement with experiment values [13, 14], it is practically only feasible for relatively small molecules (e.g., about 20 non-hydrogen atoms or fewer) due to its steep scaling of $\mathcal{O}(N^7)$, such that N is the number of basis functions.

More recently, density functional theory (DFT) [15] has emerged as another popular approach for solving the the Schrödinger Equation. In contrast to wavefunction-based methods, DFT centers around the electron probability density, whose integral determines the probability of finding any electrons within a defined volume element. The merit in this approach stems from Hohenberg and Kohn [16] proving that electron density can be used to derive ground-state properties of a molecule. Additionally, numerous studies have shown that DFT strikes a favorable balance between computational cost and accuracy, which has helped establish its popularity within computational chemistry [17, 18].

2.1.2 Statistical Thermodynamics

Crucial to modeling chemical systems is obtaining thermochemical data, such as enthalpy, entropy, Gibbs free energy, and heat capacities. Statistical thermodynamics provides the framework to accomplish this by transforming microscopic quantities, like the electronic energy, vibrations, and geometry obtained via quantum chemistry, to these relevant macroscopic properties. The key concept of statistical mechanics is to treat molecules as a collection of microscopic states. The probability of being in these states depends on the energy required to access these states–and hence the temperature of the system–and is ultimately described by a Boltzmann distribution, which allows us to calculate the partition function of a molecule

$$q(V,T) = \sum_{i=1}^{N} g_i \exp\left(-\frac{\varepsilon_i}{k_B T}\right) \tag{2.2}$$

such that V is the volume, T is the temperature, ε_i is the energy associated with each microscopic state which is determined via quantum chemistry throughout this thesis but could also be determined experimentally, g_i is the degeneracy of each state, N is the number of energy levels, and k_B is the Boltzmann constant.

Determining the molecular partition function typically requires strong assumptions. The first being the ideal gas law, which assumes individual particles have no interactions and hence are independent. The second is the rigid rotor harmonic oscillator (RRHO) approximation, which states that the modes of motion–translational, vibrational, rotational, and electronic–that contribute to the partition function are independent of one another. Thus, the energy of a molecule at state i is given by the sum of each mode:

$$\varepsilon_i = \varepsilon_i^T + \varepsilon_i^V + \varepsilon_i^R + \varepsilon_i^E \tag{2.3}$$

The RRHO approximation results in a relatively simple way to derive the molecular partition function and gives reasonably accurate results. Substituting Equation (2.3) into Equation (2.2) gives an expression for the total molecular partition function:

$$q = q_T q_V q_R q_E \tag{2.4}$$

Definitions of the individual partition functions for each mode, as well as derivations of thermodynamic properties from the molecular partition function, are detailed in several textbooks [1–8].

2.1.3 Transition State Theory

Transition state theory (TST) is a popular method to estimate rate coefficients of elementary chemical reactions from properties derived using first-principle calculations, namely energies calculated via quantum mechanics and partition functions calculated via statistical thermodynamics. Several forms of TST have been developed over the years [19–23], so only a brief summary is provided here.

TST relies on the concept of a phase space, which consists of all positions and momenta of the molecular system. Within the TST framework, an elementary reaction is defined as a minimum energy path connecting the reactants and products. Solving the Schrödinger equation for many atomic configurations along this path allows construction of a potential energy surface (PES). Reactants and products are stable species i.e., their potential energy is a local minimum on the PES. A key assumption is that the reactants and products are separated by a dividing surface, known as the transition state (TS). Mathematically, a TS is a first-order saddle point on the PES, and identifying TS structures and calculating their energies is essential for TST. Another assumption of TST is the no recrossing assumption i.e., if a reactant obtains enough energy to cross this dividing surface, it will proceed to forming the products. In reality, re-crossing occurs so the reactive flux is overestimated by TST.

The final assumption made by TST is that the reactants are in equilibrium with the TS, either on a fixed-temperature basis which is deemed canonical TST, or on a fixed-energy basis which is deemed microcanonical TST. This thesis exclusively uses canonical TST, defined as

$$k_\infty(T) = \frac{k_B T}{h} \frac{q^\ddagger}{\prod_i^{N_r} q^{r,i}} \exp\left(-\frac{E_0^\ddagger}{k_B T}\right) \tag{2.5}$$

where $k_\infty(T)$ is the high-pressure limit temperature-dependent rate coefficient, $q^\ddagger$ is the partition function of the TS, $q^{r,i}$ is the partition function of reactant i, N_r is the number of reactants in the reaction, h is Planck's constant, and $E_0^\ddagger$ is the difference in the ground state electronic energy between the TS and the reactant(s) at 0 K. The above expression is often corrected for several factors, including quantum mechanical tunneling, symmetry factors of the reaction, and the presence of optical isomers to yield

$$k(T) = \kappa \frac{k_B T}{h} \frac{\sigma^\ddagger}{\prod_i^{N_r} \sigma^{r,i}} \frac{\prod_i^{N_r} m^{r,i}}{m^\ddagger} \frac{q^\ddagger}{\prod_i^{N_r} q^{r,i}} \exp\left(-\frac{E_0^\ddagger}{k_B T}\right) \tag{2.6}$$

such that $\sigma^\ddagger$ is the external symmetry factor of the TS, and $\sigma^{r,i}$ is the external symmetry factors of reactant i, $m^\ddagger$ is the number of optical isomers of the TS, and $m^{r,i}$ is the number of optical isomers of reactant i. The tunneling correction factor κ accounts for the fact that particles do not necessarily need to have a greater energy than the dividing surface, as they can simply tunnel through the boundary.

2.2 Machine Learning

2.2.1 Overview

Machine learning (ML) models can be classified as discriminative or generative. Given some training data x and labels y, discriminative models learn the conditional probability distribution $p(y|x)$, which can then be used to predict the value of y given a new example x. This is also termed supervised learning since the labels are used to "supervise" the learning process. Generative models learn the joint probability distribution $p(x, y)$, which can be used to generate new examples of (x, y). These methods are useful for unsupervised learning tasks. This thesis focuses on using discriminative models. However, traditional unsupervised methods are used to cluster the data that is subsequently used for training and evaluating the discriminative models in Chapter 6 and Chapter 9; more broadly, the large amount of unlabeled chemistry data presents several promising avenues to augment model performance, which could be explored in future work.

Discriminative models can be used for either classification or regression tasks. The former categorizes an input x into one of many discrete output classes (e.g., binary classification of whether a molecule is toxic), while the latter predicts a continuous output value (e.g., solubility or melting point). The prediction tasks in this thesis focus on regression. However, quantitative structure-activity or structure-property relationship (QSAR or QSPR) have a long history within the field of predictive chemistry [24–26].

At a high-level, ML for chemistry workflows are comprised of two core components–representation and model architecture–though this distinction often becomes less clear for modern deep learning architectures. Section 2.2.5 will touch on this nuance in the context of graph neural networks, but the general idea is similar for convolutional neural networks and transformers. Regardless, molecules must first be converted into a machine-readable vector representation that is then passed to the model. While there are many choices for the representation and model architecture, the key concept for ML is that the model parameters are "learned" (i.e., determined algorithmically via gradient descent) as opposed to being manually determined by humans. This is described in the next section.

2.2.2 Backpropagation

At its core, ML is an optimization procedure with the goal of fitting a model to some given data such that the model's predictions minimize a loss function. Although there are numerous optimization strategies, the most popular approach–particularly for deep learning–is based upon stochastic gradient descent (SGD). Briefly, SGD randomly samples some amount of data during training, performs a forward pass of the model to make a prediction, and then calculates the gradient of the prediction with respect to the model's parameters Θ. Backpropagation, which is short for backward propagation of errors and is effectively an application of the chain rule from calculus, along with automatic differentiation, are used to automatically calculate these gradients. Finally, the model's parameters are updated by taking a step in the direction of the negative gradient using the following equation

$$\Theta_{i+1} = \Theta_i - \eta \nabla_\Theta \mathcal{L}(y, \hat{y}) \tag{2.7}$$

such that Θ_i are the model's parameters at iteration i, η is the learning rate (i.e., step-size), ∇_Θ is the gradient operator with respect to the model's parameters, and $\mathcal{L}$ is the loss function, which depends on the predicted value(s) y using the parameters from iteration i and the corresponding true value(s) $\hat{y}$. Since this thesis focuses on regression tasks, namely predicting barrier heights and reaction energies for use in quantitative chemical kinetics, the loss function used in Equation (2.7) is taken to be the mean squared error (MSE):

$$\text{MSE} = \frac{1}{n} \sum_{j=i}^{n} (y_j - \hat{y}_j)^2 \tag{2.8}$$

such that n is the number of samples, y_j is the predicted value(s) for the j^{th} sample, and $\hat{y_j}$ is the corresponding true value(s) for the j^{th} sample.

2.2.3 Featurization and Molecular Representation

The first step in many ML workflows is to create a fixed-length vector representation for the input example x. While text representations, such as SMILES [27] or InChI [28], are useful for efficiently storing information about the molecular structure in databases, these must be converted into a numerical vector before being passed to the ML model. Even language models that operate "directly on text" must first create a numeric representation for each token in the string via an embedding layer. These representations are commonly called "fingerprints" since ideally they can uniquely identify each molecule. Historically, low-dimensional vectors were manually constructed using physically informed features e.g., molecular weight, melting point, boiling point, dipole moment, etc [29]. Since some of these features require experimental measurements, obtaining feature vectors for hundreds of compounds can be rather challenging so alternative methods are desired.

Gradually, the field shifted to representations that depend only on the molecular connectivity. One simple approach is to use one-hot encoding to create a Boolean vector in which each entry represents the presence or absence of a certain functional group or other feature, though continuous representations are often more expressive in comparison to simple binary representations. Many algorithmic fingerprint approaches effectively have a similar goal of iteratively considering all substructures of a certain size and type. For example, extended-connectivity fingerprints (ECFPs) [30] start by assigning each atom in the molecule a unique integer identifier. Initially, these atom identifiers only represent information about the atom itself and its attached bonds. This representation is updated by iterating over the neighboring atoms, up to a specified number of bonds or radius, and finally applying a hashing function to generate a binary representation. The Morgan circular fingerprint [31] is likely the most popular variant, but many fingerprints exist with Avalon [32] and MACCS [33] being just a few additional examples.

Molecular representations can also be based on 3D information. Bond lengths and atomic partial charges are examples of features that can be calculated via quantum mechanics. The Coulomb matrix is another popular choice that computes the pairwise distances between each of the atoms and uses the product of their nuclear charges weighted by a function to mask long-range effects as a variation of Coulomb's Law [34, 35]. Many other similar representations exist. For example, the bag-of-bonds takes Coulomb matrix terms and organizes the elements into atom-pairwise types [36]. SLATM [37, 38] and FCHL19 [39, 40] append higher-order interaction terms (between triplets of atoms) with different potentials. SOAP considers atoms in molecules according to their neighbouring atom density [41, 42]. However, all of these 3D featurization approaches require a geometry optimization, which is often too expensive to be reasonably considered for high-throughput applications.

2.2.4 Discriminative Architectures

Once a vector representation has been obtained, it is passed to a discriminative model to make a prediction. Examples of traditional model architectures include support vector machines, random forests, and kernel ridge regression, to name just a few. Another popular example is the multilayer perceptron (MLP)–also referred to as a standard feed forward neural network (FFNN) or simply a feed forward network (FFN) for short. First developed in the 1950s [43], an FFN consists of an input layer, some number of hidden layers, and an output layer. Today, the FFN has become an primary component of modern ML, such that the term "deep" learning describes networks that have more hidden layers. Each layer consists of several stacked perceptrons, also called "neurons" since the architecture is inspired by the human brain. A single layer of an FFN receives an input vector of scalars $\mathbf{x} \in \mathbb{R}^m$ and multiplies it by a learnable weight matrix $\mathbf{W} \in \mathbb{R}^{n \times m}$. Typically, a scalar offset $\mathbf{b} \in \mathbb{R}^n$ is added as well. The result is passed through a nonlinear activation function, which gives the model better expressivity and its theoretical capacity as a universal function approximator [44]. Without this nonlinearity, a deep neural network would be equivalent to a single linear operation, albeit with more computational expense. Mathematically, these operations can be expressed as

$$\mathbf{h} = \phi(\mathbf{W}\mathbf{x} + \mathbf{b}) \tag{2.9}$$

which produces the output vector $\mathbf{h} \in \mathbb{R}^n$.

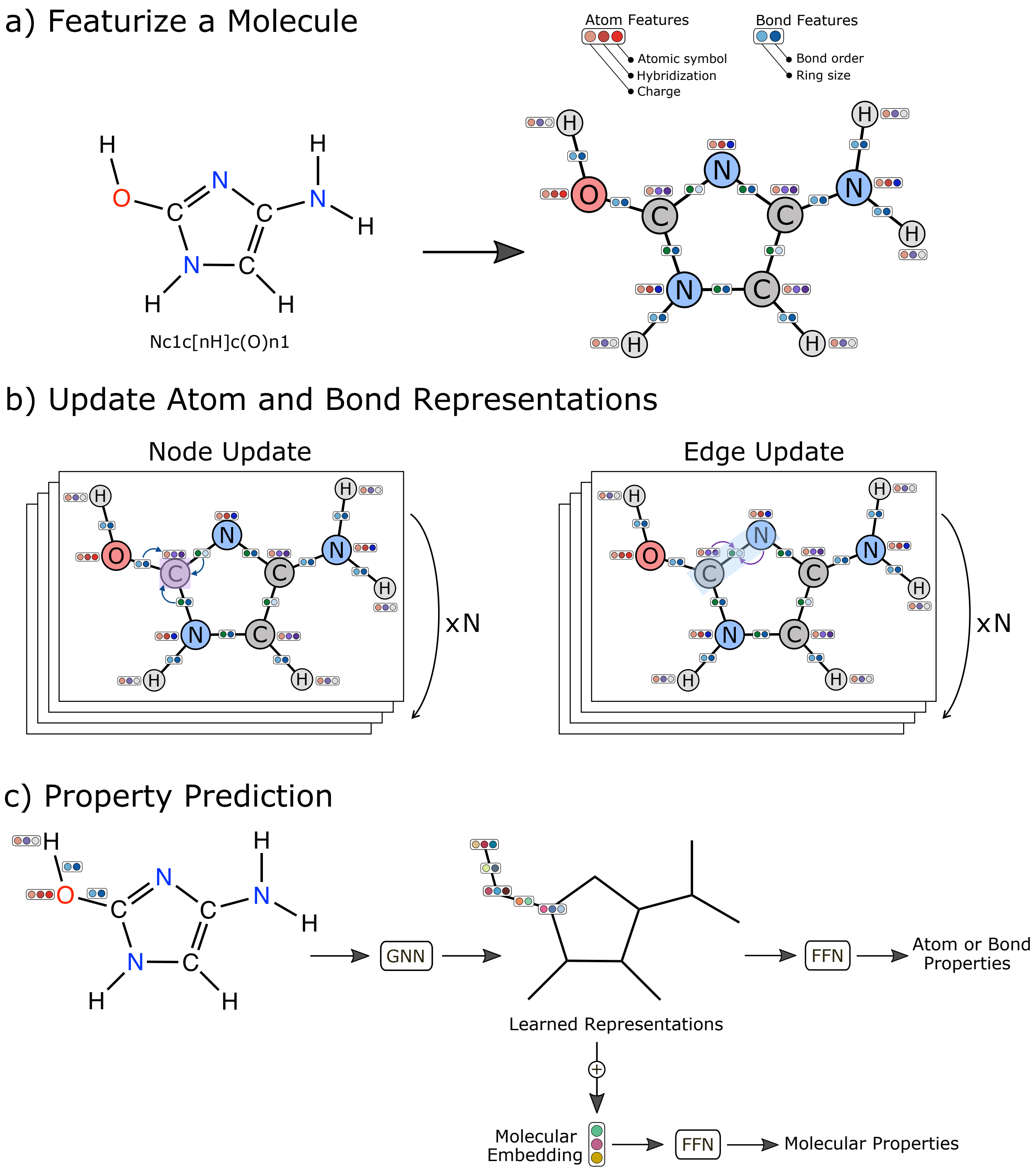
The choice of discriminative model architecture depends on the task of interest and the amount of available data. If working with relatively smaller datasets (e.g., fewer than 1,000 data points), traditional ML architectures may outperform neural networks [45]. In contrast, deep learning typically outperforms when one has access to thousands of datapoints.

2.2.5 Message Passing Neural Networks

Although many chemical prediction tasks still rely on pre-computed feature vectors that condense molecular representations into 1D vectors, many molecules are inherently described as graphical objects. Note that graph representations are not perfect as they often struggle with stereochemistry and multi-haptic bonds, but in many cases, graph-based architectures are a natural choice for molecular representation. I will loosely refer to these graph neural networks (GNNs) as "2D" methods, simply because the schematic of a molecular graph–such as that shown in Figure 2.1–is typically represented in two dimensions. However, it often does not make sense to discuss dimensions with graph theory since coordinates (i.e., atomic positions) are commonly not used when creating graph objects as input to GNNs.

The concept of message passing neural networks (MPNNs) was first formalized by Gilmer et al. [46] by building upon work from Kipf and Welling [47]. MPNNs operate on molecules by representing them as graphs $G = (V, E)$ with atoms corresponding to graph nodes or vertices (V) and covalent bonds corresponding to graph edges (E) [48, 49]. Each atom and bond is initially labeled with a corresponding feature vector, $v \in V$ and $e_{uv} \in E$ respectively. Example atomic features include its atomic symbol encoded as a one-hot vector as well as its degree, formal charge, and atomic mass. Example bond features include its bond type, the size of the ring if applicable, and whether the bond is aromatic. These features can be computed using standard cheminformatics packages, such as RDKit [50] or OpenBabel [51]. Figure 2.1 shows the main components of an MPNN applied to chemical data.

The next step is to perform message passing to iteratively update these atom and bond representations and propagate information throughout the graph. These updates are analogous to the procedure used by circular fingerprints described earlier. Several varieties of MPNNs have been developed in recent years, each claiming to improve performance on a set of chemical prediction tasks. The PyTorch Geometric [52] website compiles a list of several graph update procedures, along with their associated paper, which contains additional details for the interested reader. Despite their various nuances, the high-level goal is the same: create a learned representation. To update an atom representation, all neighboring atom and bond vectors undergo some matrix transformation–think of this as an FFN but some MPNNs use more advanced update steps. These updated vectors are then aggregated together, which entails a permutation invariant operation such as addition, averaging, or max pooling. A similar procedure is used to create updated representations for the bonds. This process is repeated for N iterations, termed the depth of the MPNN. The value of N controls how far the information propagates and is a hyperparameter that must be tuned during training. The FFN, and any other parameters used during the update steps, can be shared for both the atom and bond update or they can be allowed to vary independently.



a) Featurize a Molecule

b) Update Atom and Bond Representations

Node Update

Edge Update

c) Property Prediction

Figure 2.1. Schematic overview of the steps within a message passing neural network. The first step, shown in panel a), is to transform a molecule into an attributed graph with initial atom and bond feature vectors. The next step, shown in panel b), is to perform N iterations of message passing to update the atom and bond representations using information from neighboring atoms and bonds. Finally, panel c) shows that these updated representations can be used for various property prediction tasks. For simplicity, only a subset of the atom and bond features are displayed in pancel c).

The final step is the readout phase, which uses the updated representations for downstream property prediction. If the task of interest is an atom- or bond-level property, the updated atom or bond feature vectors can be directly passed through an FFN. If the task of interest is a molecular property, a molecular representation must be constructed, typically by aggregating the updated node representations together via mean or sum. The resulting molecular representation is then passed through an FFN. Thus, MPNNs create learned representations, which can be thought of as feature vectors that are customized for representing the specific chemical task. As a broad generalization, the purpose of a GNN, and other ML architectures such as a convolutional neural network or transformer block, is to create a learned representation (of fixed length that matches the user-specified hidden dimension) that is then passed to a FFN in the last step of the network to make a prediction. This goal is somewhat analogous to the approach described in Section 2.2.3, and it is here where the distinction between feature representations and model architectures can become blurred.

2.2.6 Directed MPNNs

Building on the formalization of MPNNs developed by Gilmer et al. [46], Yang et al. [53] introduces message passing over directed edges rather than over nodes. This new architecture is commonly referred to Chemprop. The representation of a directed edge $u - v$ is created by concatenating h_u with h_{vu} and then calculating messages based on these directed edges, which ultimately makes the D-MPNN architecture more expressive than traditional MPNNs. As explained in the publication from Yang et al. [53], the directed messages help to resolve "tottering" in which messages can oscillate back and forth between nodes, introducing additional noise in the hidden representation [54]. Gasteiger et al. [55] described the D-MPNN architecture as message passing over the *line graph* of the original molecular graph, which can be a helpful way of understanding the update steps.

2.2.7 3D MPNNs

Most of this thesis focuses on utilizing 2D MPNNs that rely in simple input representations to avoid the computational expense from obtaining 3D geometries. However, comparing the performance of 3D MPNNs with 2D MPNNs is an interesting research question, and Chapter 5 offers preliminary results towards this goal. Further, some tasks, such as predicting properties of different conformers, will obviously require information about the 3D structure so 2D MPNNs would be insufficient in this case. The main distinction between 3D and 2D

MPNNs is how graph edges are defined. While the latter typically use covalent bonds to define edges, the former uses a radius cutoff (e.g., all neighboring atoms within 5 Å are defined as connected); This former definition can create more edges than would be allowed by the physical constraints based on the valence of each atom. Another important difference is that for 3D MPNNs, the atomic featurization only includes the atom identity (atomic number), which is then passed through an embedding layer to expand it to higher dimensions.

To give a brief history, the goal of many 3D MPNNs is to act as a neural force field that can be used to predict energies and forces. However, this architecture can also be used to predict other regression properties as well. One early example of 3D MPNNs is SchNet [56], which uses learned convolutional filters that are a function of interatomic distances expanded with radial basis functions for greater expressivity and applies skip connections between node updates. PhysNet similarly considers interatomic distances [57]. Considering that the potential energy of molecules depends on bond angles, directional message passing neural network (DimeNet) builds on this by embedding distances and angles jointly using a spherical 2D Fourier-Bessel basis that allows message transforms to occur in a rotationally equivariant fashion [58]. DimeNet++ [59] leaves the core message passing unchanged, but incorporates relatively minor updates to the architecture and training procedure–replacing a bilinear layer with a Hadamard product, reducing the embedding size, using fewer hidden layers, and reducing the batch size–that ultimately result in a model that is 8x faster and 10% more accurate than the original DimeNet on the QM9 benchmark of equilibrium molecules [60]. SphereNet [61] and GemNet [62] also incorporate torsions into their models.

2.2.8 MPNNs for Reactions

Fitting models to predict properties of reactions requires some extra design considerations in comparison to property prediction of individual molecules. The most common approach is to individually featurize both the reactant(s) and product(s). Although some papers use features from the reactant and TS, identifying a TS geometry is often a challenging task that is not well-suited for high-throughput. Thus, this discussion will focus on the former. The core concept is to directly concatenate the feature vectors for the reactant(s) and product(s) or use the difference of these vectors to give insight into what changes during the reaction. This has been done with classical ML techniques for quite some time, with one example coming from Schneider et al. [63].

When it comes to more modern ML models, such as MPNNs, the main idea of concatenating representations is the same, though there are some subtle differences regarding when the concatenation occurs in the model architecture. One strategy comes from Grambow et al. [64] in which each reactant and product graph is passed through the same MPNN, whose job is to create a learned representation for a molecule. The learned atomic representations of the reactant are then subtracted from those of the product. The downstream architecture is identical to before in which the resulting atomic representations are then aggregated into a molecular (reaction) representation that is then passed to a standard FFN to predict the property of interest. Another possible approach could be to perform the subtraction after creating the aggregated molecular representation of the reactant(s) and product(s).

Another approach is to use the established condensed graph of reaction (CGR) representation [65–67], which creates a superposition of the graphs for the reactant complex and product complex. When using a CGR, only one graph is input to the MPNN. This time, the concatenation and/or subtraction occurs before message passing since the initial atom and bond features of the reactant and product complex are combined into one graph. Conceptually, the main benefit of this strategy is that it removes disjoint graphs present in multi-molecular reactions. For example, if a reaction creates two products, the CGR would allow message passing to occur between all atoms, which should make it more expressive than the approach from Grambow et al. [64] in which messages can only be passed between atoms within the same molecule (graph). This idea is supported by numerical results. As demonstrated by Heid and Green [68], CGR typically shows better performance, so this representation is used throughout this thesis. This approach requires atom-mapped reaction SMILES to compare the atoms and bonds between the reactant(s) and product(s) (i.e., we must know which reactant atoms lead to the corresponding product atoms so the concatenation or subtraction is done correctly).

2.3 References

(1) Hill, T. L., *An Introduction to Statistical Thermodynamics*; Courier Corporation: 1986.

(2) Callen, H. B. Thermodynamics and An Introduction to Thermostatistics, 1998.

(3) McQuarrie, D. A.; Simon, J. D., *Molecular Thermodynamics*; Sterling Publishing Company: 1999.

(4) Atkins, P. W.; Friedman, R. S., *Molecular Quantum Mechanics*; Oxford university press: 2011.

(5) Szabo, A.; Ostlund, N. S., *Modern quantum chemistry: introduction to advanced electronic structure theory*; Courier Corporation: 2012.

(6) Cramer, C. J., *Essentials of Computational Chemistry: Theories and Models*; John Wiley & Sons: 2013.

(7) Shell, M. S., *Thermodynamics and Statistical Mechanics: An Integrated Approach*; Cambridge University Press: 2015.

(8) Jensen, F., *Introduction to Computational Chemistry*; John wiley & sons: 2017.

(9) Slater, J. C. A Simplification of the Hartree-Fock Method. *Phys. Rev.* **1951**, *81*, 385.

(10) Slater, J. C. A Generalized Self-Consistent Field Method. *Phys. Rev.* **1953**, *91*, 528.

(11) Møller, C.; Plesset, M. S. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.* **1934**, *46*, 618.

(12) Coester, F.; Kümmel, H. Short-Range Correlations in Nuclear Wave Functions. *Nucl. Phys.* **1960**, *17*, 477–485.

(13) Tajti, A.; Szalay, P. G.; Császár, A. G.; Kállay, M.; Gauss, J.; Valeev, E. F.; Flowers, B. A.; Vázquez, J.; Stanton, J. F. HEAT: High Accuracy Extrapolated Ab initio Thermochemistry. *J. Chem. Phys.* **2004**, *121*, 11599–11613.

(14) Karton, A.; Rabinovich, E.; Martin, J. M. L.; Ruscic, B. W4 Theory for Computational Thermochemistry: In Pursuit of Confident Sub-kJ/mol Predictions. *J. Chem. Phys.* **2006**, *125*, 144108.

(15) Ziegler, T. Approximate Density Functional Theory as a Practical Tool in Molecular Energetics and Dynamics. *Chem. Rev.* **1991**, *91*, 651–667.

(16) Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *136*, B864.

(17) Becke, A. D. Perspective: Fifty Years of Density-Functional Theory in Chemical Physics. *J. Chem. Phys.* **2014**, *140*, 18A301.

(18) Mardirossian, N.; Head-Gordon, M. Thirty Years of Density Functional Theory in Computational Chemistry: An Overview and Extensive Assessment of 200 Density Functionals. *Mol. Phys.* **2017**, *115*, 2315–2372.

(19) Eyring, H. The Activated Complex in Chemical Reactions. *J. Chem. Phys.* **1935**, *3*, 107–115.

(20) Laidler, K. J.; King, M. C. The Development of Transition-State Theory. *J. Phys. Chem.* **1983**, *87*, 2657–2664.

(21) Truhlar, D. G.; Garrett, B. C.; Klippenstein, S. J. Current Status of Transition-State Theory. *J. Phys. Chem.* **1996**, *100*, 12771–12800.

(22) Yu, T.; Zheng, J.; Truhlar, D. G. Multi-Structural Variational Transition State Theory. Kinetics of the 1,4-Hydrogen Shift Isomerization of the Pentyl Radical with Torsional Anharmonicity. *Chem. Sci.* **2011**, *2*, 2199–2213.

(23) Yu, T.; Zheng, J.; Truhlar, D. G. Multipath Variational Transition State Theory: Rate Constant of the 1,4-Hydrogen Shift Isomerization of the 2-Cyclohexylethyl Radical. *J. Phys. Chem. A* **2012**, *116*, 297–308.

(24) Hammett, L. P. The Effect of Structure Upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **1937**, *59*, 96–103.

(25) Rogers, D.; Hopfinger, A. J. Application of Genetic Function Approximation to Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854–866.

(26) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. QSPR: The Correlation and Quantitative Prediction of Chemical and Physical Properties from Structure. *Chem. Soc. Rev.* **1995**, *24*, 279–287.

(27) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31–36.

(28) Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. InChI-The Worldwide Chemical Structure Identifier Standard. *J. Cheminform.* **2013**, *5*, 1–9.

(29) Selwood, D. L.; Livingstone, D. J.; Comley, J. C. W.; O'Dowd, A. B.; Hudson, A. T.; Jackson, P.; Jandu, K. S.; Rose, V. S.; Stables, J. N. Structure-Activity Relationships of Antifilarial Antimycin Analogs: A Multivariate Pattern Recognition Study. *J. Med. Chem.* **1990**, *33*, 136–142.

(30) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(31) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–113.

(32) Gedeck, P.; Rohde, B.; Bartels, C. QSAR - How Good Is It in Practice? Comparison of Descriptor Sets on an Unbiased Cross Section of Corporate Data Sets. *J. Chem. Inf. Model.* **2006**, *46*, 1924–1936.

(33) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.

(34) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.

(35) Rupp, M.; Ramakrishnan, R.; Von Lilienfeld, O. A. Machine Learning for Quantum Mechanical Properties of Atoms in Molecules. *J. Phys. Chem. Lett.* **2015**, *6*, 3309–3313.

(36) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; Von Lilienfeld, O. A.; Muller, K.-R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326–2331.

(37) Axilrod, B. M.; Teller, E. Interaction of the Van der Waals Type Between Three Atoms. *J. Chem. Phys.* **1943**, *11*, 299–300.

(38) Huang, B.; von Lilienfeld, O. A. Quantum Machine Learning using Atom-in-Molecule-Based Fragments Selected on the Fly. *Nat. Chem.* **2020**, *12*, 945–951.

(39) Faber, F. A.; Christensen, A. S.; Huang, B.; Von Lilienfeld, O. A. Alchemical and Structural Distribution Based Representation for Universal Quantum Machine Learning. *J. Chem. Phys.* **2018**, *148*.

(40) Christensen, A. S.; Bratholm, L. A.; Faber, F. A.; Anatole von Lilienfeld, O. FCHL Revisited: Faster and More Accurate Quantum Machine Learning. *J. Chem. Phys.* **2020**, *152*.

(41) Bartók, A. P.; Kondor, R.; Csányi, G. On Representing Chemical Environments. *Phys. Rev. B* **2013**, *87*, 184115.

(42) Grisafi, A.; Wilkins, D. M.; Csányi, G.; Ceriotti, M. Symmetry-Adapted Machine Learning for Tensorial Properties of Atomistic Systems. *Phys. Rev. Lett.* **2018**, *120*, 036002.

(43) Rosenblatt, F. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychol. Rev.* **1958**, *65*, 386.

(44) Hornik, K.; Stinchcombe, M.; White, H. Multilayer Feedforward Networks are Universal Approximators. *Neural Networks* **1989**, *2*, 359–366.

(45) Chuang, K. V.; Gunsalus, L. M.; Keiser, M. J. Learning Molecular Representations for Medicinal Chemistry. *J. Med. Chem.* **2020**, *63*, 8705–8722.

(46) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. In *International Conference on Machine Learning*, 2017, pp 1263–1272.

(47) Kipf, T. N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv* **2016**, *1609.02907*.

(48) Bacciu, D.; Errica, F.; Micheli, A.; Podda, M. A Gentle Introduction to Deep Learning for Graphs. *Neural Networks* **2020**, *129*, 203–221.

(49) Fioresi, R.; Zanchetta, F. Deep Learning and Geometric Deep Learning: An Introduction for Mathematicians and Physicists. *arXiv* **2023**, *2305.05601*.

(50) Landrum, G. et al. RDKit: Open-Source Cheminformatics Software, https://www.rdkit.org, 2006.

(51) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminf.* **2011**, *3*, 1–14.

(52) Fey, M.; Lenssen, J. E. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

(53) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M., et al. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388.

(54) Mahé, P.; Ueda, N.; Akutsu, T.; Perret, J.-L.; Vert, J.-P. In *Proceedings of the twenty-first international conference on Machine learning*, 2004, p 70.

(55) Gasteiger, J.; Yeshwanth, C.; Günnemann, S. Directional Message Passing on Molecular Graphs via Synthetic Coordinates. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 15421–15433.

(56) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet–A Deep Learning Architecture for Molecules and Materials. *J. Chem. Phys.* **2018**, *148*, 241722.

(57) Unke, O. T.; Meuwly, M. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *J. Chem. Theory Comput.* **2019**, *15*, 3678–3693.

(58) Klicpera, J.; Groß, J.; Günnemann, S. Directional Message Passing for Molecular Graphs. *arXiv* **2020**, *2003.03123*.

(59) Klicpera, J.; Giri, S.; Margraf, J. T.; Günnemann, S. Fast and Uncertainty-Aware Directional Message Passing for Non-Equilibrium Molecules. *arXiv* **2020**, *2011.14115*, Accessed 2022-03-08.

(60) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1*, 1–7.

(61) Liu, Y.; Wang, L.; Liu, M.; Zhang, X.; Oztekin, B.; Ji, S. Spherical Message Passing for 3D Graph Networks. *arXiv* **2021**, *2102.05013*.

(62) Gasteiger, J.; Becker, F.; Günnemann, S. Gemnet: Universal Directional Graph Neural Networks for Molecules. *Adv. Neural. Inf. Process. Syst.* **2021**, *34*, 6790–6802.

(63) Schneider, N.; Lowe, D. M.; Sayle, R. A.; Landrum, G. A. Development of a Novel Fingerprint for Chemical Reactions and Its Application to Large-Scale Reaction Classification and Similarity. *J. Chem. Inf. Model.* **2015**, *55*, 39–53.

(64) Grambow, C. A.; Pattanaik, L.; Green, W. H. Deep Learning of Activation Energies. *J. Phys. Chem. Lett.* **2020**, *11*, 2992–2997.

(65) Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V. P. Substructural Fragments: An Universal Language to Encode Reactions, Molecular and Supramolecular Structures. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 693–703.

(66) Hoonakker, F.; Lachiche, N.; Varnek, A.; Wagner, A. Condensed Graph of Reaction: Considering a Chemical Reaction as One Single Pseudo Molecule. *Int. J. Artif. Intell. Tools* **2011**, *20*, 253–270.

(67) Madzhidov, T. I.; Polishchuk, P. G.; Nugmanov, R. I.; Bodrov, A. V.; Lin, A. I.; Baskin, I. I.; Varnek, A. A.; Antipin, I. Structure-Reactivity Relationships in Terms of the Condensed Graphs of Reactions. *Russ. J. Org. Chem.* **2014**, *50*, 459–463.

(68) Heid, E.; Green, W. H. Machine Learning of Reaction Properties via Learned Representations of the Condensed Graph of Reaction. *J. Chem. Inf. Model.* **2021**, *62*, 2101–2110.

Chapter 3

High Accuracy Barrier Heights, Enthalpies, and Rate Coefficients for Chemical Reactions

Much of this work has previously appeared as Spiekermann, K. A.; Pattanaik, L.; Green, W. H. High Accuracy Barrier Heights, Enthalpies, and Rate Coefficients for Chemical Reactions. *Sci. Data* **2022**, *9*, 417 [link]. Lagnajit Pattanaik helped clean the data (e.g., update the SMILES). All data is free and publicly accessible at https://zenodo.org/record/6618262.

3.1 Introduction

Detailed reaction mechanisms are valuable tools for analyzing and predicting physical phenomena driven by chemical kinetics. Historically, kinetic model parameters were fit to a specific set of experimental results, which limited their generalizability to systems with different temperatures, pressures, or initial compositions. In recent decades, the field of chemical kinetics has transitioned from postdictive to predictive modeling approaches [1]. This shift has been motivated by advances in compute power, which make it possible to predict many kinetic parameters using ab initio calculations rather than relying on scarce experimental data [2–13]. Our research group has long been interested in the automated generation of kinetic models, which can simulate and predict the concentrations of all relevant species [14, 15]. Reliable datasets are essential for constructing such models with predictive power. A small error of a few kcal mol^{-1} in the activation energy will lead to significant errors in the final rate estimate, particularly at lower temperatures. Unfortunately, accurate barriers are often known for fewer than 10% of the reactions in kinetic models [3–5, 11–13, 16, 17].

To help address this paucity of data, here we present relatively accurate barriers for nearly 12,000 reactions.

Kinetic parameters are currently estimated using functional group and linear-free-energy (LFER) methods, [18, 19], but machine learning models are much more flexible and have broader scope. Indeed, machine learning has sparked an explosion of progress in physical and organic chemistry, especially in the areas of automated synthesis planning [20, 21], targeted molecular optimization [22, 23], and general property prediction [24, 25] from thermodynamic [26, 27] and solvation parameters [28, 29] to full infrared spectra [30]. In situations where data are plentiful, machine learning-based algorithms often provide excellent predictions and some have successfully been applied to experiments [31]. When such data is lacking, researchers generate their own datasets–both experimentally and computationally–to regress desired properties from them [32–34]. In machine learning applied to chemistry, the community has largely taken a model-driven approach, where significant effort has been devoted to refining models on a few benchmark datasets [35, 36]. As a result, the community has delivered strong architectures from advanced graph convolutional networks [37–39] to atomistic networks [40, 41]. Today, progress is limited primarily by the scarcity of large, diverse, and high-quality datasets.

Here, we report a cleaned, high-quality dataset of reaction barriers, enthalpies, and transition state theory (TST) rate coefficients. We build upon the prior work from Grambow et al. [42, 43] Briefly, their work used the single-ended growing string method [44] to automatically identify thousands of transition states (TSs) and products from a given set of reactants. Reactants were chosen by using all molecules with six or fewer heavy atoms from GDB-7 [45] as well as randomly selecting some ($\sim$430) molecules with seven heavy atoms; the molecules contain H, C, N, and O atoms. Conformer searches were performed for the reactants by embedding several hundred conformers for each molecule using RDKit [46] with the ETKDG distance geometry method [47] and relaxing their geometries using the MMFF94 force field implemented in RDKit. The lowest energy conformer was then optimized using Q-Chem [48] at both the B97-D3/def2-mSVP level of theory with Becke-Johnson damping [49] and the ωB97X-D3/def2-TZVP [50] level of theory. The reactant conformer was the starting point for the growing string search. The highest energy point in the string was used as the initial guess for a conventional saddle point search. Additional details can be found in the original publication.

Our work brings the following advances. First, we clean the SMILES [51] reported in Grambow et al. [42] The original publication treated all reactions with multiple products as containing one product complex, which does not conform well with traditional TST calculations that expect partition functions for each species. Thus, we separate the products from any product complex, recalculate the geometry optimization and frequency at either B97-D3/def2-mSVP or ωB97X-D3/def2-TZVP. Finally, we refine the single point energies for each species using explicitly correlated coupled-cluster calculations, which are expected to be much more accurate than the density functional theory methods [52–55]. We provide the updated barrier heights and reaction enthalpies from our CCSD(T)-F12a/cc-pVDZ-F12//ωB97X-D3/def2-TZVP calculations. Our higher-accuracy calculations improve the RMSE of the barrier heights by approximately 5 kcal mol^{-1} relative to those calculated at ωB97X-D3/def2-TZVP. We believe that the high-quality values in this dataset will accelerate development of automated and reliable methods for quantitative reaction prediction. Finally, we also identify a subset of reactions with rigid species that do not require a conformer search nor hindered-rotor treatment. The rigid-rotor harmonic oscillator (RRHO) TST rate coefficients $k_\infty(T)$ and fitted Arrhenius parameters for this subset are reported since these values should be accurate. We do not report $k_\infty(T)$ or Arrhenius parameters for reactions involving flexible reactants or transition states since RRHO TST is not accurate for these reactions.

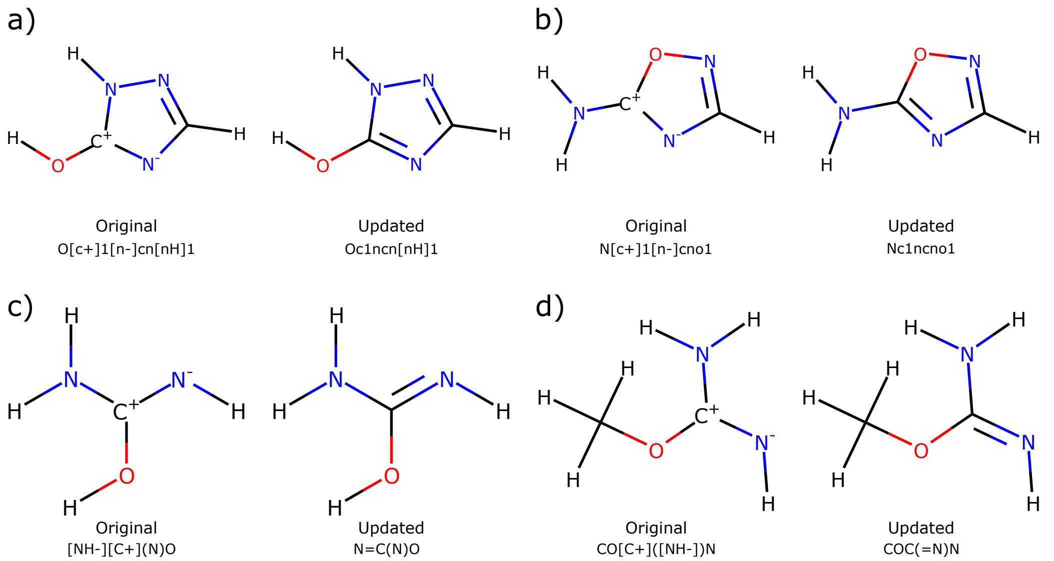
3.2 Methods

3.2.1 Overview

Dataset refinement started by cleaning the SMILES from the original dataset [43] and filtering reactions to those containing one reactant and at most three products. Next, product complexes were separated into individual species, each of which was reoptimized at the respective level of theory i.e., either B97-D3/def2-mSVP or ωB97X-D3/def2-TZVP. The single point energy of all species optimized at ωB97X-D3/def2-TZVP was computed at CCSD(T)-F12a/cc-pVDZ-F12. These energies were used to calculate updated barrier heights by adding the zero-point energies (ZPEs) from the harmonic vibrational analysis to the reactant, product, and TS energies and then computing the difference between the resulting TS and reactant energies. Similarly, enthalpies of reaction were calculated based on the difference of the ZPE-corrected product and reactant energies; bond additivity corrections (BACs) were added to each species. Finally, we identify a subset of reactions that contain rigid species, calculate high-pressure limit TST rate coefficients, and report the fitted Arrhenius parameters.

3.2.2 Cleaning SMILES

The work from Grambow et al. [42] used the single-ended growing string method [44] to generate a list of possible products from a given reactant. The input and output for the growing string method are a set of three-dimensional coordinates to describe the molecule or multi-molecule complex. Grambow et al. used Open Babel [56] to perceive connectivity and generate a SMILES for the reactant and product from each set of three-dimensional coordinates. However, in some cases, the bond-order and formal charges did not correspond to the most representative resonance structure. Here, we update the SMILES by using RDKit [46] to look for neighboring atoms with opposite formal charges, which often occurred between nitrogen and carbon atoms. Some representative examples of the updated SMILES and their impact on molecular structure are shown in Figure 3.1. Additionally, there were a handful of reactions whose reactant was neutral, but whose product was positively charged. This charge imbalance was likely due to Open Babel occasionally generating an incorrect SMILES from the molecular coordinates. Here, we update the corresponding product SMILES to conserve charge for the reaction i.e., added an electron to create a correct Lewis structure. Representative examples are shown in Figure 3.2. Atom-mapping is preserved when updating the SMILES.



a) Original O[c+]1[n-]cn[nH]1 Updated Oc1ncn[nH]1

b) Original N[c+]1[n-]cno1 Updated Nc1ncno1

c) Original [NH-][C+](N)O Updated N=C(N)O

d) Original CO[C+]([NH-])N Updated COC(=N)N

Figure 3.1. Representative examples of updating SMILES to correspond to the most representative resonance structure. For convenience, the SMILES shown here omit atom-map numbers.

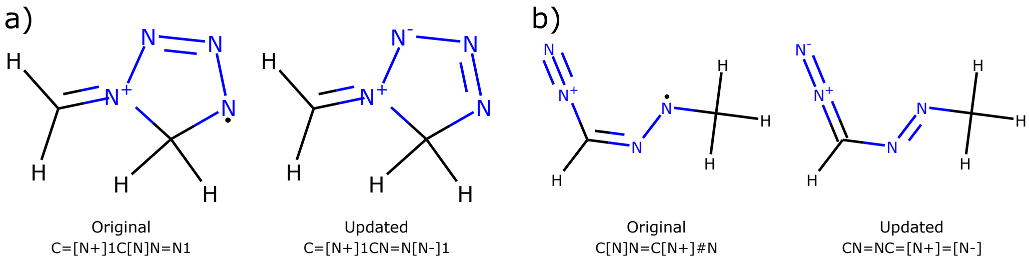


Figure 3.2. Representative examples of updating SMILES to fix incorrect charge imbalances. For convenience, the SMILES shown here omit atom-map numbers.

3.2.3 Reoptimizing Products

Of the reactions in the previously published dataset from Grambow et al. [42], approximately 30% contain two products and 2% contain three products. These were previously treated as one product complex when using the growing string method as well as during subsequent geometry optimization and frequency calculation. However, to obtain rate coefficients, conventional canonical TST calculations expect partition functions for each individual species. Here, we separate the complexes into individual products, reoptimize the geometries, and recalculate the frequencies using Q-Chem 5.3.0 [48] for both the B97-D3/def2-mSVP and ωB97X-D3/def2-TZVP datasets. The previously optimized geometries from the complex are used as the initial guess for the new optimization. The exact same settings are used in the input files as were used by Grambow et al.[42] during the original Q-Chem calculations. This ensures that the separated products are run with the same method and basis set as well as with identical convergence criteria as those used by their corresponding reactant and TS.

Consistent with the previous work, nearly all molecules are run in the singlet state and use a spin-unrestricted ansatz. For example, the ground electronic state for methylene (CH_2) is a triplet, but because the TS for all reactions was computed at the singlet state, any CH_2 products were also recalculated in the singlet state. However, upon splitting some products, there are 49 reactions unique to the larger B97-D3 dataset whose product pairs were radicals; these individual species were calculated in the doublet state since it was assumed that the lone electron on each product had opposite spins to conserve the overall multiplicity for the reaction. We verified that spin contamination was not a problem by confirming that the average value of the total spin operator was between 0.75 and 0.77 for these species.

Note that reoptimizing the separated products and then summing their energy resulted in a different product energy than that from the original product complex. In a few cases, this changed the reaction enthalpy enough such that $\Delta H > \Delta E_0$, which would cause the reverse reaction to have a negative barrier height. Although submerged barriers are possible, often a large negative barrier height is reason to be suspicious. Thus, we remove any reaction in which the explicitly correlated coupled-cluster reaction enthalpy was more than 10 kcal mol^{-1} larger than the barrier height.

3.2.4 Refining Single Point Energies

A major accomplishment of this work is providing highly accurate kinetic parameters computed at CCSD(T)-F12a/cc-pVDZ-F12//ωB97X-D3/def2-TZVP for a large and diverse set of atom-mapped gas phase reactions. Although the ωB97X-D3 method is more accurate for predicting barrier heights than many other functionals [57], coupled-cluster CCSD(T) calculations are commonly considered the gold-standard in quantum chemistry [58, 59]. Here, we refine the single point energies of each species from the ωB97X-D3/def2-TZVP dataset using the explicitly correlated CCSD(T)-F12 method since previous literature has shown that CCSD(T)-F12 can achieve similar accuracy to the standard CCSD(T) calculation while using a much smaller basis set [6, 52–55, 60]. This is notable because both coupled-cluster methods scale as $\mathcal{O}(N^7)$, such that N is the number of orbitals [61], and so using a smaller basis set offers substantial computational savings.

We next consider which basis set to use. Although triple-ζ and quadruple-ζ basis sets have shown reaction energies within 1 kcal mol^{-1}, many studies comparing basis sets use about 100 molecules or fewer. Further, such studies often focus on small molecules containing primarily three or four heavy atoms due to the steep scaling of coupled-cluster calculations. The main exception to this generalization is the recent calculation of the 133,000 molecules from QM9 [35] with the G4MP2 level of theory [62, 63]; however, that dataset only contains stable species, while our dataset has approximately 12,000 TSs. Considering that our dataset contains nearly 22,000 unique stable species and transition states, each containing up to seven heavy atoms, we chose the cc-pVDZ-F12 basis set to accommodate the large number of calculations while maintaining high accuracy. 15 reactions were also run with cc-pVTZ-F12 to validate the double-ζ accuracy. For all coupled-cluster single point calculations, we use the energy from CCSD(T)-F12a since both published literature [64] and the MOLPRO documentation [65] conclude this method offers a better approximation to the complete basis set limit for the double-ζ and triple-ζ basis sets. All calculations were run in parallel using MOLPRO 2015.1 [65] on the National Energy Research Scientific Computing Center (NERSC).

3.2.5 Calculating Reaction Barrier Heights and Enthalpies

ZPEs from the harmonic vibrational analysis are added to the electronic energy for the reactant, product, and TS. For all species, a scaling factor is applied to the computed harmonic frequencies used to compute the ZPE; the scaling factor for B97-D3/def2-mSVP and for ωB97X-D3/def2-TZVP are calculated as described by Alecu et al. [66] and found to be 1.014 and 0.984 respectively. Reaction barrier heights are computed by taking the difference of the resulting TS and reactant energies. Similarly, enthalpies of reaction at 298 K are computed by taking the difference of the resulting product and reactant energies. When calculating the values for the coupled-cluster dataset, the CCSD(T)-F12a energies are used while the ZPEs are taken from the ωB97X-D3 calculation since this was the level of theory used for the geometry optimization and vibrational analysis. Note that atom energy corrections (AECs) and bond additivity corrections (BACs) are added to the enthalpy values for each species. Although the AECs cancel out during the subtraction to obtain the reaction enthalpy since all reactions are balanced, these corrections are important when comparing the Δ_fH(298 K) to experimental values as described in the technical validation. Corrections are not used when computing reaction barriers. AECs are calculated by fitting the atomization energies of 14 small molecules. The atomization energies come from CCCBDB [67] and all have uncertainty values less than 0.2 kcal mol^{-1}. Petersson type BACs [68] were fit using a set of about 400 reference species with well-known heats of formation, primarily drawn from ATcT [69] and CCCBDB [67]. The experimental uncertainty is at most 0.55 kcal mol^{-1}, though most values are much lower with the median being just 0.14 kcal mol^{-1}. For more details on the fitting procedure, see the Reaction Mechanism Generator (RMG) documentation at `https://reactionmechanismgenerator.github.io/RMG-Py/users/arkane/input.html#atom-energy-fitting`.

Calculating Rates

Automated Reaction Kinetics and Network Exploration (Arkane) is a software package for computing thermodynamic properties and high-pressure limit rate coefficients using the results from quantum chemistry calculations. Thermodynamic properties are computed using the RRHO approximation, while kinetic parameters are computed using conventional canonical TST, also with RRHO. Arkane is developed and distributed as part of RMG-Py [14, 15]. All software is written in Python and provided as free, open source code under the terms of the MIT License.

We use Arkane to convert the single point energy from the quantum chemistry calculation to the gas-phase reference state; by default atom and spin-orbit coupling energy corrections are applied, but will cancel during the TST calculation. As before, the corresponding scaling factor is applied to the ZPE for each species. Arkane uses RRHO TST with Eckart tunneling correction to calculate the forward rate coefficient for a set of user-defined temperatures. BACs are omitted when calculating the forward rate coefficient since BACs are not present for the partial bonds in the TS. Arkane then uses a linear least-squares fitting to fit the list of reciprocal temperatures and logarithm of the rate coefficients to an Arrhenius expression, yielding the best approximation for the pre-exponential A-factor and the activation energy. We use 50 linearly spaced points in the reciprocal temperature space between 300 K and 2000 K when obtaining the Arrhenius parameters.

3.3 Data Records

All data is free and publicly accessible on Zenodo [70]. Q-Chem output files are provided for the 16,302 reactions at B97-D3/def2-mSVP and for the 11,926 reactions at ωB97X-D3/def2-TZVP level of theory. For convenience, these also include the original log files for the reactant, TS, and non-reoptimized products from Grambow et al. [42] since they are used to calculate barrier heights, enthalpies, and rate coefficients in this work. MOLPRO output files from the single point calculations are provided for the 11,926 reactions at the CCSD(T)-F12a/cc-pVDZ-F12 level of theory as well as for the 15 reactions calculated with the triple-ζ basis. Information for each reaction is organized by the level of theory and stored in a separate folder labeled as `rxn######`, such that `######` denotes the reaction number padded with zeros. The numbering matches that from the originally published dataset [43] to facilitate easy comparison. For the quantum chemistry calculations, each folder contains the log files for the reactant, TS, and product as `r######.log`, `ts######.log`, and `p######.log` respectively. An additional number is appended to the file names from the separated products. For example, the log files for any reaction containing two products are labeled as `p######_0.log` and `p######_1.log`.

The cleaned atom-mapped SMILES, as well as all values calculated in this work, are provided in the comma-separated values (csv) files `b97d3.csv`, `wb97xd3.csv`, `ccsdtf12_dz.csv`, and `ccsdtf12_tz.csv`. The columns for the csv files are described in Table 3.1. The calculated TST rate coefficients and fitted Arrhenius parameters for the rigid species are provided in `ccsdtf12_dz_rigid.csv`, whose columns are described in Table 3.2. The Arkane output files from TST calculations and Arrhenius parameter fitting are also provided. Each reaction is again stored in a separate folder labeled as `rxn######`, which contains a `rxn` folder with all information from the Arrhenius fitting. The kinetic information is stored in Chemkin [71] file format. The list of 50 temperatures (K) used during Arrhenius fitting is provided in `arkane_temperatures.csv`.

The improvement from fitting BACs at B97-D3/def2-mSVP, ωB97X-D3/def2-TZVP, CCSD(T)-F12a/cc-pVDZ-F12//ωB97X-D3/def2-TZVP, and CCSD(T)-F12a/cc-pVTZ-F12//ωB97X-D3/def2-TZVP is contained in `b97d3_def2msvp_BAC.csv`, `wb97xd3_def2tzvp_BAC.csv`, `ccsdtf12_ccpvdzf12__wb97xd3_def2tzvp_BAC.csv`, and `ccsdtf12_ccpvtzf12__wb97xd3_def2tzvp_BAC.csv` respectively. The files contain the experimental and calculated enthalpies for the reference species from RMG-database used for fitting. The atom and bond correction values are publicly stored on the RMG-database GitHub, though they are also provided in `fitted_corrections.pkl` for convenience. Further validation of the BACs at the double-ζ basis set is done by comparing to experimental values from the Pedley [72] set since over half of these molecules are not in the RMG-database training set used for fitting. The comparison is shown in `ccsdtf12_dz_vs_Pedley_experimental.csv`.

Table 3.1. Description of the columns in the main comma-separated value files for each level of theory.

Column label	**Description**
idx	Reaction index
rsmi	Reactant SMILES
psmi	Product SMILES
rinchi	Reactant InChI
pinchi	Product InChI
dE0	Barrier height $\left(\text{kcal mol}^{-1}\right)$
dHrxn298	Enthalpy of reaction $\left(\text{kcal mol}^{-1}\right)$
rmg_family	RMG reaction family

Table 3.2. Description of the columns in the comma-separated value file for rigid reactions.

Column label	Description
idx	Reaction index
rsmi	Reactant SMILES
psmi	Product SMILES
k(T0) to k(T49)	50 columns with the calculated rate coefficient (s^{-1})
lnA	Natural log of the fitted pre-exponential factor (s^{-1})
Ea	Fitted activation energy $\left(\text{kcal mol}^{-1}\right)$
percent_error	Average absolute percent error between the calculated and fitted rate coefficients

3.4 Technical Validation

The published work from Grambow et al. [42] already performed several integrity checks, such as ensuring that all TSs have exactly one imaginary frequency, whose atomic displacements matched the bond changes occurring between the reactant and product. The authors also removed any TS with an imaginary frequency less than 100 cm^{-1} in magnitude as that typically corresponds to conformational changes. In this work, we ensure that multiplicity and charge are conserved for all reactions. As described in the methods section, this is important when separating product complexes into individual product geometries for reoptimization.

We next identify whether each reaction matches a reaction template from the RMG-database. As shown in Figure 3.3, keto-enol is the most represented RMG template. However, due to the diversity of these reactions, the majority do not match any RMG template. This is consistent with the previously published work [42], which chose to characterize the reaction diversity by extracting general templates that do not necessarily match a template from RMG-database. RMG-database is frequently updated, which includes occasionally updating the reaction templates to be more broad or more specific. Thus, using reaction templates from a different version of RMG-database may capture more or less reactions for a given family (or even include different reaction families). To generate Figure 3.3 and

Table 3.3, we used the `AEC_BAC` branch of RMG-database.

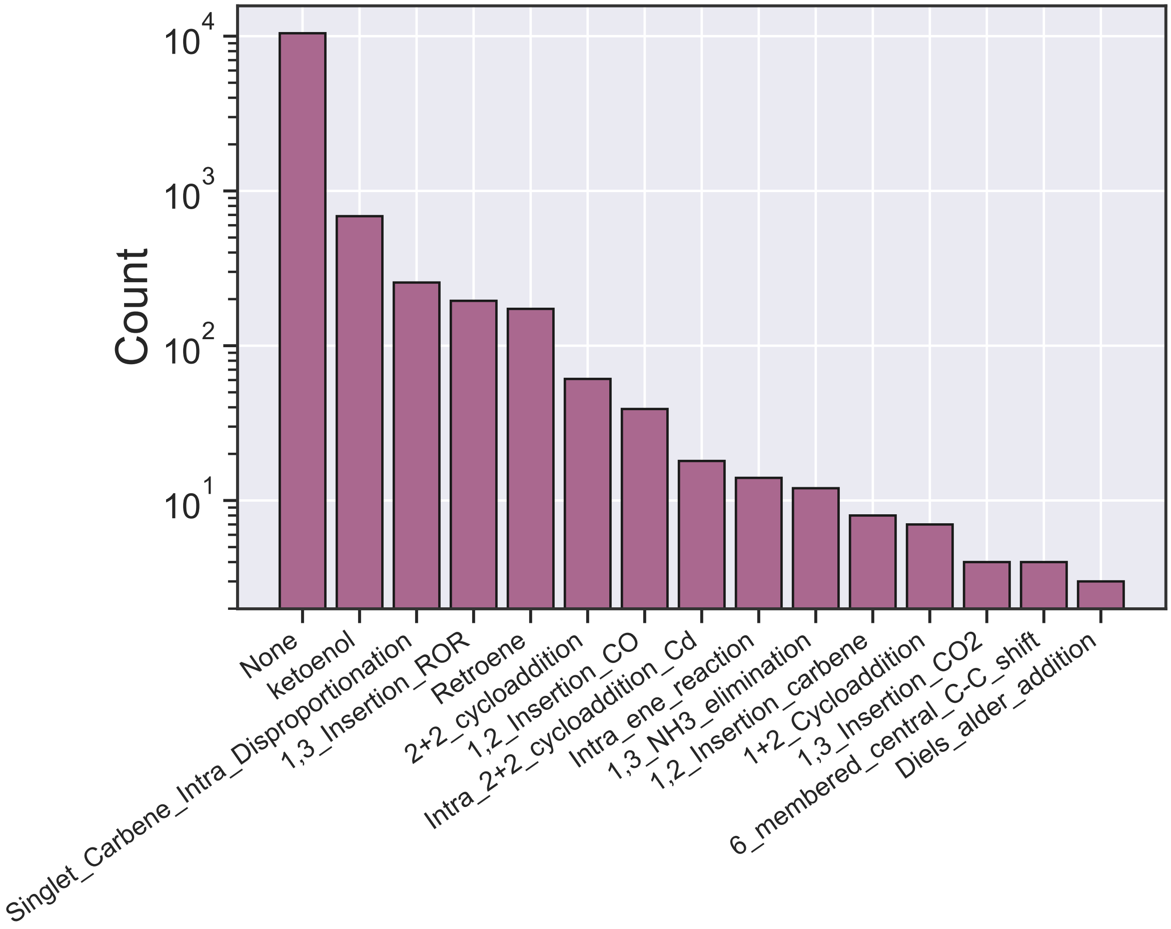
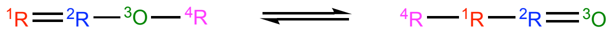
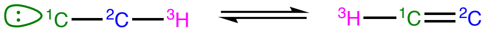
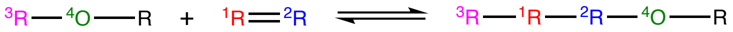
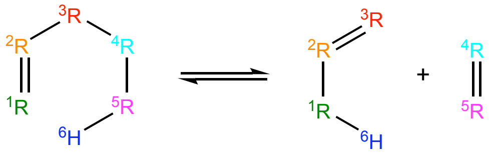
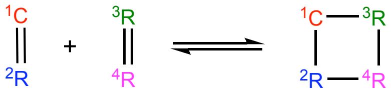
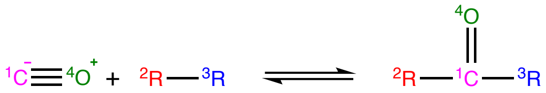
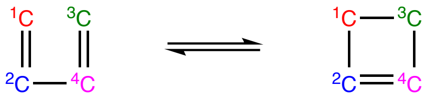
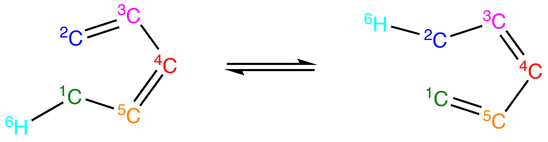
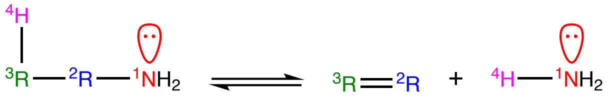
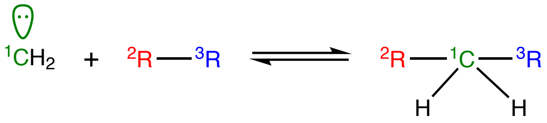
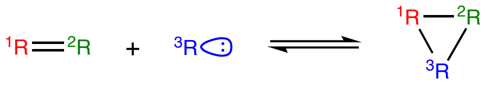
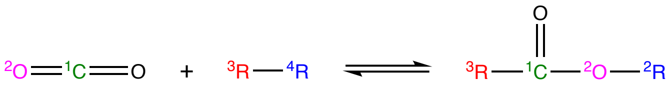
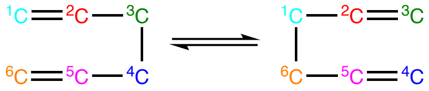
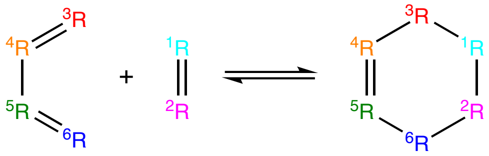


Figure 3.3. Distribution of RMG reaction families present in the CCSD(T)-F12a/cc-pVDZ-F12 dataset.

Table 3.3. RMG reaction templates present in the CCSD(T)-F12a/cc-pVDZ-F12 dataset.

RMG Reaction Family	Template
ketoenol	
Singlet_Carbene_Intra_Disproportionation	
1,3_Insertion_ROR	
Retroene	
2+2_cycloaddition	
1,2_Insertion_CO	
Intra_2+2_cycloaddition_Cd	
Intra_ene_reaction	
1,3_NH3_elimination	
1,2_Insertion_Carbene	
1+2_Cycloaddition	
1,3_Insertion_CO2	
6_membered_central_C-C_shift	
Diels_alder_addition	

To evaluate the improvement from applying the fitted atom and bond corrections to the enthalpy values, we calculate the error relative to the high-quality reference set of about 400 molecules used for fitting. For the coupled-cluster data, the training mean absolute error (MAE) and root mean squared error (RMSE) are 0.5 and 0.8 kcal mol^{-1} respectively. We also compare the corrected enthalpy values to an external test set originally published by Pedley [72] and compiled and verified by Narayanan et al. [62] to validate our enthalpy calculation approach (CCSD(T)-F12a + AEC + BAC). This set, named the Pedley test set, contains 459 species that have experimental uncertainty, defined as 95% confidence intervals [73, 74], of less than 1 kcal mol^{-1}. After removing 76 species common to both our reference set and our coupled-cluster dataset (so as to exclude training species from our test set), we measure the error of our corrected enthalpies against the Pedley test set. This evaluation returns an MAE of 0.8 kcal mol^{-1} and RMSE of 1.2 kcal mol^{-1}, indicating strong agreement of our approach with high-quality experimental data.

To compare accuracy improvements from the coupled-cluster calculations in this work, Table 3.4 shows the MAE and RMSE of the barrier height and reaction enthalpy relative to the lower levels of theory. As expected, values from the ωB97X-D3 dataset show less deviation than those from the B97-D3 dataset, yet they are still several kcals away from the explicitly correlated coupled-cluster values. The RMSE of 5 kcal mol^{-1} is significant since it implies that rate coefficients calculated at ωB97X-D3 differ on average by a factor of 12 at 1,000 K relative to those calculated at CCSD(T)-F12a; the difference increases substantially to a factor of 4,000 at 300 K. It is interesting to note that the RMSE for barrier heights reported here is more than twice as large as the RMSE reported in ref. [57]. However, this previous analysis was done with 206 reactions from just a few reaction families, whereas the data presented in Table 3.4 represent more than 10,000 reactions and a much more diverse array of chemistry. Taking the barrier height RMSE of only RMG reaction families gives 8.0 and 3.8 kcal mol^{-1} for the B97-D3 and ωB97X-D3 dataset respectively, both of which are smaller deviations than that for the entire dataset. As seen in Figure 3.4, the barrier heights calculated at DFT tended to be an overestimate relative to those calculated at CCSD(T)-F12a/cc-pVTZ-F12//ωB97X-D3/def2-TZVP. On average, the difference is a few kcal mol^{-1}, though there are a minority of reactions in which the errors are larger. Further exploration as to why this DFT functional gives such different values could be an area of future research.

Table 3.4. Errors in kcal mol^{-1} for each level of theory relative to CCSD(T)-F12a/cc-pVDZ-F12//ωB97X-D3/def2-TZVP.

	Barrier Height		**Reaction Enthalpy**	
Level of Theory	MAE	RMSE	MAE	RMSE
B97-D3/def2-mSVP	7.0	8.5	3.5	4.8
ωB97X-D3/def2-TZVP	3.5	5.0	1.8	2.5

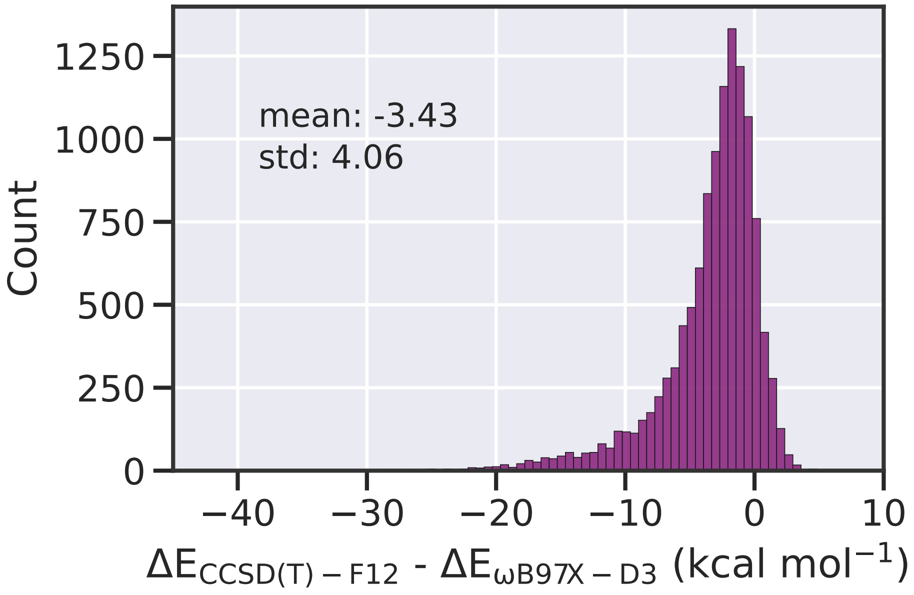


Figure 3.4. Difference in barrier height calculated at CCSD(T)-F12a/cc-pVDZ-F12//ωB97X-D3/def2-TZVP and ωB97X-D3/def2-TZVP.

We next compare some of the coupled cluster values in our dataset to those from other published works. For example, Dontgen et al. [75] reported ten keto-enol reactions calculated at DLPNO-CCSD(T)/CBS//B3LYP-D3BJ/def2-TZVP. Two of their reactions (reactions 1 and 7) are also present in our dataset. For both reactions, the barrier heights agree within 0.3 kcal mol^{-1}. Balabin [76] calculated tautomerization reactions of triazoles at CCSD(T)/CBS//MP2/aug-cc-pVT. For the reaction of 1H-1,2,3-triazole producing 2H-1,2,3-triazole, they report a reaction enthalpy of -3.98 kcal mol^{-1} compared to our calculated value of -3.96. Their reported barrier height for the reverse direction is 53.6 kcal mol^{-1} compared to our value of 49.8 kcal mol^{-1}, calculated by subtracting our reaction enthalpy from our barrier height for the forward direction.

To further evaluate the improvement from the double-ζ single point calculations, we also calculate some reactions at CCSD(T)-F12a/cc-pVTZ-F12//ωB97X-D3/def2-TZVP. The triple-ζ calculations required substantially more computational time and scratch space when compared to the double-ζ calculations. Three reactions are sampled from each of the top five most common RMG reaction families, shown in Figure 3.3. When looking at the distribution of barrier heights within each family for the double-ζ basis set, the three reactions are chosen to represent approximately the 25th, 50th, and 75th percentile. Table 3.5 and Table 3.6 show the MAE and RMSE of the barrier height and reaction enthalpy respectively from the different levels of theory with respect to the triple-ζ basis set. These trends are consistent with other previous literature that emphasizes the high fidelity of explicitly correlated coupled-cluster calculations, even with a double-ζ basis set [55, 60].

Table 3.5. Barrier height errors $\left(\mathrm{kcal\ mol^{-1}}\right)$ for each level of theory relative to CCSD(T)-F12a/cc-pVTZ-F12 for sample reactions.

	B97-D3/def2-mSVP		**ωB97X-D3/def2-TZVP**		**CCSD(T)-F12a/cc-pVDZ-F12**	
Reaction Family	MAE	RMSE	MAE	RMSE	MAE	RMSE
1,3 Insertion ROR	7.6	7.6	0.7	0.8	0.06	0.08
2+2 Cycloaddition	9.0	9.0	2.5	2.5	0.06	0.09
Keto-Enol	4.8	5.7	0.9	1.4	0.05	0.05
Retroene	14.7	15.1	1.4	1.5	0.24	0.24
Singlet Carbene Intra Disproportionation	0.5	0.5	0.3	0.4	0.04	0.05
Overall	7.3	9.0	1.2	1.5	0.09	0.12

Table 3.6. Reaction enthalpy errors $\left(\mathrm{kcal\ mol^{-1}}\right)$ for each level of theory relative to CCSD(T)-F12a/cc-pVTZ-F12 for sample reactions.

	B97-D3/def2-mSVP		**ωB97X-D3/def2-TZVP**		**CCSD(T)-F12a/cc-pVDZ-F12**	
Reaction Family	MAE	RMSE	MAE	RMSE	MAE	RMSE
1,3 Insertion ROR	1.2	1.4	0.7	0.8	0.04	0.04
2+2 Cycloaddition	1.6	1.8	1.5	1.6	0.13	0.14
Keto-Enol	2.7	2.9	0.6	0.7	0.12	0.12
Retroene	6.4	6.6	1.5	1.6	0.16	0.18
Singlet Carbene Intra Disproportionation	5.1	5.3	2.9	3.0	0.11	0.12
Overall	3.4	4.1	1.4	1.8	0.11	0.13

Finally, Grambow et al. [42] already performed a conformer search for the reactants. No additional conformer searching is done in this work. Instead, we identify rigid reactions that do not require a conformer search and report the RRHO TST rate coefficients and fitted Arrhenius parameters for this subset since the explicitly correlated coupled-cluster values should be quite reliable. Figure 3.5 summarizes the filtering workflow to identify rigid reactions. We start by using RDKit's Lipinski rotatable bond SMARTS to find reactions whose reactant and product(s) do not have any rotatable bonds. Further, if any rings are present, we filter molecules with only planar rings (either aromatic or 3-membered). With these criteria, we identify a subset of reactions from the CCSD(T)-F12a/cc-pVDZ-F12 dataset that contain rigid stable species. We next omit any reaction whose TS has a positive frequency smaller than 100 cm^{-1} since this is a common threshold for distinguishing conformational motions from vibrational modes that will be used in the rigid rotor harmonic oscillator approximation [77–80]. As the last filtering step, we visually inspect the remaining reactions to verify that RDKit correctly identified rigid species and also confirm that the TS would not require a conformer search either; this left 105 rigid reactions. Two examples are shown in Figure 3.6.

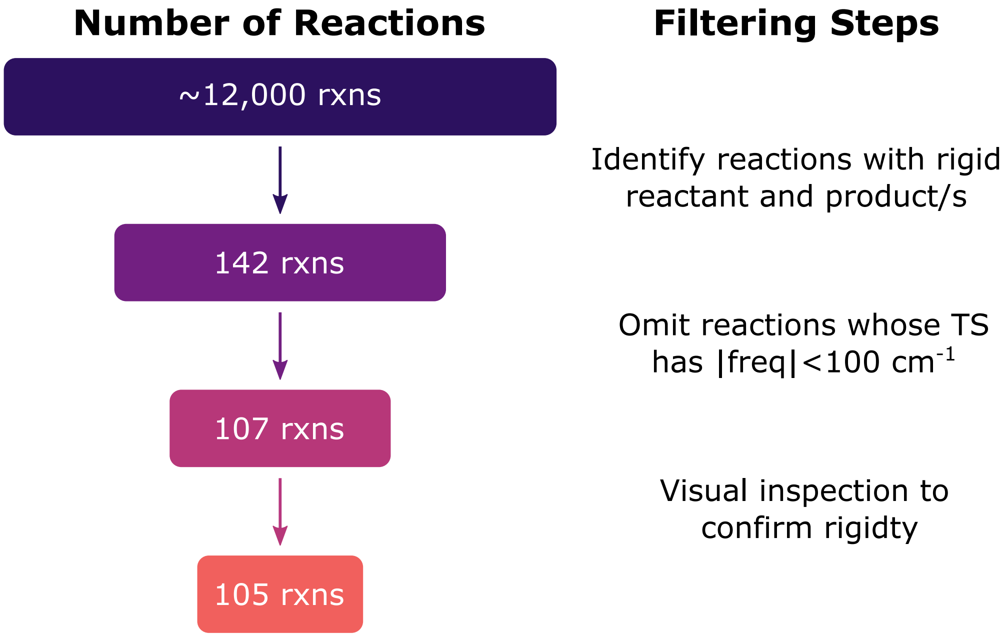

Number of Reactions
Filtering Steps
~12,000 rxns
Identify reactions with rigid reactant and product/s
142 rxns
Omit reactions whose TS has |freq|<100 cm-1
107 rxns
Visual inspection to confirm rigidty
105 rxns

Figure 3.5. Schematic workflow for identifying rigid reactions.

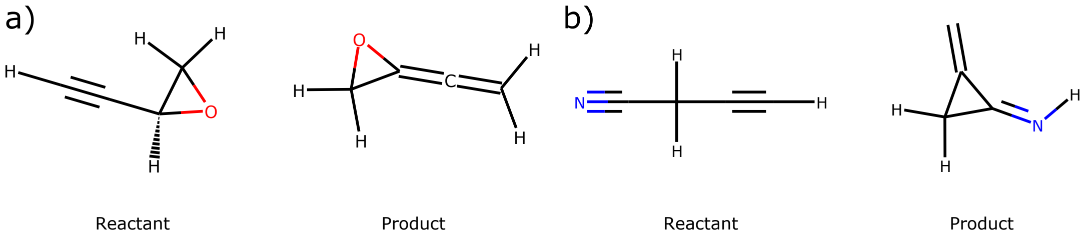


Figure 3.6. Examples of rigid reactions a) rxn001645 b) rxn002603.

To determine whether the fitted parameters give a good estimate of the rate coefficient for these rigid reactions, we examine the average absolute percentage error between the rate coefficient calculated from TST and the predicted rate coefficients using the fitted parameters. The largest value is 233% i.e., about a factor of 3 error in the rate coefficient which is often quite acceptable, though most reactions have a much lower error. For instance, 62 of these rigid reactions have average fitting errors below 20%; thus, the fitted A-factor and activation energy should be very reliable for these reactions. Residuals from the least-squares fit for reactions with the 25th and 75th percentile for average percentage error are shown in Figure 3.7. For nearly all data points, the residuals are very close to zero. Lastly, we searched for published experimental data to compare with our calculated values. Saito et al. [81] studied isomerization of acetonitrile to methyl isocyanide at 1600–2100 K behind reflected shock waves. They report $k_\infty(T) = 10^{13.5}$ exp($-$260 kJ mol^{-1} / RT) s^{-1}, which agrees quite well with the A-factor of 1.14×10^{14} s^{-1} and Ea of 262 kJ mol^{-1} from our fitted Arrhenius expression.

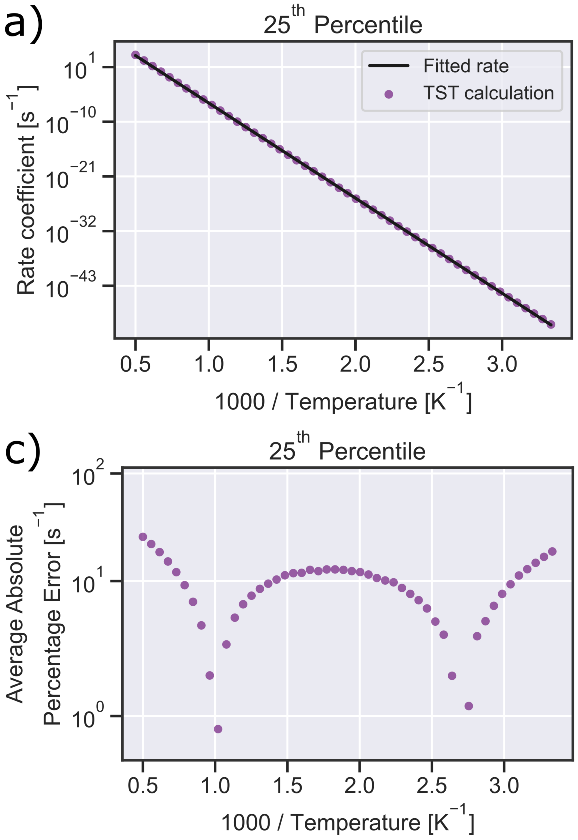


Figure 3.7. Arkane fitting for rigid reactions corresponding to a) the 25th percentile (rxn001645) and b) the 75th percentile (rxn002603) of average absolute percentage error from the CCSD(T)-F12a/cc-pVDZ-F12 dataset. Residuals between the TST rate coefficients and those calculated from the Arrhenius fit are shown in c) and d).

3.5 Usage Notes

Except for the commercial Q-Chem and MOLPRO quantum chemistry softwares, all code necessary to reproduce the generated data is available on GitHub [82]. The repository contains scripts, which should be run in the following order:

- Jupyter notebooks were used to identify potentially erroneous SMILES for the reactants and products of both the B97-D3/def2-mSVP and ωB97X-D3/def2-TZVP level of theory. The suggested SMILES were manually inspected, utilizing the interactive nature of Jupyter notebooks, to confirm that the change was chemically reasonable and preserved the atom-mapping.

- `create_qchem_input_files.py`: Parses the original Q-Chem log files from Grambow et al. [42] to separate product complexes into the individual Q-Chem input files for both the B97-D3/def2-mSVP and ωB97X-D3/def2-TZVP levels of theory.

- `create_molpro_input_files.py`: Creates MOLPRO input files for the single point calculations at CCSD(T)-F12a/cc-pVDZ-F12 using the reoptimized ωB97X-D3/def2-TZVP geometries.

- `parse_barriers_enthalpies.py`: Compiles the reactant and product SMILES into comma-separated values files and parses the reaction barrier heights and enthalpies.

- `get_enthalpies_corrected.py`: Uses the atom and bond corrections from RMG-database to obtain more accurate reaction enthalpies.

- `identify_rmg_reactions.py`: Identifies which reactions correspond to RMG reaction templates.

- `identify_rigid_species.py`: Uses RDKit to identify reactions with rigid reactant and product.

- `run_arkane.py`: Runs Arkane to obtain Arrhenius rate parameters for the rigid reactions.

- `parse_tst_rates.py`: Parses the calculated rate coefficients from the Arkane output files.

- `parse_arrhenius_parameters.py`: Parses the fitted A-factors and activation energies from Arkane output files.

- `calculate_percent_error.py`: Calculates the average percent error between the rate coefficients calculated from TST and those predicted using the fitted Arrhenius parameters.

3.6 Code Availability

The code used to generate this data is freely available on GitHub under the MIT license [82]. Details on how to use the scripts to generate the data are provided in the Usage Notes. Some of the scripts utilize helpful components of the Reaction Mechanism Generator, such as RMG-Py, RMG-database, and the Automatic Rate Calculator (ARC) [83]. All related software is open-source under the MIT license and freely accessible on GitHub. For RMG-Py, checkout the `qchem_parser` branch, and for RMG-database, checkout `AEC_BAC`. The GitHub version commit string was ea2eb625fb1dcc6892ef6ddd5d7fdc96abf477e1 for ARC on the main branch.

3.7 Follow-up Work to Refine the DFT Energies

Approximately 100 reactions (i.e., less than 1% of the data) from Figure 3.4 have ΔE_0 which differs by over 20 kcal mol^{-1} between CCSD(T)-F12a/cc-pVDZ-F12 and ωB97X-D3/def2-TZVP. This difference is much larger than expected. However, the source of these large error was initially unclear as both Spiekermann et al. and Grambow et al. had performed several validation checks, such as ensuring that all TSs have exactly one imaginary frequency, whose atomic displacements matched the bond changes occurring between the reactant and product. Any TS with an imaginary frequency less than 100 cm^{-1} in magnitude was removed as that typically corresponds to conformational changes. Multiplicity and charge are conserved for all reactions. And no species showed spin contamination or any numerical convergence errors. Finally, the updated species enthalpies and reaction barriers showed strong agreement with high-quality experimental data.

Still, the differences are large enough to warrant additional investigation. Thus, in collaboration with the Head-Gordon group at UC Berkeley, follow-up work has recalculated the single-point energy of the TS and reactant for these reactions using updated settings. All updated ωB97X-D3/def2-TZVP single-point energy calculations are done using Q-Chem [48]. The updated settings include the following changes:

1. The Geometric Direct Minimization (GDM) convergence algorithm is used because it is known to be more robust than the Direct Inversion in the Iterative Subspace (DIIS) algorithm used in the original work.

2. A spin-polarized initial guess is used by adding 10% of the lowest unoccupied molecular orbital (LUMO) to the highest occupied molecular orbital (HOMO) to break alpha/beta symmetry. This step is crucial to obtaining more accurate DFT energies.

The results of recalculating these 100 reactions is shown in Figure 3.8. The original DFT values, which were calculated by Grambow et al. [42], differed substantially from the CCSD(T)-F12a/cc-pVDZ-F12 values calculated by Spiekermann et al. [84]. However, the updated DFT values show much smaller deviations. Further investigation into these reactions is on-going.

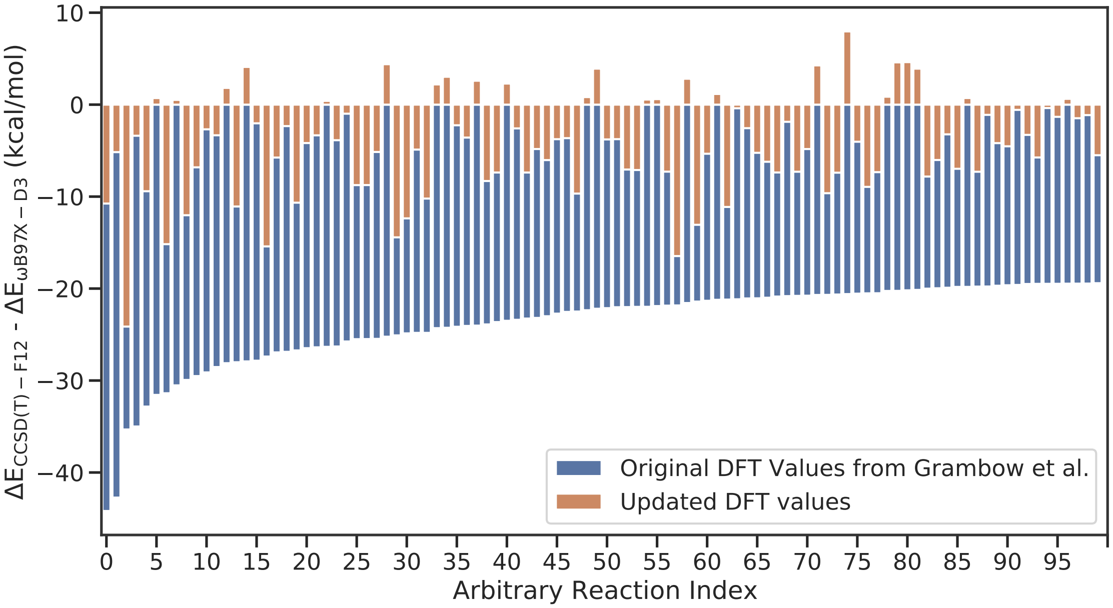


Figure 3.8. The difference between the barrier heights calculated at CCSD(T)-F12a/cc-pVDZ-F12//ωB97X-D3/def2-TZVP relative to the original barrier heights calculated at ωB97X-D3/def2-TZVP by Grambow et al. [42] is shown in blue. The difference between the barrier heights calculated at CCSD(T)-F12a/cc-pVDZ-F12//ωB97X-D3/def2-TZVP relative to the new barrier heights calculated at ωB97X-D3/def2-TZVP in a collaboration with the Head-Gordon group is shown in orange. Only 100 reactions with the biggest differences from Figure 3.4 are shown here.

Recently, I was able to recalculate the ωB97X-D3/def2-TZVP single-point energies using the updated settings for all reactants, TSs, and products of all reactions. As some aggregate statistics for ΔE_0, the MAE between the original DFT barriers and the updated DFT barriers is just 0.83 kcal mol^{-1}. The RMSE is 3.08 kcal mol^{-1} due to a few hundred outliers with massive differences. 88% of the reactions have a new DFT barrier that is within 1 kcal mol^{-1} of the original DFT barrier i.e., these ∼10,500 reactions were already estimated quite well so many compute hours were spent recalculating the single-point energies for only a small improvement. 94% of the reactions have a new DFT barrier that is within 5 kcal mol^{-1} of the original DFT barrier Approximately 10,600 reactions have an updated DFT barrier that is within 6 kcal mol^{-1} of the CCSD(T)-F12 value.

Still, it is important to fix the values for the outlier reactions since the initial values were so inaccurate that they impacted the overall statistics. Comparing Figure 3.9 below with Figure 3.4 shows that the MAE between the DFT barrier heights and those calculated using the explicitly correlated coupled cluster is lower by approximately 0.8 kcal mol^{-1}. The RMSE is lower by approximately 1.6 kcal mol^{-1}.

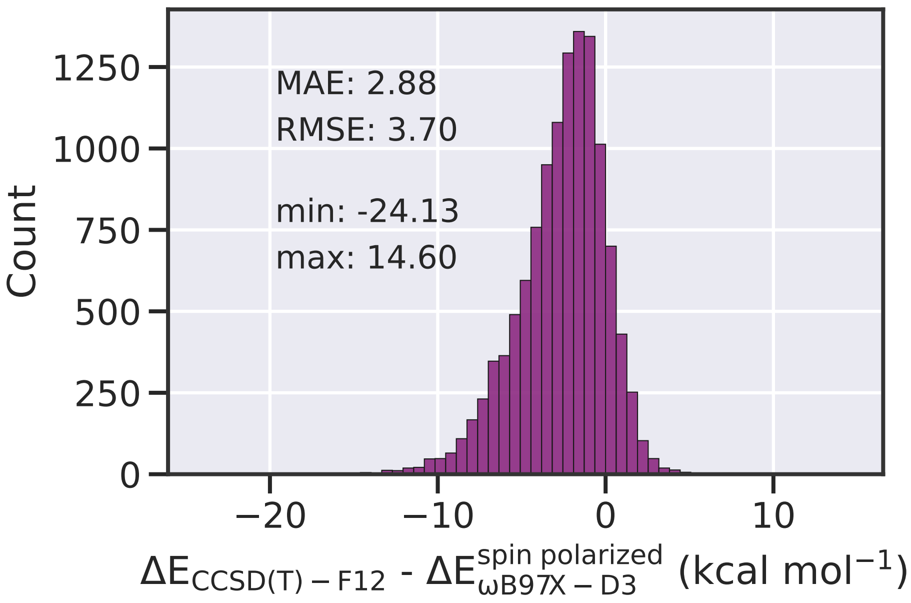


Figure 3.9. Difference in the barrier height calculated at CCSD(T)-F12a/cc-pVDZ-F12//ωB97X-D3/def2-TZVP and ωB97X-D3/def2-TZVP using the updated DFT settings.

Similar results are obtained when analyzing aggregate statistics for the reaction energy. Here, the values do not include the BAC. The MAE between the original DFT enthalpy and the updated DFT enthalpy is just 0.14 kcal mol^{-1}. The RMSE is just 0.91 kcal mol^{-1}. 96% of the reactions have a new DFT enthalpy that is within 1 kcal mol^{-1} of the original DFT enthalpy. The number of outliers for reaction enthalpy was far less than those for the barriers. i.e., there were more TSs that had issues than reactants or products. Comparing Figure 3.10 with Figure 3.11 shows that the MAE between the DFT reaction energies and those calculated using the explicitly correlated coupled cluster is lower by only 0.1 kcal mol^{-1}. The RMSE is lower by only 0.2 kcal mol^{-1}.

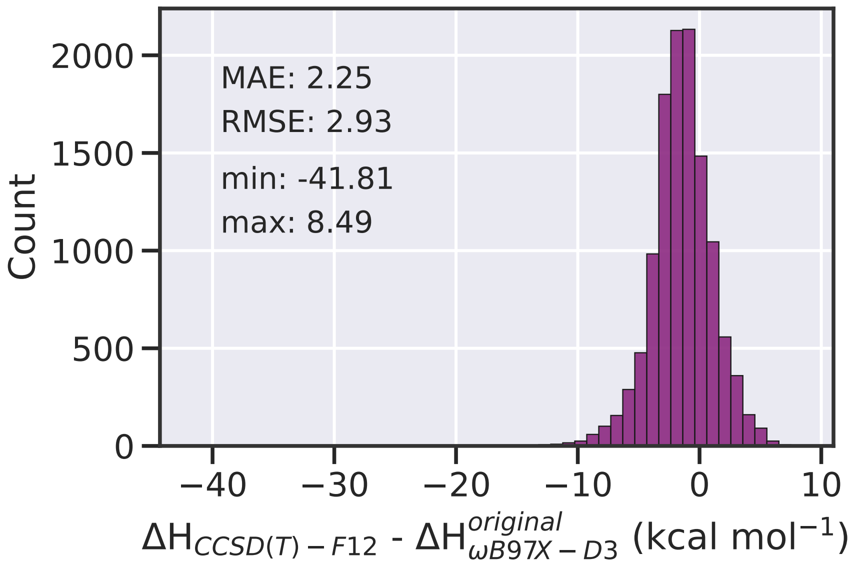


Figure 3.10. Difference in the reaction energy calculated at CCSD(T)-F12a/cc-pVDZ-F12//ωB97X-D3/def2-TZVP and ωB97X-D3/def2-TZVP using the original DFT settings.

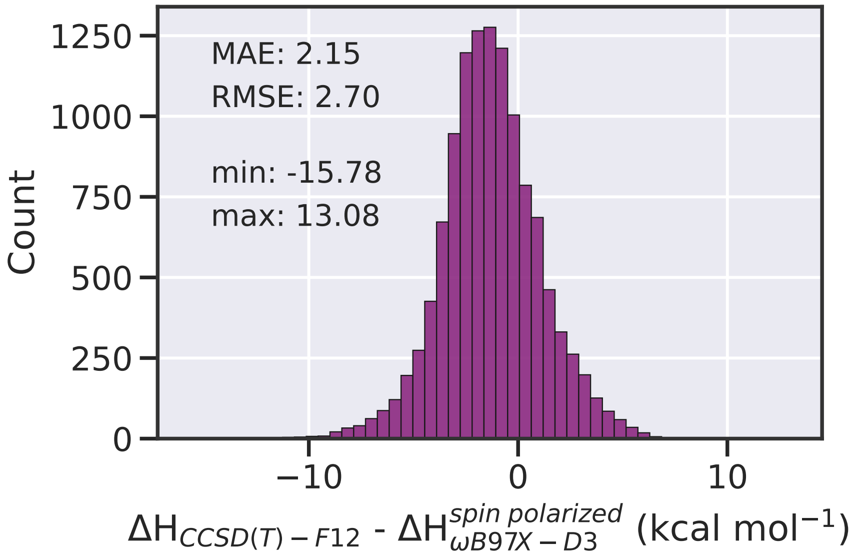


Figure 3.11. Difference in the reaction energy calculated at CCSD(T)-F12a/cc-pVDZ-F12//ωB97X-D3/def2-TZVP and ωB97X-D3/def2-TZVP using the updated DFT settings.

In conclusion, caution should be exercised with the subset of the data that initially showed large discrepancies between the barrier heights calculated at ωB97X-D3/def2-TZVP vs. those calculated at CCSD(T)-F12a/cc-pVDZ-F12. The DFT energies for this specific subset are likely not reliable. However, the explicitly correlated coupled cluster values should be very accurate.

3.8 References

(1) Green, W. H. Moving from Postdictive to Predictive Kinetics in Reaction Engineering. *AIChE Journal* **2020**, *66*, e17059.

(2) Wang, K.; Dean, A. M., Rate Rules and Reaction Classes In *Comput. Aided Chem. Eng.* Elsevier: 2019; Vol. 45, pp 203–257.

(3) Class, Caleb A. and Vasiliou, AnGayle K. and Kida, Yuko and Timko, Michael T. and Green, William H. Detailed Kinetic Model for Hexyl Sulfide Pyrolysis and its Desulfurization by Supercritical Water. *Phys. Chem. Chem. Phys.* **2019**, *21*, 10311–10324.

(4) Khanniche, S.; Lai, L.; Green, W. H. Kinetics of Intramolecular Phenyl Migration and Fused Ring Formation in Hexylbenzene Radicals. *J. Phys. Chem. A* **2018**, *122*, 9778–9791.

(5) Payne, A. M.*; Spiekermann, K. A.*; Green, W. H. Detailed Reaction Mechanism for 350–400°C Pyrolysis of an Alkane, Aromatic, and Long-Chain Alkylaromatic Mixture. *Energy & Fuels* **2022**, *36*, 1635–1646.

(6) Zheng, J.; Zhao, Y.; Truhlar, D. G. The DBH24/08 Database and its Use to Assess Electronic Structure Model Chemistries for Chemical Reaction Barrier Heights. *J. Chem. Theory Comput.* **2009**, *5*, 808–821.

(7) Krasnoukhov, V. S.; Zagidullin, M. V.; Zavershinskiy, I. P.; Mebel, A. M. Formation of Phenanthrene via Recombination of Indenyl and Cyclopentadienyl Radicals: A Theoretical Study. *J. Phys. Chem. A* **2020**, *124*, 9933–9941.

(8) Grinberg Dana, A.; Moore III, K. B.; Jasper, A. W.; Green, W. H. Large Intermediates in Hydrazine Decomposition: A Theoretical Study of the N3H5 and N4H6 Potential Energy Surfaces. *J. Phys. Chem. A* **2019**, *123*, 4679–4692.

(9) Keçeli, M.; Elliott, S. N.; Li, Y.-P.; Johnson, M. S.; Cavallotti, C.; Georgievskii, Y.; Green, W. H.; Pelucchi, M.; Wozniak, J. M.; Jasper, A. W., et al. Automated Computational Thermochemistry for Butane Oxidation: A Prelude to Predictive Automated Combustion Kinetics. *Proc. Combust. Inst.* **2019**, *37*, 363–371.

(10) Gillis, R. J.; Green, W. H. Thermochemistry Prediction and Automatic Reaction Mechanism Generation for Oxygenated Sulfur Systems: A Case Study of Dimethyl Sulfide Oxidation. *ChemSystemsChem* **2020**, *2*, e1900051.

(11) Johnson, M. S.; Nimlos, M. R.; Ninnemann, E.; Laich, A.; Fioroni, G. M.; Kang, D.; Bu, L.; Ranasinghe, D.; Khanniche, S.; Goldsborough, S. S., et al. Oxidation and Pyrolysis of Methyl Propyl Ether. *Int. J. Chem. Kinet.* **2021**, *53*, 915–938.

(12) Dong, X.; Ninnemann, E.; Ranasinghe, D. S.; Laich, A.; Greene, R.; Vasu, S. S.; Green, W. H. Revealing the Critical Role of Radical-Involved Pathways in High Temperature Cyclopentanone Pyrolysis. *Combust. Flame* **2020**, *216*, 280–292.

(13) Pio, G.; Dong, X.; Salzano, E.; Green, W. H. Automatically Generated Model for Light Alkene Combustion. *Combust. Flame* **2022**, *241*, 112080.

(14) Gao, C. W.; Allen, J. W.; Green, W. H.; West, R. H. Reaction Mechanism Generator: Automatic Construction of Chemical Kinetic Mechanisms. *Comput. Phys. Commun.* **2016**, *203*, 212–225.

(15) Liu, M.; Grinberg Dana, A.; Johnson, M. S.; Goldman, M. J.; Jocher, A.; Payne, A. M.; Grambow, C. A.; Han, K.; Yee, N. W.; Mazeau, E. J., et al. Reaction Mechanism Generator v3.0: Advances in Automatic Mechanism Generation. *J. Chem. Inf. Model.* **2021**, *61*, 2686–2696.

(16) Lai, L.; Khanniche, S.; Green, W. H. Thermochemistry and Group Additivity Values for Fused Two-Ring Species and Radicals. *J. Phys. Chem. A* **2019**, *123*, 3418–3428.

(17) Lai, L.; Pang, H.-W.; Green, W. H. Formation of Two-Ring Aromatics in Hexylbenzene Pyrolysis. *Energy & Fuels* **2020**, *34*, 1365–1377.

(18) Benson, S. W. Thermochemistry and Kinetics of Sulfur-Containing Molecules and Radicals. *Chem. Rev.* **1978**, *78*, 23–35.

(19) Carstensen, H.-H.; Dean, A. M. Rate Constant Rules for the Automated Generation of Gas-Phase Reaction Mechanisms. *J. Phys. Chem. A* **2009**, *113*, 367–380.

(20) Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. Predicting Retrosynthetic Pathways using Transformer-Based Models and a Hyper-Graph Exploration Strategy. *Chem. Sci.* **2020**, *11*, 3316–3325.

(21) Mo, Y.; Guan, Y.; Verma, P.; Guo, J.; Fortunato, M. E.; Lu, Z.; Coley, C. W.; Jensen, K. F. Evaluating and Clustering Retrosynthesis Pathways with Learned Strategy. *Chem. Sci.* **2021**, *12*, 1469–1478.

(22) Bradshaw, J.; Paige, B.; Kusner, M. J.; Segler, M. H. S.; Hernández-Lobato, J. M. Barking up the Right Tree: An Approach to Search Over Molecule Synthesis DAGs. *Adv. Neural Inf. Proc. Sys.* **2020**, *33*, 6852–6866.

(23) Coley, C. W. Defining and Exploring Chemical Spaces. *Trends in Chem.* **2021**, *3*, 133–145.

(24) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M., et al. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388.

(25) Stokes, J. M.; Yang, K.; Swanson, K.; Jin, W.; Cubillos-Ruiz, A.; Donghia, N. M.; MacNair, C. R.; French, S.; Carfrae, L. A.; Bloom-Ackermann, Z., et al. A Deep Learning Approach to Antibiotic Discovery. *Cell* **2020**, *180*, 688–702.

(26) Li, Y.-P.; Han, K.; Grambow, C. A.; Green, W. H. Self-Evolving Machine: A Continuously Improving Model for Molecular Thermochemistry. *J. Phys. Chem. A* **2019**, *123*, 2142–2152.

(27) Grambow, C. A.; Li, Y.-P.; Green, W. H. Accurate Thermochemistry with Small Data Sets: A Bond Additivity Correction and Transfer Learning Approach. *J. Phys. Chem. A* **2019**, *123*, 5826–5835.

(28) Vermeire, F. H.; Green, W. H. Transfer learning for solvation free energies: From quantum chemistry to experiments. *Chem. Eng. J.* **2021**, *418*, 129307.

(29) Chung, Y.; Vermeire, F. H.; Wu, H.; Walker, P. J.; Abraham, M. H.; Green, W. H. Group Contribution and Machine Learning Approaches to Predict Abraham Solute Parameters, Solvation Free Energy, and Solvation Enthalpy. *J. Chem. Inf. Model.* **2022**, *62*, 433–446.

(30) McGill, C.; Forsuelo, M.; Guan, Y.; Green, W. H. Predicting Infrared Spectra with Message Passing Neural Networks. *J. Chem. Inf. Model.* **2021**, *61*, 2594–2609.

(31) Coley, C. W.; Thomas, D. A.; Lummiss, J. A. M.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H., et al. A Robotic Platform for Flow Synthesis of Organic Compounds Informed by AI Planning. *Science* **2019**, *365*.

(32) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting Reaction Performance in C–N Cross-Coupling using Machine Learning. *Science* **2018**, *360*, 186–190.

(33) Grambow, C. A.; Pattanaik, L.; Green, W. H. Deep Learning of Activation Energies. *J. Phys. Chem. Lett.* **2020**, *11*, 2992–2997.

(34) Pattanaik, L.; Ingraham, J. B.; Grambow, C. A.; Green, W. H. Generating Transition States of Isomerization Reactions with Deep Learning. *Phys. Chem. Chem. Phys.* **2020**, *22*, 23618–23626.

(35) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 kilo Molecules. *Sci. Data* **2014**, *1*, 1–7.

(36) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1, A Data Set of 20 Million Calculated Off-Equilibrium Conformations for Organic Molecules. *Sci. Data* **2017**, *4*, 1–8.

(37) Kipf, T. N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv* **2016**, *1609.02907*.

(38) Pattanaik, L.; Ganea, O. E.; Coley, I.; Jensen, K. F.; Green, W. H.; Coley, C. W. Message Passing Networks for Molecules with Tetrahedral Chirality. *arXiv* **2020**, *2012.00094*.

(39) Mercado, R.; Rastemo, T.; Lindelöf, E.; Klambauer, G.; Engkvist, O.; Chen, H.; Bjerrum, E. J. Graph Networks for Molecular Design. *Mach. Learn.: Sci. Technol.* **2021**, *2*, 025023.

(40) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet–A Deep Learning Architecture for Molecules and Materials. *J. Chem. Phys.* **2018**, *148*, 241722.

(41) Klicpera, J.; Groß, J.; Günnemann, S. Directional Message Passing for Molecular Graphs. *arXiv* **2020**, *2003.03123*.

(42) Grambow, C. A.; Pattanaik, L.; Green, W. H. Reactants, Products, and Transition States of Elementary Chemical Reactions Based on Quantum Chemistry. *Sci. Data* **2020**, *7*, 1–8.

(43) Grambow, C. A.; Pattanaik, L.; Green, W. H. Reactants, Products, and Transition States of Elementary Chemical Reactions based on Quantum Chemistry, version 1.0.1, 2020.

(44) Zimmerman, P. M. Single-Ended Transition State Finding with the Growing String Method. *J. Comput. Chem.* **2015**, *36*, 601–611.

(45) 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. *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875.

(46) Landrum, G. et al. RDKit: Open-Source Cheminformatics Software, https://www.rdkit.org, 2006.

(47) Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know to Improve Conformation Generation. *J. Chem. Inf. Model.* **2015**, *55*, 2562–2574.

(48) Shao, Y.; Gan, Z.; Epifanovsky, E.; Gilbert, A. T. B.; Wormit, M.; Kussmann, J.; Lange, A. W.; Behn, A.; Deng, J.; Feng, X., et al. Advances in Molecular Quantum Chemistry Contained in the Q-Chem 4 Program Package. *Mol. Phys.* **2015**, *113*, 184–215.

(49) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011**, *32*, 1456–1465.

(50) Lin, Y.-S.; Li, G.-D.; Mao, S.-P.; Chai, J.-D. Long-range Corrected Hybrid Density Functionals with Improved Dispersion Corrections. *J. Chem. Theory Comput.* **2013**, *9*, 263–272.

(51) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31–36.

(52) Bischoff, F. A.; Wolfsegger, S.; Tew, D. P.; Klopper, W. Assessment of Basis Sets for F12 Explicitly-Correlated Molecular Electronic-Structure Methods. *Mol. Phys.* **2009**, *107*, 963–975.

(53) Knizia, G.; Adler, T. B.; Werner, H.-J. Simplified CCSD(T)-F12 Methods: Theory and Benchmarks. *J. Chem. Phys.* **2009**, *130*, 054104.

(54) Adler, T. B.; Knizia, G.; Werner, H.-J. A Simple and Efficient CCSD(T)-F12 Approximation. *J. Chem. Phys.* **2007**, *127*, 221106.

(55) Pfeiffer, F.; Rauhut, G.; Feller, D.; Peterson, K. A. Anharmonic Zero Point Vibrational Energies: Tipping the Scales in Accurate Thermochemistry Calculations? *J. Chem. Phys.* **2013**, *138*, 044311.

(56) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminf.* **2011**, *3*, 1–14.

(57) Mardirossian, N.; Head-Gordon, M. Thirty Years of Density Functional Theory in Computational Chemistry: An Overview and Extensive Assessment of 200 Density Functionals. *Mol. Phys.* **2017**, *115*, 2315–2372.

(58) Tajti, A.; Szalay, P. G.; Császár, A. G.; Kállay, M.; Gauss, J.; Valeev, E. F.; Flowers, B. A.; Vázquez, J.; Stanton, J. F. HEAT: High Accuracy Extrapolated Ab initio Thermochemistry. *J. Chem. Phys.* **2004**, *121*, 11599–11613.

(59) Karton, A.; Rabinovich, E.; Martin, J. M. L.; Ruscic, B. W4 Theory for Computational Thermochemistry: In Pursuit of Confident Sub-kJ/mol Predictions. *J. Chem. Phys.* **2006**, *125*, 144108.

(60) Shang, Y.; Ning, H.; Shi, J.; Wang, H.; Luo, S.-N. Chemical Kinetics of H-abstractions from Dimethyl Amine by H, CH_3, OH, and HO_2 Radicals with Multi-Structural Torsional Anharmonicity. *Phys. Chem. Chem. Phys.* **2019**, *21*, 12685–12696.

(61) Feller, D.; Peterson, K. A.; Dixon, D. A. A Survey of Factors Contributing to Accurate Theoretical Predictions of Atomization Energies and Molecular Structures. *J. Chem. Phys.* **2008**, *129*, 204105.

(62) Narayanan, B.; Redfern, P. C.; Assary, R. S.; Curtiss, L. A. Accurate Quantum Chemical Energies for 133000 Organic Molecules. *Chem. Sci.* **2019**, *10*, 7449–7455.

(63) Kim, H.; Park, J. Y.; Choi, S. Energy Refinement and Analysis of Structures in the QM9 Database via a Highly Accurate Quantum Chemical Method. *Sci. Data* **2019**, *6*, 1–8.

(64) Feller, D.; Peterson, K. A.; Hill, J. G. Calibration Study of the CCSD(T)-F12a/b Methods for C2 and Small Hydrocarbons. *J. Chem. Phys.* **2010**, *133*, 184102.

(65) Werner, H. J.; Knowles, P. J.; Knizia, G.; Manby, F.; Schütz, M.; Celani, P.; Györffy, W.; Kats, D.; Korona, T.; Lindh, R., et al. MOLPRO, version 2015.1, A Package of Ab initio Programs, https://www.molpro.net, 2015.

(66) Alecu, I. M.; Zheng, J.; Zhao, Y.; Truhlar, D. G. Computational Thermochemistry: Scale Factor Databases and Scale Factors for Vibrational Frequencies Obtained from Electronic Model Chemistries. *J. Chem. Theory Comput.* **2010**, *6*, 2872–2887.

(67) Johnson III, R. D. NIST Computational Chemistry Comparison and Benchmark Database, NIST Standard Reference Database Number 101. *http://cccbdb.nist.gov/* **2020**.

(68) Petersson, G. A.; Malick, D. K.; Wilson, W. G.; Ochterski, J. W.; Montgomery Jr., J. A.; Frisch, M. J. Calibration and Comparison of the Gaussian-2, Complete Basis Set, and Density Functional Methods for Computational Thermochemistry. *J. Chem. Phys.* **1998**, *109*, 10570–10579.

(69) Ruscic, B.; Bross, D. H. Active Thermochemical Tables (ATcT) Values Based on ver. 1.122d of the Thermochemical, https://atct.anl.gov/Thermochemical%20Data/version%201.122d/index.php, 2018.

(70) Spiekermann, K. A.; Pattanaik, L.; Green, W. H. High Accuracy Barrier Heights, Enthalpies, and Rate Coefficients for Chemical Reactions version 1.0.1, version 1.0.1, 2022.

(71) ANSYS Inc. Chemkin-Pro, http://www.ansys.com/products/fluids/ansys-chemkin-pro, San Diego, CA, 2017.

(72) Pedley, J. B., *Thermochemical Data and Structures of Organic Compounds*; CRC Press: 1994; Vol. 1.

(73) Ruscic, B. Uncertainty Quantification in Thermochemistry, Benchmarking Electronic Structure Computations, and Active Thermochemical Tables. *Int. J. Quantum Chem.* **2014**, *114*, 1097–1101.

(74) Ruscic, B.; Bross, D. H., Thermochemistry In *Comput. Aided Chem. Eng.* Elsevier: 2019; Vol. 45, pp 3–114.

(75) Døntgen, M.; Fenard, Y.; Heufer, K. A. Atomic Partial Charges as Descriptors for Barrier Heights. *J. Chem. Inf. Model.* **2020**, *60*, 5928–5931.

(76) Balabin, R. M. Tautomeric Equilibrium and Hydrogen Shifts in Tetrazole and triazoles: Focal-Point Analysis and Ab initio Limit. *J. Chem. Phys.* **2009**, *131*, 154307.

(77) Pratt, L. M.; Truhlar, D. G.; Cramer, C. J.; Kass, S. R.; Thompson, J. D.; Xidos, J. D. Aggregation of Alkyllithiums in Tetrahydrofuran. *J. Org. Chem.* **2007**, *72*, 2962–2966.

(78) Zhao, Y.; Truhlar, D. G. Computational Characterization and Modeling of Buckyball Tweezers: Density Functional Study of Concave–Convex $\pi \cdots \pi$ Interactions. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2813–2818.

(79) Ribeiro, R. F.; Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Use of Solution-Phase Vibrational Frequencies in Continuum Models for the Free Energy of Solvation. *J. Phys. Chem. B* **2011**, *115*, 14556–14562.

(80) Bao, J. L.; Xing, L.; Truhlar, D. G. Dual-Level Method for Estimating Multistructural Partition Functions with Torsional Anharmonicity. *J. Chem. Theory Comput.* **2017**, *13*, 2511–2522.

(81) Saito, K.; Kakumoto, T.; Murakami, I. A Study of the Isomerization of Acetonitrile at High Temperatures. *Chem. Phys. Lett.* **1984**, *110*, 478–481.

(82) Spiekermann, K. A.; Pattanaik, L. reactants_products_ts_refined release version 1.0.0, version 1.0.0, 2022.

(83) Grinberg Dana, A.; Ranasinghe, D.; Wu, H.; Grambow, C.; Dong, X.; Johnson, M.; Goldman, M.; Liu, M.; Green, W. ARC - Automated Rate Calculator, version 1.1.0, 2019.

(84) Spiekermann, K. A.; Pattanaik, L.; Green, W. H. High Accuracy Barrier Heights, Enthalpies, and Rate Coefficients for Chemical Reactions. *Sci. Data* **2022**, *9*, 417.

Chapter 4

Improvements to RMG database

This work is adapted from Johnson, M. S. et al. RMG Database for Chemical Property Prediction. *J. Chem. Inf. Model.* **2022**, *62*, 4906–4915 [link]. While the publication offers a brief summary of all features related to our group's Reaction Mechanism Generator (RMG) Database, such as thermodynamic group additivity as well as providing estimates for rate coefficients, liquid-phase diffusivity, and Lennard-Jones parameters, this chapter provides a detailed description of the original contributions put forth by this thesis. All results have been integrated into RMG database and are publicly accessible at https://github.com/ReactionMechanismGenerator/RMG-database.

4.1 Introduction

Detailed kinetic mechanisms allow improved understanding of the time evolution of many important chemical processes. In many cases, to accurately represent the chemistry involved, such mechanisms must include hundreds to thousands of species and tens of thousands of reactions. The process for constructing a kinetic mechanism requires a rate coefficient for each of these reactions. The rate for the reverse reaction must also be determined, either by explicitly calculating a rate coefficient for the reverse reaction or by computing the equilibrium constant of the reaction. For larger models, the latter is typically preferred: it guarantees thermodynamic consistency, thermochemistry is easier to estimate than kinetics, and there are typically far fewer species in a model than reactions. Several databases exist to query these values. One popular example is the National Institute of Standards and Technology (NIST) [1], which archives experimental and quantum chemical rate coefficients and thermodynamic properties. The NIST databases are very extensive, although the data quality can be highly variable. While this makes NIST a great resource to search for parameters, it also makes it not ideal for applications where there is no human in the loop. The active thermochemical tables (ATcT) database provides high-accuracy values for a limited number of species [2]. Many databases such as PriMe [3], ChemKED [4, 5], CloudFlame [6, 7] and ReSpecTh [8] exist for storing experimental measurements that may be relevant. The MolSSI QC Archive project is a quantum chemistry specific database intended to help supply homogeneous data sets [9] while CCCBDB [10] has both experimental and computed thermochemical data.

There are various ways of estimating thermochemical and kinetic parameters. By far the most common method of estimating thermochemical parameters is the Benson type group additivity method [11, 12], which is a linear model for adding enthalpies of formation of predefined groups to obtain the enthalpy of the molecule. Implementations of this method are used in THERM and Genesys [13, 14]. Graph convolutional neural network methods have been emerging as a popular alternative since they are template-free and do not require manual group definitions [15–17]. Kinetics estimators typically rely on either use of rate rules to assign specific rates to reactions matching specific templates or on reaction group additivity [18, 19]. Recent machine learning approaches using hierarchical decision trees have also been found to be effective for predicting rate coefficients [20].

The Reaction Mechanism Generator (RMG) database [21] was primarily created to supply the RMG software [22, 23] with good estimates for thermochemical and kinetic parameters. It provides kineticists with easy access to estimates of the numerous parameters needed to model and analyze kinetic systems. The database aggregates several curated thermochemistry and kinetic datasets that have been published in the literature. When possible, these values will be directly queried from the RMG database and used during reaction mechanism generation. However, more commonly, many species or reactions of interest are not directly stored in the database so instead the RMG software relies on estimators to quickly make predictions; these estimators are trained on data from the RMG database. These simple data-driven estimators are convenient since they can quickly make predictions for the tens of thousands (or even hundreds of thousands) of reactions that must commonly be considered during automatic mechanism generation. However, the accuracy of these estimations is crucial because unlike in other mechanism construction methods, the decisions of whether or not species and reactions are included in the resulting mechanism are directly based on their estimated properties. Poor thermochemistry and kinetic estimation in RMG can therefore cause model generation to miss the real chemical pathways in a kinetic system, highlighting the important role the RMG database has in providing reasonable data estimation schemes.

RMG has been a seminal contribution to the field of chemical kinetics. Still, one of the main limitations for RMG is that many thermodynamic and kinetic parameters are poorly estimated, which can cause RMG to sometimes generate models whose predicted concentration profiles differ substantially from experimental data. Although the values in RMG database are often of high-quality, most parameters in an RMG mechanism are estimated and are often of lower-quality, particularly the kinetic parameters estimated from the decision trees. The first reason for these poor estimates is that many of the decision trees are trained on a very limited dataset of ten or fewer reactions. As shown in Table 4.1, several reaction families have fewer than 25 reactions to train the decision tree estimators, which are then used to make predictions for hundreds to thousands of reactions during mechanism generation. These estimators are likely extrapolating, so the quality of the rate estimates is likely lower than we would prefer. Since the Arrhenius expression is sensitive to small changes in the activation energy due to the exponential, especially at lower temperatures, these errors can be substantial and cause RMG to explore unimportant reaction pathways and possibly ignore relevant chemistry. The second contributing factor is that although a few reaction families have several hundred or even up to 3,000 training reactions to train the decision tree estimator as shown in Table 4.2, more than 70% of the values for these training reactions come from a linear group additivity estimate rather than a reliable experimental value or from a high-quality quantum chemistry calculation used with transition state theory (TST), either of which would have ideally gone through peer-review as well. In summary, there is a pressing need to substantially improve the quality of the training data and estimators within the RMG database.

Table 4.1. Number of training reactions for select RMG families.

RMG Reaction Family	Num. Training Reactions
1+2 Cycloaddition	12
1,2 Insertion CO	7
1,2 Insertion Carbene	12
1,2 NH3 Elimination	4
1,3 Insertion CO_2	8
1,3 Insertion ROR	22
1,3 NH3 Elimination	4
1,3 Sigmatropic Rearrangement	10
2+2 Cycloaddition	6
6 Membered Central C−C Shift	7
Birad R Recombination	7
Birad Recombination	8
CO Disproportionation	13
Diels Alder Addition	23
Intra 2+2 Cycloaddition Cd	2
Intra Disproportionation	2
Intra Diels Alder Monocyclic	2
Intra-ene Reaction	21
Singlet Carbene Intra Disproportionation	4
Intra OH Migration	8

Table 4.2. Percentage of training reactions for select RMG families whose rate coefficients come from group additivity values (GAV).

RMG Reaction Family	Num. Training Reactions	% from GAV
H Abstraction	3,116	77.4%
R Addition MultipleBond	2,952	85.4%
Intra R Add Endocyclic	851	79.1%
Intra R Add Exocyclic	374	74.8%

Although software tools such as the Automated Rate Calculator (ARC) [24] and The Tandem Tool (T3) [25] can be very helpful for refining and improving the parameters in these kinetic mechanisms, it is also crucial to address these parameter estimation errors at the source by substantially improving the thermodynamic and kinetic values that RMG uses during mechanism generation in the first place. As shown in Figure 3.3, about 1,000 reactions match the RMG templates shown in Table 3.3, many of which also overlap with the sparse families shown in Table 4.1. Here, I address these parameter errors by incorporating many of these reactions and corresponding species thermochemistry into RMG database. Considering that RMG database has around 8,000 training reactions in total and about 70% of these values are from group additivity, adding high-quality calculations at the CCSD(T)-F12a/cc-pVDZ-F12//ωB97X-D3/def2-TZVP level of theory and subsequently refitting the decision tree estimators should dramatically improve RMG's rate estimates and their extrapolation capabilities to new reactions. The new thermodynamic library and updated decision trees contributed through this work should improve RMG's reliability and substantially improve the automated generation of kinetic mechanisms for all future users.

4.2 Methods

Here, I describe the procedure for computing various corrections that are necessary to improve the estimate of species thermochemistry, namely the frequency scaling factor, atom energy corrections, and bond additivity corrections. This accurate thermochemistry information for all reactant and product species is certainly useful by itself, but it is also used to accurately estimate the rate coefficient for the reverse reaction to guarantee thermodynamic consistency. Finally, this section describes the conformer search methodology as well as the software utilized to perform statistical mechanics calculations to determine thermochemistry and high-pressure limit rate coefficients.

4.2.1 Frequency Scaling Factor

The motivation for using a frequency scale factor is to reduce errors in the calculated harmonic vibrational frequencies, fundamental frequencies, zero-point energies (ZPEs), and vibrational partition functions, which are ultimately used to calculate species thermochemistry. Although calculating anharmonicity corrections is another plausible approach to improve accuracy, this requires higher-order force constants and information about torsional barriers or full potential energy surfaces, which are often unavailable (linear species cannot be treated with perturbative approaches) or computationally intractable for larger species. Instead, Alecu et al. [26] demonstrate that using a simple scaling factor is substantially more accessible while still yielding accurate results. In this chapter, the frequency scaling factor is calculated for B97-D3/def2-mSVP and ωB97X-D3/def2-TZVP levels of theory (LoT) using the method described by Alecu et al. [26].

Briefly, the optimal scale factor for the ZPE (λ^{ZPE}) is obtained by minimizing the root-mean-square deviation (RMSD) between a set of ZPEs computed from harmonic vibrational frequencies at a given level of theory ($\varepsilon_{\mathrm{vib}_m}^{\mathrm{G,input}}$) and their experimentally determined counterparts ($\varepsilon_{\mathrm{vib}_m}^{\mathrm{G}}$):

$$\mathrm{RMSD(ZPE)} = \left(\frac{\sum_{m=1}^{M} \left(\lambda^{\mathrm{ZPE}} \varepsilon_{\mathrm{vib}_m}^{\mathrm{G,input}} - \varepsilon_{\mathrm{vib}_m}^{\mathrm{G}} \right)^2}{M} \right)^{1/2} . \tag{4.1}$$

Here, the number of molecules M is 15 since the best estimates for the experimental ZPEs ($\varepsilon_{\mathrm{vib}_m}^{\mathrm{G}}$) are taken from the ZPE15 database, shown in Table 4.3, while $\varepsilon_{\mathrm{vib}_m}^{\mathrm{G,input}}$ is defined as:

$$\varepsilon_{\mathrm{vib}_m}^{\mathrm{G,input}} \equiv \frac{hc}{2} \sum_{i} \omega_{i,m} \tag{4.2}$$

such that $\omega_{i,m}$ is the computed harmonic frequency of mode i of molecule m in units of cm^{-1}, h is Planck's constant, and c is the speed of light. The value of λ^{ZPE} which minimizes the RMSD(ZPE) from Equation (4.1) is obtained analytically from

$$\lambda^{\mathrm{ZPE}} = \frac{\sum_{m=1}^{M} \left(\varepsilon_{\mathrm{vib}_m}^{\mathrm{G,input}} \varepsilon_{\mathrm{vib}_m}^{\mathrm{G}} \right)}{\left(\varepsilon_{\mathrm{vib}_m}^{\mathrm{G,input}} \right)^2} . \tag{4.3}$$

Specifically, this approach explicitly calculates just the optimal scale factor for the ZPEs (λ^{ZPE}). The scale factors for the true harmonic frequencies (λ^{H}) and fundamental frequencies (λ^{F}) can be subsequently obtained through the relationships

$$\lambda^{\mathrm{H}} = \alpha^{\mathrm{H/ZPE}} \lambda^{\mathrm{ZPE}} \tag{4.4}$$

$$\lambda^{\mathrm{F}} = \alpha^{\mathrm{F/ZPE}} \lambda^{\mathrm{ZPE}} \tag{4.5}$$

such that the proportionality constants $\alpha^{\mathrm{H/ZPE}}$ and $\alpha^{\mathrm{F/ZPE}}$ are called the universal scale factor ratios since they should be independent of the electronic model chemistry and have been determined to be 1.014 ± 0.002 and 0.974 ± 0.002 respectively based on a test set of forty electronic model chemistries [26].

Table 4.3. Experimental ZPEs in kcal mol^{-1} for the 15 molecules used for fitting frequency scaling factors. The sources for the experimental values are listed in the footnote below the table.

Molecule	ZPE (kcal mol^{-1})
C_2H_2	16.490[a]
CH_4	27.710[a]
CO_2	7.3[a]
CO	3.0929144[a]
F_2	1.302[a]
H_2CO	16.1[a]
H_2O	13.26[a]
H_2	6.231[a]
HCN	10.000[a]
HF	5.864[a]
N_2O	6.770[b]
N_2	3.3618[a]
NH_3	21.200[c]
OH	5.2915[a]
Cl_2	0.7983[a]

[a] Karl K. Irikura [27]
[b] Grev et al. [28]
[c] J. M. L. Martin [29]

The workflow put forth by Alecu et al. [26] has been incorporated into the Automated Rate Calculator (ARC) software [24], which can automatically submit various electronic structure calculations, including the frequency calculations required to calculate λ^{ZPE}. This software was developed by Professor Dana and members for the Green Group at MIT and is currently maintained by the Dana Group at the Technion. ARC is free and publicly available on GitHub at https://github.com/ReactionMechanismGenerator/ARC. The Python package FREQ developed by Yu et al. [30] could easily be integrated into other computational workflows or serve as a stand-alone option to also calculate λ^{ZPE}.

4.2.2 Atom Energy Corrections

Atom Energy Corrections (AECs) are empirical corrections to systematic errors in quantum chemistry calculations for each specific LoT that can be used to improve the accuracy of species thermochemistry calculations. Specifically, AECs are corrections to the atomization energies associated with each atom in a molecule. These corrections are necessary because atomization energies obtained from ab initio calculations are often not very accurate since atoms and standard-state forms of some elements (e.g., graphite, O_2 $\left({}^3\Sigma_g^-\right)$ have substantially different electronic states than the closed-shell organic molecules studied in this chapter.

For a given molecule M, the enthalpy of formation at 298 K, $\Delta_\mathrm{f}\mathrm{H}(M,\ 298\mathrm{K})$, is defined as the difference in enthalpy between the molecule and its constituents' elemental states. However, the electronic energy obtained from a quantum chemistry single point energy calculation typically defines the zero of energy as charges separated at infinity. Thus, it is necessary to define a thermocycle to calculate $\Delta_\mathrm{f}\mathrm{H}(M,\ 298\mathrm{K})$. Starting at the top left of Figure 4.1 and working clockwise around the outer loop, the first step is to subtract the thermal contribution to enthalpy for the atoms reported by the Gaussian thermochemistry whitepaper [31] and then add the enthalpy of formation to create the stoichiometric number of atoms at 0K, $\Delta_\mathrm{f}\mathrm{H}(A,\ 298\mathrm{K})$. The next steps are to subtract the single point energies of the atoms E^A_{SP} (E^A_{ZPE} is 0 since individual atoms have no vibrational modes), add the calculated electronic energy and ZPE for the molecule (E^M_{SP} and E^M_{ZPE} respectively), and finally add the thermal contribution for the molecule, which is calculated using rigid-rotor harmonic oscillator (RRHO) approximation. In practice, calculating E^A_{SP} often results in inaccurate values so instead AECs are used and are fit to each level of theory. The RMG developers acknowledge that the term "atom energy corrections" may be misleading since these values are not corrections applied to E^A_{SP}, but rather entirely replace E^A_{SP} in the thermocycle. In summary, the outer loop of this thermocycle has only one unknown, the AECs.

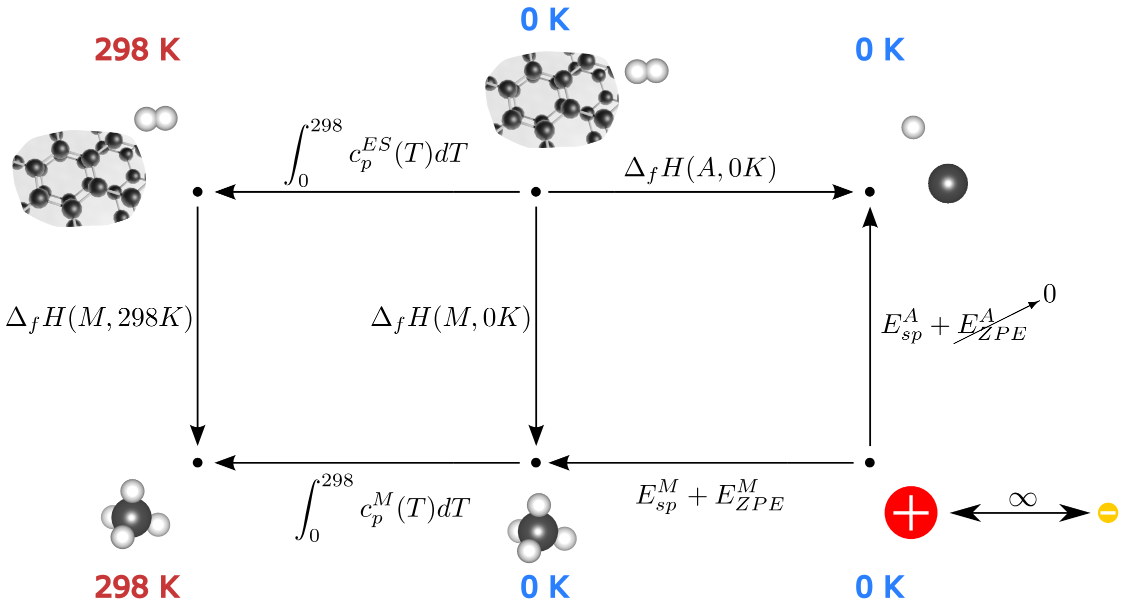


Figure 4.1. Schematic diagram outlining the thermocycle used to calculate the enthalpy of formation at 298K for a given molecule, Δ_fH(M, 298K). In this work, E_{SP}^A is replaced by AECs. The calculated estimate for Δ_fH(298K) can be further corrected by applying bond additivity corrections as described in Section 4.2.3. This figure was designed by Dr. A. Mark Payne and is adapted from his thesis.

The first step in determining the AECs is to perform a single point calculation for 16 small molecules: H_2, N_2, O_2, S_2, F_2, Cl_2, Br_2, HF, HCl, HBr, H_2S, H_2O, CH_4, NH_3, $ClCH_3$, and the methyl radical $^\bullet CH_3$. The AECs are then determined using linear least-squares fitting to minimize the error in the calculated Δ_fH(M, 298K) for these species. This set of molecules was chosen because they have precisely known heats of formation; the experimental atomization energies come from CCCBDB [10], and all have uncertainty values less than 0.2 kcal mol^{-1}. Additionally, these molecules are small enough that there is no risk of converging to the wrong conformer during the quantum chemistry calculation. To further minimize errors from the chosen level of theory, experimental geometries are used for the single point energy calculation, whose result is subsequently used for fitting. Note that for the CCSD(T)-F12a values in this work, the AECs are calculated by fitting the atomization energies to 14 molecules from this set. This is because the cc-pVDZ-F12 and cc-pVTZ-F12 basis sets only support the first three rows of the periodic table [32] so Br_2 and HBr cannot be used during the fitting procedure.

Although these corrections are crucial when comparing the computed Δ_fH(298 K) to experimental values, substantial residual error often remains if no further corrections, such as bond additivity corrections, are applied. Additionally, the AECs cancel out during the subtraction to obtain reaction enthalpies and barrier heights since all reactions considered here are balanced.

4.2.3 Bond Additivity Corrections

At most LoT, AECs alone are insufficient to estimate enthalpies of formation accurately enough for most kinetics applications. Empirically, bond additivity corrections (BACs) are required to further correct for errors in energies computed by electronic structure methods and thus obtain accurate thermochemistry. BACs are fit using a set of approximately 400 species with well-known experimental enthalpies of formation, primarily drawn from ATcT [33] and CCCBDB [10]. The experimental uncertainty is at most 0.55 kcal mol^{-1}, though most values are much lower with the median being just 0.14 kcal mol^{-1}. This reference set is stored within the RMG database. The full list of species can be found on GitHub at https://github.com/ReactionMechanismGenerator/RMG-database/tree/main/input/reference_sets/main.

Similar to AECs, BAC parameters are fit for each level of theory by minimizing the errors between the enthalpy of formation, $\Delta_f\mathrm{H}(M, 298\mathrm{K})$, calculated from quantum chemistry (including atom energy corrections) and the known experimental values for all reference species in the database. The first step in obtaining BACs is to perform a geometry optimization and frequency calculation for each species in the reference set at the desired level of theory. Originally, the experimental geometry was used as the initial guess for an optimization and frequency calculation with ωB97M-V/def2-TZVPD in Q-Chem when previous group members were calculating BACs for this level of theory. This density functional theory (DFT) optimized geometry is stored in RMG database and is used as the initial guess for the DFT optimizations with B97-D3/def2-mSVP and ωB97X-D3/def2-TZVP in this work.

The RMG database stores two popular types of BACs. First, Petersson-type BACs [34] apply a correction to $\Delta_f\mathrm{H}(298\ \mathrm{K})$ for each bond type present in the 2D representation of the molecule. These bond types are defined pairwise for every combination of elements for all integer bond orders (e.g., C–H, C–C, C–O, C=C, etc). Previous RMG developers chose the molecules in the reference set such that they would cover many of these pairwise combinations. The total BAC value for a molecule M is simply the summation of all bond type instances

$$\mathrm{BAC}(M) = \sum_{b}^{B} n_b E_b \tag{4.6}$$

such that n_b is the number of bond type b and E_b is the associated energy correction for that bond type. The total BAC value for the molecule can be applied to Δ_fH(298 K). Petersson-type BACs have a few potential shortcomings. First, this approach only considers integer bond lengths; it does not treat aromatic bonds differently so aromatic species must be represented by single and double bonds. As a related point, Petersson-type BACs do not account for variability in bond length or for the neighboring atoms, both of which are hypothesized to affect the contribution to the BAC to Δ_fH(M, 298K). Finally, by limiting the corrections to only explicitly encoded bond types, Petersson-type BACs may be sensitive to the resonance structure chosen to represent the molecule. Petersson-type BACs are derived using ordinary least-squares (OLS) fitting to minimize the difference between the reference Δ_fH(M, 298K) and the calculated values.

The second type of BAC was developed by Anantharaman and Melius [35]. These Melius-type BACs apply a correction based on bond distances from the 3D geometry of the optimized structure and considers information about neighboring atoms. Note that although Anantharaman and Melius refer to the method as "bond additivity correction", there are no parameters for specific bond types; however, we adopt their nomenclature for consistency in the literature. The equations from their paper can be rearranged to more clearly show the contributions from atom, bond, and molecule level corrections:

$$\Delta_f H^\circ_{298}(M) - \Delta_f^{\mathrm{BAC}} H^\circ_{298}(M) = \sum_{a \in M} \mathrm{Cor}^{\mathrm{atom}}(a) + \sum_{b \in M} \mathrm{Cor}^{\mathrm{bond}}\,(b) + \mathrm{Cor}^{\mathrm{molecule}}(M) \tag{4.7}$$

$$\mathrm{Cor}^{\mathrm{atom}}(a) = \alpha_a \tag{4.8}$$

$$\mathrm{Cor}^{\mathrm{bond}}(\mathrm{bond}(x,y)) = \sqrt{\beta_x \beta_y} \cdot e^{-\xi R^{xy}} + \sum_{w \in N(x) \backslash y} [\gamma_w + \gamma_x] + \sum_{z \in N(y) \backslash z} [\gamma_z + \gamma_y] \tag{4.9}$$

$$\mathrm{Cor}^{\mathrm{molecule}}(a) = K_{\mathrm{LoT}} \cdot \left(S^M - \sum_{a \in M} S^a \right) \tag{4.10}$$

Fitting Melius-type BACs involves fitting three parameters (α, β, γ) per element per level of theory as well as K_{LoT}, which depends only on the level of theory; these fitted parameters are denoted with a subscript. Variables with superscripts (e.g., S^M) are properties of the atoms, bonds, or molecule, such as spin and bond distances. ξ is a fixed parameter of 3 Å^{-1} as recommended by Anantharaman and Melius. As shown in Equation (4.8), the atom corrections for molecule M apply a correction (α_a) based on the element of atom a. Equation (4.9) shows the correction for a single bond between atoms x and y, such that R^{xy} is the bond length in Å and the fitted parameters β and γ are based on the elements. Note that this first term has a negative exponential dependence on bond distance i.e., the term will be more significant for shorter distances common in unsaturated bonds. The second and third term in Equation (4.9) sum over the neighbors of x and y, respectively, so the correction account for adjacent atoms. Finally, Equation (4.10) considers the spin of the molecule S^M and the spins of the atoms S^a. Since Equation (4.7) is non-linear in its fitted parameters, a non-linear least-squares fitting was used. 10 iterations are performed using different randomly-generated initial values for the parameters. The lowest value of this global optimization determines the best value for the fitted parameters. RMG's implementation uses the `least_squares` function in `scipy.optimize` [36] with the Trust Region Reflective method [37] and a 3-point method for calculating the Jacobian.

4.2.4 Conformer Search

The dataset from Chapter 3 is based on the geometries published by Grambow et al. [38], who had performed a conformer search for the reactants. The conformer search was performed by embedding several hundred conformers for each molecule using RDKit [39] with the ETKDG distance geometry method [40] and relaxing their geometries using the MMFF94 force field implemented in RDKit. The lowest energy conformer based on the MMFF94 energy was then optimized using Q-Chem [41] at both the B97-D3/def2-mSVP level of theory with Becke-Johnson damping [42] and the ωB97X-D3/def2-TZVP [43] level of theory. The reactant conformer was then used as the starting point for the growing string search. No conformer search was done for the transition state or product(s) of each reaction, and no additional conformer searching was done in Chapter 3.

Since only a subset of the reactions from Chapter 3 will be added to RMG database, a conformer search is done for all reactants, transition states, and products discussed in this chapter. The conformer search done here is also much more thorough and would require enormous compute resources to complete for the entire dataset of nearly 12,000 reactions. Here, the Automated Conformer Search (ACS) software tool was used to do conformer searching. This software was developed by Haoyang (Oscar) Wu and Xiaorui Dong and is made publicly available at https://github.com/oscarwumit/ACS.

The workflow for ACS starts from an initial geometry and then exhaustively rotates dihedrals to perform a thorough conformer search. Here, increments of 18° were chosen to create a relatively fine grid, and the geometry that was previously optimized at ωB97X-D3/def2-TZVP was used the initial geometry. This approach should do a good job of identifying the lowest energy conformer for molecules with ≤ 2 rotatable bonds. If a molecule has more than two rotors, ACS will rotate $\binom{N}{2}$ rotors, such that N is the number of rotors. Although coupling all rotors would be a more exhaustive search and give higher likelihood of identifying the lowest energy conformer, such an approach would scale as $\left(\frac{360°}{\text{Number of increments}}\right)^N$, which leads to far too many initial conformer structures for even medium-size molecules. Thus, $\binom{N}{2}$ presents a compromise between accuracy and computational expense.

The next step is to perform single point energy calculations using Psi4 [44] for each of the generated structures to estimate the potential energy surface. Ideally, a relatively cheap method would be used for this step. However, recent work has shown that MMFF94[45–47] does a poor job of ranking the relative energies of different conformers.[48] Although GFN2-xTB[49] single point energies show stronger agreement with accurate quantum calculations[50] than MMFF94, this work chooses to place a very high priority on accuracy and thus uses ωB97X-D3/def2-TZVP within Psi4 to calculate the energy of these un-optimized conformers. This level of theory was chosen since it is the same level used for the subsequent optimization and frequency calculations so this approach should avoid any mismatch that may be caused by using a different level of theory. The next step is to take the structure the corresponds to each local minima on this potential energy surface and use that as the initial guess for optimization and frequency calculations, which are done with ωB97X-D3/def2-TZVP in Q-Chem. Finally, single point energy refinement is at CCSD(T)-F12a/cc-pVDZ-F12 with Molpro for all unique conformers.

This chapter does not change the underlying workflow. However, I did perform various software engineering tasks to automate this workflow so it could adequately handle the high-throughput calculations required for this work. My modifications and scripts are free and publicly available on the `qchem_molpro` branch. In general, this conformer search workflow can find a lower energy conformer relative to the geometries published by Grambow et al. [38] about 60% of the time, even for the reactants, which had previously undergone a conformer search.

4.2.5 Computing Thermochemistry and Kinetics

The Automated Reaction Kinetics and Network Exploration (Arkane) software package is used to compute thermochemistry and high-pressure limit rate coefficients using the results from these quantum chemistry calculations [51]. 1D hindered rotor (HR) scans are performed for the reactants, transition states, and products, at ωB97X-D/def2-TZVP in Gaussian [52] to improve the estimate of the partition function [53–55]. 1DHR scans are used to balance accuracy and the immense computational cost of intense conformational search; future work could explore multidimensional torsion treatment or even more thorough conformer searching methods, whose resulting conformers could be used with multi-structure TST [56–58]. Arkane uses RRHO TST with Eckart tunneling correction to calculate the forward rate coefficient for a set of user-defined temperatures. Here, I use 50 linearly spaced points in the reciprocal temperature space between 300 K and 2000 K. Finally, Arkane uses a linear least-squares fitting to fit the list of reciprocal temperatures and logarithm of the rate coefficients to a 3-parameter Arrhenius expression,

$$k = AT^n e^{\frac{-E_a}{RT}} \tag{4.11}$$

yielding the best approximation for the pre-exponential A-factor, temperature exponential n, and the activation energy E_a. The individual species thermochemistry can be used to compute the Gibbs free energy of a reaction, which is related to a reaction's equilibrium constant and allows for easy estimation of the reverse reaction's rate coefficient. Additional details about Arkane can be found in the corresponding publication [51].

4.3 Results

4.3.1 Frequency Scaling Factor

The harmonic frequency scaling factor was found to be 1.014 for B97-D3/def2-mSVP. For ωB97X-D3/def2-TZVP, it was found to be 0.984. As a sanity check, these values are similar to values stored in CCCBDB [10] for similar functionals as well as to calculated values stored in RMG database. The harmonic frequency scaling factors calculated in this work, as well as for many other LoT, are publicly available in RMG's database at `RMG-database/input/quantum_corrections/data.py` and on GitHub so others can benefit from these results.

4.3.2 Atom Energy Corrections

The atom energy corrections resulting from the OLS fitting for two levels of DFT are shown in Table 4.4. The results from the two explicitly correlated coupled cluster methods are shown in Table 4.5. All values are also publicly available in RMG's database at `RMG-database/input/quantum_corrections/data.py` and on GitHub. As a sanity check, the AEC values for the LoT in this work are similar to AEC values for other LoT that are already stored in RMG database.

Table 4.4. Fitted AECs in Hartrees for two DFT methods from Q-Chem.

	Atom Energy Correction (Hartree)	
Atom	B97-D3/def2-mSVP	ωB97X-D3/def2-TZVP
H	-0.49516021903680546	-0.49991749801833063
C	-37.833878658169624	-37.84993993601866
N	-54.541699219144505	-54.58750889521559
O	-75.01524900146931	-75.07402423801669
F	-99.65948092488345	-99.7392410428872
S	-397.97912317555034	-398.10051734579775
Cl	-459.9963616463421	-460.1341841971797
Br	-2575.1192984500663	-2574.175384284091

Table 4.5. Fitted AECs in Hartrees for two explicitly correlated coupled cluster levels of theory from Molpro.

	Atom Energy Correction (Hartree)	
Atom	CCSD(T)-F12a/cc-pVDZ-F12	CCSD(T)-F12a/cc-pVTZ-F12
H	-0.50000836574607	-0.5003554055415579
C	-37.784271457731904	-37.787690380768844
N	-54.523156256858144	-54.52874573314225
O	-74.99320041804718	-75.00314974528959
F	-99.6511861642652	-99.6658140570029
S	-397.6625539051389	-397.67549377693314
Cl	-459.68886861844055	-459.70521576925796

4.3.3 Bond Additivity Corrections

The parameters for the Petersson-type BACs resulting from the OLS fitting for the two DFT methods are shown in Table 4.6. The parameters for the two composite LoT at explicitly correlated coupled cluster are shown in Table 4.7. The molecular correction parameter for Melius-type BACs for all LoT is shown in Table 4.8. The atom, bond length, and neighboring atom correction parameters for the four LoT are shown in Tables 4.9, 4.10, 4.11, and 4.12. All values are publicly available in RMG's database at `RMG-database/input/quantum_corrections/data.py` and on GitHub. The BAC values reported here are similar to values for other LoT that are already stored in RMG database.

Table 4.6. Petersson-type BACs in kcal mol^{-1} for two DFT methods from Q-Chem.

	Bond Additivity Correction (kcal mol^{-1})	
Atom	B97-D3/def2-mSVP	ωB97X-D3/def2-TZVP
Br–Br	3.3184897418778894	2.849979178650586
Br–C	-0.6447457118766096	0.7381762723415655
Br–Cl	2.5458979359916487	1.4764289266543953
Br–F	3.9321266497717513	2.7336108440861704
Br–H	3.1450731163630294	2.621866526877559
Br–O	-7.328960869354291	-2.5532834764130006
C≡C	-19.75782127331025	-7.92496166307853
C≡N	-7.499027151749764	-3.864663093893819
C≡O	-18.475225675145023	-7.484067031104203
C–C	-6.498460645920902	-1.0817871087375186
C–Cl	-1.9238440096546874	-0.5732704243120552
C–F	-1.9747353079897023	1.3719540592946495
C–H	0.01729290032079855	0.07704003746021679
C–N	-1.7433898062480928	0.5919032459443995
C–O	-3.8525377364775215	-1.2626115105817743
C–S	-5.419539134610862	-1.2805399168193157
C=C	-12.998063133432941	-3.890848609981227
C=N	-5.027572638791928	-1.5036825315665
C=O	-6.801076987753849	-2.9505302757552596
C=S	-5.654222632762431	-3.7136643912288623
Cl–Cl	1.3452682321784095	0.136962629607007
Cl–F	3.491738723755338	1.043801519750975
Cl–H	0.8075310212109841	0.3209061926667829
Cl–N	11.423744786147633	1.254243872961764
Cl–O	5.488334623555848	-1.2050377287736824
Cl–S	-3.0170592756249057	-0.353190509282856
F–F	5.002957678305598	1.2363169265407195
F–H	-8.586384858108858	-1.4703521420571022
F–O	4.485916714230829	-0.008583994665357622
F–S	-4.159578702518104	0.3093056813192921
H–H	8.526122583518188	0.16749830148486178
H–N	1.2169107867656828	0.7806156486742294
H–O	-3.8567427899343043	-1.2446586747155781
H–S	1.8235600781274208	1.186472823080999
N≡N	2.7964223466789293	-0.15105524935670453
N–N	5.170147621733188	3.437416216021716
N–O	3.5592924696925587	1.4467328418363092
N=N	8.731292427600357	2.786466016509395
N=O	5.09080676349987	-2.389849264267907
O–O	2.565372397566582	-2.27124059610634
O–S	-8.578422244244877	-2.6554640707366945
O=O	1.3114212140745158	-9.840434624890442
O=S	-14.229024507889747	-3.470457647257026
S–S	-4.258008825508542	-1.3273180651577856
S=S	-3.631460853989999	-6.864453512935457

Table 4.7. Petersson-type BACs in kcal mol^{-1} for two composite levels of theory. Optimization and frequency calculations are done with Q-Chem while single point energy refinement is done using Molpro.

	Bond Additivity Correction (kcal mol^{-1})	
	CCSD(T)-F12a/cc-pVDZ-F12//	CCSD(T)-F12a/cc-pVTZ-F12//
Atom	ωB97X-D3/def2-TZVP	ωB97X-D3/def2-TZVP
C≡C	-0.7940736288861328	-0.37319831366523176
C≡N	-0.2736082180869344	-0.04394537008104246
C≡O	1.215648303725274	0.6845507542289013
C−C	-0.006153161261658441	0.12602185168066687
C−Cl	0.7595723077901367	0.4000395071268901
C−F	0.4815204101380463	0.5386953055500939
C−H	0.010156098370800108	0.03014784912580578
C−N	0.10438905540134882	0.3022567343857892
C−O	0.09987252247930076	0.22156464774957604
C−S	0.4090396417072708	0.059041022372472614
C=C	-0.32837634513571645	-0.049584091100670837
C=N	-0.6872690836499937	-0.4428639989215071
C=O	-0.011879042720396643	0.05943900637676746
C=S	-0.45740929709171363	-0.7758177193710785
Cl−Cl	-0.014451326911147977	0.1382704625282193
Cl−F	0.6479093120532244	0.3846791604086004
Cl−H	0.37650914650193457	0.19401601074149966
Cl−N	0.5872703379266317	0.4109773959959559
Cl−O	0.5095510277409292	0.2892748705965537
Cl−S	0.646585596795355	0.21784097251125079
F−F	-0.7037812030852666	-0.6423018409407416
F−H	0.13879622147405257	0.1102321131699
F−O	-0.5823379151756342	-0.5555283888816168
F−S	0.9474443582208304	0.8514568108157713
H−H	-0.00879296155638677	-0.24061025875957348
H−N	-0.3973581320058072	-0.2458184828878194
H−O	-0.05082328074118156	0.007966375791966561
H−S	0.9006352849672502	0.6141296238226581
N≡N	0.6426432154389473	0.15991986841815756
N−N	1.0261474051290596	1.2279613861249787
N−O	-0.03642922490749102	0.20879882425699858
N=N	0.1708502331465831	0.44949205775999634
N=O	-0.8042125954300557	-1.029059282226874
O−O	-0.7325577965540944	-0.611163660061622
O−S	-0.18825122982916948	-0.43698174359431335
O=O	-2.7021030234346695	-2.8858691436809196
O=S	0.9232025676598207	0.3100869947699715
S−S	0.14102013117456597	0.10571953923531874
S=S	-0.9855656265939845	-1.6070546530751422

Table 4.8. Molecular correction parameter for Melius-type BACs for each level of theory

Level of Theory	K_{LoT} (kcal mol^{-1})
B97-D3/def2-mSVP	-3.096451394144581
ωB97X-D3/def2-TZVP	-3.782190737782739
CCSD(T)-F12a/cc-pVDZ-F12// ωB97X-D3/def2-TZVP	0.30287826305208276
CCSD(T)-F12a/cc-pVTZ-F12// ωB97X-D3/def2-TZVP	0.3057912044551125

Table 4.9. Fitted parameters in kcal mol^{-1} for Melius-type BACs for B97-D3/def2-mSVP.

Element	α	β	γ
Br	-4.262780823587865	8692.219738590926	-0.637372927288811
C	4.999999999999999	133.3489203999116	0.07497384490455655
Cl	-4.999999999999999	983.472669453573	0.412600941322612
F	-4.999999999999999	98.4144446431002	0.44818650615117767
H	-1.0450492093486716	0.9620481174299411	-0.9999999999999999
N	-4.999999999999999	1.258833913444385e-27	-0.2675044921389425
O	-4.999999999999999	356.54294878127047	-0.16343725217730817
S	-4.99999999999999	4253.83724390785	0.11474046296004109

Table 4.10. Fitted parameters in kcal mol^{-1} for Melius-type BACs for ωB97X-D3/def2-TZVP.

Element	α	β	γ
Br	-4.236702026546933	2886.4449179333255	0.27904803369228603
C	0.2049164817767285	0.05644392159534064	-0.09439333163743208
Cl	-2.5370579912419085	257.3272543651555	0.17830274429860773
F	-4.390697891358869	136.75193570503143	0.14749707198083395
H	-2.062847444255194	1.0783080754563377	-0.1790691880962216
N	-4.999999999999999	4.0393833100316303e-35	-0.3694958066667204
O	-2.485820286098204	120.00409931223447	-0.07383806558254727
S	-2.3806435571125646	301.39625436311553	-0.13458257713999408

Table 4.11. Fitted parameters in kcal mol^{-1} for Melius-type BACs for CCSD(T)-F12a/cc-pVDZ-F12//ωB97X-D3/def2-TZVP.

Element	α	β	γ
C	-0.5464261912627	64.6767878335258	-0.024587039354705877
Cl	-0.3331232529849068	6.316147607205602	-0.03540309101545116
F	0.09684535337591452	2.9969902462203326e-14	-0.016933453385969043
H	0.3093277751661718	1.8496519952345347e-16	0.046660614206200754
N	0.4773300729311916	9.624991083595463	-0.0053021340471547
O	0.6949720732175282	1.0824291132601096e-15	-0.01078514461511072
S	-0.9142689727759676	735.661291892654	-0.13988268641251655

Table 4.12. Fitted parameters in kcal mol^{-1} for Melius-type BACs for CCSD(T)-F12a/cc-pVTZ-F12//ωB97X-D3/def2-TZVP.

Element	α	β	γ
C	-0.6004599146970828	57.600230521006345	-0.02675767045779017
Cl	-0.21285368087888518	72.75494140267224	-0.026157341196276294
F	0.20136534958427602	2.7637442704545506e-39	-0.09738287356772249
H	0.1409606206795142	0.04568584404132764	0.08331791884967885
N	0.44677472396061857	7.988043181948807	-0.031799354286893795
O	0.7724579645053421	1.1170405486633287e-25	-0.04321158886089615
S	-0.1294113249415446	337.25420543898184	-0.07172725215109269

Table 4.13 summarizes the mean absolute errors (MAE) and root mean squared errors (RMSE) when applying these corrections for each level of theory. Figures 4.2, 4.3, 4.4, and 4.5 show a histogram of the residuals before and after applying BACs; errors are relative to the reliable set of experimental enthalpies used for fitting. AECs are always included. After applying BACs, the distribution of errors both centers much more closely to zero and noticeably tightens, representing a significant reduction in both the bias and variance of errors. These figures show the expected trend in which using a higher level of theory, such as explicitly correlated coupled cluster calculations, give lower errors both before and after applying the fitted BACs. Further, despite the potential limitations of Petersson-type BACs, the results for these four LoT indicate that Petersson-type BACs perform slightly better than Melius-type BACs. While all of these results report a training error, I also tested whether the fitted parameters generalize well to new molecules. As described in

the technical validation for Chapter 3, I applied the Petersson-type BACs fit at CCSD(T)-F12a/cc-pVDZ-F12//ωB97X-D3/def2-TZVP to molecules from the Pedley compilation set [59] that have experimental uncertainty of less than 1 kcal mol^{-1}. The resulting RMSE of 1.2 kcal mol^{-1} demonstrates the accuracy of our BACs.

Table 4.13. Errors in kcal mol^{-1} relative to experimental data before and after applying either Petersson-type or Melius-type BACs that were fit for each respective level of theory. AECs are always included.

	Without BACs		**Petersson**		**Melius**	
Reaction Family	MAE	RMSE	MAE	RMSE	MAE	RMSE
B97-D3/def2-mSVP	21.76	27.83	3.16	6.58	4.23	7.31
ωB97X-D3/def2-TZVP	5.12	6.79	1.33	2.22	1.43	2.22
CCSD(T)-F12a/cc-pVDZ-F12// ωB97X-D3/def2-TZVP	0.94	1.26	0.52	0.83	0.57	0.91
CCSD(T)-F12a/cc-pVTZ-F12// ωB97X-D3/def2-TZVP	0.82	1.17	0.49	0.80	0.51	0.87

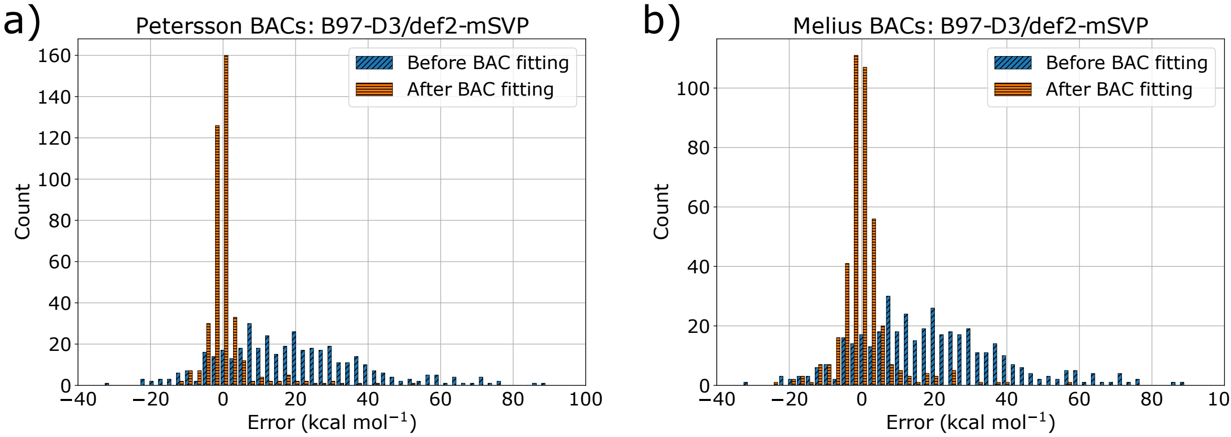


Figure 4.2. Histogram of fitting errors (corrected minus experimental enthalpies) for B97-D3/def2-mSVP when using a) Petersson-type BACs and b) Melius-type BACs.

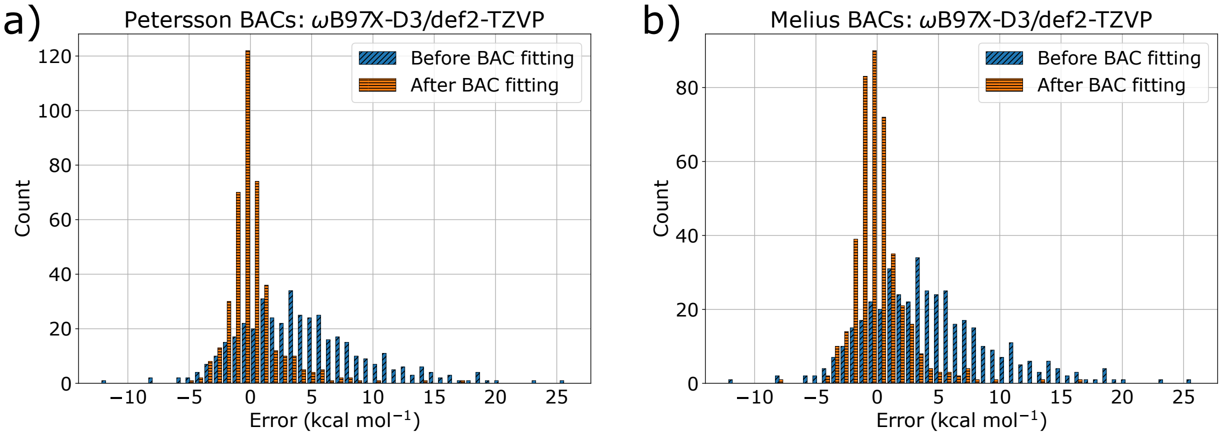


Figure 4.3. Histogram of fitting errors (corrected minus experimental enthalpies) for ωB97X-D3/def2-TZVP when using a) Petersson-type BACs and b) Melius-type BACs.

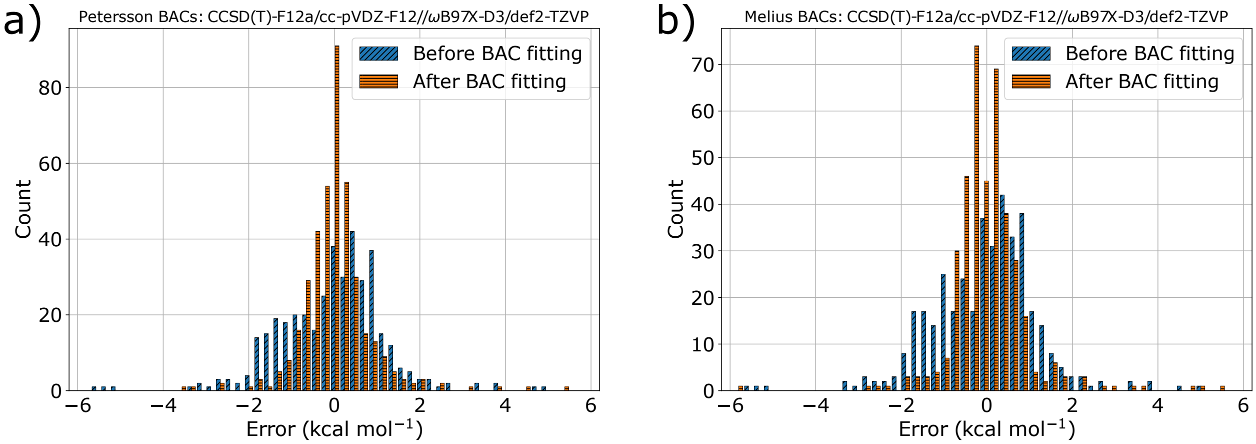


Figure 4.4. Histogram of fitting errors (corrected minus experimental enthalpies) for CCSD(T)-F12a/cc-pVDZ-F12//ωB97X-D3/def2-TZVP when using a) Petersson-type BACs and b) Melius-type BACs.

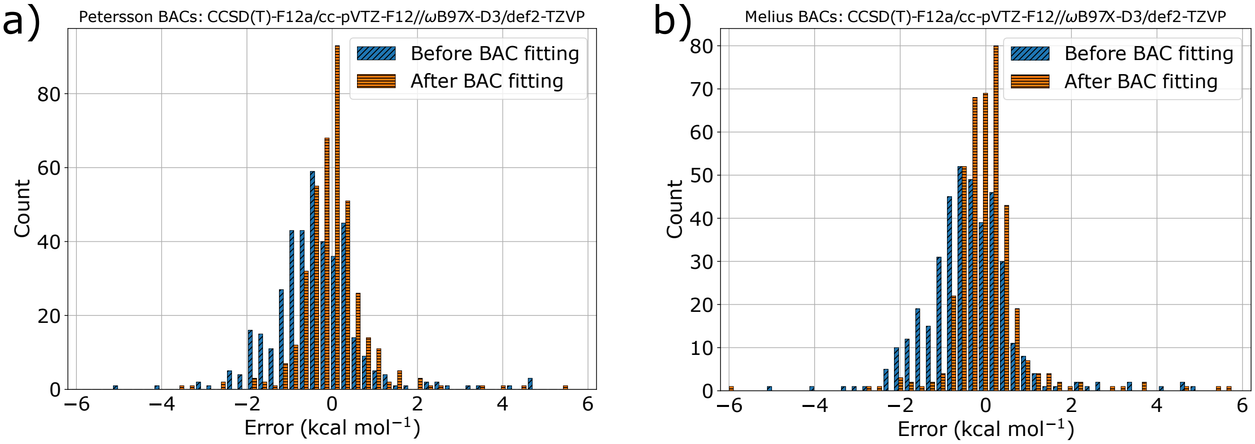


Figure 4.5. Histogram of fitting errors (corrected minus experimental enthalpies) for CCSD(T)-F12a/cc-pVTZ-F12//ωB97X-D3/def2-TZVP when using a) Petersson-type BACs and b) Melius-type BACs.

4.4 Improvements to RMG's Kinetic Estimates

Here, I add over one hundred training reactions to RMG database and refit the decision tree rate estimators. The thermochemistry information for all reactant and product species involved in these reactions is stored in `Spiekermann_refining_elementary_reactions.py`. Given the high-quality data stored in this library, future users should be encouraged to include this library when using RMG to generate kinetic mechanisms. That way, if any of these species end up in the mechanism, RMG can directly use this very accurate thermochemistry data. Showing a plot of each rotor scan and Arrhenius fit for each of the reactions would use an unnecessary amount of pages. Interested readers can find this info in the various pull requests that have been made on the RMG database GitHub. Each pull request includes a Jupyter Notebook that contains all related plots for the corresponding rotor scans and Arrhenius fits.

To show the benefit of retraining the rate trees with these new training reactions, I calculate the improvement ratio as the ratio between the new and old RMG rate estimate for these reactions using

$$\text{Improvement ratio} = \frac{\text{New rate estimate}}{\text{Prior rate estimate}} \tag{4.12}$$

If needed, the reciprocal is taken so all ratios are ≥ 1. The prior rate estimate comes from RMG rate rules. After my additions, all of these reactions now exist as training reactions within RMG database. Thus, RMG will directly query their Arrhenius expression to evaluate the rate coefficients rather than estimate the parameters. As described in the methods section, the Arrhenius expression for these training reactions comes from fitting the results of TST calculations from Arkane. The summary from Table 4.14 shows that the rate estimates are improved by at least one order of magnitude, sometimes substantially more.

Table 4.14. Median improvement ratio between the new and old RMG rate coefficient estimate for reactions from Chapter 3.

RMG Family	500K	1000K	1500K	2000K
1,3 Insertion CO_2	1.06e10	2.81e4	603	159
1,3 Sigmatropic	1.86e12	2.52e4	2.05e3	4.05e3
Diels Alder Addition	476	21.2	4.49	2.28
Ketoenol	168	10.9	7.32	12.4
Retroene	25.2	5.72	3.56	3.01

In addition to adding reactions from the dataset presented in Chapter 3, I also added two reactions as part of a separate project. The improvement ratio for these two reactions is shown in Table 4.15 and was very substantial for the Intra R Add Exocyclic reaction.

Table 4.15. Improvement ratio between the new and old RMG rate estimate for Intra R Add Exocyclic and Endocyclic reactions.

RMG Family	500K	1000K	1500K	2000K
Intra R Add Exocyclic	3.46e4	44.39	4.56	1.42
Intra R Add Endocyclic	5.48	4.52	4.34	4.30

4.5 References

(1) Manion, J. A.; Huie, R. E.; Levin, R. D.; Jr., D. R. B.; Orkin, V. L.; Tsang, W.; McGivern, W. S.; Hudgens, J. W.; Knyazev, V. D.; Atkinson, D. B.; Chai, E.; Tereza, A. M.; Lin, C.-Y.; Allison, T. C.; Mallard, W. G.; Westley, F.; Herron, J. T.; Hampson, R. F.; Frizzell, D. H. NIST Chemical Kinetics Database, NIST Standard Reference Database.

(2) Ruscic, B.; Pinzon, R. E.; Von Laszewski, G.; Kodeboyina, D.; Burcat, A.; Leahy, D.; Montoy, D.; Wagner, A. F. In *Journal of Physics: Conference Series*, 2005, pp 26–30.

(3) PrIMe Kinetics, https://primekinetics.org.

(4) Chemked, https://www.chemked.com.

(5) Weber, B. W.; Niemeyer, K. E. ChemKED: A Human- and Machine-Readable Data Standard for Chemical Kinetics Experiments. *Int. J. Chem. Kinet.* **2018**, *50*, 135–148.

(6) CloudFlame, https://cloudflame.kaust.edu.sa/.

(7) Goteng, G. L.; Nettyam, N.; Sarathy, S. M. In *Proceedings - 2013 International Conference on Information Science and Cloud Computing Companion, ISCC-C 2013*, Inst. Electr. Electron. Eng. Inc.: 2014, pp 294–299.

(8) Respecth, https://respecth.chem.elte.hu/respecth/index.php.

(9) Smith, D. G. A.; Altarawy, D.; Burns, L. A.; Welborn, M.; Naden, L. N.; Ward, L.; Ellis, S.; Pritchard, B. P.; Crawford, T. D. The MolSSI QC Archive project: An Open-Source Platform to Compute, Organize, and Share Quantum Chemistry Data. *WIREs Comput. Mol. Sci.* **2020**, *11*, 1491.

(10) Johnson III, R. D. NIST Computational Chemistry Comparison and Benchmark Database, NIST Standard Reference Database Number 101. *http://cccbdb.nist.gov/* **2020**.

(11) Benson, S. W.; Golden, D. M.; Haugen, G. R.; Shaw, R.; Cruickshank, F. R.; Rodgers, A. S.; O'neal, H. E.; Walsh, R. Additivity Rules for the Estimation of Thermochemical Properties. *Chem. Rev.* **1969**, *69*, 279–324.

(12) Eigenmann, H. K.; Golden, D. M.; Benson, S. W. Revised Group Additivity Parameters for the Enthalpies of Formation of Oxygen-Containing Organic Compounds. *J. Phys. Chem.* **1973**, *77*, 1687–1691.

(13) Vandewiele, N. M.; Van Geem, K. M.; Reyniers, M. F.; Marin, G. B. Genesys: Kinetic Model Construction using Chemo-informatics. *Chem. Eng. J.* **2012**, *207-208*, 526–538.

(14) Ritter, E. R.; Bozzelli, J. W. THERM: Thermodynamic Property Estimation for Gas Phase Radicals and Molecules. *Int. J. Chem. Kinet.* **1991**, *23*, 767–778.

(15) Li, Y. P.; Han, K.; Grambow, C. A.; Green, W. H. Self-Evolving Machine: A Continuously Improving Model for Molecular Thermochemistry. *J. Phys. Chem. A* **2019**, *123*, 2142–2152.

(16) Grambow, C. A.; Pattanaik, L.; Green, W. H. Deep Learning of Activation Energies. *J. Phys. Chem. Lett.* **2020**, *11*, 2992–2997.

(17) Sirumalla, S. K. Graph Neural Networks and High Throughput Quantum Chemistry Workflows for Detailed Kinetic Modeling, Ph.D. Thesis, 2021.

(18) Van de Vijver, R.; Sabbe, M. K.; Reyniers, M.-F.; Van Geem, K. M.; Marin, G. B. Ab Initio Derived Group Additivity Model for Intramolecular Hydrogen Abstraction Reactions. *Phys. Chem. Chem. Phys.* **2018**, *20*, 10877–10894.

(19) West, R. H.; Allen, J. W.; Green, W. H. ChemInform Abstract: Automatic Reaction Mechanism Generation with Group Additive Kinetics. *ChemInform* **2012**, *43*.

(20) Johnson, M. S.; Green, W. H. A Machine Learning Based Approach to Reaction Rate Estimation. *ChemRxiv* **2022**, DOI: 10.26434/chemrxiv-2022-c98gc.

(21) Johnson, M. S.; Dong, X.; Grinberg Dana, A.; Chung, Y.; Farina Jr., D.; Gillis, R. J.; Liu, M.; Yee, N. W.; Blondal, K.; Mazeau, E.; Grambow, C. A.; Payne, A. M.; Spiekermann, K. A.; Pang, H.-W.; Goldsmith, F. C.; West, R. H.; Green, W. H. RMG Database for Chemical Property Prediction. *J. Chem. Inf. Model.* **2022**, *62*, 4906–4915.

(22) Gao, C. W.; Allen, J. W.; Green, W. H.; West, R. H. Reaction Mechanism Generator: Automatic Construction of Chemical Kinetic Mechanisms. *Comput. Phys. Commun.* **2016**, *203*, 212–225.

(23) Liu, M.; Grinberg Dana, A.; Johnson, M. S.; Goldman, M. J.; Jocher, A.; Payne, A. M.; Grambow, C. A.; Han, K.; Yee, N. W.; Mazeau, E. J., et al. Reaction Mechanism Generator v3.0: Advances in Automatic Mechanism Generation. *J. Chem. Inf. Model.* **2021**, *61*, 2686–2696.

(24) Grinberg Dana, A.; Ranasinghe, D.; Wu, H.; Grambow, C.; Dong, X.; Johnson, M.; Goldman, M.; Liu, M.; Green, W. ARC - Automated Rate Calculator, version 1.1.0, 2019.

(25) A. Grinberg Dana K.A. Spiekermann, W. G. The Tandem Tool (T3) for automated kinetic model generation and refinement. Version 0.1.0, https://github.com/ReactionMechanismGenerator/T3, 2020.

(26) Alecu, I. M.; Zheng, J.; Zhao, Y.; Truhlar, D. G. Computational Thermochemistry: Scale Factor Databases and Scale Factors for Vibrational Frequencies Obtained from Electronic Model Chemistries. *J. Chem. Theory Comput.* **2010**, *6*, 2872–2887.

(27) Irikura, K. K. Experimental Vibrational Zero-Point Energies: Diatomic Molecules. *J. Phys. Chem. Ref. Data* **2007**, *36*, 389–397.

(28) Grev, R. S.; Janssen, C. L.; Schaefer III, H. F. Concerning Zero-Point Vibrational Energy Corrections to Electronic Energies. *J. Chem. Phys.* **1991**, *95*, 5128–5132.

(29) Martin, J. M. L. On the Performance of Large Gaussian Basis Sets for the Computation of Total Atomization Energies. *J. Chem. Phys.* **1992**, *97*, 5012–5018.

(30) Haoyu, S. Y.; Fiedler, L. J.; Alecu, I. M.; Truhlar, D. G. Computational Thermochemistry: Automated Generation of Scale Factors for Vibrational Frequencies Calculated by Electronic Structure Model Chemistries. *Comput. Phys. Commun.* **2017**, *210*, 132–138.

(31) Ochterski, J. W. Thermochemistry in Gaussian. *Gaussian Inc* **2000**, *1*, 1–19.

(32) Pritchard, B. P.; Altarawy, D.; Didier, B.; Gibson, T. D.; Windus, T. L. New Basis Set Exchange: An Open, Up-to-Date Resource for the Molecular Sciences Community. *J. Chem. Inf. Model.* **2019**, *59*, 4814–4820.

(33) Ruscic, B.; Bross, D. H. Active Thermochemical Tables (ATcT) Values Based on ver. 1.122d of the Thermochemical, https://atct.anl.gov/Thermochemical%20Data/version%201.122d/index.php, 2018.

(34) Petersson, G. A.; Malick, D. K.; Wilson, W. G.; Ochterski, J. W.; Montgomery Jr., J. A.; Frisch, M. J. Calibration and Comparison of the Gaussian-2, Complete Basis Set, and Density Functional Methods for Computational Thermochemistry. *J. Chem. Phys.* **1998**, *109*, 10570–10579.

(35) Anantharaman, B.; Melius, C. F. Bond Additivity Corrections for G3B3 and G3MP2B3 Quantum Chemistry Methods. *J. Phys. Chem. A* **2005**, *109*, 1734–1747.

(36) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J., et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261–272.

(37) Branch, M. A.; Coleman, T. F.; Li, Y. A Subspace, Interior, and Conjugate Gradient Method for Large-Scale Bound-Constrained Minimization Problems. *SIAM J. Sci. Comput.* **1999**, *21*, 1–23.

(38) Grambow, C. A.; Pattanaik, L.; Green, W. H. Reactants, Products, and Transition States of Elementary Chemical Reactions Based on Quantum Chemistry. *Sci. Data* **2020**, *7*, 1–8.

(39) Landrum, G. et al. RDKit: Open-Source Cheminformatics Software, https://www.rdkit.org, 2006.

(40) Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know to Improve Conformation Generation. *J. Chem. Inf. Model.* **2015**, *55*, 2562–2574.

(41) Shao, Y.; Gan, Z.; Epifanovsky, E.; Gilbert, A. T. B.; Wormit, M.; Kussmann, J.; Lange, A. W.; Behn, A.; Deng, J.; Feng, X., et al. Advances in Molecular Quantum Chemistry Contained in the Q-Chem 4 Program Package. *Mol. Phys.* **2015**, *113*, 184–215.

(42) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011**, *32*, 1456–1465.

(43) Lin, Y.-S.; Li, G.-D.; Mao, S.-P.; Chai, J.-D. Long-range Corrected Hybrid Density Functionals with Improved Dispersion Corrections. *J. Chem. Theory Comput.* **2013**, *9*, 263–272.

(44) Smith, D. G. A.; Burns, L. A.; Simmonett, A. C.; Parrish, R. M.; Schieber, M. C.; Galvelis, R.; Kraus, P.; Kruse, H.; Di Remigio, R.; Alenaizan, A., et al. PSI4 1.4: Open-Source Software for High-Throughput Quantum Chemistry. *J. Chem. Phys.* **2020**, *152*.

(45) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490–519.

(46) Halgren, T. A. Merck Molecular Force Field. II. MMFF94 Van der Waals and Electrostatic Parameters for Intermolecular Interactions. *J. Comput. Chem.* **1996**, *17*, 520–552.

(47) Halgren, T. A. Merck Molecular Force Field. III. Molecular Geometries and Vibrational Frequencies for MMFF94. *J. Comput. Chem.* **1996**, *17*, 553–586.

(48) Kanal, I. Y.; Keith, J. A.; Hutchison, G. R. A Sobering Assessment of Small-Molecule Force Field Methods for Low Energy Conformer Predictions. *Int. J. Quantum Chem.* **2018**, *118*, e25512.

(49) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB–An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.

(50) Sharapa, D. I.; Genaev, A.; Cavallo, L.; Minenkov, Y. A Robust and Cost-Efficient Scheme for Accurate Conformational Energies of Organic Molecules. *ChemPhysChem* **2019**, *20*, 92–102.

(51) Dana, A. G.; Johnson, M. S.; Allen, J. W.; Sharma, S.; Raman, S.; Liu, M.; Gao, C. W.; Grambow, C. A.; Goldman, M. J.; Ranasinghe, D. S.; Gillis, R. J.; Payne, A. M.; Li, Y.-P.; Dong, X.; Spiekermann, K. A.; Wu, H.; Dames, E. E.; Buras, Z. J.; Vandewiele, N. M.; Yee, N. W.; Merchant, S. S.; Buesser, B.; Class, C. A.; Goldsmith, F.; West, R. H.; Green, W. H. Automated Reaction Kinetics and Network Exploration (Arkane): A Statistical Mechanics, Thermodynamics, Transition State Theory, and Master Equation Software. *Int. J. Chem. Kinet.* **2022**, DOI: 10.1002/kin.21637.

(52) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. Gaussian 16 Revision C.01, Gaussian Inc. Wallingford CT, 2016.

(53) Vansteenkiste, P.; Van Speybroeck, V.; Pauwels, E.; Waroquier, M. How Should We Calculate Multi-Dimensional Potential Energy Surfaces for an Accurate Reproduction of Partition Functions? *Chem. Phys.* **2005**, *314*, 109–117.

(54) Sharma, S.; Raman, S.; Green, W. H. Intramolecular Hydrogen Migration in Alkylperoxy and Hydroperoxyalkylperoxy Radicals: Accurate Treatment of Hindered Rotors. *J. Phys. Chem. A* **2010**, *114*, 5689–5701.

(55) Dzib, E.; Merino, G. The Hindered Rotor Theory: A Review. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2022**, *12*, e1583.

(56) Møller, K. H.; Otkjær, R. V.; Hyttinen, N.; Kurtén, T.; Kjaergaard, H. G. Cost-Effective Implementation of Multiconformer Transition State Theory for Peroxy Radical Hydrogen Shift Reactions. *J. Phys. Chem. A* **2016**, *120*, 10072–10087.

(57) Zhao, Q.; Møller, K. H.; Chen, J.; Kjaergaard, H. G. Cost-Effective Implementation of Multiconformer Transition State Theory for Alkoxy Radical Unimolecular Reactions. *J. Phys. Chem. A* **2022**, *126*, 6483–6494.

(58) Yu, T.; Zheng, J.; Truhlar, D. G. Multi-Structural Variational Transition State Theory. Kinetics of the 1,4-Hydrogen Shift Isomerization of the Pentyl Radical with Torsional Anharmonicity. *Chem. Sci.* **2011**, *2*, 2199–2213.

(59) Pedley, J. B., *Thermochemical Data and Structures of Organic Compounds*; CRC Press: 1994; Vol. 1.

Chapter 5

Fast Predictions of Reaction Barrier Heights: Towards Coupled-Cluster Accuracy

Much of this work has previously appeared as Spiekermann, K. A.; Pattanaik, L.; Green, W. H. Fast Predictions of Reaction Barrier Heights: Toward Coupled-Cluster Accuracy. *J. Phys. Chem. A* **2022**, *126*, 3976–3986 [link]. Lagnajit Pattanaik assisted with analyzing model results. All code and model weights are freely available at https://github.com/kspieks/chemprop/tree/barrier_prediction.

5.1 Introduction

Accurately predicting the time evolution of reacting chemical systems and the yields of various products and side products has been one of the main projects of physical chemistry for more than 135 years [1, 2]. Some important systems have been very heavily studied, leading to chemical kinetic model predictions so compelling that they drive governmental and business decisions, and have even led to major international treaties (e.g., the Montreal Protocol) [3, 4]. If a large number of data are available, one can make useful predictions, e.g., regarding which organic synthesis routes are likely to succeed at specified reaction conditions [5, 6]. However, data relevant to a particular system of interest are usually scarce, and it is often impractical to do enough experiments to develop a predictive model with the desired accuracy or generalizability.

Over the last ∼25 years, it has become possible to accurately compute the rate coefficients of individual reactions using quantum chemistry and rate theory [7, 8]. This suggests it might be possible to accurately predict the behavior of a reacting system even before any experimental data are available on that system, with many ramifications [9]. Indeed, recently several research groups have constructed models to quantitatively predict the time-evolution of diverse chemical systems, based largely on parameters derived from quantum chemistry rather than experiment. For example, earlier this year our group has published models for combustion [10], pyrolysis [11, 12], and the degradation of pharmaceutical compounds [13]. However, because quantum chemistry calculations of rate coefficients are slow, only a small fraction of the numerous rate coefficients in these models have been computed accurately, which can make the predictions somewhat erratic. The vision of reliable predictive chemical kinetics can only be realized if accurate rate coefficients can be calculated much more rapidly than is possible today.

The traditional workflow to obtain kinetic parameters, shown in Figure 5.1, is quite involved [7, 8]. In summary, the workflow starts by creating 3D structures for the reactant(s) and product(s), which are optimized using quantum chemical methods. The search for a good 3D structure for the transition state (TS) often consumes the most human time, since automated saddle point finders are not yet sufficiently reliable [14, 15]. After the most promising TS structure is identified, it is optimized using a more accurate method, and often an intrinsic reaction coordinate (IRC) calculation is run to confirm it connects the reactant(s) and product(s) [16]. Then the energy of reactant(s) and TS is re-computed at a high-level of theory (e.g., coupled-cluster) with zero-point energy corrections to provide an accurate reaction barrier. If the studied structures are flexible with many rotatable bonds or other large-amplitude motions, conformational effects must also be considered. Usually, this is done by approximating each hindered internal rotor as independent [17–19], but often this approximation is not accurate, requiring some treatment of the coupling between different rotors [20]. Alternative methods which involve multiple conformational minima for reactants and transition states and couple conformational effects with low frequency anharmonic torsional modes–including multistructure methods developed by Truhlar and coworkers–are becoming popular, but it can be challenging to find all the conformers [21–24]. Finally, one can calculate the partition functions, which canonical transition state theory (TST) uses to estimate the high-pressure limit rate coefficient $k_\infty(T)$, including effects of symmetry/reaction path degeneracy, tunneling, and other corrections [8, 25–27].

Given the importance and impact of kinetic models, several efforts have been made to accelerate portions of the traditional workflow in Figure 5.1, either by accelerating individual steps, or by skipping over steps, illustrated as arrows in the interior of the figure. Many programs have been published for rapid conformer generation of stable species [24, 28–31]. Several methods also exist for generating good TS guesses, such as using hand-made templates [32–34] or deep learning [35–37].

Despite all these advances, calculations of rate coefficients are still limited by the accurate yet expensive quantum methods used to find the saddle point geometry, and the poor scaling of the even more expensive methods used to compute its energy. For example, coupled-cluster CCSD(T)-F12 calculations are commonly considered the gold-standard in quantum chemistry due to their reliable single point energies and reaction energies [38–43], yet they scale as $\mathcal{O}(N^7)$, where N is the number of orbitals [44]. For some reactions, even CCSD(T) is not accurate enough, and more expensive calculations are required [8].

Since so many computer resources and human efforts are required to obtain each reliable value, especially for reactions of large molecules, it would be advantageous to directly estimate kinetic parameters instead. Indeed, established methods for directly estimating barrier heights or $\log(k_\infty(T))$ from reactant and product identities include simple models such as Evans-Polanyi relationships [45] and the Hammett correlations [46]. In his textbook, Benson presented simple methods to predict Arrhenius A-factors [47]. Several authors have extended Benson's popular thermochemical group additivity approach [47–49] to estimate activation energies [50, 51], A-factors [52], and rates [53]. Unfortunately, linear models are an inherently limited representation, so the simple Benson-type groups that work so well for predicting the thermochemistry of many organics are not sufficient for predicting rate coefficients over a broad range of reactions. Much larger supergroups [54] or decision trees [55–57] are needed to reach acceptable accuracy. Still, the supergroups and decision trees are often manually defined for each reaction type, which is tedious and error-prone. Thus, there has been interest in automating tree construction [58, 59] to facilitate the incorporation of new training reactions, but the process is still cumbersome.

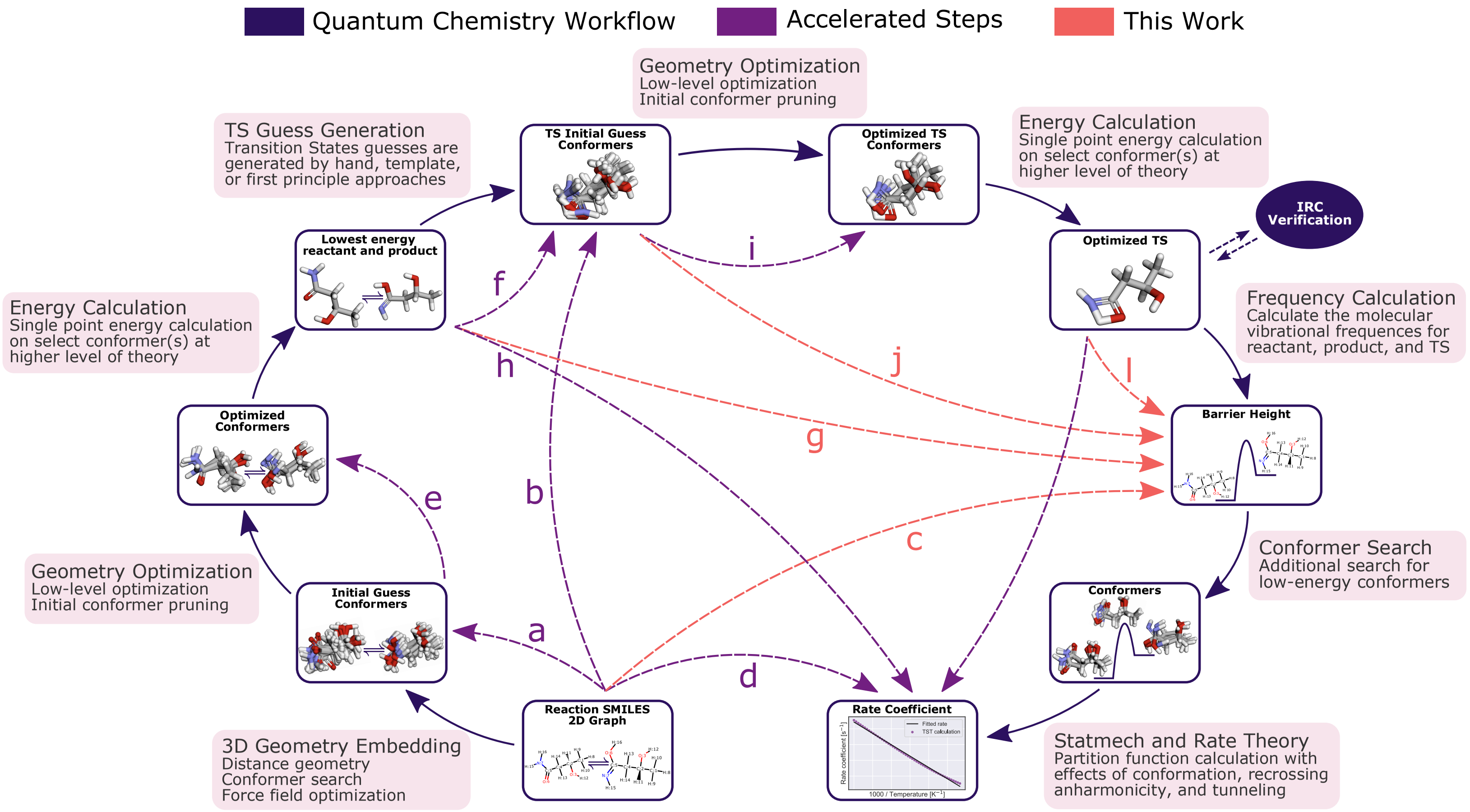

Quantum Chemistry Workflow
Accelerated Steps
This Work
Geometry Optimization
Low-level optimization
Initial conformer pruning
TS Guess Generation
Transition States guesses are generated by hand, template, or first principle approaches
TS Initial Guess Conformers
Optimized TS Conformers
Energy Calculation
Single point energy calculation on select conformer(s) at higher level of theory
IRC Verification
Optimized TS
Lowest energy reactant and product
Frequency Calculation
Calculate the molecular vibrational frequences for reactant, product, and TS
Energy Calculation
Single point energy calculation on select conformer(s) at higher level of theory
Optimized Conformers
Barrier Height
Geometry Optimization
Low-level optimization
Initial conformer pruning
Conformer Search
Additional search for low-energy conformers
Initial Guess Conformers
Conformers
Reaction SMILES 2D Graph
Rate Coefficient
Fitted rate
TST calculation
Rate coefficient [s⁻¹]
1000 / Temperature [K⁻¹]
3D Geometry Embedding
Distance geometry
Conformer search
Force field optimization
Statmech and Rate Theory
Partition function calculation with effects of conformation, recrossing anharmonicity, and tunneling
a
b
c
d
e
f
g
h
i
j
l

Figure 5.1. The conventional quantum chemistry and TST workflow for predicting a high-pressure limit rate coefficient is shown in the outer loop with dark purple arrows. Several methods have been proposed in literature (and in the present work from this chapter, shown with orange arrows) to accelerate steps in the process: a) Initial conformer generation ref. [24, 28–31, 60–62] b) TS guess generation ref. [33] c) Estimate barrier height from 2D representation ref. [45, 50, 56, 58, 63–66] d) Estimate rate coefficients from 2D representation ref. [52, 53, 67–69] e) Semi-empirical optimization ref. [70] f) TS guess generation ref. [34–37, 71–74] g) Estimate barrier height from 3D representation ref. [75, 76] h) Estimate rate coefficient from 3D representation ref. [77, 78] i) Accelerate TS optimization ref. [79–81] j) Estimate barrier height from an un-optimized TS guess k) Estimate rate coefficient from an optimized TS ref. [82] l) Estimate barrier height from an optimized TS ref. [83, 84]

Recent work has focused on other nonlinear methods to predict kinetic parameters. For example, Heinen et al. [66] trained a kernel ridge regression (KRR) model to predict activation energies. The model was fit to their previously generated dataset [83] of S_N2 and E2 reactions with single point energies computed at DF-LCCSD/cc-TZVP. Stuyver et al. [64] trained a graph neural network (GNN), augmented with quantum descriptors, on the same dataset. In both cases, random splits gave a testing MAE of $\sim$2.5 kcal mol^{-1}. The relatively low error is expected since the test set had a similar composition as the training set i.e., it is a measure of interpolation. To assess generalizability, Stuyver et al. also restricted training to three of the four nucleophiles in the dataset, reserving the fourth only for the test set. Although their augmented GNN extrapolated better than KRR to the unseen nucleophile, both models showed significantly higher testing errors relative to random splitting. Further, since all reactions start from a substituted ethyl-based scaffold, it is unclear how well these models would generalize to other reactants.

Estimators for activation energy have also been applied to catalysis. Takahashi et al. [76] trained several types of models (linear, random forest, and support vector regression) to predict the activation energy for a small dataset of heterogeneous catalytic reactions. When using a random split, testing errors were $\sim$0.90 eV ($>$20 kcal mol^{-1}) for all models, much too large to be useful. Similarly, Singh et al. [75] trained a feed forward neural network (FFN) model to predict reaction barriers for a small dataset containing dehydrogenation reactions as well as N_2 and O_2 dissociations. They report a testing error of 0.22 eV when using a random split.

Komp and Valleau [85] trained two different FFNs to predict the natural logarithm of the "partition function" at a given temperature. The first model takes the molecular geometry and inverse temperature as input to predict the "partition function" of a molecule. The second model uses the geometry and "partition function" of the reactant and product as well as the inverse temperature to predict a "partition function" of the TS. Unfortunately, the quantities used by those authors for training and testing their model were not the true partition functions that appear in rate theory, but instead rigid-rotor harmonic-oscillator (RRHO) partition functions of a single conformer of the TS, reactant or product. However, the great majority of the species in their data set have rotatable bonds and a large number of low-energy conformers, and many of the "product" structures and vibrational frequencies they used were actually properties of van der Waals complexes of two or three product molecules. Partition functions are drastically affected by conformers and rotors [20–24], which is why the refined dataset from ref. [43] published high-pressure limit TST rate coefficients for only a small subset of the reactions–those where the TS and the reactant(s) were rigid so there was only one low-lying conformer. Further, in the Komp and Valleau paper, no analysis was provided for how sensitive the model predictions are to the input 3D geometry; as shown in our results section below, this could be a significant problem.

Another focus is using machine learning (ML) methods to directly estimate $k_\infty(T)$. Houston et al. [77] used Gaussian process regression (GPR) on 13 reactions to predict bimolecular rate constants over a large temperature range. As a follow up, Nandi et al. [78] clustered the input data, retrained the GPR model on each cluster, and then predicted rate coefficients for the $O(^3P)$ + HCl reaction. In other work, Komp et al. [82] trained a FFN on ∼1.5 million data points to predict the product of $k_\infty(T)$ with the reactant partition function for 1D barrier problems. FFNs have also been fit to the rate coefficients from small datasets of hydroxyl radical reactions [67, 68] and ionic liquids [69]. In all these cases, the data sets employed did not cover much reaction space.

Still another approach is to use Δ-ML, a technique developed by Ramakrishnan et al. [86] in which a model is fit to the residuals between a high- and low-level of theory. Thus, rather than directly predicting a high-quality value for the parameter of interest, the Δ-ML model predicts the correction to the low-level value, which should be relatively inexpensive to calculate. In the context of kinetics, Bragato et al. [84] showed that using KRR for Δ-ML outperformed direct ML approaches for predicting barrier heights from von Rudorff et al.'s [83] S_N2 dataset.

In the remainder of this paper, we present a new model for estimating the zero-point energy corrected barrier height from either 2D or 3D structures of the reactants and products. Many of the ML models summarized here, as well as in recent reviews [87–89], are trained on relatively small datasets from a specific reaction family, which severely hinders their generalizability. In contrast, our goal is to develop a deep learning model that quickly predicts accurate barrier heights for a diverse set of reactions. Our model uses information from only the reactant and product so future workflows could avoid the often delicate and challenging task of finding the TS geometry. This model would be useful for directly estimating barrier heights during the automated generation of kinetic models [56]; it would also offer substantial speedup when refining existing kinetic models by avoiding the need for a TS geometry. Lastly, this model would allow for quantitative ranking of potential reactions produced from an automated enumeration [14, 15, 90, 91].

Our work brings several advances over a recent study, which trained a graph neural network to predict barrier heights on a similar dataset calculated at a lower level of theory

[92]. First, our model is trained on a refined dataset that allows barrier height predictions to approach coupled-cluster accuracy. As shown below, our new model's predictions are about 1 kcal mol^{-1} more accurate than a good density functional theory (DFT) calculation and offer a $\sim 10^5$ factor speedup. Our model also uses an improved reaction representation, which substantially improves the accuracy, and it uses 75% fewer parameters than models from previously published works [63, 93], making our lightweight model more practical to integrate into prediction workflows. Importantly, our training procedure uses proper data splits to accurately estimate the model's performance on unseen reactions. Finally, our work also compares performance of models using 2D and 3D information, highlighting opportunities for future innovation.

5.2 Methods

5.2.1 Dataset

To train our model, we leverage a recently refined gas-phase dataset of elementary reactions with atom-mapped SMILES [43, 94]. These data span a diverse set of reactions, whose neutral molecules involve carbon, hydrogen, nitrogen, and oxygen and contain up to seven heavy atoms. Reactions are available at three levels of theory: B97-D3/def2-mSVP, ωB97X-D3/def2-TZVP, and CCSD(T)-F12a/cc-pVDZ-F12//ωB97X-D3/def2-TZVP. Data from all levels of theory are used in a transfer learning approach, which first trains the model on a larger dataset with lower accuracy and then fine-tunes the model with a slightly smaller dataset with higher accuracy calculations. All regression targets are z-scored during training for numerical stability. Similar to in ref. [63] and [95], the model is initialized with the final weights from the previous run and then uses a smaller learning rate. The final model is evaluated against the coupled-cluster values. At each level, the zero-point energies from the harmonic vibrational analysis were added to the reactant, product, and TS energies. These energies used scaled vibrational frequencies to account for anharmonic effects [96, 97]. The barrier heights were calculated by subtracting the resulting TS and reactant energies. Reaction enthalpies were calculated by subtracting the resulting product and reactant energies, which include bond additivity corrections as described in ref. [43]. These reactions largely overlap with those used to train the model in ref. [63]. However, ref. [43] refined the energies of each species using coupled-cluster calculations. All of the earlier works fitting these reactions [63, 85, 93] only had access to DFT energies. Suspecting convergence errors in the quantum chemistry, we removed any reaction in which the coupled-cluster reaction enthalpy is over 10 kcal mol^{-1} above the barrier height.

All reactions contain one reactant and one to three products. All the transition states and unimolecular reactants have an even number of electrons, and were computed as singlets (S=0). Almost all of the products also have an even number of electrons. However, there are 49 reactions unique to the larger B97-D3 dataset whose product was a pair of radicals (each product had an odd number electrons). We modeled these odd-electron species as doublets (S=$\frac{1}{2}$). We augment our data by including the reverse reactions by using the following expressions

$$\Delta E_{0,\text{reverse}} = \Delta E_{0,\text{forward}} - \Delta H_{\text{forward}} \tag{5.1}$$

$$\Delta H_{\text{reverse}} = -\Delta H_{\text{forward}} \tag{5.2}$$

which results in approximately 33,000 reactions at B97-D3 and 24,000 reactions at both the ωB97X-D3 and CCSD(T)-F12a levels.

When creating training, validation, and testing sets, we use five folds, each with an 85:5:10 split. To create these sets, we use a scaffold split on the reactant SMILES, which partitions the data based on the Bemis-Murcko scaffold [98] as calculated by RDKit [99]. Scaffold splits are a better measure of generalizability compared to random splits [93, 100–104]. When assigning reactions to each set, we place each pair of forward and reverse reactions in the same set. Otherwise, if we separate forward and reverse reactions between sets, the model's testing error will be polluted by data leakage–the same transition state would appear in both the training and testing set–and so will not reflect the true performance of the model when evaluating a new reaction [93].

5.2.2 2D D-MPNN Model

When developing data-driven methods to predict barrier heights, graph neural networks (GNNs) are a natural choice; here, molecules are abstracted as graphs where atoms are graph nodes and bonds are graph edges [105]. GNNs operate by updating representations of nodes or edges with the information from neighboring nodes or edges, propagating information throughout the graph. For our work, we adapt Chemprop [100], a directed message passing neural network (D-MPNN), which is a type of GNN that passes messages across directed bonds. Like other GNNs, the D-MPNN architecture builds a learned molecular representation by aggregating atomic representations after the message passing phase and feeds this representation through a dense layer to predict the molecular property of interest.

Extending the D-MPNN architecture to reaction properties is an intriguing way to model barrier heights. Grambow et al. [63] altered Chemprop to predict reaction properties by embedding the reactant and product with the same D-MPNN, subtracting the learned atomic representations, and passing the aggregated molecular representation through a FFN to obtain the final prediction. Heid and Green [93] showed that using the established condensed graph of reaction (CGR) representation [106] improved Chemprop's performance on barrier height prediction. A key aspect of the CGR representation is that it removes disjoint graphs present in multi-molecular reactions and allows message passing between all atoms, overcoming a limitation of the Grambow et al. method. Since CGR is a superposition of the reactant and product graphs, only one graph is input to the model, leaving the rest of Chemprop's architecture unchanged.

Here, we extend the CGR version of Chemprop by incorporating additional atom and bond features. We also concatenate additional features before the dense layer readout, such as reaction enthalpy, to improve prediction performance. A schematic of the model architecture is shown in Figure 5.2. Table 5.2 shows the improvement from each modification, and Section 5.5.1 in the Appendix contains more detail about the network, training procedure, and hyperparameter optimization. Our modified code and final weights are freely available on GitHub [107].

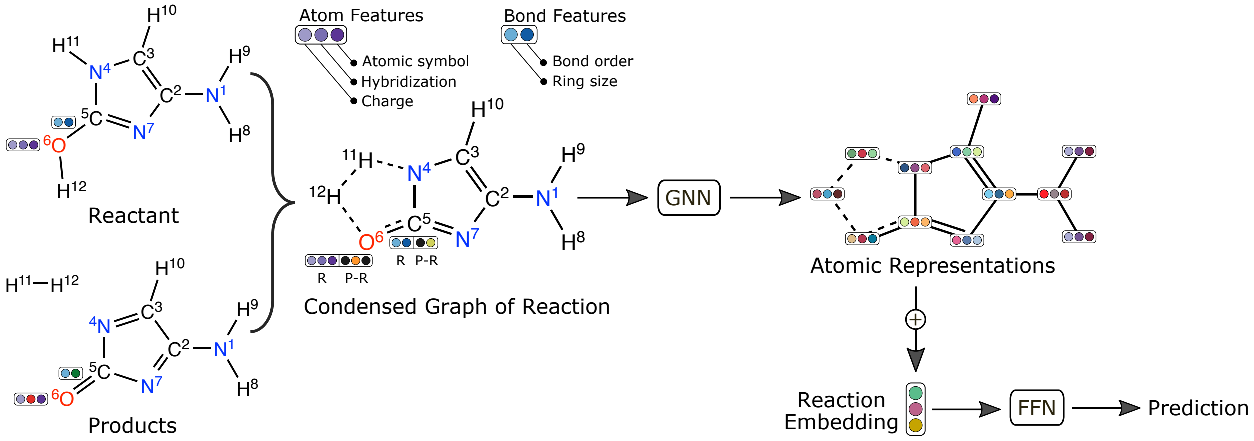


Figure 5.2. Schematic of the machine learning model architecture. Here, the GNN uses the directed message passing scheme introduced by Yang et al. [100] The condensed graph of reaction (CGR) representation [93, 106] is used to predict the reaction's barrier height and corrected enthalpy. The CGR concatenates the featurization of the reactant(s) with the difference of the featurization between the reactant(s) and product(s). A subset of the initial atom and bond features are shown for simplicity. The superscripts indicate the arbitrary atom-map numbers. Colors are for qualitative purposes only.

5.2.3 3D DimeReaction Model

Since models based on molecular geometries have worked well for property prediction [108], neural network potentials [109], and excited-state dynamics [110], we also explored 3D networks for predicting the reaction barrier. We start with DimeNet++ [111], which is a faster and more accurate version of the original DimeNet model [112]. Both are directional message passing networks that operate on 3D coordinates and have shown promising results for property prediction of individual molecules, such as on the popular QM9 dataset [113]. Here, we extend the architecture so that it can predict reaction properties. We follow a similar approach from Grambow et al. [63] by passing the reactant and product through the same DimeNet++ model, subtracting the learned molecular representation, and passing the result to a dense layer to predict the regression target. Our modified code and final weights are freely available on GitHub [114].

Our DimeReaction model is trained on the unimolecular reactions from the dataset. Since the model operates on molecular coordinates, reactions with multiple products should not be arbitrarily aligned in 3D space. However, properly translating and rotating those molecules to form a multi-product complex is beyond the scope of this work. Further, the unimolecular reactions account for about 70% of the data, giving nearly 17,000 reactions after augmenting with the reverse reactions which are also unimolecular. As before, a scaffold split is used to create the training, validation, and testing sets with an 85:5:10 split, and each pair of forward and reverse reactions is placed in the same set. The model is trained on only one fold of the data since preliminary results indicated sub-par performance relative to the 2D-MPNN model. Additional detail can be found in Section 5.5.2 the Appendix.

5.2.4 Baseline Models

We also include simpler models from Scikit-Learn [115] as baselines, such as random forest (RF), multi-layer perceptron (MLP), and support vector regression (SVR) with both linear and radial basis function (RBF) kernel. We also use extreme gradient boosting (XGB) from the XGBoost package [116]. The best hyperparameters are chosen via Optuna [117] and are listed in Section 5.5.3 in the Appendix. Since these methods operate on vector inputs, we use Morgan (ECFP) fingerprints [118, 119] with 2048 bit hashing and radius of 2 as calculated using RDKit [99]. To represent the reaction, we concatenate the fingerprint of the reactant complex with the difference of the fingerprint between the reactant and product complex; this is the same approach as the initial featurization in the CGR for the 2D D-MPNN model.

The final baseline is a model that simply predicts the mean of the training target values for all test reactions. Comparing against such a simple baseline provides helpful context, especially if dealing with relatively homogeneous datasets whose narrow target distribution inherently lead to low test errors that may be mistaken for satisfactory performance [120]; given that the barrier heights span an enormous range from 0 to 200 kcal mol^{-1} for the dataset used in this work, this is less of a concern and is simply done out of best practices.

5.3 Results and Discussion

5.3.1 Baseline Models

We first compare the performance of the classical ML methods. We also include the trivial baseline model that predicts the mean of the training target value for all test reactions. The results from Table 5.1 are generally quite disappointing. Even the best performing model, MLP, gives a test MAE of 9.93 kcal mol^{-1}. Although such a model may be useful for extremely rough qualitative analysis, it most likely will not be useful for quantitative analysis since an uncertainty on the activation energy of this magnitude implies that rate coefficients would differ on average by a factor of 150 at 1000 K and by an enormous factor of 17 million at 300 K.

Table 5.1. Testing errors (mean ± 1 standard deviation from the five folds) from various models for predicting ΔE_0 when using scaffold splits. MAE and RMSE have units of kcal mol^{-1}.

Model	MAE	RMSE	R^2
Baseline (mean)	25.08 ± 0.31	31.04 ± 0.28	0.00 ± 0.00
SVR (RBF)	16.02 ± 0.25	20.00 ± 0.21	0.58 ± 0.01
SVR (Linear)	14.20 ± 0.36	18.63 ± 0.48	0.64 ± 0.01
RF	14.14 ± 0.36	19.02 ± 0.49	0.62 ± 0.02
XGB	11.81 ± 0.25	15.72 ± 0.40	0.74 ± 0.01
MLP	9.93 ± 0.19	14.10 ± 0.42	0.79 ± 0.01

5.3.2 2D D-MPNN Model

The testing mean absolute error (MAE) is 2.89 ± 0.16 kcal mol^{-1}, the root-mean-square error (RMSE) is 4.90 ± 0.16 kcal mol^{-1}, and the coefficient of determination (R^2) is 0.975 ±

0.001, such that the bounds correspond to one standard deviation calculated across five folds. The parity plot in Figure 5.3a shows the model's predictive power across the entire range of data; accuracy is maintained even in regions where the data are sparser. The residuals are centered around zero, indicating no systematic over- or under-prediction, which is further supported by the error histogram in Figure 5.3b. 95% of the reaction barriers are predicted within 6 kcal mol^{-1} of the calculated coupled-cluster value.

The overall testing performance is improved by using ensembles of models, which is consistent with previous literature [100, 121]. For a given fold, different initializations are used when training a model on that data split. The predictions are then averaged from each ensemble and compared to the target values from the respective test set. Here, five different initializations are used for each of the five data splits, resulting in a total of 25 models. The weights for all 25 models are published on GitHub [107]. Ensembling lowers the testing MAE (RMSE) to 2.57 ± 0.13 (4.58 ± 0.16) kcal mol^{-1} and slightly increases the R^2 to 0.978 ± 0.001.

It is also important to examine model performance specifically on low barrier reactions since these are the most feasible reaction pathways included in kinetic models. For example, filtering the testing reactions to those with $\Delta E_0 < 50$ kcal mol^{-1} gives a testing MAE (RMSE) of 2.07 ± 0.18 (3.76 ± 0.33) kcal mol^{-1}, which is a slight improvement compared to the test errors for all reactions. The subsequent analysis and all figures and tables use weights from the first initial seed within each fold since the purpose is simply to explain interesting trends rather than obtain the best predictions for each intermediate case.

Figure 5.3c shows that the model strongly benefits from additional training data. The learning curve (how the model improves as more training data is added) has not yet shown asymptotic behavior, which implies further improvement may be possible with more data. The slope of the learning curve is similar to those we have observed when developing models for other chemical properties using Chemprop [122]. To ensure that the only variable changing was dataset size, each plotted point uses the exact same reactions and data splits for all three quantum chemistry datasets (one for each level of theory) for the transfer learning scheme.

As described in ref. [43], a barrier height calculated at ωB97X-D3 has an MAE of about 3.5 kcal mol^{-1} from the coupled-cluster value. Thus, this is the ideal prediction error that one could expect when training a model only on DFT data. Indeed, the previous model from Grambow et al. [63], which was trained to predict ωB97X-D3 values, has an MAE of 4.74 ± 0.10 kcal mol^{-1} relative to the coupled-cluster values (approximately twice as large

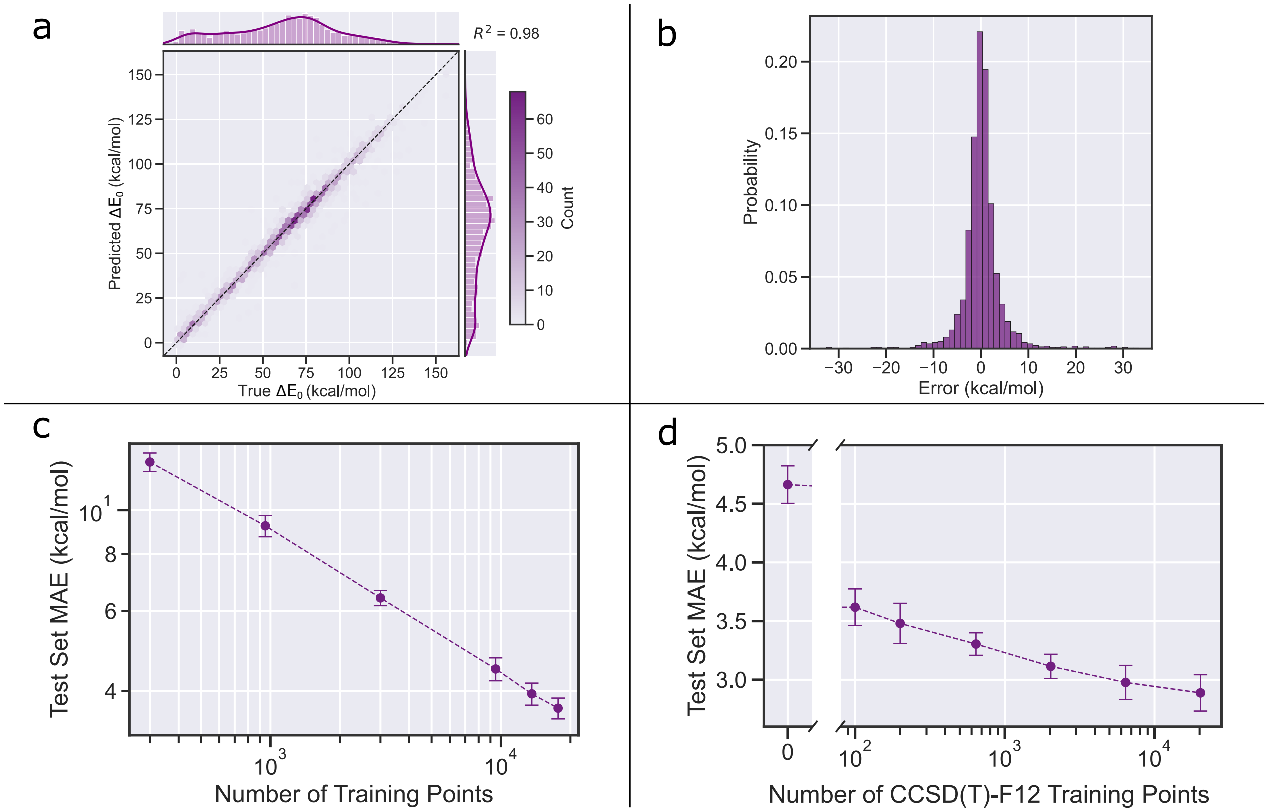


Figure 5.3. Deep learning model results for predicting reaction barriers. Error bars indicate one standard deviation calculated across the five folds. (a) Parity plot of model predictions vs "true" (CCSD(T)-F12a) barriers ΔE_0 for the first fold. (b) Histogram of testing errors (predicted minus "true") for the first fold. (c) Testing MAE vs. the number of data points at each stage of training. These reactions are present in all three level of theory datasets. (d) Testing MAE vs. number of coupled-cluster training reactions. Each point shows a model pretrained on all B97-D3 and ωB97X-D3 reactions and then fine-tuned with increasing amounts of high-accuracy data.

as the model error from this work). Thus, Figure 5.3d is particularly exciting since it shows that even a small amount of fine-tuning gives a meaningful improvement to our model's predictions. Each plotted point shows a model pretrained on both the B97-D3 and ωB97X-D3 datasets and then fine-tuned with increasing amounts of high accuracy data. The same coupled-cluster validation and test set is used during each run. Although the model certainly benefits from additional coupled-cluster values, performing such calculations, which scale as $\mathcal{O}(N^7)$, may not always be practical for large datasets or for large molecules. However, we find that using these high-accuracy calculations for even just 50 reactions (augmented with the reverse to yield 100 total training reactions), lowers the testing MAE by over 1 kcal mol^{-1}. Similar behavior was observed in some of our prior transfer learning studies, where models built on large DFT data sets were fine-tuned using a small number of more accurate data [95, 122]. We believe this works because the model trained from the DFT data on a large number of reactions has learned how a wide variety of chemical structures affect the barrier height. But the DFT numbers are rough and benefit from calibration using a few high-accuracy numbers. We suggest this type of fine-tuning may be a good strategy whenever DFT gives reasonable predictions of a chemical property, one can afford to generate a large DFT data set, and one has a smaller high-accuracy data set.

Table 5.2 demonstrates how various factors contribute to improving model performance. The biggest improvement comes from optimizing the hyperparameters (see Table 5.6 in the Appendix) and allowing enough degrees of freedom for the model to capture the diverse chemistry in this dataset. Adding ring size to both the atom and bond features further improves the model. We also use tried using RDKit [99] to generate molecular feature vectors [123] for each SMILES and pass the difference of the product and reactant vector as input to the dense layer. However, this does not improve results. Some recent work has shown that using quantum mechanical descriptors can improve model performance for reaction prediction. Similar to both Stuyver and Coley [64] and Guan et al. [102], we use the D-MPNN network developed by Guan et al. [102] without modification to predict atomic descriptors that are concatenated with the learned atomic representations before the aggregated molecular representation is passed to the dense layer. However, we observe no effect of using these descriptors. Finally, using reaction enthalpy as an input to the dense layer, rather than co-training on both the barrier height and enthalpy as in ref. [63], further improves the performance. This approach requires quantum calculations for the reactant and product to obtain the reaction enthalpy; however, this is often easier than obtaining a TS, so our model still offers substantial time savings compared to the traditional workflow outlined in Figure 5.1. Alternatively, the predicted enthalpy from our co-trained model could serve as input to the final model, which offers further time savings and only minimal increases in testing error (see Table 5.7 in the Appendix).

Table 5.2. Sequential optimization study showing the improvement in testing error (kcal mol^{-1}). Each row includes all changes from the previous row. Results are from the first fold of cross-validation.

	B97-D3		ωB97X-D3		CCSD(T)-F12a	
Description	MAE	RMSE	MAE	RMSE	MAE	RMSE
Default	6.74	10.22	5.30	8.17	5.07	8.18
+Optimize hyperparameters	5.70	9.29	3.61	6.08	3.44	5.86
+Ring features	5.28	8.79	3.33	5.77	3.24	5.66
+Input ΔH	**4.81**	**7.88**	**3.06**	**5.15**	**2.98**	**4.96**

As mentioned above, creating proper data splits is essential to evaluating model performance. We use a scaffold split on the reactants from the forward reactions to create training, validation, and testing sets. The reverse reaction is then added to the corresponding set, which ensures that each set is independent. In contrast, the previous work from Grambow et al. [63] performed a scaffold split on the reactants from the forward and reverse reactions. However, since the reactant and product of the same reaction can have different scaffolds, the forward reaction could be placed in the training set while the reverse reaction could be placed in the test set. Indeed, ∼80% of the testing reactions from Grambow et al. had their forward or reverse counterpart in the training set. This data leakage resulted in underestimating the true testing error since the regression task had accidentally been reframed as an easier problem [93]. Rather than asking the model to produce a barrier height for a new reaction, the data leakage meant that the model was mostly predicting a barrier height for a reaction it has already seen in either the forward or reverse direction. The dramatic impact on model performance is shown in Table 5.3.

Table 5.3. Importance of independent data splits on ΔE_0 testing errors (kcal mol^{-1}). Results are from the first fold of cross-validation.

		Proper	B97-D3		ωB97X-D3		CCSD(T)-F12a	
Model	**Hyperparameters**	**Splits**	MAE	RMSE	MAE	RMSE	MAE	RMSE
Grambow et al.	Grambow et al.	No	2.99	5.87	1.91	3.35	1.87	3.29
Grambow et al.	Grambow et al.	Yes	8.15	11.46	5.94	8.55	5.50	7.97
Grambow et al.	This work	Yes	7.16	10.85	5.13	7.71	4.31	6.93
This work	**This work**	**Yes**	**4.81**	**7.88**	**3.06**	**5.15**	**2.98**	**4.96**

5.3.3 3D DimeReaction Model

One hypothesis is that using 3D information may improve the predictive capabilities relative to using a 2D attributed graph. Since obtaining a TS geometry is difficult, our ideal DimeReaction model would receive an optimized structure for the reactant and product and then predict a barrier height. We find that model performance is better when only training on the coupled-cluster data, rather than performing transfer learning as described earlier. Still, results from this approach were quite poor with a barrier height testing MAE (RMSE) of 6.20 (10.10) kcal mol^{-1}. Although this large error is surprising, additional tests indicate that the DimeNet++ model functions as intended. For example, training a multi-task model that predicts both ΔE$_0$ and ΔH yields excellent performance on reaction enthalpy with a testing MAE of 1.78 kcal mol^{-1}. Thus, when the regression target is a function of the input geometries, such as using reactant and product structures to predict reaction enthalpy rather than the barrier height, the model performs well. Additional results and analysis are provided in Section 5.5.2 in the Appendix.

Another potential workflow is using the difference between the learned representations of the reactant and TS geometries (rather than the reactant and product) since these directly correspond to a barrier height. To test this idea, we first consider the ideal scenario of using optimized TS structures. As expected, this results in a very low testing error—lower than our best 2D model—shown as the first row in Table 5.4. While it is chemically satisfying to see the reactant and TS encode more information about ΔE_0 than the reactant and product, using an optimized TS geometry here is not sensible, at least not for molecules in this size range. Once one has the optimal TS geometry, it is not expensive to perform a single-point coupled-cluster calculation and then proceed with canonical TST to compute $k_\infty(T)$.

A more practical workflow involves quickly generating a TS guess and training a model with this non-optimized structure. Several methods exist for generating TS guesses, from using hand-made templates [32–34] to using deep learning [35–37], so it should be quite practical to quickly obtain a TS guess. To first get an idea of how much a non-optimized TS structure would impact the results, we add increasing amounts of Gaussian noise to the optimized structures that were then used as input to the model. As seen in Table 5.4, the model is very sensitive to the TS structure, as testing performance quickly decreases when adding noise. We next use one TS guess predictor to quickly generate TS guesses. The graph network from Pattanaik et al. [35] had achieved a testing RMSD of 0.28 Å. Using TS guesses generated by this network as input to our DimeReaction model, we obtain a testing MAE (RMSE) of 6.19 (9.26) kcal mol^{-1}, which is inline with the trend from Table 5.4.

Table 5.4. Impact of noisy TS inputs on DimeReaction's testing errors (kcal mol^{-1}). The first row corresponds to using an optimized TS structure with no added noise.

σ	RMSD (Å)	MAE	RMSE
0	0	2.25	3.91
0.05	0.086	3.62	5.66
0.10	0.173	4.81	7.16
0.15	0.260	6.28	8.83
0.20	0.346	7.42	10.42

The sensitivity of the 3D reaction model renders it impractical compared to the 2D workflow. However, it highlights an area of further research; future studies should investigate methods to quickly and accurately produce 3D TS guesses, which can subsequently be used in predicting kinetic parameters. Currently, available estimators achieve geometric errors on TS structure generation that are much too high for such modeling efforts. As another idea, the sensitivity could make DimeNet++ quite useful for screening which initial guess geometries are worth the computational expense for optimization and frequency calculation. It may also prove beneficial for screening conformers to identify low-energy structures.

5.4 Conclusion

To accurately predict the time evolution of a reacting chemical system, one needs a quick way to estimate rate coefficients to decide which of the many conceivable reactions are important enough to include in the kinetic model. Due to the high computational cost, it is usually impractical to directly compute $k_\infty(T)$ for all of the reactions in a system of interest. Methods are needed to accelerate or bypass some of the computational bottlenecks in the conventional TST workflow based on high-accuracy quantum chemistry. Even rough estimates can be helpful in deciding which reactions require expensive calculations. The need for fast estimates of rate coefficients has been recognized for many decades, and excellent estimators have been developed for certain reaction families. However, data scarcity has made it difficult to develop models that generalize well over a broad range of reactions. With recent advances in compute power, we can now routinely generate large datasets with DFT and sometimes higher-level quantum chemistry methods, opening up the possibility of constructing accurate estimators with much broader scope using machine learning techniques.

For instance, the generation of new high-quality kinetics datasets [43, 92] allowed the 2D D-MPNN model presented here to quickly predict reliable values for barrier heights across a diverse set of thousands of gas-phase reactions. The model gives more accurate barrier heights on average than good (e.g., ωB97X-D3/def2-TZVP) DFT calculations for vastly lower computational cost. By directly estimating barrier heights, we anticipate the model will be useful during the automated generation of kinetic models by reducing the amount of computational effort devoted to reactions which cannot be important. The new model can also quickly rank the most feasible (lower barrier) reactions produced from an automated enumeration of possible reactions, making the process of discovering new low-barrier reactions much more efficient. Combining this barrier height estimator with other methods for computing or estimating Arrhenius A-factors would allow rapid estimation of reasonable rate coefficients.

Going forward, existing kinetics datasets should be expanded. As shown in Figure 5.3c, additional data would be beneficial for reducing the errors in predicted barrier heights. Although the dataset from Spiekermann et al. [43] is the largest coupled-cluster reaction barrier dataset to our knowledge, it is still relatively limited; its species contain at most only seven heavy atoms, they only include the elements H, C, N, and O, all transition states are closed shell singlets, no ions are included, and there are only gas-phase reactions. The dataset used in this work also contains only a single conformer for each species, so some of the prediction error may be due to not using the lowest energy conformers when calculating the barrier height. Future dataset generation should perform thorough conformer searches. This would also facilitate computation of $k_\infty(T)$ for reactions where the reactant(s) and/or TS are non-rigid e.g., because they contain internal rotors.

When it comes to training models, using proper data splits is essential to evaluating model performance. As shown by Table 5.3, each pair of forward and reverse reactions must be placed in the same set to ensure that the training, validation, and testing sets are independent. Otherwise, the model's testing error will appear unrealistically low since it has been polluted by data leakage. Scaffold splitting is also considered a better measure of generalizability [93, 100–104]. In contrast, random splitting creates an easier prediction task that is less useful at measuring extrapolation capabilities. Finally, we emphasize the importance of fine-tuning models using whatever high-accuracy data is available. As shown in Figure 5.3d, even small amounts of high-accuracy data give meaningful improvements to model predictions.

5.5 Appendix

5.5.1 D-MPNN Description

5.5.1.1 Architecture and Featurization

For our experiments, we rely on Chemprop, a deep learning property prediction framework built by Yang et al. [100] A detailed description of the model architecture is provided in Section 2.2.5. Briefly, Chemprop takes SMILES [124] strings as input and outputs the single or multiple properties of interest. While Chemprop is really a general-use framework that can train a range of machine learning models, the heart of Chemprop's innovation is their directed message passing neural network (D-MPNN) architecture, which we adapt here. Crucially, we include the additional architectural choices made by the original authors (i.e., input and output neural layers), which makes the D-MPNN architecture and the resultant Chemprop property prediction framework so successful.

Since we are learning properties of reactions and not individual molecules, we must featurize multiple molecular graphs (i.e., reactants and products). Here, we use the established condensed graph of reaction (CGR) representation [106, 125], which was recently incorporated in Chemprop [93]. CGR is a superposition of the reactant and product graphs; thus it requires atom-mapped reactions to compare the atoms and bonds between the reactant and product (i.e., we must know which reactant atoms led to which product atoms, usually defined by an annotation within the reaction SMILES string). A key aspect of the CGR representation is that it removes disjoint graphs present in multi-molecular reactions and allows message passing between all atoms, a limitation of the Grambow et al. method. Using the CGR representation as input allows the rest of the Chemprop architecture to remain unchanged.

Upon creating the new condensed graph, initial feature vectors are created with RDKit [99] for each atom and bond for both the reactant and product. This creates three possible combinations of features for the single CGR. First, the feature vectors from the reactant and product can be concatenated together (`reac_prod`). Second, the reactant feature vector can be concatenated with the difference of the product and reactant feature vectors (`reac_diff`). Third, the product feature vector can be concatenated with the difference of the reactant and product feature vectors (`prod_diff`). More detail about CGR in Chemprop, along with performance on benchmark datasets, can be found in the original publication [93].

Heid and Green [93] reported that the `reac_diff` CGR representation usually performed best. However, the `reac_diff` representation of forward and reverse reactions is identical to the prod_diff representation of reverse and forward reactions. Since this work augments the datasets at each level of theory with the reverse reactions, we arbitrarily choose the reac_diff CGR representation. Empirically, Table 5.5 shows that all three CGR representations give very similar performance; all models use the Optimal hyperparameter values described in Table 5.8 and Table 5.9.

Table 5.5. Comparison of CGR Representations. Errors are in kcal mol^{-1}, and results are from the first fold of cross-validation.

	B97-D3		**ωB97X-D3**		**CCSD(T)-F12a**	
Representation	MAE	RMSE	MAE	RMSE	MAE	RMSE
reac_diff	4.81	7.88	3.06	5.15	2.98	4.96
prod_diff	4.75	7.78	3.06	5.20	2.84	5.00
reac_prod	4.89	7.80	3.17	5.33	2.96	5.05

We tried several modifications to the model architecture. By default, the initial features in Chemprop simply assign whether the bond is in a ring of any size. We modify this by specifying the actual ring size (mapped to a one-hot vector) and incorporating this for both the atom and bond feature vectors. As shown in Table 5.6, this addition improves the testing RMSE by about 0.4 kcal mol^{-1}. We also allow additional features to be concatenated to the learned molecular representation that is passed to the feed forward network during readout. We find that using RDKit molecular features [123] and quantum atomic descriptors from Guan et al. [102] have no effect.

Table 5.6. Sequential optimization study showing the improvement in testing error (kcal mol^{-1}). Each row includes all changes from the previous row. Results are from the first fold of cross-validation.

	B97-D3		**ωB97X-D3**		**CCSD(T)-F12a**	
Description	MAE	RMSE	MAE	RMSE	MAE	RMSE
Default	6.74	10.22	5.30	8.17	5.07	8.18
+Optimize hyperparameters	5.70	9.29	3.61	6.08	3.44	5.86
+Ring features	5.28	8.79	3.33	5.77	3.24	5.66
+Molecular RDKit features	5.41	8.90	3.42	5.90	3.22	5.66
+Atom QM Descriptors	5.25	8.72	3.39	5.83	3.13	5.60

In contrast, using reaction enthalpy as an input to the dense layer improves model performance. Thus, our model enables a workflow for predicting high-quality barrier heights without the need for a transition state (TS). Researchers only need to obtain the reaction enthalpy by running quantum calculations for the reactant(s) and product(s). One potential drawback of this approach is that quantum calculations can be expensive depending on the method and basis. An alternative workflow could instead quickly obtain a predicted value as input to the final model, an approach commonly done by other published works for property prediction [104, 126]. For example, we could instead use the model from Grambow et al. [63], which was co-trained on ΔE_0 and ΔH (without atom or bond corrections) at the ωB97X-D3/def2-TZVP level of theory, to quickly predict ΔH. Similarly, we also co-trained a model on ΔE_0 and ΔH (with atom and bond corrections) at the CCSD(T)-F12a/cc-pVDZ-F12//ωB97X-D3/def2-TZVP level of theory in this work; testing errors are shown in the third row of Table 5.6. Using the predicted enthalpy from either of these models as input to our final model would be more efficient than using quantum calculations to determine the reaction enthalpy. As seen in Table 5.7, using either of these predicted enthalpy values yields very comparable performance, with just a small 0.17 kcal mol^{-1} decline in testing MAE as compared to using the calculated enthalpy. However, these results may be biased since both the model from Grambow et al. [63] and from this work have already seen nearly all of these reactions during training; the accuracy when extrapolating to new reactions may be lower.

Table 5.7. Testing error (kcal mol^{-1}) when using reaction enthalpy calculated at the respective level of theory or predicted using a pretrained model from Grambow et al. [63] or this work. Results are from the first fold of cross-validation.

	B97-D3		**ωB97X-D3**		**CCSD(T)-F12a**	
Description	MAE	RMSE	MAE	RMSE	MAE	RMSE
Input ΔH (predicted by ref. 62)	5.18	8.26	3.40	5.59	3.18	5.36
Input ΔH (our predictions)	4.94	8.08	3.26	5.36	3.15	5.21
Input ΔH (QM calculation)	**4.81**	**7.88**	**3.06**	**5.15**	**2.98**	**4.96**

5.5.1.2 Training and Hyperparameter Selection

Training, validation, and testing sets are created using a scaffold split on the reactant SMILES, which partitions the data based on the Bemis-Murcko scaffold [98] as calculated by RDKit. The exact procedure is described in ref. [100]. Intuitively, a scaffold split is more challenging than a random split since validation and testing molecules are deliberately chosen to have some class imbalance relative to the training set. This causes the testing performance to be a better measure of extrapolation to new molecules rather than interpolation with molecules very similar to those in the training set. As described in the main text, scaffold splitting was done based on the reactants from the forward reactions. Each pair of forward and reverse reactions is placed in the same set, which ensures that each set is independent. The data is split into 85% training, 5% validation, and 10% testing data. To avoid bias that may be introduced from using just one split, we use 5-fold cross-validation. Performance on the validation set is used to determine the best model weights as well as to choose the Optimal hyperparameter values. We use the Noam learning rate scheduler from Vaswani et al. [127] during training, which starts by linearly increasing the learning rate from the initial to the maximum value over a specified number of warm-up epochs. Then the learning rate is exponentially decreased to the final value over the remaining epochs.

We utilize data at all levels of theory during training. First, the model is pretrained with the lowest level data from B97-D3/def2-mSVP. The model is then fine-tuned with the ωB97X-D3/def2-TZVP data and subsequently with the high-quality CCSD(T)-F12a/cc-pVDZ-F12//ωB97X-D3/def2-TZVP data. We find that model performance is slightly better when the message passing weights are not frozen after the first round of pretraining. Instead, the model weights are initialized using the best weights from the previous training run, and all weights are fine-tuned. We use the provided hyperparameter search code from Chemprop [100] to determine the hyperparameters controlling the model architecture and fitting procedure, which are summarized in Table 5.8 and Table 5.9 respectively. Hyperparameters not shown in the table used the default values from Chemprop.

Table 5.8. Optimal hyperparameter values for model architecture.

Hyperparameter	Value
Hidden size	900
Hidden layers	4
Activation function	Leaky ReLU
Aggregation	sum
Additional FFN Inputs	ΔH

Table 5.9. Optimal hyperparameter values for training.

Hyperparameter	Pretraining	Fine-Tuning
Epochs	65	55
Initial learning rate	10^{-4}	10^{-5}
Maximum learning rate	10^{-3}	10^{-4}
Final learning rate	10^{-5}	10^{-6}
Warm-up epochs	5	5
Gradient clip	10	10

5.5.1.3 Fitting Error vs. Level Of Theory

Table 5.10 shows the testing error from training a model on each level of theory (no transfer learning). Only reactions present in all levels are used i.e., each model had the same data splits for training, validation, and testing so the only variable changing was the level of theory for the regression target. We use the Optimal hyperparameter values from Table 5.8 and Table 5.9. Our results show that is about equally difficult to fit a model to each level of theory; of course, the predictions will be better from a model trained on the higher level dataset.

Table 5.10. Barrier height testing error (kcal mol^{-1}) from training a model on each level of theory.

Level of Theory	MAE	RMSE
B97D3	3.95	6.68
ωB97X-D3	3.94	7.04
CCSD(T)-F12a	4.07	7.01

5.5.2 DimeReaction

Our DimeReaction model is an extension of the recently published DimeNet++ [111], which is directional message passing network that operates on 3D coordinates. DimeNet++ was designed to predict energies and forces of molecular structures with the intent of speeding up molecular dynamics simulations, but the architecture has also shown promise in predicting molecular targets other than energies. More detail about the architecture as well as performance on benchmark datasets can be found in the original publication. In our DimeReaction architecture, the QM-optimized reactant and product coordinates are each passed through the same DimeNet++ model to create a learned representation of each molecule. The learned representation of the product is then subtracted from that of the product before passing through a dense layer to predict the barrier height. Although it may be interesting to explore other architectures, such as subtracting the learned atom representations before aggregating into a molecular representation, our results from the main text show that model performance is extremely sensitive to the TS geometry, which precluded efforts to further modify the architecture.

A scaffold split on the reactant SMILES is again used to create independent sets. As before, the data is split into 85% training, 5% validation, and and 10% testing, and each pair of forward and reverse reactions is placed in the same set. Unlike with our modified Chemprop model, training on only the coupled cluster data yields lower testing errors than using all three levels in a transfer learning approach. We use the Noam learning rate scheduler [127] during training and use Optuna for hyperparameter search [117]; the final hyperparameter values are shown in Table 5.11.

Table 5.11. Optimal hyperparameter values for the DimeReaction model.

Hyperparameter	Value
Epochs	100
Warm-up epochs	4
Batch size	32
Learning rate	10^{-3}
Learning rate scheduler	Noam
Hidden channels	100
Output embedding channels	100
Output channels	100
Interaction embedding size	64
Basis embedding size	8
Blocks	6
Spherical harmonics	6
Radial basis functions	6
Output layers	2
Layers (FFN)	3
Activation (FFN)	SiLU

Using the optimal hyperparameters for the network architecture and training procedure, a few combinations of inputs and training targets are tested. Two types of inputs are explored, either the optimized reactant and product geometries or the optimized reactant and TS geometries. The first strategy represents a realistic prediction strategy since reactants and products are much easier to identify and optimize than a TS. The second strategy better aligns with chemical intuition, since reaction barriers correspond to the difference in energies of the TS and the reactant. The dense layer could optionally receive the enthalpy as an additional input that is concatenated to the difference of the learned molecular representations. The regression target(s) are either just the barrier height or co-training the model on both the barrier height and enthalpy.

The results from training on the coupled cluster data for these combinations are summarized in Table 5.12. For the case when we use reactant and product geometries, using reaction enthalpy marginally improves the testing errors. However, using a 2D network, such as our modified Chemprop, yields far better performance, which is quite surprising. For the case when we use reactant and TS geometries, testing errors improve over the 2D Chemprop formulation that had used the CGR of the reactant and product graphs. Although it is satisfying to see the TS encode more information about the reaction's barrier height than the product does, the optimized TS is better used with canonical transition state theory to calculate the rate constant.

An interesting conclusion comes from row 2 and row 4 in Table 5.12, both of which train a multi-task model to predict ΔE_0 and ΔH. When using the optimized reactant and product geometries as input (row 2), the model gives very good predictions for reaction enthalpy with a test MAE of 1.78 kcal mol^{-1}. On the contrary, the model trained with optimized reactants and TSs (row 4) results in a MAE of 5.95 kcal mol^{-1} for ΔH, which is quite poor. These results indicate that the DimeNet++ model functions as intended; it was designed as a neural network potential. When the output property is a function of the input geometries (i.e., reactants and products for reaction enthalpy or reactants and TSs for barrier height), DimeNet++ works well. However, when the output property is not a function of the inputs, the model performs poorly. This is unfortunate since the more convenient use case would be to predict barrier height from reactant and product structures.

Table 5.12. DimeReaction testing errors (kcal mol^{-1}) for combinations of regression targets.

Optimized Input	Target(s)	ΔE_0 MAE	ΔE_0 RMSE	ΔH MAE	ΔH RMSE
Reactant & Product	ΔE_0	6.20	10.10	-	-
Reactant & Product	ΔE_0 & ΔH	6.03	9.84	1.78	2.78
Reactant & Product	ΔH	-	-	1.26	2.20
Reactant & TS	ΔE_0	2.25	3.91	-	-
Reactant & TS	ΔE_0 & ΔH	2.67	4.60	5.95	11.93
Reactant & TS	ΔH	-	-	6.32	11.87

5.5.3 Hyperparameters for Baseline Models

The best hyperparameters for the baseline models are chosen via Optuna [117], which is used to minimize the average root mean squared error (RMSE) on the validation sets. The optimal hyperparameter values are listed in Tables 5.13, 5.14, 5.15, 5.16, 5.17. Any other hyperparameters assumed the default value.

Table 5.13. Optimal hyperparameter values for LinearSVR i.e., support vector regression with linear kernel.

Hyperparameter	Value
`C`	0.0136

Table 5.14. Optimal hyperparameter values for SVR with RBF kernel.

Hyperparameter	Value
`C`	2.5646

Table 5.15. Optimal hyperparameter values for random forest (RF).

Hyperparameter	Value
`max_depth`	15
`n_estimators`	125

Table 5.16. Optimal hyperparameter values for extreme gradient boosting (XGB).

Hyperparameter	Value
`gamma`	4
`learning_rate`	0.1801
`max_depth`	9
`min_child_weight`	0
`n_estimators`	75
`reg_alpha`	3
`reg_lambda`	3

Table 5.17. Optimal hyperparameter values for multi-layer perceptron (MLP).

Hyperparameter	Value
`num_layers`	4
`layer_size`	500
`alpha`	0.00137

5.6 References

(1) Van't Hoff, J. H., *Etudes De Dynamique Chimique*; Muller: 1884; Vol. 1.

(2) Arrhenius, S. Über Die Reaktionsgeschwindigkeit Bei Der Inversion Von Rohrzucker Durch SÄUren. *Zeitschrift für physikalische Chemie* **1889**, *4*, a translation of the four pages in this paper that deal with temperature dependence is included in Back, M. H., and Laidler, K. J., "Selected Readings in Chemical Kinetics," Pergamon Press, Oxford, 1967, pp. 31-35., 226–248.

(3) Montreal Protocol on Substances That Deplete the Ozone Layer. *Washington, DC: US Government Printing Office* **1987**, *26*, 128–136.

(4) Velders, G. J. M.; Andersen, S. O.; Daniel, J. S.; Fahey, D. W.; McFarland, M. The Importance of the Montreal Protocol in Protecting Climate. *Proc. Natl. Acad. Sci.* **2007**, *104*, 4814–4819.

(5) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **2017**, *3*, 434–443.

(6) Gao, H.; Struble, T. J.; Coley, C. W.; Wang, Y.; Green, W. H.; Jensen, K. F. Using Machine Learning to Predict Suitable Conditions for Organic Reactions. *ACS Cent. Sci.* **2018**, *4*, 1465–1476.

(7) Truhlar, D. G.; Garrett, B. C.; Klippenstein, S. J. Current Status of Transition-State Theory. *J. Phys. Chem.* **1996**, *100*, 12771–12800.

(8) Klippenstein, S. J. From Theoretical Reaction Dynamics to Chemical Modeling of Combustion. *Proc. Combust. Inst.* **2017**, *36*, 77–111.

(9) Green, W. H. Moving from Postdictive to Predictive Kinetics in Reaction Engineering. *AIChE Journal* **2020**, *66*, e17059.

(10) Pio, G.; Dong, X.; Salzano, E.; Green, W. H. Automatically Generated Model for Light Alkene Combustion. *Combust. Flame* **2022**, *241*, 112080.

(11) Payne, A. M.*; Spiekermann, K. A.*; Green, W. H. Detailed Reaction Mechanism for 350–400°C Pyrolysis of an Alkane, Aromatic, and Long-Chain Alkylaromatic Mixture. *Energy & Fuels* **2022**, *36*, 1635–1646.

(12) Vermeire, F. H.; Aravindakshan, S. U.; Jocher, A.; Liu, M.; Chu, T.-C.; Hawtof, R. E.; Van de Vijver, R.; Prendergast, M. B.; Van Geem, K. M.; Green, W. H. Detailed Kinetic Modeling for the Pyrolysis of a Jet A Surrogate. *Energy Fuels* **2022**, *36*, 1304–1315.

(13) Wu, H.; Grinberg Dana, A.; Ranasinghe, D. S.; Pickard I. V., F. C.; Wood, G. P. F.; Zelesky, T.; Sluggett, G. W.; Mustakis, J.; Green, W. H. Kinetic Modeling of API Oxidation: 2. Imipramine Stress Testing. *Mol. Pharmaceutics* **2022**, *19*, 1526–1539.

(14) Grambow, C. A.; Jamal, A.; Li, Y.-P.; Green, W. H.; Zador, J.; Suleimanov, Y. V. Unimolecular Reaction Pathways of a γ-Ketohydroperoxide from Combined Application of Automated Reaction Discovery Methods. *J. Am. Chem. Soc.* **2018**, *140*, 1035–1048.

(15) Maeda, S.; Harabuchi, Y. On Benchmarking of Automated Methods for Performing Exhaustive Reaction Path Search. *J. Chem. Theory Comput.* **2019**, *15*, 2111–2115.

(16) Fukui, K. The Path of Chemical Reactions-The IRC Approach. *Acc. Chem. Res.* **1981**, *14*, 363–368.

(17) Pitzer, K. S.; Gwinn, W. D. Energy Levels and Thermodynamic Functions for Molecules with Internal Rotation I. Rigid Frame with Attached Tops. *J. Chem. Phys.* **1942**, *10*, 428–440.

(18) Truhlar, D. G. A Simple Approximation for the Vibrational Partition Function of a Hindered Internal Rotation. *J. Comput. Chem.* **1991**, *12*, 266–270.

(19) Pfaendtner, J.; Yu, X.; Broadbelt, L. J. The 1-D Hindered Rotor Approximation. *Theor. Chem. Acc.* **2007**, *118*, 881–898.

(20) Sharma, S.; Raman, S.; Green, W. H. Intramolecular Hydrogen Migration in Alkylperoxy and Hydroperoxyalkylperoxy Radicals: Accurate Treatment of Hindered Rotors. *J. Phys. Chem. A* **2010**, *114*, 5689–5701.

(21) Zheng, J.; Yu, T.; Papajak, E.; Alecu, I. M.; Mielke, S. L.; Truhlar, D. G. Practical Methods for Including Torsional Anharmonicity in Thermochemical Calculations on Complex Molecules: The Internal-Coordinate Multi-Structural Approximation. *Phys. Chem. Chem. Phys.* **2011**, *13*, 10885–10907.

(22) Yu, T.; Zheng, J.; Truhlar, D. G. Multi-Structural Variational Transition State Theory. Kinetics of the 1,4-Hydrogen Shift Isomerization of the Pentyl Radical with Torsional Anharmonicity. *Chem. Sci.* **2011**, *2*, 2199–2213.

(23) Zheng, J.; Truhlar, D. G. Quantum Thermochemistry: Multistructural Method with Torsional Anharmonicity Based on a Coupled Torsional Potential. *J. Chem. Theory Comput.* **2013**, *9*, 1356–1367.

(24) Zheng, J.; Meana-Pañeda, R.; Truhlar, D. G. MSTor Version 2013: A New Version of the Computer Code for the Multi-Structural Torsional Anharmonicity, Now With a Coupled Torsional Potential. *Comput. Phys. Commun.* **2013**, *184*, 2032–2033.

(25) Gonzalez-Lavado, E.; Corchado, J. C.; Suleimanov, Y. V.; Green, W. H.; Espinosa-Garcia, J. Theoretical Kinetics Study of the $O(^3P)+$ CH_4/CD_4 Hydrogen Abstraction Reaction: The Role of Anharmonicity, Recrossing Effects, and Quantum Mechanical Tunneling. *J. Phys. Chem. A* **2014**, *118*, 3243–3252.

(26) Zheng, J.; Bao, J. L.; Meana-Pañeda, R.; Zhang, S.; Lynch, B. J.; Corchado, J. C.; Chuang, Y.-Y.; Fast, P. L.; Hu, W.-P.; Liu, Y.-P.; Lynch, G. C.; Nguyen, K. A.; Jackels, C. F.; Fernandez Ramos, A.; Ellingson, B. A.; Melissas, V. S.; Villà, J.; Rossi, I.; Coitiño, E. L.; Pu, J.; Albu, T. V.; Ratkiewicz, A.; Steckler, R.; Garrett, B. C.; Isaacson, A. D.; Truhlar, D. G. *Polyrate*-version 2017-C, University of Minnesota: Minneapolis, 2017.

(27) Goldman, M. J.; Ono, S.; Green, W. H. Correct Symmetry Treatment for X+ X Reactions Prevents Large Errors in Predicted Isotope Enrichment. *J. Phys. Chem. A* **2019**, *123*, 2320–2324.

(28) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50*, 572–584.

(29) Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know to Improve Conformation Generation. *J. Chem. Inf. Model.* **2015**, *55*, 2562–2574.

(30) Pracht, P.; Bohle, F.; Grimme, S. Automated Exploration of the Low-Energy Chemical Space with Fast Quantum Chemical Methods. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169–7192.

(31) Ganea, O.; Pattanaik, L.; Coley, C.; Barzilay, R.; Jensen, K.; Green, W.; Jaakkola, T. GeoMol: Torsional Geometric Generation of Molecular 3D Conformer Ensembles. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 13757–13769.

(32) Bhoorasingh, P. L.; West, R. H. Transition State Geometry Prediction Using Molecular Group Contributions. *Phys. Chem. Chem. Phys.* **2015**, *17*, 32173–32182.

(33) Harms, N.; Underkoffler, C.; West, R. H. Advances in Automated Transition State Theory Calculations: Improvements on the AutoTST Framework. *10.26434/chemrxiv* **2020**, *13277870.v2*, Accessed 2022-03-08.

(34) Van de Vijver, R.; Zádor, J. KinBot: Automated Stationary Point Search on Potential Energy Surfaces. *Comput. Phys. Commun.* **2020**, *248*, 106947.

(35) Pattanaik, L.; Ingraham, J. B.; Grambow, C. A.; Green, W. H. Generating Transition States of Isomerization Reactions with Deep Learning. *Phys. Chem. Chem. Phys.* **2020**, *22*, 23618–23626.

(36) Jackson, R.; Zhang, W.; Pearson, J. TSNet: Predicting Transition State Structures with Tensor Field Networks and Transfer Learning. *Chem. Sci.* **2021**, *12*, 10022–10040.

(37) Makoś, M. Z.; Verma, N.; Larson, E. C.; Freindorf, M.; Kraka, E. Generative Adversarial Networks for Transition State Geometry Prediction. *J. Chem. Phys.* **2021**, *155*, 024116.

(38) Bischoff, F. A.; Wolfsegger, S.; Tew, D. P.; Klopper, W. Assessment of Basis Sets for F12 Explicitly-Correlated Molecular Electronic-Structure Methods. *Mol. Phys.* **2009**, *107*, 963–975.

(39) Knizia, G.; Adler, T. B.; Werner, H.-J. Simplified CCSD(T)-F12 Methods: Theory and Benchmarks. *J. Chem. Phys.* **2009**, *130*, 054104.

(40) Adler, T. B.; Knizia, G.; Werner, H.-J. A Simple and Efficient CCSD(T)-F12 Approximation. *J. Chem. Phys.* **2007**, *127*, 221106.

(41) Zheng, J.; Zhao, Y.; Truhlar, D. G. The DBH24/08 Database and its Use to Assess Electronic Structure Model Chemistries for Chemical Reaction Barrier Heights. *J. Chem. Theory Comput.* **2009**, *5*, 808–821.

(42) Pfeiffer, F.; Rauhut, G.; Feller, D.; Peterson, K. A. Anharmonic Zero Point Vibrational Energies: Tipping the Scales in Accurate Thermochemistry Calculations? *J. Chem. Phys.* **2013**, *138*, 044311.

(43) Spiekermann, K. A.; Pattanaik, L.; Green, W. H. High Accuracy Barrier Heights, Enthalpies, and Rate Coefficients for Chemical Reactions. *Sci. Data* **2022**, *9*, 417.

(44) Feller, D.; Peterson, K. A.; Dixon, D. A. A Survey of Factors Contributing to Accurate Theoretical Predictions of Atomization Energies and Molecular Structures. *J. Chem. Phys.* **2008**, *129*, 204105.

(45) Evans, M. G.; Polanyi, M. Inertia and Driving Force of Chemical Reactions. *Trans. Faraday Soc.* **1938**, *34*, 11–24.

(46) Hammett, L. P. The Effect of Structure Upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **1937**, *59*, 96–103.

(47) Benson, S. W., *Thermochemical Kinetics*; Wiley: 1976.

(48) Benson, S. W.; Buss, J. H. Additivity Rules for the Estimation of Molecular Properties. Thermodynamic Properties. *J. Chem. Phys.* **1958**, *29*, 546–572.

(49) Cohen, N.; Benson, S. W. Estimation of Heats of Formation of Organic Compounds by Additivity Methods. *Chem. Rev.* **1993**, *93*, 2419–2438.

(50) Saeys, M.; Reyniers, M.-F.; Marin, G. B.; Van Speybroeck, V.; Waroquier, M. Ab Initio Group Contribution Method for Activation Energies for Radical Additions. *AIChE Journal* **2004**, *50*, 426–444.

(51) Saeys, M.; Reyniers, M.-F.; Van Speybroeck, V.; Waroquier, M.; Marin, G. B. Ab Initio Group Contribution Method for Activation Energies of Hydrogen Abstraction Reactions. *ChemPhysChem* **2006**, *7*, 188–199.

(52) Adamczyk, A. J.; Reyniers, M.-F.; Marin, G. B.; Broadbelt, L. J. Exploring 1, 2-Hydrogen Shift in Silicon Nanoparticles: Reaction Kinetics from Quantum Chemical Calculations and Derivation of Transition State Group Additivity Database. *J. Phys. Chem. A* **2009**, *113*, 10933–10946.

(53) Sumathi, R.; Carstensen, H.-H.; Green, W. H. Reaction Rate Prediction via Group Additivity Part 1: H Abstraction from Alkanes by H and CH_3. *J. Phys. Chem. A* **2001**, *105*, 6910–6925.

(54) Sumathi, R.; Green, W. H. A Priori Rate Constants for Kinetic Modeling. *Theor. Chem. Acc.* **2002**, *108*, 187–213.

(55) Green, W. H. Predictive Kinetics: A New Approach for the 21st Century. *Adv. Chem. Eng.* **2007**, *32*, 1–50.

(56) Liu, M.; Grinberg Dana, A.; Johnson, M. S.; Goldman, M. J.; Jocher, A.; Payne, A. M.; Grambow, C. A.; Han, K.; Yee, N. W.; Mazeau, E. J., et al. Reaction Mechanism Generator v3.0: Advances in Automatic Mechanism Generation. *J. Chem. Inf. Model.* **2021**, *61*, 2686–2696.

(57) Wang, K.; Dean, A. M., Rate Rules and Reaction Classes In *Comput. Aided Chem. Eng.* Elsevier: 2019; Vol. 45, pp 203–257.

(58) Johnson, M. S.; Green, W. H. In *2019 AIChE Annual Meeting*, 2019.

(59) Heid, E.; Goldman, S.; Sankaranarayanan, K.; Coley, C. W.; Flamm, C.; Green, W. H. EHreact: Extended Hasse Diagrams for the Extraction and Scoring of Enzymatic Reaction Templates. *J. Chem. Inf. Model.* **2021**, *61*, 4949–4961.

(60) Hawkins, P. C. D. Conformation Generation: The State of the Art. *J. Chem. Inf. Model.* **2017**, *57*, 1747–1756.

(61) Mansimov, E.; Mahmood, O.; Kang, S.; Cho, K. Molecular Geometry Prediction Using a Deep Generative Graph Neural Network. *Sci. Rep.* **2019**, *9*, 1–13.

(62) Simm, G. N.; Hernández-Lobato, J. M. A Generative Model for Molecular Distance Geometry. *arXiv* **2019**, *1909.11459*, Accessed 2022-03-08.

(63) Grambow, C. A.; Pattanaik, L.; Green, W. H. Deep Learning of Activation Energies. *J. Phys. Chem. Lett.* **2020**, *11*, 2992–2997.

(64) Stuyver, T.; Coley, C. W. Quantum Chemistry-Augmented Neural Networks for Reactivity Prediction: Performance, Generalizability, and Explainability. *J. Chem. Phys.* **2022**, *156*, 084104.

(65) Jorner, K.; Brinck, T.; Norrby, P.-O.; Buttar, D. Machine Learning Meets Mechanistic Modelling for Accurate Prediction of Experimental Activation Energies. *Chem. Sci.* **2021**, *12*, 1163–1175.

(66) Heinen, S.; von Rudorff, G. F.; von Lilienfeld, O. A. Toward the Design of Chemical Reactions: Machine Learning Barriers of Competing Mechanisms in Reactant Space. *J. Chem. Phys.* **2021**, *155*, 064105.

(67) Zhong, S.; Hu, J.; Fan, X.; Yu, X.; Zhang, H. A Deep Neural Network Combined with Molecular Fingerprints (DNN-MF) to Develop Predictive Models for Hydroxyl Radical Rate Constants of Water Contaminants. *J. Hazard. Mater.* **2020**, *383*, 121141.

(68) Lu, J.; Zhang, H.; Yu, J.; Shan, D.; Qi, J.; Chen, J.; Song, H.; Yang, M. Predicting Rate Constants of Hydroxyl Radical Reactions with Alkanes Using Machine Learning. *J. Chem. Inf. Model.* **2021**, *61*, 4259–4265.

(69) Greaves, T. L.; McHale, K. S. S.; Burkart-Radke, R. F.; Harper, J. B.; Le, T. C. Machine Learning Approaches to Understand and Predict Rate Constants for Organic Processes in Mixtures Containing Ionic Liquids. *Phys. Chem. Chem. Phys.* **2021**, *23*, 2742–2752.

(70) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods V: Modification of NDDO Approximations and Application to 70 Elements. *J. Mol. Model.* **2007**, *13*, 1173–1213.

(71) Henkelman, G.; Uberuaga, B. P.; Jónsson, H. A Climbing Image Nudged Elastic Band Method for Finding Saddle Points and Minimum Energy Paths. *J. Chem. Phys.* **2000**, *113*, 9901–9904.

(72) Peters, B.; Heyden, A.; Bell, A. T.; Chakraborty, A. A Growing String Method for Determining Transition States: Comparison to the Nudged Elastic Band and String Methods. *J. Chem. Phys.* **2004**, *120*, 7877–7886.

(73) Behn, A.; Zimmerman, P. M.; Bell, A. T.; Head-Gordon, M. Efficient Exploration of Reaction Paths via a Freezing String Method. *J. Chem. Phys.* **2011**, *135*, 224108.

(74) Dewyer, A. L.; Argüelles, A. J.; Zimmerman, P. M. Methods for Exploring Reaction Space in Molecular Systems. *WIREs Comput. Mol. Sci.* **2018**, *8*, e1354.

(75) Singh, A. R.; Rohr, B. A.; Gauthier, J. A.; Nørskov, J. K. Predicting Chemical Reaction Barriers with a Machine Learning Model. *Catal. Lett.* **2019**, *149*, 2347–2354.

(76) Takahashi, K.; Miyazato, I. Rapid Estimation of Activation Energy in Heterogeneous Catalytic Reactions via Machine Learning. *J. Comput. Chem.* **2018**, *39*, 2405–2408.

(77) Houston, P. L.; Nandi, A.; Bowman, J. M. A Machine Learning Approach for Prediction of Rate Constants. *J. Phys. Chem. Lett.* **2019**, *10*, 5250–5258.

(78) Nandi, A.; Bowman, J. M.; Houston, P. A Machine Learning Approach for Rate Constants. II. Clustering, Training, and Predictions for the O(^{3}P) + HCl –> OH + Cl Reaction. *J. Phys. Chem. A* **2020**, *124*, 5746–5755.

(79) Peterson, A. A. Acceleration of Saddle-Point Searches with Machine Learning. *J. Chem. Phys.* **2016**, *145*, 074106.

(80) Pozun, Z. D.; Hansen, K.; Sheppard, D.; Rupp, M.; Müller, K.-R.; Henkelman, G. Optimizing Transition States via Kernel-Based Machine Learning. *J. Chem. Phys.* **2012**, *136*, 174101.

(81) Hermes, E. D.; Sargsyan, K.; Najm, H. N.; Zádor, J. Accelerated Saddle Point Refinement Through Full Exploitation of Partial Hessian Diagonalization. *J. Chem. Theory Comput.* **2019**, *15*, 6536–6549.

(82) Komp, E.; Valleau, S. Machine Learning Quantum Reaction Rate Constants. *J. Phys. Chem. A* **2020**, *124*, 8607–8613.

(83) Von Rudorff, G. F.; Heinen, S. N.; Bragato, M.; von Lilienfeld, O. A. Thousands of Reactants and Transition States for Competing E2 and S2 Reactions. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045026.

(84) Bragato, M.; von Rudorff, G. F.; von Lilienfeld, O. A. Data Enhanced Hammett-Equation: Reaction Barriers in Chemical Space. *Chem. Sci.* **2020**, *11*, 11859–11868.

(85) Komp, E.; Valleau, S. Low-Cost Prediction of Molecular and Transition State Partition Functions via Machine Learning. *Chem. Sci.* **2022**, *13*, 7900–7906.

(86) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: the Δ-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087–2096.

(87) Meuwly, M. Machine Learning for Chemical Reactions. *Chem. Rev.* **2021**, *121*, 10218–10239.

(88) Lewis-Atwell, T.; Townsend, P. A.; Grayson, M. N. Machine Learning Activation Energies of Chemical Reactions. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2021**, `https://doi.org/10.1002/wcms.1593`, e1593.

(89) Komp, E.; Janulaitis, N.; Valleau, S. Progress Towards Machine Learning Reaction Rate Constants. *Phys. Chem. Chem. Phys.* **2022**, *24*, 2692–2705.

(90) Suleimanov, Y. V.; Green, W. H. Automated Discovery of Elementary Chemical Reaction Steps Using Freezing String and Berny Optimization Methods. *J. Chem. Theory Comput.* **2015**, *11*, 4248–4259.

(91) Koerstz, M.; Rasmussen, M. H.; Jensen, J. H. Fast and Automated Identification of Reactions with Low Barriers: the Decomposition of 3-Hydroperoxypropanal. *SciPost Chemistry* **2021**, *1*, 003.

(92) Grambow, C. A.; Pattanaik, L.; Green, W. H. Reactants, Products, and Transition States of Elementary Chemical Reactions Based on Quantum Chemistry. *Sci. Data* **2020**, *7*, 1–8.

(93) Heid, E.; Green, W. H. Machine Learning of Reaction Properties via Learned Representations of the Condensed Graph of Reaction. *J. Chem. Inf. Model.* **2021**, *62*, 2101–2110.

(94) Spiekermann, K. A.; Pattanaik, L.; Green, W. H. High Accuracy Barrier Heights, Enthalpies, and Rate Coefficients for Chemical Reactions version 1.0.1, version 1.0.1, 2022.

(95) Grambow, C. A.; Li, Y.-P.; Green, W. H. Accurate Thermochemistry with Small Data Sets: A Bond Additivity Correction and Transfer Learning Approach. *J. Phys. Chem. A* **2019**, *123*, 5826–5835.

(96) Scott, A. P.; Radom, L. Harmonic Vibrational Frequencies: An Evaluation of Hartree-Fock, Møller-Plesset, Quadratic Configuration Interaction, Density Functional Theory, and Semiempirical Scale Factors. *J. Phys. Chem.* **1996**, *100*, 16502–16513.

(97) Alecu, I. M.; Zheng, J.; Zhao, Y.; Truhlar, D. G. Computational Thermochemistry: Scale Factor Databases and Scale Factors for Vibrational Frequencies Obtained from Electronic Model Chemistries. *J. Chem. Theory Comput.* **2010**, *6*, 2872–2887.

(98) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(99) Landrum, G. et al. RDKit: Open-Source Cheminformatics Software, https://www.rdkit.org, 2006.

(100) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M., et al. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388.

(101) Wang, A. Y.-T.; Murdock, R. J.; Kauwe, S. K.; Oliynyk, A. O.; Gurlo, A.; Brgoch, J.; Persson, K. A.; Sparks, T. D. Machine Learning for Materials Scientists: An Introductory Guide Toward Best Practices. *Chem. Mater.* **2020**, *32*, 4954–4965.

(102) Guan, Y.; Coley, C. W.; Wu, H.; Ranasinghe, D.; Heid, E.; Struble, T. J.; Pattanaik, L.; Green, W. H.; Jensen, K. F. Regio-Selectivity Prediction with a Machine-Learned Reaction Representation and On-The-Fly Quantum Mechanical Descriptors. *Chem. Sci.* **2021**, *12*, 2198–2208.

(103) Artrith, N.; Butler, K. T.; Coudert, F.-X.; Han, S.; Isayev, O.; Jain, A.; Walsh, A. Best Practices in Machine Learning for Chemistry. *Nat. Chem.* **2021**, *13*, 505–508.

(104) Greenman, K. P.; Green, W. H.; Gomez-Bombarelli, R. Multi-Fidelity Prediction of Molecular Optical Peaks with Deep Learning. *Chem. Sci.* **2022**, *13*, 1152–1162.

(105) Bacciu, D.; Errica, F.; Micheli, A.; Podda, M. A Gentle Introduction to Deep Learning for Graphs. *Neural Networks* **2020**, *129*, 203–221.

(106) Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V. P. Substructural Fragments: An Universal Language to Encode Reactions, Molecular and Supramolecular Structures. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 693–703.

(107) Spiekermann, K. A.; Pattanaik, L.; Green, W. H.; Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M., et al., https://github.com/kspieks/chemprop/tree/barrier_prediction. Accessed 2022-05-16.

(108) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet–A Deep Learning Architecture for Molecules and Materials. *J. Chem. Phys.* **2018**, *148*, 241722.

(109) Devereux, C.; Smith, J. S.; Davis, K. K.; Barros, K.; Zubatyuk, R.; Isayev, O.; Roitberg, A. E. Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens. *J. Chem. Theory Comput.* **2020**, *16*, 4192–4202.

(110) Westermayr, J.; Gastegger, M.; Marquetand, P. Combining SchNet and SHARC: The SchNarc Machine Learning Approach for Excited-State Dynamics. *J. Phys. Chem. Lett.* **2020**, *11*, 3828–3834.

(111) Klicpera, J.; Giri, S.; Margraf, J. T.; Günnemann, S. Fast and Uncertainty-Aware Directional Message Passing for Non-Equilibrium Molecules. *arXiv* **2020**, *2011.14115*, Accessed 2022-03-08.

(112) Klicpera, J.; Groß, J.; Günnemann, S. Directional Message Passing for Molecular Graphs. *arXiv* **2020**, *2003.03123.*

(113) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1*, 1–7.

(114) Spiekermann, K. A.; Pattanaik, L.; Green, W. H., https://github.com/kspieks/DimeReaction. Accessed 2022-04-11.

(115) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(116) Chen, T.; Guestrin, C. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM: San Francisco, California, USA, 2016, pp 785–794.

(117) Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

(118) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–113.

(119) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(120) Bender, A.; Schneider, N.; Segler, M.; Patrick Walters, W.; Engkvist, O.; Rodrigues, T. Evaluation Guidelines for Machine Learning Tools in the Chemical Sciences. *Nat. Rev. Chem.* **2022**, *6*, 428–442.

(121) Dietterich, T. G. In *Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science, vol 1857. Springer, Berlin, Heidelberg*; https://link.springer.com/chapter/10.1007/3-540-45014-9_1.pdf, 2000, pp 1–15.

(122) Vermeire, F. H.; Green, W. H. Transfer learning for solvation free energies: From quantum chemistry to experiments. *Chem. Eng. J.* **2021**, *418*, 129307.

(123) Kelley, B. Descriptor Computation (Chemistry) and (Optional) Storage for Machine Learning. DescriptaStorus, version 2.2.0, https://github.com/bp-kelley/descriptastorus. Accessed 2021-09-03.

(124) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31–36.

(125) Hoonakker, F.; Lachiche, N.; Varnek, A.; Wagner, A. Condensed Graph of Reaction: Considering a Chemical Reaction as One Single Pseudo Molecule. *Int. J. Artif. Intell. Tools* **2011**, *20*, 253–270.

(126) Chung, Y.; Vermeire, F. H.; Wu, H.; Walker, P. J.; Abraham, M. H.; Green, W. H. Group Contribution and Machine Learning Approaches to Predict Abraham Solute Parameters, Solvation Free Energy, and Solvation Enthalpy. *J. Chem. Inf. Model.* **2022**, *62*, 433–446.

(127) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.

Chapter 6

Accurately Predicting Barrier Heights for Radical Reactions in Solution

This work is in preparation as Spiekermann, K. A.; Bergeler, M.; Menon, A.; Dong, X.; Pfeifle, M.; Sandfort, F.; Welz, O.; Green, W. H. "Accurately Predicting Barrier Heights for Radical Reactions in Solution using Deep Graph Networks" 2023. Angiras Menon helped with quantum chemistry calculations. Xiaorui Dong helped parse SMILES from optimized structures. Maike Bergeler, Mark Pfeifle, Frederik Sandfort, and Oliver Welz led the dataset generation. All code and model weights are available at https://github.com/kspieks/chemprop/tree/solvated_barrier_barrier_prediction.

6.1 Introduction

Accurate kinetic parameters are crucial for modeling various chemical kinetic processes, with examples including pyrolysis [1, 2], polymerization [3–5], plastic recycling [6], and oxidative degradation of pharmaceutical compounds [7], just to name a few. Two such parameters are the Gibbs free energy of activation and of reaction, which determine a reaction's rate coefficient and equilibrium constant at a given temperature via the Eyring and Gibbs-Helmholtz equation, respectively. Gibbs free energies can be computed using quantum chemistry calculations [4]. Although they are commonly calculated in gas phase, solvent effects can have an important impact on the species energies and subsequent reaction properties when looking at reactivity in solution [8, 9]. To convert from gas phase to solution phase, one needs the solvation free energy, which is defined as the change in Gibbs free energy between a molecule in the gas phase and in solution at a given temperature. It is related to many physicochemical properties, including solubility [10], reaction equilibrium and kinetics [11], protein-ligand binding affinities [12], dissociation constants [13], lipophilicity via the alkane-water partition coefficient [14–17], and the distribution coefficient [17]. Altogether, this makes the Gibbs free energy relevant to pharmacokinetic parameters, such as absorption, distribution, metabolism, and excretion (ADME). Unfortunately, experimental solvation free energy data are relatively rare [18, 19] so researchers must rely on computational techniques to calculate values for many reactions.

There are three main approaches to calculate solvation effects [5, 20–22]. Explicit methods include an explicit solvent environment during geometry optimization and energy calculation, which provides the most descriptive and realistic model of solvent-solute interactions [23]. However, adding many solvent molecules is often too computationally expensive for high-throughput applications. Implicit methods represent solvation as a solute placed inside a cavity within an implicit solvent, which is modeled as a continuum with a constant property, such as conductivity or dielectric constant; these methods are commonly integrated into quantum mechanical workflows [24]. Many implicit models have been developed, including the polarizable continuum model (PCM) [25], solvation method based on density (SMD) [26, 27], conductor-like screening model (COSMO) [28], Poisson–Boltzmann (PB) model [29], and generalized Born (GB) model [30]. Although several studies have used implicit continuum models to calculate free energies and rate coefficients of liquid phase reactions [31–36], these methods are not the most accurate, partially because they do not consider local solvent-solute interactions [5, 22, 37]. An implicit model going beyond the continuum approximation is COSMO-RS (conductor-like screening model for real solvents) [38–42]. It considers the statistical thermodynamics of pairwise interacting molecular surface segments and is therefore able to describe interactions between different functional groups in a physically sound way. The density functional theory (DFT) methodology used within COSMO-RS is further discussed in the methods section. Due to its favorable prediction performance at relatively modest computational cost, it has gained quite some popularity in the last decades [5, 43–47]. Finally, hybrid implicit-explicit models employ a small number of explicit solvent molecules around the solute and use implicit models to describe the interaction of the solvent-solute cluster with the solvent environment [37]. This has the potential to combine the advantage of explicit and implicit models. The caveat of this approach is that the placement of solvent molecules must be chosen in an adequate way, which is a tough challenge for automated high-throughput workflows.

Despite the various methods to calculate solvation effects, all ab initio techniques to calculate thermodynamic and kinetic properties in either the gas phase or solution phase are limited by the computational expense of obtaining a 3D structure, which can be particularly challenging for a transition state (TS). Additionally, many modern kinetic mechanisms must consider tens of thousands of elementary reactions [48], or even more, so it is crucial to have estimators that can quickly make predictions. While many published models predict free energy values of individual molecules [49–61], comparatively fewer studies have focused on reactions in solution. One example comes from Ravasco and Coelho [62], who used multivariate regression to predict the Gibbs free energy of activation ($\Delta G^{\ddagger}$) for inverse-electron demand Diels-Alder reactions primarily in water using molecular descriptors calculated at the M06-2X/6-31G(d) level of theory (LoT). Migliaro and Cundari [63] use a shallow neural network to predict $\Delta G^{\ddagger}$ for methane activation reactions using descriptors calculated at the ωB97X-D/def2-TZVPP LoT. In both cases, the computational cost associated with calculating descriptors using DFT precludes these models from quickly screening large regions of reaction space. Other studies use a Δ-ML approach [64] to predict the residual between a high- and low-accuracy method. One example comes from Gómez-Flores et al. [65], who fit a shallow neural network to the energy difference between the density functional tight-binding model and other higher level QM methods for thiol-disulfide exchange reactions in water. Similarly, Farrar and Grayson [66] used seven different ML models to predict DFT-quality $\Delta G^{\ddagger}$ for a set of nitro-Michael addition reactions in toluene based on the features generated from semi-empirical quantum mechanical (SQM) methods. Although optimizing the geometries of several conformers is much faster via SQM than with DFT, the computational time is non-negligible so these models are still not feasible for high-throughput screening. Furthermore, their models are limited to a single solvent and reaction family (i.e., the template representing the transformation of reactants to products), which severely hinders their generalizability.

There are a few examples of models that predict reaction properties in multiple solvents without relying on 3D geometries. Jorner et al. [67] employed a Gaussian process regressor, along with other traditional ML models, to predict experimental quality barrier heights of 443 S_NAr reactions in various solvents. The best model presented from this work requires features from a DFT-level optimization of both the TS and minima, which is not well-suited for high-throughput applications. When using features only derived from SMILES [68], such as fingerprints from the Morgan algorithm [69, 70] in RDKit [71] or from a BERT model trained to create atom-mapped reaction SMILES [72], model performance was worse, but obtaining these fingerprints is more practical for inference. Using only SMILES as input, Heid and Green [73] trained a graph network on this same S_NAr dataset. The learned reaction representation in their model gives better barrier height predictions than the fingerprint models from Jorner et al. [67] and even gives similar performance as when Jorner et al. [67] used quantum chemical features. Although the models from both studies can successfully provide fast predictions for reactions in solution, there are still several limitations. First, these models continue to be limited by a small homogeneous training set comprised of only one reaction family. Over half of these $\mathrm{S_NAr}$ reactions took place in just two solvents: acetonitrile and methanol. Furthermore, both models used fixed descriptors to represent solvents, but previous studies have shown that molecular representations learned for the specific task often outperform fixed molecular descriptors [49, 54, 73–108]. In summary, there is a need for larger datasets that contain both increased reaction diversity and better solvent representation. Satisfying these criteria is necessary to train a general purpose estimator that offers broad utility.

Here, our work presents two important contributions for the field of kinetics in solution. First, we create a high-quality dataset with nearly 6,000 elementary radical reactions. While many computational kinetic datasets focus on gas phase [99, 109–116], we additionally calculate solvation corrections for each reaction in 40 popular solvents. These data span a diverse set of reactions, include conformer searches in the COSMO phase, and all TS geometries are validated by an intrinsic reaction coordinate (IRC) calculation. Since identifying TSs is quite challenging and computationally expensive, our second contribution is to train a general purpose deep graph neural network to directly predict the Gibbs free energy of activation and of reaction using only the SMILES for the reaction and solvent. While many other papers focus on small homogeneous datasets containing few reaction families and solvents, our model should have much broader applicability. The simple input representation avoids the need for expensive quantum chemistry calculations and makes our model well-suited for high-throughput tasks, e.g., quickly refining detailed kinetic models with tens of thousands of reactions [117] or identifying low-barrier reactions produced from an automated enumeration [118–121]. Following best practices in ML, our training procedure also uses proper data splits to estimate the model's performance in both interpolation and extrapolation regimes.

6.2 Methods

6.2.1 Dataset Generation

This dataset spans a diverse set of radical reactions. The vast majority of the reactions are generated from a closed-shell species, such as common vinylic monomer-derived esters, ethers, alcohols, and acids, reacting with a radical species and thus are bimolecular in the forward direction. These reactions may be of interest for radical polymerization [3] or degradation of active pharmaceutical ingredients [7]. A small percentage (68 reactions) started from one radical species and thus are unimolecular in the forward direction. The reaction generation logic is described further in the Supporting Information. Unless specified otherwise, all calculations are done using Turbomole [122, 123]; version 7.2.1 is used for the MGSM calculations while version 7.5.2 is used for subsequent calculations. The balanced reactions involve the elements H, C, N, O, and S and contain canonical atom-mapped SMILES. All species are neutrally charged. Each TS contains up to 19 heavy atoms and is calculated in the doublet state.

The reactions are generated using a combinatorial reaction rule (i.e., driving coordinate) enumerator in combination with the single-ended molecular growing string method (MGSM) for reaction path optimizations [124]. First, starting from the SMILES, the connectivity of the reactant structure(s) is analyzed and driving coordinates for subsequent MGSM calculations are generated. The reaction rule enumerator generates driving coordinates for the MGSM calculations based on a combinatoric enumeration of bond breakages and bond formations in the "ZStruct-like" manner [125] with constraints on the number of allowed bond changes. Since elementary reactions typically involve few bond changes, we specify that at most two bonds could be broken, at most two bonds could be formed, and a total of at most three bonds could be changed. Here, a bond only considers connections via an adjacency matrix rather than bond orders. Additional constraints are set for the reactive (radical) centers. Second, a van der Waals complex of the reactant structure(s) is generated and aligned for each driving coordinate using the workflow described in ref. [125]. This reactant complex is the starting point for the growing string search. Note that in the MGSM optimization, the driving coordinates are projected onto the nonredundant delocalized internal coordinates [126], a single tangent vector that represents all of the driving coordinates simultaneously. Importantly, this projected dimension allows other coordinates to change without constraint during the reaction path optimization, so the limits for breaking and forming bonds only apply to the initial search direction. The entire reaction path is optimized towards a path resembling a minimum energy path in the TS region using the setting specified in the Supporting Information. Finally, the reactant and product complexes are separated into individual species.

All minima and saddle-point structures are optimized using the B3LYP functional [127–131] with D3 dispersion corrections [132] and Becke-Johnson damping [133, 134] along with a triple-ζ polarized split-valence basis set [135] i.e., B3LYP-D3(BJ)/def2-TZVP. These calculations employed an m4 integration grid and tight convergence settings i.e., energy 8, gcart 4, scfconv 8 and denconv 10^{-8}. For saddle-point optimizations, the eigenvector-following algorithm as implemented in Turbomole was used. All optimizations and single-point energy calculations described in this section employed the resolution-of-identity (RI) approximation for the Coulomb integrals using matching default auxiliary basis sets [136, 137].

For each reactant, TS, and product structure, conformational searches are performed using a BASF in-house tool that iteratively scans dihedral angles around all rotatable bonds that are not part of a ring. For ring systems of all species, including those that formed on-the-fly in saddle-point structures (e.g., for an intramolecular hydrogen atom transfer), full sets of conformers (accessible through rotations around bonds) are considered. For saddle-point structures, the bonds broken and formed during the reaction are kept frozen during conformer search. This set of conformers is subsequently optimized using the meta-GGA functional [138] TPSS/def2-TZVP using the COSMO continuum solvation model [28] (with COSMOtherm version 18) with a dielectric constant of $\epsilon = \infty$. TPSS has been shown to give good geometries for organic species [139]. For structures containing -OH, -NH, or -COOH groups, the conformer search procedure was applied not only to original input structure, but also to two more conformer structures that were generated from the original input structures by random multiple rotations around all dihedral angles not part of a ring system. The conformer with the overall lowest electronic energy at the level of the conformer-search method was taken as the best conformer in all solvents. Due to computational constraints, we could not do a thorough conformer search in every solvent of interest. However, the conformer search in the COSMO phase should capture generic solvent effects and result in more realistic energy rankings than a conformer search in the gas phase.

This lowest-energy conformer is re-optimized at the B3LYP-D3(BJ)/def2-TZVP level of theory with the same settings as the pre-optimization procedure described above. Harmonic vibrational frequencies were computed based on second-order analytical derivatives [140]. We verified that the optimized minima contain no imaginary vibrational frequency. The single-point energies are refined using the M06-2X functional [141] with the quadruple-zeta def2-QZVP basis set [135], which is known to give reliable barrier heights [114, 142, 143]. Since only the lowest-energy conformer from the COSMO phase is used, there is likely some error introduced in the calculated Gibbs free energies of activation and of reaction in comparison to doing the optimization, frequency, and single-point energy calculation on all possible conformers and then performing a Boltzmann ensembling to get the conformer population Gibbs free energy.

All TS geometries are verified by a forward and reverse IRC calculation, which link the saddle-point to its adjacent reactant and product potential energy surface minima by a mass-weighted downhill optimization [144]. Gaussian's IRC algorithm is well-established, so all IRCs are run with Gaussian 16 [145] using the same level of DFT as was used during the optimization and frequency calculations. The IRC endpoints are parsed using OpenBabel [146], which infers connectivity using its "connect-the-dots" method by adding bonds to atoms closer than their combined covalent radii while maintaining minimum distance and valence constraints. The TS is considered valid if the adjacency list matches between the IRC endpoints and the original reactants and products which were optimized in the previous step. All reactions published here satisfy this criteria. We also ensure that spin and charge are conserved for all reactions. All TSs have exactly one imaginary frequency and show no substantial spin contamination i.e., the expectation value of the spin squared operator ($<S^2>$) did not deviate by more than 0.1 from the expected value of 0.75. Additional analysis can be found in Section 6.6.1.3 in the Appendix.

Finally, the Gibbs free energy of solvation (ΔG_{solv}) at 298.15 K is calculated using COSMO-RS theory[38, 39] as implemented in the COSMOtherm 18 software package[147]. COSMO-RS has proven to be a very valuable tool for predicting solvation effects on kinetic parameters [148, 149]. The `BP_TZVP_18` parameterization is employed, which is based on DFT single-point calculations at the BP86[150, 151]/def-TZVP[152] (COSMO, $\epsilon = \infty$) level of theory using Turbomole 7.5.2. Corresponding calculations with COSMO switched off are performed as well to obtain gas-phase energies which are explicitly passed to COSMOtherm for the calculation of the transfer energies from gas phase to the ideal-conductor COSMO medium. D3 dispersion corrections with Becke-Johnson damping [132] are active in both BP86 calculations, though this detail influences neither the calculated screening charge distributions nor the transfer energies. Although the `BP_TZVPD_FINE_18` parameterization should yield slightly more accurate values for individual molecules [153], any systematic errors from `BP_TZVP_18` should largely cancel when doing the subtraction to obtain $\Delta G^{\ddagger}_{\mathrm{soln}}$ and $\Delta G^{\mathrm{rxn}}_{\mathrm{soln}}$. In our experience, the `BP_TZVP_18` method is generally more robust and less likely to yield outliers. Indeed, Chung and Green[149] show that the errors in calculated $\Delta G^{\ddagger}_{\mathrm{soln}}$ relative to experimental values are actually slightly lower for `BP_TZVP_18` while the `BP_TZVPD_FINE_18` method performed only about 0.1 kcal mol^{-1} better for estimating relative Gibbs free energies between solvents.

The COSMOtherm simulations are set up with the solutes at infinite dilution in the solvent of interest. From the vapor pressure of the solute p_{vap} predicted by COSMO-RS, one obtains the Gibbs free energy of solvation for the reference state extrapolated to a mole fraction of unity,

$$\Delta G_{\mathrm{solv}}^{x=1} = RT\ ln\left(\frac{p_{vap}}{p_{ref}}\right), \tag{6.1}$$

where the reference pressure $p_{ref} = 1$ bar is identical to the ideal-gas vapor pressure that is employed for the calculation of the gas-phase thermodynamic potentials within Turbomole's "freeh" utility. A practically more useful quantity is the Gibbs free energy of solvation corresponding to a reference state of $c = 1$ mol/L in the liquid phase (i.e., the molar reference state), which is obtained from $\Delta G_{\mathrm{solv}}^{x=1}$ via the following relationship

$$\Delta G_{\mathrm{solv}}^{1\ \mathrm{mol/L}} = \Delta G_{\mathrm{solv}}^{x=1} + RT\ ln\left(\frac{1\ \mathrm{mol/L}}{c_{solvent}}\right) \tag{6.2}$$

such that $c_{solvent}$ is the molar concentration of the pure solvent. In the remainder of this paper, the notation ΔG_{solv} refers to the molar reference state, thus the superscript of 1 mol/L will be omitted. The solvent molarities have been estimated from a simple linear correlation, using the COSMO cavity volumes and molecular weights as input. The values used for $c_{solvent}$ are listed in the Appendix.

6.2.2 Regression Targets

For each minima and TS, the Gibbs free energy in solution is defined as

$$\begin{aligned} G_{\mathrm{soln}} &= E_{\mathrm{electronic}} + \mathrm{ZPE} + G_{\mathrm{RRHO}} + \Delta G_{\mathrm{solv}} && (6.3)\\ &= G_{\mathrm{gas}} + \Delta G_{\mathrm{solv}} && (6.4) \end{aligned}$$

such that $E_{\mathrm{electronic}}$ is the electronic energy, ZPE is the zero-point energy, G_{RRHO} is the Gibbs free energy due to gas-phase thermal contributions with standard rigid-rotor harmonic oscillator (RRHO) approximations, and ΔG_{solv} is the change in solvation free energy due to the solute-solvent interaction. Unscaled harmonic frequencies have been used for the ZPE and thermal corrections. Tunneling corrections are not considered in this work. The Gibbs free energy of activation and of reaction in the gas phase are defined as

$$\Delta G_{\mathrm{gas}}^{\ddagger} = G_{\mathrm{gas}}^{\mathrm{TS}} - \sum_{i=1}^{N\ \mathrm{reactants}} G_{\mathrm{gas}}^{\mathrm{R}_i} \tag{6.5}$$

$$\Delta G_{\text{gas}}^{\text{rxn}} = \sum_{j=1}^{N \text{ products}} G_{\text{gas}}^{\text{P}_j} - \sum_{i=1}^{N \text{ reactants}} G_{\text{gas}}^{\text{R}_i} \tag{6.6}$$

The solvation correction to the Gibbs free energy of activation and of reaction are defined as

$$\Delta\Delta G_{\text{solv}}^{\ddagger} = \Delta G_{\text{solv}}^{\text{TS}} - \sum_{i=1}^{N \text{ reactants}} \Delta G_{\text{solv}}^{\text{R}_i} \tag{6.7}$$

$$\Delta\Delta G_{\text{solv}}^{\text{rxn}} = \sum_{j=1}^{N \text{ products}} \Delta G_{\text{solv}}^{\text{P}_j} - \sum_{i=1}^{N \text{ reactants}} \Delta G_{\text{solv}}^{\text{R}_i} \tag{6.8}$$

such that $\Delta G_{\text{solv}}^{\text{TS}}$, $\Delta G_{\text{solv}}^{\text{R}_i}$, and $\Delta G_{\text{solv}}^{\text{P}_j}$ represent the solvation free energies of a TS, reactant i, and product j respectively. Graphically, the effect of solvation free energy is shown in Figure 6.1.

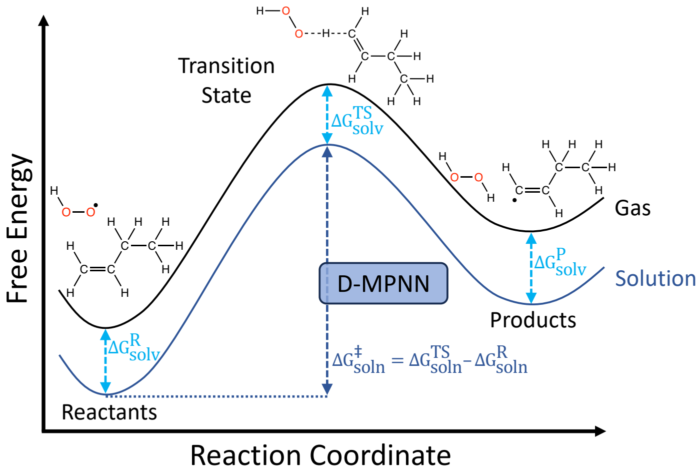

Free Energy
Reaction Coordinate
Transition State
Reactants
Products
Gas
Solution
D-MPNN
$\Delta G_{\text{solv}}^{\text{TS}}$
$\Delta G_{\text{solv}}^{\text{R}}$
$\Delta G_{\text{solv}}^{\text{P}}$
$\Delta G_{\text{soln}}^{\ddagger} = \Delta G_{\text{soln}}^{\text{TS}} - \Delta G_{\text{soln}}^{\text{R}}$

Figure 6.1. Potential energy diagram of a reaction in the gas phase and in solution phase.

Our primary interest is in $\Delta G_{\text{soln}}^{\ddagger}$ so all baseline models are trained to predict this quantity. However, our final model is co-trained to predict $\Delta G_{\text{gas}}^{\ddagger}$, $\Delta G_{\text{gas}}^{\text{rxn}}$, $\Delta\Delta G_{\text{solv}}^{\ddagger}$, and $\Delta\Delta G_{\text{solv}}^{\text{rxn}}$. These values can be added together to give the Gibbs free energy of activation and of reaction in solution, which are defined as

$$\Delta G^{\ddagger}_{\text{soln}} = G^{\text{TS}}_{\text{soln}} - \sum_{i=1}^{N\ \text{reactants}} G^{\text{R}_i}_{\text{soln}} \tag{6.9}$$

$$= \Delta G^{\ddagger}_{\text{gas}} + \Delta\Delta G^{\ddagger}_{\text{solv}} \tag{6.10}$$

$$\Delta G^{\text{rxn}}_{\text{soln}} = \sum_{j=1}^{N\ \text{products}} G^{\text{P}_j}_{\text{soln}} - \sum_{i=1}^{N\ \text{reactants}} G^{\text{R}_i}_{\text{soln}} \tag{6.11}$$

$$= \Delta G^{\text{rxn}}_{\text{gas}} + \Delta\Delta G^{\text{rxn}}_{\text{solv}} \tag{6.12}$$

6.2.3 Model Architecture

As discussed in detail in Chapter 8, it is desirable to produce models that operate on simple user-friendly inputs. Thus, all models presented here use SMILES, rather than 3D coordinates, as the source of chemical representation. Although inexpensive semi-empirical methods [154–156] offer substantial time-savings, operating on SMILES offers a few practical advantages when making predictions on new reactions. First, it lets users avoid the extra step of performing geometry optimizations as well as conformer searches, which can be quite challenging for large flexible molecules. It is also more amenable for high-throughput screening (e.g. of millions of candidate reactions) to occur on standard consumer computers since few resources are needed during inference. Language models operate only on SMILES, so in principal they could also satisfy the desire to operate on simple inputs. However, previous language models that were fine-tuned to predict solvation free energy of individual molecules require 7.6M [53] to 48.8M parameters [157]. In comparison, our final graph network used for Table 6.4 has only around 650,000 parameters, which leads to more reasonable resource requirements during model training and inference. Our baseline directed message passing neural network (D-MPNN) used for Table 6.2 and Table 6.3 has only about 340,000 parameters.

6.2.3.1 D-MPNN

When using data-driven methods to model chemistry, graph neural networks (GNNs) are a natural choice [158]. Molecules can be modeled as mathematical graphs such that atoms are graph nodes and bonds are graph edges. Each atom and bond is assigned an initial feature vector whose representation is updated with information from neighboring nodes and/or edges to create a learned molecular representation that is passed to a standard feed forward network (FFN) to predict the property of interest. Empirically, the learned representations from graph networks often outperform traditional ML models operating on fixed fingerprint representations [49, 54, 73–108], particularly when predicting kinetic properties.

Here, we adapt the directed message passing neural network (D-MPNN) developed by Yang et al. [54], which is a type of GNN that passes messages across directed bonds. To encode the reaction, we use the established condensed graph of reaction (CGR) representation as the input since it has been shown to outperform other representations for various reaction property predictions [73, 159–161]. This outperformance is partially expected since the CGR is a superposition of the reactant and product graphs and thus resembles a 2D-structure of the reaction's TS. As shown in Figure 6.2, this architecture builds a learned reaction embedding by aggregating atomic representations after the message passing phase. This involves summing the vectors and then dividing by a constant for numerical stability. As done by ref. [61], a separate D-MPNN is used to create a learned molecular embedding of the solvent, which is concatenated to the reaction embedding and passed through an FFN to predict the regression targets. For both the CGR and solvent graph, the initial atom and bond features are generated using RDKit [71] and include atomic symbol, formal charge, hybridization, bond order, etc. Section 6.6.4 in the Appendix contains additional detail about the network, training procedure, and hyperparameter optimization. Our modified code and final weights are freely available on GitHub [162] under the `solvated_barrier_heights` branch of our forked repository.

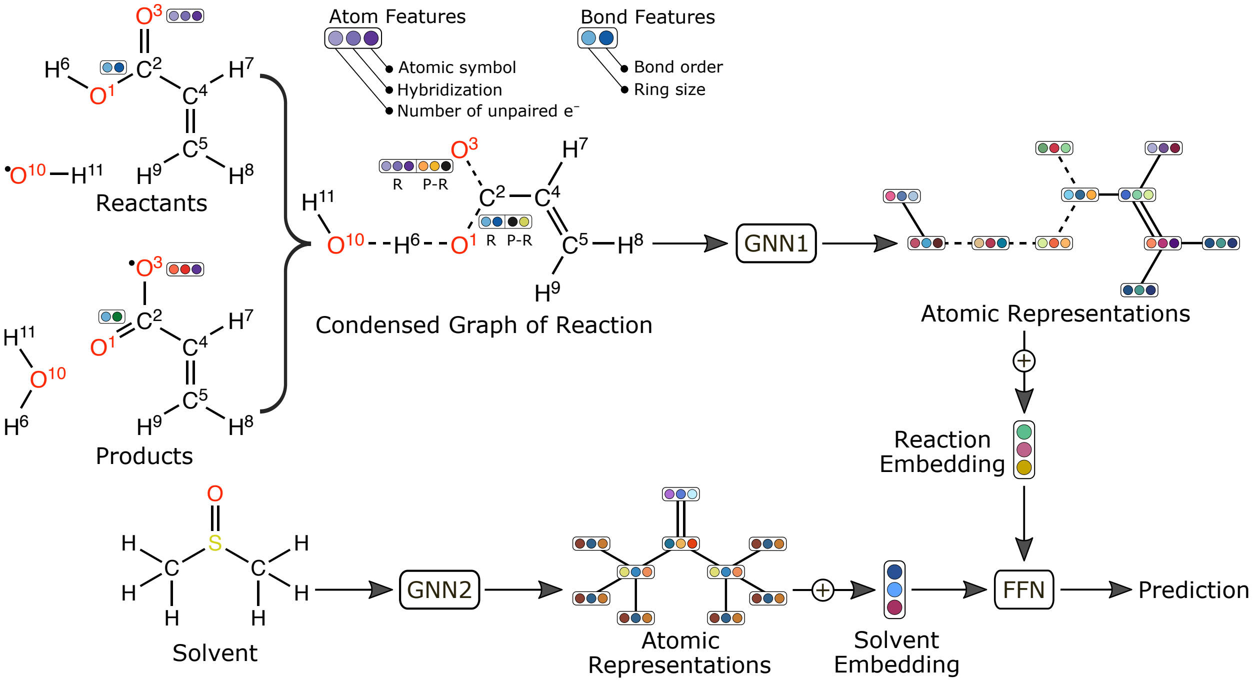


Figure 6.2. Schematic of the machine learning model architecture. Here, the GNN uses the directed message passing scheme introduced by Yang et al. [54] The condensed graph of reaction (CGR) representation concatenates the featurization of the reactant complex with the difference of the featurization between the reactant and product complex. A subset of the initial atom and bond features are shown for simplicity. The superscripts indicate the arbitrary atom-map numbers. Colors are for qualitative purposes only. Only one resonance structure for the radical product is shown here, but multiple resonance structures are included to augment the dataset while training the model.

Our final models use the full architecture from Figure 6.2, which we will call D-MPNN-2 to indicate that two D-MPNNs are present. This model is co-trained to predict the Gibbs free energy of activation and of reaction in the gas phase as well as the corresponding solvation correction i.e., $\Delta G^{\ddagger}_{\text{gas}}$, $\Delta G^{\text{rxn}}_{\text{gas}}$, $\Delta\Delta G^{\ddagger}_{\text{solv}}$, and $\Delta\Delta G^{\text{rxn}}_{\text{solv}}$. As shown by Eq. 6.10 and 6.12, these terms can added together to yield the Gibbs free energy in solution. This approach allows the model to be more interpretable in comparison to directly predicting $\Delta G^{\ddagger}_{\text{soln}}$ and $\Delta G^{\text{rxn}}_{\text{soln}}$. For example, $\Delta\Delta G^{\ddagger}_{\text{solv}}$ could be used to calculate the ratio of a gas phase rate coefficient (k_{gas}) to a liquid phase rate coefficient (k_{liq}) via the following expression:

$$\frac{k_{\text{liq}}}{k_{\text{gas}}} = \exp\left(\frac{-\Delta\Delta G^{\ddagger}_{\text{solv}}}{RT}\right) \tag{6.13}$$

such that R is the universal gas constant and T is the temperature. It could be also used to calculate the relative rate coefficient between two solvents:

$$k_{\mathrm{rel}} = \frac{k^{\mathrm{s1}}_{\mathrm{liq}}}{k^{\mathrm{s2}}_{\mathrm{liq}}} = \exp\left(-\frac{\Delta\Delta G^{\ddagger}_{\mathrm{solv},s1} - \Delta\Delta G^{\ddagger}_{\mathrm{solv,s2}}}{RT}\right) \tag{6.14}$$

such that $k^{\mathrm{s1}}_{\mathrm{liq}}$ and $k^{\mathrm{s2}}_{\mathrm{liq}}$ are the rate coefficient of a reaction in a solvent 1 and in solvent 2, respectively, and $\Delta\Delta G^{\ddagger}_{\mathrm{solv},s1}$ and $\Delta\Delta G^{\ddagger}_{\mathrm{solv,s2}}$ are the corresponding solvation Gibbs free energies of activation for the reaction in each solvent.

Our baseline comparisons will only use one solvent for simplicity, so the solvent GNN is not needed. We call this architecture D-MPNN-1 to indicate that only one D-MPNN is used. This baseline model is trained to predict $\Delta G^{\ddagger}_{\mathrm{soln}}$ in water.

6.2.3.2 Baseline Models

We also include simpler models from Scikit-Learn [163] as baselines, such as random forest (RF), multi-layer perceptron (MLP), and support vector regression (SVR) with either a linear or radial basis function (RBF) kernel. We also use extreme gradient boosting (XGB) from the XGBoost package [164]. The best hyperparameters are chosen via Optuna [165] and are listed in Section 6.6.3 in the Appendix. Since these methods operate on vector inputs, we use Morgan (ECFP) fingerprints [70] with 2048 bit hashing and radius of 2 as calculated using RDKit [71]. We also tried three other vector representations from RDKit: Molecular ACCess System (MACCS) [166], Avalon [167], and Atom-pair [168]. To represent the reaction, we concatenate the fingerprint of the reactant complex with the difference of the fingerprint between the reactant and product complex; this is the same approach as the initial featurization in the CGR. In addition to the baseline regression models, we also add a trivial approach that simply predicts the mean of the training target values for all test reactions [169].

6.2.4 Data Splits

Although random splitting, stratified sampling [170], and furthest point sampling[171] are frequently used in the literature, these all create relatively simple learning tasks that measure interpolation performance. However, given the vastness of chemical space [172–174] and its often unsmooth nature (e.g. activity cliffs), it is prudent to evaluate model performance on more challenging extrapolative splits as well. One idea is to use solvent-based splits in which the model sees molecules or reactions in some solvents during training and then performance is tested with the same molecules or reactions in new solvents. However, given that ΔG_{solv} values often have relatively small magnitudes for neutral solutes regardless of solvent [31, 36, 175, 176], solvent-based splits are likely not as challenging as they would appear. Indeed, previous work has shown that solute-based splits typically represent a more challenging task than solvent-based splits [53, 61, 177], so we focus on solute-based splits in this work.

Here, we create both interpolative splits and extrapolative splits based on the procedure in the `astartes` software package [178] so future users are informed of the likely performance in these different settings. Since data-driven methods typically benefit from additional training points, all data splits include reactions from the forward and reverse direction, which creates ~470,000 data points. To obtain the Gibbs free energy in gas phase, solution phase, and of solvation for the reverse reactions, we use the following expressions:

$$\Delta G^{\ddagger}_{\text{reverse}} = \Delta G^{\ddagger}_{\text{forward}} - \Delta G^{\text{rxn}}_{\text{forward}} \tag{6.15}$$

$$\Delta G^{\text{rxn}}_{\text{reverse}} = -\Delta G^{\text{rxn}}_{\text{forward}} \tag{6.16}$$

Splitting is done based on the forward reactions; the corresponding reverse reaction is then added to the same set. Otherwise, the same transition state would appear in both the training and testing set, which would cause the testing error to not reflect the true performance when evaluating a new reaction [73, 179]. Preliminary tests showed that the model was sensitive to the resonance structure passed as input so we additionally augment the dataset by sampling multiple resonance structures when applicable. This creates approximately 720,000 data points in total.

To measure interpolation, we use random splitting to assign 85% of the data to training, 5% to validation, and 10% to testing. To measure extrapolation, such as making predictions for new types of molecules, we cluster the data based on reactant substructure. Splitting based on the Bemis-Murcko scaffold [180] is a common approach for this task [49, 54, 73, 84, 87, 88, 100, 104, 157, 179, 181–183]. However, about 80% of the molecules in our dataset do not match any Bemis-Murcko ring scaffolds, which precludes any attempt to cluster the reactions in this way. Instead, we use a different procedure (K-means splits) with three steps to accomplish our goal of creating more challenging splits with chemically dissimilar compounds. Our approach is conceptually similar to leave-one-cluster-out cross-validation [184], though many other papers have taken similar approaches to measure performance on chemically dissimilar data splits [51, 67, 185–194] or quantify domains of model applicability [195–198].

We first create Morgan fingerprints [69, 70] with 2048 bit hashing and radius of 2 for each reactant complex from the forward reactions using RDKit [71]. Since these vectors are quite sparse, we use Principal Component Analysis (PCA) to project these vectors to a more dense subspace. We next use K-means clustering to create twelve clusters. Finally, we assign ten clusters to the training set, one cluster to the validation set, and one cluster to the testing set to closely approximate an 85:5:10 split. Additional analysis is shown in Section 6.6.1 in the Appendix. Five folds are created for both interpolative and extrapolative splits; the subsequent results report the mean and standard deviation of test predictions from the five folds. All regression targets are Z-scored (i.e., subtract the mean and divide by standard deviation) during training for numerical stability.

6.3 Results and Discussion

6.3.1 Dataset Statistics

The final dataset contains 5911 reactions occurring in 40 popular solvents summarized in Table 6.1. The provided dielectric constant is the average of all values from ref. [199], which compiles results from many sources; some values are taken from ref. [200] and [201]. The dielectric constant represents the charge-screening ability of a solvent and serves as a one-dimensional proxy for its polarity. Note that COSMO-RS does not use dielectric constants as input parameters.

Table 6.1. Solvents used from COSMO-RS. The tabulated dielectric constant is the average of all values from the corresponding reference at the specified temperature.

Solvent name	Chemical formula	Dielectric constant	Temp. (K)	Reference
1,4-dioxane	$C_4H_8O_2$	2.21	298.2	186
2-ethylhexanol	$C_8H_{18}O$	7.58	298.2	186
Acetic acid	$C_2H_4O_2$	6.19	298.2	186
Acetone	C_3H_6O	20.77	298.2	186
Acetonitrile	C_2H_3N	36.26	298.2	186
Aniline	C_6H_7N	6.91	298.2	186
Benzaldehyde	C_7H_6O	17.09	298.2	186
Benzene	C_6H_6	2.27	298.2	186
Caprolactam	$C_6H_{11}NO$	2.80	298.2	187
Carbon disulfide	CS_2	2.63	298.2	186
Chlorobenzene	C_6H_5Cl	5.61	298.2	186
Chloroform	$CHCl_3$	4.73	298.2	186
Cyclohexane	C_6H_{12}	2.02	298.2	186
Cyclopentane	C_5H_{10}	1.96	298.2	186
n-Decane	$C_{10}H_{22}$	1.99	293.2	186
Dichloromethane	CH_2Cl_2	8.93	298.0	186
Diethyl ether	$C_4H_{10}O$	4.22	298.2	186
Diethyl sulfide	$C_4H_{10}S$	5.72	298.2	186
Dimethyl formamide	C_3H_7NO	37.04	298.2	186
DMSO	C_2H_6OS	46.56	298.2	186
Ethanol	C_2H_6O	24.41	298.2	186
Ethyl acetate	$C_4H_8O_2$	6.06	293.2	186
Ethyl acrylate	$C_5H_8O_2$	5.85	301.2	186
Furan	C_4H_4O	2.95	298.2	186
Glycerol	$C_3H_8O_3$	45.72	298.2	186
Isopropanol	C_3H_8O	19.24	298.2	186
n-Hexane	C_6H_{14}	1.88	298.2	186
Methanol	CH_4O	32.51	298.2	186
Nitromethane	CH_3NO_2	37.38	298.2	186
Octanol	$C_8H_{18}O$	10.00	298.2	186
Phenol	C_6H_6O	13.40	293.2	186
Phenyl isocyanate	C_7H_5NO	8.94	293.2	186
Piperidine	$C_5H_{11}N$	4.18	298.2	186
Pyridine	C_5H_5N	13.45	298.2	186
Pyrrolidine	C_4H_9N	8.30	293.0	186
Tetrahydrofuran	C_4H_8O	7.62	298.2	186
Toluene	C_7H_8	2.39	298.2	186
Triethylamine	$C_6H_{15}N$	2.42	298.2	186
γ-valerolactone	$C_5H_8O_2$	36.47	298.2	188
Water	H_2O	78.39	298.2	186

As seen in Figure 6.3a, the reactions span a wide range of activation and reaction energies. Thus, this dataset is suitable for training a model to act as a general purpose estimator that has been exposed to both high-barrier and low-barrier reactions. The data generally follows the expected Evans-Polanyi relationship [202] in which the activation energy is positively correlated with the reaction energy. However, the correlation is mild at only 0.47 due to the bimodal distribution, with the distribution on the right side of Figure 6.3a containing many thermoneutral yet high-barrier reactions. Consistent with other published datasets [179, 203] that were generated using a combinatorial driving coordinate generator together with the MGSM method for reaction path optimization, there is large reaction diversity. Most reactions do not match any of the manually-generated templates from the Reaction Mechanism Generator (RMG) database [112]. Of the reactions that did match RMG templates, the vast majority belong to hydrogen abstraction and R-addition multiple bond (i.e., radical attacking a higher-order bond), with additional details shown in Section 6.6.1 in the Appendix.

Although Figure 6.3a only shows the ΔG_{soln} values for water, the distribution is quite similar for the gas phase as well as for the other solvents. This is because solvation corrections are often relatively small, typically just 1-5 kcal mol^{-1} regardless of the solvent [31, 36, 175, 176]. Indeed, 98% of all values for $\Delta\Delta G^{\ddagger}_{\text{solv}}$ and $\Delta\Delta G^{\text{rxn}}_{\text{solv}}$ values in this work are within ±5 kcal mol^{-1} across all solvents. Figure 6.4 shows the distribution of $\Delta\Delta G_{\text{solv}}$ for select solvents. The distributions for more polar solvents have a larger range, i.e., outliers have a larger magnitude for $\Delta\Delta G_{\text{solv}}$. However, the average values are still very close to zero.

Figure 6.3b shows the broad distribution of the number of heavy atoms. Each reaction can be classified by a general reaction template of “breaks X bonds and forms Y bonds”, which is abbreviated as “bXfY”; this analyzes changes in connectivity, not bond order. As shown in Figure 6.3c, most reactions have two bonds changing (i.e. breaking or forming a total of two bonds). The distribution of which types of bonds are changing is shown in Figure 6.3d; changes to C-H and C-C bonds are the most common bond changes, which is expected given that carbon is the most common heavy atom in this dataset. Additional dataset analysis is provided in Section 6.6.1.

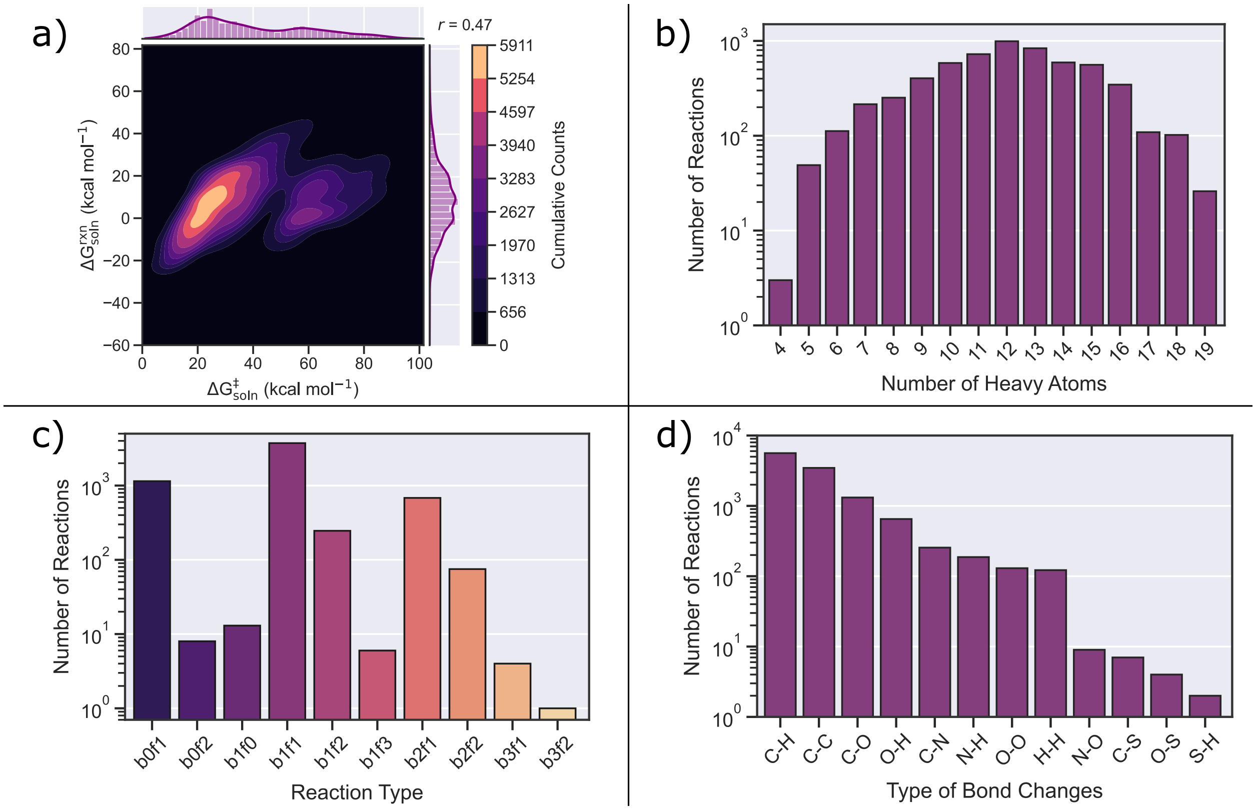


Figure 6.3. Summary of reaction characteristics. (a) Heatmap showing distribution of $\Delta G^{\ddagger}_{\text{soln}}$ and $\Delta G^{\text{rxn}}_{\text{soln}}$ for all reactions occurring in water. (b) Distribution of the number of heavy atoms. (c) Distribution of reaction type, classified as "breaking X bonds and forming Y bonds". Warmer colors generally denote a larger activation energy as breaking more bonds requires additional energy. (d) Distribution of bond changes occurring during the reactions.

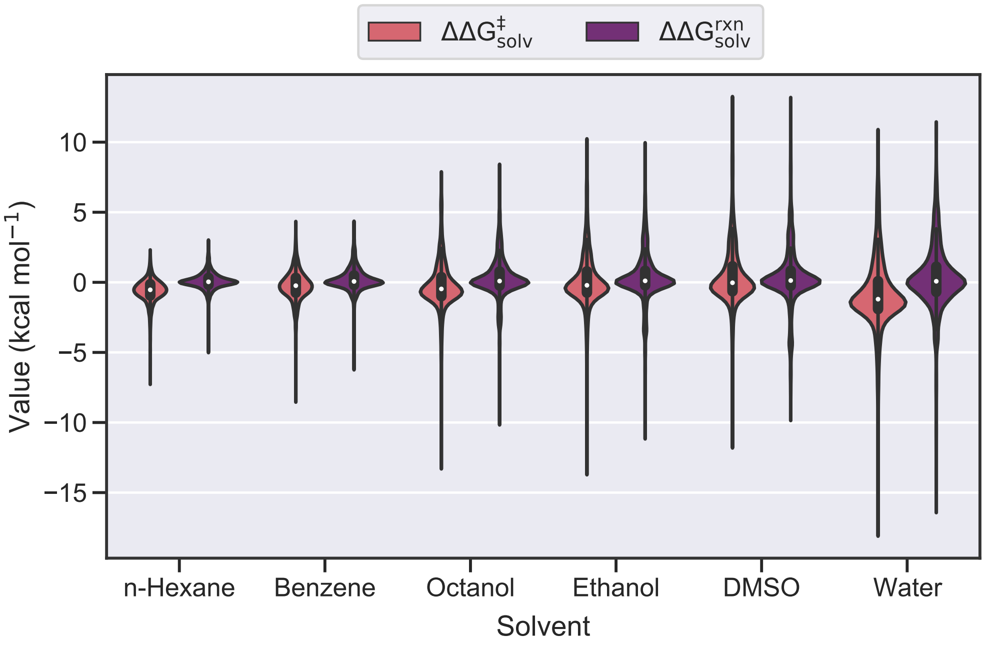


Figure 6.4. Distribution of solvation correction for select solvents. Dielectric constant increases from left to right.

6.3.2 Baseline Model Results

To determine which modeling approach to focus on, we first compare the performance of all models for predicting $\Delta G^{\ddagger}_{\text{soln}}$ in water. We also include the trivial baseline that predicts the mean value of $\Delta G^{\ddagger}_{\text{soln}}$ from the training set for all test reactions. For this preliminary task, the D-MPNN model only uses GNN1 from Figure 6.2 since the solvent is constant. Although both XGB and the MLP perform fairly well, the results from Table 6.2 and Table 6.3 show that learned representations from graphs vastly outperform all baseline models for both random splits and more challenging K-means splits; this result is consistent with many published references [49, 54, 73–108]. Further, the D-MPNN is much more robust as the average performance is very similar for both types of data splits. In contrast, the baseline models perform substantially worse when presented with more challenging data splits. The poor performance of the baseline mean predictor is expected given the relatively large variance in the target values as shown in Figure 6.3a. It is interesting to note that the Avalon and MACCS featurization consistently yield poor performance for the models and prediction task studied in this work.

Table 6.2. Testing errors (mean ± 1 standard deviation from the five folds) from various models for predicting $\Delta G^{\ddagger}_{\mathrm{soln}}$ in water when using random reaction splits. MAE and RMSE have units of kcal mol^{-1}.

Model	Featurization	MAE	RMSE	R^2
Baseline (mean)	N/A	17.15 ± 0.29	20.29 ± 0.23	0.00 ± 0.00
SVR (Linear)	AtomPair	8.29 ± 0.20	11.96 ± 0.33	0.65 ± 0.02
	Avalon	12.63 ± 0.10	16.61 ± 0.18	0.33 ± 0.02
	MACCS	12.12 ± 0.20	16.06 ± 0.37	0.37 ± 0.03
	Morgan	4.61 ± 0.24	8.08 ± 0.60	0.84 ± 0.02
SVR (RBF)	AtomPair	11.24 ± 0.58	14.31 ± 0.91	0.50 ± 0.06
	Avalon	12.85 ± 0.28	15.83 ± 0.42	0.39 ± 0.03
	MACCS	13.62 ± 0.63	16.65 ± 0.81	0.33 ± 0.06
	Morgan	10.97 ± 0.15	13.54 ± 0.37	0.56 ± 0.02
MLP	AtomPair	3.94 ± 0.22	7.07 ± 0.41	0.88 ± 0.02
	Avalon	5.30 ± 0.13	8.77 ± 0.30	0.81 ± 0.01
	MACCS	6.14 ± 0.20	10.20 ± 0.36	0.75 ± 0.02
	Morgan	3.38 ± 0.23	6.60 ± 0.67	0.89 ± 0.02
RF	AtomPair	4.90 ± 0.20	8.05 ± 0.43	0.84 ± 0.02
	Avalon	7.94 ± 0.20	11.50 ± 0.34	0.68 ± 0.02
	MACCS	8.76 ± 0.25	12.49 ± 0.22	0.62 ± 0.02
	Morgan	7.14 ± 0.11	10.26 ± 0.32	0.74 ± 0.02
XGB	AtomPair	3.82 ± 0.22	6.88 ± 0.39	0.89 ± 0.01
	Avalon	7.60 ± 0.21	10.71 ± 0.34	0.72 ± 0.02
	MACCS	6.78 ± 0.32	10.29 ± 0.46	0.74 ± 0.03
	Morgan	5.62 ± 0.16	8.23 ± 0.45	0.84 ± 0.02
D-MPNN-1	**Learned**	**2.53 ± 0.22**	**5.26 ± 0.38**	**0.93 ± 0.01**

Table 6.3. Testing errors (mean ± 1 standard deviation from the five folds) from various models for predicting $\Delta G^{\ddagger}_{\text{soln}}$ in water when using K-means splits. MAE and RMSE have units of kcal mol^{-1}.

Model	Featurization	MAE	RMSE	R^2
Baseline (mean)	N/A	19.23 ± 2.17	22.18 ± 2.04	-0.10 ± 0.10
SVR (Linear)	AtomPair	10.80 ± 1.95	13.97 ± 2.59	0.55 ± 0.19
	Avalon	17.69 ± 2.95	21.82 ± 2.99	-0.06 ± 0.18
	MACCS	15.98 ± 2.54	19.36 ± 2.38	0.17 ± 0.12
	Morgan	5.95 ± 2.80	9.96 ± 3.95	0.74 ± 0.23
SVR (RBF)	AtomPair	16.53 ± 3.32	19.75 ± 3.90	0.08 ± 0.43
	Avalon	15.90 ± 1.73	19.36 ± 1.95	0.16 ± 0.12
	MACCS	16.22 ± 1.86	19.19 ± 1.58	0.18 ± 0.06
	Morgan	14.74 ± 2.33	17.81 ± 2.60	0.29 ± 0.17
MLP	AtomPair	5.29 ± 0.96	8.43 ± 1.22	0.84 ± 0.05
	Avalon	7.62 ± 1.44	11.20 ± 1.62	0.71 ± 0.08
	MACCS	9.31 ± 1.52	13.42 ± 1.62	0.59 ± 0.09
	Morgan	4.10 ± 1.27	7.79 ± 1.65	0.86 ± 0.07
RF	AtomPair	7.03 ± 1.39	9.96 ± 1.48	0.77 ± 0.07
	Avalon	10.08 ± 0.91	13.95 ± 1.33	0.56 ± 0.07
	MACCS	11.91 ± 2.07	15.53 ± 2.05	0.45 ± 0.15
	Morgan	9.71 ± 1.33	13.03 ± 1.61	0.62 ± 0.08
XGB	AtomPair	5.49 ± 0.68	8.48 ± 0.75	0.84 ± 0.03
	Avalon	9.09 ± 1.06	12.66 ± 1.35	0.64 ± 0.07
	MACCS	11.47 ± 1.90	14.65 ± 1.85	0.51 ± 0.13
	Morgan	6.39 ± 1.22	9.11 ± 1.20	0.81 ± 0.05
D-MPNN-1	**Learned**	**2.89 ± 0.45**	**5.55 ± 1.02**	**0.93 ± 0.02**

6.3.3 D-MPNN Results

Based on the results from Section 6.3.2, we focus on training a D-MPNN using all data i.e., nearly 720,000 entries from the forward and reverse reactions in all 40 solvents. The testing errors when training on both random reaction splits and K-Means splits are shown in Table 6.4. Similar to the results observed in Tables 6.2 and 6.3, the D-MPNN performance is robust for both split types.

When analyzing the predictions for $\Delta G^{\ddagger}_{\text{soln}}$ on the test set from the K-Means splits, the MAE is 3.00 ± 0.44 kcal mol^{-1} and RMSE is 5.99 ± 0.91 kcal mol^{-1}, such that the bounds correspond to one standard deviation calculated across five folds. Thus, the prediction error of our ML model approximating M06-2X is on the same order of magnitude as the error of the DFT method itself [114]. The overall signed error of the ML prediction will be a superposition of the regression error and the intrinsic error of the DFT method, which may involve some cancellation of errors. However, our model offers a $\sim 10^5$ factor speedup so it is well-suited to handle many reactions.

Table 6.4. Testing errors (mean ± 1 standard deviation from the five folds) for predicting $\Delta G^{\ddagger}_{\text{gas}}$, $\Delta G^{\text{rxn}}_{\text{gas}}$, $\Delta\Delta G^{\ddagger}_{\text{solv}}$, and $\Delta\Delta G^{\text{rxn}}_{\text{solv}}$. Eq. 6.10 and 6.12 are used to calculate $\Delta G^{\ddagger}_{\text{soln}}$ and $\Delta G^{\text{rxn}}_{\text{soln}}$. MAE and RMSE have units of kcal mol^{-1}.

	Random			**K-Means**		
Target	MAE	RMSE	R^2	MAE	RMSE	R^2
$\Delta G^{\ddagger}_{\text{gas}}$	2.83 ± 0.32	5.82 ± 0.68	0.92 ± 0.02	2.95 ± 0.45	5.89 ± 0.94	0.92 ± 0.02
$\Delta G^{\text{rxn}}_{\text{gas}}$	3.01 ± 0.36	6.61 ± 0.72	0.87 ± 0.03	3.13 ± 0.38	6.54 ± 0.93	0.85 ± 0.02
$\Delta\Delta G^{\ddagger}_{\text{solv}}$	0.58 ± 0.01	0.97 ± 0.04	0.69 ± 0.02	0.59 ± 0.09	0.94 ± 0.17	0.65 ± 0.08
$\Delta\Delta G^{\text{rxn}}_{\text{solv}}$	0.46 ± 0.04	0.84 ± 0.05	0.71 ± 0.04	0.43 ± 0.09	0.79 ± 0.19	0.67 ± 0.09
$\Delta G^{\ddagger}_{\text{soln}}$	2.85 ± 0.34	5.89 ± 0.68	0.92 ± 0.02	3.00 ± 0.44	5.99 ± 0.91	0.92 ± 0.03
$\Delta G^{\text{rxn}}_{\text{soln}}$	3.11 ± 0.37	6.85 ± 0.72	0.86 ± 0.03	3.21 ± 0.38	6.76 ± 0.96	0.85 ± 0.02

The parity plot in Figure 6.5a shows the model's predictive power is maintained across the entire range of data, even in regions where the data are sparser. Consistent with previous literature using this model architecture [179], Figure 6.5b shows that the residuals are centered around zero, indicating no systematic over- or under-prediction. 90% of the reaction barriers are predicted within 6 kcal mol^{-1} of the value calculated via quantum chemistry.

It is also important to examine model performance specifically on low barrier reactions since these are the most feasible reaction pathways included in kinetic models. For example, reactions matching the RMG templates of R-addition multiple bond and hydrogen abstraction have a $\Delta G^{\ddagger}_{\text{soln}} < 45$ kcal mol^{-1}, which is about half as large as the dynamic range for the entire dataset. As seen in Figure 6.5c, filtering the test set reactions to this subset gives much smaller errors. For R-Addition Multiple Bond reactions, the MAE is 2.46 ± 0.54 kcal mol^{-1} and RMSE is 4.04 ± 1.14 kcal mol^{-1} for $\Delta G^{\ddagger}_{\text{soln}}$. Similarly, the MAE is 1.29 ± 0.31 kcal mol^{-1} and RMSE is 1.69 ± 0.41 kcal mol^{-1} for Hydrogen Abstraction reactions. It is promising to see the model perform even better on this relevant subset.

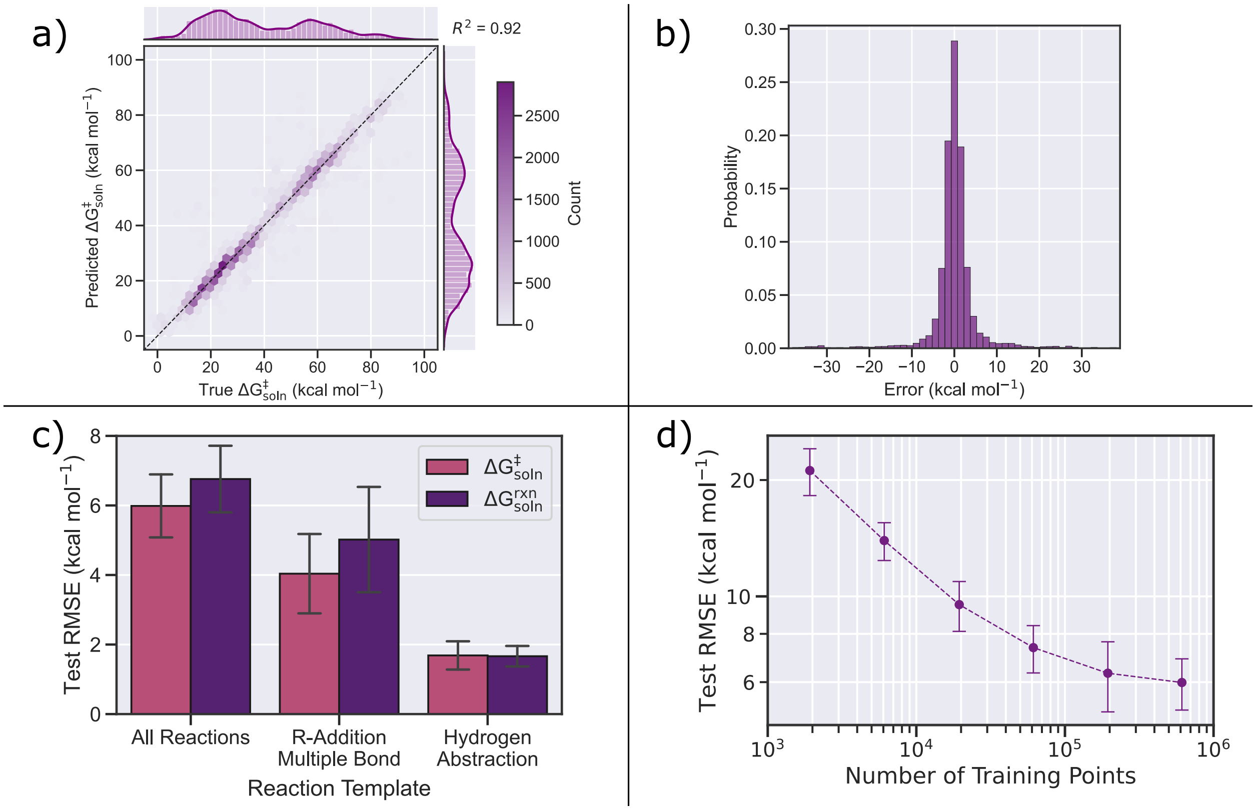


Figure 6.5. Deep learning model results for predicting $\Delta G^{\ddagger}_{\text{soln}}$ and $\Delta G^{\text{rxn}}_{\text{soln}}$ when using K-Means splits. Error bars indicate one standard deviation calculated across the five folds. (a) Parity plot of model predictions vs "true" (i.e., calculated) barriers $\Delta G^{\ddagger}_{\text{soln}}$ for the first fold. (b) Histogram of testing errors (predicted minus "true") for the first fold. (c) Testing RMSE grouped by reaction template. (d) Testing RMSE vs. the number of training data points.

To examine model sensitivity to the size of the training set, we trained models using incrementally less data. The validation and test sets are kept constant so that model performance is always evaluated on the same sets and thus training set size is the only variable changing. Figure 6.5d plots the learning curve and demonstrates that the model strongly benefits from additional training data. The test set RMSE begins to level off after about 200,000 training points. However, the curve has not fully reached an asymptote so it is expected that further improvement could be obtained if training with more data.

6.4 Conclusion

The field of quantitative predictive chemistry depends on having accurate kinetic parameters. Historically, the field has relied on regressing these values from a relatively narrow set of experimental measurements. More recently, the field of chemical kinetics has transitioned from a post-dictive to predictive modeling approach that utilizes ab initio calculations which can better extrapolate to new conditions and reactions [204]. This framework is powerful as exploring these systems computationally should be cheaper, faster, and safer in comparison to experimental approaches. Still, the standard quantum chemistry workflow is far too computationally expensive to be applied to every possible reaction that may be of interest for modeling a given system. There is a growing need to provide easy access to accurate kinetic predictions. While training ML models using high-accuracy calculations or experiments would yield the greatest value, data scarcity has made it difficult to train data-driven estimators that generalize well over a broad range of reactions.

This work presents two key contributions for the field of kinetics in solution. To address the current paucity of quantitative chemical reaction data, we first create a novel dataset with approximately 6000 elementary radical reactions. This dataset contains substantial reaction diversity, all TS geometries are validated by an IRC calculation, and all reactions are calculated in the gas phase as well as in 40 popular solvents. We anticipate this dataset will serve as a benchmark for future studies. Our second contribution is training a deep graph network to simultaneously predict the Gibbs free energy of activation and of reaction in both gas and solution phases. Importantly, our model only requires the atom-mapped SMILES of the reactant, product, and solvent as input, making it ideal for high-throughput screening of large regions of reaction space. Our model shows strong performance for both interpolation-based random splits and more challenging K-Means splits. In contrast, many other published models require more expensive representations or were only evaluated on random splits. Finally, our results corroborate existing literature that shows learned representations from graphs typically outperform traditional approaches within the field of chemical kinetics. We anticipate our model will be useful for estimating barrier heights during the automated generation of kinetic models, identifying relevant low-barrier reactions produced from an automated enumeration, and offering substantial speedup when refining existing kinetic models.

Going forward, one idea for future work is to explore passing quantum mechanical descriptors as input to the D-MPNN model. However, prior studies with select descriptors have only demonstrated improvements to model performance for small training set sizes. These descriptors often have negligible impact for any reasonably sized dataset [84, 93, 95, 193], but it is possible other descriptors exist which are more correlated with the prediction target. More broadly, the field should continue generating high-quality kinetics datasets. To our knowledge, the dataset from ref. [109] is the largest open-source kinetics dataset with nearly 12,000 reactions calculated at CCSD(T)-F12a/cc-pVDZ-F12//ωB97X-D3/def2-TZVP. However, the scope of this dataset is still relatively limited; the species only include the elements H, C, N, and O, each TS contains at most only seven heavy atoms, all TSs are closed shell singlets, and all reactions are only calculated in the gas phase. The data created in this work directly addresses several of these limitations. For example, the reactions in our dataset contain up to 19 heavy atoms, include some sulfur containing compounds, and focus on radical chemistry in both gas and solution phase. Despite this progress, most chemical space remains unexplored [173, 174, 205]. It is imperative for the field to continue developing open-source datasets and new models that can process the magnitude of reactions necessary to achieve the grand vision of quantitative predictive chemistry.

6.5 Appendix

6.5.1 Settings for the Single-Ended Molecular Growing String Method

The settings used for the single-ended molecular growing string method (MGSM) [124] are shown in Table 6.5.

Table 6.5. Settings used for the single-ended MGSM [124] to identify transition state structures.

Setting	Value
M_TYPE	SSM
MAX_OPT_ITERS	100
STEP_OPT_ITERS	30
CONV_TOL	0.0005
ADD_NODE_TOL	0.1
SSM_DQMAX	0.4
GROWTH_DIRECTION	0
INT_THRESH	2.0
MIN_SPACING	1.0
INITIAL_OPT	150
FINAL_OPT	150
PRODUCT_LIMIT	100
TS_FINAL_TYPE	1
NNODES	30

6.6 COSMO-RS Information

The molar concentrations of the pure solvents ($c_{solvent}$) used during the COSMO-RS workflow are shown in Table 6.6.

Table 6.6. Solvents used from COSMO-RS. $c_{solvent}$ is provided in units of mol/L.

Solvent	$c_{solvent}$ at 298 K
Acetic acid	17.8
Acetone	14.7
Acetonitrile	19.5
Aniline	10.4
Benzaldehyde	9.6
Benzene	11.7
ϵ-Caprolactam	9.0
Carbon disulfide	16.1
Chlorobenzene	10.2
Chloroform	13.2
Cyclohexane	10.0
Cyclopentane	11.6
n-Decane	5.3
Dichloromethane	16.7
Diethyl ether	11.2
Diethyl sulfide	9.9
Dimethyl formamide	12.7
1,4-Dioxane	12.1
DMSO	13.5
Ethanol	18.1
Ethyl acetate	11.1
Ethyl acrylate	9.8
2-Ethylhexanol	6.3
Furane	14.9
Glycerol	11.8
n-Hexane	8.4
Isopropanol	13.7
Methanol	26.0
Nitromethane	19.2
n-Octanol	6.3
Phenol	10.9
Phenyl isocyanate	9.0
Piperidine	10.5
Pyridine	12.4
Pyrrolidine	12.2
Tetrahydrofurane	13.1
Toluene	9.8
Triethylamine	7.8
γ-Valerolactone	10.5
Water	50.1

6.6.1 Supplemental Dataset Analysis

6.6.1.1 Reaction Generation

Our dataset centers around main and side reactions relevant to radical polymerization of vinyl ($H_2C{=}CH{-}R$) and methyl-substituted vinyl ($H_2C{=}C(CH_3){-}R$) monomers. The following species types are considered:

- initial radicals, which are either generated in situ from technically important radical polymerization initiators, or which occur as intermediates in autoxidation chemistry of hydrocarbons (see Table 6.7)
- a variety of vinyl and methyl-substituted vinyl monomers, which can undergo a radical polymerization (see Table 6.8)
- 2-Mercaptoethanol as a classical chain-transfer agent (CTA)
- propagating radicals derived from the respective monomers (as explained below)
- polymer repeating units derived from the respective monomers (as explained below)

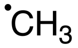
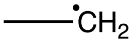
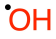
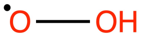
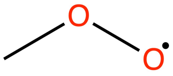
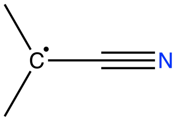
The propagating radicals are model species for the growing polymer chain end radicals. In our dataset, they are derived from the monomers by adding a CH3 radical to either the terminal or central carbon atom of the C=C double bond, as shown in Figure 6.6. CTAs are species that facilitate transfer of the radical center from the propagating radicals if possible, usually by a hydrogen abstraction reaction, and for which the formed radical center subsequently reacts with a monomer to start a new propagating chain. Transfer of the radical center from propagating radicals to the polymer backbone is technically also a chain-transfer reaction. Here, we explicitly consider transfer to a polymer repeating unit, which we represent in our dataset by adding a CH_3 to both sites of the C=C double bond, as shown in Figure 6.7 for styrene and methacrylate. We furthermore add a few radicals to our dataset, for which only unimolecular reactions are explored. These species are listed in Table 6.9.

We next define combinations of reactant molecules that are subjected to exhaustive reaction path exploration. Here, we focus on bimolecular reactions of a radical with a closed-shell molecule and left out radical-radical recombination reactions. The following reaction types are studied:

- initial radical + monomer
- propagating radical + monomer
- initial radical + CTA
- propagating radicals + CTA
- initial radical + polymer repeating unit
- propagating radicals + polymer repeating unit

In these reaction types, we considered the full set of radicals/molecules for each species type. Additionally, to cover some aspects of potential unimolecular side reactions–especially under the presence of molecular oxygen, which can lead to the formation of peroxy and hydroperoxy alkyl radicals–a few selected unimolecular reactions are studied for the species listed in Table 6.9. Note that backbiting reactions, in which a radical chain end reacts with the polymer backbone of one of the neighboring repeating units (usually by abstracting a hydrogen atom) and therefore transfers the radical to another position in the polymer) are not specifically covered in this reaction data set.

Table 6.7. Initial radicals chosen for the reaction data generation.

SMILES String	Molecular Structure
[CH3]	
C[CH2]	
[OH]	
O[O]	
CO[O]	
C[C](C)C#N	

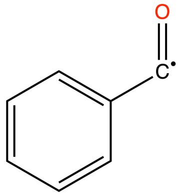
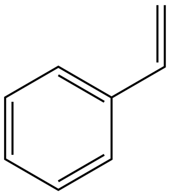
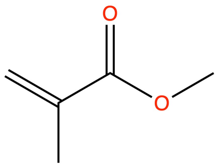
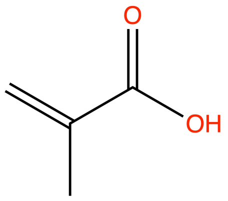
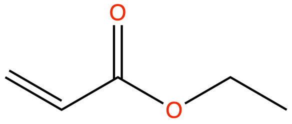
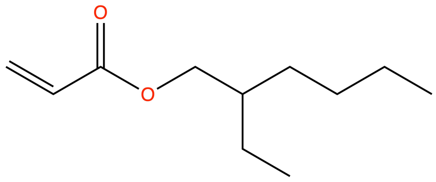
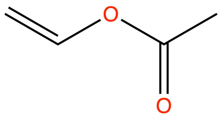
O=[C]c1ccccc1	

Table 6.8. Monomers used in the reaction data generation.

SMILES String	Molecular Structure
C=Cc1ccccc1	
COC(=O)C(=C)C	
CC(=C)C(=O)O	
CCOC(=O)C=C	
CCCCC(CC)COC(=O)C=C	
CC(=O)OC=C	

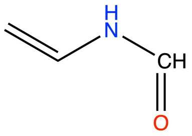
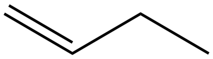
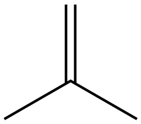
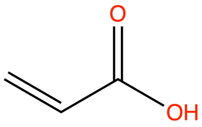
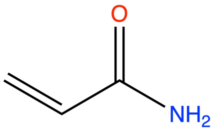
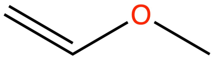
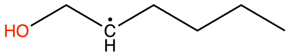
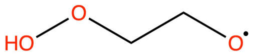
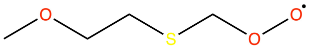
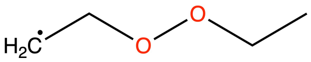
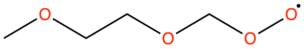
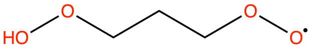
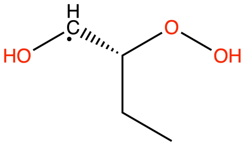
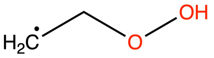
C=CNC=O	
CCC=C	
CC(=C)C	
OC(=O)C=C	
NC(=O)C=C	
COC=C	

Table 6.9. Radicals chosen for the unimolecular reaction data generation.

SMILES String	Molecular Structure
OC[CH]CCCC	
C(COO)[O]	
COCCSCO[O]	

[CH2]COOCC	
COCCOCO[O]	
C(COO)CO[O]	
O[CH][C@@H](CC)OO	
[CH2]COO	

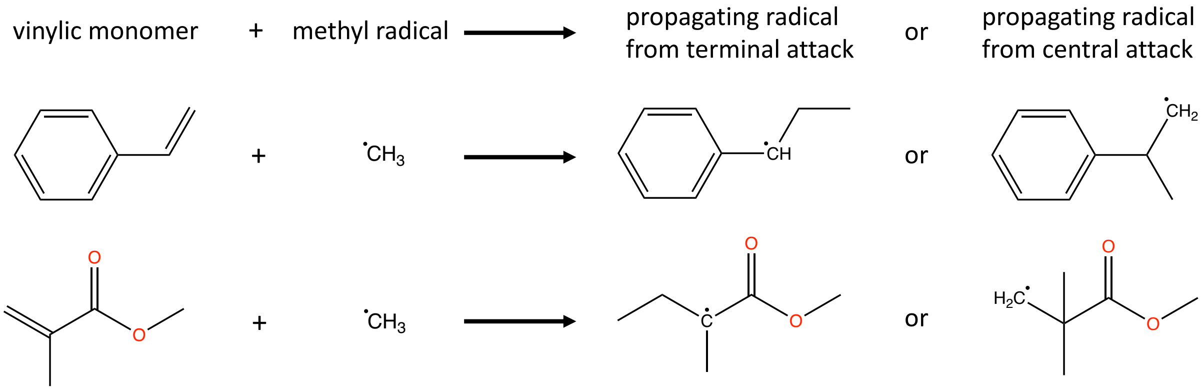

vinylic monomer
+
methyl radical
propagating radical from terminal attack
or
propagating radical from central attack
+
•CH3
•CH
or
•CH2
+
•CH3
O
O
C
O
O
or
H2C•
O
O

Figure 6.6. Examples to demonstrate how the two propagating radicals are formed from different vinyl monomers.

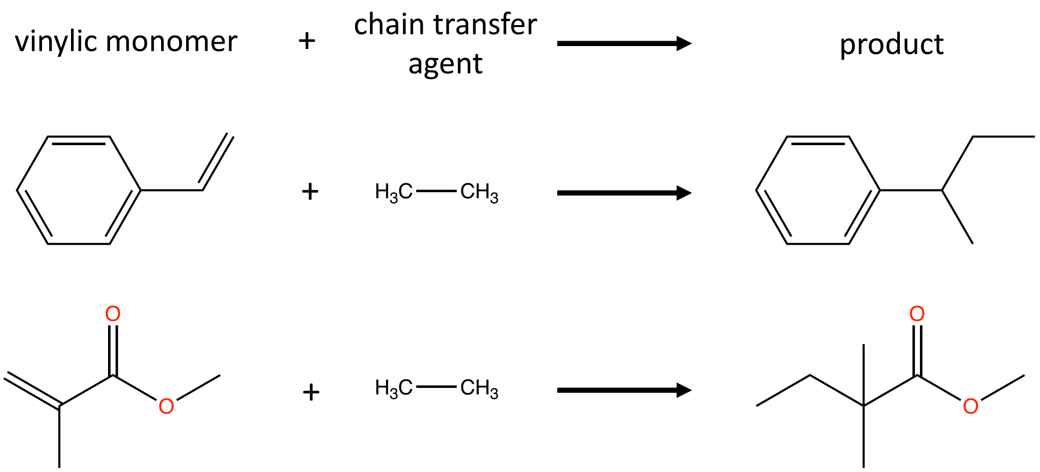


Figure 6.7. Examples to demonstrate the logic of the chain transfer agent generation from vinyl monomers.

After the exhaustive search of possible reactions–based on combinatoric bond changes as explained in the main manuscript–starting from the different reactant combinations described above, we refine the transition-state structures and perform an intrinsic reaction coordinate (IRC) calculation. This procedure is also explained in the main manuscript. In a few cases, the IRC calculations led to reactants or products that were different from the original reactants and products according to the MGSM calculations. In this case, the original structures are replaced with the structures from the IRC calculations, and the latter structures are used in the dataset. This procedure explains the presence of reactions in the data set that cannot be derived from the systematic data generation procedure described above.

6.6.1.2 Dataset Statistics

Here, we give an overview of the data generated in this work. All reactions are balanced, contain neutrally charged species, and have canonical atom-mapped SMILES. Figure 6.8 shows the broad distribution of molecular weights, with the most common value being around 170 Da. For organic molecules comprised primarily of carbon atoms, this correspond to approximately 12 heavy atoms, as shown in Figure 6.3b from the main text.

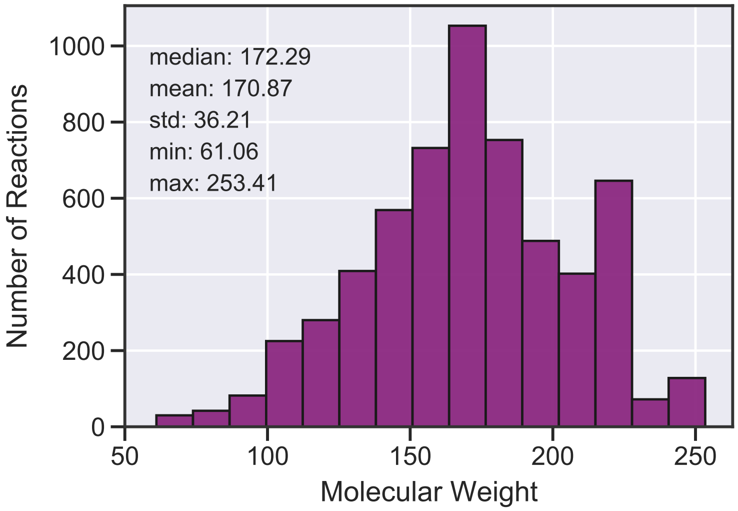


Figure 6.8. Distribution of molecular weights for the reactions in this dataset.

With the exception of 68 unimolecular reactions, all reactions start from two reactants. As seen in Figure 6.9, most reactions produce two products, making them bimolecular from both directions. Four reactions produced four products in the forward direction. Although these reactions would be quite improbable in the reverse direction, the transition state (TS) geometry of these four reactions–and of all reactions included in our dataset–was successfully verified by both a forward and reverse intrinsic reaction coordinate (IRC) calculation at B3LYP-D3(BJ)/def2-TZVP, so we included them in our final set. These four reactions have similar reaction templates as they all produce carbon dioxide and methane as two of the four products. The distribution of the Gibbs free energy of activation in water and the Gibbs free energy of reaction in water vs. the number of products is shown in Figure 6.10. The color in Figure 6.9 and Figure 6.10 generally denotes increasing Gibbs free energy of activation since forming more products requires breaking more bonds which in turn requires more energy to be input to the system.

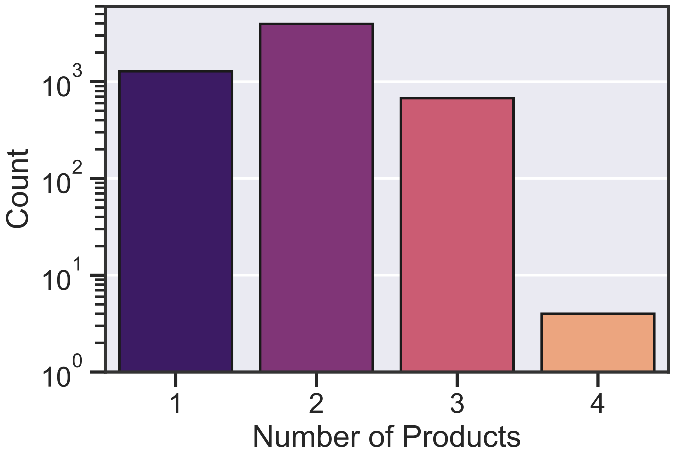


Figure 6.9. Distribution of the number of products formed by the reactions.

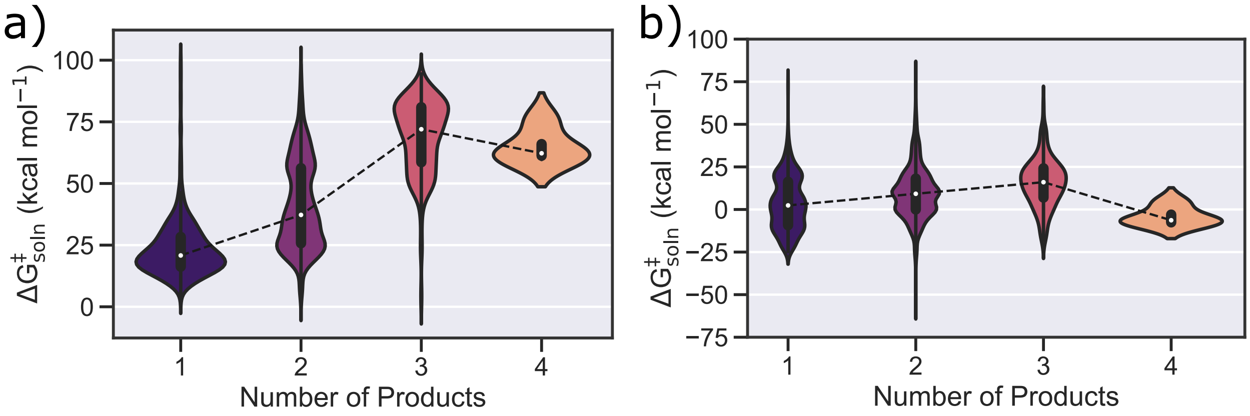


Figure 6.10. Distribution of the a) Gibbs free energy of activation in water and b) Gibbs free energy of reaction in water grouped by the number of products formed by the reaction. Although some categories appear to have negative barrier heights, this is only an artifact of the fitted kernel density distribution; all $\Delta G^{\ddagger}_{\mathrm{soln}}$ values are positive.

As another method to characterize the data, we look at the number of bond changes that occurred during the reaction. Here, we account for changes in connectivity, not changes in bond order to existing connections. Each reaction can be classified by a general reaction template of "breaks X bonds and forms Y bonds", which is abbreviated as "bXfY". As shown in Figure 6.11, most reactions have two bonds changing (i.e. breaking or forming a total of two bonds). This plot is identical to Figure 6.3c from the main text and is reproduced here for convenience. The distribution of which types of bonds are changing is shown in Figure 6.3d from the main text. The distribution of the Gibbs free energy of activation in water and the Gibbs free energy of reaction in water grouped by the number of broken and formed bonds is shown in Figure 6.12. The data generally follows the expected trend in that the Gibbs free energy of activation increases as more bonds are broken during the reaction. As before, the color generally becomes warmer to denote this larger energy requirement.

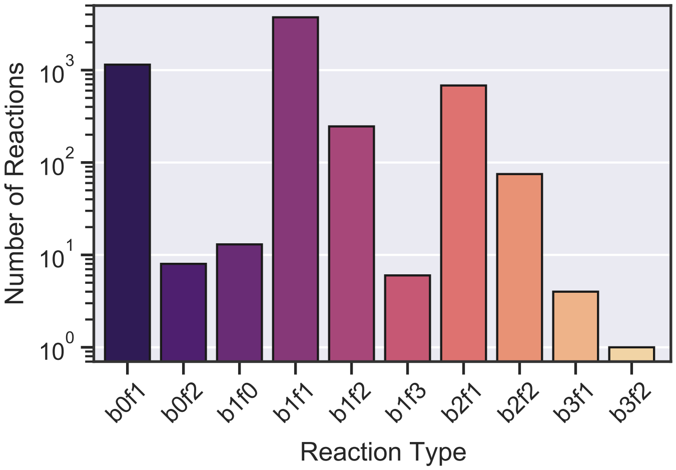


Figure 6.11. Distribution of the number of bonds broken and formed during the reaction.

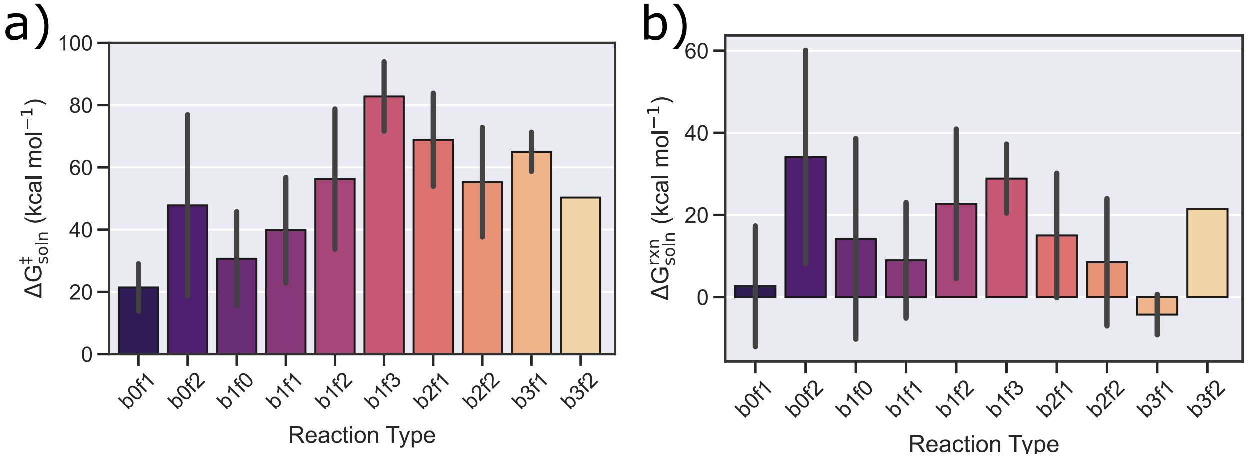
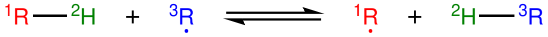
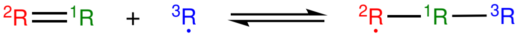


Figure 6.12. Distribution of the a) Gibbs free energy of activation in water and b) Gibbs free energy of reaction in water grouped by the number of broken and formed bonds during the reaction.

Consistent with other published datasets [179, 203] that were generated using a combinatoric driving coordinate generator in combination with the single ended molecular growing string method, most reactions do not match templates from the manually curated list of families stored in the Reaction Mechanism Generator (RMG) database [112]. This result is expected since the driving coordinates specified by combinatorics of potential bond formation and bond breakages is known to create large reaction diversity. Of the reactions that did match RMG templates, the vast majority belong to the two families shown in Table 6.10. This result is also inline with expectations since Figure 6.11 indicates that the most common reaction type involves breaking and forming one bond. As seen in Figure 6.13, the Pearson correlation coefficient is approximately 0.90 for both of these families, indicating strong agreement with the expected Evans-Polanyi relationship [202] in which the activation energy is positively correlated with the reaction energy.

Table 6.10. RMG reaction templates present in the dataset.

RMG Reaction Family	Template
H_Abstraction	
R_Addition_MultipleBond	

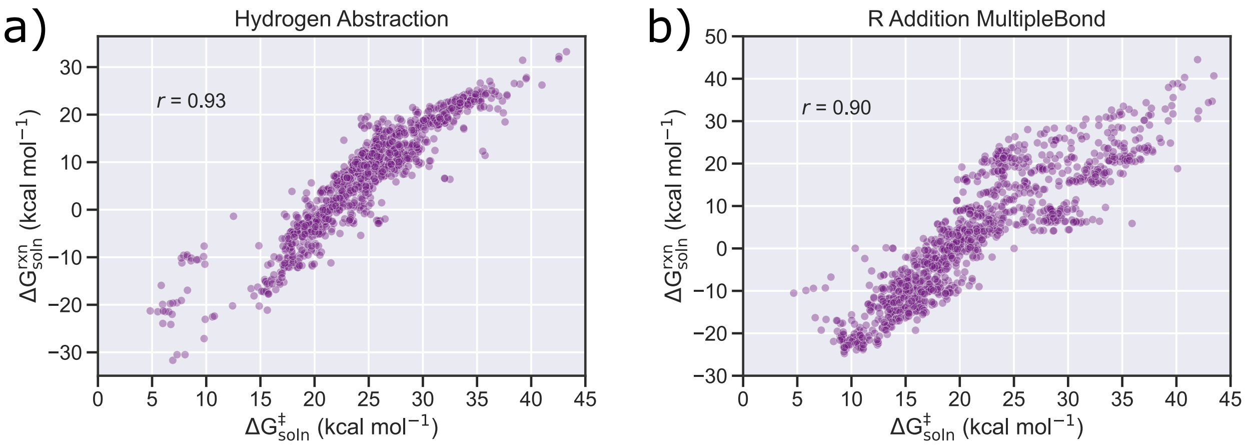


Figure 6.13. Scatter plot of the Gibbs free energy of reaction vs. the Gibbs free energy of activation in water for a) hydrogen abstraction reactions and b) R addition multiple bond reactions.

We also look at the number of rotatable bonds, which indicates how flexible a molecule is and correlates with the number of possible conformers. Here, the number of rotatable bonds is calculated using the `CalcNumRotatableBonds` function from RDKit [71] with `strict=True` when considering explicit hydrogens. The distribution for the reactants and products is shown in Figure 6.14.

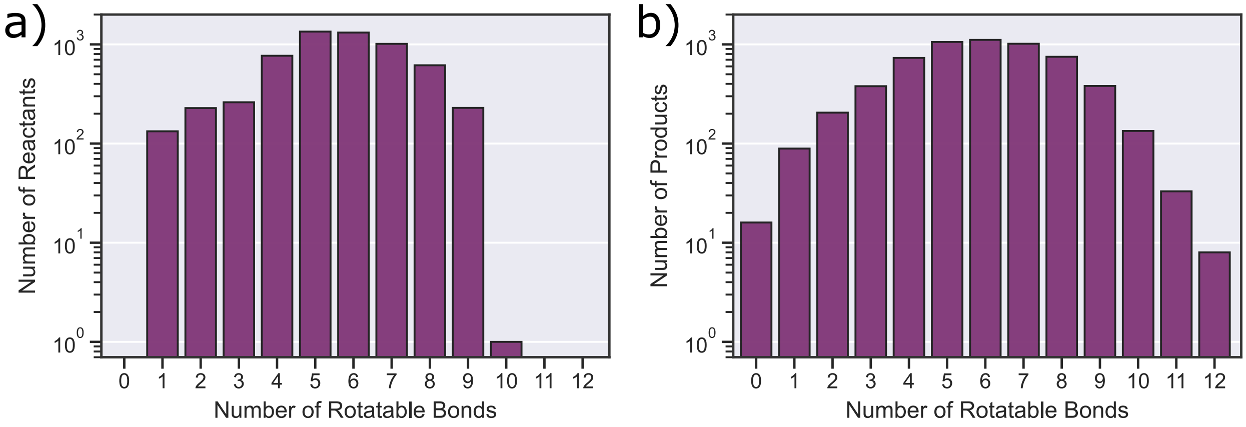


Figure 6.14. Distribution of the number of rotatable bonds in the a) reactants and b) products.

We also examine the number of hydrogen bond donors and acceptors, which can be important properties for drug design [206, 207]. These values were calculated using the `CalcNumHBD` and `CalcNumHBA` functions from RDKit. Hydrogen bonding typically arises due to the presence of electronegative atoms, and can act as a general indicator of the molecule's acidity. The distribution of donors from the reactants and products is shown in Figure 6.15, while the distribution of acceptors is shown in Figure 6.16.

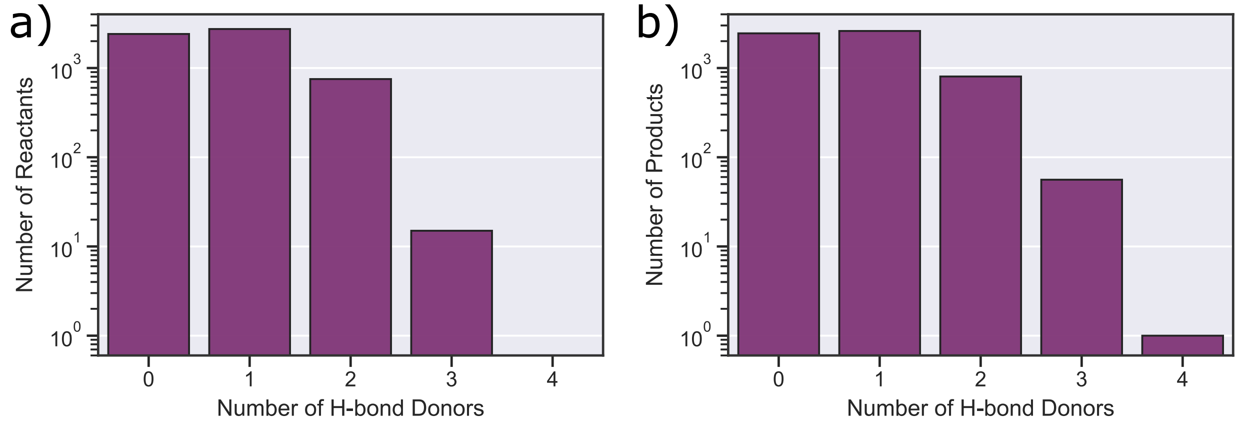


Figure 6.15. Distribution of the number of hydrogen bond donors in the a) reactants and b) products.

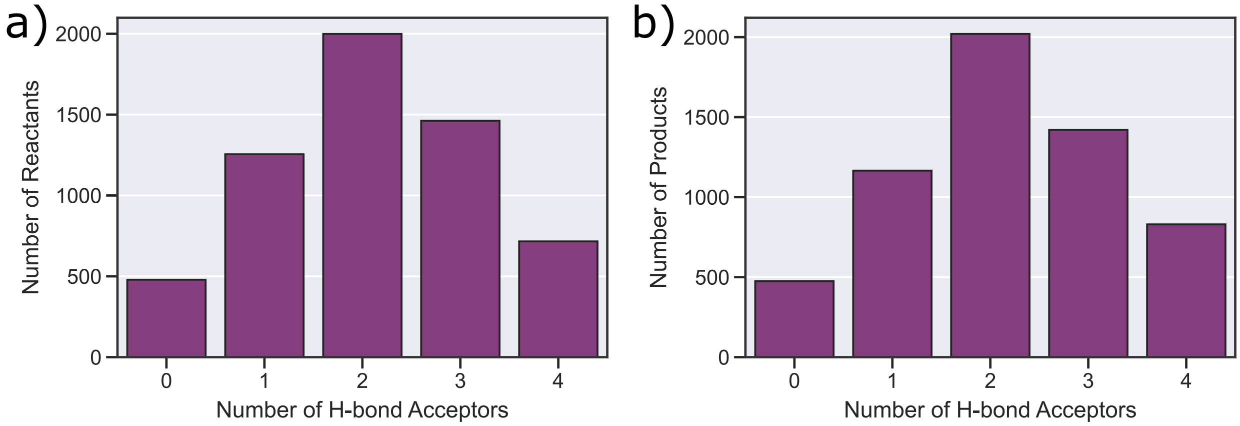


Figure 6.16. Distribution of the number of hydrogen bond acceptors in the a) reactants and b) products.

6.6.1.3 Analysis of the Spin Squared Operator

To inspect whether the quantum chemistry calculations are spin contaminated, we examine the expectation value of the spin squared operator $<S^2>$. Figure 6.17 plots the distribution of $<S^2>$ for the radical reactants, radical products, and TSs, which were all run with unrestricted methods in the doublet state. The values from the singlet reactant and singlet product species are not plotted since they were run with a restricted method so by definition are forced to have $<S^2> = 0$.

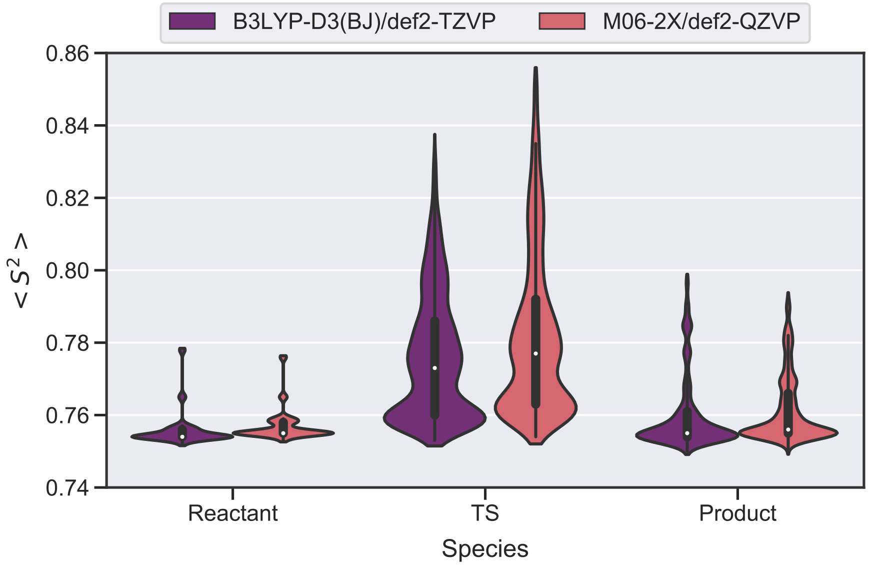


Figure 6.17. Distribution of $<S^2>$ for reactants, transition states, and products in the doublet state across B3LYP-D3(BJ)/def2-TZVP used for geometry optimization and M06-2X/def2-QZVP used for single point energy refinement.

Some of the TS calculations in the initial set had large $<S^2>$ values (e.g. larger than 1.00). When considering what threshold to set for an acceptable deviation, we consulted with Gaussian technical support, who shared that a deviation of ± 0.10 from the expected $<S^2>$ value is not necessarily substantial, especially for TS structures in which there are partial bonds being formed and/or broken. All reactant and product calculations fell within this tolerance, but we removed about 5% of the initial data to exclude reactions whose TS had a $<S^2>$ that deviated substantially from the expected value of 0.75 in this case.

We applied a spin projection method to estimate how much the spin contamination impacts the unrestricted electronic energy. One method to correct for the contamination is to project out the energetic contribution from the higher energy state (quartet in this case) causing the contamination [208, 209]. This allows us to obtain a spin-projected energy of the doublet state without the spin contamination error as follows:

$$E_{AP}^{LS} = \alpha E_{BS}^{LS} - \beta E^{HS} \tag{6.17}$$

where

$$\alpha = \frac{\left\langle \hat{S}^2 \right\rangle^{HS} - \left\langle \hat{S}^2 \right\rangle_{\text{exact}}^{LS}}{\left\langle \hat{S}^2 \right\rangle^{HS} - \left\langle \hat{S}^2 \right\rangle_{BS}^{LS}} \tag{6.18}$$

and

$$\beta = \alpha - 1 \tag{6.19}$$

such that BS stands for broken symmetry, AP stands for approximate spin projection, LS stands for low spin state, and HS stands for high spin state.

We applied this method to five transition states with $\left\langle \hat{S}^2 \right\rangle_{BS}^{LS}$ ranging from 0.80 to 0.85. In general, the spin-projected energies are lower than the original M06-2X energies by 0.9-1.4 kcal mol^{-1}. The energy difference of around 1 kcal mol^{-1} is not completely negligible, but it is also not the largest error in the entire workflow. For example, ensuring that one has the lowest energy conformer and/or the full set of conformers for multi-structure transition state theory will likely have a bigger impact on the energy value; future work with additional computational resources could explore this further. Additionally, there should be some cancellation of errors if the reactant energies are recalculated with spin projection as well.

6.6.2 Creating Clusters for Extrapolative Data Splits

Here, our goal is to create more challenging data splits to measure the likely performance in extrapolative tasks, such as making predictions on new types of molecules. Although splitting based on the Bemis-Murcko scaffold [180] is commonly used for this task, this approach relies on many ring structures. The molecules contained in our dataset happen to not be good candidates for this since about 82% of the reactants do not match any Bemis-Murcko scaffolds, which precludes any attempt to cluster molecules or reactions based on this label. Instead, we use a different procedure with three step to accomplish our goal of creating more challenging splits. The first step is to create a vector representation of the reactant complex from each reaction. We use Morgan (ECFP4) fingerprints [69, 70] with standard settings of 2048 bit hashing and radius of 2 from RDKit [71]. Several fingerprint representations exist in the literature [210], and future work could explore how the choice of initial representation impacts the clustering.

Since these Morgan fingerprint vectors are quite sparse (e.g. about 90% of the entries are zero across all vectors), we use Principal Component Analysis (PCA) to project these vectors to a more dense subspace. To determine how many principal components to use, we do a hyperparameter sweep and plot the cumulative explained variance vs. the number of principal components. We choose to use four components since this explains 80% of the variance as shown in Figure 6.18. Future work could explore other projection methods, such as t-Distributed Stochastic Neighbor Embedding (t-SNE), Uniform Manifold Approximation and Projection (UMAP) [211, 212], or others.

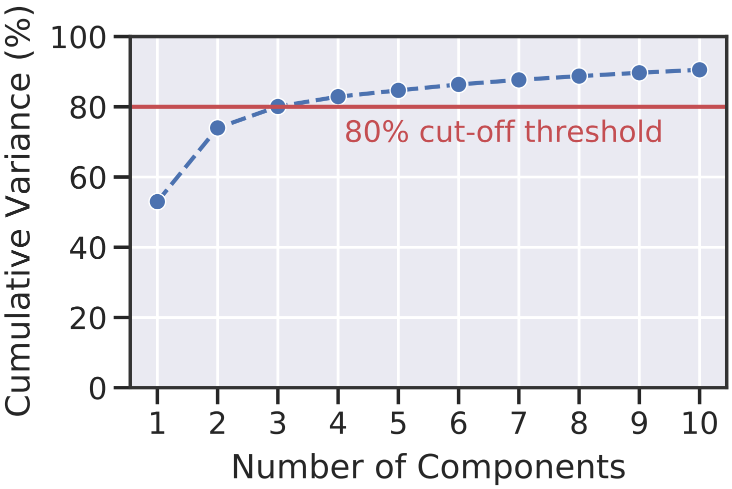


Figure 6.18. Plot of cumulative explained variance vs. number of principal components.

Clustering relies on calculating a distance matrix (or similarity matrix), of each reactant complex with respect to all other reactant complexes, which inherently scales as $\mathcal{O}(N^2)$, such that N is the number of complexes. Projecting the vectors to a smaller dimension does not change the overall scaling, but it effectively multiplies it by a smaller prefactor since fewer floating point operations need to be performed to calculate the distance in the reduced dimensional space e.g. calculating the distance between vectors of only four dimensions is faster than between vectors with 2048 dimensions.

The final step is to create clusters of similar reactant complexes. We use K-means clustering, though future work could explore other clustering algorithms as well. The algorithm randomly initializes k centroids and iteratively updates the centroids to minimize inertia–defined as the within cluster sum-of-squares–which is given by

$$\text{Inertia} = \sum_{j=0}^{k} \sum_{i=0}^{n} \left\| x_i^{(j)} - \mu_j \right\|^2 \tag{6.20}$$

such that k is the number of clusters, n is the number of samples within a cluster, $x_i^{(j)}$ is the i^{th} sample within cluster j, and μ_j is the centroid for cluster j. The entire process is repeated several times. To determine the number of clusters to use, we create an elbow plot shown in Figure 6.19 to approximate the point of diminishing marginal returns (i.e. when adding additional clusters offers negligible improvement towards lowering the inertia). Here, we choose twelve clusters.

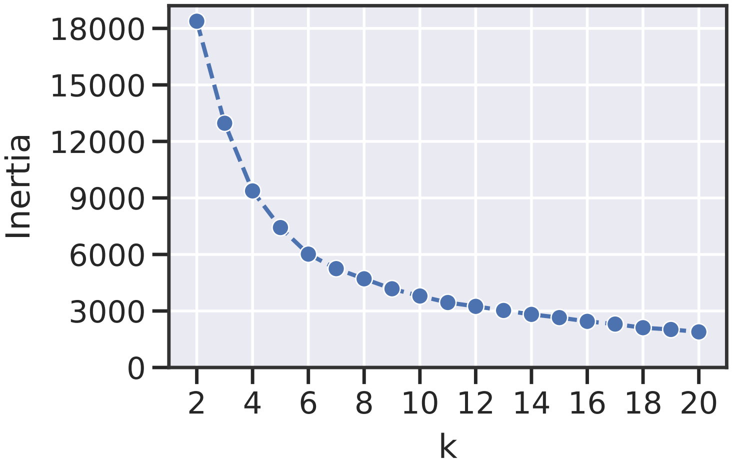


Figure 6.19. Elbow plot showing inertia vs. number of clusters (k).

Intuitively, K-means creates more challenging splits than randomly splitting since validation and testing molecules are deliberately chosen to be different relative to the training set.

This intuition is confirmed by using Tanimoto (Jaccard) [213, 214] similarity to quantitatively compare the clusters. We observe the expected trend that the average similarity of reactant complexes within a cluster is higher than the average similarity to reactant complexes for all other clusters. Finally, we inspect the clusters to confirm that they are chemically sensible. For example, reactant complexes that involve acrylic acid, radical derivatives of acrylic acid, methacrylic acid, and/or ethylacrylic acid are grouped together. Similarly, reactions involving isobutene or radical derivatives are grouped together while reactions involving smaller radicals like AIBN, HO_2, MeOO, and OH form a separate cluster. Reactions involving sytrene or radical derivatives of styrene are grouped together. Thus, this automated clustering workflow preserves chemical intuition.

6.6.3 Baseline Models

As described in the main text, we compare the performance of several baseline models from Scikit-Learn [163] and featurizations when predicting $\Delta G^{\ddagger}_{\text{soln}}$ in water. We concatenate the vector fingerprint of the reactant complex with the difference of the fingerprint between the reactant and product complex. This is the same approach as the initial featurization in the condensed graph of reaction representation [159]. We explored using Morgan (ECFP4) fingerprints [69, 70], Molecular ACCess System (MACCS) [166], Avalon [167], and Atom-pair [168]. All fingerprints are generated with RDKit [71] using the following functions respectively:

- `GetHashedMorganFingerprint` with default arguments of `radius`=2 and `nBits`=2048
- `MACCSkeys`
- `GetAvalonFP` with the default argument of `nBits`=512
- `GetHashedAtomPairFingerprint` with default arguments of `nBits`=2048, `minLength`=1, and `maxLength`=30.

Since this baseline model will only predict $\Delta G^{\ddagger}_{\text{soln}}$ in water, a representation of the solvent is not needed. The best hyperparameters for the baseline models are chosen via Optuna [165], which is used to minimize the average root mean squared error (RMSE) on the validation sets. The values are listed in Tables 6.11, 6.12, 6.13, 6.14, and 6.15. Any other hyperparameters assumed the default value.

Table 6.11. Optimal hyperparameter values for LinearSVR i.e. support vector regression with a linear kernel.

		Split Type	
Featurization	**Hyperparameter**	Random	K-means
AtomPair	`C`	0.0745	0.0174
Avalon	`C`	0.7757	0.0068
MACCS	`C`	0.9307	0.0086
Morgan	`C`	0.3618	0.1449

Table 6.12. Optimal hyperparameter values for SVR with an RBF kernel.

		Split Type	
Featurization	**Hyperparameter**	Random	K-means
AtomPair	`C`	5.6467	1.9803
Avalon	`C`	13.841	1.6795
MACCS	`C`	4.7346	3.5793
Morgan	`C`	8.2168	3.8101

Table 6.13. Optimal hyperparameter values for random forest (RF).

		Split Type	
Featurization	**Hyperparameter**	Random	K-means
AtomPair	`max_depth`	15	15
	`n_estimators`	125	250
Avalon	`max_depth`	15	15
	`n_estimators`	300	75
MACCS	`max_depth`	15	15
	`n_estimators`	125	225
Morgan	`max_depth`	15	15
	`n_estimators`	125	150

Table 6.14. Optimal hyperparameter values for extreme gradient boosting (XGB).

		Split Type	
Featurization	**Hyperparameter**	Random	K-means
AtomPair	`gamma`	0	0
	`learning_rate`	0.0901	0.0973
	`max_depth`	11	12
	`min_child_weight`	3	1
	`n_estimators`	140	90
	`reg_alpha`	2	5
	`reg_lambda`	0	2
Avalon	`gamma`	3	0
	`learning_rate`	0.0638	0.0847
	`max_depth`	15	9
	`min_child_weight`	5	4
	`n_estimators`	120	120
	`reg_alpha`	4	0
	`reg_lambda`	5	1
MACCS	`gamma`	0	2
	`learning_rate`	0.0911	0.0351
	`max_depth`	15	12
	`min_child_weight`	5	2
	`n_estimators`	200	200
	`reg_alpha`	4	4
	`reg_lambda`	0	3
Morgan	`gamma`	4	0
	`learning_rate`	0.0851	0.0948
	`max_depth`	14	10
	`min_child_weight`	4	3
	`n_estimators`	120	160
	`reg_alpha`	0	3
	`reg_lambda`	4	5

Table 6.15. Optimal hyperparameter values for multi-layer perceptron (MLP).

		Split Type	
Featurization	**Hyperparameter**	Random	K-means
AtomPair	`num_layers`	0.0005948	0.000672
	`layer_size`	500	500
	`alpha`	3	3
Avalon	`num_layers`	0.00049499	0.002489
	`layer_size`	400	450
	`alpha`	4	2
MACCS	`num_layers`	0.00157	0.001128
	`layer_size`	400	400
	`alpha`	3	3
Morgan	`num_layers`	0.0008288	0.00209
	`layer_size`	200	350
	`alpha`	3	3

6.6.4 Graph Network Description

A detailed description of the D-MPNN architecture is provided in Section 2.2.5. We use the Noam learning rate scheduler from Vaswani et al. [215] during training, which starts by linearly increasing the learning rate from the initial to the maximum value over a specified number of warm-up epochs. Then the learning rate is exponentially decreased to the final value over the remaining epochs.

We use the provided hyperparameter search code from Chemprop [54] to determine the hyperparameters controlling the model architecture and fitting procedure. Only a subset of the data is used during the hyperparameter search. The optimal hyperparameters used by D-MPNN-1 and D-MPNN-2 are summarized in Table 6.16 and Table 6.17 respectively. As described in the main text, the D-MPNN-1 notation indicates that only one graph network is used while D-MPNN-2 indicates that two graph networks are used to account of the different solvents. Hyperparameters not shown in the tables used the default values from Chemprop.

Table 6.16. Hyperparameters the define the architecture and training for D-MPNN-1. The same values are used when training on both random reaction and K-Means splits.

Hyperparameter	**Value**
MPNN hidden size	300
MPNN depth	3
Activation function	LeakyReLU
Aggregation	sum
Aggregation norm	40
Epochs	65
Batch size	16
Initial learning rate	1×10^{-4}
Maximum learning rate	1×10^{-3}
Final learning rate	1×10^{-5}
Warm-up epochs	6

Table 6.17. Hyperparameters the define the architecture and training for D-MPNN-2. The same values are used when training on both random reaction and K-Means splits.

Hyperparameter	**Value**
MPNN hidden size	300
MPNN depth	3
Activation function	LeakyReLU
Aggregation	sum
Aggregation norm	40
Epochs	65
Batch size	16
Initial learning rate	1×10^{-4}
Maximum learning rate	2×10^{-3}
Final learning rate	1×10^{-5}
Warm-up epochs	6

6.6.5 Comparison of Compute Resources

The following tests were run to estimate the computational speed-up offered by our ML model. We used the model to make predictions for 600 reactions using one CPU. This took 18 seconds, i.e., ~0.03 seconds per reaction. In contrast, a geometry optimization and frequency calculation using density functional theory commonly takes about 1-2 hours per molecule when parallelized across 8-12 cores. Dividing the two values shows that our ML model can estimate kinetic parameters approximately 1×10^5 times faster than using quantum chemistry calculations. The times reported here may vary depending on the available compute hardware.

6.7 References

(1) Vermeire, F. H.; Aravindakshan, S. U.; Jocher, A.; Liu, M.; Chu, T.-C.; Hawtof, R. E.; Van de Vijver, R.; Prendergast, M. B.; Van Geem, K. M.; Green, W. H. Detailed Kinetic Modeling for the Pyrolysis of a Jet A Surrogate. *Energy Fuels* **2022**, *36*, 1304–1315.

(2) Payne, A. M.*; Spiekermann, K. A.*; Green, W. H. Detailed Reaction Mechanism for 350–400°C Pyrolysis of an Alkane, Aromatic, and Long-Chain Alkylaromatic Mixture. *Energy & Fuels* **2022**, *36*, 1635–1646.

(3) Deglmann, P.; Müller, I.; Becker, F.; Schäfer, A.; Hungenberg, K.-D.; Weiß, H. Prediction of Propagation Rate Coefficients in Free Radical Solution Polymerization Based on Accurate Quantum Chemical Methods: Vinylic and Related Monomers, Including Acrylates and Acrylic Acid. *Macromol. React. Eng.* **2009**, *3*, 496–515.

(4) Deglmann, P.; Schäfer, A.; Lennartz, C. Application of Quantum Calculations in the Chemical Industry-—An Overview. *Int. J. Quantum Chem.* **2015**, *115*, 107–136.

(5) Edeleva, M.; Van Steenberge, P. H. M.; Sabbe, M. K.; D'hooge, D. R. Connecting Gas-Phase Computational Chemistry to Condensed Phase Kinetic Modeling: The State-of-the-Art. *Polymers* **2021**, *13*, 3027.

(6) Zhang, F.; Zeng, M.; Yappert, R. D.; Sun, J.; Lee, Y.-H.; LaPointe, A. M.; Peters, B.; Abu-Omar, M. M.; Scott, S. L. Polyethylene upcycling to long-chain alkylaromatics by tandem hydrogenolysis/aromatization. *Science* **2020**, *370*, 437–441.

(7) Grinberg Dana, A.; Wu, H.; Ranasinghe, D. S.; Pickard IV, F. C.; Wood, G. P. F.; Zelesky, T.; Sluggett, G. W.; Mustakis, J.; Green, W. H. Kinetic Modeling of API Oxidation:(1) The AIBN/H_2O/CH_3OH Radical "Soup". *Mol. Pharmaceutics* **2021**, *18*, 3037–3049.

(8) Saidi, C. N.; Mielczarek, D. C.; Paricaud, P. Predictions of Solvation Gibbs Free Energies with COSMO-SAC Approaches. *Fluid Phase Equilib.* **2020**, *517*, 112614.

(9) Wang, J.; Song, Z.; Lakerveld, R.; Zhou, T. Solvent Selection for Chemical Reactions Toward Optimal Thermodynamic and Kinetic Performances: Group Contribution and COSMO-based Modeling. *Fluid Phase Equilib.* **2023**, *564*, 113623.

(10) Fang, J.; Zhang, M.; Zhu, P.; Ouyang, J.; Gong, J.; Chen, W.; Xu, F. Solubility and Solution Thermodynamics of Sorbic Acid in Eight Pure Organic Solvents. *J. Chem. Thermodyn.* **2015**, *85*, 202–209.

(11) Walker, T. W.; Chew, A. K.; Li, H.; Demir, B.; Zhang, Z. C.; Huber, G. W.; Van Lehn, R. C.; Dumesic, J. A. Universal Kinetic Solvent Effects in Acid-Catalyzed Reactions of Biomass-Derived Oxygenates. *Energy Environ. Sci.* **2018**, *11*, 617–628.

(12) Gilson, M. K.; Zhou, H.-X. Calculation of Protein-Ligand Binding Affinities. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21–42.

(13) Klamt, A.; Eckert, F.; Diedenhofen, M.; Beck, M. E. First Principles Calculations of Aqueous pK_a Values for Organic and Inorganic Acids Using COSMO-RS Reveal an Inconsistency in the Slope of the pK_a Scale. *J. Phys. Chem. A* **2003**, *107*, 9380–9386.

(14) Mannhold, R.; Poda, G. I.; Ostermann, C.; Tetko, I. V. Calculation of Molecular Lipophilicity: State-of-the-Art and Comparison of logP Methods on more than 96,000 Compounds. *J. Pharm. Sci.* **2009**, *98*, 861–893.

(15) Loschen, C.; Reinisch, J.; Klamt, A. COSMO-RS Based Predictions for the SAMPL6 logP Challenge. *J. Comput. Aided Mol. Des.* **2020**, *34*, 385–392.

(16) Petris, P. C.; Becherer, P.; Fraaije, J. G. E. M. Alkane/Water Partition Coefficient Calculation Based on the Modified AM1 Method and Internal Hydrogen Bonding Sampling Using COSMO-RS. *J. Chem. Inf. Model.* **2021**, *61*, 3453–3462.

(17) Klamt, A.; Eckert, F.; Reinisch, J.; Wichmann, K. Prediction of Cyclohexane-Water Distribution Coefficients with COSMO-RS on the SAMPL5 Data Set. *J. Comput. Aided Mol. Des.* **2016**, *30*, 959–967.

(18) Nicholls, A.; Mobley, D. L.; Guthrie, J. P.; Chodera, J. D.; Bayly, C. I.; Cooper, M. D.; Pande, V. S. Predicting Small-Molecule Solvation Free Energies: An Informal Blind Test for Computational Chemistry. *J. Med. Chem.* **2008**, *51*, 769–779.

(19) Duarte Ramos Matos, G.; Kyu, D. Y.; Loeffler, H. H.; Chodera, J. D.; Shirts, M. R.; Mobley, D. L. Approaches for Calculating Solvation Free Energies and Enthalpies Demonstrated with an Update of the FreeSolv Database. *J. Chem. Eng. Data* **2017**, *62*, 1559–1569.

(20) Jalan, A.; Ashcraft, R. W.; West, R. H.; Green, W. H. Predicting Solvation Energies for Kinetic Modeling. *Annu. Rep. Prog. Chem., Sect. C: Phys. Chem.* **2010**, *106*, 211–258.

(21) Skyner, R. E.; McDonagh, J. L.; Groom, C. R.; van Mourik, T.; Mitchell, J. B. O. A Review of Methods for the Calculation of Solution Free Energies and the Modelling of Systems in Solution. *Phys. Chem. Chem. Phys.* **2015**, *17*, 6174–6191.

(22) Slakman, B. L.; West, R. H. Kinetic Solvent Effects in Organic Reactions. *J. Phys. Org. Chem.* **2019**, *32*, e3904.

(23) Boereboom, J. M.; Fleurat-Lessard, P.; Bulo, R. E. Explicit Solvation Matters: Performance of QM/MM Solvation Models in Nucleophilic Addition. *J. Chem. Theory Comput.* **2018**, *14*, 1841–1852.

(24) Tomasi, J.; Mennucci, B.; Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chem. Rev.* **2005**, *105*, 2999–3094.

(25) Miertuš, S.; Scrocco, E.; Tomasi, J. Electrostatic Interaction of a Solute with a Continuum. A Direct Utilizaion of AB Initio Molecular Potentials for the Prevision of Solvent Effects. *Chem. Phys.* **1981**, *55*, 117–129.

(26) 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. *J. Phys. Chem. B* **2009**, *113*, 6378–6396.

(27) Cramer, C. J.; Truhlar, D. G. A Universal Approach to Solvation Modeling. *Acc. Chem. Res.* **2008**, *41*, 760–768.

(28) Klamt, A.; Schüürmann, G. COSMO: A New Approach to Dielectric Screening in Solvents with Explicit Expressions for the Screening Energy and its Gradient. *Chem. J. Soc. Perkin Trans. 2* **1993**, 799–805.

(29) Grochowski, P.; Trylska, J. Continuum Molecular Electrostatics, Salt Effects, and Counterion Binding—-A Review of the Poisson-Boltzmann Theory and its Modifications. *Biopolymers* **2008**, *89*, 93–113.

(30) Wojciechowski, M.; Lesyng, B. Generalized Born Model: Analysis, Refinement, and Applications to Proteins. *J. Phys. Chem. B* **2004**, *108*, 18368–18376.

(31) Thickett, S. C.; Gilbert, R. G. Propagation Rate Coefficient of Acrylic Acid: Theoretical Investigation of the Solvent Effect. *Polym.* **2004**, *45*, 6993–6999.

(32) Paasche, A.; Schirmeister, T.; Engels, B. Benchmark Study for the Cysteine–Histidine Proton Transfer Reaction in a Protein Environment: Gas Phase, COSMO, QM/MM Approaches. *J. Chem. Theory Comput.* **2013**, *9*, 1765–1777.

(33) Galano, A.; Alvarez-Idaboy, J. R. Kinetics of Radical-Molecule Reactions in Aqueous Solution: A Benchmark Study of the Performance of Density Functional Methods. *J. Comput. Chem.* **2014**, *35*, 2019–2026.

(34) Diamanti, A.; Ganase, Z.; Grant, E.; Armstrong, A.; Piccione, P. M.; Rea, A. M.; Richardson, J.; Galindo, A.; Adjiman, C. S. Mechanism, Kinetics and Selectivity of a Williamson Ether Synthesis: Elucidation under Different Reaction Conditions. *React. Chem. Eng.* **2021**, *6*, 1195–1211.

(35) Nkungli, N. K.; Tasheh, S. N.; Fouegue, A. D. T.; Bine, F. K.; Ghogomu, J. N. Theoretical Insights into the Direct Radical Scavenging Activities of 8-Hydroxyquinoline: Mechanistic, Thermodynamic and Kinetic Studies. *Comput. Theor. Chem.* **2021**, *1198*, 113174.

(36) Boulebd, H. Radical Scavenging Behavior of Butylated Hydroxytoluene against Oxygenated Free Radicals in Physiological Environments: Insights from DFT Calculations. *Int. J. Chem. Kinet.* **2022**, *54*, 50–57.

(37) Sure, R.; el Mahdali, M.; Plajer, A.; Deglmann, P. Towards a Converged Strategy for Including Microsolvation in Reaction Mechanism Calculations. *J. Comput. Aided Mol. Des.* **2021**, *35*, 473–492.

(38) Klamt, A. Conductor-like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena. *J. Phys. Chem.* **1995**, *99*, 2224–2235.

(39) Klamt, A.; Jonas, V.; Bürger, T.; Lohrenz, J. C. W. Refinement and Parametrization of COSMO-RS. *J. Phys. Chem. A* **1998**, *102*, 5074–5085.

(40) Klamt, A. The COSMO and COSMO-RS Solvation Models. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 699–709.

(41) Klamt, A. The COSMO and COSMO-RS Solvation Models. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2018**, *8*, e1338.

(42) Hellweg, A.; Eckert, F. Brick by Brick Computation of the Gibbs Free Energy of Reaction in Solution using Quantum Chemistry and COSMO-RS. *AIChE J.* **2017**, *63*, 3944–3954.

(43) Klamt, A.; Eckert, F.; Diedenhofen, M. Prediction of the Free Energy of Hydration of a Challenging Set of Pesticide-Like Compounds. *J. Phys. Chem. B* **2009**, *113*, 4508–4510.

(44) Theilacker, K.; Buhrke, D.; Kaupp, M. Validation of the Direct-COSMO-RS Solvent Model for Diels–Alder Reactions in Aqueous Solution. *J. Chem. Theory Comput.* **2015**, *11*, 111–121.

(45) Sure, R.; Grimme, S. Comprehensive Benchmark of Association (Free) Energies of Realistic Host–Guest Complexes. *J. Chem. Theory Comput.* **2015**, *11*, 3785–3801.

(46) Silva, N. M.; Deglmann, P.; Pliego Jr., J. R. CMIRS Solvation Model for Methanol: Parametrization, Testing, and Comparison with SMD, SM8, and COSMO-RS. *J. Phys. Chem. B* **2016**, *120*, 12660–12668.

(47) Taylor, M.; Yu, H.; Ho, J. Predicting Solvent Effects on S_N2 Reaction Rates: Comparison of QM/MM, Implicit, and MM Explicit Solvent Models. *J. Phys. Chem. B* **2022**, *126*, 9047–9058.

(48) Yuan, W.; Li, Y.; Qi, F. Challenges and Perspectives of Combustion Chemistry Research. *Sci. China Chem.* **2017**, *60*, 1391–1401.

(49) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. S. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018**, *9*, 513–530.

(50) Ansari, R.; Ghorbani, A. Accurate Prediction of Free Solvation Energy of Organic Molecules via Graph Attention Network and Message Passing Neural Network from Pairwise Atomistic Interactions. *arXiv* **2021**, *2105.02048*.

(51) Chung, Y.; Vermeire, F. H.; Wu, H.; Walker, P. J.; Abraham, M. H.; Green, W. H. Group Contribution and Machine Learning Approaches to Predict Abraham Solute Parameters, Solvation Free Energy, and Solvation Enthalpy. *J. Chem. Inf. Model.* **2022**, *62*, 433–446.

(52) Zhang, D.; Xia, S.; Zhang, Y. Accurate Prediction of Aqueous Free Solvation Energies Using 3D Atomic Feature-Based Graph Neural Network with Transfer Learning. *J. Chem. Inf. Model.* **2022**, *62*, 1840–1848.

(53) Yu, J.; Zhang, C.; Cheng, Y.; Yang, Y.-F.; She, Y.-B.; Liu, F.; Su, W.; Su, A. SolvBERT for Solvation Free Energy and Solubility Prediction: A Demonstration of an NLP Model for Predicting the Properties of Molecular Complexes. *Digital Discovery* **2023**, *2*, 409–421.

(54) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M., et al. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388.

(55) Borhani, T. N.; Garcia-Munoz, S.; Luciani, C. V.; Galindo, A.; Adjiman, C. S. Hybrid QSPR Models for the Prediction of the Free Energy of Solvation of Organic Solute/Solvent Pairs. *Phys. Chem. Chem. Phys.* **2019**, *21*, 13706–13720.

(56) Zhang, Z.-Y.; Peng, D.; Liu, L.; Shen, L.; Fang, W.-H. Machine Learning Prediction of Hydration Free Energy with Physically Inspired Descriptors. *J. Phys. Chem. Lett.* **2023**, *14*, 1877–1884.

(57) Hutchinson, S. T.; Kobayashi, R. Solvent-Specific Featurization for Predicting Free Energies of Solvation through Machine Learning. *J. Chem. Inf. Model.* **2019**, *59*, 1338–1346.

(58) Lim, H.; Jung, Y. Delfos: Deep Learning Model for Prediction of Solvation Free Energies in Generic Organic Solvents. *Chem. Sci.* **2019**, *10*, 8306–8315.

(59) Pathak, Y.; Laghuvarapu, S.; Mehta, S.; Priyakumar, U. D. In *Proc. of the AAAI Conf. on Artif. Intell.* 2020; Vol. 34, pp 873–880.

(60) Weinreich, J.; Browning, N. J.; von Lilienfeld, O. A. Machine Learning of Free Energies in Chemical Compound Space using Ensemble Representations: Reaching Experimental Uncertainty for Solvation. *J. Chem. Phys.* **2021**, *154*, 134113.

(61) Vermeire, F. H.; Green, W. H. Transfer learning for solvation free energies: From quantum chemistry to experiments. *Chem. Eng. J.* **2021**, *418*, 129307.

(62) Ravasco, J. M. J. M.; Coelho, J. A. S. Predictive Multivariate Models for Bioorthogonal Inverse-Electron Demand Diels–Alder Reactions. *J. Am. Chem. Soc.* **2020**, *142*, 4235–4241.

(63) Migliaro, I.; Cundari, T. R. Density Functional Study of Methane Activation by Frustrated Lewis Pairs with Group 13 Trihalides and Group 15 Pentahalides and a Machine Learning Analysis of Their Barrier Heights. *J. Chem. Inf. Model.* **2020**, *60*, 4958–4966.

(64) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: the Δ-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087–2096.

(65) Gómez-Flores, C. L.; Maag, D.; Kansari, M.; Vuong, V.-Q.; Irle, S.; Gräter, F.; Kubar, T.; Elstner, M. Accurate Free Energies for Complex Condensed-Phase Reactions Using an Artificial Neural Network Corrected DFTB/MM Methodology. *J. Chem. Theory Comput.* **2022**, *18*, 1213–1226.

(66) Farrar, E. H. E.; Grayson, M. N. Machine Learning and Semi-empirical Calculations: A Synergistic Approach to Rapid, Accurate, and Mechanism-Based Reaction Barrier Prediction. *Chem. Sci.* **2022**, *13*, 7594–7603.

(67) Jorner, K.; Brinck, T.; Norrby, P.-O.; Buttar, D. Machine Learning Meets Mechanistic Modelling for Accurate Prediction of Experimental Activation Energies. *Chem. Sci.* **2021**, *12*, 1163–1175.

(68) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31–36.

(69) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–113.

(70) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(71) Landrum, G. et al. RDKit: Open-Source Cheminformatics Software, https://www.rdkit.org, 2006.

(72) Schwaller, P.; Probst, D.; Vaucher, A. C.; Nair, V. H.; Kreutter, D.; Laino, T.; Reymond, J.-L. Mapping the Space of Chemical Reactions using Attention-Based Neural Networks. *Nat. Mach. Intell.* **2021**, *3*, 144–152.

(73) Heid, E.; Green, W. H. Machine Learning of Reaction Properties via Learned Representations of the Condensed Graph of Reaction. *J. Chem. Inf. Model.* **2021**, *62*, 2101–2110.

(74) Montanari, F.; Kuhnke, L.; Ter Laak, A.; Clevert, D.-A. Modeling Physico-Chemical ADMET Endpoints with Multitask Graph Convolutional Networks. *Molecules* **2019**, *25*, 44.

(75) Xiong, Z.; Wang, D.; Liu, X.; Zhong, F.; Wan, X.; Li, X.; Li, Z.; Luo, X.; Chen, K.; Jiang, H.; Zheng, M. Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. *J. Med. Chem.* **2019**, *63*, 8749–8760.

(76) Song, Y.; Zheng, S.; Niu, Z.; Fu, Z.-H.; Lu, Y.; Yang, Y. In *IJCAI*, 2020; Vol. 2020, pp 2831–2838.

(77) Chuang, K. V.; Gunsalus, L. M.; Keiser, M. J. Learning Molecular Representations for Medicinal Chemistry. *J. Med. Chem.* **2020**, *63*, 8705–8722.

(78) Feinberg, E. N.; Joshi, E.; Pande, V. S.; Cheng, A. C. Improvement in ADMET Prediction with Multitask Deep Featurization. *J. Med. Chem.* **2020**, *63*, 8835–8848.

(79) Cáceres, E. L.; Tudor, M.; Cheng, A. C. Deep Learning Approaches in Predicting ADMET Properties. *Future Med. Chem.* **2020**, *12*, 1995–1999.

(80) Maziarka, Ł.; Danel, T.; Mucha, S.; Rataj, K.; Tabor, J.; Jastrzębski, S. Molecule attention transformer. *arXiv* **2020**, *2002.08264*.

(81) Struble, T. J.; Coley, C. W.; Jensen, K. F. Multitask Prediction of Site Selectivity in Aromatic C–H Functionalization Reactions. *React. Chem. Eng.* **2020**, *5*, 896–902.

(82) Tang, B.; Kramer, S. T.; Fang, M.; Qiu, Y.; Wu, Z.; Xu, D. A Self-Attention Based Message Passing Neural Network for Predicting Molecular Lipophilicity and Aqueous Solubility. *J. Cheminform.* **2020**, *12*, 1–9.

(83) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *Advances in Neural Information Processing Systems* **2015**, *28*.

(84) Guan, Y.; Coley, C. W.; Wu, H.; Ranasinghe, D.; Heid, E.; Struble, T. J.; Pattanaik, L.; Green, W. H.; Jensen, K. F. Regio-Selectivity Prediction with a Machine-Learned Reaction Representation and On-The-Fly Quantum Mechanical Descriptors. *Chem. Sci.* **2021**, *12*, 2198–2208.

(85) Wieder, O.; Kuenemann, M.; Wieder, M.; Seidel, T.; Meyer, C.; Bryant, S. D.; Langer, T. Improved Lipophilicity and Aqueous Solubility Prediction with Composite Graph Neural Networks. *Molecules* **2021**, *26*, 6185.

(86) Li, C.; Wang, J.; Niu, Z.; Yao, J.; Zeng, X. A Spatial-Temporal Gated Attention Module for Molecular Property Prediction based on Molecular Geometry. *Brief. Bioinform.* **2021**, *22*, bbab078.

(87) Magar, R.; Wang, Y.; Lorsung, C.; Liang, C.; Ramasubramanian, H.; Li, P.; Farimani, A. B. AugLiChem: Data Augmentation Library of Chemical Structures for Machine Learning. *Mach. Learn.: Sci. Technol.* **2022**, *3*, 045015.

(88) Greenman, K. P.; Green, W. H.; Gomez-Bombarelli, R. Multi-Fidelity Prediction of Molecular Optical Peaks with Deep Learning. *Chem. Sci.* **2022**, *13*, 1152–1162.

(89) Wang, Y.; Wang, J.; Cao, Z.; Barati Farimani, A. Molecular Contrastive Learning of Representations via Graph Neural Networks. *Nat. Mach. Intell.* **2022**, *4*, 279–287.

(90) Aldeghi, M.; Coley, C. W. A Graph Representation of Molecular Ensembles for Polymer Property Prediction. *Chem. Sci.* **2022**, *13*, 10486–10498.

(91) Kang, B.; Seok, C.; Lee, J. A Benchmark Study of Machine Learning Methods for Molecular Electronic Transition: Tree-based Ensemble Learning versus Graph Neural Network. *Bull. Korean Chem. Soc.* **2022**, *43*, 328–335.

(92) Han, X.; Jia, M.; Chang, Y.; Li, Y.; Wu, S. Directed Message Passing Neural Network (D-MPNN) with Graph Edge Attention (GEA) for Property Prediction of Biofuel-Relevant Species. *Energy and AI* **2022**, *10*, 100201.

(93) Stuyver, T.; Coley, C. W. Machine Learning-Guided Computational Screening of New Candidate Reactions with High Bioorthogonal Click Potential. *Chemistry–A European Journal* **2023**, e202300387.

(94) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; Von Lilienfeld, O. A. Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *J. Chem. Theory Comput.* **2017**, *13*, 5255–5264.

(95) Guan, Y.; Lee, T.; Wang, K.; Yu, S.; McWilliams, J. C. S_NAr Regioselectivity Predictions: Machine Learning Triggering DFT Reaction Modeling through Statistical Threshold. *J. Chem. Inf. Model.* **2023**, *63*, 3751–3760.

(96) Liu, Y.; Li, Z. Predict Ionization Energy of Molecules Using Conventional and Graph-Based Machine Learning Models. *J. Chem. Inf. Model.* **2023**, *63*, 806–814.

(97) Guo, J.; Sun, M.; Zhao, X.; Shi, C.; Su, H.; Guo, Y.; Pu, X. General Graph Neural Network-Based Model To Accurately Predict Cocrystal Density and Insight from Data Quality and Feature Representation. *J. Chem. Inf. Model.* **2023**, *63*, 1143–1156.

(98) Isert, C.; Kromann, J. C.; Stiefl, N.; Schneider, G.; Lewis, R. A. Machine Learning for Fast, Quantum Mechanics-Based Approximation of Drug Lipophilicity. *ACS Omega* **2023**, *8*, 2046–2056.

(99) Riedmiller, K.; Reiser, P.; Bobkova, E.; Maltsev, K.; Gryn'ova, G.; Friederich, P.; Gräter, F. Predicting Reaction Barriers of Hydrogen Atom Transfer in Proteins. *10.26434/chemrxiv-2023-7hntk* **2023**.

(100) Duan, Y.-J.; Fu, L.; Zhang, X.-C.; Long, T.-Z.; He, Y.-H.; Liu, Z.-Q.; Lu, A.-P.; Deng, Y.-F.; Hsieh, C.-Y.; Hou, T.-J.; Cao, D.-S. Improved GNNs for Log $D_{7.4}$ Prediction by Transferring Knowledge from Low-Fidelity Data. *J. Chem. Inf. Model.* **2023**, DOI: https://doi.org/10.1021/acs.jcim.2c01564.

(101) Liu, C.; Sun, Y.; Davis, R.; Cardona, S. T.; Hu, P. ABT-MPNN: An Atom-Bond Transformer-Based Message-Passing Neural Network for Molecular Property Prediction. *J. Cheminformatics* **2023**, *15*, 29.

(102) Ren, G.-P.; Wu, K.-J.; He, Y. Enhancing Molecular Representations Via Graph Transformation Layers. *J. Chem. Inf. Model.* **2023**, *63*, 2679–2688.

(103) AbdulHameed, M. D. M.; Liu, R.; Wallqvist, A. Using a Graph Convolutional Neural Network Model to Identify Bile Salt Export Pump Inhibitors. *ACS Omega* **2023**, *8*, 21853–21861.

(104) Li, J.; Cai, D.; He, X. Learning Graph-Level Representation for Drug Discovery. *arXiv* **2017**, *1709.03741*.

(105) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. In *International Conference on Machine Learning*, 2017, pp 1263–1272.

(106) Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; Jensen, K. F. Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction. *J. Chem. Inf. Model.* **2017**, *57*, 1757–1772.

(107) Hop, P.; Allgood, B.; Yu, J. Geometric Deep Learning Autonomously Learns Chemical Features That Outperform Those Engineered by Domain Experts. *Mol. Pharmaceutics* **2018**, *15*, 4371–4377.

(108) Feinberg, E. N.; Sur, D.; Wu, Z.; Husic, B. E.; Mai, H.; Li, Y.; Sun, S.; Yang, J.; Ramsundar, B.; Pande, V. S. PotentialNet for Molecular Property Prediction. *ACS Cent. Sci.* **2018**, *4*, 1520–1530.

(109) Spiekermann, K. A.; Pattanaik, L.; Green, W. H. High Accuracy Barrier Heights, Enthalpies, and Rate Coefficients for Chemical Reactions. *Sci. Data* **2022**, *9*, 417.

(110) Von Rudorff, G. F.; Heinen, S. N.; Bragato, M.; von Lilienfeld, O. A. Thousands of Reactants and Transition States for Competing E2 and S2 Reactions. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045026.

(111) Harms, N.; Underkoffler, C.; West, R. H. Advances in Automated Transition State Theory Calculations: Improvements on the AutoTST Framework. *10.26434/chemrxiv* **2020**, *13277870.v2*, Accessed 2022-03-08.

(112) Johnson, M. S.; Dong, X.; Grinberg Dana, A.; Chung, Y.; Farina Jr., D.; Gillis, R. J.; Liu, M.; Yee, N. W.; Blondal, K.; Mazeau, E.; Grambow, C. A.; Payne, A. M.; Spiekermann, K. A.; Pang, H.-W.; Goldsmith, F. C.; West, R. H.; Green, W. H. RMG Database for Chemical Property Prediction. *J. Chem. Inf. Model.* **2022**, *62*, 4906–4915.

(113) Zimmerman, P. Reliable Transition State Searches Integrated with the Growing String Method. *J. Chem. Theory Comput.* **2013**, *9*, 3043–3050.

(114) Prasad, V. K.; Pei, Z.; Edelmann, S.; Otero-de-la-Roza, A.; DiLabio, G. A. BH9, a New Comprehensive Benchmark Data Set for Barrier Heights and Reaction Energies: Assessment of Density Functional Approximations and Basis Set Incompleteness Potentials. *J. Chem. Theory Comput.* **2022**, *18*, 151–166.

(115) Zhao, Q.; Vaddadi, S. M.; Woulfe, M.; Ogunfowora, L.; Garimella, S.; Isayev, O.; Savoie, B. M. Comprehensive Exploration of Graphically Defined Reaction Spaces. *Sci. Data* **2023**, *10*, 145.

(116) Tavakoli, M.; Chiu, Y. T. T.; Baldi, P.; Carlton, A. M.; Van Vranken, D. RMechDB: A Public Database of Elementary Radical Reaction Steps. *J. Chem. Inf. Model.* **2023**, *63*, 1114–1123.

(117) Liu, M.; Grinberg Dana, A.; Johnson, M. S.; Goldman, M. J.; Jocher, A.; Payne, A. M.; Grambow, C. A.; Han, K.; Yee, N. W.; Mazeau, E. J., et al. Reaction Mechanism Generator v3.0: Advances in Automatic Mechanism Generation. *J. Chem. Inf. Model.* **2021**, *61*, 2686–2696.

(118) Suleimanov, Y. V.; Green, W. H. Automated Discovery of Elementary Chemical Reaction Steps Using Freezing String and Berny Optimization Methods. *J. Chem. Theory Comput.* **2015**, *11*, 4248–4259.

(119) Grambow, C. A.; Jamal, A.; Li, Y.-P.; Green, W. H.; Zador, J.; Suleimanov, Y. V. Unimolecular Reaction Pathways of a γ-Ketohydroperoxide from Combined Application of Automated Reaction Discovery Methods. *J. Am. Chem. Soc.* **2018**, *140*, 1035–1048.

(120) Maeda, S.; Harabuchi, Y. On Benchmarking of Automated Methods for Performing Exhaustive Reaction Path Search. *J. Chem. Theory Comput.* **2019**, *15*, 2111–2115.

(121) Koerstz, M.; Rasmussen, M. H.; Jensen, J. H. Fast and Automated Identification of Reactions with Low Barriers: the Decomposition of 3-Hydroperoxypropanal. *SciPost Chemistry* **2021**, *1*, 003.

(122) TURBOMOLE V7.5 2020, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989-2007, TURBOMOLE GmbH, since 2007; available from https://www.turbomole.org.

(123) Balasubramani, S. G.; Chen, G. P.; Coriani, S.; Diedenhofen, M.; Frank, M. S.; Franzke, Y. J.; Furche, F.; Grotjahn, R.; Harding, M. E.; Hättig, C.; Hellweg, A.; Helmich-Paris, B.; Holzer, C.; Huniar, U.; Kaupp, M.; Marefat Khah, A.; Karbalaei Khani, S.; Müller, T.; Mack, F.; Nguyen, B. D.; Parker, S. M.; Perlt, E.; Rappoport, D.; Reiter, K.; Roy, S.; Rückert, M.; Schmitz, G.; Sierka, M.; Tapavicza, E.; Tew, D. P.; van Wüllen, C.; Voora, V. K.; Weigend, F.; Wodyński, A.; Yu, J. M. TURBOMOLE: Modular program suite for ab initio quantum-chemical and condensed-matter simulations. *J. Chem. Phys.* **2020**, *152*, 184107.

(124) Zimmerman, P. M. Single-Ended Transition State Finding with the Growing String Method. *J. Comput. Chem.* **2015**, *36*, 601–611.

(125) Geiger, J.; Settels, V.; Deglmann, P.; Schäfer, A.; Bergeler, M. Automated Input Structure Generation for Single-Ended Reaction Path Optimizations. *J. Comput. Chem.* **2022**, *43*, 1662–1674.

(126) Baker, J.; Kessi, A.; Delley, B. The Generation and Use of Delocalized Internal Coordinates in Geometry Optimization. *J. Chem. Phys.* **1996**, *105*, 192–212.

(127) Becke, A. D. Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys* **1993**, *98*, 5648–5652.

(128) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B* **1988**, *37*, 785–789.

(129) Kim, K.; Jordan, K. D. Comparison of Density Functional and MP2 Calculations on the Water Monomer and Dimer. *J. Phys. Chem.* **1994**, *98*, 10089–10094.

(130) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra using Density Functional Force Fields. *J. Phys. Chem.* **1994**, *98*, 11623–11627.

(131) Vosko, S. H.; Wilk, L.; Nusair, M. Accurate Spin-Dependent Electron Liquid Correlation Energies for Local Spin Density Calculations: A Critical Analysis. *Can. J. Phys.* **1980**, *58*, 1200–1211.

(132) Grimme, S. Density Functional Theory with London Dispersion Corrections. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 211–228.

(133) Becke, A. D.; Johnson, E. R. A Density-Functional Model of the Dispersion Interaction. *J. Chem. Phys.* **2005**, *123*.

(134) Johnson, E. R.; Becke, A. D. A Post-Hartree–Fock Model of Intermolecular Interactions. *J. Chem. Phys.* **2005**, *123*.

(135) Weigend, F.; Ahlrichs, R. Balanced Basis Sets of Split Valence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to Rn: Design and Assessment of Accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

(136) Eichkorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R. Auxiliary Basis Sets to Approximate Coulomb Potentials. *Chem. Phys. Lett.* **1995**, *240*, 283–290.

(137) Weigend, F. Accurate Coulomb-Fitting Basis Sets for H to Rn. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1057–1065.

(138) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. Climbing the Density Functional Ladder: Nonempirical Meta–Generalized Gradient Approximation Designed for Molecules and Solids. *Phys. Rev. Lett.* **2003**, *91*, 146401.

(139) Kanai, Y.; Wang, X.; Selloni, A.; Car, R. Testing the TPSS Meta-Generalized-Gradient-Approximation Exchange-Correlation Functional in Calculations of Transition States and Reaction Barriers. *J. Chem. Phys.* **2006**, *125*, 234104.

(140) Deglmann, P.; May, K.; Furche, F.; Ahlrichs, R. Nuclear Second Analytical Derivative Calculations using Auxiliary Basis Set Expansions. *Chem. Phys. Lett.* **2004**, *384*, 103–107.

(141) Zhao, Y.; Truhlar, D. G. The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent Interactions, Excited States, and Transition Elements: Two New Functionals and Systematic Testing of Four M06-Class Functionals and 12 Other Functionals. *Theor. Chem. Acc.* **2008**, *120*, 215–241.

(142) Mardirossian, N.; Head-Gordon, M. Thirty Years of Density Functional Theory in Computational Chemistry: An Overview and Extensive Assessment of 200 Density Functionals. *Mol. Phys.* **2017**, *115*, 2315–2372.

(143) Bursch, M.; Mewes, J.-M.; Hansen, A.; Grimme, S. Best-Practice DFT Protocols for Basic Molecular Computational Chemistry. *Angewandte Chemie* **2022**, e202205735.

(144) Fukui, K. The Path of Chemical Reactions-The IRC Approach. *Acc. Chem. Res.* **1981**, *14*, 363–368.

(145) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. Gaussian 16 Revision C.01, Gaussian Inc. Wallingford CT, 2016.

(146) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminf.* **2011**, *3*, 1–14.

(147) Dassault Systèmes. BIOVIA COSMOtherm, Release 2020.

(148) Deglmann, P.; Hungenberg, K.-D.; Vale, H. M. Dependence of Propagation Rate Coefficients in Radical Polymerization on Solution Properties: A Quantitative Thermodynamic Interpretation. *Macromol. React. Eng.* **2018**, *12*, 1800010.

(149) Chung, Y.; Green, W. H. Computing Kinetic Solvent Effects and Liquid Phase Rate Constants Using Quantum Chemistry and COSMO-RS Methods. *J. Phys. Chem. A* **2023**, *127*, 5637–5651.

(150) Perdew, J. P. Density-Functional Approximation for the Correlation Energy of the Inhomogeneous Electron Gas. *Phys. Rev. B* **1986**, *33*, 8822.

(151) Perdew, J. P. Erratum: Density-Functional Approximation for the Correlation Energy of the Inhomogeneous Electron Gas. *Phys. Rev. B* **1986**, *34*, 7406.

(152) Schäfer, A.; Huber, C.; Ahlrichs, R. Fully Optimized Contracted Gaussian Basis Sets of Triple Zeta Valence Quality for Atoms Li to Kr. *J. Chem. Phys.* **1994**, *100*, 5829–5835.

(153) Sülzner, N.; Haberhauer, J.; Hättig, C.; Hellweg, A. Prediction of Acid pK_a Values in the Solvent Acetone Based on COSMO-RS. *J. Comput. Chem.* **2022**, *43*, 1011–1022.

(154) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.

(155) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods VI: More Modifications to the NDDO Approximations and Re-Optimization of Parameters. *J. Mol. Model.* **2013**, *19*, 1–32.

(156) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB–An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.

(157) Rong, Y.; Bian, Y.; Xu, T.; Xie, W.; Wei, Y.; Huang, W.; Huang, J. Self-Supervised Graph Transformer on Large-Scale Molecular Data. *Advances in Neural Information Processing Systems* **2020**, *33*, 12559–12571.

(158) Bacciu, D.; Errica, F.; Micheli, A.; Podda, M. A Gentle Introduction to Deep Learning for Graphs. *Neural Networks* **2020**, *129*, 203–221.

(159) Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V. P. Substructural Fragments: An Universal Language to Encode Reactions, Molecular and Supramolecular Structures. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 693–703.

(160) Hoonakker, F.; Lachiche, N.; Varnek, A.; Wagner, A. Condensed Graph of Reaction: Considering a Chemical Reaction as One Single Pseudo Molecule. *Int. J. Artif. Intell. Tools* **2011**, *20*, 253–270.

(161) Madzhidov, T. I.; Polishchuk, P. G.; Nugmanov, R. I.; Bodrov, A. V.; Lin, A. I.; Baskin, I. I.; Varnek, A. A.; Antipin, I. Structure-Reactivity Relationships in Terms of the Condensed Graphs of Reactions. *Russ. J. Org. Chem.* **2014**, *50*, 459–463.

(162) Spiekermann, K. A.; Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M., et al., Accessed 2023-05-30.

(163) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(164) Chen, T.; Guestrin, C. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM: San Francisco, California, USA, 2016, pp 785–794.

(165) Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

(166) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.

(167) Gedeck, P.; Rohde, B.; Bartels, C. QSAR - How Good Is It in Practice? Comparison of Descriptor Sets on an Unbiased Cross Section of Corporate Data Sets. *J. Chem. Inf. Model.* **2006**, *46*, 1924–1936.

(168) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.

(169) Bender, A.; Schneider, N.; Segler, M.; Patrick Walters, W.; Engkvist, O.; Rodrigues, T. Evaluation Guidelines for Machine Learning Tools in the Chemical Sciences. *Nat. Rev. Chem.* **2022**, *6*, 428–442.

(170) Qian, J., Sampling In *International Encyclopedia of Education (Third Edition)*, Peterson, P., Baker, E., McGaw, B., Eds., Third Edition; Elsevier: Oxford, 2010, pp 390–395.

(171) Bartók, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R.; Csányi, G.; Ceriotti, M. Machine Learning Unifies the Modeling of Materials and Molecules. *Sci. Adv.* **2017**, *3*, e1701816.

(172) Dobson, C. M. Chemical Space and Biology. *Nature* **2004**, *432*, 824–828.

(173) 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. *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875.

(174) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the Size of Drug-like Chemical Space Based on GDB-17 Data. *J. Comput. Aided Mol. Des.* **2013**, *27*, 675–679.

(175) Izgorodina, E. I.; Coote, M. L. Accurate Ab Initio Prediction of Propagation Rate Coefficients in Free-Radical Polymerization: Acrylonitrile and Vinyl Chloride. *Chem. Phys.* **2006**, *324*, 96–110.

(176) Udvarhelyi, A.; Rodde, S.; Wilcken, R. ReSCoSS: A Flexible Quantum Chemistry Workflow Identifying Relevant Solution Conformers of Drug-Like Molecules. *J. Comput. Aided Mol. Des.* **2021**, *35*, 399–415.

(177) Low, K.; Coote, M. L.; Izgorodina, E. I. Explainable Solvation Free Energy Prediction Combining Graph Neural Networks with Chemical Intuition. *J. Chem. Inf. Model.* **2022**, *62*, 5457–5470.

(178) Burns, J. W.; Spiekermann, K. A.; Bhattacharjee, H.; Vlachos, D. G.; Green, W. H. astartes.

(179) Spiekermann, K. A.; Pattanaik, L.; Green, W. H. Fast Predictions of Reaction Barrier Heights: Toward Coupled-Cluster Accuracy. *J. Phys. Chem. A* **2022**, *126*, 3976–3986.

(180) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(181) Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; Leskovec, J. Strategies for Pre-training Graph Neural Networks. *arXiv* **2019**, *1905.12265*.

(182) Li, X.; Fourches, D. Inductive Transfer Learning for Molecular Activity Prediction: Next-Gen QSAR Models with MolPMoFiT. *J. Cheminform.* **2020**, *12*, 1–15.

(183) Hwang, D.; Yang, S.; Kwon, Y.; Lee, K. H.; Lee, G.; Jo, H.; Yoon, S.; Ryu, S. Comprehensive Study on Molecular Supervised Learning with Graph Neural Networks. *J. Chem. Inf. Model.* **2020**, *60*, 5936–5945.

(184) Meredig, B.; Antono, E.; Church, C.; Hutchinson, M.; Ling, J.; Paradiso, S.; Blaiszik, B.; Foster, I.; Gibbons, B.; Hattrick-Simpers, J.; Mehta, A.; Ward, L. Can Machine Learning Identify the Next High-Temperature Superconductor? Examining Extrapolation Performance for Materials Discovery. *Mol. Syst. Des. Eng.* **2018**, *3*, 819–825.

(185) Durdy, S.; Gaultois, M. W.; Gusev, V. V.; Bollegala, D.; Rosseinsky, M. J. Random Projections and Kernelised Leave One Cluster Out Cross Validation: Universal Baselines and Evaluation Tools for Supervised Machine Learning of Material Properties. *Digital Discovery* **2022**, *1*, 763–778.

(186) Fitzner, M.; Wuitschik, G.; Koller, R.; Adam, J.-M.; Schindler, T. Machine Learning C–N Couplings: Obstacles for a General-Purpose Reaction Yield Prediction. *ACS Omega* **2023**, *8*, 3017–3025.

(187) Yarish, D.; Garkot, S.; Grygorenko, O. O.; Radchenko, D. S.; Moroz, Y. S.; Gurbych, O. Advancing Molecular Graphs with Descriptors for the Prediction of Chemical Reaction Yields. *J. Comput. Chem.* **2023**, *44*, 76–92.

(188) Espley, S. G.; Farrar, E. H. E.; Buttar, D.; Tomasi, S.; Grayson, M. N. Machine Learning Reaction Barriers in Low Data Regimes: A Horizontal and Diagonal Transfer Learning Approach. *Digital Discovery* **2023**, DOI: 10.1039/D3DD00085K.

(189) Tricarico, G. A.; Hofmans, J.; Lenselink, E. B.; Ramos, M. L.; Dréanic, M.-P.; Stouten, P. F. W. Construction of Balanced, Chemically Dissimilar Training, Validation and Test Sets for Machine Learning on Molecular Datasets. *10.26434/chemrxiv-2022-m8l33* **2022**.

(190) Terrones, G. G.; Duan, C.; Nandy, A.; Kulik, H. J. Low-Cost Machine Learning Prediction of Excited State Properties of Iridium-Centered Phosphors. *Chem. Sci.* **2023**, *14*, 1419–1433.

(191) Xiong, Z.; Cui, Y.; Liu, Z.; Zhao, Y.; Hu, M.; Hu, J. Evaluating Explorative Prediction Power of Machine Learning Algorithms for Materials Discovery using k-fold Forward Cross-Validation. *Comp. Mater. Sci.* **2020**, *171*, 109203.

(192) Kovács, D. P.; McCorkindale, W.; Lee, A. A. Quantitative Interpretation Explains Machine Learning Models for Chemical Reaction Prediction and Uncovers Bias. *Nat. Commun.* **2021**, *12*, 1–9.

(193) Stuyver, T.; Coley, C. W. Quantum Chemistry-Augmented Neural Networks for Reactivity Prediction: Performance, Generalizability, and Explainability. *J. Chem. Phys.* **2022**, *156*, 084104.

(194) Artrith, N.; Butler, K. T.; Coudert, F.-X.; Han, S.; Isayev, O.; Jain, A.; Walsh, A. Best Practices in Machine Learning for Chemistry. *Nat. Chem.* **2021**, *13*, 505–508.

(195) Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Oberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical Assessment of QSAR Models of Environmental Toxicity against *Tetrahymena pyriformis*: Focusing on Applicability Domain and Overfitting by Variable Selection. *J. Chem. Inf. Model.* **2008**, *48*, 1733–1746.

(196) Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Molecules* **2012**, *17*, 4791–4810.

(197) Zang, Q.; Mansouri, K.; Williams, A. J.; Judson, R. S.; Allen, D. G.; Casey, W. M.; Kleinstreuer, N. C. In Silico Prediction of Physicochemical Properties of Environmental Chemicals Using Molecular Fingerprints and Machine Learning. *J. Chem. Inf. Model* **2017**, *57*, 36–49.

(198) Nakajima, M.; Nemoto, T. Machine Learning Enabling Prediction of the Bond Dissociation Enthalpy of Hypervalent Iodine from SMILES. *Sci. Rep.* **2021**, *11*, 20207.

(199) Wohlfahrt, C. Static Dielectric Constants of Pure Liquids and Binary Liquid Mixtures; Pure Liquids: Data. *Landolt-Bornstein. Group IV Physical Chemistry* **1991**, *6*.

(200) Fang, X.; Hutcheon, R.; Scola, D. A. Microwave Syntheses of Poly (ε-caprolactam-co-ε-Caprolactone). *J. Polym. Sci., Part A: Polym. Chem.* **2000**, *38*, 1379–1390.

(201) Aparicio, S.; Alcalde, R. Characterization of Two Lactones in Liquid Phase: An Experimental and Computational Approach. *Phys. Chem. Chem. Phys.* **2009**, *11*, 6455–6467.

(202) Evans, M. G.; Polanyi, M. Inertia and Driving Force of Chemical Reactions. *Trans. Faraday Soc.* **1938**, *34*, 11–24.

(203) Grambow, C. A.; Pattanaik, L.; Green, W. H. Reactants, Products, and Transition States of Elementary Chemical Reactions Based on Quantum Chemistry. *Sci. Data* **2020**, *7*, 1–8.

(204) Green, W. H. Moving from Postdictive to Predictive Kinetics in Reaction Engineering. *AIChE Journal* **2020**, *66*, e17059.

(205) Hoffmann, T.; Gastreich, M. The Next Level in Chemical Space Navigation: Going Far Beyond Enumerable Compound Libraries. *Drug Discovery Today* **2019**, *24*, 1148–1156.

(206) Coimbra, J. T. S.; Feghali, R.; Ribeiro, R. P.; Ramos, M. J.; Fernandes, P. A. The Importance of Intramolecular Hydrogen Bonds on the Translocation of the Small Drug Piracetam through a Lipid Bilayer. *RSC Adv.* **2021**, *11*, 899–908.

(207) Kenny, P. W. Hydrogen-Bond Donors in Drug Design. *J. Med. Chem.* **2022**, *65*, 14261–14275.

(208) Kitagawa, Y.; Saito, T.; Nakanishi, Y.; Kataoka, Y.; Matsui, T.; Kawakami, T.; Okumura, M.; Yamaguchi, K. Spin Contamination Error in Optimized Geometry of Singlet Carbene (^{1}A$_1$) by Broken-Symmetry Method. *J. Phys. Chem. A* **2009**, *113*, 15041–15046.

(209) Kitagawa, Y.; Saito, T.; Yamaguchi, K., Approximate Spin Projection for Broken-Symmetry Method and Its Application In *Symmetry (Group Theory) and Mathematical Treatment in Chemistry*, Akitsu, T., Ed.; IntechOpen: Rijeka, 2018; Chapter 7.

(210) Molfeat, 2023.

(211) McInnes, L.; Healy, J.; Melville, J. Umap: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* **2018**, *1802.03426*.

(212) McInnes, L.; Healy, J.; Saul, N.; Grossberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **2018**, *3*, 861.

(213) Jaccard, P. The Distribution of the Flora in the Alpine Zone. *New Phytologist* **1912**, *11*, 37–50.

(214) Rogers, D. J.; Tanimoto, T. T. A Computer Program for Classifying Plants: The Computer is Programmed to Simulate the Taxonomic Process of Comparing Each Case with Every Other Case. *Science* **1960**, *132*, 1115–1118.

(215) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.

Chapter 7

Towards a Foundation Model to Predict Kinetic Properties

This work is in preparation as Spiekermann, K. A.; Pang, H. W.; Dong, X.; Wu, H.; Pattanaik, L.; Green, W. H. "Enumerating Reaction Space for Small Molecule Chemistry: Towards a Foundation Model to Predict Kinetic Properties" 2023. Hao-Wei Pang, Xiaorui Dong, and Haoyang Wu helped predict values via RMG. Lagnajit Pattanaik assisted with analyzing model results. All code and model weights are freely available at https://github.com/kspieks/NLP_barriers/tree/main as well as https://github.com/kspieks/chemprop/tree/foundation_rxn_model.

7.1 Introduction

Quantitative estimates for kinetic properties, namely reaction barrier heights and reaction energies, are essential for developing kinetic mechanisms, predicting reaction outcomes, and optimizing chemical processes. However, obtaining kinetic parameters is not trivial. The field has historically relied on regressing these parameters from a relatively narrow set of experimental measurements. More recently, ab initio methods, such as quantum chemistry, have become a popular alternative since this approach should be cheaper, faster, and safer in comparison to doing experiments. Still, quantum chemistry calculations exhibit high computational cost that scales anywhere from $\mathcal{O}(N^4)$ for density functional methods with exchange correlation, such as B3LYP [1–4], to $\mathcal{O}(N^7)$ for highly accurate CCSD(T) methods [5, 6], such that N is the number of basis functions.

This cost becomes particularly prohibitive when considering the vastness of chemical space. Modern kinetic mechanisms must often consider hundreds of thousands of possible reactions [7]. More broadly, even when considering only organic species comprised of the elements H, C, N, O, S, and halogens, there exist at least 166 billion organic molecules with up to 17 heavy atoms [8]; this size range is still rather limited as many modern medicines and materials contain much larger species. The scope of possible reactions is intuitively much larger since each molecule can contain multiple reaction sites and participate in multiple reactions, leading to a combinatorial explosion. For example, naively listing all pairwise combinations (e.g., potential bimolecular reactions) in this set is equivalent to 166×10^9 choose 2, which is 1.37×10^{22}. It is imperative to develop tools that can handle numbers of this magnitude in order to achieve the grand vision of quantitative predictive chemistry.

The need for fast estimators has created a ripe opportunity for machine learning (ML). Quantitative structure-activity or structure-property relationships (QSAR or QSPR) have a long history of utilizing traditional ML methods that operate on low-dimensional feature vectors to quickly make predictions [9–11]. Perhaps the most popular example is Benson's group additivity approach to estimate species thermochemistry [12–15], though a myriad of QSAR and QSPR papers have been published over the past decades. Today, more modern techniques are emerging that often outperform traditional ML approaches. For instance, the transformer architecture [16] has revolutionized sequence-to-sequence modeling, and fields such as language processing have rapidly shifted towards creating foundation models [17], which typically refers to models trained on a large amounts of unlabeled data that can be fine-tuned for many specific applications, with one example being the Generative Pre-trained Transformer (GPT) series [18–21]. However, these more complex model architectures are data hungry (e.g., requiring billions of examples), which has proven to be a significant hurdle within the field of chemistry. Although certain chemical tasks have utilized the transformer [22–27], they have yet to have a similarly profound impact. Many published models for predicting kinetic properties have very limited applicability since they are trained using datasets that contain just one reaction class [28–36]. Thus, the main obstacle for quantitative predictive chemistry is the lack of publicly available data in comparison to other fields.

Although several datasets have been published for molecular property prediction [37–53], there are comparatively fewer reaction datasets with atom-mapped SMILES [54–57] and high quality datasets are especially rare [58]. One example comes from von Rudorff et al. [54], who presents approximately 4,500 reactions belonging to the E2 and S_N2 templates. Although this dataset has provided an opportunity to benchmark various modeling approaches, this dataset is fairly narrow in focus; it contains just two reaction templates, all reactions start from a substituted ethyl-based scaffold, and only four nucleophiles are present. To generate datasets that contain more reaction diversity, the field has also utilized more exhaustive approaches that enumerate many possible bond forming and breaking steps with only a few constraints. Examples come from Grambow et al. [55], Zhao et al. [56], and Chapter 6, all of which use the single ended growing string method (GSM) [59] to search for TS geometries. This exhaustive enumeration inherently leads to a very diverse set of reaction templates. Despite the success with using the GSM, it can sometimes identify TSs for relatively high-barrier reactions whose transformations may not be intuitive to chemists and whose high-barriers often preclude them from being added to kinetic mechanisms in the first place. In some cases, it may be beneficial to focus on templates that generally have lower barriers and whose transformations have already proven important for modeling radical chemistry.

Here, we seek to address this paucity of reaction data by providing two new datasets for the field. The first is RMG-DB-11, which enumerates 750 million atom-mapped reaction SMILES [60] that belong to 22 elementary reaction templates from the Reaction Mechanism Generator (RMG) database [61]. These transformations have been selected since they are relevant for radical gas-phase chemistry. This dataset should serve as a reaction equivalent to the popular GDB11, which enumerates organic molecules with up to 11 heavy atoms [62, 63]. We believe this will be valuable for various pre-training tasks within ML and push the field towards creating foundation models. It also provides a starting point for identifying transition state (TS) geometries to eventually calculate kinetic parameters at scale. The second is RDB7-CCSD(T)-F12, which builds upon prior work from Grambow [64] and contains 1,143 radical reactions calculated at CCSD(T)-F12a/cc-pVDZ-F12//ωB97X-D3/def2-TZVP. The explicitly correlated coupled-cluster method is known to be very accurate [5, 65–69] so this data should be useful for radical gas-phase chemistry. Finally, we use the new RMG-DB-11 to pre-train both a graph network and language model, which are then fine-tuned on RDB7-CCSD(T)-F12 and on another published dataset. These new datasets should be valuable to the kinetics community. All data is free and publicly accessible, and we are excited about the future that data-driven techniques can enable.

7.2 Methods

7.2.1 Dataset Generation

A total of three datasets are used in this work, which we refer to as RMG-DB-11, RDB7-CCSD(T)-F12, and RDB19-M06-2X. The first two are created as novel contributions from this work while the latter has already been published in the literature. All reactions contain neutrally charged species and atom-mapped reaction SMILES. Our main contribution is RMG-DB-11, which contains 750 million reactions with up to 11 heavy atoms that match templates from the RMG database. This dataset is useful for defining a subset of reaction space and can be used to pre-train ML models. RDB7-CCSD(T)-F12 contains radical reactions with up to 7 heavy atoms whose barrier heights and reaction energies are calculated using the CCSD(T)-F12 method. This dataset is used to fine-tune pre-trained models. RDB19-M06-2X contains radical reactions with up to 19 heavy atoms whose kinetic properties are calculated using the M06-2X method. It comes from Chapter 6 and is used here as another fine-tuning task. The overall workflow is shown in Figure 7.1.

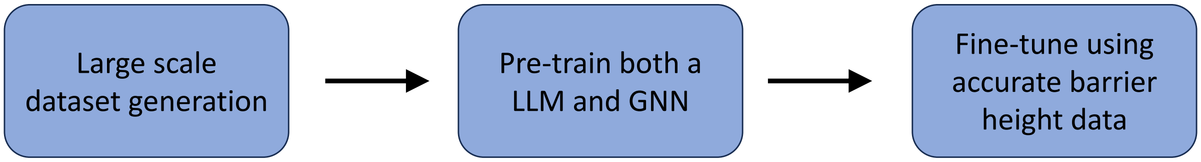
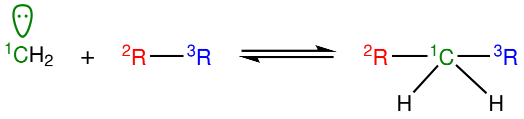
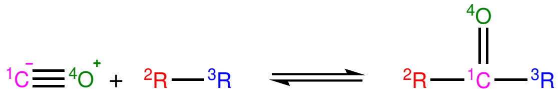
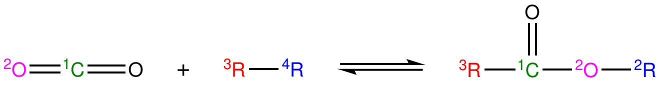
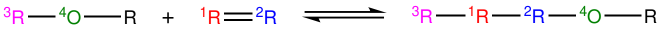
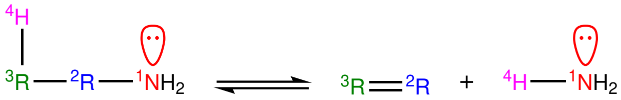
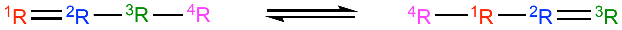
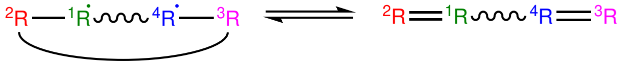
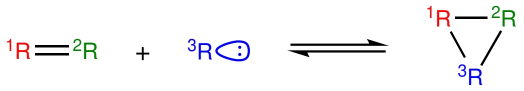
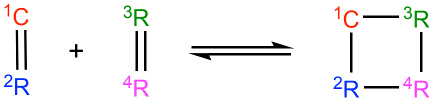
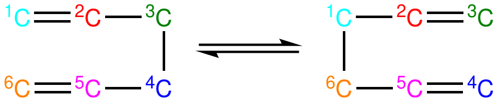
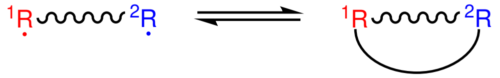
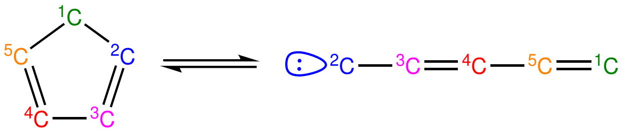
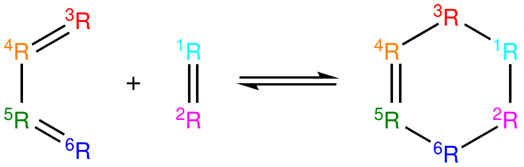

Large scale dataset generation
Pre-train both a LLM and GNN
Fine-tune using accurate barrier height data

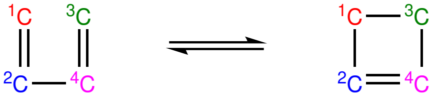
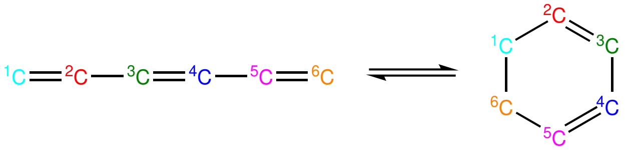
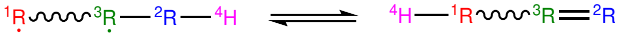
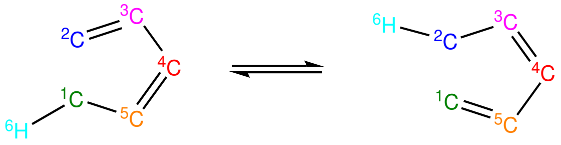
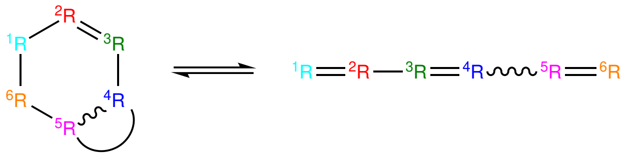
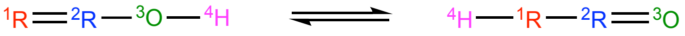
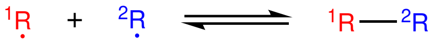
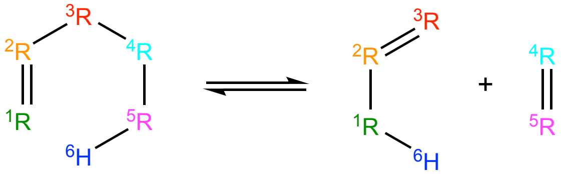
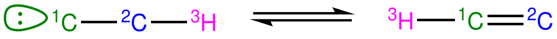
Figure 7.1. Schematic outlining the high-level workflow in this work. First, we generate a large dataset of reactions that belong to RMG templates. This new dataset is used to pre-train both a large language model (LLM) and graph neural network (GNN), each of which is fine-tuned using two different quantum chemistry datasets with highly accurate barrier height values.

7.2.1.1 RMG-DB-11

Dataset generation starts from GDB-11 [62, 63], which enumerates 26.4 million small, organic, stable molecules containing up to 11 heavy atoms and comprised of the elements H, C, N, O, and F. Analysis of the chemical diversity of GDB11 is presented in the original papers. We downsample this dataset by removing all Fluorine-containing species, which leaves approximately 17 million starting molecules. We then apply several reaction templates from RMG to exhaustively generate a list of potential reactions and their corresponding atom-mapped reaction SMILES. Table 7.1 provides the list of 22 elementary reaction templates that define the transformation from the reactant to the product. Dataset validation includes ensuring that all SMILES can be sanitized via RDKit [70] without errors i.e., the SMILES satisfy basic valence rules. Each SMILES is also canonicalized with RDKit. All work is done using the GitHub version commit string of e3d0617fbcc43bdf58c770def18baef0dacb7bf0 on the main branch of RMG-Py and bd1319070f83447e0951d583b4e3a4b770d67ce5 on the main branch of RMG-database. This dataset is accessible on Zenodo [71].

Table 7.1. RMG reaction templates used to generate the RMG-DB-11 dataset. "R" refers to part of the molecule not directly participating in the reaction.

RMG Reaction Family	**Template**
1,2_Insertion_Carbene	
1,2_Insertion_CO	
1,3_Insertion_CO2	
1,3_Insertion_ROR	
1,3_NH3_Elimination	
1,3_Sigmatropic_Rearrangement	
1,4_Cyclic_Birad_Scission	
1+2_Cycloaddition	
2+2_Cycloaddition	
6_Membered_Central_C-C_Shift	
Birad_recombination	
Cyclopentadiene_Scission	
Diels_Alder_Addition	

Intra_2+2_Cycloaddition_Cd	
Intra_Diels_alder_monocyclic	
Intra_Disproportionation	
Intra_ene_reaction	
Intra_Retro_Diels_alder_bicyclic	
Ketoenol	
R_Recombination	
Retroene	
Singlet_Carbene_Intra_Disproportionation	

Pre-training on all 750 million reactions (1.5 billion if including reverse reactions), would require substantial compute resources. Thus, we create a downsampled version of RMG-DB-11. The first filtering criteria is to only use reactions that contain at most 8 heavy atoms. However, this still leaves approximately 3 million reactions when considering the forward and reverse direction, so the second filtering criteria only allows a reaction template to be used once per starting SMILES from GDB8. If a given RMG template matches multiple reaction sites on a molecule, one is chosen at random and the remaining are discarded when creating this subset. This step produces a pre-training set that has more even representation of the reaction templates than is shown in Figure 7.3 for the entire dataset. The downsampled set contains approximately 800,000 reactions when accounting for the forward and reverse directions. This unlabeled data is used to pre-train a language model, though future work could explore using more data for pre-training.

Although this same unsupervised approach could be used to pre-train a graph network [72], here we use supervised pre-training. To create labels for the data, we use RMG to estimate the activation energy (E_a) and the enthalpy of reaction ($\Delta \mathrm{H}_{\mathrm{rxn}}$) at 0 K for each reaction. The former is estimated using RMG's decision trees while the latter is estimated via linear group additivity. However, some species contain groups which are not stored in the RMG database, so this subset is omitted since an estimate could not be obtained. The final regression pre-training set contains just over 700,000 reactions when accounting for the forward and reverse directions. Note that E_a is simply a fitting parameter from an Arrhenius expression that does not offer the same physical interpretation as the reaction barrier height ($\Delta \mathrm{E}_0$), which is defined as the difference in energy–electronic energy plus corrected zero point energy–of a reaction's TS and reactants at 0 K in the gas phase. However, our hypothesis is that the correlation between RMG's estimated E_a and the calculated $\Delta \mathrm{E}_0$ will be close enough to still be useful as a regression pre-training task.

7.2.1.2 RDB7-CCSD(T)-F12

The original source of this dataset is around 1300 radical reactions identified by Colin Grambow [64] and optimized at ωB97X-D3/def2-TZVP [73, 74] using Q-Chem [75]. Similar to previous work from Grambow et al. [55, 76], the TS geometries were identified via the single ended growing string method [59] from a given set of reactants. All reactions start from one species and produce one, two, or three products. Similar to in ref. [58], we first clean the SMILES by fixing a few reactions that initially had charge imbalances due to occasional errors from using Open Babel [77] to perceive connectivity and generate a SMILES from the 3D coordinates. Since all reactions with multiple products contained one van der Waals product complex, we separate the product complexes into individual product geometries and recalculate the optimization and frequency at ωB97X-D3/def2-TZVP. Finally, we refine the single point energies using the explicitly correlated coupled cluster method from Molpro [78]. The R-CCSD(T)-F12 method is used for closed shell species while RO-CCSD(T) is used for radical species to avoid any spin contamination. All single-point energy calculations use the cc-pVDZ-F12 basis set. Although the energy of an individual species from an unrestricted calculation will be less than or equal to the energy from a restricted calculation, these differences should mostly cancel when doing the subtraction to obtain a barrier height or reaction enthalpy. Additional analysis is shown in the Appendix. Some species failed during these various steps so our final dataset contains 1143 reactions with cleaned atom-mapped reaction SMILES.

Zero-point energies (ZPEs) from the harmonic vibrational analysis are added to the electronic energy for the reactant, product, and TS. For each species, a scaling factor is applied to the computed harmonic frequencies used to compute the ZPE. The scaling factor for ωB97X-D3/def2-TZVP was previously calculated as 0.984 from ref. [58] by using the procedure described by Alecu et al. [79] The regression targets used during fine-tuning are the barrier height (ΔE_0) and reaction energy (ΔH_0). The former is determined by computing the difference between the ZPE-corrected TS and reactant energies while the latter is calculated based on the difference of the ZPE-corrected product and reactant energies. All data is accessible on Zenodo [80].

7.2.1.3 RDB19-M06-2X

This dataset comes from Chapter 6. Briefly, this work also uses the growing string method [59] to exhaustively enumerate nearly 6,000 elementary reactions calculated at the M06-2X/def2-QZVP//B3LYP-D3(BJ)/def2-TZVP level of theory. While Chapter 6 focuses on predicting kinetic values in solution, here we fine-tune the models to predict the Gibbs free energy of activation and of reaction in the gas-phase defined as:

$$\Delta G^{\ddagger} = G^{\mathrm{TS}} - \sum_{i=1}^{N\ \text{reactants}} G^{\mathrm{R}_i} \tag{7.1}$$

$$\Delta G^{\mathrm{rxn}} = \sum_{j=1}^{N\ \text{products}} G^{\mathrm{P}_j} - \sum_{i=1}^{N\ \text{reactants}} G^{\mathrm{R}_i} \tag{7.2}$$

Unscaled harmonic frequencies have been used for the ZPE and thermal corrections. A thorough analysis of this dataset is presented in Chapter 6.

7.2.2 Machine Learning Models

7.2.2.1 D-MPNN

When using data-driven methods to model chemistry, graph neural networks (GNNs) are a natural choice [81]. Molecules can be modeled as graphs such that atoms are graph nodes and bonds are graph edges. Each atom and bond is assigned an initial feature vector whose representation is updated with information from neighboring nodes and/or edges to create a learned molecular representation that is passed to a standard feed forward network (FFN) to predict the property of interest. This type of model is advantageous since the input only requires the atom-mapped SMILES of the reactant(s) and product(s), which are converted to molecular graphs using RDKit [70]. No expensive quantum chemistry calculations are required, which makes this model amenable to high-throughput.

Here, we adapt the directed message passing neural network (D-MPNN) developed by Yang et al. [82, 83], which is a type of GNN that passes messages across directed bonds. To encode the reaction, we use the established condensed graph of reaction (CGR) representation as the input [84–87]. The CGR creates a superposition of the reactant and product graphs by concatenating the features of the reactant with the difference of the feature vectors between the reactant and product and thus resembles a 2D-structure of the reaction's TS. Figure 7.2 shows a schematic of the model architecture, which builds a learned reaction embedding by aggregating atomic representations after the message passing phase. The Appendix lists the hyperparameters used when training the model. Our modified code and final weights are freely available on GitHub [88] under the `foundation_rxn_model` branch of the forked repository.

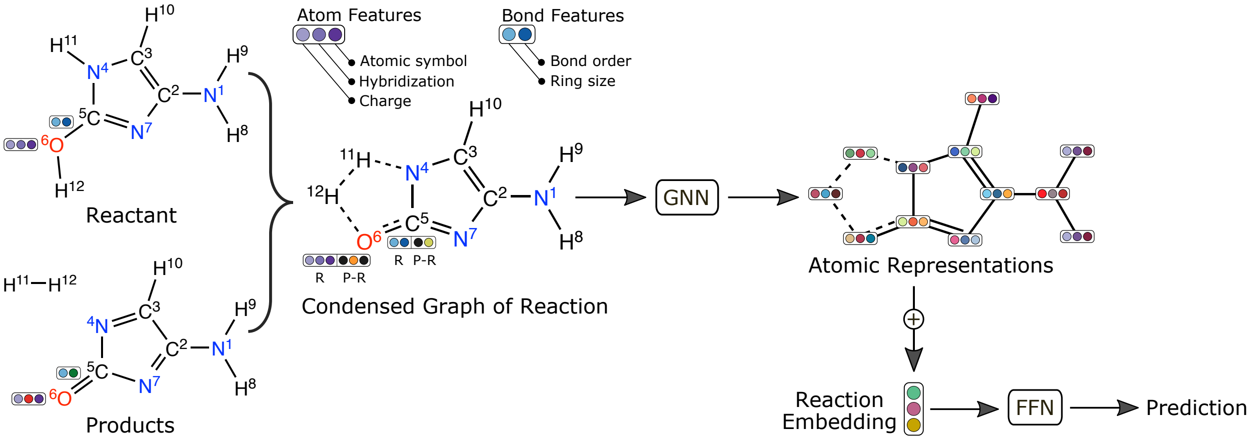


Figure 7.2. Schematic of the machine learning model architecture. Here, the GNN uses the directed message passing scheme introduced by Yang et al. [82] The condensed graph of reaction (CGR) representation [84, 87] is used to predict the reaction's barrier height and corrected enthalpy. The CGR concatenates the featurization of the reactant(s) with the difference of the featurization between the reactant(s) and product(s). A subset of the initial atom and bond features are shown for simplicity. The superscripts indicate the arbitrary atom-map numbers. Colors are for qualitative purposes only.

7.2.2.2 ALBERT

We also use the ALBERT model architecture [89] from the popular Huggingface package [90]. In contrast to its precedent BERT [16], ALBERT shares weights across layers, resulting in smaller models that should be more practical to incorporate into prediction workflows. Similar to the D-MPNN, ALBERT only requires a SMILES string as input. However, the language model does not require atom-mapped SMILES as input, which may offer advantages in some cases. Still, applying reaction templates, such as those from RMG, preserves atom-mapping and recent work [91] should make atom-mapped SMILES easy to obtain even if the template is unknown so this likely presents only a slight benefit for language models.

Unlike graph networks which first convert the SMILES into an attributed graph, the transformer architecture operates directly on the SMILES string as a sequence of tokens. To convert a SMILES string into tokens, we use a regular expression introduced by Schwaller et al. [92] that converts each character (element or bond) into a token. The total vocab size is only 72. Example tokens include C, N, O, (, and). The ALBERT model has 6 layers, 8 attention heads, embedding size of 128, hidden size of 512, and intermediate size of 512. The hyperparameters are listed in the Supporting Information. Our code and model weights are freely available on GitHub [93].

The model is pre-trained using standard masked language modeling (MLM) using the canonicalized SMILES from the downsampled version of RMG-DB-11. This approach randomly masks 15% of the input tokens. The model must then predict the true token, thus learning the vocabulary of the SMILES representation. However, SMILES is a non-unique identifier as there are often many SMILES strings that correspond to the same molecule. Future work could explore data augmentation, which has been shown to slightly improve accuracy for reaction yield predictions [94, 95].

7.2.2.3 Baseline Models

We also include simpler models from Scikit-Learn [96] as baselines, such as random forest (RF) and support vector regression (SVR) using a linear kernel. The best hyperparameters are chosen via Optuna [97] and are listed in the Supporting Information. Since these methods operate on vector inputs, we use Morgan (ECFP) fingerprints [98] with 2048 bit hashing and radius of 2 as calculated using RDKit [70]. To represent the reaction, we concatenate the fingerprint of the reactant complex with the difference of the fingerprint between the reactant and product complex; this is the same approach as the initial featurization in the CGR. The final baseline simply predicts the mean of the training target values for all test reactions [99].

7.2.2.4 Data Splits

Since data-driven methods typically benefit from additional training points, all data splits include reactions from the forward and reverse direction. Splitting is done based on the forward reactions; the corresponding reverse reaction is then added to the same set. Otherwise, the same TS would appear in both the training and testing set, which would cause the testing error to not reflect the true performance when evaluating a new reaction [87, 100].

The pre-training sets are randomly partitioned using a 94:3:3 split (i.e., 94% of the data is assigned to training, 3% to validation, and 3% to testing). The fine-tuning data is partitioned using an 85:5:10 split. For the regression tasks, the values for the reverse reactions are obtained using the following expressions:

$$\Delta \mathrm{E}_{0,\mathrm{reverse}} = \Delta \mathrm{E}_{0,\mathrm{forward}} - \Delta \mathrm{H}_{\mathrm{forward}} \tag{7.3}$$

$$\Delta \mathrm{H}_{0,\mathrm{reverse}} = -\Delta \mathrm{H}_{0,\mathrm{forward}} \tag{7.4}$$

We create both interpolative splits and extrapolative splits based on the procedure in the `astartes` software package [101] so future users are informed of the likely performance in these different settings. Splitting is done based on the forward reactions; the corresponding reverse reaction is then added to the same set. Otherwise, the same transition state would appear in both the training and testing set, which would cause the testing error to not reflect the true performance when evaluating a new reaction [87, 100]. To measure interpolation, we use random splitting to assign 85% of the data to training, 5% to validation, and 10% to testing. To measure extrapolation, such as making predictions for new types of molecules, we cluster the data based on reactant substructure. We first create Morgan fingerprints [98, 102] with 2048 bit hashing and radius of 2 for each reactant complex from the forward reactions using RDKit [70]. We next use K-Means clustering to create 15 clusters. Finally, we assign 13 clusters to the training set, one cluster to the validation set, and one cluster to the testing set to closely approximate an 85:5:10 split.

For fine-tuning, we create both interpolative splits and extrapolative splits based on the procedure in the `astartes` software package [101] so future users are informed of the likely performance in these different settings. To measure interpolation, we use random splitting. To measure extrapolation, such as making predictions for new types of molecules, we cluster the data based on reactant substructure. We first create Morgan fingerprints [98, 102] with 2048 bit hashing and radius of 2 for each reactant complex from the forward reactions using RDKit [70]. For RDB7-CCSD(T)-F12, we use K-Means clustering to create 15 clusters.

Finally, we assign 13 clusters to the training set, one cluster to the validation set, and one cluster to the testing set to closely approximate an 85:5:10 split. For RDB19-M06-2X, we use the exact same splits as published in Chapter 6 which created 12 clusters and then assigned ten clusters to the training set, one cluster to the validation set, and one cluster to the testing set to closely approximate an 85:5:10 split.

7.3 Results & Discussion

7.3.1 Dataset Statistics

7.3.1.1 RMG-DB-11

All data is free and publicly accessible on Zenodo [71]. A summary of the number of generated reactions grouped by the number of heavy atoms is provided in Table 7.2. The combinatorial explosion is evident. Even starting from relatively small species quickly results in millions of reactions. Dividing the number of reactions by the number of starting species shows the expected trend that larger molecules have more possible reaction sites and thus more total reactions can be generated; we anticipate that trillions of reactions could be enumerated by applying RMG templates to the 166 billion species from GDB17 [8]. The large magnitude from Table 7.2 is impressive considering that the scope of this dataset is still relatively limited. For example, all species are neutrally charged and contain up to only 11 heavy atoms that are restricted to the elements C, N, and O. Furthermore, all reactions are unimolecular from at least one direction. The distribution of RMG reaction templates is shown in Figure 7.3.

Table 7.2. Summary of the number of total elementary reactions enumerated from the corresponding number of starting species. Results are grouped by the number of heavy atoms.

Num. of Heavy Atoms	**Num. of Starting Species**	**Num. of Reactions**	**Num. Reactions / Num. Starting Species**
1	3	4	1.3
2	6	17	2.8
3	17	89	5.2
4	64	475	7.4
5	276	3,244	11.8
6	1,409	22,750	16.1
7	7,846	169,500	21.6
8	48,222	1,305,198	27.1
9	312,827	10,266,077	32.8
10	2,132,245	82,304,042	38.6
11	14,648,825	655,009,337	44.7
Total	**17,151,740**	**749,080,733**	**43.7**

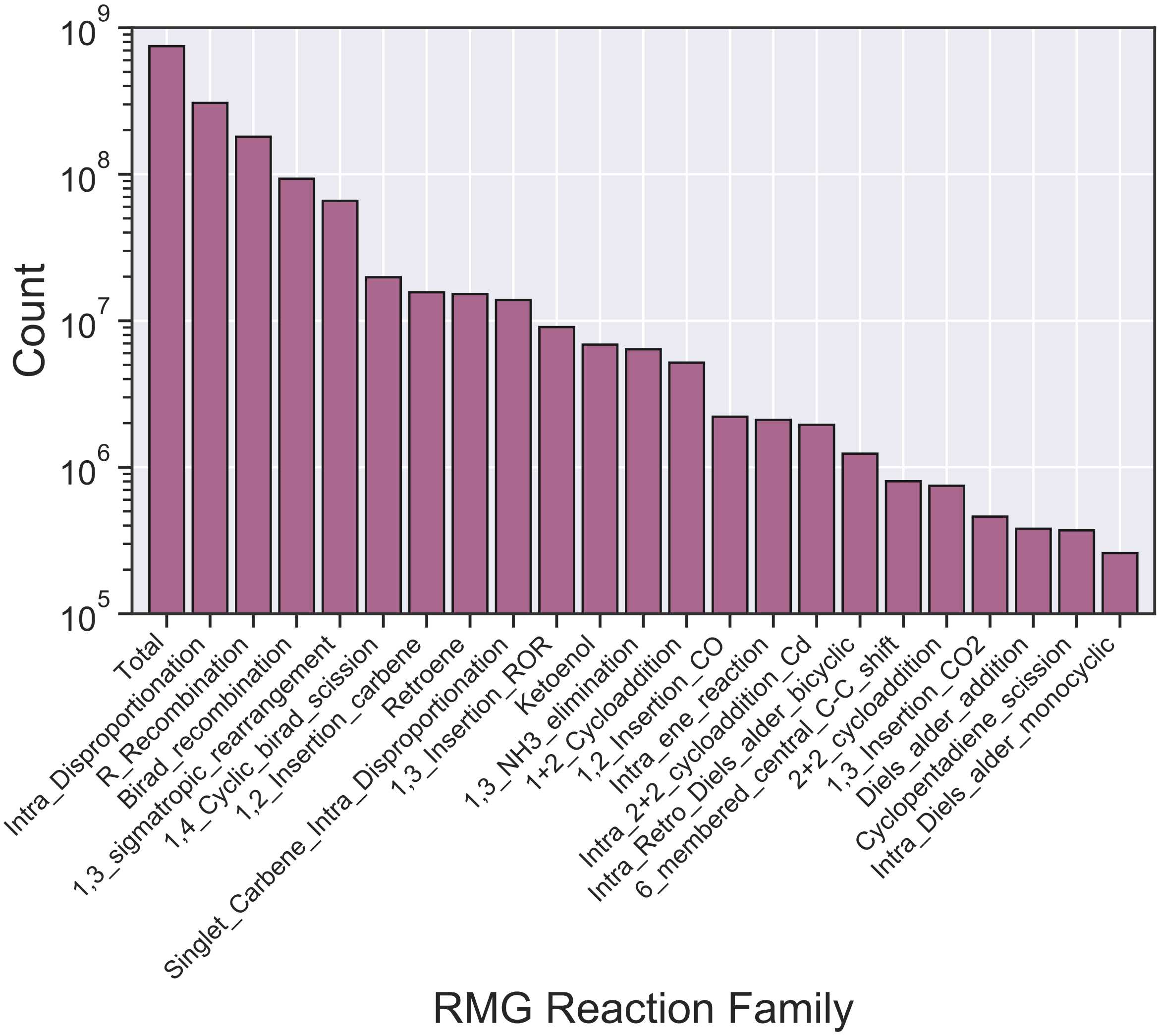


Figure 7.3. Distribution of RMG reaction families present in our RMG-DB-11 dataset.

7.3.1.2 RDB7-CCSD(T)-F12

Our RDB7-CCSD(T)-F12 dataset contains 1,143 reactions with cleaned atom-mapped reaction SMILES. All data is free and publicly accessible on Zenodo [80]. 657 reactions (i.e. ∼57%) do not match any RMG templates. This is an expected result since the TS geometries were identified by Grambow [64] via the single ended growing string method [59], which is known to give large reaction diversity that often extends beyond the hand-curated list of templates from RMG. For example, of the approximately 12,000 reactions originally published by Grambow et al. [55], whose energies were further refined by Spiekermann et al. [58], only about 12% of the reactions match templates from the RMG database. 43% of reactions in RDB7-CCSD(T)-F12 match RMG templates; this higher percentage is expected since RMG, and the corresponding reaction templates, primarily focuses on radical gas-phase chemistry. The distribution of reaction templates is shown in Figure 7.4. Intra hydrogen migration, in which the radical reactant abstracts a hydrogen from itself, is the most common RMG template. R-addition multiple bond is the next most common. Table 7.3 shows the reaction templates.

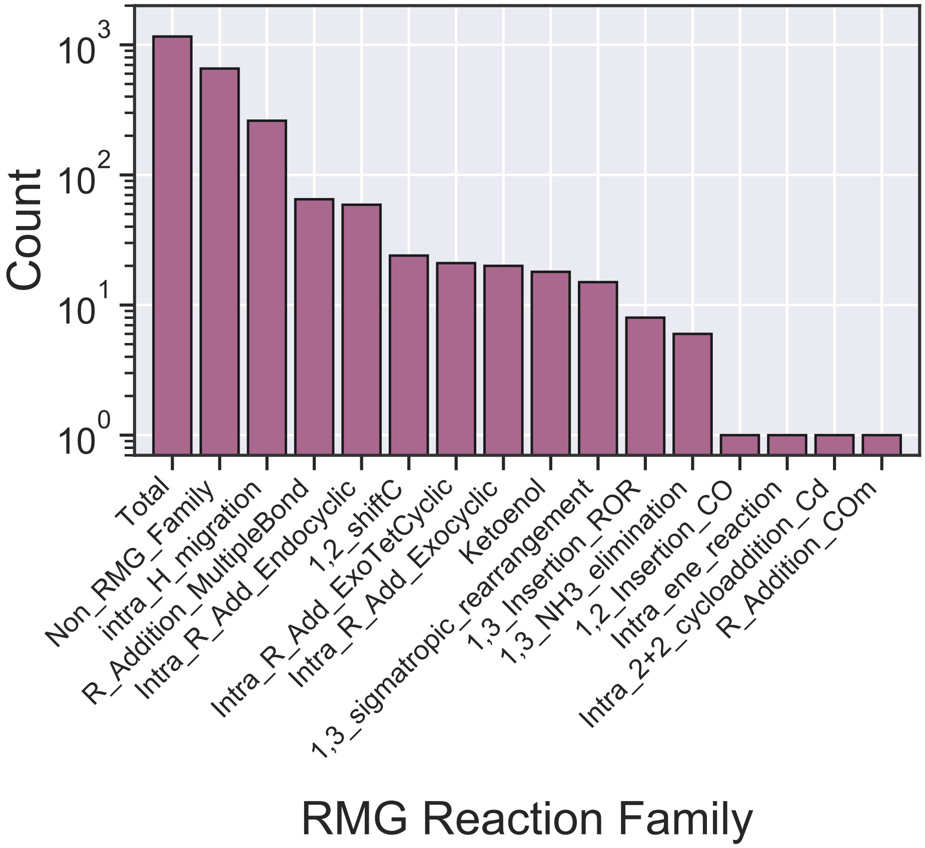
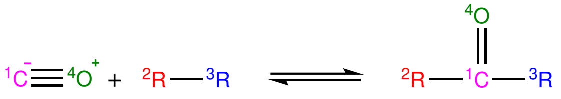
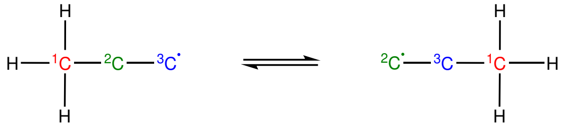
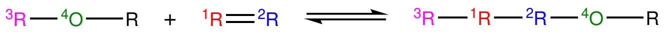
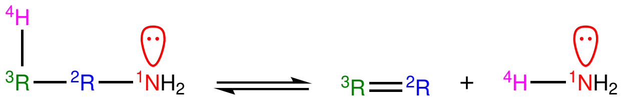
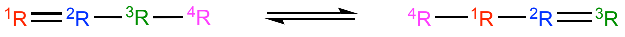
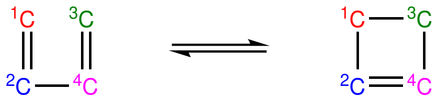
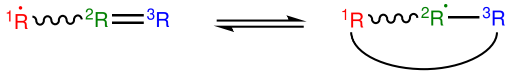
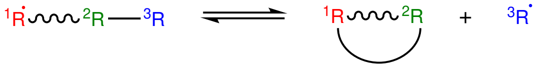
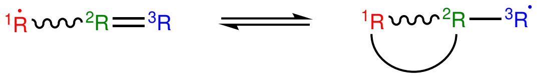
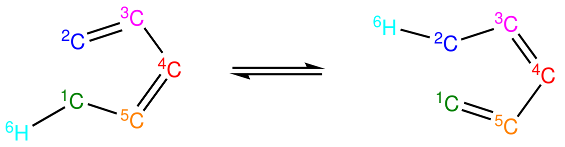
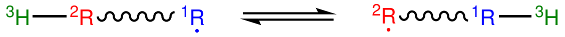
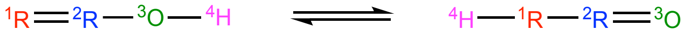
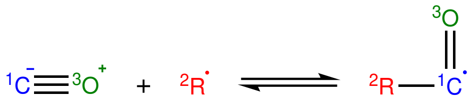
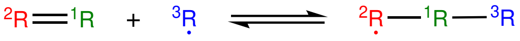


Figure 7.4. Distribution of RMG reaction families present in the RDB7-CCSD(T)-F12 dataset.

Table 7.3. RMG reaction templates present in the RDB7-CCSD(T)-F12 dataset. “R” refers to part of the molecule not directly participating in the reaction.

RMG Reaction Family	Template
1,2_Insertion_CO	
1,2_shiftC	 H H 1C 2C 3C H 2C 3C 1C H H H

1,3_Insertion_ROR	
1,3_NH3_Elimination	
1,3_Sigmatropic_Rearrangement	
Intra_2+2_Cycloaddition_Cd	
Intra_R_Add_Endocyclic	
Intra_R_Add_ExoTetCyclic	
Intra_R_Add_Exocyclic	
Intra_ene_reaction	
Intra_H_migration	
Ketoenol	
R_Addition_COm	
R_Addition_MultipleBond	

As shown in Figure 7.5a, the data generally follows the expected Evans-Polanyi relationship [103] in which the barrier height is positively correlated with the reaction energy. Figure 7.5b shows that the barrier heights calculated at density functional theory by Grambow [64] have a slight systematic offset relative to the explicitly correlated coupled cluster values. On average, the bias is relatively small as the difference is only a couple of kcal mol^{-1}. A few reactions show larger differences between the two methods; further investigation into these reactions could present an area of future research.

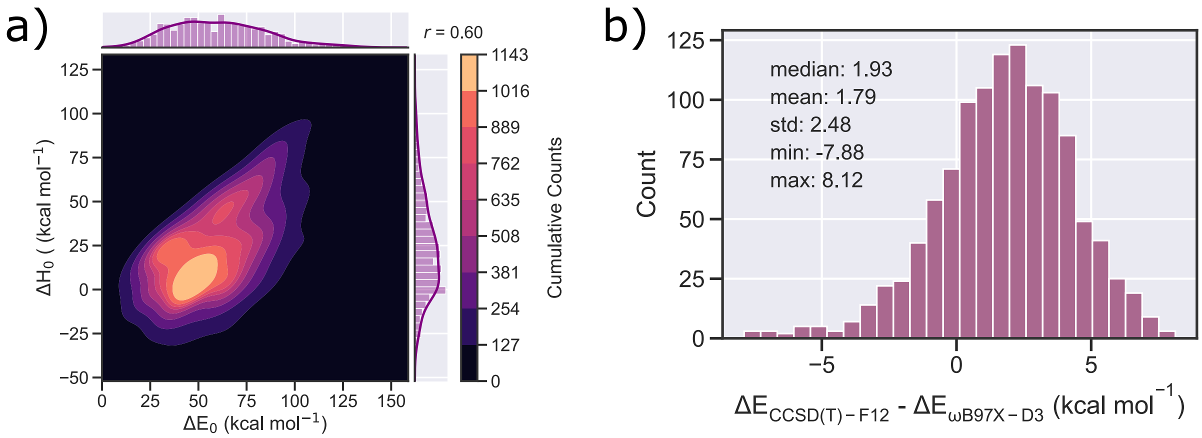


Figure 7.5. Dataset statistics. Panel a) shows the correlation between ΔE_0 and ΔH_0 for the CCSD(T)-F12 values. Panel b) shows the difference in barrier height (ΔE_0) calculated at CCSD(T)-F12a/cc-pVDZ-F12//ωB97X-D3/def2-TZVP and ωB97X-D3/def2-TZVP

7.3.2 Model Results

7.3.2.1 RDB7-CCSD(T)-F12

As the main contribution of this work is the RMG-DB-11 and RDB7-CCSD(T)-F12 datasets, only a preliminary study is done with machine learning. As expected, the baseline model which simply predicts the mean value of the training set gives poor performance with a coefficient of determination (R^2) near 0. As seen in Table 7.4, the D-MPNN does better than the SVR and RF, which is consistent with many published results that empirically demonstrate the learned representations from graph networks often outperform traditional ML models operating on fixed fingerprint representations [22, 23, 30, 35, 36, 44, 52, 82, 87, 104–132], particularly when predicting kinetic properties.

Pre-training the D-MPNN only gives slightly improvements when fine-tuning with random reaction splits. One interpretation of this is that random splits present a relatively simple learning task such that the benefit of pre-training is small. In contrast, pre-training the D-MPNN substantially improves performance on the more challenging K-Means splits. The mean absolute error (MAE) and root-mean squared error (RMSE) are much lower compared to directly training the D-MPNN on RDB7-CCSD(T)-F12 without pre-training. This is a promising result, especially when considering that only 4% of the reactions in the RDB7-CCSD(T)-F12 fine-tuning set match the RMG reaction templates present in the RMG-DB-11 pre-training set. Furthermore, the pre-training regression target included an E_a from an Arrhenius expression estimated by RMG, which does not capture the same physical significance as the ΔE_0 in the fine-tuning set, which was calculated via quantum chemistry.

Table 7.4 also shows that the D-MPNN outperforms the ALBERT model. The outperformance is particularly impressive when considering that the D-MPNN has only about 330,000 parameters, about six times fewer than the nearly two million parameters in our ALBERT model. The D-MPNN's smaller size leads to faster training times and more reasonable computational requirements. Thus, these specific results indicate that representing molecules using a physically-motivated graph structure provides better barrier height predictions in comparison to representing molecules as a sequence of tokens. However, it is possible that additional pre-training could improve performance, and in general there are examples of language models succeeding for certain chemistry-related tasks [133–140].

Still, it is equally important to emphasize that the MAE and RMSE values from all models are too large to be useful. The results from Table 7.4 indicate that these radical reactions present a challenging benchmark for ML models, and future work should continue trying to improve predictions on this set.

Table 7.4. Testing errors (mean ± 1 standard deviation from the five folds) from various models for predicting ΔE_0 from RDB7-CCSD(T)-F12. MAE and RMSE have units of kcal mol^{-1}.

Model	Split	MAE	RMSE	R^2
Baseline (mean)	Random	21.37 ± 1.58	26.69 ± 1.70	0.00 ± 0.00
	K-Means	22.50 ± 2.37	27.90 ± 2.97	-0.01 ± 0.01
SVR (Linear)	Random	16.32 ± 0.83	21.35 ± 1.23	0.36 ± 0.03
	K-Means	18.82 ± 1.97	24.07 ± 2.60	0.25 ± 0.04
Random Forest	Random	16.93 ± 1.46	21.72 ± 1.79	0.34 ± 0.03
	K-Means	20.03 ± 2.49	25.33 ± 3.39	0.17 ± 0.08
D-MPNN without pretraining	Random	11.05 ± 1.04	15.37 ± 1.46	0.66 ± 0.07
	K-Means	14.50 ± 1.84	18.75 ± 2.46	0.53 ± 0.13
D-MPNN with pre-training	Random	10.72 ± 0.88	14.95 ± 1.35	0.68 ± 0.06
	K-Means	12.18 ± 1.83	16.37 ± 2.31	0.65 ± 0.08
ALBERT with pre-training	Random	14.59 ± 0.90	19.67 ± 1.39	0.45 ± 0.05
	K-Means	16.48 ± 1.52	21.79 ± 2.05	0.38 ± 0.08

7.3.2.2 RDB19-M06-2X

Similar results are observed when fine-tuning on RDB19-M06-2X as with RDB7-CCSD(T)-F12 above. As seen in Table 7.5, the D-MPNN does much better than the various baseline models and again outperforms the ALBERT model. For this dataset, pre-training on the regression subset of RMG-DB-11 offers no benefit when examining results from random reaction splits. However, like before, pre-training does improve performance on the more challenging K-Means splits.

The D-MPNN results are promising for several reasons. First, the species in the fine-tuning set are approximately twice as large as those in the pre-training set. Other papers have identified molar mass-based splits as more challenging [40] so it is encouraging to see pre-training improve performance when extrapolating to reactions containing larger species. Second, only 0.1% of the reactions in the RDB19-M06-2X fine-tuning set match the RMG reaction templates present in the RMG-DB-11 pre-training set i.e., pre-training on general radical chemistry appears quite helpful even when extrapolating to entirely new reaction templates. Additionally, 66% of the reactions in RDB19-M06-2X are bimolecular in both directions, while all reactions from RMG-DB-11 are unimolecular from at least one direction. Thus, the benefit of pre-training on our new set of radical reactions is evident.

Table 7.5. Testing errors (mean ± 1 standard deviation from the five folds) from various models for predicting $\Delta G^{\ddagger}$ from RDB19-M06-2X. MAE and RMSE have units of kcal mol^{-1}.

Model	Split	MAE	RMSE	R^2
Baseline (mean)	Random	17.67 ± 0.30	20.78 ± 0.26	0.00 ± 0.00
	K-Means	19.75 ± 2.21	22.70 ± 2.12	-0.11 ± 0.11
SVR (Linear)	Random	4.30 ± 0.23	8.06 ± 0.67	0.85 ± 0.03
	K-Means	5.88 ± 2.90	9.97 ± 4.02	0.75 ± 0.22
Random Forest	Random	7.33 ± 0.15	10.52 ± 0.23	0.74 ± 0.01
	K-Means	10.34 ± 1.45	13.69 ± 1.75	0.60 ± 0.09
D-MPNN without pretraining	Random	2.52 ± 0.07	5.09 ± 0.35	0.94 ± 0.01
	K-Means	3.15 ± 0.36	5.68 ± 0.96	0.93 ± 0.02
D-MPNN with pre-training	Random	2.57 ± 0.10	5.14 ± 0.27	0.94 ± 0.01
	K-Means	2.76 ± 0.42	5.44 ± 0.96	0.93 ± 0.02
ALBERT with pre-training	Random	3.27 ± 0.52	6.58 ± 0.70	0.90 ± 0.02
	K-Means	3.96 ± 0.93	7.42 ± 1.56	0.88 ± 0.05

7.4 Conclusions and Future Directions

Quickly obtaining reliable estimates for kinetic properties remains one of the grand challenges in quantitative predictive chemistry. To date, the lack of suitable data has been the main limitation towards accomplishing this aim. While many previous works have trained data-driven estimators on very specific reaction templates [28–36], the goal of this work is to push the fields towards creating more generalizable models.

To accomplish this, we report two new datasets. The first is RMG-DB-11, which enumerates 750 million atom-mapped reaction SMILES. This dataset acts as a first step towards defining an enormous reaction space and should serve as a reaction equivalent to the popular GDB11 [62, 63]. The second is RDB7-CCSD(T)-F12, which contains about 1,100 radical reactions calculated at CCSD(T)-F12/cc-pVDZ-F12//ωB97X-D3/def2-TZVP. Although some reaction databases exist in the literature [55, 56, 58], these do not focus on radical reactions. Thus, the highly accurate barrier heights in our refined dataset should be valuable.

To demonstrate the value of our new RMG-DB-11 dataset, we pre-train both a graph network and language model and then fine-tune each on two datasets with barrier heights calculated using highly accurate methods. The results in this specific study indicate that graph networks are better suited than language models for predicting kinetic properties. However, future work should continue to explore this further. Identifying good model architectures and generalizable reaction representations remains an important task for computational kinetics.

Despite the utility of RMG-DB-11, there are still several opportunities to further expand the scope of this work. Regarding dataset generation, future work could use these atom-mapped SMILES to identify transition states using a myriad of methods–nudged elastic band [141], the doubled ended growing string method [142], AutoTST [143], or various ML methods [144–147]–and then calculate rate coefficients using established statistical mechanics software [148, 149]. Future work could also explore larger molecules, additional elements and functional groups, or more reaction templates e.g., including those that are bimolecular from both directions. Regarding the ML results, a more thorough hyperparameter search could be beneficial. It is also possible that other architectures from Huggingface could demonstrate better performance. Specifically for training the language model, it would also be interesting to explore other tokenization strategies [150] or try using IUPAC names as input rather than SMILES [151]. Lastly, it's important to continue exploring various methods to measure uncertainty and assess out-of-domain generalizability [152–155].

7.5 Appendix

7.5.1 RDB7-CCSD(T)-F12

7.5.1.1 Analysis of Unrestricted Calculations

All reactions in this dataset are calculated using RO-CCSD(T)-F12a/cc-pVDZ-F12//ωB97X-D3/def2-TZVP to avoid spin contamination. To investigate the magnitude of any inaccuracies caused by using a restricted method, we perform a preliminary test by recalculating the single-point energy of the first two reactions using UCCSD(T)-F12a/cc-pVDZ-F12. The indices for these reactions are `rxn000000` and `rxn000001`. Note that the reactant is identical for these two reactions. For `r000000`, `ts000000`, and `ts000001`, the expectation value of the spin squared operator $<S^2>$ is $<< 0.01$ which would suggest minimal spin contamination. As expected, the energies of the individual reactant and transition state are each lower when using an unrestricted method. However, as shown in Table 7.6, the difference in electronic energy (i.e., the reaction's barrier height) is essentially unchanged. The difference in energy for the individual species largely cancels when doing the subtraction to obtain ΔE_0. A more rigorous study could be done in the future.

Table 7.6. Comparison of the barrier height between using a restricted or unrestricted method for the single-point energy calculation for two sample reactions. ΔE_0 values have units of kcal mol^{-1}.

	$\mathbf{\Delta E_0}$	
Reaction Index	RCCSD(T)-F12	UCCSD(T)-F12
`rxn000000`	122.94	122.68
`rxn000001`	76.35	76.35

7.5.2 Model Hyperparameters

The best hyperparameters for the baseline models from Scikit-Learn [96] are chosen via Optuna [97], which is used to minimize the average root mean squared error (RMSE) on the validation sets. The values are listed in Tables 7.7 and Table 7.8. Any other hyperparameters assumed the default value.

Table 7.7. Optimal hyperparameter values for LinearSVR i.e. support vector regression with a linear kernel.

	Split Type	
Hyperparameter	Random	K-Means
`C`	0.01330	0.01253

Table 7.8. Optimal hyperparameter values for random forest (RF)

	Split Type	
Hyperparameter	Random	K-Means
`max_depth`	15	15
`n_estimators`	125	100

The hyperparameters used for the D-MPNN are shown in Table 7.9 and Table 7.10. Hyperparameters not shown in the tables used the default values from Chemprop [82]. The same values are used when training on both random reaction and K-Means splits.

Table 7.9. Hyperparameters the define the architecture for D-MPNN.

Hyperparameter	**Value**
MPNN hidden size	300
MPNN depth	3
Activation function	LeakyReLU
Aggregation	sum
FFN layers	2

Table 7.10. Hyperparameters used for training the D-MPNN. The "Training" columns refers to hyperparameters used to train directly on either RDB7-CCSD(T)-F12 or RDB19-M06-2X, which acts as a baseline to compare performance to the pre-training + fine-tuning approach.

Hyperparameter	Pre-training	Fine-Tuning	Training
Epochs	2	55	65
Warm-up epochs	1	5	5
Initial learning rate	2×10^{-4}	1×10^{-5}	1×10^{-4}
Maximum learning rate	2×10^{-3}	1×10^{-4}	1×10^{-3}
Final learning rate	2×10^{-5}	1×10^{-6}	1×10^{-5}
Batch size	64	50	32
Gradient clip	10	10	10

The hyperparameters used for the ALBERT model are shown in Table 7.11 and Table 7.12. Hyperparameters not shown in the tables assumed the default values. The same values are used when training on both random reactions and K-Means splits.

Table 7.11. Hyperparameters the define the architecture for ALBERT.

Hyperparameter	**Value**
`attention_probs_dropout_prob`	0.1
`embedding_size`	128
`hidden_act`	gelu new
`hidden_dropout_prob`	0.1
`hidden_size`	512
`initializer_range`	0.02
`inner_group_num`	1
`intermediate_size`	512
`layer_norm_eps`	1×10^{-12}
`max_position_embeddings`	128
`num_attention_heads`	8
`num_hidden_layers`	6
`num_hidden_groups`	1
`position_embedding_type`	relative key

Table 7.12. Hyperparameters used for pre-training and fine-tuning the ALBERT model.

Hyperparameter	Pre-training	Fine-Tuning
Epochs	5	30
Warm-up ratio	0.05	0.05
Max grad norm	3	3
Learning rate	1×10^{-4}	1×10^{-4}
Learning rate scheduler	cosine with restarts	cosine
Batch size	16	16

7.6 References

(1) Becke, A. D. Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys* **1993**, *98*, 5648–5652.

(2) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B* **1988**, *37*, 785–789.

(3) Kim, K.; Jordan, K. D. Comparison of Density Functional and MP2 Calculations on the Water Monomer and Dimer. *J. Phys. Chem.* **1994**, *98*, 10089–10094.

(4) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra using Density Functional Force Fields. *J. Phys. Chem.* **1994**, *98*, 11623–11627.

(5) Adler, T. B.; Knizia, G.; Werner, H.-J. A Simple and Efficient CCSD(T)-F12 Approximation. *J. Chem. Phys.* **2007**, *127*, 221106.

(6) Feller, D.; Peterson, K. A.; Dixon, D. A. A Survey of Factors Contributing to Accurate Theoretical Predictions of Atomization Energies and Molecular Structures. *J. Chem. Phys.* **2008**, *129*, 204105.

(7) Payne, A. M.*; Spiekermann, K. A.*; Green, W. H. Detailed Reaction Mechanism for 350–400°C Pyrolysis of an Alkane, Aromatic, and Long-Chain Alkylaromatic Mixture. *Energy & Fuels* **2022**, *36*, 1635–1646.

(8) 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. *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875.

(9) Hammett, L. P. The Effect of Structure Upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **1937**, *59*, 96–103.

(10) Rogers, D.; Hopfinger, A. J. Application of Genetic Function Approximation to Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854–866.

(11) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. QSPR: The Correlation and Quantitative Prediction of Chemical and Physical Properties from Structure. *Chem. Soc. Rev.* **1995**, *24*, 279–287.

(12) Benson, S. W.; Buss, J. H. Additivity Rules for the Estimation of Molecular Properties. Thermodynamic Properties. *J. Chem. Phys.* **1958**, *29*, 546–572.

(13) Benson, S. W., *Thermochemical Kinetics: Methods for the Estimation of Thermochemical Data and Rate Parameters*; Wiley: 1968.

(14) Benson, S. W., *Thermochemical Kinetics*; Wiley: 1976.

(15) Cohen, N.; Benson, S. W. Estimation of Heats of Formation of Organic Compounds by Additivity Methods. *Chem. Rev.* **1993**, *93*, 2419–2438.

(16) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.

(17) Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E., et al. On the Opportunities and Risks of Foundation Models. *arXiv* **2021**, *2108.07258*.

(18) Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I., et al. Improving Language Understanding by Generative Pre-training. **2018**.

(19) Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I., et al. Language Models are Unsupervised Multitask Learners. *OpenAI blog* **2019**, *1*, 9.

(20) Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A., et al. Language Models are Few-Shot Learners. *Adv. Neural Inf. Process.* **2020**, *33*, 1877–1901.

(21) OpenAI. GPT-4 Technical Report. *arXiv* **2023**, *2303.08774*.

(22) Xiong, Z.; Wang, D.; Liu, X.; Zhong, F.; Wan, X.; Li, X.; Li, Z.; Luo, X.; Chen, K.; Jiang, H.; Zheng, M. Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. *J. Med. Chem.* **2019**, *63*, 8749–8760.

(23) Maziarka, Ł.; Danel, T.; Mucha, S.; Rataj, K.; Tabor, J.; Jastrzębski, S. Molecule attention transformer. *arXiv* **2020**, *2002.08264*.

(24) Jablonka, K. M.; Ai, Q.; Al-Feghali, A.; Badhwar, S.; Bran, J. D.; Bringuier, S.; Brinson, L. C.; Choudhary, K.; Circi, D.; Cox, S., et al. 14 Examples of How LLMs Can Transform Materials Science and Chemistry: A Reflection on a Large Language Model Hackathon. *arXiv* **2023**, *2306.06283*.

(25) Moret, M.; Pachon Angona, I.; Cotos, L.; Yan, S.; Atz, K.; Brunner, C.; Baumgartner, M.; Grisoni, F.; Schneider, G. Leveraging Molecular Structure and Bioactivity with Chemical Language Models for De Novo Drug Design. *Nat. Commun.* **2023**, *14*, 114.

(26) Flam-Shepherd, D.; Aspuru-Guzik, A. Language Models can Generate Molecules, Materials, and Protein Binding Sites Directly in Three Dimensions as XYZ, CIF, and PDB Files. *arXiv* **2023**, *2305.05708*.

(27) White, A. D. The Future of Chemistry is Language. *Nat. Rev. Chem.* **2023**, 1–2.

(28) Ravasco, J. M. J. M.; Coelho, J. A. S. Predictive Multivariate Models for Bioorthogonal Inverse-Electron Demand Diels–Alder Reactions. *J. Am. Chem. Soc.* **2020**, *142*, 4235–4241.

(29) Migliaro, I.; Cundari, T. R. Density Functional Study of Methane Activation by Frustrated Lewis Pairs with Group 13 Trihalides and Group 15 Pentahalides and a Machine Learning Analysis of Their Barrier Heights. *J. Chem. Inf. Model.* **2020**, *60*, 4958–4966.

(30) Guan, Y.; Coley, C. W.; Wu, H.; Ranasinghe, D.; Heid, E.; Struble, T. J.; Pattanaik, L.; Green, W. H.; Jensen, K. F. Regio-Selectivity Prediction with a Machine-Learned Reaction Representation and On-The-Fly Quantum Mechanical Descriptors. *Chem. Sci.* **2021**, *12*, 2198–2208.

(31) Gómez-Flores, C. L.; Maag, D.; Kansari, M.; Vuong, V.-Q.; Irle, S.; Gräter, F.; Kubar, T.; Elstner, M. Accurate Free Energies for Complex Condensed-Phase Reactions Using an Artificial Neural Network Corrected DFTB/MM Methodology. *J. Chem. Theory Comput.* **2022**, *18*, 1213–1226.

(32) Farrar, E. H. E.; Grayson, M. N. Machine Learning and Semi-empirical Calculations: A Synergistic Approach to Rapid, Accurate, and Mechanism-Based Reaction Barrier Prediction. *Chem. Sci.* **2022**, *13*, 7594–7603.

(33) Jorner, K.; Brinck, T.; Norrby, P.-O.; Buttar, D. Machine Learning Meets Mechanistic Modelling for Accurate Prediction of Experimental Activation Energies. *Chem. Sci.* **2021**, *12*, 1163–1175.

(34) Stuyver, T.; Coley, C. W. Quantum Chemistry-Augmented Neural Networks for Reactivity Prediction: Performance, Generalizability, and Explainability. *J. Chem. Phys.* **2022**, *156*, 084104.

(35) Guan, Y.; Lee, T.; Wang, K.; Yu, S.; McWilliams, J. C. S_NAr Regioselectivity Predictions: Machine Learning Triggering DFT Reaction Modeling through Statistical Threshold. *J. Chem. Inf. Model.* **2023**, *63*, 3751–3760.

(36) Riedmiller, K.; Reiser, P.; Bobkova, E.; Maltsev, K.; Gryn'ova, G.; Friederich, P.; Gräter, F. Predicting Reaction Barriers of Hydrogen Atom Transfer in Proteins. *10.26434/chemrxiv-2023-7hntk* **2023**.

(37) Delaney, J. S. ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1000–1005.

(38) Sorkun, M. C.; Khetan, A.; Er, S. AqSolDB, A Curated Reference Set of Aqueous Solubility and 2D Descriptors for a Diverse Set of Compounds. *Sci. Data* **2019**, *6*, 143.

(39) Donchev, A. G.; Taube, A. G.; Decolvenaere, E.; Hargus, C.; McGibbon, R. T.; Law, K.-H.; Gregersen, B. A.; Li, J.-L.; Palmo, K.; Siva, K.; Bergdorf, M.; Klepeis, J. L.; Shaw, D. E. Quantum Chemical Benchmark Databases of Gold-Standard Dimer Interaction Energies. *Sci. data* **2021**, *8*, 55.

(40) Vermeire, F. H.; Green, W. H. Transfer learning for solvation free energies: From quantum chemistry to experiments. *Chem. Eng. J.* **2021**, *418*, 129307.

(41) Axelrod, S.; Gomez-Bombarelli, R. GEOM, Energy-Annotated Molecular Conformations for Property Prediction and Molecular Generation. *Sci. Data* **2022**, *9*, 185.

(42) Zhang, D.; Xia, S.; Zhang, Y. Accurate Prediction of Aqueous Free Solvation Energies Using 3D Atomic Feature-Based Graph Neural Network with Transfer Learning. *J. Chem. Inf. Model.* **2022**, *62*, 1840–1848.

(43) Eastman, P.; Behara, P. K.; Dotson, D. L.; Galvelis, R.; Herr, J. E.; Horton, J. T.; Mao, Y.; Chodera, J. D.; Pritchard, B. P.; Wang, Y.; Fabritiis, D. G.; Markland, T. E. SPICE, A Dataset of Drug-like Molecules and Peptides for Training Machine Learning Potentials. *Sci. Data* **2023**, *10*, 11.

(44) Liu, Y.; Li, Z. Predict Ionization Energy of Molecules Using Conventional and Graph-Based Machine Learning Models. *J. Chem. Inf. Model.* **2023**, *63*, 806–814.

(45) Biswas, S.; Chung, Y.; Ramirez, J.; Wu, H.; Green, W. H. Predicting Critical Properties and Acentric Factors of Fluids Using Multitask Machine Learning. *J. Chem. Inf. Model.* **2023**, DOI: 10.1021/acs.jcim.3c00546.

(46) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757–1768.

(47) Mobley, D. L.; Guthrie, J. P. FreeSolv: A Database of Experimental and Calculated Hydration Free Energies, with Input Files. *J. Comput. Aided Mol. Des.* **2014**, *28*, 711–720.

(48) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42*, D1083–D1090.

(49) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1*, 1–7.

(50) Kim, H.; Park, J. Y.; Choi, S. Energy Refinement and Analysis of Structures in the QM9 Database via a Highly Accurate Quantum Chemical Method. *Sci. Data* **2019**, *6*, 1–8.

(51) Narayanan, B.; Redfern, P. C.; Assary, R. S.; Curtiss, L. A. Accurate Quantum Chemical Energies for 133000 Organic Molecules. *Chem. Sci.* **2019**, *10*, 7449–7455.

(52) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. S. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018**, *9*, 513–530.

(53) Podlewska, S.; Kafel, R. MetStabOn—Online Platform for Metabolic Stability Predictions. *Int. J. Mol. Sci.* **2018**, *19*, 1040.

(54) Von Rudorff, G. F.; Heinen, S. N.; Bragato, M.; von Lilienfeld, O. A. Thousands of Reactants and Transition States for Competing E2 and S2 Reactions. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045026.

(55) Grambow, C. A.; Pattanaik, L.; Green, W. H. Reactants, Products, and Transition States of Elementary Chemical Reactions Based on Quantum Chemistry. *Sci. Data* **2020**, *7*, 1–8.

(56) Zhao, Q.; Vaddadi, S. M.; Woulfe, M.; Ogunfowora, L.; Garimella, S.; Isayev, O.; Savoie, B. M. Comprehensive Exploration of Graphically Defined Reaction Spaces. *10.26434/chemrxiv-2022-1vmwv* **2022**.

(57) Tavakoli, M.; Chiu, Y. T. T.; Baldi, P.; Carlton, A. M.; Van Vranken, D. RMechDB: A Public Database of Elementary Radical Reaction Steps. *J. Chem. Inf. Model.* **2023**, *63*, 1114–1123.

(58) Spiekermann, K. A.; Pattanaik, L.; Green, W. H. High Accuracy Barrier Heights, Enthalpies, and Rate Coefficients for Chemical Reactions. *Sci. Data* **2022**, *9*, 417.

(59) Zimmerman, P. M. Single-Ended Transition State Finding with the Growing String Method. *J. Comput. Chem.* **2015**, *36*, 601–611.

(60) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31–36.

(61) Johnson, M. S.; Dong, X.; Grinberg Dana, A.; Chung, Y.; Farina Jr., D.; Gillis, R. J.; Liu, M.; Yee, N. W.; Blondal, K.; Mazeau, E.; Grambow, C. A.; Payne, A. M.; Spiekermann, K. A.; Pang, H.-W.; Goldsmith, F. C.; West, R. H.; Green, W. H. RMG Database for Chemical Property Prediction. *J. Chem. Inf. Model.* **2022**, *62*, 4906–4915.

(62) Fink, T.; Bruggesser, H.; Reymond, J.-L. Virtual Exploration of the Small-Molecule Chemical Universe below 160 Daltons. *Angewandte Chemie International Edition* **2005**, *44*, 1504–1508.

(63) Fink, T.; Reymond, J.-L. Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discovery. *J. Chem. Inf. Model.* **2007**, *47*, 342–353.

(64) Grambow, C. A. Reactants, Products, and Transition States of Radical Reactions, version 1.0.0, 2020.

(65) Bischoff, F. A.; Wolfsegger, S.; Tew, D. P.; Klopper, W. Assessment of Basis Sets for F12 Explicitly-Correlated Molecular Electronic-Structure Methods. *Mol. Phys.* **2009**, *107*, 963–975.

(66) Knizia, G.; Adler, T. B.; Werner, H.-J. Simplified CCSD(T)-F12 Methods: Theory and Benchmarks. *J. Chem. Phys.* **2009**, *130*, 054104.

(67) Zheng, J.; Zhao, Y.; Truhlar, D. G. The DBH24/08 Database and its Use to Assess Electronic Structure Model Chemistries for Chemical Reaction Barrier Heights. *J. Chem. Theory Comput.* **2009**, *5*, 808–821.

(68) Pfeiffer, F.; Rauhut, G.; Feller, D.; Peterson, K. A. Anharmonic Zero Point Vibrational Energies: Tipping the Scales in Accurate Thermochemistry Calculations? *J. Chem. Phys.* **2013**, *138*, 044311.

(69) Shang, Y.; Ning, H.; Shi, J.; Wang, H.; Luo, S.-N. Chemical Kinetics of H-abstractions from Dimethyl Amine by H, CH_3, OH, and HO_2 Radicals with Multi-Structural Torsional Anharmonicity. *Phys. Chem. Chem. Phys.* **2019**, *21*, 12685–12696.

(70) Landrum, G. et al. RDKit: Open-Source Cheminformatics Software, https://www.rdkit.org, 2006.

(71) Spiekermann, K. A. RMG-DB-11: Enumerating Reaction Space for Small Molecule Chemistry, version 1.0.0, 2023.

(72) Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; Leskovec, J. Strategies for Pre-training Graph Neural Networks. *arXiv* **2019**, *1905.12265*.

(73) Lin, Y.-S.; Li, G.-D.; Mao, S.-P.; Chai, J.-D. Long-range Corrected Hybrid Density Functionals with Improved Dispersion Corrections. *J. Chem. Theory Comput.* **2013**, *9*, 263–272.

(74) Weigend, F.; Ahlrichs, R. Balanced Basis Sets of Split Valence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to Rn: Design and Assessment of Accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

(75) Shao, Y.; Gan, Z.; Epifanovsky, E.; Gilbert, A. T. B.; Wormit, M.; Kussmann, J.; Lange, A. W.; Behn, A.; Deng, J.; Feng, X., et al. Advances in Molecular Quantum Chemistry Contained in the Q-Chem 4 Program Package. *Mol. Phys.* **2015**, *113*, 184–215.

(76) Grambow, C. A.; Pattanaik, L.; Green, W. H. Reactants, Products, and Transition States of Elementary Chemical Reactions based on Quantum Chemistry, version 1.0.1, 2020.

(77) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminf.* **2011**, *3*, 1–14.

(78) Werner, H. J.; Knowles, P. J.; Knizia, G.; Manby, F.; Schütz, M.; Celani, P.; Györffy, W.; Kats, D.; Korona, T.; Lindh, R., et al. MOLPRO, version 2015.1, A Package of Ab initio Programs, https://www.molpro.net, 2015.

(79) Alecu, I. M.; Zheng, J.; Zhao, Y.; Truhlar, D. G. Computational Thermochemistry: Scale Factor Databases and Scale Factors for Vibrational Frequencies Obtained from Electronic Model Chemistries. *J. Chem. Theory Comput.* **2010**, *6*, 2872–2887.

(80) Spiekermann, K. A. RDB7-Rad, version 1.1.0, 2023.

(81) Bacciu, D.; Errica, F.; Micheli, A.; Podda, M. A Gentle Introduction to Deep Learning for Graphs. *Neural Networks* **2020**, *129*, 203–221.

(82) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M., et al. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388.

(83) 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. *10.26434/chemrxiv-2023-3zcfl-v2* **2023**.

(84) Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V. P. Substructural Fragments: An Universal Language to Encode Reactions, Molecular and Supramolecular Structures. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 693–703.

(85) Hoonakker, F.; Lachiche, N.; Varnek, A.; Wagner, A. Condensed Graph of Reaction: Considering a Chemical Reaction as One Single Pseudo Molecule. *Int. J. Artif. Intell. Tools* **2011**, *20*, 253–270.

(86) Madzhidov, T. I.; Polishchuk, P. G.; Nugmanov, R. I.; Bodrov, A. V.; Lin, A. I.; Baskin, I. I.; Varnek, A. A.; Antipin, I. Structure-Reactivity Relationships in Terms of the Condensed Graphs of Reactions. *Russ. J. Org. Chem.* **2014**, *50*, 459–463.

(87) Heid, E.; Green, W. H. Machine Learning of Reaction Properties via Learned Representations of the Condensed Graph of Reaction. *J. Chem. Inf. Model.* **2021**, *62*, 2101–2110.

(88) Spiekermann, K. A.; Pattanaik, L.; Green, W. H.; Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M., et al., https://github.com/kspieks/chemprop/tree/barrier_prediction. Accessed 2022-05-16.

(89) Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations. *arXiv* **2019**, *1909.11942*.

(90) Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M., et al. Huggingface's Transformers: State-of-the-Art Natural Language Processing. *arXiv* **2019**, *1910.03771*.

(91) Schwaller, P.; Hoover, B.; Reymond, J.-L.; Strobelt, H.; Laino, T. Extraction of Organic Chemistry Grammar from Unsupervised Learning of Chemical Reactions. *Sci. Adv.* **2021**, *7*, eabe4166.

(92) Schwaller, P.; Gaudin, T.; Lanyi, D.; Bekas, C.; Laino, T. "Found in Translation"': Predicting Outcomes of Complex Organic Chemistry Reactions using Neural Sequence-to-Sequence Models. *Chem. Sci* **2018**, *9*, 6091–6098.

(93) Spiekermann, K. A.

(94) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019**, *5*, 1572–1583.

(95) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. Data Augmentation Strategies to Improve Reaction Yield Predictions and Estimate Uncertainty. *10.26434/chemrxiv* **2020**, *13286741.v1*.

(96) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(97) Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

(98) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(99) Bender, A.; Schneider, N.; Segler, M.; Patrick Walters, W.; Engkvist, O.; Rodrigues, T. Evaluation Guidelines for Machine Learning Tools in the Chemical Sciences. *Nat. Rev. Chem.* **2022**, *6*, 428–442.

(100) Spiekermann, K. A.; Pattanaik, L.; Green, W. H. Fast Predictions of Reaction Barrier Heights: Toward Coupled-Cluster Accuracy. *J. Phys. Chem. A* **2022**, *126*, 3976–3986.

(101) Burns, J. W.; Spiekermann, K. A.; Bhattacharjee, H.; Vlachos, D. G.; Green, W. H. astartes.

(102) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–113.

(103) Evans, M. G.; Polanyi, M. Inertia and Driving Force of Chemical Reactions. *Trans. Faraday Soc.* **1938**, *34*, 11–24.

(104) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *Advances in Neural Information Processing Systems* **2015**, *28*.

(105) Song, Y.; Zheng, S.; Niu, Z.; Fu, Z.-H.; Lu, Y.; Yang, Y. In *IJCAI*, 2020; Vol. 2020, pp 2831–2838.

(106) Chuang, K. V.; Gunsalus, L. M.; Keiser, M. J. Learning Molecular Representations for Medicinal Chemistry. *J. Med. Chem.* **2020**, *63*, 8705–8722.

(107) Feinberg, E. N.; Joshi, E.; Pande, V. S.; Cheng, A. C. Improvement in ADMET Prediction with Multitask Deep Featurization. *J. Med. Chem.* **2020**, *63*, 8835–8848.

(108) Cáceres, E. L.; Tudor, M.; Cheng, A. C. Deep Learning Approaches in Predicting ADMET Properties. *Future Med. Chem.* **2020**, *12*, 1995–1999.

(109) Struble, T. J.; Coley, C. W.; Jensen, K. F. Multitask Prediction of Site Selectivity in Aromatic C–H Functionalization Reactions. *React. Chem. Eng.* **2020**, *5*, 896–902.

(110) Tang, B.; Kramer, S. T.; Fang, M.; Qiu, Y.; Wu, Z.; Xu, D. A Self-Attention Based Message Passing Neural Network for Predicting Molecular Lipophilicity and Aqueous Solubility. *J. Cheminform.* **2020**, *12*, 1–9.

(111) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; Von Lilienfeld, O. A. Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *J. Chem. Theory Comput.* **2017**, *13*, 5255–5264.

(112) Wieder, O.; Kuenemann, M.; Wieder, M.; Seidel, T.; Meyer, C.; Bryant, S. D.; Langer, T. Improved Lipophilicity and Aqueous Solubility Prediction with Composite Graph Neural Networks. *Molecules* **2021**, *26*, 6185.

(113) Li, C.; Wang, J.; Niu, Z.; Yao, J.; Zeng, X. A Spatial-Temporal Gated Attention Module for Molecular Property Prediction based on Molecular Geometry. *Brief. Bioinform.* **2021**, *22*, bbab078.

(114) Magar, R.; Wang, Y.; Lorsung, C.; Liang, C.; Ramasubramanian, H.; Li, P.; Farimani, A. B. AugLiChem: Data Augmentation Library of Chemical Structures for Machine Learning. *Mach. Learn.: Sci. Technol.* **2022**, *3*, 045015.

(115) Greenman, K. P.; Green, W. H.; Gomez-Bombarelli, R. Multi-Fidelity Prediction of Molecular Optical Peaks with Deep Learning. *Chem. Sci.* **2022**, *13*, 1152–1162.

(116) Wang, Y.; Wang, J.; Cao, Z.; Barati Farimani, A. Molecular Contrastive Learning of Representations via Graph Neural Networks. *Nat. Mach. Intell.* **2022**, *4*, 279–287.

(117) Aldeghi, M.; Coley, C. W. A Graph Representation of Molecular Ensembles for Polymer Property Prediction. *Chem. Sci.* **2022**, *13*, 10486–10498.

(118) Kang, B.; Seok, C.; Lee, J. A Benchmark Study of Machine Learning Methods for Molecular Electronic Transition: Tree-based Ensemble Learning versus Graph Neural Network. *Bull. Korean Chem. Soc.* **2022**, *43*, 328–335.

(119) Han, X.; Jia, M.; Chang, Y.; Li, Y.; Wu, S. Directed Message Passing Neural Network (D-MPNN) with Graph Edge Attention (GEA) for Property Prediction of Biofuel-Relevant Species. *Energy and AI* **2022**, *10*, 100201.

(120) Stuyver, T.; Coley, C. W. Machine Learning-Guided Computational Screening of New Candidate Reactions with High Bioorthogonal Click Potential. *Chemistry–A European Journal* **2023**, e202300387.

(121) Li, J.; Cai, D.; He, X. Learning Graph-Level Representation for Drug Discovery. *arXiv* **2017**, *1709.03741*.

(122) Guo, J.; Sun, M.; Zhao, X.; Shi, C.; Su, H.; Guo, Y.; Pu, X. General Graph Neural Network-Based Model To Accurately Predict Cocrystal Density and Insight from Data Quality and Feature Representation. *J. Chem. Inf. Model.* **2023**, *63*, 1143–1156.

(123) Isert, C.; Kromann, J. C.; Stiefl, N.; Schneider, G.; Lewis, R. A. Machine Learning for Fast, Quantum Mechanics-Based Approximation of Drug Lipophilicity. *ACS Omega* **2023**, *8*, 2046–2056.

(124) Duan, Y.-J.; Fu, L.; Zhang, X.-C.; Long, T.-Z.; He, Y.-H.; Liu, Z.-Q.; Lu, A.-P.; Deng, Y.-F.; Hsieh, C.-Y.; Hou, T.-J.; Cao, D.-S. Improved GNNs for Log $D_{7.4}$ Prediction by Transferring Knowledge from Low-Fidelity Data. *J. Chem. Inf. Model.* **2023**, DOI: https://doi.org/10.1021/acs.jcim.2c01564.

(125) Liu, C.; Sun, Y.; Davis, R.; Cardona, S. T.; Hu, P. ABT-MPNN: An Atom-Bond Transformer-Based Message-Passing Neural Network for Molecular Property Prediction. *J. Cheminformatics* **2023**, *15*, 29.

(126) Ren, G.-P.; Wu, K.-J.; He, Y. Enhancing Molecular Representations Via Graph Transformation Layers. *J. Chem. Inf. Model.* **2023**, *63*, 2679–2688.

(127) AbdulHameed, M. D. M.; Liu, R.; Wallqvist, A. Using a Graph Convolutional Neural Network Model to Identify Bile Salt Export Pump Inhibitors. *ACS Omega* **2023**, *8*, 21853–21861.

(128) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. In *International Conference on Machine Learning*, 2017, pp 1263–1272.

(129) Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; Jensen, K. F. Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction. *J. Chem. Inf. Model.* **2017**, *57*, 1757–1772.

(130) Hop, P.; Allgood, B.; Yu, J. Geometric Deep Learning Autonomously Learns Chemical Features That Outperform Those Engineered by Domain Experts. *Mol. Pharmaceutics* **2018**, *15*, 4371–4377.

(131) Feinberg, E. N.; Sur, D.; Wu, Z.; Husic, B. E.; Mai, H.; Li, Y.; Sun, S.; Yang, J.; Ramsundar, B.; Pande, V. S. PotentialNet for Molecular Property Prediction. *ACS Cent. Sci.* **2018**, *4*, 1520–1530.

(132) Montanari, F.; Kuhnke, L.; Ter Laak, A.; Clevert, D.-A. Modeling Physico-Chemical ADMET Endpoints with Multitask Graph Convolutional Networks. *Molecules* **2019**, *25*, 44.

(133) Honda, S.; Shi, S.; Ueda, H. R. Smiles Transformer: Pre-trained Molecular Fingerprint for Low Data Drug Discovery. *arXiv* **2019**, *1911.04738.*

(134) Shin, B.; Park, S.; Kang, K.; Ho, J. C. In *Mach. Learn. Healthcare Conference*, 2019, pp 230–248.

(135) Fabian, B.; Edlich, T.; Gaspar, H.; Segler, M.; Meyers, J.; Fiscato, M.; Ahmed, M. Molecular Representation Learning with Language Models and Domain-Relevant Auxiliary Tasks. *arXiv* **2020**, *2011.13230.*

(136) Li, J.; Jiang, X. Mol-BERT: An Effective Molecular Representation with BERT for Molecular Property Prediction. *Wirel. Commun. Mob. Comput.* **2021**, *2021*, 1–7.

(137) Chithrananda, S.; Grand, G.; Ramsundar, B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. *arXiv* **2020**, *2010.09885.*

(138) Ahmad, W.; Simon, E.; Chithrananda, S.; Grand, G.; Ramsundar, B. Chemberta-2: Towards Chemical Foundation Models. *arXiv* **2022**, *2209.01712.*

(139) Winter, B.; Winter, C.; Schilling, J.; Bardow, A. A Smile is All You Need: Predicting Limiting Activity Coefficients from SMILES with Natural Language Processing. *Digital Discovery* **2022**, *1*, 859–869.

(140) Kang, H.; Goo, S.; Lee, H.; Chae, J.-w.; Yun, H.-y.; Jung, S. Fine-Tuning of BERT Model to Accurately Predict Drug–Target Interactions. *Pharmaceutics* **2022**, *14*, 1710.

(141) Henkelman, G.; Uberuaga, B. P.; Jónsson, H. A Climbing Image Nudged Elastic Band Method for Finding Saddle Points and Minimum Energy Paths. *J. Chem. Phys.* **2000**, *113*, 9901–9904.

(142) Zimmerman, P. M. Automated Discovery of Chemically Reasonable Elementary Reaction Steps. *J. Comput. Chem.* **2013**, *34*, 1385–1392.

(143) Harms, N.; Underkoffler, C.; West, R. H. Advances in Automated Transition State Theory Calculations: Improvements on the AutoTST Framework. *10.26434/chemrxiv* **2020**, *13277870.v2*, Accessed 2022-03-08.

(144) Pattanaik, L.; Ingraham, J. B.; Grambow, C. A.; Green, W. H. Generating Transition States of Isomerization Reactions with Deep Learning. *Phys. Chem. Chem. Phys.* **2020**, *22*, 23618–23626.

(145) Jackson, R.; Zhang, W.; Pearson, J. TSNet: Predicting Transition State Structures with Tensor Field Networks and Transfer Learning. *Chem. Sci.* **2021**, *12*, 10022–10040.

(146) Makoś, M. Z.; Verma, N.; Larson, E. C.; Freindorf, M.; Kraka, E. Generative Adversarial Networks for Transition State Geometry Prediction. *J. Chem. Phys.* **2021**, *155*, 024116.

(147) Duan, C.; Du, Y.; Jia, H.; Kulik, H. J. Accurate Transition State Generation with an Object-Aware Equivariant Elementary Reaction Diffusion Model. *arXiv* **2023**, *2304.06174*.

(148) Dana, A. G.; Johnson, M. S.; Allen, J. W.; Sharma, S.; Raman, S.; Liu, M.; Gao, C. W.; Grambow, C. A.; Goldman, M. J.; Ranasinghe, D. S.; Gillis, R. J.; Payne, A. M.; Li, Y.-P.; Dong, X.; Spiekermann, K. A.; Wu, H.; Dames, E. E.; Buras, Z. J.; Vandewiele, N. M.; Yee, N. W.; Merchant, S. S.; Buesser, B.; Class, C. A.; Goldsmith, F.; West, R. H.; Green, W. H. Automated Reaction Kinetics and Network Exploration (Arkane): A Statistical Mechanics, Thermodynamics, Transition State Theory, and Master Equation Software. *Int. J. Chem. Kinet.* **2022**, DOI: 10.1002/kin.21637.

(149) Cavallotti, C.; Pelucchi, M.; Georgievskii, Y.; Klippenstein, S. J. EStokTP: Electronic Structure to Temperature–And Pressure-Dependent Rate Constants–A Code for Automatically Predicting the Thermal Kinetics of Reactions. *J. Chem. Theory Comput.* **2018**, *15*, 1122–1145.

(150) Ucak, U. V.; Ashyrmamatov, I.; Lee, J. Improving the Quality of Chemical Language Model Outcomes with Atom-in-SMILES Tokenization. *J. Cheminformatics* **2023**, *15*, 55.

(151) Mao, J.; Wang, J.; Cho, K.-H.; No, K. T. iupacGPT: IUPAC-based Large-Scale Molecular Pre-trained Model for Property Prediction and Molecule Generation. *10.26434/chemrxiv-2023-5kjvh* **2023**.

(152) Meredig, B.; Antono, E.; Church, C.; Hutchinson, M.; Ling, J.; Paradiso, S.; Blaiszik, B.; Foster, I.; Gibbons, B.; Hattrick-Simpers, J.; Mehta, A.; Ward, L. Can Machine Learning Identify the Next High-Temperature Superconductor? Examining Extrapolation Performance for Materials Discovery. *Mol. Syst. Des. Eng.* **2018**, *3*, 819–825.

(153) Nakajima, M.; Nemoto, T. Machine Learning Enabling Prediction of the Bond Dissociation Enthalpy of Hypervalent Iodine from SMILES. *Sci. Rep.* **2021**, *11*, 20207.

(154) Espley, S. G.; Farrar, E. H. E.; Buttar, D.; Tomasi, S.; Grayson, M. N. Machine Learning Reaction Barriers in Low Data Regimes: A Horizontal and Diagonal Transfer Learning Approach. *Digital Discovery* **2023**, DOI: 10.1039/D3DD00085K.

(155) Ng, N. H.; Hulkund, N.; Cho, K.; Ghassemi, M. Predicting Out-of-Domain Generalization with Neighborhood Invariance. *Transactions on Machine Learning Research* **2023**.

Chapter 8

Best Practices in Machine Learning for Chemistry

This work has been submitted as Spiekermann, K. A.; Stuyver, T.; Pattanaik, L.; Green, W. H. Comment on "Physics-based Representations for Machine Learning Properties of Chemical Reactions". *Mach. Learn.: Sci. Technol.* 2023. Thijs Stuyver and Lagnajit Pattanaik helped with the analysis.

8.1 Introduction

Mechanistic kinetic models are invaluable tools to acquire understanding in—and potentially engineer—chemical processes, with applications ranging from advanced pharmaceutical design, high-performance material development, and renewable energy research [1]. Such kinetic models commonly require accurate thermodynamic and kinetic parameters for hundreds of species and several thousands of reactions. Unfortunately, there are relatively few experimental data, and the traditional computational workflow to obtain kinetic parameters—namely barrier heights, A-factors, and rate coefficients—is complex and expensive [2, 3]. The grand vision of predictive kinetics is to quickly and accurately estimate these parameters. As such, much effort has been devoted to developing models to directly estimate kinetic parameters, rather than computing them explicitly with computational chemistry methods. Several methods for estimating rate coefficients or barriers have been developed [4]. In recent years, large datasets of reaction barriers have become available [5–9], and several research groups have used these to train machine learning models. A summary of recent advancements on this front is provided in our feature article [10], and others have similarly commented on the state of affairs of applying machine learning (ML) to chemical kinetics

[11–14].

Recently, van Gerwen et al. [15] presented a new ML framework using kernel ridge regression (KRR) models to estimate reaction properties. In their work, the authors introduced a set of reaction representations based on quantum machine learning (QML) in which reactant and product representations are subtracted, similar to the well-documented approach of representing a reaction as a difference fingerprint for both graph neural network (GNN) architectures [16, 17] and classical ML techniques [18]. Other highlights include the release of a new dataset focused on alkene insertion in the catalytic cycle of olefin hydroformylation and the introduction of a new bond-based reaction representation, B^2R^2.

Despite the novelty of these contributions, the presented work also raises some important and fundamental questions which are relevant to the field as a whole. Specifically, we strongly believe that applicability should be front and center when developing and publishing a new ML model. We provide a point-by-point overview of our main concerns below.

With this specific Comment and broader perspective, we encourage the community to reflect on the use-cases of their models by asking pertinent and fundamental questions such as “Which problem are we trying to solve with this model?” and “How could our models be usefully employed?” before presenting a new ML approach.

8.1.1 Providing Context for Model Performance

To gauge whether a model will actually be useful to the kinetics community, it is instructive to place the model’s errors into context. For example, some recent papers have used ML models to quickly predict energy values for individual molecules [19, 20] or reactions [10] that are more accurate than those obtained through a density functional theory (DFT) calculation, which has the potential to offer considerable time savings. The test MAE reported by van Gerwen and co-workers for the DFT dataset from ref. [5] however amounts to about 10 kcal mol^{-1}. While a model with such errors may be useful for very rough qualitative analysis, it most likely won’t be useful for quantitative analysis, since an uncertainty on the activation energy of this magnitude implies that rate coefficients would differ on average by a factor of 150 at 1000 K and by a factor of 19 million at 300 K.

Another way to place model performance into context is to compare to some baselines. The most trivial baseline model is one that simply predicts the mean value. For the dataset from Grambow et al. [5], always predicting the mean barrier height results in an MAE of 18 kcal mol^{-1}. Remarkably, none of the bag-of-bonds (BOB) based reaction representations outperform this trivial baseline by more than a few kcal mol^{-1}, and only when SLATM/SOAP is used in combination with the X_d reaction representation does the accuracy improve. For the Proparg-21-TS dataset [21], the X_r representation actually does worse than always predicting the mean of 3 kcal mol^{-1}, and the X_p representation does only marginally better. For the Hydroform-22-TS dataset, predicting the barrier heights based on the means for each of the three catalyst groups (Co, Rh, and Ir) achieves an MAE of 1.6 kcal mol^{-1} for the entire dataset. Using the best reaction representation (X_d), the various models presented by van Gerwen et al. achieve a test MAE of 0.8 to 1.1 kcal mol^{-1} on this dataset, which offers only a modest improvement of 0.5-0.8 kcal mol^{-1} over the trivial baseline model. Thus, comparing against such a simple baseline provides helpful context [22], particularly when dealing with relatively homogeneous datasets (e.g., Proparg-21-TS and Hydroform-22-TS) whose narrow ranges inherently lead to low test MAEs that might be mistaken for satisfactory performance.

A final way to contextualize prediction errors is to compare performance to published papers that train models on the same datasets. For the Grambow et al. [5] dataset in particular, test MAEs 2-3x lower than what is reported by van Gerwen et al. have been obtained in recent years [10, 16, 17]. Similarly, in a related task of predicting barrier heights for S$_\mathrm{N}$2 and E2 reactions, Stuyver et al. [23] show that a QM augmented GNN extrapolates better to unseen nucleophiles than KRR with the SLATM kernel.

At several instances throughout their paper, van Gerwen et al. suggest that KRR is superior to GNN for problems involving chemical structures. But this assertion has already been tested in the literature and found to be false, not just for barrier heights [10, 17, 23], but also for other molecular properties. For example, both Yang et al. [24] and Faber et al. [25] compared several representations and modeling approaches for property prediction on the famous QM9 dataset [26] and found that graph convolutions often outperformed the alternatives. Feinberg et al. [27] similarly found that graph networks outperformed random forest when predicting various drug properties, especially on more challenging splits, for the vast majority of the 30 datasets analyzed.

Despite the demonstrated superior performance of GNNs for predicting many molecular properties, including reaction barriers, our argument is not that researchers should exclusively use deep learning. For example, sometimes classical ML techniques as simple as decision trees or random forest can be quite effective at estimating kinetic parameters [28] or ranking reaction conformations [29], and we do not doubt that there are some molecular problems and/or datasets where KRR or other architectures may have some advantages over GNN-based approaches. Our argument instead is that comparing model performance to relevant baselines—and to experimental data when available [30–42]—should never be "beyond the scope" of work for a model-developer. Rather, it is an essential first step at convincing readers that the model is useful.

8.1.2 Reconsidering Model Inputs

Given the myriad of ways to represent a molecule [43–45], a second crucial consideration is determining whether the model's inputs are user-friendly and economical. The model architecture presented by van Gerwen et al. requires DFT-optimized geometries as input and infers DFT quality (electronic) activation or reaction energies as output.

When reaction energies are selected as the target, as in the first application presented with the SN2-20 dataset, this issue is particularly clear and pressing. To put it bluntly, the presented ML workflow becomes an exercise in futility in this case. The aim of this model is to predict energy differences between optimized structures of reactants and products, but as one generates the input for the model, the model's output (i.e., electronic energies) are automatically computed at no additional cost, which begs the question of why an ML model would be needed in the first place? In all fairness, van Gerwen et al. briefly acknowledge that low testing errors are expected for this specific task since the output is already contained in the input. At the same time, one can think of many opportunities to make this application more sensible, for example by including energy corrections in the predicted targets through the computation of zero-point energies and partition functions with statistical mechanics. Applying these energy corrections is necessary to compute practically useful reaction enthalpies. Or, as discussed below, one could build a model based on some simpler or cheaper input than DFT calculations.

Of course, the situation is different when activation energies are selected as the target, which is the case for applications 2-4 in their paper. In this setup, the developed model makes it possible to circumvent the need to find a TS, which is often very challenging and resource intensive [10, 46–48]. However, when cheaper and equally-accurate (or even more accurate) alternatives are available (*vide supra/infra*), the appeal of this specific KRR model is arguably limited.

It should be noted that the van Gerwen et al. manuscript leaves some questions related to this issue unaddressed, which could potentially shine a different light on the appeal and user-friendliness of their presented models. More specifically, if low-quality geometries or conformers are sufficient to get accurate predictions, and if the attained accuracy would be a chemically meaningful improvement over cheaper baselines and alternative approaches, then a clearer motivation to prefer these KRR-based models could be envisioned. However, the authors do not provide evidence for the former point, and our own analysis (*vide infra*) suggests that—at least for one of the datasets considered—the latter point is not fulfilled. Finally, if optimized geometries do prove beneficial (or are necessary for certain tasks such as quickly ranking conformer energies), comparison to state-of-the-art models such as SchNet [49], DimeNet++ [50], PaiNN [51], EGNN [52], E3NN [53], etc would be informative.

Overall, we argue that in order to convince the kinetics community that a geometry-based ML model is robust and useful, a detailed analysis regarding the sensitivity of the results toward the input geometry, as well as an analysis of the sensitivity toward conformer identity, is needed. Several precedents exist for this kind of analysis. For example, other groups have analyzed performance on geometries from semi-empirical methods or even force fields [54], which are much more practical for users to obtain than *ab initio* geometries. Other studies have used a ΔML technique [55], such as using semi-empirical geometries to predict DFT quality energies [56–58] or barrier heights [59, 60], bringing a potential energy surface from B3LYP up to CCSD(T) [61], or predicting the effect of perturbatively included triples [62]. Regardless, we want to conclude this section by stressing that any model operating on 3D coordinates will always require an extra step compared to models operating on simpler representations. Given that modern kinetic mechanisms must often consider hundreds of thousands of possible reactions [63], it is imperative to have estimators that can quickly make predictions at scale. We encourage the community to consider, and verify on a case-by-case basis, whether the compute time spent optimizing geometries, and likely also performing conformer searches, actually offers meaningful improvements to the model's predictions.

8.1.3 Interpolation vs. Extrapolation

Another important point regarding model applicability is whether we expect the model to be used for interpolation or extrapolation. Since models can sometimes learn irrelevant patterns in the data when training on interpolation (i.e., random) splits that hinder their generalizability [64–67], end users should always understand the limitations of such models and apply them within appropriate bounds. However, given the vastness of chemical space [68, 69] and its often unsmooth nature (e.g., activity cliffs), it seems unlikely that users will want to be restricted to exclusively operate in an interpolation regime.

We see a few different ways to test extrapolation capabilities. If we are interested in assessing model performance on new molecules, we can train a model with many reaction templates but use substructure splitting to create training, validation, and testing sets. Bemis-Murcko scaffolds [70] are commonly used to partition the data for this purpose, though clustering based on other input features or chemical similarity to measure extrapolation has also been explored [23, 71–88] as has quantifying domains of model applicability [89–93]. Scaffold splitting is not perfect, but by ensuring that molecules in the testing set are structurally different than those in the training set, it offers a much better assessment of generalizability than splitting randomly [17, 24, 67, 94–110]. This is also evident when examining the learning curves for the different splitting approaches, as presented in Figure 8.1; the test error is higher for scaffold splitting because it is a more challenging task. As a concrete example of extrapolation error for the popular S_N2 and E2 dataset that is also analyzed by van Gerwen et al., Stuyver et al. [23] show that both regular GNNs and KRR do not generalize well to different nucleophiles, even when the reaction family and substituted ethyl-based scaffold are held constant. As another example, substantially worse performance on out-of-sample test sets compared to random splitting has been observed for reaction yield prediction tasks as well [111, 112]. Another option is to train on many molecular scaffolds but split by reaction type to measure performance on new reaction families [113, 114]. Yet another example comes from Chen et al. [115] who trained a graph network on molecules with up to 11 heavy atoms and then obtained a testing MAE of 2 kcal mol^{-1} relative to experimental data on molecules with up to 42 heavy atoms. Still other papers show that time-based splits offer a better measure of generalization than random splits [27, 116–118], though it is likely more rigorous to directly split on chemical information rather than a proxy such as a time which may still leak substructures and/or reaction templates. If working with solvated systems, some papers claim that solvent-based splits–in which the model sees molecules or reactions in some solvents during training and then performance is tested with the same molecules or reactions in new solvents–can measure extrapolation [119]. However, given that $\Delta G_{solvation}$ values often have relatively small magnitudes [120–122], solvent-based splits are likely not as challenging as they would appear, and it is likely that a trivial baseline model would perform only slightly worse; in contrast, solute-based splits represent a harder task [123–125].

Nevertheless, many papers use random splits [64, 126–147], which ultimately measures an interpolation error. For example, when doing a random split of 85:5:10 to create training, validation, and testing sets for the Grambow et al. [16] dataset, all scaffolds end up in the training set and 78% of all scaffolds end up in the test set. Given the similarity in composition, good performance on random splits is unsurprising.

For their Hydroform-22-TS dataset, the authors go one step further by using furthest point sampling (FPS) when creating splits. This greedy algorithm places the most diverse points in the training set, which guarantees all testing points will be contained within the span of the training set. FPS is commonly used within the KRR community [148–154]. However, given that the field has established that random splitting overestimates the extrapolation capabilities, testing errors from the greedy FPS splits are likely to also be unrealistically low and will not reflect the true model performance when evaluating a new sample. Additionally, due to its dependence on the kernel space, the split cannot be expanded straightforwardly to e.g., graph neural networks. As such, we encourage future work on KRR models to also measure performance against more challenging splits so that readers can better understand the scope of applicability and make fair comparisons with alternative model architectures.

8.2 Methods

8.2.1 Dataset

We analyze the performance of predicting barrier heights from the DFT dataset created by Grambow et al. [5], which has been used in several other modeling papers [10, 16, 17, 155–158]. van Gerwen et al. refer to this dataset as GDB7-20-TS. Of the four datasets studied by van Gerwen et al., GDB7-20-TS is the most diverse, spanning the largest range of barrier heights and covering many different reaction templates; thus, this dataset arguably offers the most challenging benchmark for barrier predictions. The other datasets include some larger species, but they tend to focus exclusively on specific reaction classes. For these reasons, and given our personal interest in developing broad purpose models capable of distinguishing high-barrier reactions from more plausible low-barrier reactions, we decided to focus our analysis on the GDB7-20-TS dataset.

When creating training, validation, and testing sets, we use five folds, each with an 85:5:10 split. Our results report the mean and standard deviation of each fold to give some sense of model uncertainty; future work could more rigorously examine uncertainty estimation [159–168], though it's sometimes unclear which method is best [169]. We use both random splitting and scaffold splitting on the reactant SMILES, which partitions the data based on Bemis-Murcko scaffold splits [70], as calculated by RDKit [170]. To make this a fair comparison, we use the exact same data splits for both model types described below. Once the training, validation, and testing sets are established, we randomly sample from the training set to create smaller training sets, which allows us to examine how model performance depends on the size of the training dataset. Thus, in total we have five training set sizes of 1016, 1807, 3214, 5716, and 10165 reactions.

8.2.2 Models

We analyze the introduced reaction representation SLATM$_d^{(2)}$ kernel from van Gerwen et al. [15] All code related to our analysis is freely available on GitHub [171]. To facilitate comparison, we use the KRR implementation and corresponding hyperparameters provided by van Gerwen et al. However, we want to briefly comment on their procedure used to determine these hyperparameters. The authors state that, "the best hyperparameters are those with a minimal MAE on the test set across the five folds." Best practices in data science however dictate that, in order to minimize the risk of hyperparameter overfitting, one ought to only optimize hyperparameters with a validation set and use a held-out test set to accurately measure performance on unseen data [107, 109, 172–180] i.e., if doing five-fold cross validation, it is essential to first create five separate training, validation, and testing splits. It is unclear to what extent a more rigorous hyperparameter optimization procedure would have affected the results for their specific experiments, but one can reasonably expect that, if anything, the approach taken will overstate the actual accuracy of the KRR model on unseen data.

We compare the KRR performance to models built by Chemprop [24] – a type of GNN – which uses a directed message passing neural network (DMPNN) to create a learned fingerprint, followed by feed-forward neural network. We use the established condensed graph of reaction (CGR) representation as the input when predicting barrier heights [17, 181–183]. Since we use the exact same forked version and hyperparameters as were used in ref. [10], we refer readers to our previous work for additional details.

8.3 Results & Discussion

Both Table 8.1 and Table 8.2 show that the testing errors obtained from our DMPNN is lower than that from KRR when using the exact same splits across all training set sizes for the Grambow et al. [5] dataset. Figure 8.1 plots the test MAEs for easy visual inspection. These results explicitly and unequivocally refute the claim from van Gerwen et al. [15] that "representations derived from graphs in deep learning models...tend to perform well only for large dataset sizes". These results are corroborated by several papers that have found the learned representations from graphs often outperform other approaches [10, 17, 24, 25, 27, 85, 97, 99, 100, 104–106, 110, 118, 140, 184–206]. While the DMPNN performs better than

KRR for all dataset sizes tested, Figure 8.1 shows that the DMPNN takes better advantage of additional data, with mean absolute errors over a factor of two lower than KRR for cases with many training data points. The errors from scaffold splitting are higher than those from splitting randomly, especially in the low-data regimes, which is an expected result since scaffold splitting is a more challenging extrapolative task as discussed in Section 8.1.3.

Table 8.1. Testing errors in kcal mol^{-1} (mean ± 1 standard deviation from the five folds) for predicting barrier heights from the DFT dataset from Grambow et al. [5] when using random splitting. The exact same data splits are used between the KRR with SLATM$_d^{(2)}$ representation and DMPNN with CGR representation.

	SLATM$_d^{(2)}$		**DMPNN**	
Training Points	MAE	RMSE	MAE	RMSE
1016	15.47 ± 0.42	22.66 ± 0.85	9.31 ± 0.42	13.65 ± 0.64
1807	13.49 ± 0.63	20.40 ± 1.29	7.59 ± 0.28	11.50 ± 0.65
3214	12.04 ± 0.42	17.71 ± 0.66	6.14 ± 0.17	9.79 ± 0.42
5716	10.92 ± 0.35	16.34 ± 0.73	5.06 ± 0.09	8.44 ± 0.34
10165	10.01 ± 0.16	14.90 ± 0.36	4.11 ± 0.07	7.15 ± 0.25

Table 8.2. Testing errors in kcal mol^{-1} (mean ± 1 standard deviation from the five folds) for predicting barrier heights from the DFT dataset from Grambow et al. [5] when using scaffold splitting. The exact same data splits are used between the KRR with SLATM$_d^{(2)}$ representation and DMPNN with CGR representation.

	SLATM$_d^{(2)}$		**DMPNN**	
Training Points	MAE	RMSE	MAE	RMSE
1016	25.06 ± 3.60	35.26 ± 4.54	14.34 ± 1.18	19.04 ± 1.24
1807	22.71 ± 1.03	31.35 ± 1.56	13.48 ± 0.35	18.28 ± 0.58
3214	20.83 ± 1.29	30.68 ± 4.06	13.27 ± 1.28	18.05 ± 1.86
5716	19.09 ± 1.18	28.68 ± 3.80	12.97 ± 0.68	17.76 ± 0.94
10165	13.32 ± 0.73	19.08 ± 0.94	6.56 ± 0.30	9.99 ± 0.59

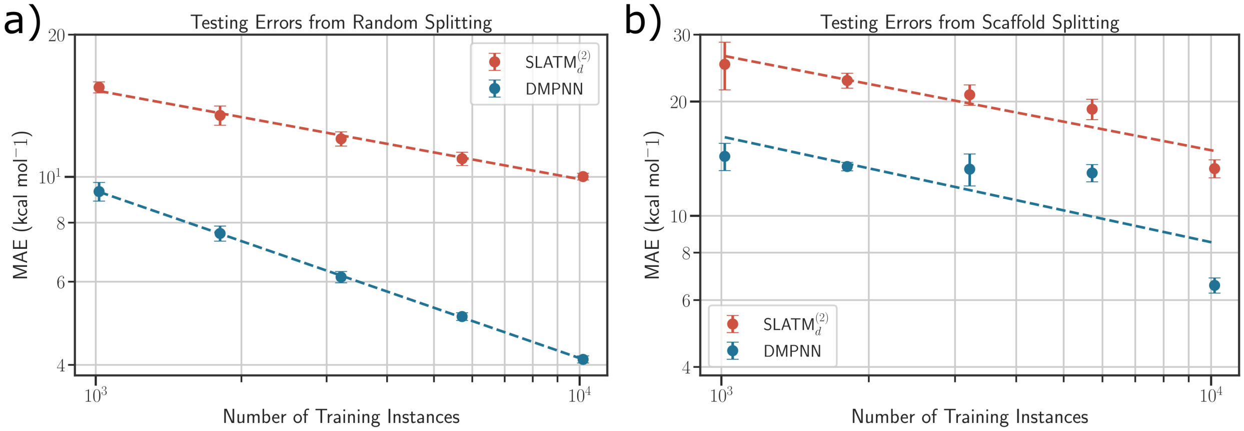


Figure 8.1. Testing errors for predicting barrier heights from the DFT dataset from Grambow et al. [5] when using either a) random splitting or b) scaffold splitting. Dashed lines are the best fit line to the plotted points. Error bars indicate one standard deviation calculated across the five folds. The exact same data splits are used between the KRR with SLATM$_d^{(2)}$ representation and DMPNN with CGR representation.

Several approaches could be taken to further reduce the DMPNN errors presented here. First, if suitable pretraining data is available, transfer learning is an established technique to improve model performance [123, 207–213]. Second, we do not use ensembling here; however, this is another established method to improve model predictions [24, 214]. Finally, one could augment the dataset with the reverse reactions, which doubles the dataset size. This is crucial because the learning curve shown here does not plateau (i.e., all models shown are operating in a data-limited regime and should yield better testing performance if trained on additional reactions). Usually reaction training data is scarce, so it is common practice to use both the forward and reverse barriers to increase the number of training data [10, 17]. However, van Gerwen et al. only used the forward reactions during training, so here we present results consistent with their study.

More broadly, the findings presented here fit into a growing body of evidence that the usefulness of GNNs is not necessarily restricted to very big datasets. For example, Stuyver et al. [23] demonstrated in a recent contribution that QM-augmented GNN models already reach similar accuracy as regular KRR models in barrier prediction for E2 and S_N2 reactions with training set sizes as small as 100 data points. Non-QM-augmented DMPNNs were shown to become competitive with KRR for training sets as small as ∼500 data points.

We are not aware of any reactivity dataset with more than ∼500 reactions where KRR models have been demonstrated to outperform GNNs. It is definitely possible that application areas will emerge in the near future where KRR models do show improved accuracy. In the absence of evidence however, caution should be exerted when discussing the performance of one model architecture relative to the other. Ideally, any statement of this type should be corroborated with unambiguous performance comparisons.

As an aside, we recommend that future work use the updated version of this dataset, which is publicly available on Zenodo [215]. We also recommend people refer to this dataset as RDB7 i.e., a diverse reaction database whose transition states contain up to seven heavy atoms. This newer work substantially improves the quality of the tabulated barrier heights and enthalpies for these reactions by performing highly accurate single-point energy refinements at CCSD(T)-F12a/cc-pVDZ-F12 level of theory and accounting for zero-point-energies along with the corresponding frequency scaling factor [216]. The updated reaction enthalpies include atom energy corrections [217], which are important for individual heats of formation but actually cancel when calculating the reaction enthalpy since all reactions are balanced, as well as bond additivity corrections [218, 219] from Arkane [220], which are essential to calculating a true thermodynamic energy. This work also cleans the reaction SMILES and re-optimizes the product complexes. Additional details can be found in ref. [6].

8.4 Conclusion

Machine learning for chemistry is a rapidly growing field that offers great potential to accelerate steps in a broader research workflow. For example, the ability to quickly predict values of interest could allow researchers to screen thousands of candidate molecules or reactions faster than they could with traditional quantum chemistry approaches. However, we encourage the community to critically think about which problem their model is trying to solve. Model performance should always be placed in context, and comparisons to other architectures should be supported by data. To facilitate faster adoption of these models, it is important to operate on user-friendly inputs or offer evidence that justifies more expensive chemical representations. Models which do not require the user to provide 3D geometries as input have many advantages, including avoiding the need to identify the lowest-energy conformer, which can be challenging for large flexible molecules. Finally, reporting errors from both interpolative and extrapolative data splits gives readers a better understanding of model applicability and likely performance in extrapolative prediction tasks. We welcome the community to contribute additional thoughts and are excited about the future that machine learning techniques can enable.

8.5 References

(1) Green, W. H. Moving from Postdictive to Predictive Kinetics in Reaction Engineering. *AIChE Journal* **2020**, *66*, e17059.

(2) Truhlar, D. G.; Garrett, B. C.; Klippenstein, S. J. Current Status of Transition-State Theory. *J. Phys. Chem.* **1996**, *100*, 12771–12800.

(3) Klippenstein, S. J. From Theoretical Reaction Dynamics to Chemical Modeling of Combustion. *Proc. Combust. Inst.* **2017**, *36*, 77–111.

(4) Wang, K.; Dean, A. M., Rate Rules and Reaction Classes In *Comput. Aided Chem. Eng.* Elsevier: 2019; Vol. 45, pp 203–257.

(5) Grambow, C. A.; Pattanaik, L.; Green, W. H. Reactants, Products, and Transition States of Elementary Chemical Reactions Based on Quantum Chemistry. *Sci. Data* **2020**, *7*, 1–8.

(6) Spiekermann, K. A.; Pattanaik, L.; Green, W. H. High Accuracy Barrier Heights, Enthalpies, and Rate Coefficients for Chemical Reactions. *Sci. Data* **2022**, *9*, 417.

(7) Von Rudorff, G. F.; Heinen, S. N.; Bragato, M.; von Lilienfeld, O. A. Thousands of Reactants and Transition States for Competing E2 and S2 Reactions. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045026.

(8) Zhao, Q.; Vaddadi, S. M.; Woulfe, M.; Ogunfowora, L.; Garimella, S.; Isayev, O.; Savoie, B. M. Comprehensive Exploration of Graphically Defined Reaction Spaces. *10.26434/chemrxiv-2022-1vmwv* **2022**.

(9) Tavakoli, M.; Chiu, Y. T. T.; Baldi, P.; Carlton, A. M.; Van Vranken, D. RMechDB: A Public Database of Elementary Radical Reaction Steps. *J. Chem. Inf. Model.* **2023**, *63*, 1114–1123.

(10) Spiekermann, K. A.; Pattanaik, L.; Green, W. H. Fast Predictions of Reaction Barrier Heights: Toward Coupled-Cluster Accuracy. *J. Phys. Chem. A* **2022**, *126*, 3976–3986.

(11) Meuwly, M. Machine Learning for Chemical Reactions. *Chem. Rev.* **2021**, *121*, 10218–10239.

(12) Lewis-Atwell, T.; Townsend, P. A.; Grayson, M. N. Machine Learning Activation Energies of Chemical Reactions. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2021**, `https://doi.org/10.1002/wcms.1593`, e1593.

(13) Komp, E.; Janulaitis, N.; Valleau, S. Progress Towards Machine Learning Reaction Rate Constants. *Phys. Chem. Chem. Phys.* **2022**, *24*, 2692–2705.

(14) Park, S.; Han, H.; Kim, H.; Choi, S. Machine Learning Applications for Chemical Reactions. *Chemistry–An Asian Journal* **2022**, *17*, e202200203.

(15) Van Gerwen, P.; Fabrizio, A.; Wodrich, M.; Corminboeuf, C. Physics-based Representations for Machine Learning Properties of Chemical Reactions. *Mach. Learn.: Sci. Technol.* **2022**, *3*, 045005.

(16) Grambow, C. A.; Pattanaik, L.; Green, W. H. Deep Learning of Activation Energies. *J. Phys. Chem. Lett.* **2020**, *11*, 2992–2997.

(17) Heid, E.; Green, W. H. Machine Learning of Reaction Properties via Learned Representations of the Condensed Graph of Reaction. *J. Chem. Inf. Model.* **2021**, *62*, 2101–2110.

(18) Schneider, N.; Lowe, D. M.; Sayle, R. A.; Landrum, G. A. Development of a Novel Fingerprint for Chemical Reactions and Its Application to Large-Scale Reaction Classification and Similarity. *J. Chem. Inf. Model.* **2015**, *55*, 39–53.

(19) Sinitskiy, A. V.; Pande, V. S. Physical Machine Learning Outperforms "Human Learning" in Quantum Chemistry. *arXiv* **2019**, *1908.00971*.

(20) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching Coupled Cluster Accuracy with a General-Purpose Neural Network Potential through Transfer Learning. *Nat. Commun.* **2019**, *10*, 2903.

(21) Doney, A. C.; Rooks, B. J.; Lu, T.; Wheeler, S. E. Design of Organocatalysts for Asymmetric Propargylations through Computational Screening. *ACS Catal.* **2016**, *6*, 7948–7955.

(22) Bender, A.; Schneider, N.; Segler, M.; Patrick Walters, W.; Engkvist, O.; Rodrigues, T. Evaluation Guidelines for Machine Learning Tools in the Chemical Sciences. *Nat. Rev. Chem.* **2022**, *6*, 428–442.

(23) Stuyver, T.; Coley, C. W. Quantum Chemistry-Augmented Neural Networks for Reactivity Prediction: Performance, Generalizability, and Explainability. *J. Chem. Phys.* **2022**, *156*, 084104.

(24) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M., et al. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388.

(25) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; Von Lilienfeld, O. A. Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *J. Chem. Theory Comput.* **2017**, *13*, 5255–5264.

(26) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1*, 1–7.

(27) Feinberg, E. N.; Joshi, E.; Pande, V. S.; Cheng, A. C. Improvement in ADMET Prediction with Multitask Deep Featurization. *J. Med. Chem.* **2020**, *63*, 8835–8848.

(28) Johnson, M. S.; Dong, X.; Grinberg Dana, A.; Chung, Y.; Farina Jr., D.; Gillis, R. J.; Liu, M.; Yee, N. W.; Blondal, K.; Mazeau, E.; Grambow, C. A.; Payne, A. M.; Spiekermann, K. A.; Pang, H.-W.; Goldsmith, F. C.; West, R. H.; Green, W. H. RMG Database for Chemical Property Prediction. *J. Chem. Inf. Model.* **2022**, *62*, 4906–4915.

(29) Zhao, Q.; Hsu, H.-H.; Savoie, B. M. Conformational Sampling for Transition State Searches on a Computational Budget. *J. Chem. Theory Comput.* **2022**, *18*, 3006–3016.

(30) Nagasawa, S.; Al-Naamani, E.; Saeki, A. Computer-Aided Screening of Conjugated Polymers for Organic Solar Cell: Classification by Random Forest. *J. Phys. Chem. Lett.* **2018**, *9*, 2639–2646.

(31) Zuranski, A. M.; Gandhi, S. S.; Doyle, A. G. A Machine Learning Approach to Model Interaction Effects: Development and Application to Alcohol Deoxyfluorination. *J. Am. Chem. Soc.* **2023**, DOI: https://doi.org/10.1021/jacs.2c13093.

(32) Nandy, A.; Duan, C.; Kulik, H. J. Using Machine Learning and Data Mining to Leverage Community Knowledge for the Engineering of Stable Metal–Organic Frameworks. *J. Am. Chem. Soc.* **2021**, *143*, 17535–17547.

(33) Kang, Y.; Park, H.; Smit, B.; Kim, J. A Multi-Modal Pre-Training Transformer for Universal Transfer Learning in Metal–Organic Frameworks. *Nat. Mach. Intell.* **2022**, *5*, 309–318.

(34) Chen, X.; Li, P.; Hruska, E.; Liu, F. Δ-Machine Learning for Quantum Chemistry Prediction of Solution-phase Molecular Properties at the Ground and Excited States. *10.26434/chemrxiv-2023-ddcr1* **2023**.

(35) Beker, W.; Gajewska, E. P.; Badowski, T.; Grzybowski, B. A. Prediction of Major Regio-, Site-, and Diastereoisomers in Diels–Alder Reactions by Using Machine-Learning: The Importance of Physically Meaningful Descriptors. *Angew. Chem., Int. Ed.* **2019**, *58*, 4515–4519.

(36) Zhong, S.; Hu, J.; Fan, X.; Yu, X.; Zhang, H. A Deep Neural Network Combined with Molecular Fingerprints (DNN-MF) to Develop Predictive Models for Hydroxyl Radical Rate Constants of Water Contaminants. *J. Hazard. Mater.* **2020**, *383*, 121141.

(37) Greaves, T. L.; McHale, K. S. S.; Burkart-Radke, R. F.; Harper, J. B.; Le, T. C. Machine Learning Approaches to Understand and Predict Rate Constants for Organic Processes in Mixtures Containing Ionic Liquids. *Phys. Chem. Chem. Phys.* **2021**, *23*, 2742–2752.

(38) Weinreich, J.; Browning, N. J.; von Lilienfeld, O. A. Machine Learning of Free Energies in Chemical Compound Space using Ensemble Representations: Reaching Experimental Uncertainty for Solvation. *J. Chem. Phys.* **2021**, *154*, 134113.

(39) Yang, L.-C.; Li, X.; Zhang, S.-Q.; Hong, X. Machine Learning Prediction of Hydrogen Atom Transfer Reactivity in Photoredox-Mediated C–H Functionalization. *Org. Chem. Front.* **2021**, *8*, 6187–6195.

(40) Wu, J.; Wang, J.; Wu, Z.; Zhang, S.; Deng, Y.; Kang, Y.; Cao, D.; Hsieh, C.-Y.; Hou, T. ALipSol: An Attention-Driven Mixture-of-Experts Model for Lipophilicity and Solubility Prediction. *J. Chem. Inf. Model.* **2022**, *62*, 5975–5987.

(41) Huoyu, R.; Zhiqiang, Z.; Zhanggao, L.; Zhenzhen, X. Quantitative Structure-Property Relationship for the Critical Temperature of Saturated Monobasic Ketones, Aldehydes, and Ethers with Molecular Descriptors. *Int. J. Quantum Chem.* **2022**, *122*, e26950.

(42) Win, Z.-M.; Cheong, A. M. Y.; Hopkins, W. S. Using Machine Learning To Predict Partition Coefficient (Log P) and Distribution Coefficient (Log D) with Molecular Descriptors and Liquid Chromatography Retention Time. *J. Chem. Inf. Model.* **2023**, DOI: https://doi.org/10.1021/acs.jcim.2c01373.

(43) Pattanaik, L.; Coley, C. W. Molecular Representation: Going Long on Fingerprints. *Chem* **2020**, *6*, 1204–1207.

(44) Wigh, D. S.; Goodman, J. M.; Lapkin, A. A. A Review of Molecular Representation in the Age of Machine Learning. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2022**, *12*, e1603.

(45) Molfeat, 2023.

(46) Grambow, C. A.; Jamal, A.; Li, Y.-P.; Green, W. H.; Zador, J.; Suleimanov, Y. V. Unimolecular Reaction Pathways of a γ-Ketohydroperoxide from Combined Application of Automated Reaction Discovery Methods. *J. Am. Chem. Soc.* **2018**, *140*, 1035–1048.

(47) Maeda, S.; Harabuchi, Y. On Benchmarking of Automated Methods for Performing Exhaustive Reaction Path Search. *J. Chem. Theory Comput.* **2019**, *15*, 2111–2115.

(48) Dewyer, A. L.; Argüelles, A. J.; Zimmerman, P. M. Methods for Exploring Reaction Space in Molecular Systems. *WIREs Comput. Mol. Sci.* **2018**, *8*, e1354.

(49) Schütt, K.; Kindermans, P.-J.; Sauceda Felix, H. E.; Chmiela, S.; Tkatchenko, A.; Müller, K.-R. SchNet: A Continuous-Filter Convolutional Neural Network for Modeling Quantum Interactions. *Adv. Neural Inf. Process Syst.* **2017**, *30.*

(50) Klicpera, J.; Giri, S.; Margraf, J. T.; Günnemann, S. Fast and Uncertainty-Aware Directional Message Passing for Non-Equilibrium Molecules. *arXiv* **2020**, *2011.14115*, Accessed 2022-03-08.

(51) Schütt, K.; Unke, O.; Gastegger, M. In *Proc. Int. Conf. Mach. Learn.* 2021, pp 9377–9388.

(52) Satorras, V. G.; Hoogeboom, E.; Welling, M. In *Proc. Int. Conf. Mach. Learn.* 2021, pp 9323–9332.

(53) Geiger, M.; Smidt, T. e3nn: Euclidean Neural Networks. *arXiv* **2022**, *2207.09453.*

(54) Li, X.; Zhang, S.-Q.; Xu, L.-C.; Hong, X. Predicting Regioselectivity in Radical C–H Functionalization of Heterocycles through Machine Learning. *Angew. Chem., Int. Ed.* **2020**, *59*, 13253–13259.

(55) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: the Δ-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087–2096.

(56) Zhu, J.; Vuong, V. Q.; Sumpter, B. G.; Irle, S. Artificial Neural Network Correction for Density-Functional Tight-Binding Molecular Dynamics Simulations. *MRS Commun.* **2019**, *9*, 867–873.

(57) Qiao, Z.; Welborn, M.; Anandkumar, A.; Manby, F. R.; Miller III, T. F. OrbNet: Deep Learning for Quantum Chemistry using Symmetry-Adapted Atomic-Orbital Features. *J. Chem. Phys.* **2020**, *153*, 124111.

(58) Atz, K.; Isert, C.; Böcker, M. N.; Jiménez-Luna, J.; Schneider, G. Δ-Quantum Machine-Learning for Medicinal Chemistry. *Phys. Chem. Chem. Phys.* **2022**, *24*, 10775–10783.

(59) Farrar, E. H. E.; Grayson, M. N. Machine Learning and Semi-empirical Calculations: A Synergistic Approach to Rapid, Accurate, and Mechanism-Based Reaction Barrier Prediction. *Chem. Sci.* **2022**, *13*, 7594–7603.

(60) García-Andrade, X.; Tahoces, P. G.; Pérez-Ríos, J.; Núnez, E. M. Barrier Height Prediction by Machine Learning Correction of Semiempirical Calculations. *arXiv* **2022**, *2208.02289*.

(61) Nandi, A.; Qu, C.; Houston, P. L.; Conte, R.; Bowman, J. M. Δ-Machine Learning for Potential Energy Surfaces: A PIP Approach to bring a DFT-based PES to CCSD(T) Level of Theory. *J. Chem. Phys.* **2021**, *154*, 051102.

(62) Ruth, M.; Gerbig, D.; Schreiner, P. R. Machine learning of coupled cluster (T)-Energy Corrections via Delta (Δ)-Learning. *J. Chem. Theory Comput.* **2022**, *18*, 4846–4855.

(63) Payne, A. M.*; Spiekermann, K. A.*; Green, W. H. Detailed Reaction Mechanism for 350–400°C Pyrolysis of an Alkane, Aromatic, and Long-Chain Alkylaromatic Mixture. *Energy & Fuels* **2022**, *36*, 1635–1646.

(64) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting Reaction Performance in C–N Cross-Coupling using Machine Learning. *Science* **2018**, *360*, 186–190.

(65) Chuang, K. V.; Keiser, M. J. Comment on "Predicting Reaction Performance in C–N Cross-Coupling using Machine Learning". *Science* **2018**, *362*, eaat8603.

(66) Estrada, J. G.; Ahneman, D. T.; Sheridan, R. P.; Dreher, S. D.; Doyle, A. G. Response to Comment on "Predicting Reaction Performance in C–N Cross-Coupling using Machine Learning". *Science* **2018**, *362*, eaat8763.

(67) Zahrt, A. F.; Henle, J. J.; Denmark, S. E. Cautionary Guidelines for Machine Learning Studies with Combinatorial Datasets. *ACS Comb. Sci.* **2020**, *22*, 586–591.

(68) 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. *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875.

(69) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the Size of Drug-like Chemical Space Based on GDB-17 Data. *J. Comput. Aided Mol. Des.* **2013**, *27*, 675–679.

(70) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(71) Terrones, G. G.; Duan, C.; Nandy, A.; Kulik, H. J. Low-Cost Machine Learning Prediction of Excited State Properties of Iridium-Centered Phosphors. *Chem. Sci.* **2023**, *14*, 1419–1433.

(72) Honrao, S.; Anthonio, B. E.; Ramanathan, R.; Gabriel, J. J.; Hennig, R. G. Machine Learning of Ab-initio Energy Landscapes for Crystal Structure Predictions. *Comput. Mater. Sci.* **2019**, *158*, 414–419.

(73) Xiong, Z.; Cui, Y.; Liu, Z.; Zhao, Y.; Hu, M.; Hu, J. Evaluating Explorative Prediction Power of Machine Learning Algorithms for Materials Discovery using k-fold Forward Cross-Validation. *Comp. Mater. Sci.* **2020**, *171*, 109203.

(74) Jorner, K.; Brinck, T.; Norrby, P.-O.; Buttar, D. Machine Learning Meets Mechanistic Modelling for Accurate Prediction of Experimental Activation Energies. *Chem. Sci.* **2021**, *12*, 1163–1175.

(75) Heinen, S.; von Rudorff, G. F.; von Lilienfeld, O. A. Toward the Design of Chemical Reactions: Machine Learning Barriers of Competing Mechanisms in Reactant Space. *J. Chem. Phys.* **2021**, *155*, 064105.

(76) Dunn, T. B.; Seabra, G. M.; Kim, T. D.; Juárez-Mercado, K. E.; Li, C.; Medina-Franco, J. L.; Miranda-Quintana, R. A. Diversity and Chemical Library Networks of Large Data Sets. *J. Chem. Inf. Model.* **2021**, *62*, 2186–2201.

(77) Bilodeau, C.; Kazakov, A.; Mukhopadhyay, S.; Emerson, J.; Kalantar, T.; Muzny, C.; Jensen, K. Machine Learning for Predicting the Viscosity of Binary Liquid Mixtures. *Chem. Eng. J.* **2023**, 142454.

(78) Yarish, D.; Garkot, S.; Grygorenko, O. O.; Radchenko, D. S.; Moroz, Y. S.; Gurbych, O. Advancing Molecular Graphs with Descriptors for the Prediction of Chemical Reaction Yields. *J. Comput. Chem.* **2023**, *44*, 76–92.

(79) Espley, S. G.; Farrar, E. H. E.; Buttar, D.; Tomasi, S.; Grayson, M. N. Machine Learning Reaction Barriers in Low Data Regimes: A Horizontal and Diagonal Transfer Learning Approach. *Digital Discovery* **2023**, DOI: 10.1039/D3DD00085K.

(80) Biswas, S.; Chung, Y.; Ramirez, J.; Wu, H.; Green, W. H. Predicting Critical Properties and Acentric Factors of Fluids Using Multitask Machine Learning. *J. Chem. Inf. Model.* **2023**, DOI: 10.1021/acs.jcim.3c00546.

(81) Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747–750.

(82) Martin, E. J.; Polyakov, V. R.; Tian, L.; Perez, R. C. Profile-QSAR 2.0: Kinase Virtual Screening Accuracy Comparable to Four-Concentration IC_{50}s for Realistically Novel Compounds. *J. Chem. Inf. Model.* **2017**, *57*, 2077–2088.

(83) Zhu, X.; Polyakov, V. R.; Bajjuri, K.; Hu, H.; Maderna, A.; Tovee, C. A.; Ward, S. C. Building Machine Learning Small Molecule Melting Points and Solubility Models Using CCDC Melting Points Dataset. *J. Chem. Inf. Model.* **2023**, DOI: 10.1021/acs.jcim.3c00308.

(84) Meredig, B.; Antono, E.; Church, C.; Hutchinson, M.; Ling, J.; Paradiso, S.; Blaiszik, B.; Foster, I.; Gibbons, B.; Hattrick-Simpers, J.; Mehta, A.; Ward, L. Can Machine Learning Identify the Next High-Temperature Superconductor? Examining Extrapolation Performance for Materials Discovery. *Mol. Syst. Des. Eng.* **2018**, *3*, 819–825.

(85) Montanari, F.; Kuhnke, L.; Ter Laak, A.; Clevert, D.-A. Modeling Physico-Chemical ADMET Endpoints with Multitask Graph Convolutional Networks. *Molecules* **2019**, *25*, 44.

(86) Lim, H.; Jung, Y. Delfos: Deep Learning Model for Prediction of Solvation Free Energies in Generic Organic Solvents. *Chem. Sci.* **2019**, *10*, 8306–8315.

(87) Durdy, S.; Gaultois, M. W.; Gusev, V. V.; Bollegala, D.; Rosseinsky, M. J. Random Projections and Kernelised Leave One Cluster Out Cross Validation: Universal Baselines and Evaluation Tools for Supervised Machine Learning of Material Properties. *Digital Discovery* **2022**, *1*, 763–778.

(88) Tricarico, G. A.; Hofmans, J.; Lenselink, E. B.; Ramos, M. L.; Dréanic, M.-P.; Stouten, P. F. W. Construction of Balanced, Chemically Dissimilar Training, Validation and Test Sets for Machine Learning on Molecular Datasets. *10.26434/chemrxiv-2022-m8l33* **2022**.

(89) Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Oberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical Assessment of QSAR Models of Environmental Toxicity against *Tetrahymena pyriformis*: Focusing on Applicability Domain and Overfitting by Variable Selection. *J. Chem. Inf. Model.* **2008**, *48*, 1733–1746.

(90) Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Molecules* **2012**, *17*, 4791–4810.

(91) Zang, Q.; Mansouri, K.; Williams, A. J.; Judson, R. S.; Allen, D. G.; Casey, W. M.; Kleinstreuer, N. C. In Silico Prediction of Physicochemical Properties of Environmental Chemicals Using Molecular Fingerprints and Machine Learning. *J. Chem. Inf. Model* **2017**, *57*, 36–49.

(92) Nakajima, M.; Nemoto, T. Machine Learning Enabling Prediction of the Bond Dissociation Enthalpy of Hypervalent Iodine from SMILES. *Sci. Rep.* **2021**, *11*, 20207.

(93) Fang, C.; Wang, Y.; Grater, R.; Kapadnis, S.; Black, C.; Trapa, P.; Sciabola, S. Prospective Validation of Machine Learning Algorithms for Absorption, Distribution, Metabolism, and Excretion Prediction: An Industrial Perspective. *J. Chem. Inf. Model.* **2023**, *63*, 3263–3274.

(94) Hwang, D.; Yang, S.; Kwon, Y.; Lee, K. H.; Lee, G.; Jo, H.; Yoon, S.; Ryu, S. Comprehensive Study on Molecular Supervised Learning with Graph Neural Networks. *J. Chem. Inf. Model.* **2020**, *60*, 5936–5945.

(95) Li, X.; Fourches, D. Inductive Transfer Learning for Molecular Activity Prediction: Next-Gen QSAR Models with MolPMoFiT. *J. Cheminform.* **2020**, *12*, 1–15.

(96) Rong, Y.; Bian, Y.; Xu, T.; Xie, W.; Wei, Y.; Huang, W.; Huang, J. Self-Supervised Graph Transformer on Large-Scale Molecular Data. *Advances in Neural Information Processing Systems* **2020**, *33*, 12559–12571.

(97) Guan, Y.; Coley, C. W.; Wu, H.; Ranasinghe, D.; Heid, E.; Struble, T. J.; Pattanaik, L.; Green, W. H.; Jensen, K. F. Regio-Selectivity Prediction with a Machine-Learned Reaction Representation and On-The-Fly Quantum Mechanical Descriptors. *Chem. Sci.* **2021**, *12*, 2198–2208.

(98) Artrith, N.; Butler, K. T.; Coudert, F.-X.; Han, S.; Isayev, O.; Jain, A.; Walsh, A. Best Practices in Machine Learning for Chemistry. *Nat. Chem.* **2021**, *13*, 505–508.

(99) Magar, R.; Wang, Y.; Lorsung, C.; Liang, C.; Ramasubramanian, H.; Li, P.; Farimani, A. B. AugLiChem: Data Augmentation Library of Chemical Structures for Machine Learning. *Mach. Learn.: Sci. Technol.* **2022**, *3*, 045015.

(100) Greenman, K. P.; Green, W. H.; Gomez-Bombarelli, R. Multi-Fidelity Prediction of Molecular Optical Peaks with Deep Learning. *Chem. Sci.* **2022**, *13*, 1152–1162.

(101) Chung, Y.; Vermeire, F. H.; Wu, H.; Walker, P. J.; Abraham, M. H.; Green, W. H. Group Contribution and Machine Learning Approaches to Predict Abraham Solute Parameters, Solvation Free Energy, and Solvation Enthalpy. *J. Chem. Inf. Model.* **2022**, *62*, 433–446.

(102) Simm, J.; Humbeck, L.; Zalewski, A.; Sturm, N.; Heyndrickx, W.; Moreau, Y.; Beck, B.; Schuffenhauer, A. Splitting Chemical Structure Data Sets for Federated Privacy-Preserving Machine Learning. *J. Cheminform.* **2021**, *13*, 1–14.

(103) Kovács, D. P.; McCorkindale, W.; Lee, A. A. Quantitative Interpretation Explains Machine Learning Models for Chemical Reaction Prediction and Uncovers Bias. *Nat. Commun.* **2021**, *12*, 1–9.

(104) Duan, Y.-J.; Fu, L.; Zhang, X.-C.; Long, T.-Z.; He, Y.-H.; Liu, Z.-Q.; Lu, A.-P.; Deng, Y.-F.; Hsieh, C.-Y.; Hou, T.-J.; Cao, D.-S. Improved GNNs for Log $D_{7.4}$ Prediction by Transferring Knowledge from Low-Fidelity Data. *J. Chem. Inf. Model.* **2023**, DOI: https://doi.org/10.1021/acs.jcim.2c01564.

(105) Li, J.; Cai, D.; He, X. Learning Graph-Level Representation for Drug Discovery. *arXiv* **2017**, *1709.03741.*

(106) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. S. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018**, *9*, 513–530.

(107) Wang, A. Y.-T.; Murdock, R. J.; Kauwe, S. K.; Oliynyk, A. O.; Gurlo, A.; Brgoch, J.; Persson, K. A.; Sparks, T. D. Machine Learning for Materials Scientists: An Introductory Guide Toward Best Practices. *Chem. Mater.* **2020**, *32*, 4954–4965.

(108) Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; Leskovec, J. Strategies for Pre-training Graph Neural Networks. *arXiv* **2019**, *1905.12265.*

(109) Ramsundar, B.; Eastman, P.; Walters, P.; Pande, V., *Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More*; O'Reilly Media, Inc.: 2019.

(110) Cáceres, E. L.; Tudor, M.; Cheng, A. C. Deep Learning Approaches in Predicting ADMET Properties. *Future Med. Chem.* **2020**, *12*, 1995–1999.

(111) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. Prediction of Chemical Reaction Yields using Deep Learning. *Mach. Learn.: Sci. Technol.* **2021**, *2*, 015016.

(112) Probst, D.; Schwaller, P.; Reymond, J.-L. Reaction Classification and Yield Prediction using the Differential Reaction Fingerprint DRFP. *Digital Discovery* **2022**, *1*, 91–97.

(113) Zahrt, A. F.; Mo, Y.; Nandiwale, K. Y.; Shprints, R.; Heid, E.; Jensen, K. F. Machine-Learning-Guided Discovery of Electrochemical Reactions. *J. Am. Chem. Soc.* **2022**, *144*, 22599–22610.

(114) Marques, E.; de Gendt, S.; Pourtois, G.; van Setten, M. J. Improving Accuracy and Transferability of Machine Learning Chemical Activation Energies by Adding Electronic Structure Information. *J. Chem. Inf. Model.* **2023**, *63*, 1454–1461.

(115) Chen, L.-Y.; Hsu, T.-W.; Hsiung, T.-C.; Li, Y.-P. Deep Learning-Based Increment Theory for Formation Enthalpy Predictions. *J. Phys. Chem. A* **2022**, *126*, 7548–7556.

(116) Chen, B.; Sheridan, R. P.; Hornak, V.; Voigt, J. H. Comparison of Random Forest and Pipeline Pilot Naive Bayes in Prospective QSAR Predictions. *J. Chem. Inf. Model.* **2012**, *52*, 792–803.

(117) Sheridan, R. P. Time-Split Cross-Validation as a Method for Estimating the Goodness of Prospective Prediction. *J. Chem. Inf. Model.* **2013**, *53*, 783–790.

(118) Struble, T. J.; Coley, C. W.; Jensen, K. F. Multitask Prediction of Site Selectivity in Aromatic C–H Functionalization Reactions. *React. Chem. Eng.* **2020**, *5*, 896–902.

(119) Pathak, Y.; Laghuvarapu, S.; Mehta, S.; Priyakumar, U. D. In *Proc. of the AAAI Conf. on Artif. Intell.* 2020; Vol. 34, pp 873–880.

(120) Thickett, S. C.; Gilbert, R. G. Propagation Rate Coefficient of Acrylic Acid: Theoretical Investigation of the Solvent Effect. *Polym.* **2004**, *45*, 6993–6999.

(121) Izgorodina, E. I.; Coote, M. L. Accurate Ab Initio Prediction of Propagation Rate Coefficients in Free-Radical Polymerization: Acrylonitrile and Vinyl Chloride. *Chem. Phys.* **2006**, *324*, 96–110.

(122) Boulebd, H. Radical Scavenging Behavior of Butylated Hydroxytoluene against Oxygenated Free Radicals in Physiological Environments: Insights from DFT Calculations. *Int. J. Chem. Kinet.* **2022**, *54*, 50–57.

(123) Vermeire, F. H.; Green, W. H. Transfer learning for solvation free energies: From quantum chemistry to experiments. *Chem. Eng. J.* **2021**, *418*, 129307.

(124) Low, K.; Coote, M. L.; Izgorodina, E. I. Explainable Solvation Free Energy Prediction Combining Graph Neural Networks with Chemical Intuition. *J. Chem. Inf. Model.* **2022**, *62*, 5457–5470.

(125) Yu, J.; Zhang, C.; Cheng, Y.; Yang, Y.-F.; She, Y.-B.; Liu, F.; Su, W.; Su, A. SolvBERT for Solvation Free Energy and Solubility Prediction: A Demonstration of an NLP Model for Predicting the Properties of Molecular Complexes. *Digital Discovery* **2023**, *2*, 409–421.

(126) Jiang, W.; Xing, X.; Zhang, X.; Mi, M. Prediction of Combustion Activation Energy of NaOH/KOH Catalyzed Straw Pyrolytic Carbon based on Machine Learning. *Renew. Energy* **2019**, *130*, 1216–1225.

(127) Xu, J.; Cao, X.-M.; Hu, P. Improved Prediction for the Methane Activation Mechanism on Rutile Metal Oxides by a Machine Learning Model with Geometrical Descriptors. *J. Phys. Chem. C* **2019**, *123*, 28802–28810.

(128) Yalamanchi, K. K.; Monge-Palacios, M.; van Oudenhoven, V. C. O.; Gao, X.; Sarathy, S. M. Data Science Approach to Estimate Enthalpy of Formation of Cyclic Hydrocarbons. *J. Phys. Chem. A* **2020**, *124*, 6270–6276.

(129) Dobbelaere, M. R.; Plehiers, P. P.; Van de Vijver, R.; Stevens, C. V.; Van Geem, K. M. Learning Molecular Representations for Thermochemistry Prediction of Cyclic Hydrocarbons and Oxygenates. *J. Phys. Chem. A* **2021**, *125*, 5166–5179.

(130) Ghanekar, P. G.; Deshpande, S.; Greeley, J. Adsorbate Chemical Environment-based Machine Learning Framework for Heterogeneous Catalysis. *Nat. Commun.* **2022**, *13*, 1–12.

(131) Zhang, X.-C.; Wu, C.-K.; Yang, Z.-J.; Wu, Z.-X.; Yi, J.-C.; Hsieh, C.-Y.; Hou, T.-J.; Cao, D.-S. MG-BERT: Leveraging Unsupervised Atomic Representation Learning for Molecular Property Prediction. *Brief. Bioinform.* **2021**, *22*, bbab152.

(132) Ismail, I.; Robertson, C.; Habershon, S. Successes and Challenges in Using Machine-Learned Activation Energies in Kinetic Simulations. *J. Chem. Phys.* **2022**, *157*, 014109.

(133) Ma, M.; Lei, X. A Dual Graph Neural Network for Drug–Drug Interactions Prediction Based on Molecular Structure and Interactions. *PLOS Comput. Biol.* **2023**, *19*, e1010812.

(134) Faber, F.; Lindmaa, A.; Von Lilienfeld, O. A.; Armiento, R. Crystal Structure Representations for Machine Learning Models of Formation Energies. *Int. J. Quantum Chem.* **2015**, *115*, 1094–1101.

(135) Okamoto, Y.; Kubo, Y. Ab Initio Calculations of the Redox Potentials of Additives for Lithium-Ion Batteries and Their Prediction through Machine Learning. *ACS Omega* **2018**, *3*, 7868–7874.

(136) Choi, S.; Kim, Y.; Kim, J. W.; Kim, Z.; Kim, W. Y. Feasibility of Activation Energy Prediction of Gas-Phase Reactions by Machine Learning. *Chem. Eur. J.* **2018**, *24*, 12354–12358.

(137) Hoffmann, G.; Balcilar, M.; Tognetti, V.; Héroux, P.; Gaüzère, B.; Adam, S.; Joubert, L. Predicting Experimental Electrophilicities from Quantum and Topological Descriptors: A Machine Learning Approach. *J. Comput. Chem.* **2020**, *41*, 2124–2136.

(138) Boobier, S.; Liu, Y.; Sharma, K.; Hose, D. R. J.; Blacker, A. J.; Kapur, N.; Nguyen, B. N. Predicting Solvent-Dependent Nucleophilicity Parameter with a Causal Structure Property Relationship. *J. Chem. Inf. Model.* **2021**, *61*, 4890–4899.

(139) Saini, V.; Sharma, A.; Nivatia, D. A Machine Learning Approach for Predicting the Nucleophilicity of Organic Molecules. *Phys. Chem. Chem. Phys.* **2022**, *24*, 1821–1829.

(140) Riedmiller, K.; Reiser, P.; Bobkova, E.; Maltsev, K.; Gryn'ova, G.; Friederich, P.; Gräter, F. Predicting Reaction Barriers of Hydrogen Atom Transfer in Proteins. *10.26434/chemrxiv-2023-7hntk* **2023**.

(141) Takahashi, K.; Miyazato, I. Rapid Estimation of Activation Energy in Heterogeneous Catalytic Reactions via Machine Learning. *J. Comput. Chem.* **2018**, *39*, 2405–2408.

(142) Stuke, A.; Todorović, M.; Rupp, M.; Kunkel, C.; Ghosh, K.; Himanen, L.; Rinke, P. Chemical Diversity in Molecular Orbital Energy Predictions with Kernel Ridge Regression. *J. Chem. Phys.* **2019**, *150*, 204121.

(143) Singh, A. R.; Rohr, B. A.; Gauthier, J. A.; Nørskov, J. K. Predicting Chemical Reaction Barriers with a Machine Learning Model. *Catal. Lett.* **2019**, *149*, 2347–2354.

(144) Cho, H.; Choi, I. S. Enhanced Deep-Learning Prediction of Molecular Properties via Augmentation of Bond Topology. *ChemMedChem* **2019**, *14*, 1604–1609.

(145) Komp, E.; Valleau, S. Machine Learning Quantum Reaction Rate Constants. *J. Phys. Chem. A* **2020**, *124*, 8607–8613.

(146) Lu, J.; Zhang, H.; Yu, J.; Shan, D.; Qi, J.; Chen, J.; Song, H.; Yang, M. Predicting Rate Constants of Hydroxyl Radical Reactions with Alkanes Using Machine Learning. *J. Chem. Inf. Model.* **2021**, *61*, 4259–4265.

(147) Abarbanel, O. D.; Hutchison, G. R. Machine Learning to Accelerate Screening for Marcus Reorganization Energies. *J. Chem. Phys.* **2021**, *155*, 054106.

(148) Bartók, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R.; Csányi, G.; Ceriotti, M. Machine Learning Unifies the Modeling of Materials and Molecules. *Sci. Adv.* **2017**, *3*, e1701816.

(149) Imbalzano, G.; Anelli, A.; Giofré, D.; Klees, S.; Behler, J.; Ceriotti, M. Automatic Selection of Atomic Fingerprints and Reference Configurations for Machine-Learning Potentials. *J. Chem. Phys.* **2018**, *148*, 241730.

(150) Stocker, S.; Csányi, G.; Reuter, K.; Margraf, J. T. Machine Learning in Chemical Reaction Space. *Nat. Commun.* **2020**, *11*, 1–11.

(151) Fabregat, R.; Fabrizio, A.; Meyer, B.; Hollas, D.; Corminboeuf, C. Hamiltonian-Reservoir Replica Exchange and Machine Learning Potentials for Computational Organic Chemistry. *J. Chem. Theory Comput.* **2020**, *16*, 3084–3094.

(152) Cersonsky, R. K.; Helfrecht, B. A.; Engel, E. A.; Kliavinek, S.; Ceriotti, M. Improving Sample and Feature Selection with Principal Covariates Regression. *Mach. Learn.: Sci. Technol.* **2021**, *2*, 035038.

(153) Dral, P. O.; Ge, F.; Xue, B. X.; Hou, Y.-F.; Pinheiro, M.; Huang, J.; Barbatti, M. MLatom 2: An Integrative Platform for Atomistic Machine Learning. *New Horizons in Comp. Chem. Software* **2022**, 13–53.

(154) Cordova, M.; Engel, E. A.; Stefaniuk, A.; Paruzzo, F.; Hofstetter, A.; Ceriotti, M.; Emsley, L. A Machine Learning Model of Chemical Shifts for Chemically and Structurally Diverse Molecular Solids. *J. Phys. Chem. C* **2022**, *126*, 16710–16720.

(155) Pattanaik, L.; Ingraham, J. B.; Grambow, C. A.; Green, W. H. Generating Transition States of Isomerization Reactions with Deep Learning. *Phys. Chem. Chem. Phys.* **2020**, *22*, 23618–23626.

(156) Jackson, R.; Zhang, W.; Pearson, J. TSNet: Predicting Transition State Structures with Tensor Field Networks and Transfer Learning. *Chem. Sci.* **2021**, *12*, 10022–10040.

(157) Makoś, M. Z.; Verma, N.; Larson, E. C.; Freindorf, M.; Kraka, E. Generative Adversarial Networks for Transition State Geometry Prediction. *J. Chem. Phys.* **2021**, *155*, 024116.

(158) Choi, S. Prediction of Transition State Structures of General Chemical Reactions via Machine Learning. **2022**, DOI: `https://doi.org/10.21203/rs.3.rs-2082595/v1`.

(159) Kendall, A.; Gal, Y. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? *Advances in Neural Information Processing Systems* **2017**, *30.*

(160) Dutschmann, T.-M.; Kinzel, L.; Ter Laak, A.; Baumann, K. Large-Scale Evaluation of k-fold Cross-Validation Ensembles for Uncertainty Estimation. *J. Cheminformatics* **2023**, *15*, 49.

(161) Scalia, G.; Grambow, C. A.; Pernici, B.; Li, Y.-P.; Green, W. H. Evaluating Scalable Uncertainty Estimation Methods for Deep Learning-Based Molecular Property Prediction. *J. Chem. Inf. Model.* **2020**, *60*, 2697–2717.

(162) Tran, K.; Neiswanger, W.; Yoon, J.; Zhang, Q.; Xing, E.; Ulissi, Z. W. Methods for Comparing Uncertainty Quantifications for Material Property Predictions. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 025006.

(163) Soleimany, A. P.; Amini, A.; Goldman, S.; Rus, D.; Bhatia, S. N.; Coley, C. W. Evidential Deep Learning for Guided Molecular Property Prediction and Discovery. *ACS Cent. Sci.* **2021**, *7*, 1356–1367.

(164) Schwalbe-Koda, D.; Tan, A. R.; Gómez-Bombarelli, R. Differentiable Sampling of Molecular Geometries with Uncertainty-Based Adversarial Attacks. *Nat. Commun.* **2021**, *12*, 1–12.

(165) McGill, C.; Forsuelo, M.; Guan, Y.; Green, W. H. Predicting Infrared Spectra with Message Passing Neural Networks. *J. Chem. Inf. Model.* **2021**, *61*, 2594–2609.

(166) Busk, J.; Jørgensen, P. B.; Bhowmik, A.; Schmidt, M. N.; Winther, O.; Vegge, T. Calibrated Uncertainty for Molecular Property Prediction using Ensembles of Message Passing Neural nNetworks. *Mach. Learn.: Sci. Technol.* **2021**, *3*, 015012.

(167) Palmer, G.; Du, S.; Politowicz, A.; Emory, J. P.; Yang, X.; Gautam, A.; Gupta, G.; Li, Z.; Jacobs, R.; Morgan, D. Calibration After Bootstrap for Accurate Uncertainty Quantification in Regression Models. *Npj Comput. Mater.* **2022**, *8*, 115.

(168) Heid, E.; McGill, C. J.; Vermeire, F. H.; Green, W. H. Characterizing Uncertainty in Machine Learning for Chemistry. *10.26434/chemrxiv-2023-00vcg-v2* **2023**.

(169) Hirschfeld, L.; Swanson, K.; Yang, K.; Barzilay, R.; Coley, C. W. Uncertainty Quantification Using Neural Networks for Molecular Property Prediction. *J. Chem. Inf. Model.* **2020**, *60*, 3770–3780.

(170) Landrum, G. et al. RDKit: Open-Source Cheminformatics Software, https://www.rdkit.org, 2006.

(171) Spiekermann, K. A.; Stuyver, T.; Pattanaik, L.; Green, W. H., https://github.com/kspieks/b2r2-reaction-rep/tree/analysis.

(172) Chollet, F., *Deep Learning with Python*; Simon and Schuster: 2021.

(173) Huyen, C., *Designing Machine Learning Systems: An Iterative Process for Production-Ready Applications*; O'Reilly Media, Inc.: 2022.

(174) Murphy, K. P., *Machine Learning: A Probabilistic Perspective*; MIT press: 2012.

(175) Müller, A. C.; Guido, S., *Introduction to Machine Learning with Python: A Guide for Data Scientists*; "O'Reilly Media, Inc.": 2016.

(176) Géron, A., *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*; O'Reilly Media, Inc.: 2019.

(177) Burkov, A., *The Hundred-Page Machine Learning Book*; Andriy Burkov Quebec City, QC, Canada: 2019; Vol. 1.

(178) Bishop, C. M.; Nasrabadi, N. M., *Pattern Recognition and Machine Learning*; 4; Springer: 2006; Vol. 4.

(179) Fregly, C.; Barth, A., *Data Science on AWS: Implementing End-to-End, Continuous AI and Machine Learning Pipelines*; O'Reilly Media, Inc.: 2021.

(180) Lakshmanan, V.; Robinson, S.; Munn, M., *Machine Learning Design Patterns: Solutions to Common Challenges in Data Preparation, Model Building, and MLOps*; O'Reilly Media, Inc.: 2020.

(181) Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V. P. Substructural Fragments: An Universal Language to Encode Reactions, Molecular and Supramolecular Structures. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 693–703.

(182) Hoonakker, F.; Lachiche, N.; Varnek, A.; Wagner, A. Condensed Graph of Reaction: Considering a Chemical Reaction as One Single Pseudo Molecule. *Int. J. Artif. Intell. Tools* **2011**, *20*, 253–270.

(183) Madzhidov, T. I.; Polishchuk, P. G.; Nugmanov, R. I.; Bodrov, A. V.; Lin, A. I.; Baskin, I. I.; Varnek, A. A.; Antipin, I. Structure-Reactivity Relationships in Terms of the Condensed Graphs of Reactions. *Russ. J. Org. Chem.* **2014**, *50*, 459–463.

(184) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *Advances in Neural Information Processing Systems* **2015**, *28*.

(185) Feinberg, E. N.; Sur, D.; Wu, Z.; Husic, B. E.; Mai, H.; Li, Y.; Sun, S.; Yang, J.; Ramsundar, B.; Pande, V. S. PotentialNet for Molecular Property Prediction. *ACS Cent. Sci.* **2018**, *4*, 1520–1530.

(186) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. In *International Conference on Machine Learning*, 2017, pp 1263–1272.

(187) Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; Jensen, K. F. Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction. *J. Chem. Inf. Model.* **2017**, *57*, 1757–1772.

(188) Hop, P.; Allgood, B.; Yu, J. Geometric Deep Learning Autonomously Learns Chemical Features That Outperform Those Engineered by Domain Experts. *Mol. Pharmaceutics* **2018**, *15*, 4371–4377.

(189) Xiong, Z.; Wang, D.; Liu, X.; Zhong, F.; Wan, X.; Li, X.; Li, Z.; Luo, X.; Chen, K.; Jiang, H.; Zheng, M. Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. *J. Med. Chem.* **2019**, *63*, 8749–8760.

(190) Sanchez-Lengeling, B.; Wei, J. N.; Lee, B. K.; Gerkin, R. C.; Aspuru-Guzik, A.; Wiltschko, A. B. Machine Learning for Scent: Learning Generalizable Perceptual Representations of Small Molecules. *arXiv* **2019**, *1910.10685*.

(191) Song, Y.; Zheng, S.; Niu, Z.; Fu, Z.-H.; Lu, Y.; Yang, Y. In *IJCAI*, 2020; Vol. 2020, pp 2831–2838.

(192) Chuang, K. V.; Gunsalus, L. M.; Keiser, M. J. Learning Molecular Representations for Medicinal Chemistry. *J. Med. Chem.* **2020**, *63*, 8705–8722.

(193) Maziarka, Ł.; Danel, T.; Mucha, S.; Rataj, K.; Tabor, J.; Jastrzębski, S. Molecule attention transformer. *arXiv* **2020**, *2002.08264*.

(194) Tang, B.; Kramer, S. T.; Fang, M.; Qiu, Y.; Wu, Z.; Xu, D. A Self-Attention Based Message Passing Neural Network for Predicting Molecular Lipophilicity and Aqueous Solubility. *J. Cheminform.* **2020**, *12*, 1–9.

(195) Wieder, O.; Kuenemann, M.; Wieder, M.; Seidel, T.; Meyer, C.; Bryant, S. D.; Langer, T. Improved Lipophilicity and Aqueous Solubility Prediction with Composite Graph Neural Networks. *Molecules* **2021**, *26*, 6185.

(196) Qian, C.; Xiong, Y.; Chen, X. Directed Graph Attention Neural Network utilizing 3D Coordinates for Molecular Property Prediction. *Comput. Mater. Sci.* **2021**, *200*, 110761.

(197) Wang, Y.; Wang, J.; Cao, Z.; Barati Farimani, A. Molecular Contrastive Learning of Representations via Graph Neural Networks. *Nat. Mach. Intell.* **2022**, *4*, 279–287.

(198) Aldeghi, M.; Coley, C. W. A Graph Representation of Molecular Ensembles for Polymer Property Prediction. *Chem. Sci.* **2022**, *13*, 10486–10498.

(199) Kang, B.; Seok, C.; Lee, J. A Benchmark Study of Machine Learning Methods for Molecular Electronic Transition: Tree-based Ensemble Learning versus Graph Neural Network. *Bull. Korean Chem. Soc.* **2022**, *43*, 328–335.

(200) Han, X.; Jia, M.; Chang, Y.; Li, Y.; Wu, S. Directed Message Passing Neural Network (D-MPNN) with Graph Edge Attention (GEA) for Property Prediction of Biofuel-Relevant Species. *Energy and AI* **2022**, *10*, 100201.

(201) Stuyver, T.; Coley, C. W. Machine Learning-Guided Computational Screening of New Candidate Reactions with High Bioorthogonal Click Potential. *Chemistry–A European Journal* **2023**, e202300387.

(202) Guan, Y.; Lee, T.; Wang, K.; Yu, S.; McWilliams, J. C. S_NAr Regioselectivity Predictions: Machine Learning Triggering DFT Reaction Modeling through Statistical Threshold. *J. Chem. Inf. Model.* **2023**, *63*, 3751–3760.

(203) Liu, Y.; Li, Z. Predict Ionization Energy of Molecules Using Conventional and Graph-Based Machine Learning Models. *J. Chem. Inf. Model.* **2023**, *63*, 806–814.

(204) Guo, J.; Sun, M.; Zhao, X.; Shi, C.; Su, H.; Guo, Y.; Pu, X. General Graph Neural Network-Based Model To Accurately Predict Cocrystal Density and Insight from Data Quality and Feature Representation. *J. Chem. Inf. Model.* **2023**, *63*, 1143–1156.

(205) Isert, C.; Kromann, J. C.; Stiefl, N.; Schneider, G.; Lewis, R. A. Machine Learning for Fast, Quantum Mechanics-Based Approximation of Drug Lipophilicity. *ACS Omega* **2023**, *8*, 2046–2056.

(206) Liu, C.; Sun, Y.; Davis, R.; Cardona, S. T.; Hu, P. ABT-MPNN: An Atom-Bond Transformer-Based Message-Passing Neural Network for Molecular Property Prediction. *J. Cheminformatics* **2023**, *15*, 29.

(207) Pan, S.; Yang, Q. A Survey on Transfer Learning. IEEE Transaction on Knowledge Discovery and Data Engineering, 2010.

(208) Hutchinson, M. L.; Antono, E.; Gibbons, B. M.; Paradiso, S.; Ling, J.; Meredig, B. Overcoming Data Scarcity with Transfer Learning. *arXiv* **2017**, *1711.05099.*

(209) Paul, A.; Jha, D.; Al-Bahrani, R.; Liao, W.-k.; Choudhary, A.; Agrawal, A. In *2019 International Joint Conference on Neural Networks (IJCNN)*, 2019, pp 1–8.

(210) Cai, C.; Wang, S.; Xu, Y.; Zhang, W.; Tang, K.; Ouyang, Q.; Lai, L.; Pei, J. Transfer Learning for Drug Discovery. *J. Med. Chem.* **2020**, *63*, 8683–8694.

(211) Grambow, C. A.; Li, Y.-P.; Green, W. H. Accurate Thermochemistry with Small Data Sets: A Bond Additivity Correction and Transfer Learning Approach. *J. Phys. Chem. A* **2019**, *123*, 5826–5835.

(212) Lu, J.; Xia, S.; Lu, J.; Zhang, Y. Dataset Construction to Explore Chemical Space with 3D Geometry and Deep Learning. *J. Chem. Inf. Model.* **2021**, *61*, 1095–1104.

(213) Zhang, D.; Xia, S.; Zhang, Y. Accurate Prediction of Aqueous Free Solvation Energies Using 3D Atomic Feature-Based Graph Neural Network with Transfer Learning. *J. Chem. Inf. Model.* **2022**, *62*, 1840–1848.

(214) Dietterich, T. G. In *Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science, vol 1857. Springer, Berlin, Heidelberg*; https://link.springer.com/chapter/10.1007/3-540-45014-9_1.pdf, 2000, pp 1–15.

(215) Spiekermann, K. A.; Pattanaik, L.; Green, W. H. High Accuracy Barrier Heights, Enthalpies, and Rate Coefficients for Chemical Reactions version 1.0.1, version 1.0.1, 2022.

(216) Alecu, I. M.; Zheng, J.; Zhao, Y.; Truhlar, D. G. Computational Thermochemistry: Scale Factor Databases and Scale Factors for Vibrational Frequencies Obtained from Electronic Model Chemistries. *J. Chem. Theory Comput.* **2010**, *6*, 2872–2887.

(217) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. Assessment of Gaussian-2 and Density Functional Theories for the Computation of Enthalpies of Formation. *J. Chem. Phys.* **1997**, *106*, 1063–1079.

(218) Petersson, G. A.; Malick, D. K.; Wilson, W. G.; Ochterski, J. W.; Montgomery Jr., J. A.; Frisch, M. J. Calibration and Comparison of the Gaussian-2, Complete Basis Set, and Density Functional Methods for Computational Thermochemistry. *J. Chem. Phys.* **1998**, *109*, 10570–10579.

(219) Anantharaman, B.; Melius, C. F. Bond Additivity Corrections for G3B3 and G3MP2B3 Quantum Chemistry Methods. *J. Phys. Chem. A* **2005**, *109*, 1734–1747.

(220) Dana, A. G.; Johnson, M. S.; Allen, J. W.; Sharma, S.; Raman, S.; Liu, M.; Gao, C. W.; Grambow, C. A.; Goldman, M. J.; Ranasinghe, D. S.; Gillis, R. J.; Payne, A. M.; Li, Y.-P.; Dong, X.; Spiekermann, K. A.; Wu, H.; Dames, E. E.; Buras, Z. J.; Vandewiele, N. M.; Yee, N. W.; Merchant, S. S.; Buesser, B.; Class, C. A.; Goldsmith, F.; West, R. H.; Green, W. H. Automated Reaction Kinetics and Network Exploration (Arkane): A Statistical Mechanics, Thermodynamics, Transition State Theory, and Master Equation Software. *Int. J. Chem. Kinet.* **2022**, DOI: 10.1002/kin.21637.

Chapter 9

Better Data Splits for Machine Learning

This work is in preparation as Burns, J. W.; Spiekermann, K. A.; Bhattacharjee, H.; Vlachos, D. G.; Green, W. H. Machine Learning Validation via Rational Dataset Sampling with astartes. *JOSS* **2023**. Jackson Burns and I designed and implemented much of this work in tandem. Jackson focused more on software engineering and algorithm development while I focused more on applications, though we each contributed to both aspects. All code is freely available on GitHub.

9.1 Introduction

Machine learning has sparked an explosion of progress in chemical kinetics [1, 2], drug discovery [3, 4], materials science [5], and energy storage [6] as researchers use data-driven methods to accelerate steps in traditional workflows within some acceptable error tolerance. To facilitate adoption of these models, researchers must critically think about several topics, such as comparing model performance to relevant baselines, operating on user-friendly inputs, and reporting performance on both interpolative and extrapolative tasks. `astartes` aims to make it straightforward for machine learning scientists and researchers to focus on two important points: rigorous hyperparameter optimization and accurate performance evaluation.

First, `astartes`' key function `train_val_test_split` returns splits for training, validation, and testing sets using an `sklearn`-like interface. These splits can then separately be used with any chosen ML model. This partitioning is crucial since best practices in data science dictate that, in order to minimize the risk of hyperparameter overfitting, one must only optimize hyperparameters with a validation set and use a held-out test set to accurately measure performance on unseen data [7–11]. Unfortunately, many published papers only mention training and testing sets but do not mention validation sets, implying that they optimize the hyperparameters to the test set, which would be blatant data leakage that leads to overly optimistic results. For researchers interested in quickly obtaining preliminary results without using a validation set to optimize hyperparameters, `astartes` also implements an `sklearn`-compatible `train_test_split` function.

Second, it is crucial to evaluate model performance in both interpolation and extrapolation settings so future users are informed of any potential limitations. Although random splits are frequently used in the cheminformatics literature, this simply measures interpolation performance. However, given the vastness of chemical space [12] and its often unsmooth nature (e.g., activity cliffs), it seems unlikely that users will want to be restricted to exclusively operate in an interpolation regime. Thus, to encourage adoption of these models, it is crucial to measure performance on more challenging splits as well. The general workflow is:

1. Convert each molecule into a vector representation.
2. Cluster the molecules based on similarity.
3. Train the model on some clusters and then evaluate performance on unseen clusters that should be dissimilar to the clusters used for training.

Although measuring performance on chemically dissimilar compounds/clusters is not a new concept [13–20], there are a myriad of choices for the first two steps; the `astartes` software incorporates many popular representations and similarity metrics to give users freedom to easily explore which combination is suitable for their needs.

9.2 Example Use-Case in Cheminformatics

To demonstrate the difference in performance between interpolation and extrapolation, `astartes` is used to generate interpolative and extrapolative data splits for two relevant cheminformatics datasets. The impact of these data splits on model performance could be analyzed with any ML model. Here, I train a modified version of Chemprop [21]–a deep message passing neural network–to predict the regression targets of interest. I use the hyperparameters reported by ref. [2] as implemented in the `barrier_prediction` branch, which is publicly available on GitHub at github.com/kspieks/chemprop/tree/barrier_prediction [22]. First is property prediction with QM9 [23], a dataset containing approximately 133,000 small organic molecules, each containing 12 relevant chemical properties calculated at B3LYP/6-31G(2df,p). I train a multi-task model to predict all properties, with the arithmetic mean of all predictions tabulated below. Second is a single-task model to predict a reaction's barrier height using the RDB7 dataset [24, 25]. This reaction database contains a diverse set of 12,000 organic reactions calculated at CCSD(T)-F12 that is relevant to the field of chemical kinetics.

For each dataset, a typical interpolative split is generated using random sampling. Two extrapolative splits are also created for comparison. The first uses the cheminformatics-specific Bemis-Murcko scaffold [26] as calculated by RDKit [27]. The second uses the more general-purpose K-means clustering based on the Euclidean distance of Morgan (ECFP4) fingerprints using 2048 bit hashing and radius of 2 [28, 29]. The QM9 dataset and RDB7 datasets were organized into 100 and 20 clusters, respectively. For each split, five different folds are created by changing the random seed. This allows the results to report the mean ± one standard deviation of the mean absolute error (MAE) and root-mean-squared error (RMSE).

Table 9.1. Average testing errors for predicting the 12 regression targets from QM9 [23].

Split	MAE	RMSE
Random	2.02 ± 0.06	3.63 ± 0.21
Scaffold	2.20 ± 0.27	3.46 ± 0.49
K-means	2.48 ± 0.33	4.47 ± 0.81

Table 9.2. Testing errors in kcal mol^{-1} for predicting a reaction's barrier height from RDB7 [24].

Split	MAE	RMSE
Random	3.87 ± 0.05	6.81 ± 0.28
Scaffold	6.28 ± 0.43	9.49 ± 0.50
K-means	5.47 ± 1.14	8.77 ± 1.85

Table 9.1 and Table 9.2 show the expected trend in which the average testing errors are higher for the extrapolation tasks than they are for the interpolation task. The results from random splitting are informative if the model will be primarily used in interpolation settings, i.e., the model will not see any molecules or substructures beyond what is contained in the training set. However, these errors are likely unrealistically low if the model is intended to make predictions on new molecules that are chemically dissimilar to those in the training

set. Performance is worse on the extrapolative data splits, which present a more challenging task, but these errors should be more representative of evaluating a new sample that is out-of-scope. Together, these tables demonstrate the utility of `astartes` in allowing users to better understand the likely performance of their model in different settings.

Briefly, it is important to point out that this analysis focuses on the general trends when comparing Table 9.1 and Table 9.2 rather than the magnitude of the errors themselves. Several approaches could be taken to further reduce the errors presented in both tables. For example, transfer learning is a popular technique [30, 31] that has found many applications in chemistry [2, 32–38]. Thus, one could pre-train on additional data or fine-tune the model with experimental values. Ensembling is another established method to improve model predictions.

9.3 References

(1) Komp, E.; Janulaitis, N.; Valleau, S. Progress Towards Machine Learning Reaction Rate Constants. *Phys. Chem. Chem. Phys.* **2022**, *24*, 2692–2705.

(2) Spiekermann, K. A.; Pattanaik, L.; Green, W. H. Fast Predictions of Reaction Barrier Heights: Toward Coupled-Cluster Accuracy. *J. Phys. Chem. A* **2022**, *126*, 3976–3986.

(3) Yang, X.; Wang, Y.; Byrne, R.; Schneider, G.; Yang, S. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. *Chem. Rev.* **2019**, *119*, 10520–10594.

(4) Bannigan, P.; Aldeghi, M.; Bao, Z.; Häse, F.; Aspuru-Guzik, A.; Allen, C. Machine Learning Directed Drug Formulation Development. *Adv. Drug Deliv. Rev.* **2021**, *175*, 113806.

(5) Wei, J.; Chu, X.; Sun, X.-Y.; Xu, K.; Deng, H.-X.; Chen, J.; Wei, Z.; Lei, M. Machine Learning in Materials Science. *InfoMat* **2019**, *1*, 338–358.

(6) Jha, S.; Yen, M.; Salinas, Y.; Palmer, E.; Villafuerte, J.; Liang, H. Learning-Assisted Materials Development and Device Management in Batteries and Supercapacitors: Performance Comparison and Challenges. *J. Mater. Chem. A* **2023**, *11*, 3904–3936.

(7) Ramsundar, B.; Eastman, P.; Walters, P.; Pande, V., *Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More*; O'Reilly Media, Inc.: 2019.

(8) Géron, A., *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*; O'Reilly Media, Inc.: 2019.

(9) Lakshmanan, V.; Robinson, S.; Munn, M., *Machine Learning Design Patterns: Solutions to Common Challenges in Data Preparation, Model Building, and MLOps*; O'Reilly Media, Inc.: 2020.

(10) Huyen, C., *Designing Machine Learning Systems: An Iterative Process for Production-Ready Applications*; O'Reilly Media, Inc.: 2022.

(11) Wang, A. Y.-T.; Murdock, R. J.; Kauwe, S. K.; Oliynyk, A. O.; Gurlo, A.; Brgoch, J.; Persson, K. A.; Sparks, T. D. Machine Learning for Materials Scientists: An Introductory Guide Toward Best Practices. *Chem. Mater.* **2020**, *32*, 4954–4965.

(12) 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. *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875.

(13) Meredig, B.; Antono, E.; Church, C.; Hutchinson, M.; Ling, J.; Paradiso, S.; Blaiszik, B.; Foster, I.; Gibbons, B.; Hattrick-Simpers, J.; Mehta, A.; Ward, L. Can Machine Learning Identify the Next High-Temperature Superconductor? Examining Extrapolation Performance for Materials Discovery. *Mol. Syst. Des. Eng.* **2018**, *3*, 819–825.

(14) Durdy, S.; Gaultois, M. W.; Gusev, V. V.; Bollegala, D.; Rosseinsky, M. J. Random Projections and Kernelised Leave One Cluster Out Cross Validation: Universal Baselines and Evaluation Tools for Supervised Machine Learning of Material Properties. *Digital Discovery* **2022**, *1*, 763–778.

(15) Jorner, K.; Brinck, T.; Norrby, P.-O.; Buttar, D. Machine Learning Meets Mechanistic Modelling for Accurate Prediction of Experimental Activation Energies. *Chem. Sci.* **2021**, *12*, 1163–1175.

(16) Stuyver, T.; Coley, C. W. Quantum Chemistry-Augmented Neural Networks for Reactivity Prediction: Performance, Generalizability, and Explainability. *J. Chem. Phys.* **2022**, *156*, 084104.

(17) Tricarico, G. A.; Hofmans, J.; Lenselink, E. B.; Ramos, M. L.; Dréanic, M.-P.; Stouten, P. F. W. Construction of Balanced, Chemically Dissimilar Training, Validation and Test Sets for Machine Learning on Molecular Datasets. *10.26434/chemrxiv-2022-m8l33* **2022**.

(18) Terrones, G. G.; Duan, C.; Nandy, A.; Kulik, H. J. Low-Cost Machine Learning Prediction of Excited State Properties of Iridium-Centered Phosphors. *Chem. Sci.* **2023**, *14*, 1419–1433.

(19) Heinen, S.; von Rudorff, G. F.; von Lilienfeld, O. A. Toward the Design of Chemical Reactions: Machine Learning Barriers of Competing Mechanisms in Reactant Space. *J. Chem. Phys.* **2021**, *155*, 064105.

(20) Bilodeau, C.; Kazakov, A.; Mukhopadhyay, S.; Emerson, J.; Kalantar, T.; Muzny, C.; Jensen, K. Machine Learning for Predicting the Viscosity of Binary Liquid Mixtures. *Chem. Eng. J.* **2023**, 142454.

(21) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M., et al. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388.

(22) Spiekermann, K. A.; Pattanaik, L.; Green, W. H.; Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M., et al., https://github.com/kspieks/chemprop/tree/barrier_prediction. Accessed 2022-05-16.

(23) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1*, 1–7.

(24) Spiekermann, K. A.; Pattanaik, L.; Green, W. H. High Accuracy Barrier Heights, Enthalpies, and Rate Coefficients for Chemical Reactions. *Sci. Data* **2022**, *9*, 417.

(25) Spiekermann, K. A.; Pattanaik, L.; Green, W. H. High Accuracy Barrier Heights, Enthalpies, and Rate Coefficients for Chemical Reactions version 1.0.1, version 1.0.1, 2022.

(26) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(27) Landrum, G. et al. RDKit: Open-Source Cheminformatics Software, `https://www.rdkit.org`, 2006.

(28) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–113.

(29) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(30) Pan, S.; Yang, Q. A Survey on Transfer Learning. IEEE Transaction on Knowledge Discovery and Data Engineering, 2010.

(31) Weiss, K.; Khoshgoftaar, T. M.; Wang, D. A Survey of Transfer Learning. *J. Big Data* **2016**, *3*, 1–40.

(32) Simoes, R. S.; Maltarollo, V. G.; Oliveira, P. R.; Honorio, K. M. Transfer and Multitask Learning in QSAR Modeling: Advances and Challenges. *Front. Pharmacol.* **2018**, *9*, 74.

(33) Grambow, C. A.; Li, Y.-P.; Green, W. H. Accurate Thermochemistry with Small Data Sets: A Bond Additivity Correction and Transfer Learning Approach. *J. Phys. Chem. A* **2019**, *123*, 5826–5835.

(34) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching Coupled Cluster Accuracy with a General-Purpose Neural Network Potential through Transfer Learning. *Nat. Commun.* **2019**, *10*, 2903.

(35) Jha, D.; Choudhary, K.; Tavazza, F.; Liao, W.-k.; Choudhary, A.; Campbell, C.; Agrawal, A. Enhancing Materials Property Prediction by Leveraging Computational and Experimental Data using Deep Transfer Learning. *Nat. Commun.* **2019**, *10*, 5316.

(36) Cai, C.; Wang, S.; Xu, Y.; Zhang, W.; Tang, K.; Ouyang, Q.; Lai, L.; Pei, J. Transfer Learning for Drug Discovery. *J. Med. Chem.* **2020**, *63*, 8683–8694.

(37) Li, X.; Fourches, D. Inductive Transfer Learning for Molecular Activity Prediction: Next-Gen QSAR Models with MolPMoFiT. *J. Cheminform.* **2020**, *12*, 1–15.

(38) Hu, H.; Bai, Y.; Yuan, Z. Improved Graph-Based Multitask Learning Model with Sparse Sharing for Quantitative Structure–Property Relationship Prediction of Drug Molecules. *AIChE Journal* **2023**, *69*, e17968.

Chapter 10

Recommendations for Future Work

This thesis has contributed several novel bodies of work towards improving the field of chemical kinetics, specifically for quantitative reaction prediction. The main theme centers around combining dataset generation via quantum chemistry with machine learning to provide easy access to accurate kinetic predictions. Sample applications of this work include automating high-throughput screening of large regions of reaction space (e.g., quickly identify relevant low-barrier reactions produced from an automated enumeration that may justify more expensive first-principles calculations if even higher accuracy is needed) and offering substantial speedup when refining existing kinetic mechanisms. Although deep graph networks are the data-driven model of choice in this thesis, it is certainly possible that future research can reveal better architectures. Despite the progress advanced by this thesis, much work remains for the field to realize the full potential of transitioning from a post-dictive to predictive modeling approach [1]. The sections below summarize some recommended future directions, such as ways to improve our group's RMG database–an open-source database that has been a seminal contribution to the field–as well as ideas to improve deep learning applications within science and engineering more generally.

10.1 Further Improving RMG Database

10.1.1 Adding High-Quality Data

Reasonably accurate estimates for thermochemistry and rate coefficients are essential for RMG to automatically construct high-fidelity kinetic mechanisms. In contrast, poor estimates can wreak havoc by causing RMG to explore irrelevant chemistry, which both wastes compute resources and often results in kinetic mechanisms that are either incomplete (i.e., missing important species and/or reactions) or contain many irrelevant reaction pathways that only slow the numerical integration when studying the time evolution of the system. As shown in Table 4.1, the training data for numerous RMG families is still quite sparse; many families have fewer than twenty reactions to train the decision tree estimators, which are then used to make predictions for many reactions during mechanism generation. It is very likely that these estimators are extrapolating; as discussed in Chapter 8 and Chapter 9, this is known to be a challenging task for machine learning models and commonly produces more erratic predictions, whose average errors are not only larger but also the standard deviation across five folds is typically much larger for extrapolation-based splits in comparison to interpolation-based splits as shown in Chapter 5, Chapter 6, Chapter 7, Chapter 8, and Chapter 9. Since the Arrhenius expression is very sensitive to small changes in the activation energy, especially at lower temperatures, these errors can be substantial. Even for the RMG reaction families that have hundreds or thousands of training reactions, as shown in Table 4.2, more than 70% of these come from a linear group additivity estimate rather than from a high-quality first principles approach that combines quantum chemistry with transition state theory or even better, a reliable experimental value.

It is essential to continue improving RMG database to address parameter errors at the source. These improvements will have broad impact by benefiting all current and future users within the kinetic community. One option is to add more gas-phase reactions from the datasets presented in Chapter 3 and Chapter 6. Although most reactions in these two datasets do not match existing RMG templates, approximately 4000 reactions total do match, which presents a prime opportunity to continue adding more high-quality training reactions. The reactions from Chapter 6 seem especially promising since all TS structures were verified by an IRC and a conformer search was done for all reactant, product, and TS structures. Importantly, these reactions contain solvation effects, which are currently not stored in RMG's database. Adding solution-phase data would be greatly beneficial to expand the scope of chemistry that RMG can model. The thermochemistry for the species involved in these reactions should be added to `Spiekermann_refining_elementary_reactions.py` on GitHub. Adding these reactions and species will increase the probability that RMG can directly query a high-quality kinetics or thermochemical value rather than relying on an estimate. Of course, it would also be beneficial to use these new data to retrain RMG's decision trees and group additivity values to improve all future estimates for kinetics and thermochemistry respectively. The hydrogen transfer reactions published in ref. [2] may also be useful. Another promising direction could be to find transition states for some of the approximately 750 million RMG reactions presented in Chapter 7. Finally, one could utilize recent work from ref. [3] to automatically parse kinetics parameters from papers published in the literature.

Another option could be to extract reaction templates from the datasets mentioned above to potentially create new RMG families. Currently, RMG relies on manually generated reaction templates, which requires human-time to review and create forbidden groups to try avoiding conflicts between reaction families. It would be greatly beneficial to develop a user-friendly workflow for automatically extracting reaction templates and integrate this data into RMG. In this case, the goal would be to expand the coverage of RMG so it can identify additional reactions that may be relevant to the kinetic mechanism of interest. If pursuing this option, a priority should be given for low-barrier reactions since those would likely be the most relevant when constructing a detailed kinetic model.

10.1.2 Improving Data Storage

A discussion about improving RMG database would be remiss without mentioning that the process of accessing current data and adding new data is not particularly user-friendly. The workflow involves several steps, with the main issue being that data is stored as a text file containing Python objects, i.e., it must be parsed via tools within RMG-Py. For example, published works in the literature have identified that RMG database stores many duplicate reactions [4]. However, this is not trivial to identify due to the text file structure that also requires installation of the entire RMG software suite. Ideally, RMG would utilize modern database architectures, such as structured query language (SQL). Although the task of updating the status quo would be very ambitions, the benefits would be equally substantial. Data storage should be more efficient, using less disk space. Querying data should be faster as well, which has benefits for kineticists seeking to interact with the database but it should also offer speed-up during RMG's automatic mechanism construction since all calls to RMG database should be faster. A tabular data format would also be more organized. For example, each row could denote a species or reaction and each column could store results from calculations at various levels of theory (e.g., electronic energy, zero point energy, and frequency scaling factor) or experimental sources (e.g., enthalpy of formation). Altogether, these improvements would substantially improve the utility and further cement RMG database as a crucial contribution to the field of chemical kinetics.

10.1.3 Use Modern Data-Driven Techniques

RMG relies on decision trees to quickly estimate kinetic parameters during mechanism generation. Historically, these trees were manually curated, which resulted in very shallow trees whose nodes were generally interpretable. The introduction of automated tree generation [5] substantially reduces the required human-time, though manual definition of the root node is still required. Although this algorithmic approach commonly produces better estimations that their hand-generated counterparts, it is prudent to ask if incorporating more modern data-driven methods would be beneficial. The first motivation stems from the compute time required to train these models. As more research groups continue to publish kinetic results, access to larger datasets that we wish to incorporate into RMG database will only grow over time. However, the current algorithm for training RMG's decision trees is relatively slow and does not scale well with these large datasets. For example, even when parallelized across many CPUs, it takes roughly 10 hours to retrain these decision trees on approximately 2000 datapoints, which is still relatively small when considering data-driven modeling techniques. In contrast, it takes only 7 minutes to train a basic directed message passing neural network [6] on approximately 10,000 reactions when using a standard GPU, such as an NVIDIA GeForce RTX 2080 Ti which was released in 2018.

The second motivation comes from a maintenance perspective. Currently, RMG supports approximately 100 reaction families. Incorporating reactions from a new diverse dataset means that each decision tree must be retrained individually. Given the computational costs discussed earlier, this is less than ideal. Having just one, or perhaps a few, deep graph network(s) would allow for more efficient maintenance. Further, some reaction families may have similar mechanisms and thus have similar kinetics. Consolidating these estimators would allow data-driven models to identify these similar patterns. If one reaction family has far fewer training reactions, consolidation could result in better predictions since it could rely on data from reactions with a similar template. One example of this is shown in Figure 6.5c in which a D-MPNN trained on a very diverse set of radical reactions was able to show very good performance for the subset of reactions that matched the hydrogen abstraction reaction template from RMG.

Finally, modern deep learning approaches can benefit from pre-training on larger amounts of data and then fine-tuning the model's weights with higher-quality data. For instance, both Table 7.4 and Table 7.5 show that pre-training a D-MPNN improves performance on downstream tasks, which is particularly impressive given how little overlap there is between the pre-training and fine-tuning sets. It is less obvious how decision trees could utilize this technique (e.g., partial learning [7] based on specific batches of data is only supported for certain classical model types within Scikit Learn [8] and the decision tree is not one of them). In short, the combination of computational speed-up, easier maintenance, and potentially better predictions present compelling reasons to explore this option in the near future.

10.2 Addressing Limitations of Machine Learning in Chemistry

10.2.1 Overcoming Data Scarcity

Data scarcity plagues chemistry. While fields such as natural language processing and computer vision routinely have access to millions or even billions of examples, many papers published within the field of predictive chemical kinetics commonly use a few hundred or maybe a few thousand data points. Obtaining new data is both challenging and expensive. Validating computational results with experiments can prove even more so.

Chapter 3, Chapter 6, and Chapter 7 introduce new datasets with approximately 12,000, 6,000, and 1,100 reactions respectively, each of which are calculated using quantum chemical methods that have been established to be very accurate relative to experimental values. Outside of this thesis, other studies have released larger datasets on the order of 100,000 reactions, albeit calculated with methods that enjoy only medium accuracy [2, 9]. Still, these numbers are dwarfed by the vastness of chemical space. There exist at least 166 billion organic molecules with up to 17 heavy atoms [10], even when applying several constraints to limit the search space, such as only considering the elements H, C, N, O, S, and halogens. The scope of possible unimolecular reactions that could be enumerated from this set is enormous since each molecule has multiple reaction sites and can participate in multiple reaction templates. The number of possible multi-molecule reactions results in a combinatorial explosion (e.g., 166×10^9 choose 2 is 1.37×10^{22}).

It is imperative for the field of predictive chemistry to continue creating more open-source kinetics datasets. One option could be to start searching for TSs for the ~750 million reactions presented in Chapter 7. The traditional workflow from Figure 1.3 could be used, which would include generating initial guesses for the TS structure via nudged elastic band [11], the doubled ended growing string method [12], or AutoTST [13] just to name a few. More recently published ML methods could also be used to quickly generate initial TS structures, though the success rate for converging to a true TS structure will likely be slightly lower [14–17]. Including solvation effects for these reactions would be another viable option to increase the scope. So far, only neutral reactions have been considered in this thesis. However, ionic reactions are prevalent in solution phase chemistry and are strongly affected by the solvent environment. Future work could explore reactions containing ionic species.

As a final point regarding dataset generation, it would also be beneficial to create standards when sharing this data. Since many SMILES strings can correspond to the same molecule, one example would be to publish data that includes InChI keys. This would make it is easier to combine data from different sources by joining on unique identifiers. Attempts to aggregate many published datasets into one database would also be helpful [18].

Regarding modeling, transfer learning is one option to try dealing with data scarcity. Transfer learning is a popular technique [19, 20] that has found many applications in chemistry [21–28]. Both Chapter 5 and Chapter 7 contain examples of transfer learning being used to improve model performance. Active learning has similarly proven useful [29].

Another promising direction is to utilize unsupervised techniques to pre-train models at scale. This has become particularly popular for the field of natural language processing (NLP) to train large language models (LLMs), e.g., the Generative Pre-trained Transformer (GPT) series [30–33]. Similar trends have been observed within text-to-image generation, e.g., DALL-E 2 from OpenAI [34] or Imagen from Google [35]. Although the cost to train such models is likely prohibitive for academic labs and often the datasets in chemistry are much smaller than those in other fields, the success of LLMs and image generation models stems from pretraining with unsupervised learning. To my knowledge, this has been explored less with graph-based networks, but it would be interesting to try pre-training GNNs to obtain a continuous molecular representation. Ideally, these pre-trained models, or the latent representations themselves, would be transferable and could be used as input for subsequent supervised tasks. This was briefly explored in Chapter 7, but future work should investigate this more thoroughly with an ultimate goal of creating foundation models for chemistry.

10.2.2 Improving Kinetics Estimation Workflows

Chapter 5 and Chapter 6 provide an important first step towards temperature-dependent kinetics estimation by training models that can quickly predict reaction barrier heights. However, actually obtaining a rate coefficient would also require fast prediction of an Arrhenius pre-exponential factor. Alternatively, models could be trained to predict the molecular partition functions of the reactants and transition state. To my knowledge, these ideas have only been briefly explored in the literature [36, 37] with far fewer papers published in comparison to models for barrier height prediction. Thus, this could present an opportunity for future research.

Regardless of the method to quickly predict kinetic parameters, there is an opportunity to integrate these models into the pipeline of automatic mechanism generation. Currently, the status quo of analyzing kinetic parameters is to perform sensitivity analysis to identify which parameters are most important to the time-evolution of the system and then use the workflow from Figure 1.3 to determine updated estimates for the rate coefficient. This first-principles approach of taking derivatives to measure rates of change is very sensible. However, it is also computationally expensive. It would be interesting to use a machine learning model to quickly predict kinetic parameters for all reactions. Then, any predictions that are wildly different than the current estimate could warrant additional analysis. Uncertainty quantification is another important topic that this thesis did not have time to explore; however, it could certainly be a useful area of future research [38–40]. Finally, various data splitting techniques, such as those aggregated within the `astartes` [41] Python package, could be integrated into popular graph network packages like Chemprop [6].

10.3 References

(1) Green, W. H. Moving from Postdictive to Predictive Kinetics in Reaction Engineering. *AIChE Journal* **2020**, *66*, e17059.

(2) Riedmiller, K.; Reiser, P.; Bobkova, E.; Maltsev, K.; Gryn'ova, G.; Friederich, P.; Gräter, F. Predicting Reaction Barriers of Hydrogen Atom Transfer in Proteins. *10.26434/chemrxiv-2023-7hntk* **2023**.

(3) Qian, Y.; Guo, J.; Tu, Z.; Coley, C. W.; Barzilay, R. RxnScribe: A Sequence Generation Model for Reaction Diagram Parsing. *arXiv* **2023**, *2305.11845*.

(4) Choi, S.; Kim, Y.; Kim, J. W.; Kim, Z.; Kim, W. Y. Feasibility of Activation Energy Prediction of Gas-Phase Reactions by Machine Learning. *Chem. Eur. J.* **2018**, *24*, 12354–12358.

(5) Johnson, M. S.; Green, W. H. In *2019 AIChE Annual Meeting*, 2019.

(6) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M., et al. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388.

(7) Scikit-learn: Machine Learning in Python.

(8) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(9) Zhao, Q.; Vaddadi, S. M.; Woulfe, M.; Ogunfowora, L.; Garimella, S.; Isayev, O.; Savoie, B. M. Comprehensive Exploration of Graphically Defined Reaction Spaces. *Sci. Data* **2023**, *10*, 145.

(10) 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. *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875.

(11) Henkelman, G.; Uberuaga, B. P.; Jónsson, H. A Climbing Image Nudged Elastic Band Method for Finding Saddle Points and Minimum Energy Paths. *J. Chem. Phys.* **2000**, *113*, 9901–9904.

(12) Zimmerman, P. M. Automated Discovery of Chemically Reasonable Elementary Reaction Steps. *J. Comput. Chem.* **2013**, *34*, 1385–1392.

(13) Harms, N.; Underkoffler, C.; West, R. H. Advances in Automated Transition State Theory Calculations: Improvements on the AutoTST Framework. *10.26434/chemrxiv* **2020**, *13277870.v2*, Accessed 2022-03-08.

(14) Pattanaik, L.; Ingraham, J. B.; Grambow, C. A.; Green, W. H. Generating Transition States of Isomerization Reactions with Deep Learning. *Phys. Chem. Chem. Phys.* **2020**, *22*, 23618–23626.

(15) Jackson, R.; Zhang, W.; Pearson, J. TSNet: Predicting Transition State Structures with Tensor Field Networks and Transfer Learning. *Chem. Sci.* **2021**, *12*, 10022–10040.

(16) Makoś, M. Z.; Verma, N.; Larson, E. C.; Freindorf, M.; Kraka, E. Generative Adversarial Networks for Transition State Geometry Prediction. *J. Chem. Phys.* **2021**, *155*, 024116.

(17) Duan, C.; Du, Y.; Jia, H.; Kulik, H. J. Accurate Transition State Generation with an Object-Aware Equivariant Elementary Reaction Diffusion Model. *arXiv* **2023**, *2304.06174*.

(18) Grinberg Dana, A. TCKDB - Theoretical Chemical Kinetics Database, 2021.

(19) Pan, S.; Yang, Q. A Survey on Transfer Learning. IEEE Transaction on Knowledge Discovery and Data Engineering, 2010.

(20) Weiss, K.; Khoshgoftaar, T. M.; Wang, D. A Survey of Transfer Learning. *J. Big Data* **2016**, *3*, 1–40.

(21) Simoes, R. S.; Maltarollo, V. G.; Oliveira, P. R.; Honorio, K. M. Transfer and Multitask Learning in QSAR Modeling: Advances and Challenges. *Front. Pharmacol.* **2018**, *9*, 74.

(22) Grambow, C. A.; Li, Y.-P.; Green, W. H. Accurate Thermochemistry with Small Data Sets: A Bond Additivity Correction and Transfer Learning Approach. *J. Phys. Chem. A* **2019**, *123*, 5826–5835.

(23) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching Coupled Cluster Accuracy with a General-Purpose Neural Network Potential through Transfer Learning. *Nat. Commun.* **2019**, *10*, 2903.

(24) Jha, D.; Choudhary, K.; Tavazza, F.; Liao, W.-k.; Choudhary, A.; Campbell, C.; Agrawal, A. Enhancing Materials Property Prediction by Leveraging Computational and Experimental Data using Deep Transfer Learning. *Nat. Commun.* **2019**, *10*, 5316.

(25) Cai, C.; Wang, S.; Xu, Y.; Zhang, W.; Tang, K.; Ouyang, Q.; Lai, L.; Pei, J. Transfer Learning for Drug Discovery. *J. Med. Chem.* **2020**, *63*, 8683–8694.

(26) Li, X.; Fourches, D. Inductive Transfer Learning for Molecular Activity Prediction: Next-Gen QSAR Models with MolPMoFiT. *J. Cheminform.* **2020**, *12*, 1–15.

(27) Spiekermann, K. A.; Pattanaik, L.; Green, W. H. Fast Predictions of Reaction Barrier Heights: Toward Coupled-Cluster Accuracy. *J. Phys. Chem. A* **2022**, *126*, 3976–3986.

(28) Hu, H.; Bai, Y.; Yuan, Z. Improved Graph-Based Multitask Learning Model with Sparse Sharing for Quantitative Structure–Property Relationship Prediction of Drug Molecules. *AIChE Journal* **2023**, *69*, e17968.

(29) Shim, E.; Tewari, A.; Cernak, T.; Zimmerman, P. M. Machine Learning Strategies for Reaction Development: Toward the Low-Data Limit. *J. Chem. Inf. Model.* **2023**.

(30) Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I., et al. Improving Language Understanding by Generative Pre-training. **2018**.

(31) Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I., et al. Language Models are Unsupervised Multitask Learners. *OpenAI blog* **2019**, *1*, 9.

(32) Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A., et al. Language Models are Few-Shot Learners. *Adv. Neural Inf. Process.* **2020**, *33*, 1877–1901.

(33) OpenAI. GPT-4 Technical Report. *arXiv* **2023**, *2303.08774*.

(34) Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical Text-Conditional Image Generation with Clip Latents. *arXiv* **2022**, *2204.06125*.

(35) Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T., et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *Adv. Neural Inf. Process.* **2022**, *35*, 36479–36494.

(36) Desgranges, C.; Delhommelle, J. A New Approach for the Prediction of Partition Functions using Machine Learning Techniques. *J. Chem. Phys.* **2018**, *149*.

(37) Komp, E.; Valleau, S. Low-Cost Prediction of Molecular and Transition State Partition Functions via Machine Learning. *Chem. Sci.* **2022**, *13*, 7900–7906.

(38) Hirschfeld, L.; Swanson, K.; Yang, K.; Barzilay, R.; Coley, C. W. Uncertainty Quantification Using Neural Networks for Molecular Property Prediction. *J. Chem. Inf. Model.* **2020**, *60*, 3770–3780.

(39) Scalia, G.; Grambow, C. A.; Pernici, B.; Li, Y.-P.; Green, W. H. Evaluating Scalable Uncertainty Estimation Methods for Deep Learning-Based Molecular Property Prediction. *J. Chem. Inf. Model.* **2020**, *60*, 2697–2717.

(40) Heid, E.; McGill, C. J.; Vermeire, F. H.; Green, W. H. Characterizing Uncertainty in Machine Learning for Chemistry. *10.26434/chemrxiv-2023-00vcg-v2* **2023**.

(41) Burns, J. W.; Spiekermann, K. A.; Bhattacharjee, H.; Vlachos, D. G.; Green, W. H. astartes.