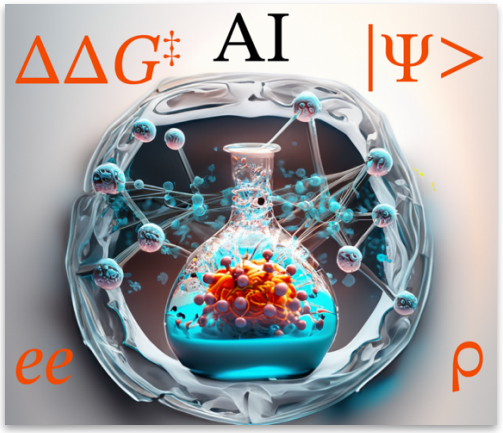
Accurate Coupled Cluster Energies via Machine Learning:

Delta Learning Extrapolated from Wavefunction and Density-Functional Theory

---

Inauguraldissertation zur Erlangung des Doktorgrades der naturwissenschaftlichen
Fachbereiche im Fachgebiet Organische Chemie (Fachbereich 08)
der Justus-Liebig-Universität Gießen



vorgelegt von
Marcel Ruth
aus Rodenbach

angefertigt im Zeitraum von Oktober 2021 bis September 2023
am Institut für Organische Chemie
der Justus-Liebig-Universität Gießen

Betreuer: Prof. Dr. Peter. R. Schreiner, PhD

“Computers make excellent and efficient servants,
but I have no wish to serve under them.”
*–**Spock** Star Trek, season 2, The Ultimate Computer **1968***

Eidesstattliche Erklärung

Hiermit versichere ich, die vorgelegte Dissertation selbständig und ohne unerlaubte fremde Hilfe und nur mit den Hilfen angefertigt zu haben, die ich in der Dissertation angegeben habe. Alle Textstellen, die wörtlich oder sinngemäß aus veröffentlichten Schriften entnommen sind, und alle Angaben, die auf mündlichen Auskünften beruhen, sind als solche kenntlich gemacht.

Bei den von mir durchgeführten und in der Dissertation erwähnten Untersuchungen habe ich die Grundsätze guter wissenschaftlicher Praxis, wie sie in der „Satzung der Justus-Liebig-Universität zur Sicherung guter wissenschaftlicher Praxis“ niedergelegt sind, eingehalten.

Ort, Datum

Unterschrift

Dekan: Prof. Dr. Thomas Wilke
Prodekan: Prof. Dr. Klaus Müller-Buschbaum
Studiendekan: Prof. Dr. Richard Dammann

Erstgutachter: Prof. Dr. Peter R. Schreiner, PhD
Zweitgutachterin: Prof. Dr. Doreen Mollenhauer

Zusammenfassung

Absolute Energien von Molekülen sind in vielen Bereichen, wie z.B. in der Atmosphärenchemie, Thermochemie, Kinetik, Katalyse, Reaktionsvorhersage und bei der Untersuchung reaktiver Intermediate essentiell. Traditionell wurden Energien durch aufwendige quantenmechanische Rechnungen erhalten. Abhängig von der Molekülgröße können dabei nur Rechnungen auf niedrigem Theorieniveau durchgeführt werden, wie z.B. Methoden die auf der Dichtefunktionaltheorie beruhen. Kleine Moleküle können mit auf Wellenfunktionen basierenden Methode, wie der Coupled Cluster Theorie untersucht werden. Das CCSD(T)/cc-pVTZ Theorieniveau wird dabei als goldener Standard der Computerchemie bezeichnet.

Durch die exponentielle Entwicklung der Rechenleistung spezieller Beschleunigerkarten (Grafikkarten) hat das maschinelle Lernen einen enormen Aufschwung erfahren und ist im alltäglichen Sprachgebrauch oft unter dem Buzzword künstliche Intelligenz zu lesen.

Im Zuge dieser Arbeit wurden zwei Methoden entwickelt, um genaue absolute Energien von Molekülen mithilfe statistischer Modelle vorherzusagen. Ausgehend von einem niedrigeren Theorieniveau, welches deutlich weniger Rechenleistung benötigt, konnten die Modelle die Energiedifferenz zum höheren Theorieniveau vorhersagen; dabei handelt es sich um einen sogenannten *Delta-Learning* Ansatz. Dies ermöglicht nicht nur eine Zeitersparnis, sondern ermöglicht auch die Vorhersage von Energien für große Moleküle, welche quantenmechanisch nicht oder nur mit hohem Zeitaufwand berechnet werden könnten.

In der ersten Veröffentlichung wurde eine Datenbank von 540 Molekülen mit der CCSD(T) Methode erzeugt, um ein Modell zu trainieren, welches den Energieunterschied zwischen der CCSD- und der CCSD(T)-Methode vorhersagen kann und dabei eine Genauigkeit von 0.25 kcal mol^{-1} aufweist. Die nachfolgende Arbeit erreichte es mit einer Datenbankgröße von 8000 Molekülen die CCSD(T)/cc-pVTZ Energie auf Grundlage von Dichtefunktional-eigenschaften mit einem mittleren absoluten Fehler von <1 kcal mol^{-1} vorherzusagen und zwar mit einer zwanzigfachen Zeitersparnis.

Abstract

The absolute energies of molecules are essential in many areas, such as atmospheric chemistry, thermochemistry, kinetics, catalysis, reaction predictions, and the study of reactive intermediates. Traditionally, energies were determined through elaborate quantum mechanical computations. Depending on the size of the molecule, only computations at a low level of theory can be carried out, such as methods based on density functional theory. Small molecules can be calculated using a wave function-based method, like the coupled cluster theory, often referred to as the gold standard of computational chemistry, especially when the CCSD(T)/cc-pVTZ theory level is used.

With the exponential development of the computing power of special accelerator cards (graphics cards), machine learning has experienced a real upswing and is often found in the everyday language under the buzzword "artificial intelligence".

In this work, two methods were developed to predict accurate molecular energies of molecules using statistical models. Starting from a lower theory level, which requires significantly less computing power, the models were able to predict the differences in energies to the higher theory level. Such an approach is known as *Delta-Learning*. This not only saves time but also enables the prediction of energies for large molecules, which could not be calculated quantum mechanically.

In the first publication, a database of 540 molecules was generated using the CCSD(T) method to train a model that can predict from the CCSD method to the CCSD(T) method and has an accuracy of 0.25 kcal mol^{-1}. The subsequent work achieved with a database size of 8000 molecules the prediction of the CCSD(T)/cc-pVTZ energy based on density functional based properties with a mean absolute error of <1 kcal mol^{-1} with twentyfold time saving.

Preface

This dissertation explores the application of machine learning techniques to problems in chemistry, representing a new avenue of research within the Institute of Organic Chemistry at Justus-Liebig University. This work aims to offer a thorough, yet accessible, overview of machine learning and its potential applications in computational chemistry, along with a concise introduction to the methods commonly employed in computational chemistry.

After establishing the foundational elements of our research, I will present the key motivations and essential components of our work. Following this, we will explore various perspectives and conclude with a summary of our findings. Detailed information on our work can be assessed from the respective peer-reviewed publications found in Chapter 2.1, where the articles are reproduced with permission from the publisher. Details of the model development and training can be found on the publishers' websites in the corresponding *Supporting Information*, which are publicly available.

Ongoing projects that are already put into manuscripts can be found in Chapter 2.2. Prepared Manuscripts. Discussing the use of machine learning in modeling asymmetric organocatalytic reactions in general and how we utilized ML to improve the Corey-Bakshi-Shibata reduction for challenging cases such as butanone.

As this work would not be possible without the help of many, I close this thesis with a few warm words towards my mentors, friends, and colleagues, who motivated and supported me over the last years.

Table of Contents

1. Introduction

1.1 Theoretical Background

1.1.1 Machine Learning

Machine Learning (ML) is a subfield of Artificial Intelligence (AI), and involves using algorithms to learn from data and make predictions or decisions based on that. The ML field can be split into supervised, unsupervised, semi-supervised, and reinforcement learning.[1]

The resurgence of ML in recent years can be attributed to advancements in specialized computer hardware like graphics processing units and accelerator cards, which are designed to expedite matrix computations. This has led to the emergence of various approaches utilizing ML. As a broader field, AI encompasses a wide range of topics and techniques, including ML. It is often used as a buzzword in media and literature to generate interest. The scope of the AI field is extensive, covering everything from brain chips, e.g., Neuralink,[2] driven by AI, to complex algorithms like linear regression. However, there are more conservative definitions of AI that may not include linear regression.[3]

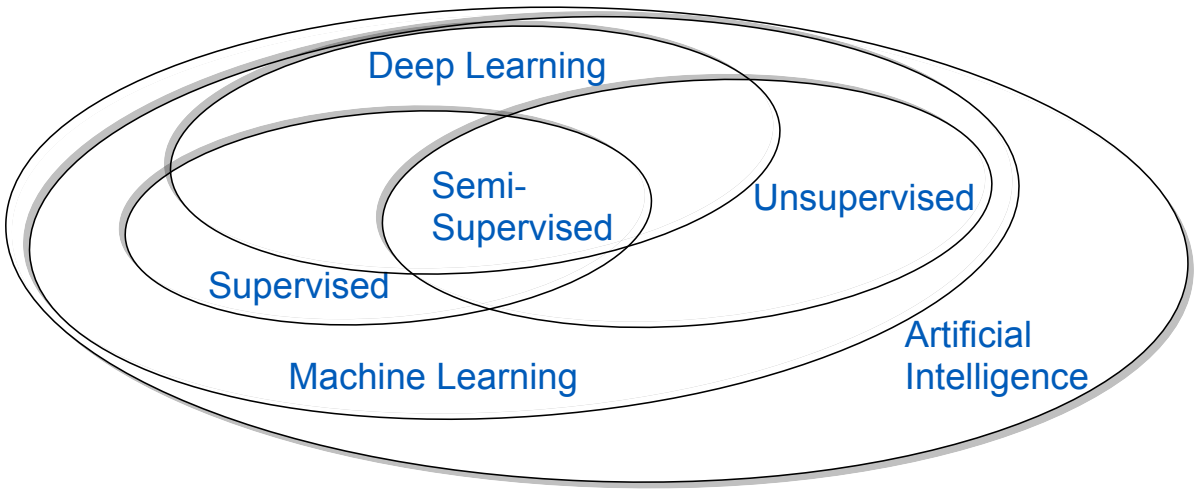


Figure 1. Venn diagram describing a simplified relation between AI, ML (part of AI), and the most common training techniques in ML, e.g., supervised, semi-supervised, and unsupervised learning. Deep learning is a big chunk of ML, which mainly consists of Neural Networks (NNs).

Supervised learning is the methodology where a model is trained on a dataset containing labeled data, which must be provided explicitly during training. These labels can be numerical values in the case of regression tasks, or categorical class labels when the task is focused on classification. A commonly used everyday analogy to describe supervised learning is the traditional educational system, where a teacher imparts knowledge to students. The students then aim to recall and apply this information in exams, a widespread teaching method for our

children. This method of instruction emphasizes guided learning, and it has unique advantages, such as direct feedback and targeted instruction.[4, 5]

In contrast, there is also unsupervised learning, which entails learning without any labels or direct supervision. An example from human learning that parallels this is the way babies, who initially don't understand any language or concepts, learn by recognizing and interpreting patterns in their surrounding environment. Without explicit instruction, they gradually learn to discern shapes, colors, and other attributes. In unsupervised learning, the ML model is not provided with labeled data but instead sifts through the data to discern abstract patterns or structures. For example, it may cluster similar data points based on their inherent relationships.

Semi-supervised learning strikes a balance between the two, leveraging both labeled and unlabeled data during the training phase. This technique capitalizes on the vast amounts of unlabeled data available to bolster the performance, accuracy, and effectiveness of models initially trained with a limited set of labeled data.[3] In this approach, the overall structure of the data space and the relationships between different data points are gleaned from the unlabeled data. This foundational knowledge is then applied to make more precise predictions on the labeled dataset. The process can be likened to a more advanced form of education where the student not only memorizes facts for rote repetition during an exam but also employs that acquired knowledge to answer questions they've never seen before—often referred to as transfer learning questions.[6]

Deep learning, a subset of ML, has revolutionized many areas of AI, including image recognition, natural language processing, and autonomous driving, by enabling computational models of multiple processing layers to learn and represent data with various levels of abstraction.[7] These algorithms discover intricate structures in large datasets by using the backpropagation algorithm to adjust internal parameters.[8] Autoencoders, a type of artificial NN used for learning efficient encodings of input data, represent an important class of unsupervised deep learning models. These models have shown impressive performance in reducing the data dimensionality.[9] While the power of deep learning is immense, it comes with challenges, such as understanding its interpretability and robustness.[10]

As this thesis primarily focuses on Delta (Δ)-learning—which will be discussed in more detail below—it focuses on supervised learning. As mentioned earlier, the main objective in a supervised learning context is to generate a prediction for a given target and then compare that prediction to what is referred to as the "ground truth" or the actual observed value. Regardless of the specific prediction task at hand, the prediction is always a numerical value denoted by $\pi \in \mathbb{R}^n$, with $n \in \mathbb{N}$, which will be predicted. This is illustrated in Figure 2 for regression and classification tasks.

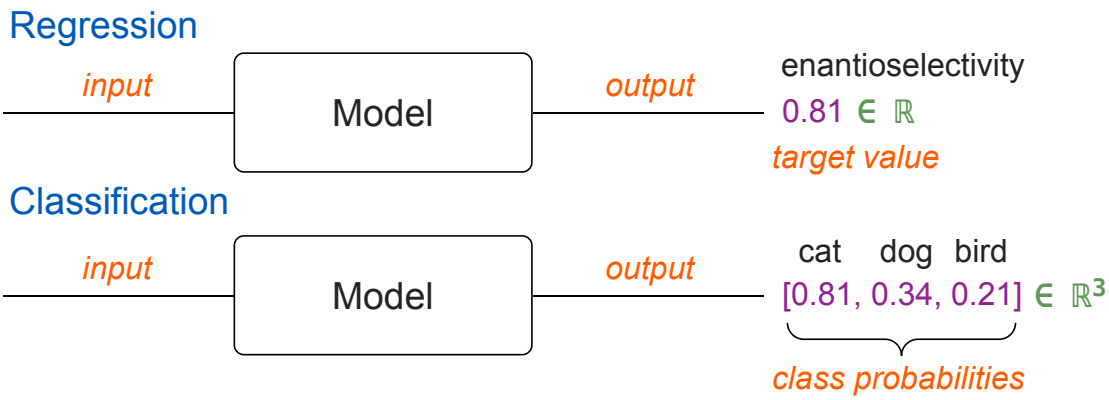


Figure 2. Shown are two flow charts describing the basic principle of an ML model for regression and classification. The model receives input and produces an output, which is a number $\pi \in \mathbb{R}^n$, with $n \in \mathbb{N}$. In the examples given, the regression, i.e., prediction of enantioselectivity, produces a number in the $\mathbb{R}$ space. In contrast, the classification, i.e., of animals, results in an output number in $\mathbb{R}^3$ space.

Now given a number that was predicted and the ground truth value, it is possible to determine how well or poorly the model performed based on a *performance metric*. This metric is chosen based on the underlying task. It can range from a simple Mean Absolute Error (MAE) for regression over negative log-likelihood (for probabilistic cases) to cross-entropy loss in classification tasks.[11] Irrespective of the actual metric and its general sign, the question can be asked in a way that the metric has to be minimized, resulting in a minimization problem. The focus of this discussion will be set on parameterized models, meaning models with trainable parameters. The metric that the model should minimize is called *loss* (*L*). It can be considered the loss of information compared to the ground truth. How such a simplified loss space with one model-parameter θ could look is illustrated in Figure 3.

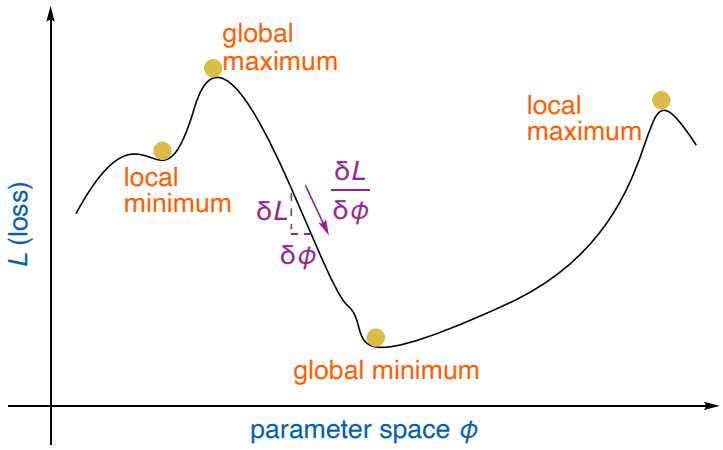


Figure 3. Artificially generated loss for a given parameter space ϕ. Extreme points are highlighted by gold disks and annotated accordingly. An exemplary slope determination at a given point is illustrated in purple.

The loss landscape in ML models can be complex and high-dimensional, making it challenging to find the global minimum. Various optimization algorithms, such as Gradient Descent (GD), Stochastic GD,[12] Adam,[13] and AdamW,[14] are available to minimize the loss. These algorithms are commonly implemented in open-source ML packages like PyTorch.[15] Besides gradient-based methods, there also exist genetic algorithms,[16] which mimic natural selection to find the optimal set of model parameters; these methods are outside of the scope of this thesis as I did not use them in any real-world project.

The gradient of the loss function with respect to the parameters, denoted as θ, plays a crucial role in guiding the parameters toward the nearest minima. This gradient is the guiding force, like gravity is for a ball rolling down a steep hill. However, it is essential to note that reaching the global minimum is not guaranteed due to the presence of local minima in the loss space.[14]

The tradeoffs in optimization algorithms differ for small-scale and large-scale learning problems. Stochastic gradient descent has shown impressive performance for large-scale problems, where the computational complexity of the optimization algorithm becomes a significant factor.[12] In large-scale ML, the limitations are often imposed by the computing time rather than the sample size.[12]

The optimization techniques mentioned are paramount to the successful training of NNs,[17] which can be used to tackle optimization problems.[1] These techniques ensure that the NN models can be reliably used within the training range and outside of it.[17] As the work described in this thesis was primarily performed using Feed-Forward NNs (FFNNs) and Graph NNs (GNNs), an overview from low- to high-level of these networks is shown in Figure 4.[8, 18]

An FFNN consists of many neurons combined with an activation function to make the model non-linear. Many of these neurons are connected and built together to form an NN. The underlying process of a neuron can be described by the simple equation Eq. 1.

$$f(\mathbf{x}) = \sigma(\mathbf{x}\mathbf{w}^{\mathrm{T}} + \mathbf{b}) \qquad \text{(Eq. 1)}$$

- $f(\mathbf{x})$: Output of a neuron, taking an input vector $\mathbf{x}$—technically $\mathbf{w}$ and $\mathbf{b}$ are also inputs to the neuron function f but are not changed once the model was trained and are therefore omitted in the simplified Eq. 1
- $\mathbf{x}$: Input vector (actual data)
- $\mathbf{w}$**:** Weight vector (optimized during training)
- $\mathbf{b}$: Bias vector (optimized during training)

Based on the activation function, the neuron might only “fire” when a certain threshold is reached. This threshold corresponds to the values inside the brackets of (Eq. 1), similar to the function of biological neurons in the human brain.[19]

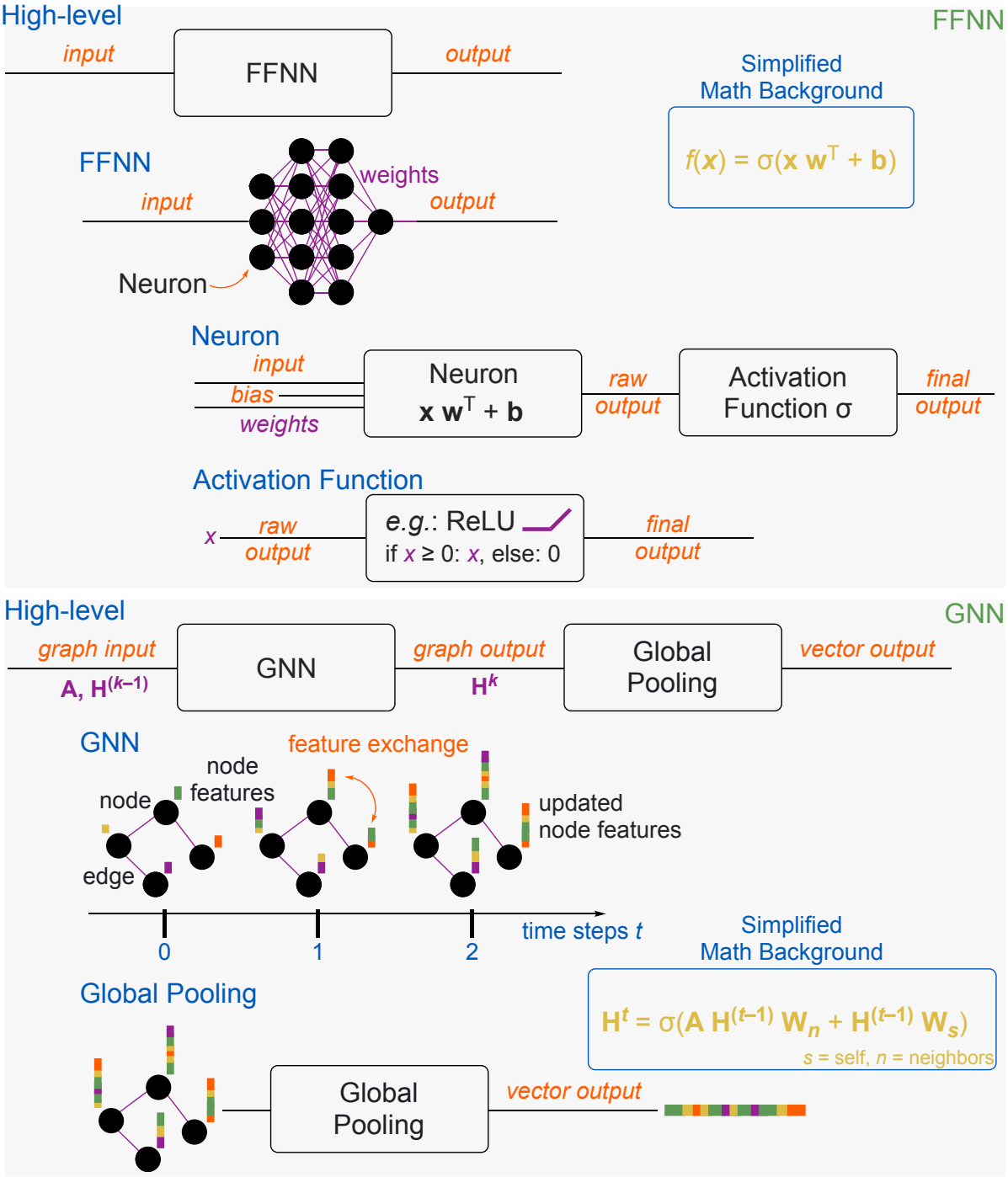


Figure 4. Overview of FFNNs (top) and GNNs (bottom) with a concise mathematical backdrop. Top: An FFNN has densely interconnected neurons, each conducting a vector calculation using the input vector **x**, weight vector **w**, and bias vector **b**. The result passes through an activation function like ReLU for non-linearity.[20] Bottom: GNNs deal with graphs with node/edge features. Features are exchanged among neighbors in each time step *t*. After a set number of steps, the updated graph is derived and reduced to a vector via global pooling.[21] This vector can be used as input in an FFNN. The connectivity of the graph is set by the adjacency matrix **A**.

The math behind GNNs is rather similar to FFNNs, which becomes clear when dissecting (Eq. 2), which describes one update step in a GNN.

$$\mathbf{H}^t = \sigma(\mathbf{A}\mathbf{H}^{(t-1)}\mathbf{W}_n + \mathbf{H}^{(t-1)}\mathbf{W}_s) \qquad \text{(Eq. 2)}$$

$$\mathbf{x}\mathbf{w}^{\mathrm{T}} \text{ is similar to } \mathbf{H}^{(t-1)}\mathbf{W}_p \qquad p = s, n$$

$\mathbf{A}$ only adds local information

- $\mathbf{H}$: Feature Matrix
- $\mathbf{W}$: Weight Matrix
- $\mathbf{A}$: Adjacency Matrix
- t: Time step

The update of the node features at the time step t, all stacked in the feature matrix $\mathbf{H}^t$, consists of the part that describes the update from adjacent nodes $\mathbf{A}\mathbf{H}^{(t-1)}\mathbf{W}_n$, and the internal node update (can also be turned off, depending on the algorithm) $\mathbf{H}^{(t-1)}\mathbf{W}_s$. The only real difference to a neuron is using the adjacency matrix $\mathbf{A} \in \mathbb{R}^{a \times a}$, with a being the number of atoms of the used graph.

The construction of an adjacency matrix from a given molecule is quite intuitive, as the column and row indices correspond directly to the atom indices. See Figure 5 for an illustration for acetyl salicylic acid† as an exemplary molecule.

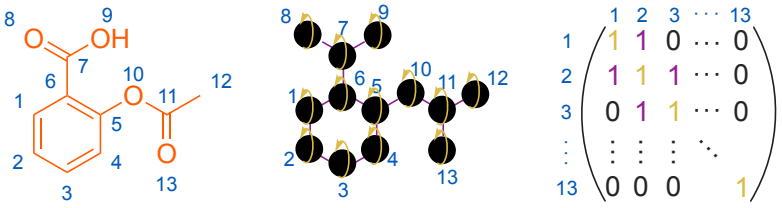


Figure 5. The molecule, e.g., acetyl salicylic acid is shown as its molecular graph and adjacency representation. In the molecular graph, the nodes (black dots) are connected via edges (purple lines). The golden arrows indicate a so-called self-loop with no direct chemical meaning. The most straightforward adjacency matrix, the binary one, is then constructed based on the molecular constitution and the respective atom numbers. Note that adjacency matrices are usually symmetric for molecular graphs because atom i is connected to atom j, and atom j is connected to atom i. The diagonal elements are 0 if no self-loop is used; otherwise, 1.

† While acetyl salicylic acid alleviates headaches, Figure 5 aspires to break down adjacency matrices without giving the reader one in the process.

The values of the adjacency matrix encode either in a binary fashion (0, 1), which atom a_i is next to atom a_j, or when weighted connections are employed, a floating-point number is used to describe the "strength" of the connection. It is also possible to use edge features, which would result in an attributed adjacency tensor $\mathbf{A} \in \mathbb{R}^{a \times a \times e}$, with e being the edge feature dimension.

A GNN is used for graph-structured data, such as molecules (Figure 6), while an FFNN is excellent at identifying connections and correlations between inputs and target values. Thus, combining both GNNs and FFNNs, makes them excel under challenging tasks. A molecule can be naturally depicted as a graph, where the nodes are represented by atoms and the edges by bonds. How a molecule can be translated to its corresponding molecular graph is shown below, along with rotation and permutation invariance, which makes graphs especially useful for chemistry-related tasks.[22, 23]

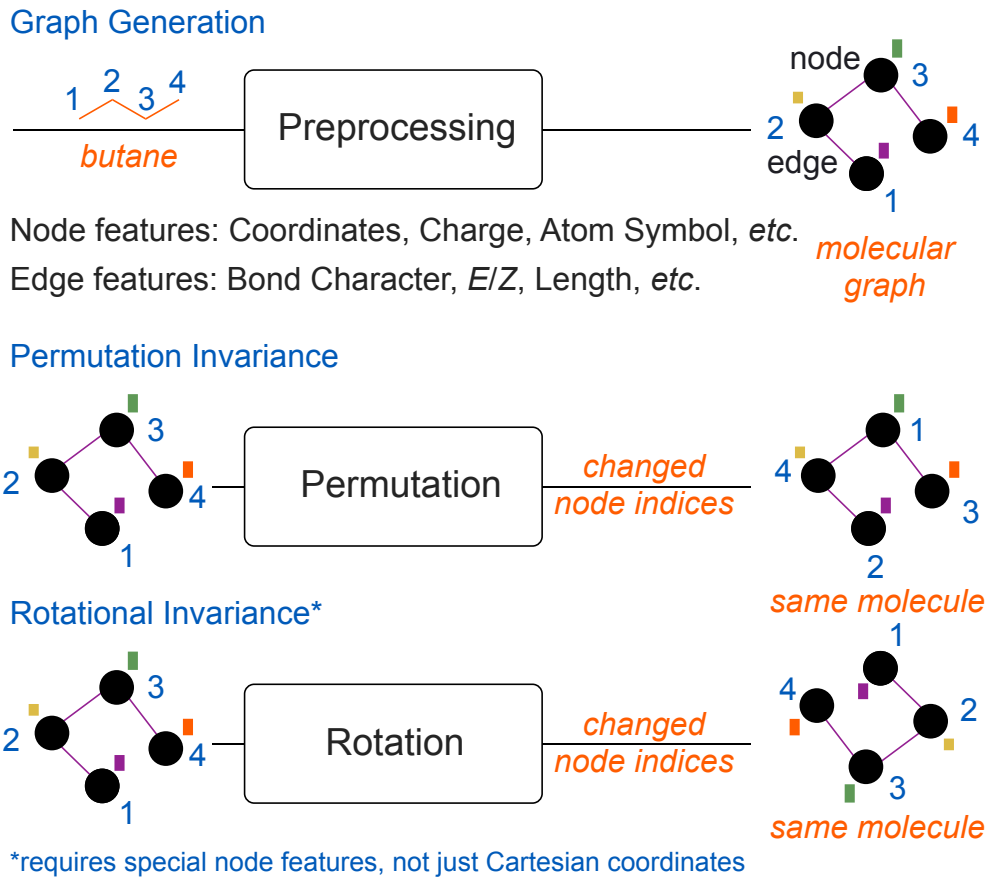


Figure 6. Shown is the transformation of butane to its corresponding molecular graph. Usually, hydrogens are encoded implicitly, and only heavy atoms are represented as nodes. Each node is enumerated in blue to keep track of the changes happening during each transition. Each node and edge can have a feature vector assigned to it. These features can be as simple as the atom symbol and as sophisticated as quantum mechanical computed properties. Graphs are always permutation invariant and can be constructed to achieve rotational invariance.

Without permutation invariance, the model that works with the graphs would predict different outcomes depending on which atom is first in the list of atoms, which makes no sense from a chemical perspective.[24, 25] Rotational invariance is useful when the three-dimensional structure of a molecule is paramount for a successful prediction of a property of interest, e.g., NN potentials that predict molecular energies under consideration of the actual conformation, not just the minimum structure.[26-33]

1.1.2 Computational Chemistry

Computational chemistry is a branch of chemistry that utilizes computers to solve chemical problems.[34] It integrates theoretical chemistry with efficient computer programs to compute the structures and properties of molecules.[34] This multidisciplinary field combines physics, computer science, and chemistry to understand the mechanisms and reactions of complex systems.[35] Over the past 50 years, computational chemistry has transitioned from a niche field to an essential component of modern chemical research.[36] Computational chemistry is now widely employed in various areas of chemistry, including organic synthesis,[37] biochemistry,[38] and materials development.[39] The increasing significance of computational chemistry in research has led to incorporating computational experiences in the undergraduate curriculum.[40] Additionally, computational chemistry plays a crucial role in drug discovery efforts, where it is utilized for compound property analysis, rationalizing drug-likeness, predicting pharmacokinetics, and designing new compounds for synthesis and biological evaluation.[41] Overall, computational chemistry has become an indispensable tool in modern chemical research, providing valuable insights into the structures, properties, and reactions of molecules and solids.

In computational chemistry, a fundamental task is to solve the time-independent Schrödinger equation (Eq. 3).

$$\widehat{H}|\Psi\rangle = E|\Psi\rangle \qquad \text{(Eq. 3)}$$

- E: Energy eigenvalue
- $|\Psi\rangle$: Wavefunction
- $\widehat{H}$: Hamiltonian Operator

Solving the Schrödinger equation for a molecule means finding the allowed energy states of the molecule and the corresponding wavefunctions. The wavefunctions can be used to calculate various properties of the molecule like electron density, charge distribution, bond lengths, angles, vibrational frequencies, and more.[42]

However, the Schrödinger equation for any atom with more than one electron (i.e., anything more complex than a hydrogen atom) has no exact analytical solution. Therefore, various approximation methods to this equation have been developed, each with its strengths and

weaknesses. Two such methods, namely Coupled Cluster (CC)[43] Theory and Density Functional Theory (DFT),[44, 45] have been used intensively during our work and will be explained below. A powerful tool for tackling electronic structure problems in computational chemistry is the CC theory.[46] It provides an efficient approximation by representing the wavefunction in an exponential ansatz shown in (Eq. 4).

$$|\Psi_{\mathrm{CC}}\rangle = e^{\hat{T}}|\Phi\rangle \qquad \text{(Eq .4)}$$

- $|\Phi\rangle$: Reference Wavefunction
- $|\Psi_{\mathrm{CC}}\rangle$: CC Wavefunction
- $\widehat{\mathrm{T}}$: Cluster Operator

The cluster operator is usually expressed as a sum of one-body, two-body, three-body, etc., excitation operators, shown in (Eq. 5).

$$\hat{T} = \hat{T}_1 + \hat{T}_2 + \hat{T}_3 + \cdots \qquad \text{(Eq. 5)}$$

Each operator corresponds to the rise of the system to an excited state (singles, doubles, triples, etc.). For practical reasons, this series is truncated at the double excitations level, and the triples excitations are only included by an approximation via perturbation theory. When used with the cc-pVTZ basis set,[47] the CCSD(T)/cc-pVTZ level of theory is known as the gold standard in computational chemistry.[48, 49]

Unlike methods that operate via wavefunctions, DFT considers the electronic density as the fundamental property. The Hohenberg-Kohn theorem legitimizes DFT, stating that its electron density uniquely determines the ground-state properties of a many-electron system;[44] the central Kohn-Sham equation is shown in (Eq. 6).[45, 50]

$$\left[-\frac{1}{2}\nabla^2 + V_{\mathrm{ext}}(\mathbf{r}) + V_{\mathrm{H}}[\rho(\mathbf{r})] + V_{\mathrm{XC}}[\rho(\mathbf{r})]\right]\Psi_i(\mathbf{r}) = E_i(\mathbf{r})\Psi_i(\mathbf{r}) \qquad \text{(Eq. 6)}$$

- $\psi_i(\boldsymbol{r})$: Kohn-Sham Orbitals (Single Electron Wavefunctions)
- ∇: Kinetic Energy Operator
- $V_{\mathrm{ext}}(\boldsymbol{r})$: External Potential
- $V_{\mathrm{H}}[\rho(\boldsymbol{r})]$: Coulomb Potential
- $V_{\mathrm{XC}}[\rho(\boldsymbol{r})]$: Exchange Correlation Potential
- $\varepsilon_i(\boldsymbol{r})$: Energy Eigenvalues
- $\boldsymbol{r}$: Position Vector

The exchange-correlation potential $V_{\mathrm{XC}}[\rho(\mathbf{r})]$ is the only unknown in equation (Eq. 6) and must be approximated. Different approximations lead to the various "flavors" of DFT such as local density approximation, generalized gradient approximation, or hybrid functionals that mix generalized gradient approximation with a portion of exact exchange from Hartree-Fock theory.[43]

1.2 Motivation

Computational chemistry based methods have revolutionized the study of many systems. One of the main challenges of computational chemistry is to balance the trade-off between accuracy and computational cost. High-level methods, such as CC theory or multireference methods, can provide reliable results but are often too expensive to apply to large or complex systems. Low-level methods, such as DFT or semi-empirical methods, can handle larger systems but may suffer from systematic errors or lack of transferability. For more details on the sparks of computations in chemical discovery, I recommend the recent editorial article of *Nature Communications*.[51]

The CC method, recognized as one of the most accurate techniques for approximating solutions to the Schrödinger equation, has found considerable applications in quantum chemistry and condensed matter physics.[46] However, the method's widespread implementation is hampered due to its high computational cost, which is based on its adverse scaling of $O(N^x)$ with N being the number of basis functions and x being 6, 8, and 10 for CCSD,[52-55] CCSDT,[56, 57] and CCSDTQ.[46, 58-60] The perturbatively approximated triples correction CCSD(T)[49, 61, 62] achieves close to CCSDT accuracy and scales with $O(N^7)$. Various approaches have been developed to address the steep computational scaling in Coupled Cluster (CC) methods, such as Divide-Expand-Consolidate DEC-CCSD(T),[63] Cluster In Molecule CIM-CC,[64] and Domain-Based Local Pair Natural Orbital DLPNO-CCSD(T).[65-67] For an in-depth examination of linear-scaling methodologies in quantum chemistry, I recommend the review by Ochsenfeld, Kussmann, and Lambrecht.[68] Complementing this is the insightful analysis by Bowler and Miyazaki.[69] It is this intersection where accuracy meets computational efficiency, which we aim to address through the application of ML.

Given these challenges and computational demands in quantum chemistry, a paradigm shift is necessary to tap into the unexplored potential of predicting molecular properties accurately without incurring excessive computational costs. Such a promise is offered by ML, acting as a bridge between detailed quantum mechanical methods and efficient predictions.

Recently, ML models, especially those based on NNs and decision trees, have been employed to predict properties that traditionally necessitate expensive quantum mechanical computations.[26, 70, 71] The efficacy of ML is apparent in its ability to map high-dimensional and nonlinear spaces and, therefore, holds promise for significantly improving computational efficiency without compromising accuracy.[72]

This promise has already begun to be realized, as recent literature has reported the successful use of ML in predicting outcomes from DFT and time-dependent DFT.[33, 73-77] Prediction of CCSD energies performed on a wide range of molecules has been performed by Townsend Vogiatzis.[78, 79] However, the application of ML in predicting CCSD(T) energies has been shown

only for exemplary cases, such as the adsorption of CO_2 in porous materials,[80] or liquid water.[81] This thesis aims to address this gap by exploring and evaluating ML algorithms' suitability for predicting CCSD(T) energies for a wide range of molecules using Δ-learning approaches.[82, 83]

The motivation for the thesis comes from the potential to explore new territories in thermochemistry, atmospheric chemistry, and prebiotic chemistry, which require precise computations of molecular energies and properties that are computationally expensive.[83-85]

In thermochemistry, accurate electronic energies are crucial for studying heat capacities, enthalpies, and free energies of molecules.[86-89] Similarly, in atmospheric chemistry, understanding chemical kinetics and photochemical reactions necessitates precise energy computations.[90, 91] Prebiotic chemistry, focusing on studying extraterrestrial atmospheres and interstellar medium, also relies heavily on these energies.[92, 93]

Furthermore, this work will contribute to the burgeoning field of quantum ML, which combines the principles of quantum physics and ML.[94] In addition, this work will provide valuable insights into the application of ML in predicting complex physical properties.

1.3 Machine Learning in Chemistry

Incorporating ML techniques in computational chemistry has sparked a considerable shift, driving the advancement of chemical research. ML algorithms are proficient in deciphering patterns and making predictions from large and complex datasets, a trait beneficial for several aspects of computational chemistry, such as chemical structure prediction, reaction optimization, property prediction, and drug discovery.[24]

In chemical structure prediction, ML techniques like NNs have shown remarkable potential. The prediction of quantum mechanical properties for new, not-yet-synthesized molecules is a computational bottleneck, and NNs have demonstrated the potential to expedite this process. The NNs learn representations of molecular structures and are trained to predict quantum mechanical properties, significantly reducing the computational time and resources compared to traditional methods.[72]

Reaction optimization is another area benefiting from ML techniques. Here, the algorithms are trained to predict the best conditions for a chemical reaction by considering a variety of parameters, such as temperature, pressure, catalyst, and solvent. The models are trained on experimental reaction data and can suggest optimal reaction conditions, thereby minimizing the need for trial-and-error experimentation.[95]

In the context of property prediction, ML can help streamline the process of determining the physical, chemical, and biological properties of compounds. One primary area where this is employed is in the prediction of solubility, a fundamental property in drug design and

environmental science. Using ML, the solubility of a vast array of compounds can be predicted based on their molecular structure.[96]

Furthermore, ML has shown its prowess in drug discovery, helping identify potential drug candidates among billions of compounds. ML models are trained on databases of known drug compounds and their biological targets and can predict potential new drug-target interactions, greatly accelerating the drug discovery process.[96]

The integration of ML in computational chemistry is challenging. As with all ML applications, the quality and quantity of training data are key to the performance of these models. This is especially relevant when working with experimental rather than computational data. Additionally, the interpretability of the results provided by ML algorithms, particularly deep learning models, is difficult due to the “black box” type architecture, as mentioned earlier.[10]

Nevertheless, the potential benefits of ML in computational chemistry are immense. As the field matures, these techniques will continue transforming chemical research, enabling more efficient and precise predictions and streamlining the processes of developing new materials and drugs. The acceleration in this field will ultimately result in more environmentally and economically friendly chemical research and production.

1.4 Machine Learning Approaches for Energy Predictions

1.4.1 Neural Network Potentials

Neural Network Potentials (NNPs) have emerged as a powerful tool in computational chemistry for the accurate and efficient prediction of potential energy surfaces (PESs) of molecular and condensed matter systems. The NNP approach builds upon the fundamental concept of representing the PES as a high-dimensional function, which is modeled using NNs.[32] An overview of the history and development of NNPs is shown in a recent review by Behler, Ko, and Kocer.[97] Four of the main NNP generations and their differences are depicted in Figure 7.

In constructing an NNP, the system's total energy is typically partitioned into atomic contributions. Each atomic energy is then modeled as a function of its local environment, described by a set of symmetry functions capturing the radial and angular distribution of neighboring atoms. These symmetry functions serve as inputs to the NN, which is trained to learn the mapping between these inputs and the atomic energy based on reference data derived from high-level quantum mechanical computations.[98] An excellent tutorial review by Behler discusses how to construct an NN in detail.[99]

A key advantage of NNPs is their ability to achieve quantum mechanical accuracy at a computational cost comparable to classical force fields. They are capable of modeling

complex, high-dimensional PESs with exceptional flexibility, given the universal approximation properties of NNs. Furthermore, once trained, NNPs can almost instantaneously predict energies and forces, making them highly scalable for large systems and extended simulations, like, e.g., the ANI-1 NNP.[30] The ANI-X NNPs were recently implemented in Pytorch in the TorchANI Python package and are now easy to access.[100, 101]

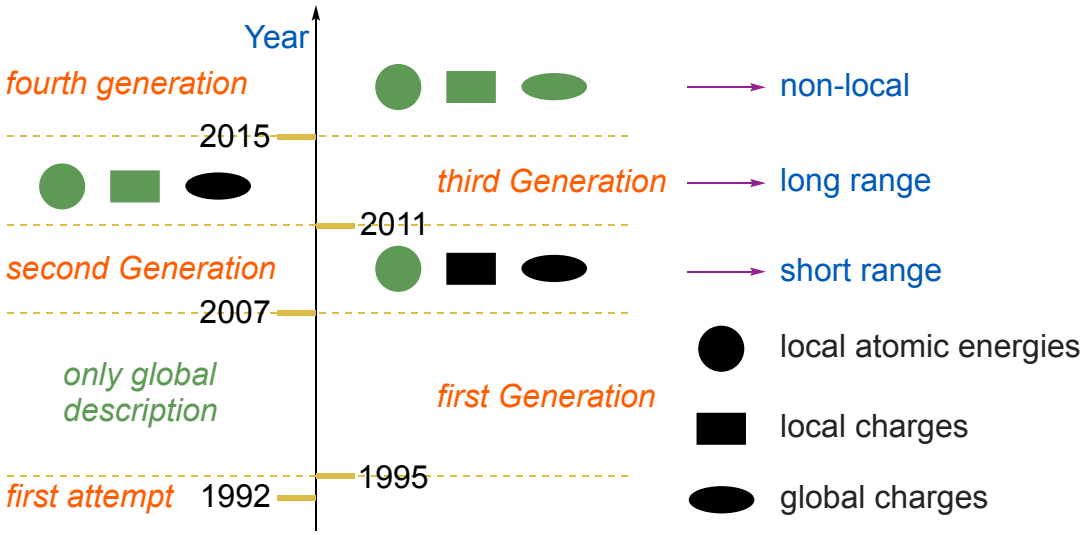


Figure 7. Depiction of four NNP generations, adapted from the mentioned review by Behler, Ko, and Kocer and an independent review by Behler.[102] The “first attempt”[103] by Sumpter and Noid does not count as the first generation based on Behlers definition.[102] The green color of the symbols indicates if this feature is available in that particular generation.

However, the construction and training of NNPs require careful consideration. For instance, the choice of symmetry functions, the size and architecture of the NN, and the composition of the training set can significantly influence the accuracy and transferability of the NNP. Strategies such as active learning have been proposed to iteratively refine the training set and improve the NNP's performance.[104]

NNPs represent a potent approach for modeling PESs in computational chemistry. They offer the prospect of combining the accuracy of quantum mechanical methods with the efficiency of classical potentials, thereby enabling the exploration of complex chemical systems on length and time scales previously unattainable.

1.4.2 Δ-Learning via Machine Learning

Correction learning, formally known as Δ-learning, represents a significant innovation in the realm of computational chemistry, particularly for its proficiency in molecular property prediction, such as the critical domain of energy calculation. At the core of Δ-learning lies the

forecasting of the “delta” or difference between two levels of quantum chemical computations: one achieved through a computationally economical method, often characterized as “low-fidelity” or “low-level”, and the other through a more precise but computationally demanding technique, referred to as “high-fidelity” or “high-level”. An intuitive and schematic illustration of Δ-learning is given in Figure 8.

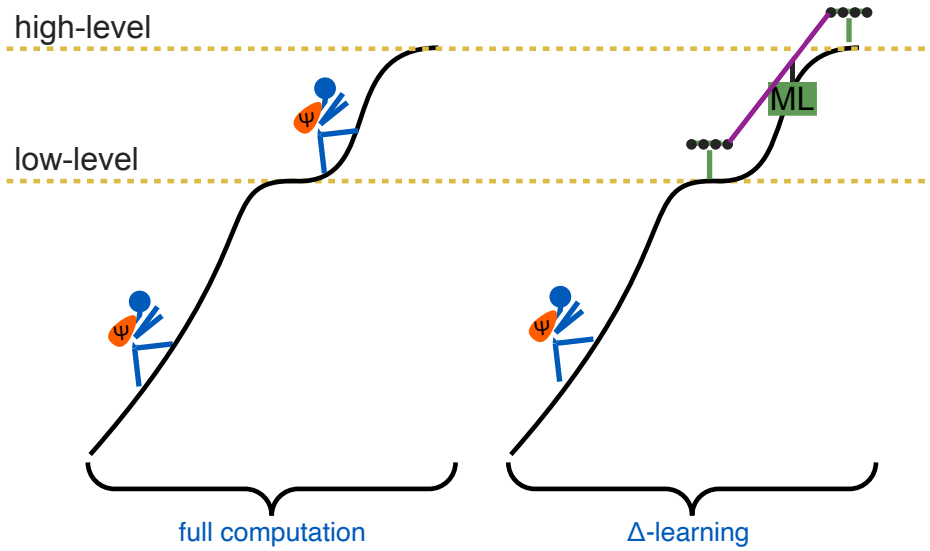


Figure 8. Schematic representation to illustrate Δ-learning, inspired by the concept of Jacob’s Ladder in DFT.[105] On the left side, the computational effort towards the high-level theory has to be performed solely by quantum chemical computations, denoted by the Ψ on the backpack of the “climber”. On the right side, the low-level theory is reached by quantum chemical computations, while the last part of the “mountain” is traveled by the “ML-Lift”. From this depiction, it is clear that it is possible to save computational time by choosing the lowest level theory, but a masterpiece of ML engineering will be needed to construct a practical and working “ML-Lift” to compensate for a larger distance to the “summit”.

The “delta”, in essence, constitutes a correction factor. It represents the offset between a low-level quantum chemistry method, such as DFT or Hartree-Fock, and a high-level method, like CC theory with single, double, and perturbative triple excitations (CCSD(T)), which often serves as the benchmark method for computations.

An intriguing aspect of this correction factor is its significantly lower complexity and magnitude than absolute property values. These characteristics render the correction factor highly tractable for modeling and prediction through ML algorithms. This relatively more straightforward problem structure can help mitigate issues commonly associated with high-dimensional data, such as the curse of dimensionality coined by Bellman in 1957.[106] Firstly, the *exponential growth of volume*:[10] As the dimensionality increases, the volume of the space increases so fast that the available data becomes sparse. This sparsity is problematic for any

method that requires statistical significance. This problem also means that the data needed often grow exponentially with the dimensionality to obtain a statistically sound and reliable result. Secondly, the *increased distance*:[107] In high-dimensional spaces, points tend to be far apart. This dispersal can make grouping similar data points in clusters challenging as the notion of "closeness" becomes less clear. Thirdly, the *increased computational complexity*:[108] Algorithms that run efficiently in low dimensions can become intractable when the input is high-dimensional. And, lastly, *overfitting*:[10] High dimensionality can lead to overfitting in models, especially in ML. With more features (dimensions), a model might create a complex decision boundary that overfits to the noise or outliers in the training data, leading to poor performance on unseen data.

To address these challenges, various dimensionality reduction techniques, such as principal component analysis,[109] *t*-distributed stochastic neighbor embedding,[110] and autoencoders[9] in deep learning, are used. These techniques aim to reduce the number of random variables under consideration by obtaining a set of principal variables. An illustration of dimensionality reduction is shown in Figure 9.

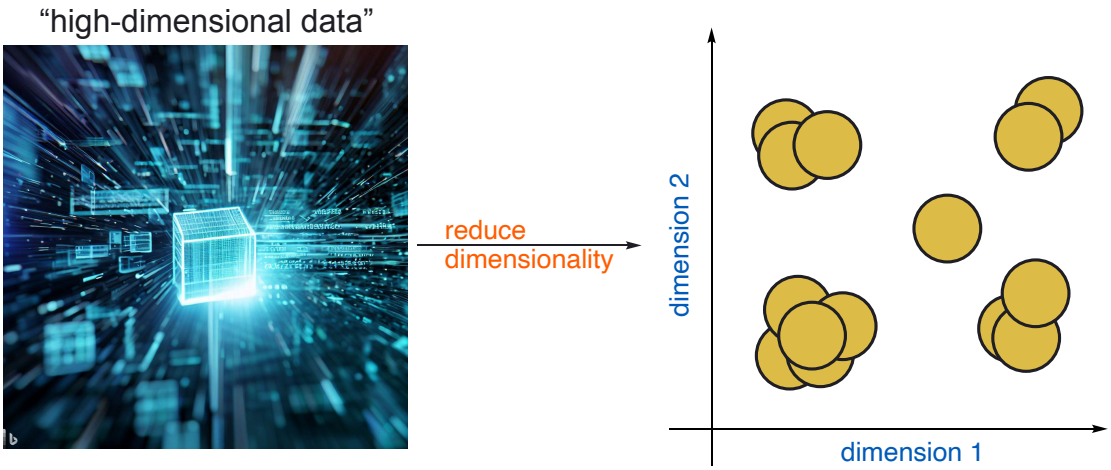


Figure 9. On the left is an illustration of high-dimensional data (created with Bing Image Creator)[111] that is reduced into two-dimensional space by a dimensionality reduction method of choice. This improves intuitive understanding of the data, reduces computational complexity, and helps identify specific data clusters.

In a typical Δ-learning workflow, the initial approximation of the molecular property (e.g., energy) is carried out via a low-level quantum mechanical method. This choice is primarily driven by the desire to reduce computational overhead while maintaining reasonable accuracy. Subsequently, an ML model is trained to predict the “delta” or correction term. This term effectively refines the initial approximation, bridging it towards the high-level quantum mechanical result.

The interplay between quantum mechanical methods and ML is central to the Δ-learning approach. The former promises high-accuracy results, albeit often at a substantial computational cost. The latter, armed with its capacity to generalize from large datasets and to make swift predictions for new instances, brings scalability and computational efficiency. Therefore, by leveraging a low-level quantum mechanical method for the initial approximation and using an ML model for refinement, the Δ-learning approach achieves a balance of accuracy and computational efficiency, allowing for high-quality predictions on a much larger scale. Similarly, the benefit of Δ-learning was shown in a recent study by Sun et al. for image recognition.[112]

The advantages of Δ-learning are particularly apparent in the realm of large molecular databases and extensive to compute molecules.[113] Here, computational efficiency and the ability to make high-throughput predictions are paramount. In these contexts, Δ-learning offers a feasible and efficient solution, leading to its adoption in high-throughput virtual screening and database-driven materials design.[113] Approaches based on Δ-learning present an exciting frontier in computational chemistry, offering a strategy that is not only accurate and efficient but also scalable. As computational resources continue to be a limiting factor in high-level quantum chemical computations, techniques such as Δ-learning will become invaluable for the field.

Many Δ-learning approaches have emerged in the past decade. In 2015 von Lilienfeld et al. used Δ-learning to predict the atomization energies of molecules with an accuracy comparable to CC theory but with a speedup of six orders of magnitude.[113] In the same year Lilienfeld et al. predicted electronics spectra from time-dependent DFT.[114] Correction of DFT errors for prediction of molecular forces and vibrational frequencies was conducted by Marquetand et al. in 2017.[98] One year later Mei et al. Δ-learned the free energy potential at *ab initio* accuracy from a semi-empirical reference potential.[115] The research of Yang et al. illustrates in 2019 the effectiveness of a basic Δ-learning model in enhancing the precision of solvent-free energy computations.[116] They achieved a level of accuracy similar to hybrid DFT, using a semi-empirical Density Functional Tight-binding as a foundation. In 2020 Aspuru-Guzik et al. used Δ-learning to improve the accuracy of semi-empirical methods for predicting reaction energies and barriers.[98] A comparable approach to the strategy by Yang et al. was utilized in 2021 by Riniker et al. in their simulation of organic compound interactions in water.[117] We performed the correction of perturbatively included triples from CCSD to CCSD(T) levels of theory with various Dunning basis sets in 2022.[83] A year later Li. et al. predicted a CCSD(T)-Quality potential energy surface for the CH_3OH + OH reaction.[118] Later this year Grimme et al. used DFT properties to predict NMR shifts at CCSD(T)-Quality.[119] One month later, we predicted the CCSD(T)/cc-pVTZ and the DLPNO approximated CC theory, namely

DLPNO-CCSD(T)/cc-pVTZ, levels of theory based on DFT properties, with various exchange-correlation approximations.[82] As Δ-learning is still an active area of research there are many opportunities for further development and application in the future. Recently Savoie et al. performed—to the best of my knowledge—the first Δ^2-learning approach, where they made reaction property predictions based on Δ-learning in geometry and energy.[120]

1.5 Remarks About Our Published Work

We were able to predict the energies at various CCSD(T) levels of theory using double-ζ and triple-ζ basis sets; this is shown in Figure 10. We first used CCSD computations that provided us with HF,[121] MP2,[122-126] and CCSD energies to accurately predict energies at the CCSD(T)/*X* levels of theory, with *X* being cc-pVDZ, aug-cc-pVDZ, and cc-pVTZ.[47, 127] Our model supported the most common elements in organic chemistry: hydrogen, carbon, oxygen, and sulfur. Though the dimer results showed a larger error, predictions were possible for monomers and dimers. We validated our models with challenging molecules, such as highly conjugated and atmospherically relevant molecules. How minor differences in the molecular constitution were handled by our model was tested with a set of constitutional isomers. To check if the model could predict non-covalent interactions without being trained on them, we used the S22 dataset for validation. For more details, see Chapter 2.1.1 Machine Learning of Coupled Cluster (T)-Energy Corrections via Delta (Δ)-Learning.

In our current model iteration, we were able to predict energies in a probabilistic fashion at the CCSD(T)/ and DLPNO-CCSD(T)/cc-pVTZ levels of theory based on DFT properties with an MAE <1 kcal mol^{-1}. We tested the PBE1PBE,[128, 129] ωB97X,[130] M06-2X,[131] revTPSS,[132] B3LYP,[133, 134] and BP86[135, 136] functionals to ascertain the generalizability of our approach. In this iteration, our model was also trained on dimers to make capturing non-covalent interactions feasible. Using DFT instead of CCSD/X levels of theory computations makes the time and energy saving even more significant compared to our first iteration. Additionally, larger molecules can be predicted when based on DFT computations instead of CCSD/*X* levels of theory due to the beneficial scaling of DFT O(N^4), with the number of basis functions *N*.[137] Our second iteration models were validated with the same set of challenging molecules as the vanilla models (*vide supra*). One example use case of our current model was the $(CH)_{12}$ system.[138-140] Our model was able to predict the relative energy trend and magnitude far better than the B3LYP-D3(BJ)[133]/cc-pVTZ level of theory did compared to the “ground truth” CCSD(T)/cc-pVTZ level of theory. We achieved a five-fold decrease in MAE with our model compared to the stated level of DFT. See Chapter 2.1.2 Machine Learning for Bridging the Gap Between Density Functional Theory and Coupled Cluster Energies for more information.

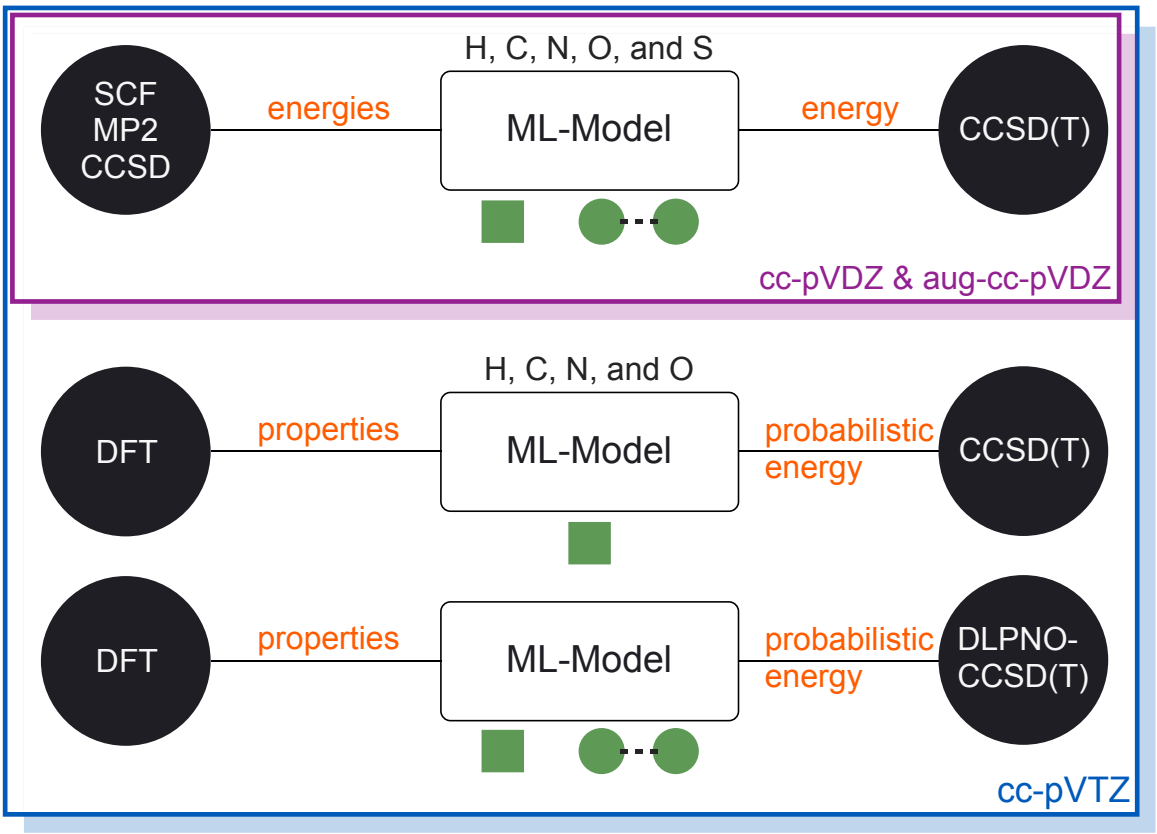


Figure 10. Summary of our recent works regarding the prediction of energies at the CCSD(T) level of theory.[82, 83] The supported basis-set "spaces" are drawn around the respective model. The double ζ basis sets cc-pVDZ and aug-cc-pVDZ in purple, while the triple ζ basis set cc-pVTZ is in blue. The general input levels of theory are shown on the left, and the output levels CCSD(T) and DLPNO-CCSD(T) on the right. The supported elements are listed above the first model in each basis set "space". The green box denotes monomer support, while the connected circles indicate support for non-covalent bound dimer prediction. Probabilistic energy refers to the probabilistic output head of the NN, which predicts a Gaussian distribution with mean μ and variance σ^2.

1.6 Remarks About Current Projects

Besides the computational chemistry related ML projects discussed so far, we also used ML to accelerate the discovery of novel organocatalysts. The first project—the manuscript can be found in Chapter 2.2.1 Rewriting the Rules: Contrasting Historic and Physical Perspectives in Asymmetric Catalysis—discusses how asymmetric catalysis should be modeled correctly. Here we discuss the history of enantiomeric excess (*ee*) and the difference in the Transition State (TS) energy between two competing TS's $\Delta\Delta G^{\ddagger}$. We show that to model asymmetric catalysis, and it is paramount to use the physically meaningful $\Delta\Delta G^{\ddagger}$ domain instead of the more practically friendly *ee* domain. Reasons for this are based on the non-linear relationship

between *ee* and $\Delta\Delta G^{\ddagger}$ paired with the limited physically meaningful space of *ee*, which is limited to $-100 \leq ee \leq 100$. Additionally, $\Delta\Delta G^{\ddagger}$ inherently includes temperature variations, while *ee* does not. Modeling in the *ee* domain might mislead into using a useless model, which we show in a pretty intuitive example when comparing a "good" model that was trained with *ee* and then calculated back to the $\Delta\Delta G^{\ddagger}$ domain. As a preliminary illustration of the potential implications of this case, the example was replicated and is presented in Figure 11.

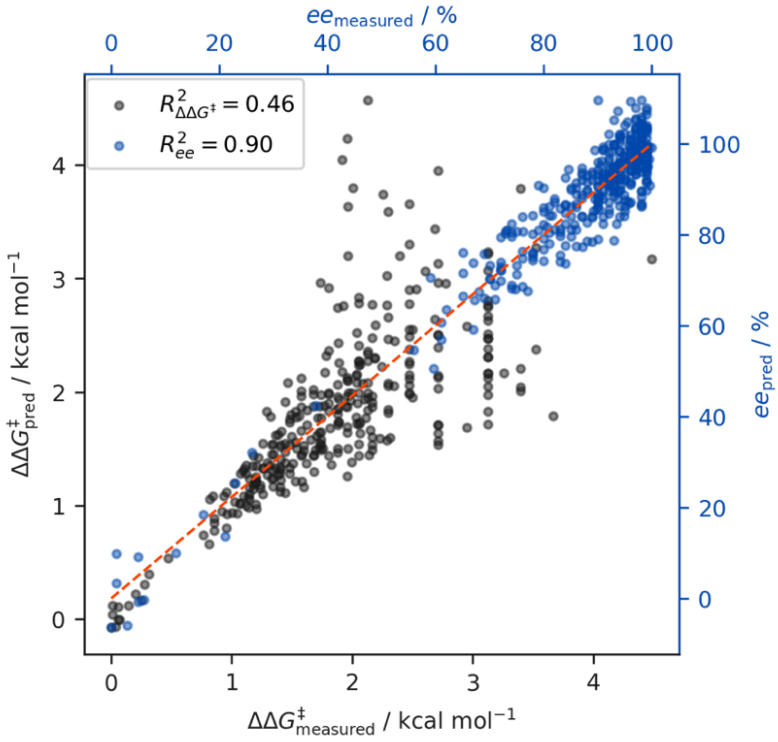


Figure 11. This scatter plot displays data from Sunoj et al.,[141] with predictions made by adding ±5% noise in the *ee* domain (blue) and translating to the $\Delta\Delta G^{\ddagger}$ domain (grey). An orange diagonal line marks optimal prediction. An imaginary model with a 5% MAE in the *ee* domain would be unhelpful, as its poor fit in the $\Delta\Delta G^{\ddagger}$ domain would not indicate a viable model. This is highlighted by the R^2 score difference and the spread of the $\Delta\Delta G^{\ddagger}$ points.

Besides the physically meaningful discussions, we evaluated if it is also beneficial to use $\Delta\Delta G^{\ddagger}$ from a pure metric-oriented perspective or if the inclusion of temperature leads to a different effect when modeling in *ee* or the $\Delta\Delta G^{\ddagger}$ domain.

With the best practices for modeling asymmetric catalysis set, we tackle the Corey-Bakshi-Shibata (CBS) reduction to improve the *ee* for the challenging butanone system. In applying ML, the goal is to address the methyl/ethyl problem, which stems from the challenges posed by the similar stereoelectronic properties of these two groups. The CBS reduction excels with certain prochiral ketones like acetophenone, yielding optimal results.[142] Traditional selectivity models, based solely on steric hindrance, proved insufficient. We

discovered that catalyst optimization requires balancing steric hindrance and London dispersion attraction, a non-trivial task addressed using Design of Experiment (DoE)[143] and computational chemistry.[142] The CBS reduction, sensitive to dispersion energy donors and widely applicable, was our model reaction for this study.[133, 142, 144] With ML, we overcame the current limit of DoE-based approaches,[142] which led to an increase in *ee* from 72% to 80% for the CBS reduction of butanone. In the next iteration of predicted catalysts, we plan to further increase the resulting *ee* in this challenging case. With this work, we show that it is possible to do ML with a small dataset (currently 90 data points) when the quality is sufficient, see Figure 12.

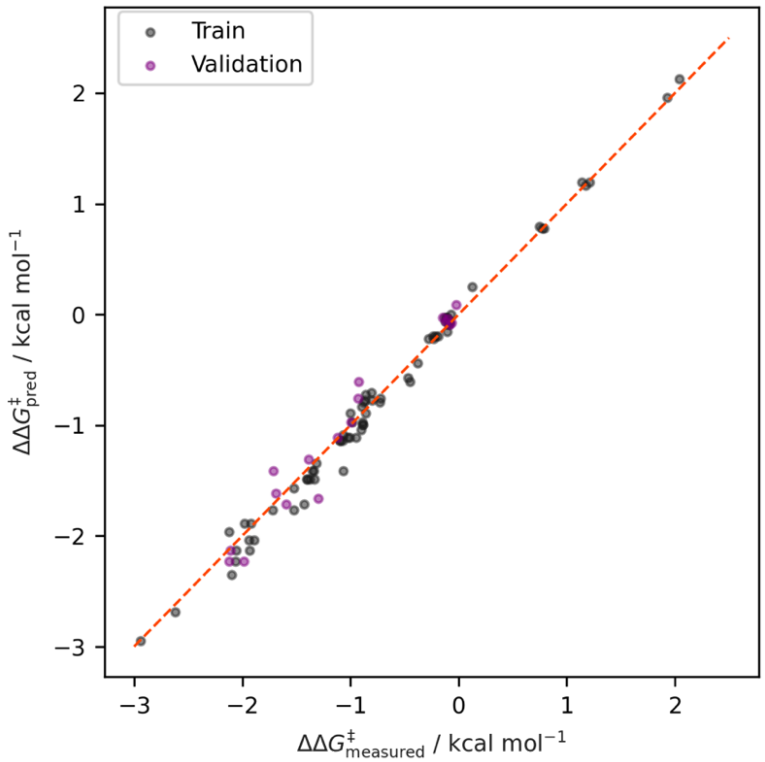


Figure 12. Scatter plot of our current ML model for prediction of $\Delta\Delta G^{\ddagger}$ for the CBS reduction. The orange diagonal indicates optimal prediction. Training points are shown in grey, while validation points are colored purple. On the validation set, we achieved an MAE of 0.027 kcal mol^{-1} with an R^2-score of 0.95.

Initial attempts to use literature data to construct an applicable ML model failed due to issues related to the publication bias, which describes the statistical shift towards “positive” results in the literature because “negative” results are rarely published.[145-147] For a more detailed introduction and additional information, see Chapter 2.2.2 Designing the Next Better Catalyst Utilizing Machine Learning with a *Key-Intermediate* Graph: Differentiating a Methyl from an Ethyl Group.

1.7 References

1. Goodfellow, I.; Bengio, Y.; Courville, A., *Deep Learning*. MIT Press: London, 2016.
2. Neuralink. https://neuralink.com/careers/ (accessed 2023-07-21).
3. Lu, Z. Q. J., The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *J. Roy. Stat. Soc. Ser. A. (Stat. Soc.)* **2010,** *173*, 693–694.
4. Singh, A.; Thakur, N.; Sharma, A. A Review of Supervised Machine Learning Algorithms. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 2016; 2016; pp 1310–1315.
5. Kotsiantis, S. B., Supervised Machine Learning: A Review of Classification Techniques. In *Proceedings of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, IOS Press: 2007; pp 3–24.
6. van Engelen, J. E.; Hoos, H. H., A Survey on Semi-Supervised Learning. *Mach. Learn.* **2020,** *109* (2), 373–440.
7. LeCun, Y.; Bengio, Y.; Hinton, G., Deep Learning. *Nature* **2015,** *521* (7553), 436–444.
8. Rumelhart, D. E.; Hinton, G. E.; Williams, R. J., Learning Representations by Back-Propagating Errors. *Nature* **1986,** *323* (6088), 533–536.
9. Hinton, G. E.; Salakhutdinov, R. R., Reducing the Dimensionality of Data With Neural Networks. *Science* **2006,** *313* (5786), 504–507.
10. Domingos, P., A Few Useful Things to Know About Machine Learning. *Commun. ACM* **2012,** *55* (10), 78–87.
11. Rasmussen, C. E., Gaussian Processes in Machine Learning. In *Advanced Lectures on Machine Learning*, Springer Berlin Heidelberg, 2004; pp 63–71.
12. Bottou, L. Large-Scale Machine Learning With Stochastic Gradient Descent. In *Proceedings of COMPSTAT'2010*, Heidelberg, 2010; Lechevallier, Y.; Saporta, G., Eds. Physica-Verlag HD: Heidelberg, 2010; pp 177–186.
13. Kingma, D. P.; Ba, J., Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv* **2014**.
14. Loshchilov, I., Decoupled Weight Decay Regularization. *arXiV preprint arXiv* **2017**.
15. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steinerand, B.; Fang, L.; Bai, J.; Chintala, S., PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Adv. Neural Inf. Process. Syst.* **2019,** *32*.
16. Katoch, S.; Chauhan, S. S.; Kumar, V., A Review on Genetic Algorithm: Past, Present, and Future. *Multimed. Tools Appl.* **2021,** *80* (5), 8091–8126.
17. Na, W.; Yan, S.; Feng, F.; Liu, W.; Zhu, L.; Zhang, Q., Recent Advances In Knowledge-Based Model Structure Optimization and Extrapolation Techniques for Microwave Applications. *Int. J. Numer. Model.: Electron. Netw. Devices Fields* **2021**.
18. Bronstein, M. M.; Bruna, J.; LeCun, Y.; Szlam, A.; Vandergheynst, P., Geometric Deep Learning: Going Beyond Euclidean Data. *ISPM* **2017,** *34* (4), 18–42.
19. McCulloch, W. S.; Pitts, W., A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bull. math. biophys.* **1943,** *5* (4), 115–133.

20. Nair, V.; Hinton, G. E., Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on Machine Learning*, Fürnkranz, J.; Joachims, T., Eds. Omnipress: Haifa, Israel, 2010; pp 807–814.
21. Lin, M.; Chen, Q.; Yan, S., Network In Network. *arXiv preprint arXiv* **2013**.
22. Kipf, T. N.; Welling, M., Semi-Supervised Classification With Graph Convolutional Networks. *arXiv preprint arXiv* **2016**.
23. Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E., Neural Message Passing for Quantum Chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, Precup, D.; Teh, Y. W., Eds. JMLR: Sydney, Australia, 2017; Vol. 70, pp 1263–1272.
24. 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* (5), 058301.
25. Behler, J., Atom-Centered Symmetry Functions for Constructing High-Dimensional Neural Network Potentials. *J. Chem. Phys.* **2011,** *134* (7), 074106.
26. Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K.-R.; Tkatchenko, A., Quantum-Chemical Insights From Deep Tensor Neural Networks. *Nat. Commun.* **2017,** *8*, 13890.
27. Chmiela, S.; Sauceda, H. E.; Müller, K.-R.; Tkatchenko, A., Towards Exact Molecular Dynamics Simulations With Machine-Learned Force Fields. *Nat. Commun.* **2018,** *9*, 3887.
28. Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E., Less Is More: Sampling Chemical Space With Active Learning. *J. Chem. Phys.* **2018,** *148*, 241733.
29. Gastegger, M.; Schwiedrzik, L.; Bittermann, M. R.; Berzsenyi, F.; Marquetand, P., wACSF—Weighted Atom-Centered Symmetry Functions as Descriptors in Machine Learning Potentials. *J. Chem. Phys.* **2018,** *148*, 241709.
30. Smith, J. S.; Isayev, O.; Roitberg, A. E., ANI-1: An Extensible Neural Network Potential With DFT Accuracy at Force Field Computational Cost. *Chem. Sci.* **2017,** *8*, 3192–3203.
31. Gastegger, M.; Marquetand, P., High-Dimensional Neural Network Potentials for Organic Reactions and an Improved Training Algorithm. *J. Chem. Theory Comput.* **2015,** *11* (5), 2187–2198.
32. Behler, J.; Parrinello, M., Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007,** *98* (14), 146401.
33. Zhang, L.; Han, J.; Wang, H.; Car, R.; E, W., Deep Potential Molecular Dynamics: A Scalable Model With the Accuracy of Quantum Mechanics. *Phys. Rev. Lett.* **2018,** *120* (14), 143001.
34. Wolf, M. J. P.; Norris, J. W.; Fynewever, H.; Turney, J. M.; Schaefer, H. F., An Undergraduate Chemistry Lab Exploring Computational Cost and Accuracy: Methane Combustion Energy. *J. Chem. Educ.* **2022,** *99* (3), 1479–1487.
35. Hocquet, A.; Wieber, F., “Only the Initiates Will Have the Secrets Revealed”: Computational Chemists and the Openness of Scientific Software. *IEEE Ann. Hist. Comput.* **2017,** *39* (4), 40–58.
36. Hu, W.; Chen, M., Editorial: Advances in Density Functional Theory and Beyond for Computational Chemistry. *Frontiers in Chemistry* **2021,** *9*.
37. Cheng, G.-J.; Zhang, X.; Chung, L. W.; Xu, L.; Wu, Y.-D., Computational Organic Chemistry: Bridging Theory and Experiment in Establishing the Mechanisms of Chemical Reactions. *J. Am. Chem. Soc.* **2015,** *137* (5), 1706–1725.

38. Kiss, G.; Çelebi-Ölçüm, N.; Moretti, R.; Baker, D.; Houk, K. N., Computational Enzyme Design. *Angew. Chem. Int. Ed.* **2013,** *52* (22), 5700–5725.
39. Hafner, J., Atomic-Scale Computational Materials Science. *Acta Materialia* **2000,** *48* (1), 71–92.
40. Miller, C. L.; Ellison, M. D., Walsh Diagrams: Molecular Orbital and Structure Computational Chemistry Exercise for Physical Chemistry. *J. Chem. Educ.* **2015,** *92* (6), 1040–1043.
41. Bajorath, J., Pushing the Boundaries of Computational Approaches: Special Focus Issue on Computational Chemistry and Computer-Aided Drug Discovery. *Future Med. Chem.* **2015,** *7* (18), 2415–2417.
42. Hinde, R. J., Quantum Chemistry, 5th Edition (by Ira N. Levine). *J. Chem. Educ.* **2000,** *77* (12), 1564.
43. Čížek, J., On the Correlation Problem in Atomic and Molecular Systems. Calculation of Wavefunction Components in Ursell-Type Expansion Using Quantum-Field Theoretical Methods. *J. Chem. Phys.* **1966,** *45* (11), 4256–4266.
44. Hohenberg, P.; Kohn, W., Inhomogeneous Electron Gas. *Phys. Rev.* **1964,** *136* (3B), 864–871.
45. Kohn, W.; Sham, L. J., Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965,** *140* (4A), 1133–1138.
46. Bartlett, R. J.; Musiał, M., Coupled-Cluster Theory in Quantum Chemistry. *Rev. Mod. Phys.* **2007,** *79* (1), 291–352.
47. Jr., T. H. D., Gaussian Basis Sets for Use in Correlated Molecular Calculations. I. The Atoms Boron Through Neon and Hydrogen. *J. Chem. Phys.* **1989,** *90* (2), 1007–1023.
48. Crawford, T. D.; Schaefer III, H. F., An Introduction to Coupled Cluster Theory for Computational Chemists. In 2007; pp 33–136.
49. Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M., A Fifth-Order Perturbation Comparison of Electron Correlation Theories. *Chem. Phys. Lett.* **1989,** *157* (6), 479–483.
50. Nam, S.; McCarty, R. J.; Park, H.; Sim, E., KS-pies: Kohn–Sham Inversion Toolkit. *J. Chem. Phys.* **2021,** *154* (12).
51. Computation Sparks Chemical Discovery. *Nat. Commun.* **2020,** *11* (1), 4811.
52. Stanton, J. F.; Gauss, J.; Watts, J. D.; Bartlett, R. J., A Direct Product Decomposition Approach for Symmetry Exploitation in Many-Body Methods. I. Energy Calculations. *J. Chem. Phys.* **1991,** *94* (6), 4334–4345.
53. Scuseria, G. E.; Scheiner, A. C.; Lee, T. J.; Rice, J. E.; Schaefer III, H. F., The Closed-Shell Coupled Cluster Single and Double Excitation (CCSD) Model for the Description of Electron Correlation. A Comparison With Configuration Interaction (CISD) Results. *J. Chem. Phys.* **1987,** *86* (5), 2881–2890.
54. Purvis III, G. D.; Bartlett, R. J., A Full Coupled-Cluster Singles and Doubles Model: The Inclusion of Disconnected Triples. *J. Chem. Phys.* **1982,** *76* (4), 1910–1918.
55. Hampel, C.; Peterson, K. A.; Werner, H.-J., A Comparison of the Efficiency and Accuracy of the Quadratic Configuration Interaction (QCISD), Coupled Cluster (CCSD), and Brueckner Coupled Cluster (BCCD) Methods. *Chem. Phys. Lett.* **1992,** *190* (1), 1–12.
56. Noga, J.; Bartlett, R. J., The Full CCSDT Model for Molecular Electronic Structure. *J. Chem. Phys.* **1987,** *86* (12), 7041–7050.

57. Noga, J.; Bartlett, R. J., Erratum: The Full CCSDT Model for Molecular Electronic Structure [J. Chem. Phys. 86, 7041 (1987)]. *J. Chem. Phys.* **1988,** *89* (5), 3401–3401.
58. Oliphant, N.; Adamowicz, L., Coupled-Cluster Method Truncated at Quadruples. *J. Chem. Phys.* **1991,** *95* (9), 6645–6651.
59. Kucharski, S. A.; Bartlett, R. J., The Coupled-Cluster Single, Double, Triple, and Quadruple Excitation Method. *J. Chem. Phys.* **1992,** *97* (6), 4282–4288.
60. Kucharski, S. A.; Bartlett, R. J., Recursive Intermediate Factorization and Complete Computational Linearization of the Coupled-Cluster Single, Double, Triple, and Quadruple Excitation Equations. *Theor. Chim. Acta* **1991,** *80* (4), 387–405.
61. Stanton, J. F., Why CCSD(T) Works: A Different Perspective. *Chem. Phys. Lett.* **1997,** *281* (1), 130–134.
62. Bartlett, R. J.; Watts, J. D.; Kucharski, S. A.; Noga, J., Non-Iterative Fifth-Order Triple and Quadruple Excitation Energy Corrections in Correlated Methods. *Chem. Phys. Lett.* **1990,** *165* (6), 513–522.
63. Eriksen, J. J.; Baudin, P.; Ettenhuber, P.; Kristensen, K.; Kjærgaard, T.; Jørgensen, P., Linear-Scaling Coupled Cluster with Perturbative Triple Excitations: The Divide–Expand–Consolidate CCSD(T) Model. *J. Chem. Theory Comput.* **2015,** *11* (7), 2984–2993.
64. Li, W.; Piecuch, P.; Gour, J. R.; Li, S., Local Correlation Calculations Using Standard and Renormalized Coupled-Cluster Approaches. *J. Chem. Phys.* **2009,** *131* (11), 114109.
65. Riplinger, C.; Neese, F., An Efficient and Near Linear Scaling Pair Natural Orbital Based Local Coupled Cluster Method. *J. Chem. Phys.* **2013,** *138* (3), 034106.
66. Liakos, D. G.; Guo, Y.; Neese, F., Comprehensive Benchmark Results for the Domain Based Local Pair Natural Orbital Coupled Cluster Method (DLPNO-CCSD(T)) for Closed- and Open-Shell Systems. *J. Phys. Chem. A* **2020,** *124* (1), 90–100.
67. Guo, Y.; Riplinger, C.; Becker, U.; Liakos, D. G.; Minenkov, Y.; Cavallo, L.; Neese, F., Communication: An Improved Linear Scaling Perturbative Triples Correction for the Domain Based Local Pair-Natural Orbital Based Singles and Doubles Coupled Cluster Method [DLPNO-CCSD(T)]. *J. Chem. Phys.* **2018,** *148* (1), 011101.
68. Ochsenfeld, C.; Kussmann, J.; Lambrecht, D. S., Linear-Scaling Methods in Quantum Chemistry. In *Rev. Comput. Chem.*, 2007; pp 1–82.
69. Bowler, D. R.; Miyazaki, T., O(*N*) Methods in Electronic Structure Calculations. *Rep. Prog. Phys.* **2012,** *75* (3), 036503.
70. Hansen, K.; Montavon, G.; Fazli, S.; Rupp, M.; Scheffler, M.; von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K.-R., Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies. *J. Chem. Theory Comput.* **2013,** *9* (8), 3404–3419.
71. Schütt, K. T.; Gastegger, M.; Tkatchenko, A.; Müller, K.-R.; Maurer, R. J., Unifying Machine Learning and Quantum Chemistry With a Deep Neural Network for Molecular Wavefunctions. *Nat. Commun.* **2019,** *10*, 5024.
72. Brockherde, F.; Vogt, L.; Li, L.; Tuckerman, M. E.; Burke, K.; Müller, K.-R., Bypassing the Kohn-Sham Equations With Machine Learning. *Nat. Commun.* **2017,** *8*, 872.
73. Grisafi, A.; Wilkins, D. M.; Csányi, G.; Ceriotti, M., Symmetry-Adapted Machine Learning for Tensorial Properties of Atomistic Systems. *Phys. Rev. Lett.* **2018,** *120* (3), 036002.

74. Jha, D.; Choudhary, K.; Tavazza, F.; Liao, W.-k.; Choudhary, A.; Campbell, C. E.; Agrawal, A., Enhancing Materials Property Prediction by Leveraging Computational and Experimental Data Using Deep Transfer Learning. *Nat. Commun.* **2019,** (10), 5316.
75. Gong, S. K.; Wang, S.; Xie, T.; Chae, W. R.; Liu, R.-Z.; Grossman, J. C.; Grossman, J. C., Calibrating DFT Formation Enthalpy Calculations by Multifidelity Machine Learning. *J. Am. Chem. Soc. Au* **2022,** *2* (9), 1964–1977.
76. Lee, S.-H.; Shostak, S. N.; Filatov, M.; Choi, C. H., Conical Intersections in Organic Molecules: Benchmarking Mixed-Reference Spin–Flip Time-Dependent DFT (MRSF-TD-DFT) vs Spin–Flip TD-DFT. *J. Phys. Chem. A* **2019,** *123* (30), 6455–6462.
77. Babuji, Y.; Woodard, A.; Li, Z.; Katz, D. S.; Clifford, B.; Kumar, R.; Lacinski, L.; Chard, R.; Wozniak, J. M.; Foster, I.; Wilde, M.; Chard, K., Parsl: Pervasive Parallel Programming in Python. *arXiv preprint arXiv* **2019**.
78. Townsend, J.; Vogiatzis, K. D., Data-Driven Acceleration of the Coupled-Cluster Singles and Doubles Iterative Solver. *J. Phys. Chem. Lett.* **2019,** *10* (14), 4129–4135.
79. Townsend, J.; Vogiatzis, K. D., Transferable MP2-Based Machine Learning for Accurate Coupled-Cluster Energies. *J. Chem. Theory Comput.* **2020,** *16* (12), 7453–7461.
80. Herzog, B.; Gallo, A.; Hummel, F.; Badawi, M.; Bučko, T.; Lebègue, S.; Grüneis, A.; Rocca, D., Coupled Cluster Finite Temperature Simulations of Periodic Materials via Machine Learning. *ChemRxiv preprint ChemRXiv* **2023**.
81. Chen, M. S.; Lee, J.; Ye, H.-Z.; Berkelbach, T. C.; Reichman, D. R.; Markland, T. E., Data-Efficient Machine Learning Potentials From Transfer Learning of Periodic Correlated Electronic Structure Methods: Liquid Water at AFQMC, CCSD, and CCSD(T) Accuracy. *J. Chem. Theory Comput.* **2023,** *19* (14).
82. Ruth, M.; Gerbig, D.; Schreiner, P. R., Machine Learning for Bridging the Gap Between Density Functional Theory and Coupled Cluster Energies. *J. Chem. Theory Comput.* **2023,** *19* (15), 4912–4920.
83. 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* (8), 4846–4855.
84. Hörst, S. M., Titan's Atmosphere and Climate. *J. Geophys. Res. Planets* **2017,** *122*, 432–482.
85. Csontos, J.; Rolik, Z.; Das, S.; Kállay, M., High-Accuracy Thermochemistry of Atmospherically Important Fluorinated and Chlorinated Methane Derivatives. *J. Phys. Chem. A* **2010,** *114* (50), 13093–13103.
86. Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Lilienfeld, O. A. v., Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014,** *1*, 140022.
87. 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* (23), 11599–11613.
88. 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* (27), 5826–5835.
89. Goldsmith, C. F.; Magoon, G. R.; Green, W. H., Database of Small Molecule Thermochemistry for Combustion. *J. Phys. Chem. A* **2012,** *116* (36), 9033–9057.
90. Mellouki, A.; Ammann, M.; Cox, R. A.; Crowley, J. N.; Herrmann, H.; Jenkin, M. E.; McNeill, V. F.; Troe, J.; Wallington, T. J., Evaluated Kinetic and Photochemical Data for Atmospheric Chemistry: Volume VIII – Gas-Phase Reactions of Organic

Species With Four, or More, Carbon Atoms (≥ C_4). *Atmos. Chem. Phys.* **2021,** *21* (6), 4797–4808.
91. Cox, R. A.; Ammann, M.; Crowley, J. N.; Herrmann, H.; Jenkin, M. E.; McNeill, V. F.; Mellouki, A.; Troe, J.; Wallington, T. J., Evaluated Kinetic and Photochemical Data for Atmospheric Chemistry: Volume VII – Criegee Intermediates. *Atmos. Chem. Phys.* **2020,** *20* (21), 13497–13519.
92. Cleaves, H. J., Prebiotic Chemistry: What We Know, What We Don't. *Evol.: Educ. Outreach* **2012,** *5* (3), 342–360.
93. Arumainayagam, C. R.; Garrod, R. T.; Boyer, M. C.; Hay, A. K.; Bao, S. T.; Campbell, J. S.; Wang, J.; Nowak, C. M.; Arumainayagam, M. R.; Hodge, P. J., Extraterrestrial Prebiotic Molecules: Photochemistry vs. Radiation Chemistry of Interstellar Ices. *Chem. Soc. Rev.* **2019,** *48* (8), 2293–2314.
94. Dunjko, V.; Briegel, H. J., Machine Learning & Artificial Intelligence in the Quantum Domain: A Review of Recent Progress. *Rep. Prog. Phys.* **2018,** *81* (7), 074001.
95. Lusci, A.; Pollastri, G.; Baldi, P., Deep Architectures and Deep Learning in Chemoinformatics: The Prediction of Aqueous Solubility for Drug-Like Molecules. *J. Chem. Inf. Model.* **2013,** *53* (7), 1563–1575.
96. Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T., The Rise of Deep Learning in Drug Discovery. *Drug Discov. Today* **2018,** *23* (6), 1241–1250.
97. Kocer, E.; Ko, T. W.; Behler, J., Neural Network Potentials: A Concise Overview of Methods. *Annu. Rev. Phys. Chem.* **2022,** *73* (1), 163–186.
98. Gastegger, M.; Behler, J.; Marquetand, P., Machine Learning Molecular Dynamics for the Simulation of Infrared Spectra. *Chem. Sci.* **2017,** *8* (10), 6924–6935.
99. Behler, J., Constructing High-Dimensional Neural Network Potentials: A Tutorial Review. *Int. J. Quantum Chem.* **2015,** *115* (16), 1032–1050.
100. Gao, X.; Ramezanghorbani, F.; Isayev, O.; Smith, J. S.; Roitberg, A. E., TorchANI: A Free and Open Source PyTorch-Based Deep Learning Implementation of the ANI Neural Network Potentials. *J. Chem. Inf. Model.* **2020,** *60* (7), 3408–3415.
101. Rossum, G. V.; Drake, F. L., Python 3 Reference Manual. *Python 3 Reference Manual.* **2009**.
102. Behler, J., Four Generations of High-Dimensional Neural Network Potentials. *Chem. Rev.* **2021,** *121* (16), 10037–10072.
103. Sumpter, B. G.; Noid, D. W., Potential Energy Surfaces for Macromolecules. A Neural Network Technique. *Chem. Phys. Lett.* **1992,** *192* (5), 455–462.
104. Podryabinkin, E. V.; Shapeev, A. V., Active Learning of Linearly Parametrized Interatomic Potentials. *Comput. Mater. Sci.* **2017,** *140*, 171–180.
105. Perdew, J. P.; Schmidt, K., Jacob's Ladder of Density Functional Approximations for the Exchange-Correlation Energy. *AIP Conf. Proc.* **2001,** *577* (1), 1–20.
106. Bellman, R., *Dynamic Programming*. Princeton University Press: 1957.
107. Aggarwal, C. C.; Hinneburg, A.; Keim, D. A. On the Surprising Behavior of Distance Metrics in High Dimensional Space. In Berlin, Heidelberg, 2001; Springer Berlin Heidelberg: Berlin, Heidelberg, 2001; pp 420–434.
108. Indyk, P.; Motwani, R., Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, Association for Computing Machinery: Dallas, Texas, USA, 1998; pp 604–613.
109. Pearson, K., LIII. On Lines and Planes of Closest Fit to Systems of Points in Space. *Lond. Edinb. Dublin philos. mag. j. sci.* **1901,** *2* (11), 559–572.

110. Maaten, L. v. d.; Hinton, G., Visualizing Data Using t-SNE. *J. Mach. Learn.* **2008,** *9* (86), 2579–2605.
111. Bing Image Creator. https://www.bing.com/create (accessed 2023-08-01).
112. He, K.; Zhang, X.; Ren, S.; Sun, J., Deep Residual Learning for Image Recognition. *arXiv preprint arXiv* **2016**.
113. 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* (5), 2087–2096.
114. Ramakrishnan, R.; Hartmann, M.; Tapavicza, E.; von Lilienfeld, O. A., Electronic Spectra From TDDFT and Machine Learning in Chemical Space. *J. Chem. Phys.* **2015,** *143* (8), 084111.
115. Li, P.; Jia, X.; Pan, X.; Shao, Y.; Mei, Y., Accelerated Computation of Free Energy Profile at ab Initio Quantum Mechanical/Molecular Mechanics Accuracy via a Semi-Empirical Reference Potential. I. Weighted Thermodynamics Perturbation. *J. Chem. Theory Comput.* **2018,** *14* (11), 5583–5596.
116. Zhang, P.; Shen, L.; Yang, W., Solvation Free Energy Calculations With Quantum Mechanics/Molecular Mechanics and Machine Learning Models. *J. Phys. Chem. B* **2019,** *123* (4), 901–908.
117. Böselt, L.; Thürlemann, M.; Riniker, S., Machine Learning in QM/MM Molecular Dynamics Simulations of Condensed-Phase Systems. *J. Chem. Theory Comput.* **2021,** *17* (5), 2641–2658.
118. Song, K.; Li, J., The Neural Network Based Δ-Machine Learning Approach Efficiently Brings the DFT Potential Energy Surface to the CCSD(T) Quality: a Case for the OH + CH_3OH Reaction. *Phys. Chem. Chem. Phys.* **2023,** *25* (16), 11192–11204.
119. Kleine Büning, J. B.; Grimme, S., Computation of CCSD(T)-Quality NMR Chemical Shifts via Δ-Machine Learning from DFT. *J. Chem. Theory Comput.* **2023,** *19* (12), 3601–3615.
120. Zhao, Q.; Anstine, D. M.; Isayev, O.; Savoie, B. M., Δ^2 Machine Learning for Reaction Property Prediction. *Chem. Sci.* **2023**.
121. Slater, J. C., A Simplification of the Hartree-Fock Method. *Phys. Rev.* **1951,** *81* (3), 385–390.
122. Sæbø, S.; Almlöf, J., Avoiding the Integral Storage Bottleneck in LCAO Calculations of Electron Correlation. *Chem. Phys. Lett.* **1989,** *154* (1), 83–89.
123. Head-Gordon, M.; Pople, J. A.; Frisch, M. J., MP2 Energy Evaluation by Direct Methods. *Chem. Phys. Lett.* **1988,** *153* (6), 503–506.
124. Head-Gordon, M.; Head-Gordon, T., Analytic MP2 Frequencies Without Fifth-Order Storage. Theory and Application to Bifurcated Hydrogen Bonds in the Water Hexamer. *Chem. Phys. Lett.* **1994,** *220* (1), 122–128.
125. Frisch, M. J.; Head-Gordon, M.; Pople, J. A., Semi-Direct Algorithms for the MP2 Energy and Gradient. *Chem. Phys. Lett.* **1990,** *166* (3), 281–289.
126. Frisch, M. J.; Head-Gordon, M.; Pople, J. A., A Direct MP2 Gradient Method. *Chem. Phys. Lett.* **1990,** *166* (3), 275–280.
127. Kendall, R. A.; Jr., T. H. D.; Harrison, R. J., Electron Affinities of the First-Row Atoms Revisited. Systematic Basis Sets and Wave Functions. *J. Chem. Phys.* **1992,** *96* (9), 6796–6806.
128. Ernzerhof, M.; Scuseria, G. E., Assessment of the Perdew–Burke–Ernzerhof Exchange-Correlation Functional. *J. Chem. Phys.* **1999,** *110* (11), 5029–5036.
129. Adamo, C.; Barone, V., Toward Reliable Density Functional Methods Without Adjustable Parameters: The PBE0 Model. *J. Chem. Phys.* **1999,** *110* (13), 6158–6170.

130. Chai, J.-D.; Head-Gordon, M., Systematic Optimization of Long-Range Corrected Hybrid Density Functionals. *J. Chem. Phys.* **2008,** *128* (8), 084106.
131. 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* (1), 215–241.
132. Perdew, J. P.; Ruzsinszky, A.; Csonka, G. I.; Constantin, L. A.; Sun, J., Workhorse Semilocal Density Functional for Condensed Matter Physics and Quantum Chemistry. *Phys. Rev. Lett.* **2009,** *103* (2), 026403.
133. Grimme, S.; Ehrlich, S.; Goerigk, L., Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011,** *32* (7), 1456–1465.
134. Becke, A. D., A New Mixing of Hartree–Fock and Local Density-Functional Theories. *J. Chem. Phys.* **1993,** *98* (2), 1372–1377.
135. Perdew, J. P., Density-Functional Approximation for the Correlation Energy of the Inhomogeneous Electron Gas. *Phys. Rev. B* **1986,** *33* (12), 8822–8824.
136. 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* (2), 785–789.
137. Goerigk, L.; Grimme, S., A Thorough Benchmark of Density Functional Methods for General Main Group Thermochemistry, Kinetics, and Noncovalent Interactions. *Phys. Chem. Chem. Phys.* **2011,** *13* (14), 6670–6688.
138. Schreiner, P. R.; Fokin, A. A.; Pascal, R. A.; de Meijere, A., Many Density Functional Theory Approaches Fail To Give Reliable Large Hydrocarbon Isomer Energy Differences. *Org. Lett.* **2006,** *8* (17), 3635–3638.
139. Karton, A.; Schreiner, P. R.; Martin, J. M. L., Heats of Formation of Platonic Hydrocarbon Cages by Means of High-Level Thermochemical Procedures. *J. Comput. Chem.* **2016,** *37* (1), 49–58.
140. de Meijere, A.; Lee, C.-H.; Kuznetsov, M. A.; Gusev, D. V.; Kozhushkov, S. I.; Fokin, A. A.; Schreiner, P. R., Preparation and Reactivity of [D_{3d}]-Octahedrane: The Most Stable $(CH)_{12}$ Hydrocarbon. *Chem. Eur. J.* **2005,** *11* (21), 6175–6184.
141. Singh, S.; Pareek, M.; Changotra, A.; Banerjee, S.; Bhaskararao, B.; Balamurugan, P.; Sunoj, R. B., A Unified Machine-Learning Protocol for Asymmetric Catalysis as a Proof of Concept Demonstration Using Asymmetric Hydrogenation. *PNAS* **2020,** *117* (3), 1339–1345.
142. Eschmann, C.; Song, L.; Schreiner, P. R., London Dispersion Interactions Rather than Steric Hindrance Determine the Enantioselectivity of the Corey–Bakshi–Shibata Reduction. *Angew. Chem. Int. Ed.* **2021,** *60* (9), 4823–4832.
143. Fisher, R. A., *The Design of Experiments*. Oliver & Boyd: Oxford, England, 1935.
144. Wagner, J. P.; Schreiner, P. R., London Dispersion in Molecular Chemistry—Reconsidering Steric Effects. *Angew. Chem. Int. Ed.* **2015,** *54* (42), 12274–12296.
145. Griffiths, R.-R.; Schwaller, P.; Lee, A., Dataset Bias in the Natural Sciences: A Case Study in Chemical Reaction Prediction and Synthesis Design. *ChemRxiv preprint ChemRxiv* **2018**.
146. Jia, X.; Lynch, A.; Huang, Y.; Danielson, M.; Lang'at, I.; Milder, A.; Ruby, A. E.; Wang, H.; Friedler, S. A.; Norquist, A. J.; Schrier, J., Anthropogenic Biases in Chemical Reaction Data Hinder Exploratory Inorganic Synthesis. *Nature* **2019,** *573* (7773), 251–255.

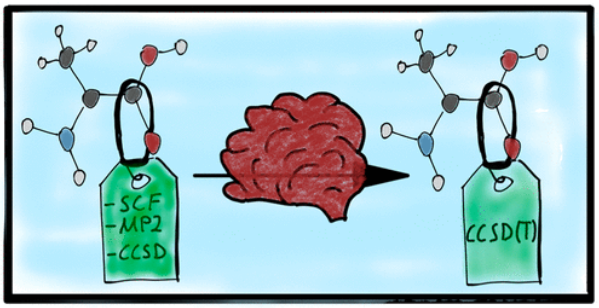
147. Strieth-Kalthoff, F.; Sandfort, F.; Kühnemund, M.; Schäfer, F. R.; Kuchen, H.; Glorius, F., Machine Learning for Chemical Reactivity: The Importance of Failed Experiments. *Angew. Chem. Int. Ed.* **2022,** *61* (29), e202204647.

2. Publications

2.1 Peer-Reviewed Publications

2.1.1 Machine Learning of Coupled Cluster (T)-Energy Corrections via Delta (Δ)-Learning

Abstract:



Accurate thermochemistry is essential in many chemical disciplines, such as astro-, atmospheric, or combustion chemistry. These areas often involve fleetingly existent intermediates whose thermo-chemistry is difficult to assess. Whenever direct calorimetric experiments are infeasible, accurate computational estimates of relative molecular energies are required. However, high-level computations, often using coupled cluster theory, are generally resource-intensive. To expedite the process using machine learning techniques, we generated a database of energies for small organic molecules at the CCSD(T)/cc-pVDZ, CCSD(T)/aug-cc-pVDZ, and CCSD(T)/cc-pVTZ levels of theory. Leveraging the power of deep learning by employing graph neural networks, we are able to predict the effect of perturbatively included triples (T), that is, the difference between CCSD and CCSD(T) energies, with a mean absolute error of 0.25, 0.25, and 0.28 kcal mol–1 (R^2 of 0.998, 0.997, and 0.998) with the cc-pVDZ, aug-cc-pVDZ, and cc-pVTZ basis sets, respectively. Our models were further validated by application to three validation sets taken from the S22 Database as well as to a selection of known theoretically challenging cases.

Highlighted: (DOI: 10.1002/nadc.20224131471)
Notizen aus der Chemie. Nachrichten aus der Chemie **2022**, *70* (10), 48–51.

Reference: (DOI: 10.1021/acs.jctc.2c00501)
M. Ruth, D. Gerbig, P. R. Schreiner, Machine Learning of Coupled Cluster (T)-Energy Corrections via Delta (Δ)-Learning, *J. Chem. Theory Comput*. **2022**, *18*, 8, 4846–4855.

JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

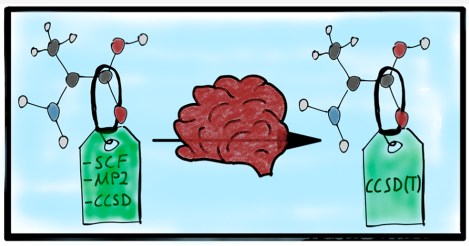
Machine Learning of Coupled Cluster (T)-Energy Corrections via Delta (Δ)-Learning

Marcel Ruth, Dennis Gerbig, and Peter R. Schreiner*

Cite This: J. Chem. Theory Comput. 2022, 18, 4846−4855 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Accurate thermochemistry is essential in many chemical disciplines, such as astro-, atmospheric, or combustion chemistry. These areas often involve fleetingly existent intermediates whose thermochemistry is difficult to assess. Whenever direct calorimetric experiments are infeasible, accurate computational estimates of relative molecular energies are required. However, high-level computations, often using coupled cluster theory, are generally resource-intensive. To expedite the process using machine learning techniques, we generated a database of energies for small organic molecules at the CCSD(T)/cc-pVDZ, CCSD(T)/aug-cc-pVDZ, and CCSD(T)/cc-pVTZ levels of theory. Leveraging the power of deep learning by employing graph neural networks, we are able to predict the effect of perturbatively included triples (T), that is, the difference between CCSD and CCSD(T) energies, with a mean absolute error of 0.25, 0.25, and 0.28 kcal mol^{-1} (R^2 of 0.998, 0.997, and 0.998) with the cc-pVDZ, aug-cc-pVDZ, and cc-pVTZ basis sets, respectively. Our models were further validated by application to three validation sets taken from the S22 Database as well as to a selection of known theoretically challenging cases.



INTRODUCTION

Machine learning (ML) is a way to utilize computational resources to predict, understand, and interpret data. In the field of ML, computers can learn without being explicitly programmed.[1] ML in chemistry is an emerging technique that has already been successfully applied to predict various physical, chemical, and materials properties.[2−4] Material properties are often described with continuum mechanics, to which ML can be applied to discover and develop new materials.[5] The recent excitement about ML is due to the enormous increase in computing power by accelerators, such as graphical processing units, and access to more complex and more extensive data sets.[6]

The prediction of coupled cluster[7] (CC) energies and molecular properties has been accomplished in various ways. ML-assisted approaches encompass the training of force-fields from ab initio data, prediction of CC amplitudes, and direct energy or property learning using mean-field, correlated, or methods based on density functional theory (DFT).[8] Using DFT densities, it is possible to predict CC energies by leveraging ML:[9] With an iterative approach, Townsend and Vogiatzis et al. were able to predict the converged CC amplitudes—hence the CC wave function—by utilizing theoretical properties inherent to Møller−Plesset perturbation theory.[10,11] Another iterative hybrid approach by Maitra et al. divides the amplitudes into significant and less significant contributions and reduces computational time without loss of accuracy.[12] Often, so-called Δ-learning is employed, in which the objective is not the prediction of the total energy, but rather that of an increment or difference between property values determined at low and high levels of theory.[13] The recent study by Nandi et al. presents an example for Δ-learning of potential energy surfaces (PES) from the DFT to the CCSD(T) levels of theory. The authors applied this approach to H_3O^+, CH_4, and *N*-methylacetamide.[14] Predicting CCSD(T) results from DFT can also be accomplished via general-purpose neural network (NN) potentials. The accuracy of this approach for thermochemistry, isomerization energies, and molecular torsion potentials compares favorably to complete basis set extrapolations.[15] Besides amplitudes and energies, approaches exist for predicting accurate anharmonic frequencies at the CCSD(T) level of theory, based on transfer learning from a low-level theory, such as MP2.[16] Besides molecular energies and frequencies, it is possible to compute accurate molecular polarizabilities using CC theory.[17]

Classic (non-ML) computational chemistry today is mostly dominated by DFT[18] due to its $O(N^3)$ to $O(N^4)$ scaling behavior with the number of basis functions N, which often

Received: May 12, 2022
Published: July 11, 2022

https://doi.org/10.1021/acs.jctc.2c00501
J. Chem. Theory Comput. 2022, 18, 4846−4855

makes DFT a reasonable choice from a performance and stability perspective.[19,20] Among the shortcomings of DFT are, for example, the lack of a proper description of dispersion interactions and the fact that the exact exchange−correlation energy functional is generally unknown.[21] These effects can lead to significant errors, for example, for thermochemistry of simple hydrocarbons[22,23] and optical spectra of transition-metal complexes.[20,24−26] Due to the stated difficulties, it is desirable to use more sophisticated methods, ideally such based on CC theory, to approximate the overall electron correlation. The included excitations depend on the truncation of the cluster operator. Including more excitations would result in a more accurate, but also more resource-intensive computation. Including only singles and doubles (CCSD) scales as $O(N^6)$, and it generally accepted that triples corrections (i.e., triples excitations) are necessary for chemically accurate results.[27] With the full triples correction, CCSDT scales as $O(N^8)$ and is therefore not feasible for larger molecules. A common approach is to include the triples correction based on perturbation theory; this level of theory is called CCSD(T), shows $O(N^7)$ scaling, and is—in conjunction with a triple-ζ basis set—considered the gold standard in quantum chemistry.[28] However, computations at the CCSD(T) level of theory with triple-ζ basis sets are also often too demanding for larger molecules.

Approximations to the CC method, resulting in more favorable scaling behavior, were developed over the past 50 years. Coupled-electron pair approximations (CEPA) were of particular interest in the 1970s, but soon lost public interest, despite initially promising accuracies.[20,29−42] These approaches were picked up in the last decades and resulted initially in the local pair natural orbital (LPNO)−CEPA approach, which has a measured scaling of $O(N^{3.5})$.[20] The LPNO−CCSD approach was further improved to the domain-based local pair orbital (DLPNO)−CCSD method, which enables the computation of molecules with several hundred atoms—several thousand basis functions—with near-linear scaling.[43] Recently, a linear scaling DLPNO−CCSD(T) method was developed,[44,45] which shows good accuracies in benchmark studies.[46−48]

Alternative approaches to CC approximations besides DLPNO are available via ML (vide supra), but none attempts to predict the triples correction based on molecular structure and electronic energy—which are implicit to a regular computation with the CCSD method—alone. This approach would result in CCSD(T) energies at the cost of CCSD energies and could, for example, be used to compute the thermochemistry of a system with high accuracy.

One possible application for high accuracy thermochemistry (i.e., accurate energies) is the domain of atmospheric chemistry, also including large scale climate models that consider chemical processes; a precise knowledge of reaction thermodynamics and kinetic properties is needed for predictive power.[49] Atmospheric chemistry encompasses a multitude of gas-phase radical reactions, most of which are not amenable to experiment. Therefore, a precise prediction of their relative energies is paramount. This work aims to predict the difference between electronic molecular energies at the CCSD and CCSD(T) levels of theory with various Dunning basis sets utilizing a Δ-learning approach. To apply ML techniques to this problem, the electronic energies of small organic molecules and their radicals were computed at the CCSD(T)/X (X = cc-pVDZ, aug-cc-pVDZ, and cc-pVTZ) levels of theory and the results compiled into a database. The molecular data consist of radicals **R•** and their hydrogen-terminated counterparts **R−H**. Manually selected atmospherically relevant molecules were collected in an additional database.[50−52] All models were also validated with a collection of theoretically challenging structures, such as highly conjugated molecules and non-covalent dimers. To demonstrate its ability to capture changes in molecular constitution, we tested our model on a range of isomeric structures. The prediction of the effect of perturbatively included triples was carried out by utilizing graph NNs (GNNs) in conjunction with molecular graphs, the general architecture of which is depicted in Figure 1.

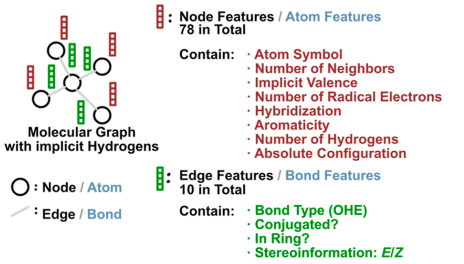


Figure 1. Depiction of a molecular graph used in this study. Each molecule is encoded into a molecular graph, consisting of nodes (assigned to atoms) and edges (assigned to bonds). Every node and edge has an assigned feature vector attached to it, which includes various chemical and physical properties. Based on the graph, a machine learnable molecular representation is generated.

Computing accurate (T) contributions at the CCSD(T)/aug-cc-pVDZ level of theory required a total of 68 d wall time for the 540 molecules in the training set—a value that illustrates the possible time savings when using our ML model for near-instantaneous (T) prediction.

MODELS AND METHODS

Initially, we built a database consisting of simplified molecular input line entry specification (SMILES) strings of 349 small organic molecules (radicals and their hydrogen-terminated counterparts) as taken from a public database.[53] As this database contains only molecules consisting of hydrogen, carbon, nitrogen, and oxygen, we created an additional set of 124 molecules, some of which also contain sulfur. To capture bicyclic and cage-like structures we generated a set of 67 cage hydrocarbons. These sets were used for initial training and testing in a fivefold splitting procedure. External validation of our models was carried out with four validation sets made up from highly conjugated molecules, atmospherically relevant molecules (including radicals), non-covalent dimers (S22 Database[54]), and constitutional isomers.

The initial geometries of all molecules were computed with the Merck molecular force field (MMFF[55−61]) as implemented in ChemML with the use of molecular objects of the RDKit.[62−67] Input files for the Gaussian16[68] quantum chemistry package were then automatically generated and used for optimization of the molecular geometries at the MP2/aug-cc-pVDZ level of theory. Minima on the PES were confirmed by checking the final Hessians.[69−73] Subsequent high-level single point computations at the CCSD(T)/cc-pVDZ, CCSD(T)/aug-cc-pVDZ, and CCSD(T)/cc-pVTZ levels of theory[74−85] were computed on the optimized

https://doi.org/10.1021/acs.jctc.2c00501
J. Chem. Theory Comput. 2022, 18, 4846−4855

structures (the published geometries in case of molecules of the S22 Database) with the quantum chemistry packages CFOUR[86,87] and ORCA.[88,89] We used the frozen-core approach along with unrestricted Hartree−Fock reference wavefunctions where required. We extracted the molecular energies at SCF, MP2, CCSD, and CCSD(T) and combined these in a database along with the SMILES strings. This training and testing database contained only CCSD(T) molecular energies of computations that were previously converged successfully with all three basis sets (540 total). The same procedure was applied to the validation sets, resulting in a total of 95 molecules for validation. Additional computational details and a listing of the molecules in the validation set can be found in the Supporting Information.

The molecular graphs were generated with RDKit in conjunction with the DeepChem node/edge featurizers.[90] Train/test splits were generated with the scikit-learn package in a stratified fashion to ensure balanced splitting.[91] As a model framework for training and testing, we chose PyTorch.[92] We used all GNN layers as implemented in PyTorch Geometric.[93] Model development was separated into four categories: Optimization, training, testing, and validation. Optimization included the perturbation of model hyperparameters, which was carried out with the Optuna[94] framework in conjunction with the Tree-structured Parzen estimator algorithm.[95] As the objective value—the metric of the model to be optimized—we chose the root mean squared error (RMSE). Additionally, we calculated the mean absolute error (MAE) for comparison. With the optimal hyperparameters in hand, we trained our models in combination with an early stopping mechanism to prevent overfitting. The model was then tested on the hitherto unseen test data. This train/test procedure was run for five different random states. The final validation was conducted with the external (out-of-sample) validation database (vide supra) to assess our model's generalization capability and its applicability to challenging systems. An overview of the entire workflow is given in Figure 2.

The GNN part consists of Crystal Graph Convolutional (CGConv[96]) layers, on which a non-linearity is applied by an activation function (Figure 3). The result is recursively fed back to the next CGConv layer in one model branch, while the other branch goes through a global pooling layer (mean-, addition-, and maximum-pooling), resulting in a temporary tensor representation of the data. After every cycle, this representation is concatenated with the molecular embedding to give the final embedding. Owing to this approach, the representation contains information from different steps of the recursion.

The general architecture of our model is depicted in Figure 4. The molecular embedding is learned by the GNN part, then concatenated with two additional molecular fingerprints—Morgan fingerprint (MFP[98]) and Attentive fingerprint (AFP[99])—as well as the SCF, MP2, and CCSD electronic energies. This complete molecular embedding is then channeled through a FFNN to predict the difference between the CCSD and CCSD(T) energies. Expectedly, an alternative approach without CCSD energies gave worse results. More details on model training and optimization and the Python code for the model architecture as well as an exemplary evaluation pipeline is given in the Supporting Information.

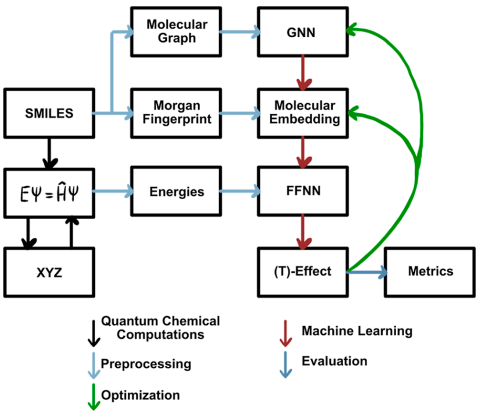


Figure 2. Workflow of data preparation, model generation, optimization, and training. Starting from a list of SMILES strings, inputs for quantum chemistry packages are generated. First, the starting geometry is computed with the MMFF implementation in ChemML. This geometry is then optimized at the MP2/aug-cc-pVDZ level of theory, followed by the respective CC computations. Molecular graphs are generated from the SMILES strings and then processed by our GNN model (Figure 3). The output of the GNN part is then combined with the computed electronic energies and passed through an feed-forward NN (FFNN) (Figure 4) to predict the effect of perturbatively included triples (T).

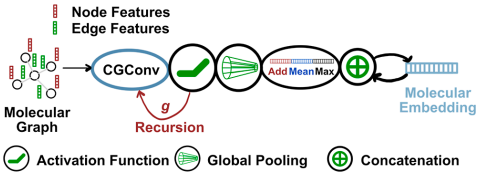


Figure 3. Depiction of the architecture of the GNN model. Consisting of CGConv layers, the model recursively feeds the output of the activation function back to the next CGConv layer (no shared parameters between the different CGConv layers), while the initial output undergoes global pooling to result in a tensorial representation (constructed with respect to mean, addition, and maximum of the graph). The pooled representations are then concatenated with the molecular embedding from the previous recursive step, similar to the graph isomorphism network.[97]

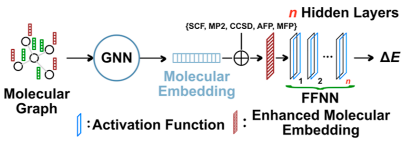


Figure 4. Architecture of our model consists to a large extent of the GNN part (Figure 3) generating the molecular embedding. The latter is then concatenated with the electronic energies (SCF, MP2, and CCSD), along with the AFP and MFP. This concatenated feature vector is then processed by an FFNN to predict the effect of perturbatively included triples (T).

https://doi.org/10.1021/acs.jctc.2c00501
J. Chem. Theory Comput. 2022, 18, 4846−4855

RESULTS AND DISCUSSION

To obtain an initial performance overview of our models, we made kernel density estimations (KDEs) for each model that was trained on our train/test database. The deviations (predicted minus true value) of all five testing phases were combined and then used for the KDE, which was carried out with the default parameters of the pandas library.[100] The KDE for each model is depicted in Figure 5.

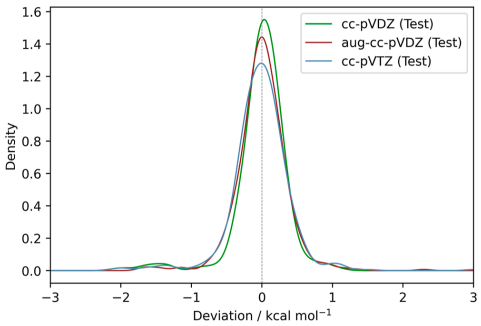


Figure 5. Depiction of the KDE of all test set deviations (predicted minus true value) over a fivefold splitting. The KDE is given for the prediction of the effect of perturbatively included triples with the cc-pVDZ, aug-cc-pVDZ, and cc-pVTZ basis sets.

The KDEs show that the predictions of our models (1) have mostly absolute deviations below 1.0 kcal mol^{-1} and (2) are symmetrically distributed around 0 kcal mol^{-1} indicating well-balanced and generalizing models. The symmetry is most pronounced for the triple-ζ basis set and the augmented double-ζ basis set, whereas the double-ζ basis sets show a small shift towards positive deviations for predictions of the effect of perturbatively included triples. The average test MAEs (RMSEs) over the fivefold splitting are 0.24 (0.39), 0.24 (0.35), and 0.28 (0.43) kcal mol^{-1} for cc-pVDZ, aug-cc-pVDZ, and cc-pVTZ, respectively. Plots showing the predicted value against the true value for all splits can be found in the Supporting Information.

To further investigate the strengths and weaknesses of our approach, we trained our models on the full train/test set (all data points used for training) and proceeded to predictions on external validation basis (vide supra). The KDEs on the deviations with a procedure analogous as used for the initial train/test models for predictions on the different validation sets in shown in Figure 6.

The KDEs are symmetrical around 0 kcal mol^{-1} for all validation sets with all three basis sets. The absolute deviations for predictions on the Atmos, S22, Isomers, and Conjugated sets are mostly below 1.5, 2.0, 1.0, and 1.5 kcal mol^{-1}, respectively. The average MAEs (RMSEs) over the three basis sets are 0.53 (0.76), 0.81 (1.00), 0.15 (0.17), and 0.34 (0.39) kcal mol^{-1} for atmos, S22, isomers, and conjugated, respectively. These metrics indicate that our models behave very well on data that are unrelated to the training data and not only have a high capacity but also generalize well. The effect of perturbatively included triples for radicals as well as for closed-shell molecules can be predicted with high precision by our models. The higher error for the prediction of non-covalent dimers (S22) is anticipated given the absence of training

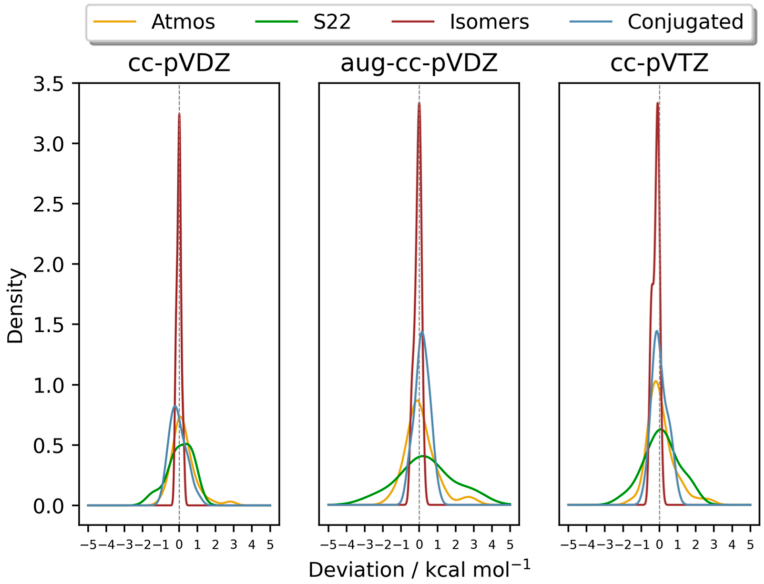


Figure 6. KDEs of the deviations (predicted minus true value) of our fully trained models on the prediction of perturbatively included triples with the cc-pVDZ, aug-cc-pVDZ, and cc-pVTZ basis sets. The KDEs for each basis set are split into contributions of the individual validation sets.

https://doi.org/10.1021/acs.jctc.2c00501
J. Chem. Theory Comput. 2022, 18, 4846−4855

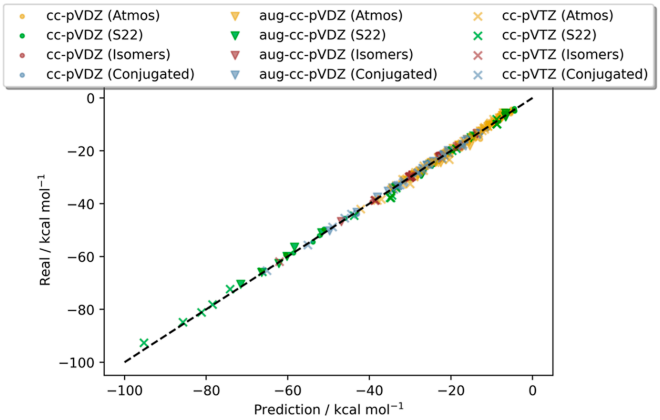


Figure 7. Performance plot of our models for the prediction of perturbatively included triples with the cc-pVDZ, aug-cc-pVDZ, and cc-pVTZ basis sets. The black-dashed diagonal indicates optimal prediction.

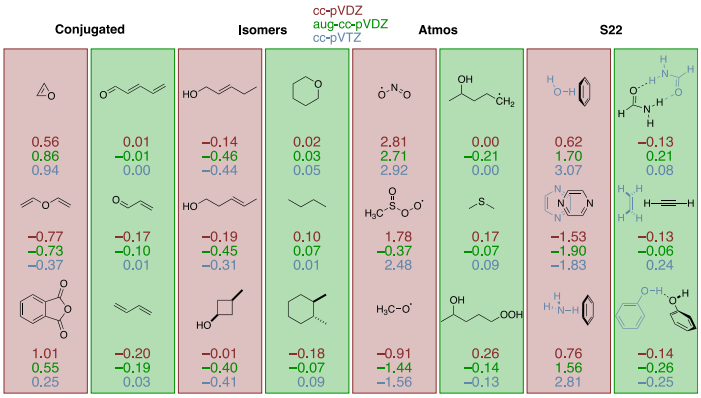


Figure 8. Overview of the average three best (green background) and worst (red background) predicted molecules of each validation set. The deviations (predicted minus true value) of the effect of perturbatively included triples are given in kcal mol^{-1} for the cc-pVDZ, aug-cc-pVDZ, and cc-pVTZ basis sets.

examples; it is, however, surprising that we do not observe a systematic shift for the S22 set considering that the models were only trained on monomers. Moreover, due to the current model architecture, the dimers are represented as one joint overall graph containing two disconnected molecular graphs without edges (i.e., discrete bonds) between them. The models hence learn the effect of perturbatively included triples for two infinitely distant particles (molecules), and thus miss the stabilizing long−range interaction energies. We originally expected our models to predict a (T) correction for dimers that is systematically too low (i.e., predicts the dimers to be bound too weakly[101]), because of the missing exposure to stabilizing long−range interactions during training. Thus, explicitly including non-covalent dimers into the training process could enable the models to "naturally" pick up such stabilizing contributions, but a more sophisticated approach with graph representations including edges between both monomers would be required to achieve even better prediction precision. The performance of our models on the external validation set depicted as computed versus predicted values is shown in Figure 7.

The performance plot illustrates again the high accuracy of our models, reaching a mean (over the three stated basis sets) R^2-score of 0.987, 0.998, 0.999, and 0.998 for the atmos, S22, isomers, and conjugated validation sets, respectively. All predictions lie close to the diagonal, which indicates near-optimal prediction. Indeed, predictions are even accurate far beyond the scope of the training data, considering that the absolute (T) values range between −39 and −12 kcal mol^{-1} for the training set, but even data points <−80 kcal mol^{-1} are predicted in good agreement with computed values.

Our working hypothesis is that our model learns the trend of the effect of perturbatively included triples (T) from the difference between the HF and MP2 energy, the general energy scale from the CCSD energy, and subsequently fine-tunes the prediction with the molecular embeddings. This has

https://doi.org/10.1021/acs.jctc.2c00501
J. Chem. Theory Comput. 2022, 18, 4846−4855

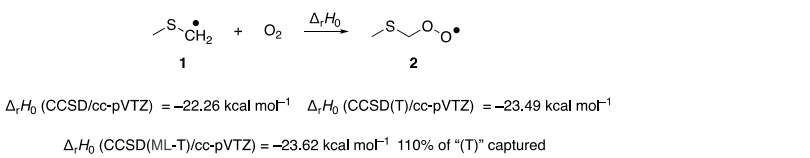


Figure 9. One of the key steps in the DMS oxidation under marine boundary layer conditions, the oxidation of the (methylthio)methyl radical H_3CSCH_2 (**1**) to the (methylthio)methylperoxy radical $H_3CSCH_2OO\cdot$ (**3**). The reaction enthalpy is shown for the CCSD/ and CCSD(T)/cc-pVTZ//MP2/aug-cc-pVDZ levels of theory, as well as the predicted CCSD(T)/cc-pVTZ reaction enthalpy.

proven essential for successful predictions (e.g., see Figure S71 in the Supporting Information). As accurate molecular embeddings are paramount for an exact prediction it becomes clear why molecules with a very small molecular graph (with less than three nodes) are predicted worse throughout than those with more nodes. As GNNs utilize message-passing between different nodes via edges, a small network would limit such feature exchanges and therefore limit the inherent functionality of the GNN. This effect can be seen in Figure 8, where the best and worst (averaged over the three basis sets) three molecules are depicted along with their respective deviations between predicted and computed (T) values. Small atmospherically relevant molecules including radicals such as $\bullet NO_2$ or $H_3CS(O)OO\bullet$ are predicted less accurately compared to larger ones with more common functional groups, for example, 4-hydroxypentanyl radical or dimethyl sulfide (DMS). Generally, Figure 8 shows that an arbitrarily exact prediction of (T) is possible, while the largest deviations are still around computational accuracy. The predictions with our models are therefore statically well suited to predict the effect of perturbatively included triples with the cost of a regular computation at the CCSD/X level of theory.

As DMS oxidation plays a major role in atmospheric chemistry,[50] we want to showcase one of the key steps of reactivity in Figure 9: The (methylthio)methyl radical $H_3CSCH_2\bullet$ **1** is oxidized by triplet dioxygen to the (methylthio)methyleneperoxy radical **2** with a reaction enthalpy Δ_rH_0 of−23.49 kcal mol^{-1} at the CCSD(T)/cc-pVTZ level of theory. Calculating Δ_rH_0 with energies predicted by our model (CCSD(ML-T)/cc-pVTZ), results in a value of −23.62 kcal mol^{-1} and therefore captures slightly more than 100% of the effect of perturbatively included triples. Computing Δ_rH_0 at the CCSD/cc-pVTZ level of theory results in a value of −22.26 kcal mol^{-1} and thus underestimates the driving force of the reaction with respect to CCSD(T).

Similarly to the reaction shown in Figure 9, it is possible to augment CCSD computations to CCSD(T) at virtually no cost with our model. An exemplary case for this workflow is picrotoxinin ($C_{15}H_{16}O_6$)—an asymmetric molecule (Figure 10) consisting of too many atoms for a regular computation at the CCSD(T)/cc-pVTZ level of theory. But with our model, it is possible to predict the relative energy of the active compound (picrotoxinin) and its inactive acylated derivative with results similar to DLPNO-CCSD(T)/cc-pVTZ. Our model predicts an $\Delta_rH_0 < 0$, which is in good agreement with the results computed at the DLPNO-CCSD(T)/cc-pVTZ//MP2/aug-cc-pVDZ level of theory; in contrast, lower-level computations at the CCSD/cc-pVDZ//MP2/aug-cc-pVDZ level and DLPNO-CCSD(T)/cc-pVDZ//MP2/aug-cc-pVDZ level of theory yield $\Delta_rH_0 > 0$.

Furthermore, we demonstrate that our model can also tackle applications where DFT fails to give accurate isomer energy differences. An example are the isomer energy differences of large hydrocarbons, for example, $(CH)_{12}$,[23,102] which are also impractical to compute at high levels of theory. To further demonstrate the predictive accuracy of our models, we predicted the isomer energy differences for three large hydrocarbons depicted in Figure 11. Note that, perhaps counterintuitive to some, octahedrane is the most stable $(CH)_{12}$ hydrocarbon, despite considerable strain.[103]

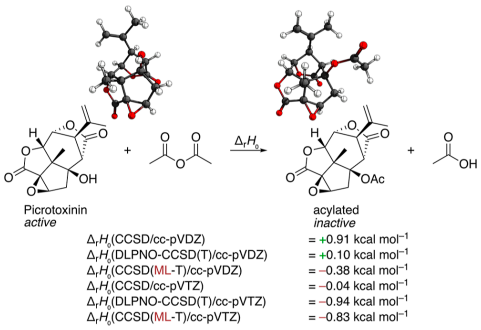


Figure 10. Depiction of an exemplary use case of machine-learned CC energy. The relative energies between picrotoxinin ($C_{15}H_{16}O_6$) and its inactive acylated analog ($C_{12}H_{18}O_7$)—both asymmetric molecules that consist of too many atoms for a regular computation at the CCSD(T) level of theory. Shown are the computed structures of picrotoxinin and its stated analog at the MP2/aug-cc-pVDZ level of theory, as well as the predicted reaction enthalpies at the CCSD(ML-T)/X//MP2/aug-cc-pVDZ level of theory. Enthalpies at the CCSD/X//MP2/aug-cc-pVDZ and DLPNO-CCSD(T)/X//MP2/aug-cc-pVDZ levels of theory are given as a comparison; X = cc-pVDZ and cc-pVTZ. All energies were corrected for the zero-point vibrational energy computed at the MP2/aug-cc-pVDZ level of theory. Atoms are encoded as red: oxygen, gray: carbon, and white: hydrogen.

The relative isomer energies shown above indicate that even for complex molecules such as the $(CH)_{12}$ isomers, relative energies can be predicted well with our models. The relative energies at the CCSD(ML-T)/X level of theory represent the relative energies closer to the more exact CCSD(T) relative energies than lower-level wavefunction-based methods, such as CCSD or MP2 with various basis sets. Many DFT methods fail to map the relative energies correctly which indicates that exact isomer energy differences at minimal computational cost require a data-centric approach.

CONCLUSION

We demonstrated that the prediction of the CCSD and CCSD(T) energy difference, that is, Δ-learning of the effect of

https://doi.org/10.1021/acs.jctc.2c00501
J. Chem. Theory Comput. 2022, 18, 4846−4855

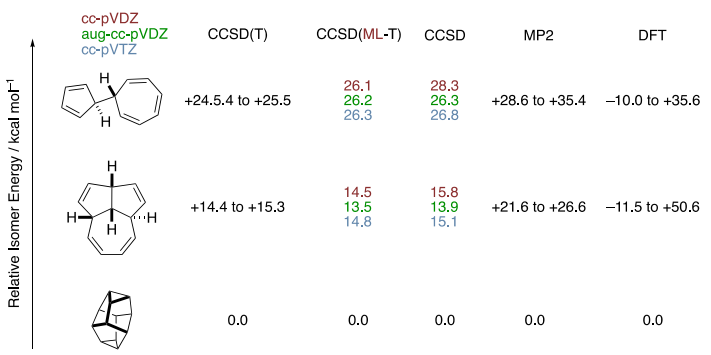


Figure 11. Depiction of the isomer energy differences for the three lowest-lying $(CH)_{12}$ hydrocarbons. The differences are given as a range for the CCSD(T), MP2, and DFT methods with various basis functions,[102] computed relative energies at the CCSD/X//MP2/aug-cc-pVDZ (X = cc-pVDZ, aug-cc-pVDZ, and cc-pVTZ) levels of theory, as well as the predicted CCSD(ML-T)/X relative energies in kcal mol^{-1}.

perturbatively included triples (T) is possible and provides excellent accuracy with an average MAE below 0.28 kcal mol^{-1} for our models. Validation procedures reveal that our model can distinguish between constitutional isomers, accurately predict the energies of highly conjugated molecules, small organic molecules containing diverse functional groups, organic radicals, and even non-covalent dimers. The results emphasize the validity of our model for the underlying task: even with comparatively small databases, it is possible to achieve quite reasonable accuracy. While the current data sets only cover a small part of chemical space, enlarging the number of included molecular species will further increase the generalizability of the approach. Accurate predictions are possible for all three tested basis sets (cc-pVDZ, aug-cc-pVDZ, and cc-pVTZ) and no significant difference in accuracy can be observed.

Our results are of comparable quality to the fragment-graph approach by Collins and Raghavachari,[104] who were able to achieve an MAE of 0.48 kcal mol^{-1} on DFT energies with a dataset size of 1000 on the GDB-9 Database[105] (MAE of 0.12−0.16 kcal mol^{-1} with the entire database). More data points would very likely improve our model even further, as it is well known that a model's performance generally increases with more data until it converges to the Bayes error.[106] The effect of the database size can also be viewed in a chemical context.[13] This work shows that a combination of several molecular representations is beneficial for an ML approach. Our model learns which input features are relevant for the prediction and no performance decrease ensues due to numerous input features or additional molecular fingerprints. The next iterations of this model will focus on a more sophisticated graph representation to capture intermolecular long−range interactions with even higher precision. Furthermore, we plan to expand our database with computations at the DLPNO-CCSD(T)/cc-pVTZ level of theory to utilize the excellent scaling behavior of DLPNO for expedited computation of input features, which the next generation of our models should use to approximate even more accurate increments of CC energies.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.2c00501.

Supporting Information, including architecture, training, and optimization of the models, as well as computational details, are available free of charge (PDF). Additionally, the databases containing electronic energies (SCF, MP2, CCSD, and CCSD(T)) along with SMILES are provided as CSV files. The geometries of hitherto unpublished molecules are available as TXT files (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Peter R. Schreiner − *Institute of Organic Chemistry, Justus Liebig University, 35392 Giessen, Germany;* orcid.org/0000-0002-3608-5515; Email: prs@uni-giessen.de

Authors

Marcel Ruth − *Institute of Organic Chemistry, Justus Liebig University, 35392 Giessen, Germany;* orcid.org/0000-0001-9880-9956

Dennis Gerbig − *Institute of Organic Chemistry, Justus Liebig University, 35392 Giessen, Germany;* orcid.org/0000-0002-7023-8298

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.2c00501

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding

M.R. thanks the Fonds der Chemischen Industrie for a doctoral scholarship.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Moritz R. Schäfer (ITC Stuttgart) for fruitful discussions.

REFERENCES

(1) Samuel, A. L. Some Studies in Machine Learning Using the Game of Checkers. *IBM J. Res. Dev.* **1959**, *3*, 210−229.
(2) 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.
(3) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559*, 547−555.
(4) Aspuru-Guzik, A.; Baik, M.-H.; Balasubramanian, S.; Banerjee, R.; Bart, S.; Borduas-Dedekind, N.; Chang, S.; Chen, P.; Corminboeuf, C.; Coudert, F.-X.; Cronin, L.; Crudden, C.; Cuk, T.; Doyle, A. G.; Fan, C.; Feng, X.; Freedman, D.; Furukawa, S.; Ghosh, S.; Glorius, F.; Jeffries-El, M.; Katsonis, N.; Li, A.; Linse, S. S.; Marchesan, S.; Maulide, N.; Milo, A.; Narayan, A. R. H.; Naumov, P.; Nevado, C.; Nyokong, T.; Palacin, R.; Reid, M.; Robinson, C.; Robinson, G.; Sarpong, R.; Schindler, C.; Schlau-Cohen, G. S.; Schmidt, T. W.; Sessoli, R.; Shao-Horn, Y.; Sleiman, H.; Sutherland, J.; Taylor, A.; Tezcan, A.; Tortosa, M.; Walsh, A.; Watson, A. J. B.; Weckhuysen, B. M.; Weiss, E.; Wilson, D.; Yam, V. W.-W.; Yang, X.; Ying, J. Y.; Yoon, T.; You, S.-L.; Zarbin, A. J. G.; Zhang, H. Charting a Course for Chemistry. *Nat. Chem.* **2019**, *11*, 286−294.
(5) Bock, F. E.; Aydin, R. C.; Cyron, C. J.; Huber, N.; Kalidindi, S. R.; Klusemann, B. A Review of the Application of Machine Learning and Data Mining Approaches in Continuum Materials Mechanics. *Front. Mater.* **2019**, *6*, 10.
(6) Pyzer-Knapp, E. O.; Laino, T. Preface. *Machine Learning in Chemistry: Data-Driven Algorithms, Learning Systems, and Predictions*; American Chemical Society: Washington, DC, 2019; Vol. *1326*, pp ix−x.
(7) Čížek, J. On the Correlation Problem in Atomic and Molecular Systems. Calculation of Wavefunction Components in Ursell-Type Expansion Using Quantum-Field Theoretical Methods. *J. Chem. Phys.* **1966**, *45*, 4256−4266.
(8) Peyton, B. G.; Briggs, C.; D'Cunha, R.; Margraf, J. T.; Crawford, T. D. Machine-Learning Coupled Cluster Properties through a Density Tensor Representation. *J. Phys. Chem. A* **2020**, *124*, 4861−4871.
(9) Bogojeski, M.; Vogt-Maranto, L.; Tuckerman, M. E.; Müller, K.-R.; Burke, K. Quantum Chemical Accuracy from Density Functional Approximations via Machine Learning. *Nat. Commun.* **2020**, *11*, 5223.
(10) Townsend, J.; Vogiatzis, K. D. Data-Driven Acceleration of the Coupled-Cluster Singles and Doubles Iterative Solver. *J. Phys. Chem. Lett.* **2019**, *10*, 4129−4135.
(11) Townsend, J.; Vogiatzis, K. D. Transferable MP2-Based Machine Learning for Accurate Coupled-Cluster Energies. *J. Chem. Theory Comput.* **2020**, *16*, 7453−7461.
(12) Agarawal, V.; Roy, S.; Chakraborty, A.; Maitra, R. Accelerating Coupled Cluster Calculations with Nonlinear Dynamics and Supervised Machine Learning. *J. Chem. Phys.* **2021**, *154*, 044110.
(13) 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.
(14) 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.
(15) 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.
(16) Käser, S.; Boittier, E. D.; Upadhyay, M.; Meuwly, M. Transfer Learning to CCSD(T): Accurate Anharmonic Frequencies from Machine Learning Models. *J. Chem. Theory Comput.* **2021**, *17*, 3687−3699.
(17) Wilkins, D. M.; Grisafi, A.; Yang, Y.; Lao, K. U.; DiStasio, R. A.; Ceriotti, M. Accurate Molecular Polarizabilities with Coupled Cluster Theory and Machine Learning. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 3401−3406.
(18) Sherrill, C. D. Frontiers in Electronic Structure Theory. *J. Chem. Phys.* **2010**, *132*, 110902.
(19) Bowler, D. R.; Miyazaki, T. Calculations for Millions of Atoms with Density Functional Theory: Linear Scaling Shows its Potential. *J. Phys. Condens. Matter* **2010**, *22*, 074207.
(20) Neese, F.; Wennmohs, F.; Hansen, A. Efficient and Accurate Local Approximations to Coupled-Electron Pair Approaches: An Attempt to Revive the Pair Natural Orbital Method. *J. Chem. Phys.* **2009**, *130*, 114108.
(21) Bao, J. L.; Gagliardi, L.; Truhlar, D. G. Self-Interaction Error in Density Functional Theory: An Appraisal. *J. Phys. Chem. Lett.* **2018**, *9*, 2353−2358.
(22) Woodcock, H. L.; Schaefer, H. F.; Schreiner, P. R. Problematic Energy Differences between Cumulenes and Poly-ynes: Does This Point to a Systematic Improvement of Density Functional Theory? *J. Phys. Chem. A* **2002**, *106*, 11923−11931.
(23) Zhao, Y.; Truhlar, D. G. A Density Functional That Accounts for Medium-Range Correlation Energies in Organic Chemistry. *Org. Lett.* **2006**, *8*, 5753−5755.
(24) Schwabe, T.; Grimme, S. Towards Chemical Accuracy for the Thermodynamics of Large Molecules: New Hybrid Density Functionals Including Non-Local Correlation Effects. *Phys. Chem. Chem. Phys.* **2006**, *8*, 4398−4401.
(25) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Challenges for Density Functional Theory. *Chem. Rev.* **2012**, *112*, 289−320.
(26) Neese, F. A Critical Evaluation of DFT, Including Time-Dependent DFT, Applied to Bioinorganic Chemistry. *J. Biol. Inorg. Chem.* **2006**, *11*, 702−711.
(27) Paldus, J.; Čížek, J.; Shavitt, I. Correlation Problems in Atomic and Molecular Systems. IV. Extended Coupled-Pair Many-Electron Theory and Its Application to the BH_3 Molecule. *Phys. Rev. A* **1972**, *5*, 50−67.
(28) Crawford, T. D.; Schaefer, H. F. *Rev. Comput. Chem.*; John Wiley & Sons, Inc., 2007; pp 33−136.
(29) Meyer, W. Ionization Energies of Water From PNO-CI Calculations. *Int. J. Quantum Chem.* **1971**, *5*, 341−348.
(30) Meyer, W. PNO−CI Studies of Electron Correlation Effects. I. Configuration Expansion by Means of Nonorthogonal Orbitals, and Application to the Ground State and Ionized States of Methane. *J. Chem. Phys.* **1973**, *58*, 1017−1035.
(31) Meyer, W. PNO-CI and CEPA Studies of Electron Correlation Effects. *Theor. Chim. Acta* **1974**, *35*, 277−292.
(32) Ahlrichs, R.; Driessler, F.; Lischka, H.; Staemmler, V.; Kutzelnigg, W. PNO−CI (Pair Natural Orbital Configuration Interaction) and CEPA−PNO (Coupled Electron Pair Approximation with Pair Natural Orbitals) Calculations of Molecular Systems. II. The Molecules BeH_2, BH, BH_3, CH_4, CH_3, NH_3 (Planar and Pyramidal), H_2O, OH_3^+, HF and the Ne Atom. *J. Chem. Phys.* **1975**, *62*, 1235−1247.
(33) Taylor, P. R.; Bacskay, G. B.; Hush, N. S.; Hurley, A. C. The Coupled-Pair Approximation in a Basis of Independent-Pair Natural Orbitals. *Chem. Phys. Lett.* **1976**, *41*, 444−449.
(34) Werner, H.-J.; Meyer, W. PNO-CI and PNO-CEPA Studies of Electron Correlation Effects. *Mol. Phys.* **1976**, *31*, 855−872.
(35) Rosmus, P.; Meyer, W. PNO-CI and CEPA Studies of Electron Correlation Effects. VI. Electron Affinities of the First-Row and Second-Row Diatomic Hydrides and the Spectroscopic Constants of Their Negative Ions. *J. Chem. Phys.* **1978**, *69*, 2745−2751.
(36) Taylor, P. R.; Bacskay, G. B.; Hush, N. S.; Hurley, A. C. Unlinked Cluster Effects in Molecular Electronic Structure. I. The HCN and HNC Molecules. *J. Chem. Phys.* **1978**, *69*, 1971−1979.
(37) Ahlrichs, R. Many Body Perturbation Calculations and Coupled Electron Pair Models. *Comput. Phys. Commun.* **1979**, *17*, 31−45.
(38) Koch, S.; Kutzelnigg, W. Comparison of CEPA and CP-MET methods. *Theor. Chim. Acta* **1981**, *59*, 387−411.

https://doi.org/10.1021/acs.jctc.2c00501
J. Chem. Theory Comput. 2022, 18, 4846−4855

(39) Staemmler, V.; Jaquet, R. CEPA Calculations on Open-Shell Molecules. I. Outline of the Method. *Theor. Chim. Acta* **1981**, *59*, 487−500.
(40) Taylor, P. R. A Rapidly Convergent CI Expansion Based on Several Reference Configurations, Using Optimized Correlating Orbitals. *J. Chem. Phys.* **1981**, *74*, 1256−1270.
(41) Ahlrichs, R.; Scharf, P.; Ehrhardt, C. The Coupled Pair Functional (CPF). A Size Consistent Modification of the CI(SD) Based on an Energy Functional. *J. Chem. Phys.* **1985**, *82*, 890−898.
(42) Pulay, P.; Sæbø, S. Variational CEPA: Comparison with Different Many-Body Methods. *Chem. Phys. Lett.* **1985**, *117*, 37−41.
(43) Riplinger, C.; Neese, F. An Efficient and Near Linear Scaling Pair Natural Orbital Based Local Coupled Cluster Method. *J. Chem. Phys.* **2013**, *138*, 034106.
(44) Riplinger, C.; Pinski, P.; Becker, U.; Valeev, E. F.; Neese, F. Sparse Maps—A Systematic Infrastructure for Reduced-Scaling Electronic Structure Methods. II. Linear Scaling Domain Based Pair Natural Orbital Coupled Cluster Theory. *J. Chem. Phys.* **2016**, *144*, 024109.
(45) Guo, Y.; Riplinger, C.; Becker, U.; Liakos, D. G.; Minenkov, Y.; Cavallo, L.; Neese, F. Communication: An Improved Linear Scaling Perturbative Triples Correction for the Domain Based Local Pair-Natural Orbital Based Singles and Doubles Coupled Cluster Method [DLPNO-CCSD(T)]. *J. Chem. Phys.* **2018**, *148*, 011101.
(46) Liakos, D. G.; Sparta, M.; Kesharwani, M. K.; Martin, J. M. L.; Neese, F. Exploring the Accuracy Limits of Local Pair Natural Orbital Coupled-Cluster Theory. *J. Chem. Theory Comput.* **2015**, *11*, 1525−1539.
(47) Sandler, I.; Chen, J.; Taylor, M.; Sharma, S.; Ho, J. Accuracy of DLPNO-CCSD(T): Effect of Basis Set and System Size. *J. Phys. Chem. A* **2021**, *125*, 1553−1563.
(48) Liakos, D. G.; Guo, Y.; Neese, F. Comprehensive Benchmark Results for the Domain Based Local Pair Natural Orbital Coupled Cluster Method (DLPNO-CCSD(T)) for Closed- and Open-Shell Systems. *J. Phys. Chem. A* **2020**, *124*, 90−100.
(49) Csontos, J.; Rolik, Z.; Das, S.; Kállay, M. High-Accuracy Thermochemistry of Atmospherically Important Fluorinated and Chlorinated Methane Derivatives. *J. Phys. Chem. A* **2010**, *114*, 13093−13103.
(50) Mardyukov, A.; Schreiner, P. R. Atmospherically Relevant Radicals Derived from the Oxidation of Dimethyl Sulfide. *Acc. Chem. Res.* **2018**, *51*, 475−483.
(51) Chen, C.; Wang, L.; Zhao, X.; Wu, Z.; Bernhardt, B.; Eckhardt, A. K.; Schreiner, P. R.; Zeng, X. Photochemistry of $HNSO_2$ in Cryogenic Matrices: Spectroscopic Identification of the Intermediates and Mechanism. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7975−7983.
(52) Gerbig, D.; Bernhardt, B.; Wende, R. C.; Schreiner, P. R. Capture and Reactivity of an Elusive Carbon−Sulfur Centered Biradical. *J. Phys. Chem. A* **2020**, *124*, 2014−2018.
(53) St. John, P. C.; Kim, Y.; Etz, B. D.; Kim, S.; Paton, R. S.; Paton, R. S. Quantum Chemical Calculations for Over 200,000 Organic Radical Species and 40,000 Associated Closed-Shell Molecules. *Sci. Data* **2020**, *7*, 244.
(54) Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. Benchmark Database of Accurate (MP2 and CCSD(T) Complete Basis Set Limit) Interaction Energies of Small Model Complexes, DNA Base Pairs, and Amino Acid Pairs. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985−1993.
(55) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.
(56) 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.
(57) Halgren, T. A. Merck Molecular Force Field. III. Molecular Geometries and Vibrational Frequencies for MMFF94. *J. Comput. Chem.* **1996**, *17*, 553−586.
(58) Halgren, T. A. Merck Molecular Force Field. V. Extension of MMFF94 Using Experimental Data, Additional Computational Data, and Empirical Rules. *J. Comput. Chem.* **1996**, *17*, 616−641.
(59) Halgren, T. A.; Nachbar, R. B. Merck Molecular Force Field. IV. Conformational Energies and Geometries for MMFF94. *J. Comput. Chem.* **1996**, *17*, 587−615.
(60) Halgren, T. A. MMFF VI. MMFF94s Option for Energy Minimization Studies. *J. Comput. Chem.* **1999**, *20*, 720−729.
(61) Halgren, T. A. MMFF VII. Characterization of MMFF94, MMFF94s, and Other Widely Available Force Fields for Conformational Energies and for Intermolecular-Interaction Energies and Geometries. *J. Comput. Chem.* **1999**, *20*, 730−748.
(62) Tosco, P.; Stiefl, N.; Landrum, G. Bringing the MMFF Force Field to the RDKit: Implementation and Validation. *J. Cheminf.* **2014**, *6*, 37.
(63) Hachmann, J.; Afzal, M. A. F.; Haghighatlari, M.; Pal, Y. Building and Deploying a Cyberinfrastructure for the Data-Driven Design of Chemical Systems and the Exploration of Chemical Space. *Mol. Simul.* **2018**, *44*, 921−929.
(64) Haghighatlari, M.; Hachmann, J. Advances of Machine Learning in Molecular Modeling and Simulation. *Curr. Opin. Chem. Eng.* **2019**, *23*, 51−57.
(65) Haghighatlari, M.; Vishwakarma, G.; Altarawy, D.; Subramanian, R.; Kota, B. U.; Sonpal, A.; Setlur, S.; Hachmann, J., ChemML: A Machine Learning and Informatics Program Package for the Analysis, Mining, and Modeling of Chemical and Materials Data. **2019**, ChemRxiv:8323271.
(66) Vishwakarma, G.; Haghighatlari, M.; Hachmann, J. Towards Autonomous Machine Learning in Chemistry via Evolutionary Algorithms, **2019**, ChemRxiv:9782387.
(67) RDKit: Open-source cheminformatics. https://www.rdkit.org.
(68) 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; 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, J. A., Jr.; 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*, Rev. C.01: Wallingford, CT, 2016.
(69) Head-Gordon, M.; Pople, J. A.; Frisch, M. J. MP2 Energy Evaluation by Direct Methods. *Chem. Phys. Lett.* **1988**, *153*, 503−506.
(70) Sæbø, S.; Almlöf, J. Avoiding the Integral Storage Bottleneck in LCAO Calculations of Electron Correlation. *Chem. Phys. Lett.* **1989**, *154*, 83−89.
(71) Frisch, M. J.; Head-Gordon, M.; Pople, J. A. A Direct MP2 Gradient Method. *Chem. Phys. Lett.* **1990**, *166*, 275−280.
(72) Frisch, M. J.; Head-Gordon, M.; Pople, J. A. Semi-Direct Algorithms for the MP2 Energy and Gradient. *Chem. Phys. Lett.* **1990**, *166*, 281−289.
(73) Head-Gordon, M.; Head-Gordon, T. Analytic MP2 Frequencies Without Fifth-Order Storage. Theory and Application to Bifurcated Hydrogen Bonds in the Water Hexamer. *Chem. Phys. Lett.* **1994**, *220*, 122−128.
(74) Hampel, C.; Peterson, K. A.; Werner, H.-J. A Comparison of the Efficiency and Accuracy of the Quadratic Configuration Interaction (QCISD), Coupled Cluster (CCSD), and Brueckner Coupled Cluster (BCCD) Methods. *Chem. Phys. Lett.* **1992**, *190*, 1−12.
(75) Purvis, G. D., III; Bartlett, R. J. A Full Coupled-Cluster Singles and Doubles Model: The Inclusion of Disconnected Triples. *J. Chem. Phys.* **1982**, *76*, 1910−1918.

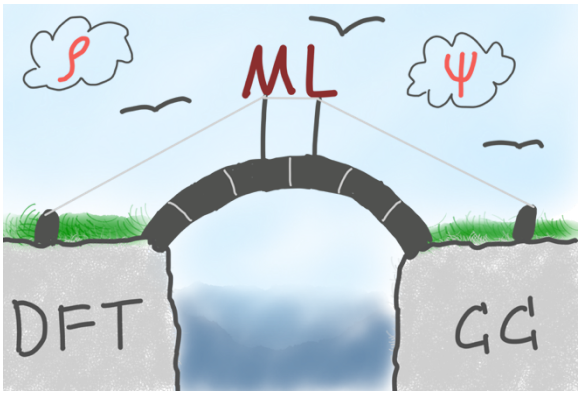
https://doi.org/10.1021/acs.jctc.2c00501
J. Chem. Theory Comput. 2022, 18, 4846−4855

(76) Scuseria, G. E.; Scheiner, A. C.; Lee, T. J.; Rice, J. E.; Schaefer, G. D., III The Closed-Shell Coupled Cluster Single and Double Excitation (CCSD) Model for the Description of Electron Correlation. A Comparison with Configuration Interaction (CISD) Results. *J. Chem. Phys.* **1987**, *86*, 2881−2890.
(77) Stanton, J. F.; Gauss, J.; Watts, J. D.; Bartlett, R. J. A Direct Product Decomposition Approach for Symmetry Exploitation in Many-Body Methods. I. Energy Calculations. *J. Chem. Phys.* **1991**, *94*, 4334−4345.
(78) Bartlett, R. J.; Watts, J. D.; Kucharski, S. A.; Noga, J. Non-Iterative Fifth-Order Triple and Quadruple Excitation Energy Corrections in Correlated Methods. *Chem. Phys. Lett.* **1990**, *165*, 513−522.
(79) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. A Fifth-Order Perturbation Comparison of Electron Correlation Theories. *Chem. Phys. Lett.* **1989**, *157*, 479−483.
(80) Stanton, J. F. Why CCSD(T) Works: A Different Perspective. *Chem. Phys. Lett.* **1997**, *281*, 130−134.
(81) Peterson, K. A.; Woon, D. E.; Dunning, T. H., Jr. Benchmark Calculations With Correlated Molecular Wave Functions. IV. The Classical Barrier Height of the $H+H_2 \rightarrow H_2+H$ Reaction. *J. Chem. Phys.* **1994**, *100*, 7410−7415.
(82) Wilson, A. K.; van Mourik, T.; Dunning, T. H. Gaussian Basis Sets for Use in Correlated Molecular Calculations. VI. Sextuple Zeta Correlation Consistent Basis Sets for Boron Through Neon. *J. Mol. Struct. THEOCHEM* **1996**, *388*, 339−349.
(83) Woon, D. E.; Dunning, T. H., Jr. Gaussian Basis Sets for Use in Correlated Molecular Calculations. III. The Atoms Aluminum Through Argon. *J. Chem. Phys.* **1993**, *98*, 1358−1371.
(84) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. Electron Affinities of the First-Row Atoms Revisited. Systematic Basis Sets and Wave Functions. *J. Chem. Phys.* **1992**, *96*, 6796−6806.
(85) Dunning, T. H., Jr. Gaussian Basis Sets for Use in Correlated Molecular Calculations. I. The Atoms Boron Through Neon and Hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007−1023.
(86) CFOUR. a quantum chemical program package written by J.F. Stanton, J. Gauss, L. Cheng, M.E. Harding, D.A. Matthews, P.G. Szalay with contributions from A.A. Auer, R.J. Bartlett, U. Benedikt, C. Berger, D.E. Bernholdt, S. Blaschke, Y.J. Bomble, S. Burger, O. Christiansen, D. Datta, F. Engel, R. Faber, J. Greiner, M. Heckert, O. Heun, M. Hilgenberg, C. Huber, T.-C. Jagau, D. Jonsson, J. Jusélius, T. Kirsch, K. Klein, G.M. Kopper, W.J. Lauderdale, F. Lipparini, T. Metzroth, L.A. Mück, T. Nottoli, D.P. O'Neill, D.R. Price, E. Prochnow, C. Puzzarini, K. Ruud, F. Schiffmann, W. Schwalbach, C. Simmons, S. Stopkowicz, A. Tajti, J. Vázquez, F. Wang, J.D. Watts and the integral packages MOLECULE (J. Almlöf and P.R. Taylor), PROPS (P.R. Taylor), ABACUS (T. Helgaker, H.J. Aa. Jensen, P. Jørgensen, and J. Olsen), and ECP routines by A. V. Mitin and C. van Wüllen. For the current version, see http://www.cfour.de.
(87) Harding, M. E.; Metzroth, T.; Gauss, J.; Auer, A. A. Parallel Calculation of CCSD and CCSD(T) Analytic First and Second Derivatives. *J. Chem. Theory Comput.* **2008**, *4*, 64−74.
(88) Izsák, R. Single-Reference Coupled Cluster Methods for Computing Excitation Energies in Large Molecules: The Efficiency and Accuracy of Approximations. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2020**, *10*, No. e1445.
(89) Neese, F. The ORCA Program System. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012**, *2*, 73−78.
(90) Ramsundar, B.; Eastman, P.; Walters, P.; Pande, V.; Leswing, K.; Wu, Z. *Deep Learning for the Life Sciences*; O'Reilly Media, 2019.
(91) 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.* **2011**, *12*, 2825−2830.
(92) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steinerand, B.; Fang, L.; Bai, J.; Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems 32*, 2019; Vol. *32*.
(93) Fey, M.; Lenssen, J. E. Fast Graph Representation Learning with PyTorch Geometric. In *International Conference on Learning Representations*; New Orleans: USA, 2019.
(94) Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: Anchorage, AK, USA, 2019; pp 2623−2631.
(95) Bergstra, J.; Bardenet, R.; Bengio, Y.; Kégl, B. Algorithms for Hyper-Parameter Optimization. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Granada, Spain, 2011; pp 2546−2554.
(96) Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **2018**, *120*, 145301.
(97) Xu, K.; Hu, W.; Leskovec, J.; Jegelka, S. How Powerful are Graph Neural Networks? In *International Conference on Learning Representations*; New Orleans: USA, 2019.
(98) 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.
(99) 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.* **2020**, *63*, 8749−8760.
(100) pandas-dev/pandas: Pandas, Zenodo. The Pandas Development Team, 2020.
(101) Hopkins, B. W.; Tschumper, G. S. Ab Initio Studies of $\pi\cdots\pi$ Interactions: The Effects of Quadruple Excitations. *J. Phys. Chem. A* **2004**, *108*, 2941−2948.
(102) Schreiner, P. R.; Fokin, A. A.; Pascal, R. A.; de Meijere, A. Many Density Functional Theory Approaches Fail To Give Reliable Large Hydrocarbon Isomer Energy Differences. *Org. Lett.* **2006**, *8*, 3635−3638.
(103) de Meijere, A.; Lee, C.-H.; Kuznetsov, M. A.; Gusev, D. V.; Kozhushkov, S. I.; Fokin, A. A.; Schreiner, P. R. Preparation and Reactivity of $[D_{3d}]$-Octahedrane: The Most Stable $(CH)_{12}$ Hydrocarbon. *Chem. Eur. J.* **2005**, *11*, 6175−6184.
(104) Collins, E. M.; Raghavachari, K. A Fragmentation-Based Graph Embedding Framework for QM/ML. *J. Phys. Chem. A* **2021**, *125*, 6872−6880.
(105) Narayanan, B.; Redfern, P. C.; Assary, R. S.; Curtiss, L. A. Accurate Quantum Chemical Energies for 133 000 Organic Molecules. *Chem. Sci.* **2019**, *10*, 7449−7455.
(106) Tumer, K.; Ghosh, J. Estimating the Bayes Error Rate Through Classifier Combining. In *Proceedings of 13th International Conference on Pattern Recognition*, 1996; Vol. 2, pp 695−699.

https://doi.org/10.1021/acs.jctc.2c00501
J. Chem. Theory Comput. 2022, 18, 4846−4855

2.1.2 Machine Learning for Bridging the Gap Between Density Functional Theory and Coupled Cluster Energies

Abstract:



Accurate electronic energies and properties are crucial for successful reaction design and mechanistic investigations. Computing energies and properties of molecular structures has proven extremely useful, and, with increasing computational power, the limits of high-level approaches (such as coupled-cluster theory) are expanding to ever larger systems. However, because scaling is highly unfavorable, these methods are still not universally applicable to larger systems. To address the need for fast and accurate electronic energies of larger systems, we created a database of around 8000 small organic monomers (2000 dimers) optimized at the B3LYP-D3(BJ)/cc-pVTZ level of theory. This database also includes single-point energies computed at various levels of theory, including PBE1PBE, ωB97X, M06-2X, revTPSS, B3LYP, BP86, for density functional theory as well as DLPNO-CCSD(T) and CCSD(T) for coupled cluster, all in conjunction with a cc-pVTZ basis. We used this database to train machine learning models based on graph neural networks using two different graph representations. Our models are able to make energy predictions from B3LYP-D3(BJ)/cc-pVTZ inputs to CCSD(T)/cc-pVTZ outputs with an MAE of 0.78 and to DLPNO-CCSD(T)/cc-pVTZ with an MAE of 0.50 and 0.18 kcal mol^{-1} for monomers and dimers, respectively. The model for dimers was further validated on the S22 database, and the monomer model was tested on challenging systems, including such with highly conjugated or functionally complex molecules.

Reference: (DOI: 10.1021/acs.jctc.3c00274)

M. Ruth, D. Gerbig, P. R. Schreiner, Machine Learning for Bridging the Gap between Density Functional Theory and Coupled Cluster Energies, *J. Chem. Theory Comput.* **2023**, *19*, 15, 4912–4920.

JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

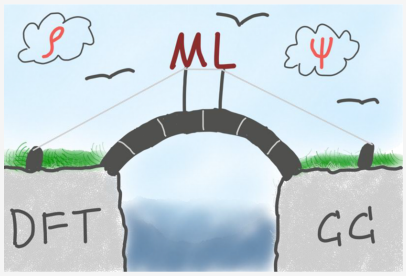
Machine Learning for Bridging the Gap between Density Functional Theory and Coupled Cluster Energies

Marcel Ruth, Dennis Gerbig, and Peter R. Schreiner*

Cite This: J. Chem. Theory Comput. 2023, 19, 4912−4920 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Accurate electronic energies and properties are crucial for successful reaction design and mechanistic investigations. Computing energies and properties of molecular structures has proven extremely useful, and, with increasing computational power, the limits of high-level approaches (such as coupled cluster theory) are expanding to ever larger systems. However, because scaling is highly unfavorable, these methods are still not universally applicable to larger systems. To address the need for fast and accurate electronic energies of larger systems, we created a database of around 8000 small organic monomers (2000 dimers) optimized at the B3LYP-D3(BJ)/cc-pVTZ level of theory. This database also includes single-point energies computed at various levels of theory, including PBE1PBE, ωB97X, M06-2X, revTPSS, B3LYP, and BP86, for density functional theory as well as DLPNO-CCSD(T) and CCSD(T) for coupled cluster theory, all in conjunction with a cc-pVTZ basis. We used this database to train machine learning models based on graph neural networks using two different graph representations. Our models are able to make energy predictions from B3LYP-D3(BJ)/cc-pVTZ inputs to CCSD(T)/cc-pVTZ outputs with a mean absolute error of 0.78 and to DLPNO-CCSD(T)/cc-pVTZ with an mean absolute error of 0.50 and 0.18 kcal mol^{-1} for monomers and dimers, respectively. The model for dimers was further validated on the S22 database, and the monomer model was tested on challenging systems, including those with highly conjugated or functionally complex molecules.



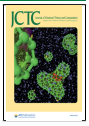
INTRODUCTION

Machine learning (ML) is a rapidly growing field within the broader field of artificial intelligence that involves the use of algorithms and statistical models to enable computers to learn and make decisions based on data. ML algorithms can be trained to perform a variety of tasks, including classification, regression, and clustering, by being fed large amounts of labeled data and adjusting their internal parameters to optimize their performance on the task at hand. Graph neural networks (GNNs) represent a rapidly expanding category within the domain of ML algorithms. Due to the inherent graph-like depiction of molecules in chemistry, GNNs hold the potential to extract vital information necessary for the accurate characterization of materials and molecular species.[1−3] A GNN operates by propagating information through the nodes and edges of a graph, iteratively updating the node and edge representations. This process captures the complex relationships within the graph structure, allowing the GNN to make informed predictions or classifications.[4]

In the field of chemistry, ML has been applied to a wide range of problems, including, e.g., the prediction of toxicity of chemical compounds,[5−8] the stability of metal−organic frameworks,[9] and the binding affinity of small molecules to protein targets.[10] In addition to these applications, ML has also been used in the interpretation of chemical simulations and experiments.[11,12] The use of ML in chemistry has the potential to greatly enhance our understanding of chemical systems and accelerate the discovery of new materials and drugs.[13−15]

Due to the adverse scaling of computational methods, especially for the "gold standard",[16] namely, coupled cluster (CC) theory including single and double excitations as well as perturbatively included triple excitations, CCSD(T), which scales as $O^2(N^8)$ with the number of basis functions N and occupied orbitals O, it is often impossible to compute accurate energies and properties of a target molecule. Many approximations have been employed to overcome this scaling problem.[17] A popular approximation for CC theory is the domain-based local pair natural orbital (DLPNO) approximation, which reduces the computational cost by a factor of two to four compared to the actual CC computation.[18−20] Still, the computational time remains prohibitive for large systems, e.g., those of biological importance for which computations

Received: March 9, 2023
Published: July 7, 2023



https://doi.org/10.1021/acs.jctc.3c00274
J. Chem. Theory Comput. 2023, 19, 4912−4920

would take years, decades, or even centuries on current hardware. One solution to overcome the computational effort is the use of ML. In recent years, there has been a growing interest in the use of ML techniques in computational chemistry to predict properties of chemical structures and to understand chemical reactions.[21,22] One of the main advantages of ML in computational chemistry is its ability to handle large and complex datasets, which can be difficult to analyze using traditional methods, e.g., design of experiments or linear free energy relationships. For example, ML algorithms have been used to predict the chemical structures of compounds[23,24] as well as the thermodynamic and spectroscopic properties of chemical systems.[25]

A common ML approach is the use of neural network potentials (NNPs), which have several advantages over empirical potentials[26] and are based on fitting functional forms to experimental or computational data to predict the energy and forces of atoms in molecules and solids.[27] These potentials are trained on large datasets of ab initio computations and use a multilayer neural network architecture to learn the underlying relationships between the atomic coordinates and the corresponding energies and forces. These potentials achieve accuracy comparable to density functional theory (DFT) methods, while being significantly faster and more scalable.[28] NNPs are quite versatile and have been applied for the prediction of a variety of chemical properties, including reaction energies,[29] vibrational frequencies,[30] and excited states.[31] In addition, NNPs have been used to study the structure and dynamics of materials at different length scales, including nanostructures[32] as well as disordered systems,[33] and are shown to be particularly effective in modeling systems with noncovalent interactions, such as those found in organic molecules and materials.[34,35] Even in the condensed phase, the accuracy of CCSD(T) has been achieved via NNPs for molecular simulations in water.[36,37] Overall, the use of NNPs has been growing in recent years due to their ability to provide accurate and efficient predictions of chemical and materials properties.

In addition to NNPs, there is the Δ-learning approach to learn molecular properties.[38] In this approach, a property determined at a low level of theory is used as an anchor-point for prediction of the difference between the low-level and the high-level results. It is common to use DFT as a low-level starting point due to its $O(N^4)$ scaling behavior. As the target-level theory, a wavefunction-based method such as CCSD(T) is often chosen.[39−42] We recently applied Δ-learning to evaluate the perturbatively included triples correction of the CCSD(T) approach utilizing correlation-consistent Dunning basis sets.[43] In the present work, we significantly expand our approach to bridge the gap between DFT and wavefunction theory. Other approaches to Δ-learn the difference between DFT and CCSD(T) shown by Bowman et al. were performed on potential energy surfaces for small organic molecules.[44] By combining the good DFT scaling behavior with the accuracy of CC theory, accurate single-point (SP) energies can be obtained at low computational cost. We chose B3LYP-D3(BJ)[45−47] as our main functional and (a) CCSD(T) for monomers (b) DLPNO-CCSD(T) for dimers as the target levels of theory; all approaches employed the cc-pVTZ basis set. For those tasks, we built a validation and train/test database.[43] For dimers, we generated a set of 2000 molecules by recursively combing a list of monomers, including derivatives of the S22[48] benchmark set to cover a broad range of functional groups and interactions. The total computational effort for B3LYP-D3(BJ)/cc-pVTZ optimizations required 3456 h wall time, compared to 77,020 h for DLPNO-CCSD(T) and 30,643 h for CCSD(T)/cc-pVTZ SP energies.[49] This illustrates the time and energy that could be saved by leveraging ML to obtain high-level electronic energies. As there is no general DFT functional that works well for all applications, we decided to use one of the most popular functionals, B3LYP—with Grimme's third-generation empirical dispersion corrections and Becke−Johnson damping D3(BJ)—to benchmark our approach. We tested various DFT functionals, such as PBE1PBE,[50,51] ωB97X,[52] M06-2X,[53] revTPSS,[54] B3LYP, and BP86[55,45] to ascertain the generalizability of our approach. Multiple features were extracted from DFT computations to provide a key signature for each molecule (cf. Supporting Information for feature importance). These features, along with the two graph types that were used in conjunction with state-of-the-art GNNs, are depicted in Figure 1.

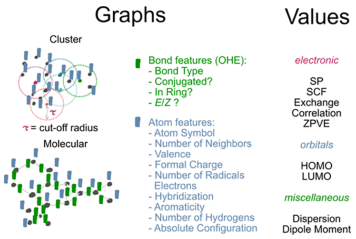


Figure 1. Overview of molecular inputs used in our models. Left: cluster and molecular graph representations; both encode hydrogens implicitly. The cluster graph is built based on the Euclidian distance between nodes. An edge between nodes is established if the distance is below a certain cut-off radius τ. The molecular graph is built based on covalent bonds in the molecule. Categorical features are one-hot encoded. Right: all properties which were extracted from DFT computations. These are final and SCF energies, DFT exchange and correlation, and frontier molecular orbital (i.e., HOMO/LUMO) energies. Additionally, the magnitudes of empirical dispersion and dipole moment as well as zero-point-vibrational-energies are included for each molecule as an additional *molecular signature*.

Models and Methods. We created a database by collecting around 8000 simplified molecular input line entry specification (SMILES) strings of small organic molecules (radicals and their hydrogen-terminated counterparts) from a public database.[56] This database contains only molecules consisting of hydrogen, carbon, nitrogen, and oxygen (cf. Supporting Information regarding our choice of elemental composition). For dimers, we generated around 2000 SMILES strings by combining all monomers from a hand-made set; details of the procedure can be found in the Supporting Information. The stated databases were used to initially train and test our models using a tenfold splitting procedure. We then validated our models using four external validation sets containing highly conjugated molecules, atmospherically relevant species (including radicals), noncovalent dimers (from the S22 Database), and constitutional isomers.

We used ChemML's[57] implementation of the Merck molecular force field (MMFF)[58] to compute the initial geometries of all the molecules contained within RDKit's[59] molecular objects. These geometries were then used for a

https://doi.org/10.1021/acs.jctc.3c00274
J. Chem. Theory Comput. 2023, 19, 4912−4920

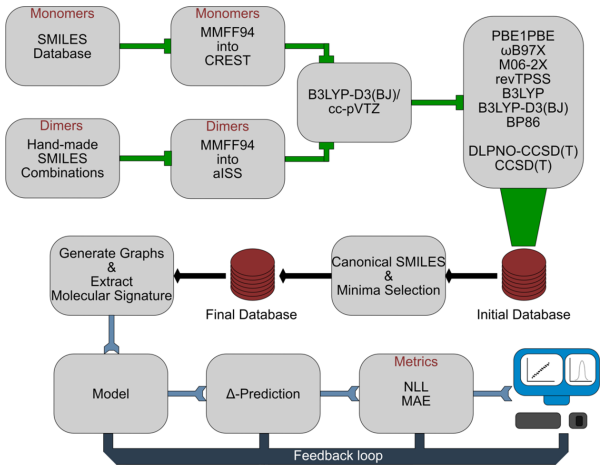


Figure 2. Schematic overview of our workflow, which starts with SMILES strings from a public database in the case of monomers and hand-crafted SMILES strings of dimers. An initial conformer is generated via MMFF94, and then, the optimal conformer is chosen by a CREST (aISS) run for the monomers (dimers). Each lowest-lying conformer is then optimized at B3LYP-D3(BJ)/cc-pVTZ, followed by an SP computation with various DFT functionals, DLPNO-CCSD(T), and CCSD(T), using the cc-pVTZ basis set. From the initial database, a selection of the lowest-energy geometries is made, from which the molecular inputs (graphs and computed properties) are extracted and used to train our models. The models' predictions are examined using the NLLLoss and MAE metrics. The results are visualized and used, in conjunction with the metrics, to optimize our general model architecture.

conformer search via CREST[60−64] in the case of monomers and aISS[65] in the case of dimers as implemented in the *bleeding edge* version of xTB.[66] Each lowest-lying conformer was subsequently optimized at B3LYP-D3(BJ)/cc-pVTZ, followed by an SP computation at either the DLPNO-CCSD(T)/cc-pVTZ (dimers and constituent monomers) or CCSD(T)/cc-pVTZ (monomers) level of theory—which was performed on 75% of the database—with the ORCA quantum chemical package.[67]

With the computed values extracted, we preprocessed our database by first generating canonical SMILES strings from the optimized Cartesian coordinates using RDKit. This resulted in a final database of 1163 monomers at the CCSD(T)/cc-pVTZ level of theory and 3126 monomers (2005 dimers) at the DLPNO-CCSD(T)/cc-pVTZ level of theory.

The process of generating molecular graphs involved using RDKit and DeepChem's node/edge featurizers.[68] The train/test splits for the data were created with scikit-learn[69] using stratified sampling to ensure balanced splits. PyTorch[70] was chosen as the framework for training and testing the model, and all of the graph neural network layers were implemented using PyTorch Geometric.[71] Model development was divided into four steps: optimization, training, testing, and validation. The optimization phase involved adjusting each model's hyperparameters using the Optuna[72] framework and the Tree-structured Parzen estimator algorithm.[73] All runs were monitored via "Weights & Biases" to adjusted hyperparameter ranges after multiple runs.[74]

As our probabilistic model architecture (*vide infra*) allows the prediction of mean μ and variance σ^2, we selected the Gaussian negative log likelihood loss (NLLLoss) as the minimized objective value. This probabilistic approach makes it possible to assess the aleatoric uncertainty (often called *data uncertainty*). Additionally, using ten different models (by using different data training; controlled via the random state), we simulated a deep ensemble to calculate the epistemic uncertainty (the intrinsic *model uncertainty*). We also calculated the mean absolute error (MAE) for comparison. Once we had determined the optimal hyperparameters, we trained the models using an early stopping mechanism to prevent overfitting. Finally, we conducted external validation using an out-of-sample dataset to assess each model's ability to generalize to new data and its effectiveness on challenging systems. A model trained on the whole train/test set, augmented with a set of 60 polycyclic molecules to provide additional polycyclic molecules for training purposes, was then used to evaluate a potential application to challenging $(CH)_{12}$ isomer scaffolds.[75,76] The incorporation of select polycyclic structures was deemed essential for achieving successful application to the aforementioned challenging scaffolds. A representation of our preprocessing and model development pipeline is depicted in Figure 2.

Our model architecture consists mainly of three different building blocks: two GNNs and one feed-forward neural network (FFNN). Each graph representation (cluster and molecular) is processed by one GNN module to result in two distinctive molecular embeddings. One GNN module therefore only learns covalent connections, while the other also includes connections that correspond to noncovalent intramolecular (or intermolecular, in the case of dimers) interactions. The embeddings are concatenated with features extracted from the DFT computations (*vide supra*, Figure 1) as well as two additional molecular fingerprints—Morgan fingerprint[77] and attentive fingerprint.[78] All combined features are channeled through an FFNN module that has two output neurons in the final layer: one to predict the mean μ and one for variance σ^2, respectively. Figure 3 depicts the schematic logic of the architecture. Full details about the exact layer

https://doi.org/10.1021/acs.jctc.3c00274
J. Chem. Theory Comput. 2023, 19, 4912−4920

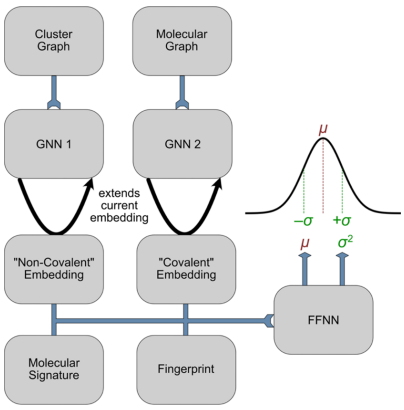


Figure 3. High-level representation of our model architecture. Each graph input (cluster and molecular) is fed into a GNN in which the embedding—generated by various pooling operations—is extended in each time step to incorporate the current embedding. The combined embeddings as well as the molecular signature values and fingerprints are subsequently concatenated and fed into the FFNN. The FFNN layers are densely connected and contain a dropout layer followed by an activation function to map nonlinear relationships. The output layer contains two neurons, one to predict the mean μ and one for variance σ^2, resulting in the prediction of a Gaussian distribution.

architecture, number of layers, etc., can be found in the Supporting Information in the hyperparameter section.

RESULTS AND DISCUSSION

To assess the performance of our models, we created kernel density estimation (KDE) plots for each model that we trained on our train/test data. The kernel density plots were created by calculating the deviations between the predicted and actual values for all ten models in each ensemble, combining these deviations, and then using the default parameters in the pandas library.[79] The resulting plots for each model are shown in Figure 4.

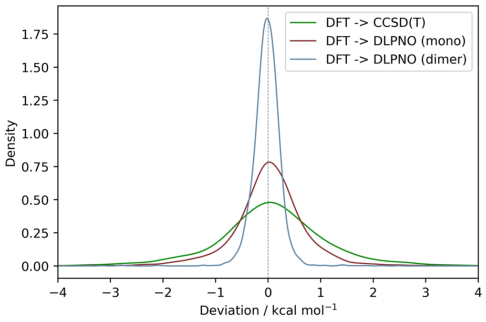


Figure 4. KDE plot of the differences between predicted and computed values for our models predicting the change in energy between B3LYP-D3(BJ)/cc-pVTZ and CCSD(T)/cc-pVTZ for monomers, and between B3LYP-D3(BJ)/cc-pVTZ and DLPNO-CCSD(T)/cc-pVTZ for monomers and dimers on the test set.

The error distribution is symmetric around 0 kcal mol^{-1}, which indicates that no systematic shift in the predicted increment is present. Prediction of DLPNO-CCSD(T)/cc-pVTZ energies seemingly works better than the prediction of true CCSD(T)/cc-pVTZ energies. An obvious reason for this could be the larger number of datapoints available for DLPNO-CCSD(T)/cc-pVTZ energies. The predictions of dimer energies show a sharper and narrower KDE compared to the monomers at DLPNO-CCSD(T)/cc-pVTZ, which is probably due to the more limited molecular space that is covered in our dimer training set. Most predictions have an error below 1 kcal mol^{-1} and are thus within reasonable chemical accuracy. The MAE for the prediction of CCSD(T)/cc-pVTZ energies is 0.78 kcal mol^{-1}, whereas that for the prediction of the DLPNO-CCSD(T)/cc-pVTZ energies is 0.50 (0.18) kcal mol^{-1} for the monomers (dimers). As the shown KDEs are for data that were taken from the same data source as the training set, information about the generalizability of our models is expected to be somewhat limited. Hence, we also performed KDEs for the external validation sets (Figure 5).

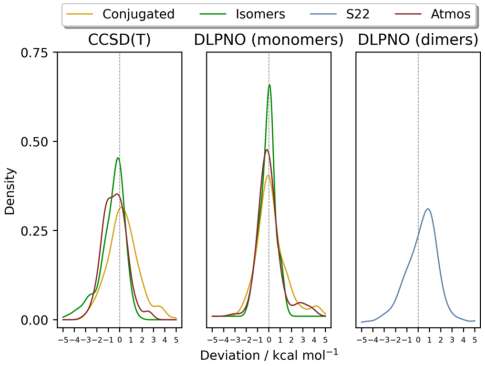


Figure 5. KDE plot of the differences between predicted and computed values for the changes in energies from B3LYP-D3(BJ)/cc-pVTZ to CCSD(T)/cc-pVTZ for monomers as well as from B3LYP-D3(BJ)/cc-pVTZ to DLPNO-CCSD(T)/cc-pVTZ for monomers and dimers. Results presented are from the external validation sets, which include highly conjugated (Conjugated), atmospherically relevant molecules (Atmos) as well as three different hydrocarbon isomers (Isomers) to evaluate the performance on monomers. The S22 set is used to assess the models' performances on dimers.

The error distributions are uniform around 0 kcal mol^{-1} for all validation sets except for the Atmos set, which shows a shift toward negative deviations, thus indicating systematic underprediction. As expected, the MAE on the validation sets is higher than that on the train/test set, resulting—over tenfold splitting—in 0.90 (1.06), 0.52 (0.87), 0.77 (0.88), and 1.17 kcal mol^{-1} for conjugated molecules, isomers, atmospherically relevant species, and the S22 dimers, respectively, for the prediction of DLPNO-CCSD(T)/cc-pVTZ (CCSD(T)/cc-pVTZ) energies. Even though most of the molecules contained in the validation sets are vastly different from the training data, an overall MAE below 1 kcal mol^{-1} was obtained. Figure 6 shows the overall increment span as a performance plot paired with an error assessment for the validation sets.

https://doi.org/10.1021/acs.jctc.3c00274
J. Chem. Theory Comput. 2023, 19, 4912−4920

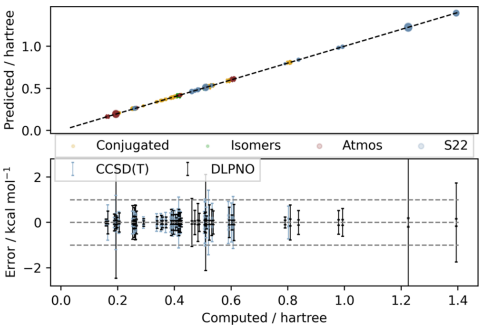


Figure 6. Depiction to capture the prediction performance along with the uncertainty of every prediction of our ensemble models. Top: performance plot (predicted vs computed) of our models on the external validation sets for predicting the energy increment between DFT and CC theory. The black-dashed diagonal indicates optimal prediction. Bottom: each data point from the performance plot at the top is mapped on the abscissa, i.e., located at its computed value. Shown are aleatoric (small error bar) and epistemic (larger error bar) uncertainties of our models for predicting the difference between B3LYP-D3(BJ)/cc-pVTZ and CCSD(T)- and DLPNO-CCSD(T)/cc-pVTZ energies in blue and black, respectively. The dashed gray lines indicate the ±1 kcal mol^{-1} error margin.

As shown in Figure 6, the predictions lie on the dashed diagonal, which is optimal. The low epistemic uncertainty (mostly below 1 kcal mol^{-1}), which can be seen in the lower part of the figure, indicates the high certainty of our models; in other words, the models within an ensemble predict similar energies.

The two large uncertainty bars—at 1.2 and 1.4 hartree on the abscissa in Figure 6—are worth noting. These belong, from left to right, to the uracil-uracil (3.1 kcal mol^{-1} error bar) and adenine-thymine dimers (1.7 kcal mol^{-1} error bar). As the increments between DFT and CC theory are quite different from those in the training set and their chemical interactions are rather intricate, a high variance in the error prediction seems reasonable. The large uncertainty around 0.5 hartree corresponds to the formic acid dimer (2.1 kcal mol^{-1} error bar), which was predicted accurately by all but one of the ten models, which results in a large variance. The only monomeric species for which our models was uncertain was the hydroperoxyl radical (2.5 kcal mol^{-1}), located at 0.2 hartree, but only for our DLPNO-CCSD(T) ensemble. Overall, aleatoric uncertainty is always smaller than epistemic uncertainty, which indicates that our models still remain somewhat uncertain about the parameters of the underlying data distribution. This usually occurs when a model has not seen enough data to learn their underlying patterns completely. In this case, providing more data or using an even more expressive model architecture may help reduce the epistemic uncertainty. The overall reliability of a prediction for a given molecule can be assessed by examining the standard deviation of the ensemble predictions (representing the epistemic uncertainty). In addition, it is also crucial to consider the ensemble variance, as larger molecules have more substantial increments to predict, which may lead to increased variance within the ensemble. Our model predicts not only the mean value but also the Gaussian distribution, making it essential to differentiate and evaluate the predicted Gaussian variance as well.

In addition, we highlight the three best and worst predictions for each validation set and theory level in Figure 7 to provide a more approachable representation of the capacity and limitations of our architecture. Energies of saturated molecules are predicted well throughout, while those of unsaturated and conjugated molecules are predicted less accurately. Note that the two predictions of nucleotide dimers with high uncertainties mentioned above are not the worst predictions of the ensemble. This means that the constituent models compensate each other to result in an uncertain, but accurate prediction for those two cases. A high ensemble error only occurs if all models predict the energy increment systematically wrong. This is especially prominent for 1*H*-azirine and cyclobutadiene (singlet, C_{2h}) at both the DLPNO-CCSD(T)/cc-pVTZ and CCSD(T)/cc-pVTZ levels of theory. Predictions that are found to be difficult also are challenging for both high-level theories.

In Figure 8, we provide a cherry-picked example of one potential use case of our CCSD(T)/cc-pVTZ model. As shown previously, many DFT functionals are unreliable in accurately describing the energies of the $(CH)_{12}$ isomers.[75,76,80] We previously showed that our CCSD(ML-T) model can also provide accurate isomer energies, and this is related to its CCSD anchor point. Now, using the DFT results—which are not only wrong with respect to the magnitude of energy differences, but also to the relative ordering of energy levels—we can even *correct* the faulty DFT values via ML.

As shown in Figure 8, the most stable $(CH)_{12}$ isomer is octahedrane. B3LYP-D3(BJ)/cc-pVTZ and many other DFT approaches wrongly predict the displayed tetraene to be the lowest-lying isomer.[81] Also, the relative energy difference between the three shown isomers is only around 4 kcal mol^{-1} at B3LYP-D3(BJ)/cc-pVTZ for the lowest- and highest-lying isomers, whereas CCSD(T)/cc-pVTZ SP energies indicate an energy difference of over 20 kcal mol^{-1} (cf. Supporting Information for a depiction of the relative isomer energy of all 38 molecules). Thus, many DFT approaches for such systems and very likely related hydrocarbons are highly inaccurate, both qualitatively and quantitatively. Fortunately, our ML-CCSD(T)/cc-pVTZ model corrects the relative energy trend and the magnitude of the relative energy differences. The MAE of the relative isomer energies for all 38 $(CH)_{12}$ isomers relative to CCSD(T)/cc-pVTZ is 12.9 kcal mol^{-1} for B3LYP-D3(BJ)/cc-pVTZ and 2.3 kcal mol^{-1} for the ML-corrected energies. This indicates that our model can truly work with the molecular structure, embedded as molecular and cluster graphs, to accurately achieve results close to CCSD(T)/cc-pVTZ quality.

CONCLUSIONS

We demonstrate that it is possible to bridge the gap between DFT and wavefunction-based methods by employing an ML model to estimate the incremental energy differences. Our models display an MAE of 0.78 kcal mol^{-1} for predicting CCSD(T)/cc-pVTZ monomers and 0.50 kcal mol^{-1} for DLPNO-CCSD(T)/cc-pVTZ monomers, with an MAE of 0.18 kcal mol^{-1} for dimers, all based on B3LYP-D3(BJ)/cc-pVTZ computations. The generalizability of our models was shown by testing them on a diverse set of molecules. Our

https://doi.org/10.1021/acs.jctc.3c00274
J. Chem. Theory Comput. 2023, 19, 4912−4920

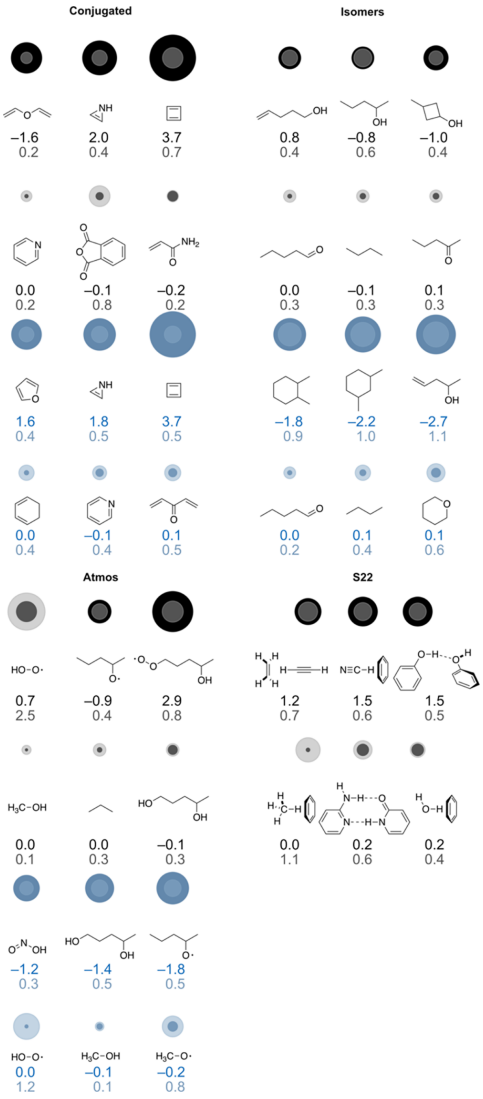


Figure 7. Three best and worst predictions of our deep ensemble for each validation set. The darker color corresponds to the ensemble error, while each lighter color denotes the epistemic error of the prediction. Disk radii are made proportional to the absolute errors to provide a visual representation of the deviations. Blue denotes predictions to the DLPNO-CCSD(T)/cc-pVTZ and black denotes predictions to the CCSD(T)/ccpVTZ level of theory. All values are given in kcal mol^{-1}.

models can correctly identify energetic trends and magnitudes even for difficult hydrocarbon scaffolds from DFT results.

The achieved accuracy is similar to our previous approach in which we predicted only the effect of perturbatively included triples.[43] The current model iteration now comes at a much lower computational cost due to the excellent scaling of DFT methods, and it successfully captures the noncovalent interactions of, e.g., molecular dimers.

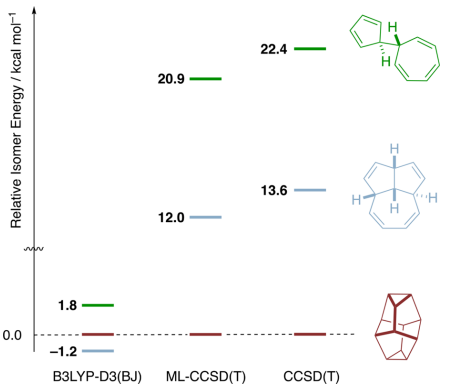


Figure 8. Relative isomer energies of three $(CH)_{12}$ isomers.[75,76,80] Not only does B3LYP-D3(BJ)/cc-pVTZ underestimate the isomer energy differences compared to CCSD(T)/cc-pVTZ, but it also results in an incorrect relative energy ordering. Using our model, denoted as ML-CCSD(T), it is possible to restore the correct the relative trend and magnitude of the relative isomer energies.

As compared to the Δ-learning approach by von Lilienfeld and co-workers, which even uses a narrower chemical space but works with HF energies, we achieved comparable accuracy with fewer training data.[38] The prominent transfer learning approach by Smith et al., ANI-1ccx, working with the ωB97X functional to predict CCSD(T)*/CBS energies, has also resulted in less accurate predictions but also works for multiple conformers of the same molecular species.[82]

Expansion of this work could encompass increasing the number of supported elements, e.g., to include halogens and the third row chalcogens and pnictogens. Additionally, as our models were only trained on minima, an approach to predict non-stationary points would give rise to an even more powerful model. Developing a more expressive model, via an even better graph representation and molecular signature, as indicated and discussed in Figure 6, would also be a reasonable evolutionary step toward a more powerful model and could in principle be universally applicable.

■ ASSOCIATED CONTENT

Data Availability Statement

Databases containing electronic energies and other input features along with SMILES and geometries are provided as CSV files. The trained models, along with a suitable Python script for predictions, will also be available free of charge at http://dx.doi.org/10.22029/jlupub-9418.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.3c00274.

Architecture, training, and optimization of the models, and computational details (PDF)

https://doi.org/10.1021/acs.jctc.3c00274
J. Chem. Theory Comput. 2023, 19, 4912−4920

AUTHOR INFORMATION

Corresponding Author

Peter R. Schreiner − *Institute of Organic Chemistry, Justus Liebig University, 35392 Giessen, Germany;* orcid.org/0000-0002-3608-5515; Email: prs@uni-giessen.de

Authors

Marcel Ruth − *Institute of Organic Chemistry, Justus Liebig University, 35392 Giessen, Germany*

Dennis Gerbig − *Institute of Organic Chemistry, Justus Liebig University, 35392 Giessen, Germany;* orcid.org/0000-0002-7023-8298

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.3c00274

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft within the priority program "Utilization and Development of Machine Learning for Molecular Applications − Molecular Machine Learning" (SPP 2363, Schr 597/41-1). M.R. thanks the Fonds der Chemischen Industrie for a doctoral scholarship.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Wieder, O.; Kohlbacher, S.; Kuenemann, M.; Garon, A.; Ducrot, P.; Seidel, T.; Langer, T. A Compact Review of Molecular Property Prediction with Graph Neural Networks. *Drug Discov. Today Technol.* **2020**, *37*, 1−12.

(2) Fung, V.; Zhang, J.; Juarez, E.; Sumpter, B. G. Benchmarking Graph Neural Networks for Materials Chemistry. *npj Comput. Mater.* **2021**, *7*, 84.

(3) Reiser, P.; Neubert, M.; Eberhard, A.; Torresi, L.; Zhou, C.; Shao, C.; Metni, H.; van Hoesel, C.; Schopmans, H.; Sommer, T.; Friederich, P. Graph Neural Networks for Materials Science and Chemistry. *Commun. Mater.* **2022**, *3*, 93.

(4) Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; Monfardini, G. The Graph Neural Network Model. *IEEE Trans. Neural Netw.* **2009**, *20*, 61−80.

(5) Raies, A. B.; Bajic, V. B. In Silico Toxicology: Computational Methods for the Prediction of Chemical Toxicity. *WIREs Comput. Mol. Sci.* **2016**, *6*, 147−172.

(6) Hemmerich, J.; Ecker, G. F. In Silico Toxicology: From Structure−Activity Relationships Towards Deep Learning and Adverse Outcome Pathways. *WIREs Comput. Mol. Sci.* **2020**, *10*, No. e1475.

(7) Pérez Santín, E.; Rodríguez Solana, R.; González García, M.; García Suárez, M. D. M.; Blanco Díaz, G. D.; Cima Cabal, M. D.; Moreno Rojas, J. M.; López Sánchez, J. I. Toxicity Prediction Based on Artificial Intelligence: A Multidisciplinary Overview. *WIREs Comput. Mol. Sci.* **2021**, *11*, No. e1516.

(8) Wang, M. W. H.; Goodman, J. M.; Allen, T. E. H. Machine Learning in Predictive Toxicology: Recent Applications and Future Directions for Classification Models. *Chem. Res. Toxicol.* **2021**, *34*, 217−239.

(9) Pétuya, R.; Durdy, S.; Antypov, D.; Gaultois, M. W.; Berry, N. G.; Darling, G. R.; Katsoulidis, A. P.; Dyer, M. S.; Rosseinsky, M. J. Machine-Learning Prediction of Metal−Organic Framework Guest Accessibility from Linker and Metal Chemistry. *Angew. Chem., Int. Ed.* **2022**, *61*, No. e202114573.

(10) Gentile, F.; Agrawal, V.; Hsing, M.; Ton, A.-T.; Ban, F.; Norinder, U.; Gleave, M. E.; Cherkasov, A. Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery. *ACS Cent. Sci.* **2020**, *6*, 939−949.

(11) Häse, F.; Galván, I. F.; Aspuru-Guzik, A.; Lindh, R.; Vacher, M. How Machine Learning Can Assist the Interpretation of Ab Initio Molecular Dynamics Simulations and Conceptual Understanding of Chemistry. *Chem. Sci.* **2019**, *10*, 2298−2307.

(12) 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*, 1695.

(13) Schmidt, J.; Marques, M. R. G.; Botti, S.; Marques, M. A. L. Recent Advances and Applications of Machine Learning in Solid-State Materials Science. *npj Comput. Mater.* **2019**, *5*, 83.

(14) Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; Zhao, S. Applications of Machine Learning in Drug Discovery and Development. *Nat. Rev. Drug Discov.* **2019**, *18*, 463−477.

(15) Tkatchenko, A. Machine Learning for Chemical Discovery. *Nat. Commun.* **2020**, *11*, 4125.

(16) Crawford, T.; Schaefer III, H. F., *An Introduction to Coupled Cluster Theory for Computational Chemists*. 2007, *14*, 33−136.

(17) Izsák, R. Single-Reference Coupled Cluster Methods for Computing Excitation Energies in Large Molecules: The Efficiency and Accuracy of Approximations. *WIREs Comput. Mol. Sci.* **2020**, *10*, No. e1445.

(18) Guo, Y.; Riplinger, C.; Becker, U.; Liakos, D. G.; Minenkov, Y.; Cavallo, L.; Neese, F. Communication: An Improved Linear Scaling Perturbative Triples Correction for the Domain Based Local Pair-Natural Orbital Based Singles and Doubles Coupled Cluster Method [DLPNO-CCSD(T)]. *J. Chem. Phys.* **2018**, *148*, No. 011101.

(19) Liakos, D. G.; Guo, Y.; Neese, F. Comprehensive Benchmark Results for the Domain Based Local Pair Natural Orbital Coupled Cluster Method (DLPNO-CCSD(T)) for Closed- and Open-Shell Systems. *J. Phys. Chem. A* **2020**, *124*, 90−100.

(20) Sandler, I.; Chen, J.; Taylor, M.; Sharma, S.; Ho, J. Accuracy of DLPNO-CCSD(T): Effect of Basis Set and System Size. *J. Phys. Chem. A* **2021**, *125*, 1553−1563.

(21) Stocker, S.; Csányi, G.; Reuter, K.; Margraf, J. T. Machine learning in chemical reaction space. *Nat. Commun.* **2020**, *11*, 5505.

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

(23) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559*, 547−555.

(24) 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.

(25) Joung, J. F.; Han, M.; Hwang, J.; Jeong, M.; Choi, D. H.; Park, S. Deep Learning Optical Spectroscopy Based on Experimental Database: Potential Applications to Molecular Design. *JACS Au* **2021**, *1*, 427−438.

(26) Gastegger, M.; Marquetand, P. High-Dimensional Neural Network Potentials for Organic Reactions and an Improved Training Algorithm. *J. Chem. Theory Comput.* **2015**, *11*, 2187−2198.

(27) Kocer, E.; Ko, T. W.; Behler, J. Neural Network Potentials: A Concise Overview of Methods. *Annu. Rev. Phys. Chem.* **2022**, *73*, 163−186.

(28) Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K.-R.; Anatole von Lilienfeld, O. Machine Learning of Molecular Electronic Properties in Chemical Compound Space. *New J. Phys.* **2013**, *15*, No. 095003.

(29) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chem. Sci.* **2017**, *8*, 3192−3203.

(30) Gastegger, M.; Behler, J.; Marquetand, P. Machine Learning Molecular Dynamics for the Simulation of Infrared Spectra. *Chem. Sci.* **2017**, *8*, 6924−6935.

https://doi.org/10.1021/acs.jctc.3c00274
J. Chem. Theory Comput. 2023, 19, 4912−4920

(31) Behler, J. Neural Network Potential-Energy Surfaces in Chemistry: A Tool for Large-Scale Simulations. *Phys. Chem. Chem. Phys.* **2011**, *13*, 17930−17955.
(32) Álvarez-Zapatero, P.; Vega, A.; Aguado, A. A Neural Network Potential for Searching the Atomic Structures of Pure and Mixed Nanoparticles. Application to ZnMg Nanoalloys with an Eye on Their Anticorrosive Properties. *Acta Mater.* **2021**, *220*, No. 117341.
(33) Takamoto, S.; Shinagawa, C.; Motoki, D.; Nakago, K.; Li, W.; Kurata, I.; Watanabe, T.; Yayama, Y.; Iriguchi, H.; Asano, Y.; Onodera, T.; Ishii, T.; Kudo, T.; Ono, H.; Sawada, R.; Ishitani, R.; Ong, M.; Yamaguchi, T.; Kataoka, T.; Hayashi, A.; Charoenphakdee, N.; Ibuka, T. Towards Universal Neural Network Potential for Material Discovery Applicable to Arbitrary Combination of 45 Elements. *Nat. Commun.* **2022**, *13*, 2991.
(34) Behler, J. First Principles Neural Network Potentials for Reactive Simulations of Large Molecular and Condensed Systems. *Angew. Chem., Int. Ed.* **2017**, *56*, 12828−12840.
(35) Gokcan, H.; Isayev, O. Learning Molecular Potentials with Neural Networks. *WIREs Comput. Mol. Sci.* **2022**, *12*, No. e1564.
(36) Daru, J.; Forbert, H.; Behler, J.; Marx, D. Coupled Cluster Molecular Dynamics of Condensed Phase Systems Enabled by Machine Learning Potentials: Liquid Water Benchmark. *Phys. Rev. Lett.* **2022**, *129*, No. 226001.
(37) Chen, M. S.; Lee, J.; Ye, H.-Z.; Berkelbach, T. C.; Reichman, D. R.; Markland, T. E. Data-Efficient Machine Learning Potentials from Transfer Learning of Periodic Correlated Electronic Structure Methods: Liquid Water at AFQMC, CCSD, and CCSD(T) Accuracy. *J. Chem. Theory Comput.* **2023**, DOI: 10.1021/acs.jctc.2c01203.
(38) 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.
(39) Townsend, J.; Vogiatzis, K. D. Data-Driven Acceleration of the Coupled-Cluster Singles and Doubles Iterative Solver. *J. Phys. Chem. Lett.* **2019**, *10*, 4129−4135.
(40) Dick, S.; Fernandez-Serra, M. Machine Learning Accurate Exchange and Correlation Functionals of the Electronic Density. *Nat. Commun.* **2020**, *11*, 3509.
(41) Ikabata, Y.; Fujisawa, R.; Seino, J.; Yoshikawa, T.; Nakai, H. Machine-Learned Electron Correlation Model Based on Frozen Core Approximation. *J. Chem. Phys.* **2020**, *153*, 184108.
(42) 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.
(43) 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.
(44) Bowman, J. M.; Qu, C.; Conte, R.; Nandi, A.; Houston, P. L.; Yu, Q. Δ-Machine Learned Potential Energy Surfaces and Force Fields. *J. Chem. Theory Comput.* **2023**, *19*, 1−17.
(45) 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.
(46) Becke, A. D. A New Mixing of Hartree−Fock and Local Density-Functional Theories. *J. Chem. Phys.* **1993**, *98*, 1372−1377.
(47) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011**, *32*, 1456−1465.
(48) Takatani, T.; Hohenstein, E. G.; Malagoli, M.; Marshall, M. S.; Sherrill, C. D. Basis Set Consistent Revision of the S22 Test Set of Noncovalent Interaction Energies. *J. Chem. Phys.* **2010**, *132*, 144104.
(49) Typically on Intel Xeon processors of the Cascade Lake or Skylake architecture families.
(50) Adamo, C.; Barone, V. Toward Reliable Density Functional Methods Without Adjustable Parameters: The PBE0 Model. *J. Chem. Phys.* **1999**, *110*, 6158−6170.
(51) Ernzerhof, M.; Scuseria, G. E. Assessment of the Perdew−Burke−Ernzerhof Exchange-Correlation Functional. *J. Chem. Phys.* **1999**, *110*, 5029−5036.
(52) Chai, J.-D.; Head-Gordon, M. Systematic Optimization of Long-Range Corrected Hybrid Density Functionals. *J. Chem. Phys.* **2008**, *128*, No. 084106.
(53) 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.
(54) Perdew, J. P.; Ruzsinszky, A.; Csonka, G. I.; Constantin, L. A.; Sun, J. Workhorse Semilocal Density Functional for Condensed Matter Physics and Quantum Chemistry. *Phys. Rev. Lett.* **2009**, *103*, No. 026403.
(55) Perdew, J. P. Density-Functional Approximation for the Correlation Energy of the Inhomogeneous Electron Gas. *Phys. Rev. B* **1986**, *33*, 8822−8824.
(56) John, P. C.; Guan, Y.; Kim, Y.; Etz, B. D.; Kim, S.; Paton, R. S. Quantum Chemical Calculations for Over 200,000 Organic Radical Species and 40,000 Associated Closed-Shell Molecules. *Sci. Data* **2020**, *7*, 244.
(57) Haghighatlari, M.; Vishwakarma, G.; Altarawy, D.; Subramanian, R.; Kota, B. U.; Sonpal, A.; Setlur, S.; Hachmann, J. ChemML: A Machine Learning and Informatics Program Package for the Analysis, Mining, and Modeling of Chemical and Materials Data. *WIREs Comput. Mol. Sci.* **2020**, *10*, No. e1458.
(58) Tosco, P.; Stiefl, N.; Landrum, G. Bringing the MMFF Force Field to the RDKit: Implementation and Validation. *J. Cheminform.* **2014**, *6*, 37.
(59) Landrum, G. *RDKit: Open-Source Cheminformatics Software*. https://www.rdkit.org. https://zenodo.org/record/6961488#.Y9znXezMKPS, (accessed 10.22.22).
(60) Pracht, P.; Bauer, C. A.; Grimme, S. Automated and Efficient Quantum Chemical Determination and Energetic Ranking of Molecular Protonation Sites. *J. Comput. Chem.* **2017**, *38*, 2618−2631.
(61) Grimme, S. Exploration of Chemical Compound, Conformer, and Reaction Space with Meta-Dynamics Simulations Based on Tight-Binding Quantum Chemical Calculations. *J. Chem. Theory Comput.* **2019**, *15*, 2847−2862.
(62) 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.
(63) Pracht, P.; Grimme, S. Calculation of Absolute Molecular Entropies and Heat Capacities Made Simple. *Chem. Sci.* **2021**, *12*, 6551−6568.
(64) Spicher, S.; Plett, C.; Pracht, P.; Hansen, A.; Grimme, S. Automated Molecular Cluster Growing for Explicit Solvation by Efficient Force Field and Tight Binding Methods. *J. Chem. Theory Comput.* **2022**, *18*, 3174−3189.
(65) Plett, C.; Grimme, S. Automated and Efficient Generation of General Molecular Aggregate Structures. *Angew. Chem., Int. Ed.* **2023**, *62*, No. e202214477.
(66) Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S. Extended Tight-Binding Quantum Chemistry Methods. *WIREs Comput. Mol. Sci.* **2021**, *11*, No. e1493.
(67) Neese, F. Software update: The ORCA program system—Version 5.0. *WIREs Comput. Mol. Sci.* **2022**, *12*, No. e1606.
(68) Ramsundar, B.; Eastman, P.; Walters, P.; Pande, V.; Leswing, K.; Wu, Z., *Deep Learning for the Life Sciences*; O'Reilly Media: 2019.
(69) 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.* **2011**, *12*, 2825−2830.
(70) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.;

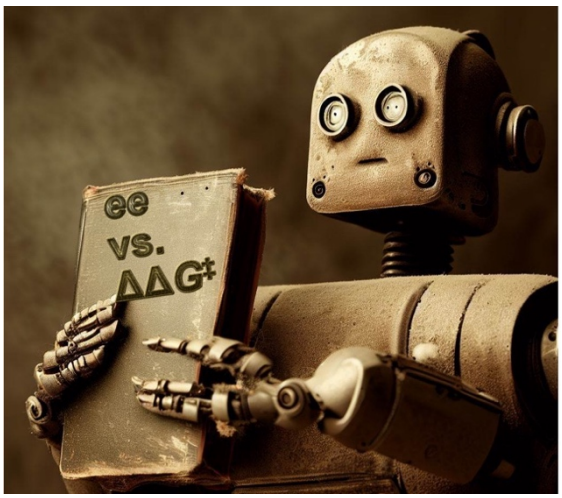
Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steinerand, B.; Fang, L.; Bai, J.; Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 8026−8037.
(71) Fey, M.; Lenssen, J. E., Fast Graph Representation Learning with PyTorch Geometric. In *International Conference on Learning Representations*; New Orleans, USA, 2019.
(72) Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M., Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: Anchorage, AK, USA, 2019, 2623−2631.
(73) Bergstra, J.; Bardenet, R.; Bengio, Y.; Kégl, B., Algorithms for Hyper-Parameter Optimization. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Granada, Spain, 2011, 2546−2554.
(74) Biewald, L. *Experiment Tracking with Weights and Biases*; Weights & Biases. http://wandb.com (accessed 10.22.22).
(75) de Meijere, A.; Lee, C.-H.; Kuznetsov, M. A.; Gusev, D. V.; Kozhushkov, S. I.; Fokin, A. A.; Schreiner, P. R. Preparation and Reactivity of $[D_{3d}]$-Octahedrane: The Most Stable $(CH)_{12}$ Hydrocarbon. *Chem. − Eur. J.* **2005**, *11*, 6175−6184.
(76) Schreiner, P. R.; Fokin, A. A.; Pascal, R. A., Jr.; de Meijere, A. Many Density Functional Theory Approaches Fail To Give Reliable Large Hydrocarbon Isomer Energy Differences. *Org. Lett.* **2006**, *8*, 3635−3638.
(77) 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.
(78) 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.* **2020**, *63*, 8749−8760.
(79) *pandas-dev/pandas: Pandas*; Zenodo: The Pandas Development Team, 2020.
(80) Karton, A.; Schreiner, P. R.; Martin, J. M. L. Heats of Formation of Platonic Hydrocarbon Cages by Means of High-Level Thermochemical Procedures. *J. Comput. Chem.* **2016**, *37*, 49−58.
(81) Zhao, Y.; Truhlar, D. G. A Density Functional That Accounts for Medium-Range Correlation Energies in Organic Chemistry. *Org. Lett.* **2006**, *8*, 5753−5755.
(82) Smith, J. S.; Zubatyuk, R.; Nebgen, B.; Lubbers, N.; Barros, K.; Roitberg, A. E.; Isayev, O.; Tretiak, S. The ANI-1ccx and ANI-1x Data Sets, Coupled-Cluster and Density Functional Theory Properties for Molecules. *Sci. Data* **2020**, *7*, 134.

https://doi.org/10.1021/acs.jctc.3c00274
J. Chem. Theory Comput. 2023, 19, 4912−4920

2.2 Prepared Manuscripts

2.2.1 Rewriting the Rules: Contrasting Historical and Physical Perspectives in Asymmetric Catalysis

Abstract:



Modeling of catalyzed enantioselective reactions has a strong history, starting with ligand-based and quantitative structure-activity relationships. With the rise of ML due to the increased power of graphic processing units, the modeling of chemical systems has reached a new era. We briefly dive into the history of ee and discuss the benefits of using physically-sound targets (i.e., Gibbs free energy difference of enantio-determining structures $\Delta\Delta G^{\ddagger}$) over historically based ones like enantiomeric excess. Ranging from slight performance increase when choosing $\Delta\Delta G^{\ddagger}$, escaping physical limitations, mitigating possible temperature effects on modeling, non-constant error conversions, data distributions based on domain, and how to deal with unphysical predictions in the ee domain. For this endeavor, we gathered eleven datasets from literature covering different reaction types, e.g., hydrogenation, Suzuki-, and Heck-reactions for 2761 data points. We evaluated fingerprint, descriptor, and graph neural network based models to decipher between different model complexities.

Version: September 18, 2023

Rewriting the Rules: Contrasting Historical and Physical Perspectives in Asymmetric Catalysis

Marcel Ruth,[‡[a]] Tobias Gensch,[‡,*] and Peter R. Schreiner*

[*] Marcel Ruth, Prof. Dr. Peter R. Schreiner
Institute of Organic Chemistry, Justus-Liebig-University
Heinrich-Buff-Ring 58, 35392 Giessen (Germany)
E-mail: prs@uni-giessen.de

Dr. Tobias Gensch
Institute of Chemistry, TU Berlin
Straße des 17. Juni 135, 10623 Berlin (Germany)
E-mail: tobias.gensch@tu-berlin.de

[‡] These authors contributed equally to this work.

Supporting information for this article is given via a link at the end of the document.

Abstract: Modeling of catalyzed enantioselective reactions has a strong history, starting with ligand-based and quantitative structure-activity relationships. With the rise of ML due to the increased power of graphic processing units, the modeling of chemical systems has reached a new era. We briefly dive into the history of ee and discuss the benefits of using physically-sound targets (*i.e.*, Gibbs free energy difference of enantio-determining structures $\Delta\Delta G^{\ddagger}$) over historically based ones like enantiomeric excess. Ranging from slight performance increase when choosing $\Delta\Delta G^{\ddagger}$, escaping physical limitations, mitigating possible temperature effects on modeling, non-constant error conversions, data distributions based on domain, and how to deal with unphysical predictions in the ee domain. For this endeavor, we gathered eleven datasets from literature covering different reaction types, *e.g.*, hydrogenation, Suzuki-, and Heck-reactions for 2761 data points. We evaluated fingerprint, descriptor, and graph neural network based models to decipher between different model complexities.

1. Introduction

Enantioselective catalysis, a challenging yet essential aspect of synthesizing pharmaceutically active compounds, represents a critical hurdle in developing and marketing these substrates. The precision required in the process makes the quest for effective catalysts a formidable task. However, emerging methodologies such as computational modeling offer promising avenues to streamline the process and expedite discovery and optimization cycles. Modeling, a sophisticated tool for unraveling complex catalytic mechanisms, acts as an innovative conduit for better understanding these phenomena, subsequently aiding the development of improved catalysts. This method is not simply about digitizing and abstracting chemical reactions; it serves as a digital twin, capable of mimicking real-life chemistry, thereby allowing for efficient, virtual trials of different catalysts under varying conditions.

Although the incorporation of Machine Learning (ML) in catalyst discovery and optimization is a comparatively newer trend, it should be noted that the underlying principles are only partially novel. Both ligand-based and quantitative structure-activity relationship models have been used for over three decades, setting a solid foundation for the integration of AI in the field.[1] The “newness” or revival can be attributed more to the sophisticated, data-driven approaches introduced by ML, which significantly enhance the accuracy and predictive power of these models. These improvements are possible thanks to the significant rise in processing power from devices like graphics processing units and access to more extensive and detailed data sets.[2]

However, as the merge of ML and catalysis is in its infancy, best practices for its successful application are yet to be firmly established. While preliminary guidelines exist for implementing ML in other fields, careful validation and methodological clarity will be needed to ensure its reliable application in catalyst discovery and optimization. Scientists are actively contributing to this field, setting out to establish these best practices and expand the potential of this exciting interdisciplinary approach.[3-5]

Enantiomeric excess (ee) and enantiomeric ratio (er) are both experimental quantities used to describe the selectivity and efficiency of asymmetric catalysts in producing one enantiomer over the other in a chiral compound synthesis.[6] However, they have different origins, historical usage, and reasons for their application as ee has been historically used in asymmetric catalysis and has its roots in the early days of enantiomer separation and analysis. It is the difference between the mole fractions of the two enantiomers in the product mixture. The ee is typically expressed as a percentage ranging from 0% (racemic mixture) to 100% (single enantiomer). The concept of enantiomeric excess was historically introduced—by measuring optical activity, *e.g.*, rotation of linear-polarized light—as a simple and intuitive way to describe the selectivity of a chiral synthesis or resolution process, making it easy to understand and communicate. Using er emerged as an alternative to ee due to its direct measurability in high-performance liquid chromatography, particularly in kinetic resolutions and asymmetric catalysis, where the two enantiomers' formation rate is more relevant.[7] The enantiomeric ratio is defined as the ratio of the rate constants for

forming each enantiomer or the ratio of their respective concentrations at a specific point in time during the reaction.

Enantioselective reactions are typically subject to the Curtin−Hammett principle, which means the enantiomeric products are formed from diastereomeric intermediates (such as substrate-catalyst complexes) through an irreversible (i.e., stereo-defining) transition structure. If the intermediates can interconvert quickly, the product ratio only depends on the difference of transition structure free energies. Thus, $\Delta\Delta G^{\ddagger}$ as the relative free energy of activation of the enantio-determining pathways quantifies the product's direct underlying cause for chiral purity. In contrast, ee and er are the experimental observables that derive from how chiral purity can be measured analytically. By using a quantity measuring the underlying cause of a process as the target in regression modeling, physically more meaningful models may be obtained. In reality, other factors can influence the chiral purity of reaction products, such as processes that lead to the racemization of the products or the formation of enantiomeric products by different mechanisms (such as a competition between enantiospecific reactions on enantiomerically enriched starting materials).

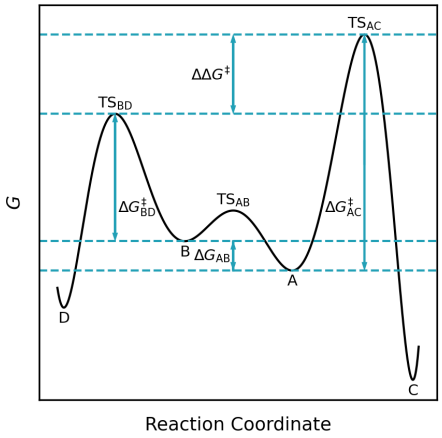


Figure 1. Exemplary potential free energy surface showing two intermediates **A** and **B,** which are in equilibrium with each other. Intermediate **A** can react further over **TS2** with a barrier of $\Delta G_2^{\ddagger}$ to the thermodynamic product **C**, while **A** could also interconvert to **B**, with a barrier of $\Delta G°$, which then further reacts to the kinetic product **C** *via* **TS1** with a barrier of $\Delta G_1^{\ddagger}$. The Curtin-Hammett principle states that the product ratio between **C** and **D** is proportional to the difference in the transition state energies $\Delta\Delta G^{\ddagger}$.

Utilizing $\Delta\Delta G^{\ddagger}$ values rather than ee values in molecular modeling can be rationalized based on the Linear Free Energy Relationships (LFERs) and the Bell−Evans−Polanyi Principle, as these concepts provide a more comprehensive understanding of the thermodynamic and kinetic factors influencing enantioselective reactions. The Bell−Evans−Polanyi Principle asserts that the difference in activation energies between two reactions is proportional to the difference in their reaction energies. In the context of enantioselective processes, this principle implies that the difference in activation energies between the formation of two enantiomers (represented by $\Delta\Delta G^{\ddagger}$) is related to the difference in their reaction energies. By incorporating $\Delta\Delta G^{\ddagger}$ values into molecular modeling, a more holistic representation of enantioselective reactions' thermodynamic and kinetic aspects can be achieved. Alongside this development, the principle of LFERs comes into play. LFERs posit that the free energy associated with a reaction correlates linearly with specific attributes of the reactants, whether steric or electronic. Historically, these steric and electronic factors were treated as separate entities in chemical interactions. However, contemporary understanding has advanced to integrate these two factors, leading to the emergence of the stereo-electronic effects concept. This convergence represents a significant stride in our current understanding and analysis of chemical reactions.

This concludes our outline between ee and $\Delta\Delta G^{\ddagger}$ with a clear suggestion to use $\Delta\Delta G^{\ddagger}$ as the target variable for modeling, as it is advantageous for several reasons. Firstly, $\Delta\Delta G^{\ddagger}$ possesses direct physical significance as it is the resulting factor behind the selectivities, making it a more pertinent choice for understanding the underlying processes and mechanisms, while using ee has only historical reasons. Secondly, $\Delta\Delta G^{\ddagger}$ incorporates temperature as a variable, which has proven to be an essential factor affecting selectivity in most catalyzed enantioselective reactions.[8-9] By using $\Delta\Delta G^{\ddagger}$, the model could capture the effect of temperature on the selectivity better, presumably resulting in more accurate predictions. Additionally, the target transformations can have an important effect on model performance and different choices may be right depending on the context.[10-13] Conversely, ee does not include the inherent temperature dependence, which could lead to a less accurate model, because the temperature has to be provided as an extra input feature and an additional context has to be learned by the model. Thirdly, a key consideration when selecting between ee and $\Delta\Delta G^{\ddagger}$ as target variables is their respective ranges of applicability. While ee is constrained to a limited range of –100 to +100, this restriction does not apply to $\Delta\Delta G^{\ddagger}$ as shown in Figure 2.

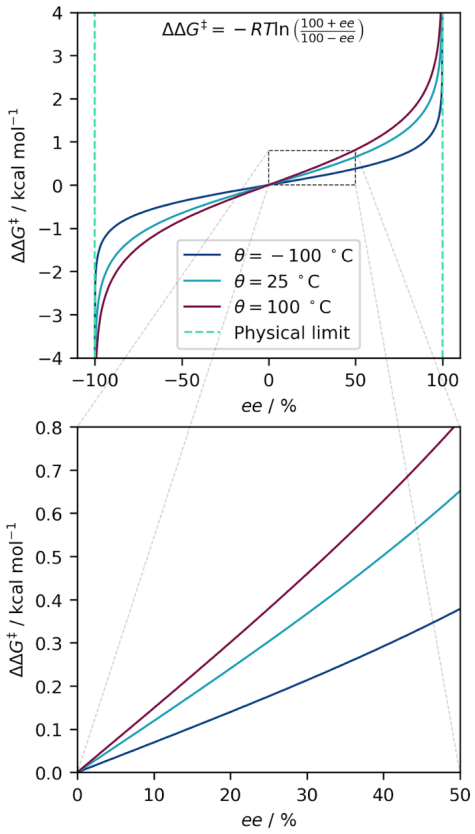


Figure 2. The plot on the left illustrates the non-linear association between ee and the change in $\Delta\Delta G^{\ddagger}$ across three distinct temperature conditions. The yellow demarcations represent the physical constraints for ee, indicating that employing a model based on ee may yield implausible predictions. In contrast, such restrictions are absent when utilizing $\Delta\Delta G^{\ddagger}$ as the modeling target. The right portion of the figure provides an amplified view of the linear domain at low ee values (<50%), demonstrating that the logarithmic correlation between ee and $\Delta\Delta G^{\ddagger}$ becomes particularly prominent at elevated ee levels.

Although the model could cover the entire range for both variables, predictions for ee might extend beyond its meaningful range (if not explicitly constrained). Consequently, these predictions would have no real-world significance, as values below –100 or above +100 are not physically possible. In contrast, $\Delta\Delta G^{\ddagger}$'s absence of such limitations allows the model to make predictions across a wider range of values, enhancing its generalizability and ensuring that its predictions maintain relevance in real-world scenarios. Furthermore, it's important to acknowledge that $\Delta\Delta G^{\ddagger}$ can be derived from ee, and vice versa, at a particular temperature. However, the conversion error is not constant and can fluctuate depending on the temperature. Nevertheless, using $\Delta\Delta G^{\ddagger}$ for data representation offers more comprehensive insights, factoring in temperature variations and lacking the range constraints that come with ee. This could contribute to improved model performance and predictive accuracy. One upside of using ee is that it is intuitive to experimentalists, which can also be reached by modeling in $\Delta\Delta G^{\ddagger}$ and then transforming to the more intuitive ee domain for visual representation.[14] As a consequence of the non-linear relationship between ee and $\Delta\Delta G^{\ddagger}$, the distribution of a data set is changed when converting between these quantities (see Figure 3), which can influence the training-testing split unfavorably. Unbalanced datasets can lead to clustering effects, which can artificially inflate the R^2-score, see Anscombe's quartet.[15] Furthermore, common metrics such as R^2 and RMSE are used to evaluate the quality of a model fit and the ability of a model to make predictions on unseen data. Both scores are affected by the transformation between ee and $\Delta\Delta G^{\ddagger}$, shown in Figure 4. In practice, minor errors in the 95–99 %ee range are critical, while even large errors in the 1–50 %ee range are negligible.

2. Benchmarking Approach

We collected data from the literature to investigate various aspects of the target transformation on real-world results in enantioselective catalysis that have been used for data-driven modeling. A focus was on publications that utilized molecular descriptor-based models to compare the influence of featurization. Furthermore, we aimed for a representative selection of data with different structures (*i.e.*, combinatorial screening or traditional linear optimization) across a range of data set sizes commonly encountered in modeling enantioselectivity (about 20–1000 data points) and featuring examples from various types of catalysis.

For descriptor-based models, molecular features were utilized unchanged where available in the original publications. In data sets **DS1**–**DS5** and **DS7**–**DS9**, various steric and electronic descriptors were available from DFT calculations. In **DS9** and **DS4**-THF, 2D topological descriptors were used. Morgan fingerprints (radius 2) were generated using the RDKit from the SMILES representations and folded to 1024 dimensions for fingerprint-based models. In data sets with several variable molecules, a concatenated SMILES containing each variation in one string (i.e., catalyst and substrate) was used. Models were trained with various standard machine learning regressors as implemented in Scikit-learn[16] on 50 different random 80:20 train:test splits of the data to obtain consistent model scores. Hyperparameter optimization was performed once for each model-feature-set combination using repeated *k*-fold (4 folds, 5 repetitions) cross-validation in the training set on the first 80:20 random split. The following regressors were used: Linear regression, ridge regression, Lasso, LassoLars, elastic net, random forest, gradient boosting regression, extra trees, kernel ridge regression, Gaussian processes, and k-nearest neighbors.[17-26]

Table 1. Overview of data sets utilized in this work. CPA = chiral phosphoric acid.

Dataset	Groups	Reaction	Samples	Available features
DS1[27]	Sigman, Biscoe	Pd-phosphine cat. Alkyl-Suzuki coupling	24	Descriptors
DS2[28]	Doyle	Ni-BiOx/BiIm, photo-cat. Cross-electrophile coupling	29	Descriptors
DS3[29]	Sigman, Toste	DAP-cat. allenoate-Claisen rearrangement	37	Descriptors
DS4[30]	Tsuji, Sidorov, Varnek, List	IDPi-cat. Hydroalkoxylation	80	SMILES, Descriptors
DS5[31]	Ackermann, Hong	Pd+TDG, electrocat. oxidative Heck reaction	127	SMILES, Descriptors
DS6[32]	Toste, Sigman	Triazole-PA-cat. cross-dehydrogenative coupling	159	SMILES, Descriptors
DS7[33]	Sunoj	Asymmetric hydrogenation	371	SMILES, Descriptors
DS8[34]	Sunoj	Pd-PyrOx-cat. relay-Heck reaction	398	Descriptors
DS9[35]	Belyk, Sherer	Phase transfer-cat. aza-Michael addition	471	SMILES, Descriptors
DS10[36]	Denmark	CPA-cat. nucleophilic thiol addition	1075	SMILES

Graph-based models were trained with molecular graphs that we generated from the provided SMILES representations, which were converted into molecular objects using RDKit and subsequently constructed into molecular graphs with standard node and edge features (refer to Supplementary Information for a comprehensive list). In instances with *n* SMILES, we generated a graph comprising *n* distinct, non-interconnected subgraphs embedded within the overall molecular graph.

These graphs were then processed using a Graph Neural Network (GNN) to produce a molecular embedding after global pooling operation, which was fed into a feed-forward neural network (FFNN) (refer to Supplementary Information for architectural details). For the GNN and FFNN, we used Pytorch Geometric and Pytorch.[37-38] We conducted training for the model both with and without incorporating temperature as an extra input feature in the FFNN. This was done to account for the potential effects of temperature dependency, which we were also interested in. We employed a Bayesian optimization approach via Optuna to optimize the hyperparameters of the GNN model.[39] As an optimizer we used Adam with the Mean Squared Error (MSE) as a training metric. To avoid overfitting, we employed the early stopping technique. While cross-validation is a widely used method for model evaluation, it may not be the most suitable choice for small datasets and GNNs due to the constraints posed by limited data, high computational cost, and potential structural bias. Instead, assessing the performance using multiple random states can provide a more reliable estimate of the model's generalization ability while overcoming these limitations. This approach involves randomly splitting the data into training and test sets multiple times (*i.e.*, 500 times) and calculating the performance metrics for each split. The average performance across all random states provides a more robust estimate of the model's generalization capability while mitigating the challenges associated with cross-validation.

We performed our models' performance evaluation, divided into four categories: fingerprint-based, descriptor-based (linear and non-linear), and GNN. We computed the MAE on each dataset for each model and variation (including temperature included or not). To scrutinize the disparity in performance between $\Delta\Delta G^{\ddagger}$ and ee modeling, we transformed both metrics in the er domain, which is mathematically possible without applying a capping threshold that could have led to potential beneficial transformations during the computation into a different domain. To ensure values between 0 and 1, we normalized the MAE by the MAE of an all-mean prediction.

3. Results

The transformation between ee and $\Delta\Delta G^{\ddagger}$ is non-linear and thus affects the distribution of data sets. We investigated the effect of this transformation on the data structure using the skew score and the Kolmogorov–Smirnov (KS) test. The skew score quantifies the skewness, that is, the degree of asymmetry in a distribution with a positive value indicating a right-skewed distribution, *i.e.*, leaning towards lower values. The Kolmogorov–Smirnov test tests the probability that samples were drawn from a specific distribution. Here, we test the experimental selectivities against a normal distribution. It should be noted that data from traditional “linear” reaction optimization and reaction scope tables tends to be biased towards higher selectivities because experimenters focus on more selective reactions when pursuing a new method. Screening methods such as combinatorial evaluation of all substrate/catalyst pairs tend to generate data with a relatively higher portion of less selective or lower-yielding results, which is favorable for ML

modeling.[40] Accordingly, nearly all data sets display a negative skew (towards higher values) when measured as ee. Averaged over all 10 sources, the absolute skew score decreases from 1.2 to 0.35 by transforming from ee to $\Delta\Delta G^{‡}$, indicating a clear overall shift towards more symmetric data distributions in the free energy domain. Likewise, the KS indicates a decrease from 17% to 11% probability that the data was not drawn from a normal distribution upon transformation from ee to $\Delta\Delta G^{‡}$. The differences are most pronounced in more extensive data sets (>100 samples) of traditional "linear" optimization and scope results. In smaller data sets (<100 samples), the differences are less pronounced but also less significant due to the low sample size. The Denmark data set (**DS10**) is the biggest screening-type data set, containing the complete combinatorial evaluation of substrate/catalyst pairs. Interestingly, while the absolute skewness increases and switches from negative to positive by transformation to $\Delta\Delta G^{‡}$, the similarity to a normal distribution is still higher in the free energy domain (17% vs. 8%). A summary of all computed skews and tests is shown in Table 2

Table 2. Summary of skewing and KS-test scores in the ee and the $\Delta\Delta G^{‡}$ domain. *N* = Number of samples.

Dataset	ee		$\Delta\Delta G^{‡}$		Difference		*N*
	skew	KS	skew	KS	skew	KS	
DS1	–0.81	0.22	0.29	0.16	0.52	0.06	24
DS2	0.03	0.19	0.31	0.22	–0.27	–0.02	29
DS3	–0.15	0.11	0.40	0.11	–0.26	0.00	37
DS4	0.18	0.12	0.78	0.12	–0.60	0.00	80
DS5	–1.49	0.29	–0.54	0.17	0.95	0.12	127
DS6	–0.46	0.08	0.39	0.07	0.07	0.02	159
DS7	–3.24	0.23	–0.04	0.05	3.20	0.18	371
DS8	–3.54	0.22	–0.39	0.09	3.15	0.14	398
DS9	–1.05	0.14	–0.60	0.09	0.44	0.05	471
DS10	–0.54	0.17	0.86	0.08	–0.33	0.09	1075

A visual representation of the skew can be seen in Figure 3, in which a small skew (top, **DS3**) and a large skew (bottom, **DS7**) are shown between the ee and $\Delta\Delta G^{‡}$ domain. For ease of comparison, the axes are normalized to each domain's minimum (MIN) and maximum (MAX). In the case of a small skew, the distributions look similar and well-distributed, while the large skew shows a well-distributed dataset in $\Delta\Delta G^{‡}$ while being completely unbalanced in ee, as apparent from the large distribution density in the high ee range close to the MAX.

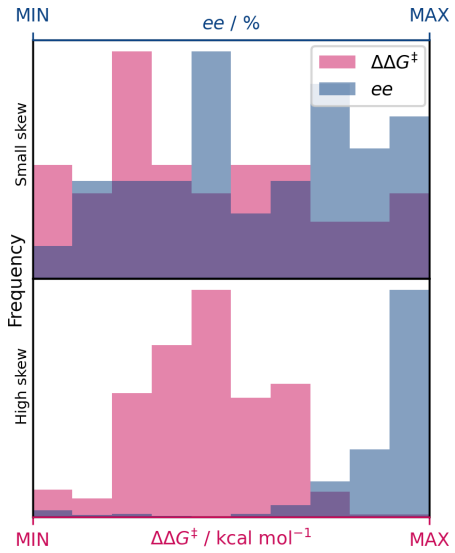


Figure 3. Histograms showing **DS3** (small skew, top) and **DS7** (high skew, bottom) in the ee (blue) and $\Delta\Delta G^{‡}$ (purple) domain. Top: The dataset is well distributed in both domains. Bottom: Only in the $\Delta\Delta G^{‡}$ domain does the dataset render balanced.

The influence and downsides of an unbalanced dataset, such as **DS7**, are apparent in Figure 4, in which artificial predictions with an MAE of 5% (based on the measured values) in ee are plotted. Such a potential model would have an excellent fit (R^2-score of 0.92, blue) in the ee domain, and the end-user might find it a 'useful' model. But the truth is that when converting to the physically meaningful $\Delta\Delta G^{‡}$ domain, the model appears useless, as indicated by the scatter plot with the bad fit (R^2-score of 0.49, purple).

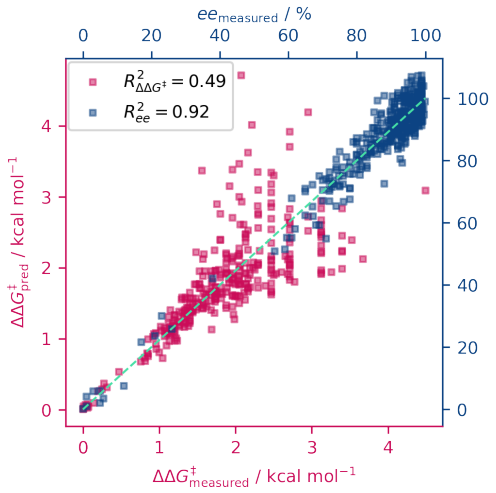


Figure 4. Scatter plot showing the data from **DS7** as measured and predicted by adding random noise of ±5% in ee (purple) and then calculating to the ΔΔ*G*‡ domain (dark blue). The light-green diagonal indicates optimal prediction. Even though an imaginary model has an excellent fit, with an artificial MAE of 5% in the ee domain, it would render useless, as in the physically relevant ΔΔ*G*‡ domain, the fit is bad and would not indicate a productive model, which is illustrated by the indicative difference in the R^2 score and visually by the spread of the ΔΔ*G*‡ points.

Figure 5 underscores the significance of maintaining predictions within the physically meaningful domain of ΔΔ*G*‡. The magnitude of the error in ee largely hinges on the specific ΔΔ*G*‡ range where the prediction error transpires. Due to the non-linear relationships, it's challenging to directly compare the MAEs of a model trained on ee and another on ΔΔ*G*‡. Consider a scenario with two models trained on ee each having an MAE of 10%. One model could be effective if the primary errors occurred in the low ΔΔ*G*‡ range, leading to an overall highly predictive model. Conversely, the other model, which happened to make incorrect predictions in the higher ΔΔ*G*‡ range, would result in a random dispersion of data points in the high ΔΔ*G*‡ domain and thus be rendered useless. This is also indicated by the dark-red error bars in Figure 5.

While predictions exceeding 100% may seem infrequent, occurring, *e.g.*, as rarely as 3.9% of all predictions, selecting an appropriate capping threshold remains a crucial factor. This scenario can significantly impact the model's metrics despite its seeming rarity. The frequency of predictions exceeding 100% is largely influenced by the number of data points with extremely high ee values. To ensure the validity of predictions, these are often capped at an arbitrary threshold, such as 99%. However, the chosen capping threshold can significantly influence the model's metrics, potentially leading to misconceptions about its true performance. To investigate this effect, we varied the capping thresholds while calculating ΔΔ*G*‡ derived from the predicted ee, and subsequently computed MAE in kcal mol^{-1}. The results on the **DS4**—depicted in Figure 6—indicate an exponential growth in error with respect to the chosen threshold, illustrating the considerable influence of this seemingly innocuous parameter on the overall model evaluation. We suggest that when calculating ee back to ΔΔ*G*‡, a cut-off threshold of 99.9% would be appropriate based on the elbow method.

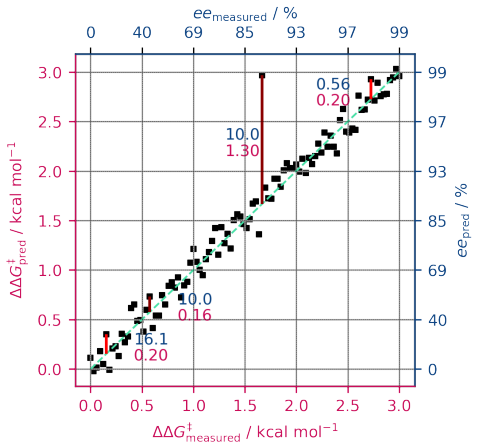


Figure 5. Scatter plot showing an artificial dataset (measured and predicted) in ΔΔ*G*‡ along with its conversion to the ee domain (dark blue axis labels). The dashed light-green line indicates optimal prediction. Four errors are showcased, two with constant ΔΔ*G*‡ - and two with constant ee-value: on the left side in a small ΔΔ*G*‡ regime, an error (marked in red) of 0.2 kcal mol^{-1} would resemble an error of 16.1% ee. The same error magnitude in a high ΔΔ*G*‡ regime would result in a much smaller error of 0.56% ee. Additionally, a constant ee error of, *e.g.*, 10.0% is shown, and the resulting difference is immediately apparent by the vastly different error bars in dark red.

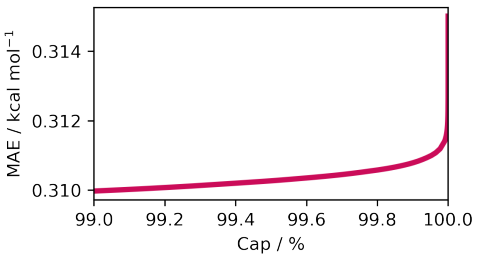


Figure 6. The error was measured as MAE in kcal mol^{-1} with respect to the capping threshold (Cap) in % during the ΔΔ*G*‡ calculation based on predicted ee. The exponential growth in MAE is evident when getting closer to the physical limit of 100%. The calculations were performed on **DS4** with temperature included as a modeling feature in conjunction with our GNN model. Of all 175 k predictions, only 62 (0.35%) were predicted over 100% or below –100% ee and subsequently capped.

The performance evaluations in the er domain shown in Figure 7 suggest that the discrepancy between modeling in the ee and ΔΔ*G*‡ domain is not negligible. For nearly all sections (model type and dataset), modeling in ΔΔ*G*‡ appears to be superior to modeling in ee; in the few cases where ee was superior over

$\Delta\Delta G^{\ddagger}$, the differences were only marginal. Generally, the differences between both domains were very prominent for some datasets, *e.g.*, **DS1**, **DS3**, and **DS6**. This renders $\Delta\Delta G^{\ddagger}$ to be the better choice of modeling domain compared to ee. Besides the effect of ee vs. $\Delta\Delta G^{\ddagger}$, we were also interested in the aspect of incorporation of temperature as an extra input feature.

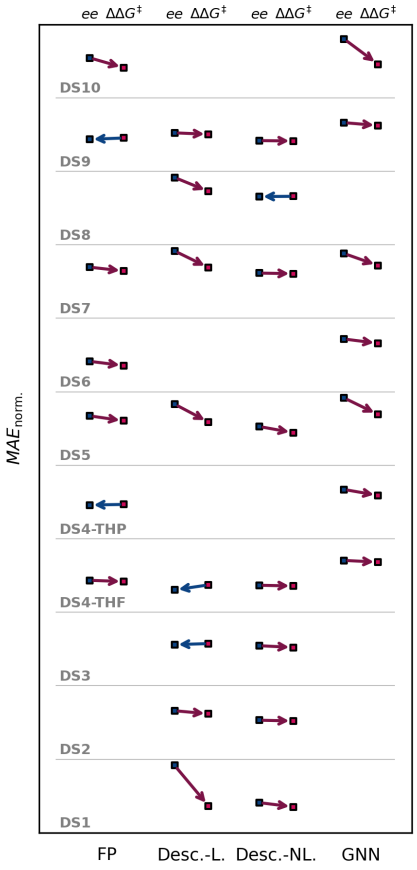


Figure 7. Shown are the performances for the best models for each dataset (stacked along the Ordinate) in each method class (appended along the Abscissa) with the respective modeling domain ee or $\Delta\Delta G^{\ddagger}$. As a common performance metric between ee and $\Delta\Delta G^{\ddagger}$, we chose the MAE in the er domain, normalized on the MAE of an all-mean-prediction to ensure values between 0 and 1. Performance points for each dataset and method are connected by an arrow, pointing to the lower-lying error, hence the better performing model; an arrow pointing towards the ee modeling (left) is colored in blue, while arrows pointing to modeling $\Delta\Delta G^{\ddagger}$ (right) are dark-purple colorized.

As $\Delta\Delta G^{\ddagger}$ has only a small linear temperature dependent via $\Delta\Delta S^{\ddagger}$, signifying that once $\Delta\Delta G^{\ddagger}$ is known, the resulting ee can be calculated at any temperature. However, when only ee is known, it is possible to determine the necessary $\Delta\Delta G^{\ddagger}$ for such an ee at a specific temperature. Still, no information about the difference in TS energies (actual $\Delta\Delta G^{\ddagger}$) can be gleaned. This necessity is illustrated in the plot shown in Figure 8, which is similarly constructed as Figure 7, meaning that the best model was picked in each respective category (cf. SI for bar plots on the temperature dependence of each dataset).

As Figure 8 indicates, incorporating temperature as an additional modeling feature is often not of major importance, but the differences are often subtle when excluding temperature, which is not statistically relevant. However, for GNNs using temperature as an extra input feature seems superior in every case, and for **DS5** a clear preference to include the temperature is observed.

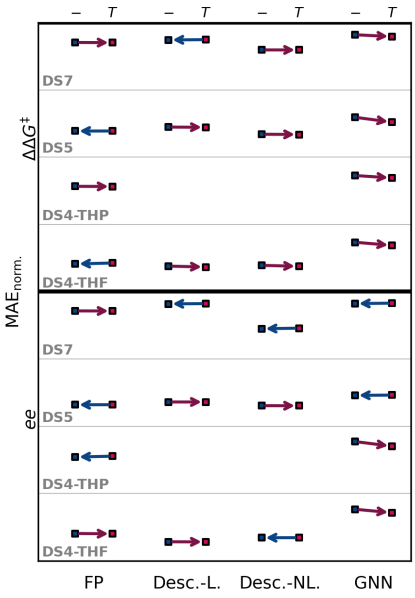


Figure 8. The performances for the best models for each dataset (for modeling in ee and $\Delta\Delta G^{\ddagger}$ stacked along the Ordinate) in each method class (appended along the Abscissa) with differentiation between inclusion or exclusion of temperature as an additional modeling feature. The MAE used for evaluation is normalized on the MAE of an all-mean prediction to ensure values between 0 and 1. Performance points for each dataset and method are connected by an arrow, pointing to the lower-lying error. Hence the better performing model; an arrow pointing towards the without temperature modeling (left) is colored in blue, while arrows pointing to with temperature (right) are dark-purple colorized.

4. Conclusion

In organic chemistry, particularly in asymmetric organocatalysis, we face a reporting bias (publication bias), which refers to the selective publication and reporting of results that exhibit high yields, enantioselectivities, and reaction rates. This bias may arise from various factors, such as the desire to present new, impactful findings or the pressure to publish positive results to secure funding and advance careers. Consequently, less

favorable or less exciting results may be underreported or not reported, leading to a skewed understanding of catalysts' true scope and limitations in asymmetric organocatalysis. This leads to an incomplete understanding: A biased representation of results in the literature can hinder the development of a comprehensive understanding of catalysts, their mechanisms, and their limitations, especially when using ML. This can lead to a distorted view and biased models. The lack of transparency and reporting negative or less favorable results can slow scientific progress, as such data might be invaluable to model training. In data modeling, this reporting bias leads to data markedly skewed to higher selectivities (or yields). This is unfortunate because an objective distribution would even be expected to lead to lower selectivities, as experience shows that achieving high selectivity is difficult and most "random" combinations of catalyst and substrate will result in no/low selectivity. However, for a model to truly learn the structure-selectivity relationships of a reaction, the reasons why specific catalysts result in low selectivity are equally as important as those leading to high selectivity. However, as shown in Figure 3, the resulting unbalanced datasets can be compensated when staying in the $\Delta\Delta G^{\ddagger}$ domain.

To truly identify a model of potential use, the domain has to be considered, as models resulting in a good fit in the ee domain do not necessarily render useful in the $\Delta\Delta G^{\ddagger}$ domain. As the error transformation between both discussed domains is non-linear, difficulties like comparison would arise between models of the ee and the $\Delta\Delta G^{\ddagger}$ domain. It is thus advisable to directly model in the physically sound $\Delta\Delta G^{\ddagger}$ domain, which aligns with the superiority in modeling performance of $\Delta\Delta G^{\ddagger}$ domain to the more intuitive and experimentalist-friendly ee domain. Using one domain and not relying on transformations between domains also dodges the need for a cutoff threshold, which we discussed in Figure 6. The effect of including temperature as an additional input feature is minor, but we still recommend incorporating temperature in the modeling process.

The process of target transformations, as discussed in this work, is crucial not only in asymmetric catalysis but also in kinetics, specifically in reaction rates. Recently, Votsmeier *et al.* used the hyperbolic sine function to model chemical kinetics through transformation.[41] A compelling extension of our research could involve examining the variations in model explanations such as attention maps, saliency maps, integrated gradients, and layer-wise relevance propagation methods, all contingent on the specific modeling domain. We hypothesize that a model trained with physically sound principles would yield more reliable and practically beneficial explanations, which could, for example, be leveraged for catalyst optimization.

Supporting Information

The authors have cited additional references within the Supporting Information.[30, 31] A more detailed description of the datasets, modeling details, and selected plots for specific datasets are included.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft within the priority program "Utilization and Development of Machine Learning for Molecular Applications – Molecular Machine Learning" (SPP 2363, Schr 597/41-1). M.R. thanks the Fonds der Chemischen Industrie for doctoral scholarship.

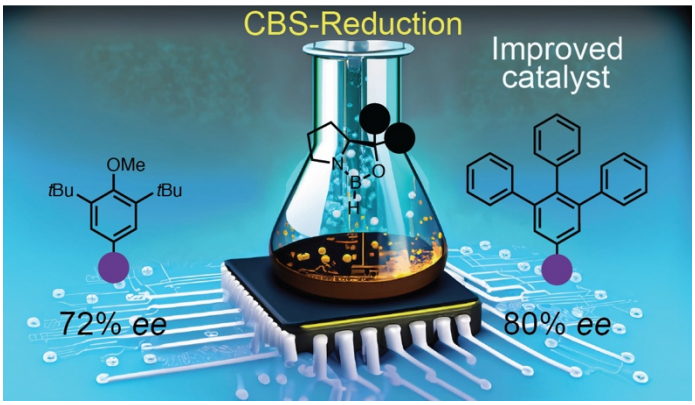
Keywords: Machine Learning • Catalysis • Modeling • Target Transformations

[1] W. L. Williams, L. Zeng, T. Gensch, M. S. Sigman, A. G. Doyle, E. V. Anslyn, *ACS Cent. Sci.* **2021**, *7*, 1622–1637.
[2] E. O. Pyzer-Knapp, T. Laino, in *Machine Learning in Chemistry: Data-Driven Algorithms, Learning Systems, and Predictions, Vol. 1326*, American Chemical Society, Washington, DC, **2019**, pp. ix–x.
[3] N. Artrith, K. T. Butler, F.-X. Coudert, S. Han, O. Isayev, A. Jain, A. Walsh, *Nat. Chem.* **2021**, *13*, 505–508.
[4] A. Bender, N. Schneider, M. Segler, W. Patrick Walters, O. Engkvist, T. Rodrigues, *Nat. Rev. Chem.* **2022**, *6*, 428–442.
[5] M. P. Maloney, C. W. Coley, S. Genheden, N. Carson, P. Helquist, P.-O. Norrby, O. Wiest, *Org. Lett.* **2023**, *25*, 2945–2947.
[6] R. E. Gawley, *J. Org. Chem.* **2006**, *71*, 2411–2416.
[7] M. Wernerova, T. Hudlicky, *Synlett* **2010**, *2010*, 2701–2707.
[8] H. Zhang, K. Shing Chan, *J. Chem. Soc., Perkin Trans. 1* **1999**, 381–382.
[9] A. Matusmoto, S. Fujiwara, Y. Hiyoshi, K. Zawatzky, A. A. Makarov, C. J. Welch, K. Soai, *Org. Biomol. Chem.* **2017**, *15*, 555–558.
[10] B. Li, J. Tang, Q. Yang, X. Cui, S. Li, S. Chen, Q. Cao, W. Xue, N. Chen, F. Zhu, *Scientific Reports* **2016**, *6*, 38881.
[11] H. S. Obaid, S. A. Dheyab, S. S. Sabry, in *2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON)*, **2019**, pp. 279–283.
[12] D. Singh, B. Singh, *Appl. Soft Comput.* **2020**, *97*, 105524.
[13] L. Huang, J. Qin, Y. Zhou, F. Zhu, L. Liu, L. Shao, *ITPAM* **2023**, 1–20.
[14] B. T. Rose, J. C. Timmerman, S. A. Bawel, S. Chin, H. Zhang, S. E. Denmark, *J. Am. Chem. Soc.* **2022**, *144*, 22950–22964.
[15] F. J. Anscombe, *Am. Stat.* **1973**, *27*, 17–21.
[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, *J. Mach. Learn.* **2011**, *12*, 2825–2830.
[17] A. E. Hoerl, R. W. Kennard, *Technometrics* **1970**, *12*, 55–67.
[18] R. Tibshirani, *Journal of the Royal Statistical Society. Series B (Methodological)* **1996**, *58*, 267–288.
[19] E. Bradley, H. Trevor, J. Iain, T. Robert, *Ann. Stat.* **2004**, *32*, 407–499.
[20] H. Zou, T. Hastie, *Journal of the Royal Statistical Society Series B: Statistical Methodology* **2005**, *67*, 301–320.
[21] L. Breiman, *Mach. Learn.* **2001**, *45*, 5–32.
[22] J. H. Friedman, *Ann. Stat.* **2001**, *29*, 1189–1232.
[23] P. Geurts, D. Ernst, L. Wehenkel, *Mach. Learn.* **2006**, *63*, 3–42.
[24] V. Vovk, in *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik* (Eds.: B. Schölkopf, Z. Luo, V. Vovk), Springer Berlin Heidelberg, Berlin, Heidelberg, **2013**, pp. 105–116.
[25] C. Williams, C. Rasmussen, in *NeurIPS*, **1995**.
[26] A. Mucherino, P. J. Papajorgji, P. M. Pardalos, in *Data Mining in Agriculture* (Eds.: A. Mucherino, P. J. Papajorgji, P. M. Pardalos), Springer New York, New York, NY, **2009**, pp. 83–106.

[27] S. Zhao, T. Gensch, B. Murray, Z. L. Niemeyer, M. S. Sigman, M. R. Biscoe, *Science* **2018**, *362*, 670–674.
[28] S. H. Lau, M. A. Borden, T. J. Steiman, L. S. Wang, M. Parasram, A. G. Doyle, *J. Am. Chem. Soc.* **2021**, *143*, 15873–15881.
[29] N. Tsuji, P. Sidorov, C. Zhu, Y. Nagata, T. Gimadiev, A. Varnek, B. List, *Angew. Chem. Int. Ed.* **2023**, *62*, e202218659.
[30] J. Miró, T. Gensch, M. Ellwart, S.-J. Han, H.-H. Lin, M. S. Sigman, F. D. Toste, *J. Am. Chem. Soc.* **2020**, *142*, 6390–6399.
[31] L.-C. Xu, J. Frey, X. Hou, S.-Q. Zhang, Y.-Y. Li, J. C. A. Oliveira, S.-W. Li, L. Ackermann, X. Hong, *Nat. Synth.* **2023**, *2*, 321–330.
[32] A. Milo, A. J. Neel, F. D. Toste, M. S. Sigman, *Science* **2015**, *347*, 737–743.
[33] S. Singh, M. Pareek, A. Changotra, S. Banerjee, B. Bhaskararao, P. Balamurugan, R. B. Sunoj, *PNAS* **2020**, *117*, 1339–1345.
[34] M. Das, P. Sharma, R. B. Sunoj, *J. Chem. Phys.* **2022**, *156*.
[35] K. W. Lexa, K. M. Belyk, J. Henle, B. Xiang, R. P. Sheridan, S. E. Denmark, R. T. Ruck, E. C. Sherer, *Org. Process Res. Dev.* **2022**, *26*, 670–682.
[36] A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow, S. E. Denmark, *Science* **2019**, *363*, eaau5631.
[37] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steinerand, L. Fang, J. Bai, S. Chintala, *Adv. Neural Inf. Process. Syst.* **2019**, *32*.
[38] M. Fey, J. E. Lenssen, in *International Conference on Learning Representations*, New Orleans, USA, **2019**.
[39] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Anchorage, AK, USA, **2019**, pp. 2623–2631.
[40] T. Gensch, S. R. Smith, T. J. Colacot, Y. N. Timsina, G. Xu, B. W. Glasspoole, M. S. Sigman, *ACS Catal.* **2022**, *12*, 7773–7780.
[41] F. A. Döppel, M. Votsmeier, *React. Chem. Eng.* **2023**.

2.2.2 Designing the Next Better Catalyst Utilizing Machine Learning with a *Key-Intermediate* Graph: Differentiating a Methyl from an Ethyl Group

Abstract:



The acceleration of catalyst design is paramount for attaining higher selectivities, which consequently reduces waste and by-products, thereby promoting more sustainable chemical processes. Established approaches, such as design of experiments or computational studies, face difficulties with the “methyl/ethyl problem” in the Corey-Bakshi-Shibata reduction. We leveraged the power of deep learning to surmount this challenge by constructing a small, albeit high-quality dataset that we used to train a model in a supervised fashion to predict the difference in Gibbs activation energies $\Delta\Delta G^{\ddagger}$ of reaction paths leading to either enantiomer. With the help of this model, we were able to select and subsequently screen multiple possible catalysts and consequently were able to increase the selectivity for the Corey-Bakshi-Shibata reduction of butanone to 80% enantiomeric excess. We underscore the transformative potential of deep learning in accelerating catalyst design for sustainable chemical processes. Our results not only champion the synergy of synthetic chemistry and computational methods but also provide a robust blueprint for future endeavors in catalysis optimization.

Version: September 18, 2023

Designing the Next Better Catalyst Utilizing Machine Learning with a *Key-Intermediate* Graph: Differentiating a Methyl from an Ethyl Group

Oliver Pereira,[#] Marcel Ruth,[#] Dennis Gerbig, Raffael C. Wende, and Peter R. Schreiner*

Institute of Organic Chemistry, Justus Liebig University, Heinrich-Buff-Ring 17, 35392 Giessen, Germany

* prs@uni-giessen.de

# These authors contributed equally

Abstract

The acceleration of catalyst design is paramount for attaining higher selectivities, which consequently reduces waste and by-products, thereby promoting more sustainable chemical processes. Established approaches, such as design of experiments or computational studies, face difficulties with the "methyl/ethyl problem" in the Corey-Bakshi-Shibata reduction. We leveraged the power of deep learning to surmount this challenge by constructing a small, albeit high-quality dataset that we used to train a model in a supervised fashion to predict the difference in Gibbs activation energies $\Delta\Delta G^{\ddagger}$ of reaction paths leading to either enantiomer. With the help of this model, we were able to select and subsequently screen multiple possible catalysts and consequently were able to increase the selectivity for the Corey-Bakshi-Shibata reduction of butanone to 80% enantiomeric excess. We underscore the transformative potential of deep learning in accelerating catalyst design for sustainable chemical processes. Our results not only champion the synergy of synthetic chemistry and computational methods but also provide a robust blueprint for future endeavors in catalysis optimization.

Introduction

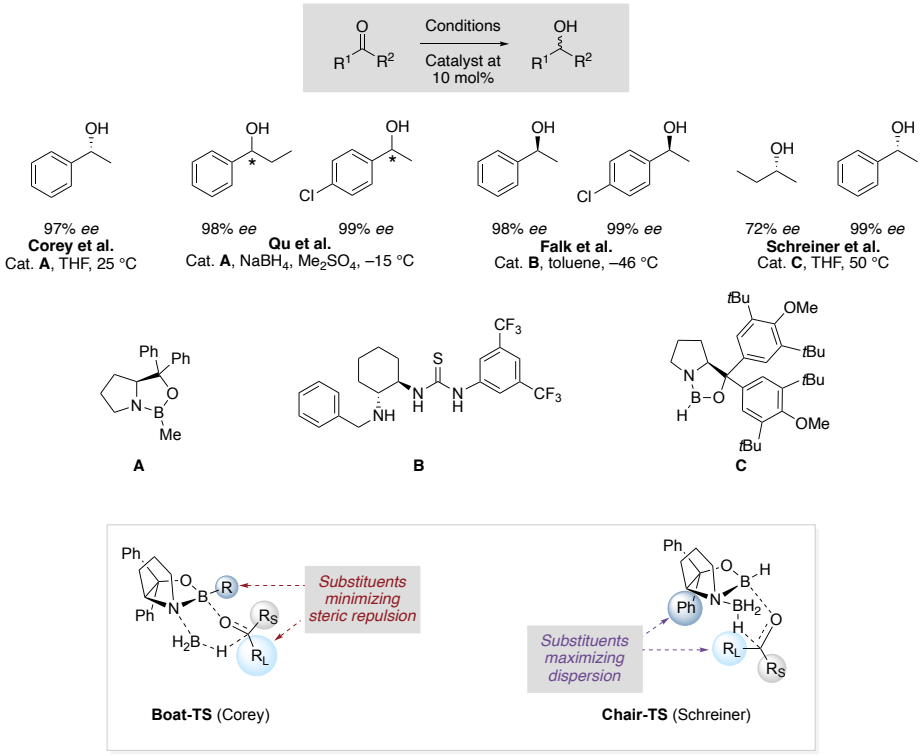
The differentiation of a methyl versus an ethyl group, for instance, in the catalytic reduction of butanone with a hydride source, is a longstanding and formidable challenge because the *re* and *si* faces of the substrate just differ by one methylene unit. One common reaction for this transformation is the Corey-Bakshi-Shibata (CBS)[1] reduction, which employs a chiral boron-based catalyst, specifically an oxazaborolidine derived from L-proline or another chiral amine. This

catalyst coordinates to borane (BH_3) a the reducing agent, and together they facilitate the reduction of the ketone in an enantioselective manner. The lessons learned from such a challenging transformation could potentially be applied to other, chemically diverse instances. The root cause of the methyl/ethyl problem –or generally two alkyl moieties– lies in the very subtle differences in stereoelectronic properties between different alkyl groups. In terms of catalyst design, this leaves very few options other than steric size differences (as judged by van-der-Waals surfaces) utilized, e.g., in "confined" catalysts[2] or London dispersion (LD) interactions.[3-5] It has been demonstrated in the last few years that LD interactions indeed can be a decisive element in catalyst design[6-10] by employing dispersion energy donors (DEDs)[5, 11] but would these be sufficient to meet the methyl/ethyl challenge? If so, which DEDs are most suitable, and how can they be identified without having to generate a large catalyst library? Here we address this issue using a small catalyst library that forms the basis for a machine learning (ML) approach that suggests potentially better-performing catalysts. We demonstrate the efficacy of our data-centric approach by the CBS reduction of several prochiral ketones, and in particular, for butanone with 80% *ee*, which is by far the highest selectivity for this substrate reported to date. Using actively fermenting yeast selectivities of 64% *ee* could be achieved.[12] The best results for the CBS reduction of butanone as the substrate were limited to 60% *ee* with the original Corey catalyst.[10, 13] The use of DED-equipped catalysts improved the *ee* to 72%, after employing a laborious design of experiment (DoE) approach and using detailed descriptors from employing computational chemistry methods.[10]

The CBS reduction of prochiral ketones has been widely employed in organic syntheses due to its generally high enantioselectivity and broad substrate scope using, *e.g.*, lactones, alkaloids, and steroids.[14-16] For instance, it has been employed in the preparation of chiral isoxazole carbinols,[17] the synthesis of (*R*)-phenylephrine,[18] as well as the enantioselective syntheses of massadine,[19] FR901464, and spliceostatin A.[20]

The CBS reduction (and many other additions to carbonyls) works well for prochiral ketones that bear stereoelectronically very different groups (Scheme 1), but it remains challenging to achieve high selectivity when the substituents are structurally very similar. As a consequence, the best results are obtained with aryl-alkyl groups, with acetophenone being the prime example that gives 97% *ee* in 99% yield.[14] This has led to a classic selectivity model that builds on the concept of "steric hindrance" alone, placing the large substituents on the substrate and the boron atom on

opposing sides of a boat-like transition state (Scheme 1).[13, 14, 21] Some of us demonstrated recently that such a model is highly insufficient by showing that the substituent at boron is largely inactive in the catalytic process and that an optimal balance between steric hindrance and LD attraction must be met for catalyst optimization in a more favorable chair-like transition structure (Scheme 1). How to strike this balance is, however, anything but trivial. Hence, we set out to use ML approaches, even though they are agnostic to molecular interactions, which can be utilized to move forward with catalyst design where our human ingenuity does not offer obvious choices. The CBS reduction was chosen as a model reaction here since it does react quite sensitively to DEDs in both catalyst and substrate, and because it is a widely used reaction of high utility.



Scheme 1. Selection of CBS reductions for the best (aryl/ alkyl) and most challenging (butanone) substrates.

While DoE and other types of optimizations can indeed help determine the optimal experimental conditions, these techniques need to address the inherent complexity of finding the optimal catalyst structure, which remains a distinct and significant challenge.

Several ML approaches have emerged to tackle catalysis design.[22-42] Drawing from an array of ML approaches to innovate and optimize catalysis, we use ML to address the methyl/ethyl challenge and thus offer a complementary alternative to DoE and dedicated computational approaches. As a result, the efficiency and selectivity of the CBS reduction of the challenging butane substrate have been improved by the selective preparation of ML-prioritized catalysts.

Results and Discussion

To ensure the reliability of ML approaches, high-quality datasets are a precondition. Since data extracted from the literature proved to vary in quality and experimental conditions, we conducted 90 reactions and averaged the results of multiple measurements. This process enables us to construct a small but robust database, providing a foundation for implementing our ML strategy. We employed our domain-specific knowledge, which is based on the structure of the key intermediate (before the rate-determining step) of the CBS reduction to develop a "key-intermediate" graph,[10] which integrates both substrate and catalyst into a single unified graph representation (Figure 1). To the best of our knowledge, this approach is new but easy to understand for the practicing chemist because we can readily envision what intermediates or transition structures may be particularly relevant for a particular reaction. This informed choice of a graph is based on nodes and edges, similar to a molecule, where atoms are represented by nodes and bonds by edges. The process of enhancing the mechanistically agnostic ML model with chemical knowledge is illustrated in Figure 1. The combined graph is connected at the boron atom of the catalyst and the carbonyl oxygen atom in the substrate, which undergoes reduction during the CBS process. Each node and edge in the graph is assigned certain molecular properties, which are listed in Figure 1. This graph thus offers a comprehensive representation of an essential “intermediate” in the CBS reduction, providing a more information-rich input for our graph neural network (GNN)-based model compared to two disjoint graphs, which we initially tested. This key-intermediate graph representation proved beneficial for our task compared to the disconnected

graph approaches, hence the catalyst and substrate each as an individual graph (cf. Supporting Information).

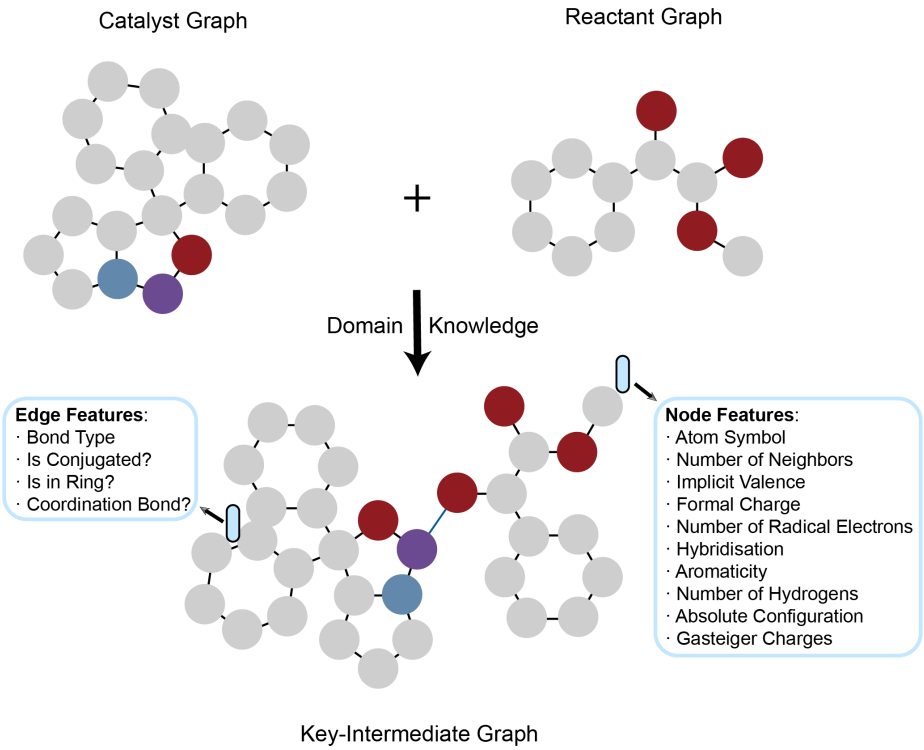


Figure 1. Combining the reactant and the catalyst graph to create a graph representing the CBS reduction's key-intermediate graph. The combination of both molecules occurs between the boron of the catalyst and the oxygen of the carbonyl. The nodes are colored based on the atom type, gray = carbon, blue = boron, purple = nitrogen, and red = oxygen.

Utilizing our custom-designed key-intermediate graph, we trained a GNN in a supervised manner to predict the Gibbs activation energy $\Delta\Delta G^{\ddagger}$ of the reaction paths leading to either enantiomer.[43] We randomly explored commercially available aryl bromo compounds, which could be attached to the L-proline scaffold via Grignard reaction. We used our trained model to determine the

optimal choice based on the predicted $\Delta\Delta G^{\ddagger}$ values. Remarkably, this approach immediately resulted in a catalyst that improved the *ee* from 72% to 80% for the CBS reduction of butanone. The time-saving benefits provided by this method are depicted schematically in Figure 2. Using ML to prioritize the next catalyst to test potentially saves months of experimental work. Once a well-distributed dataset is constructed, a better catalyst could probably be predicted immediately.

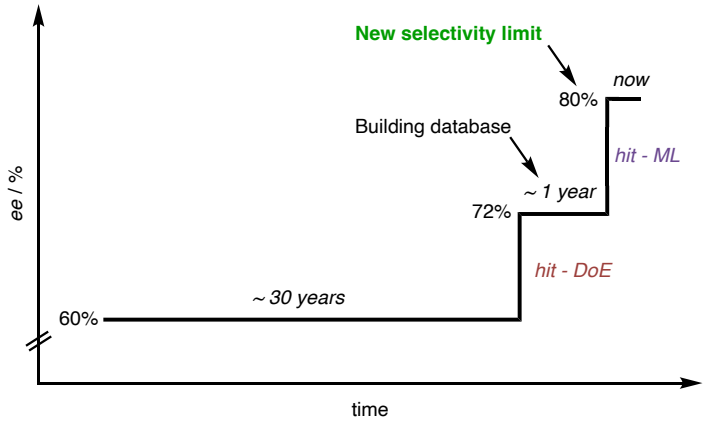


Figure 2. Schematic representation of the discovery timeline for new CBS catalysts using ML.

While time savings are possible with high-quality, widely distributed data, as shown in Figure 3, it is clear that this requires a combination of skilled synthetic chemists and data-centric approaches to generate robust data to feed and optimize the algorithm effectively. To overcome the long-standing issue of publication bias, which has been known since the 17th century and is still an ongoing issue,[44] it is paramount to validate each reaction yield or selectivity metric multiple times. The publication bias describes the statistical trend towards "positive" rather than "negative" results in the literature, hence resulting in an artificially shifted database towards "positive" results. Figure 3 highlights how a database suitable for ML looks like when including "negative results" that would be lost due to publication bias in the literature. We show that by using such high-quality data we can utilize the power of deep learning via GNNs, even with less than 100 data points.

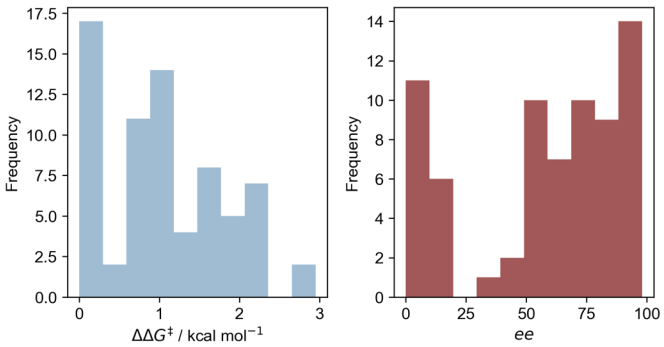


Figure 3. The histograms show the distribution of our database in $\Delta\Delta G^{\ddagger}$ (blue) and *ee* (red) in both domains; the distribution is near-uniform. Such a database is well suited for ML, in contrast to databases that could be constructed from biased literature data, which we have experienced ourselves in an initial approach using literature data.

Models and Methods. Ensuring that our model enables accurate predictions necessitated careful consideration of the purity of catalysts, reducing agents, and substrates, as well as conducting multiple runs for each reduction. We implemented this rigor to ensure the integrity of the experimental data and to boost their trustworthiness. To achieve a balanced stoichiometry, we used 0.1 equiv of the precursor and 1.1 equiv of the reductant, *i.e.*, borane dimethylsulfide (BMS), for catalyst formation. The rate of substrate addition proved to be important since fast addition of substrate could compromise enantioselectivity because of an uncatalyzed background reaction. To ascertain the optimal conditions (illustrated in Figure 5), we varied the addition rate and found that a minimal injection speed of 0.033 mL min^{-1} is needed for consistent results. Additionally, we conducted temperature trials to identify the optimal temperature for catalyst-substrate reactivity. As illustrated in Scheme 1, when performing reductions with the CBS catalyst on ketones, the ideal reaction temperature can vary based on the conditions and catalysts. However, it is known that lower than ambient temperatures tend to diminish selectivity.[45, 46] Therefore, we opted for a temperature of 50 °C for all subsequent reactions, which was shown to be optimal previously and was also confirmed in our condition screening;[46] for the rate of addition, we chose 0.017 mL min^{-1}. Determinations of enantioselectivities were conducted using chiral stationary phase (β-TBDAc

or β-6-TBDM column) gas chromatography with a flame ionization detector, *cf.* Supporting Information for details.

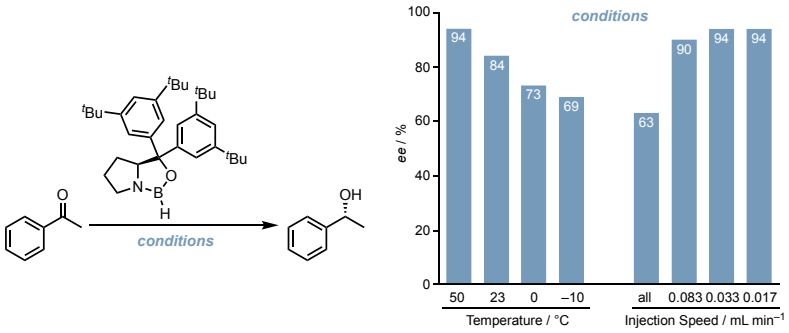


Figure 5. Tests were carried out using acetophenone to establish the reaction conditions at the start of the overall analysis with the shown catalyst. All injection speed tests were carried out at 50 °C.

For our ML modeling approach, we combined the reactant and catalyst graph to create a key-intermediate graph (*vide supra*, Figure 1). The reactant and catalyst graphs were initially created with the help of the RDKit framework[47] and a graph feature encoder, which was built upon one of DeepChem's[48] feature encoders. Both individual graphs were combined by locating the node index of boron in the catalyst graph, and the carbonyl oxygen node index of the substrate, and creating a virtual edge between both nodes.

We used Pytorch[49] and Pytorch Geometric[50] (PyG) for learning graph representations. Our model consists of a GNN part utilizing a graph attention layer[51] as implemented in PyG. The node embeddings are then used to generate the molecular embedding via global pooling operations. The molecular embedding is then fed into a feed-forward neural network (FFNN) to predict $\Delta\Delta G^{\ddagger}$. An overview of our workflow is depicted in Figure 6.

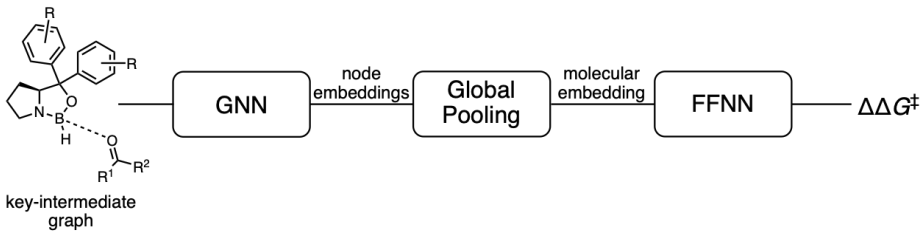


Figure 6. The key-intermediate graph is processed by a GNN that produces node embeddings. Using global pooling operations, molecular embedding is computed and used as input in a feed-forward neural network to predict $\Delta\Delta G^{\ddagger}$.

We optimized the model parameters to minimize the mean squared error (MSE) between predicted and experimental (calculated from gas chromatography measurements) $\Delta\Delta G^{\ddagger}$ values. Due to the limited data available, we performed leave-one-out cross-validation (LOOCV) with the scikit-learn[52] package to assess our model's performance reliably and to determine the uncertainty of our model for each data point. In addition to LOOCV, we checked whether the model "learned the chemistry" behind the CBS reduction by completely excluding a substrate or catalyst from training and then validating. With this approach, we can dissect for which substrates and catalysts the model can accurately predict $\Delta\Delta G^{\ddagger}$, hinting towards certain strengths and weaknesses of both our model and approach.

Using 90 data points for the LOOCV, we achieved an R^2-score, root MSE, and mean absolute error (MAE) of 0.95, 0.12, and 0.16 kcal mol^{-1}, respectively, for the combined validation set (Figure 7). The bottom part of the performance plot indicates that the uncertainty is uniformly distributed over the range of possible $\Delta\Delta G^{\ddagger}$ values, which arguably is due to optimally distributed training data.

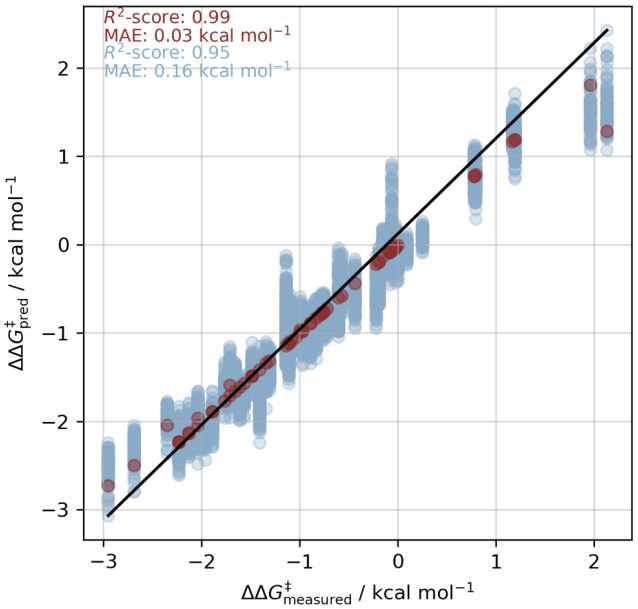


Figure 7. Scatter plot showing the predicted against the experimental $\Delta\Delta G^{\ddagger}$ values throughout the LOOCV. The black diagonal indicates optimal prediction. Validation points are colored in red, while training points are colored in blue.

The results for excluding substrates and catalysts are shown in the scatter plot depicted in Figure 8. It is apparent that the only catalyst with inverted stereochemistry is predicted incorrectly when excluded from the training data. In this case, the model predicts values close to the mean of the dataset. Indeed, the model is designed to associate stereochemistry with the sign of $\Delta\Delta G^{\ddagger}$, given appropriate training data. However, when training data only include examples with $\Delta\Delta G^{\ddagger} < 0$, it becomes challenging for the model to accurately predict $\Delta\Delta G^{\ddagger} > 0$ values for data points that were not included in the training set, and *vice versa*.

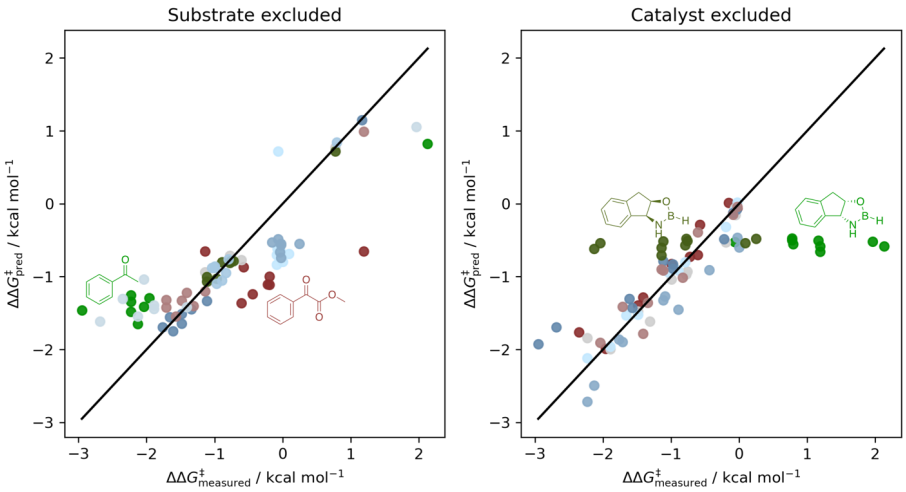


Figure 8. Scatter plots show the validation by excluding specific substrates (left) and catalysts (right). Molecules that led to a biased prediction are shown. Both substrates show a biased prediction, overprediction (green), and underprediction (red). Both depicted catalysts have nearly the same predicted *ee* for all substrates.

We initially let our model predict the selectivity of the catalysts depicted in Figure 9, for which only the catalyst highlighted by the green disk (top part, first iteration) was predicted to result in higher selectivity than the best literature-known catalyst for the CBS reduction of butanone.[10]

11

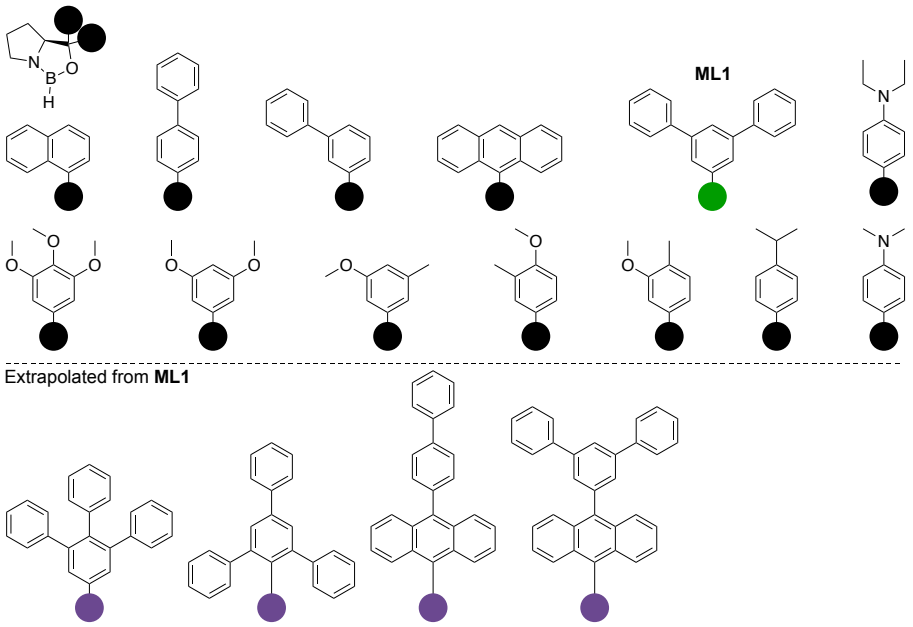


Figure 9. Depiction of our selection of catalyst candidates that we constructed and filtered based on commercial availability. The catalyst highlighted with the green circle was predicted to result in higher selectivity than the best literature-known catalyst.[10] Possible catalyst candidates extrapolated from **ML1** are highlighted with purple circles.

Experimental Realization. Using the hitherto unreported **ML1**, we consistently achieved 80% *ee* in the asymmetric reduction of butanone to (*S*)-butane-2-ol. We used **ML1** on all other substrates in our database and extended the database with the new results. One prominent feature of **ML1** is the addition of two phenyl groups to the aromatic ring, which is attached to the L-proline scaffold. To the best of our knowledge such an extension, *i.e.*, additional aromatic substituents on the phenyl rings that are attached to the L-proline scaffold, has not been performed before. The two phenyl rings in *meta*-position to the L-proline scaffold are excellent DEDs, and the resulting high selectivity of **ML1** in the CBS reduction is in accordance with our previous study.[10] Extrapolating from **ML1** even more DED rich catalysts can be envisioned (Figure 9). The downside of the even

higher performing catalysts is their challenging preparation, which thus far bears serious limitations in the Grignard step in our hands.

Conclusions

We demonstrate that ML can overcome the very challenging "methyl/ethyl-problem" in the CBS reduction with an enantioselectivity of 80%. Even a small database like the one used in this work can leverage the power of ML, provided the experimental data are tightly curated through careful optimization, monitoring, and repetition. As our catalyst was still hand-made, we suggest the next step to reach an even better catalyst would be to develop a generative modeling approach or use reinforcement learning to develop a better idea in the "selectivity game."

LOOCV-based analyses show an R^2 score of 0.95 for prediction accuracy. The achieved root MSE and MAE values of 0.12 and 0.17 kcal mol^{-1}, respectively, demonstrate our method's high precision and reliability. Moreover, by analyzing the scatter plots, we found that the uncertainty of our predictions is uniformly distributed over the entire range of possible $\Delta\Delta G^{\ddagger}$ values. This is primarily because our training data are well-distributed and representative. Our approach was validated by including a broad range of substrates and catalysts. In particular, we identified a promising catalyst (highlighted in green in Figure 9) that exhibits higher selectivity than the previously known best catalyst for the CBS reduction of butanone.

Our results highlight the effectiveness of ML algorithms in catalysis optimization, demonstrating the powerful synergy of synthetic chemistry and ML. Even with a small but high-quality dataset, the iterative cycle of synthesis, informed by domain knowledge and synthetic ingenuity, coupled with ML, can lead to substantial advancements. This integrated approach provides a promising foundation for future developments in the field and can complement or even replace the traditional design of experiment approaches when high-quality data is available. Future directions to expand on our fundamentally simple GNN-based approach would be the development of a framework that can be universally applied to any catalytic reaction. We envision a suitable (chemically meaningful) graph, which would be constructed automatically (in the optimal case) and then processed by a GNN to produce a molecular embedding. This embedding would then –combined with additional physical information (temperature, pressure, concentration, etc.)– be fed into an FFNN.

ASSOCIATED CONTENT

Supporting Information containing model architecture is available free of charge *via* the Internet at *tbd*.

Acknowledgment

This work was supported by the Deutsche Forschungsgemeinschaft within the priority program "Utilization and Development of Machine Learning for Molecular Applications – Molecular Machine Learning" (SPP 2363, Schr 597/41-1). M.R. thank the Fonds der Chemischen Industrie for a doctoral scholarship.

References

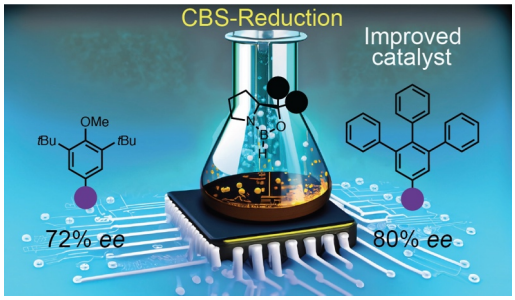
1. Corey, E. J.; Bakshi, R. K.; Shibata, S.; Chen, C. P.; Singh, V. K., A Stable and Easily Prepared Catalyst for the Enantioselective Reduction of Ketones. Applications to Multistep Syntheses. *J. Am. Chem. Soc.* **1987,** *109* (25), 7925–7926.
2. Mitschke, B.; Turberg, M.; List, B., Confinement as a Unifying Element in Selective Catalysis. *Chem* **2020,** *6* (10), 2515–2532.
3. Bursch, M.; Caldeweyher, E.; Hansen, A.; Neugebauer, H.; Ehlert, S.; Grimme, S., Understanding and Quantifying London Dispersion Effects in Organometallic Complexes. *Acc. Chem. Res.* **2019,** *52* (1), 258–266.
4. Liptrot, D. J.; Power, P. P., London Dispersion Forces in Sterically Crowded Inorganic and Organometallic Molecules. *Nat. Rev. Chem.* **2017,** *1* (1), 0004.
5. Wagner, J. P.; Schreiner, P. R., London Dispersion in Molecular Chemistry—Reconsidering Steric Effects. *Angew. Chem. Int. Ed.* **2015,** *54* (42), 12274–12296.
6. Li, B.; Xu, H.; Dang, Y.; Houk, K. N., Dispersion and Steric Effects on Enantio-/Diastereoselectivities in Synergistic Dual Transition-Metal Catalysis. *J. Am. Chem. Soc.* **2022,** *144* (4), 1971–1985.
7. Yang, L.; Li, B.; Houk, K. N., The Role of Attractive Dispersion Interaction in Promoting the Catalytic Activity of Asymmetric Hydrogenation. *Org. Chem. Front.* **2023,** *10* (14), 3485–3490.
8. Singha, S.; Buchsteiner, M.; Bistoni, G.; Goddard, R.; Fürstner, A., A New Ligand Design Based on London Dispersion Empowers Chiral Bismuth–Rhodium Paddlewheel Catalysts. *J. Am. Chem. Soc.* **2021,** *143* (15), 5666–5673.
9. Gramüller, J.; Franta, M.; Gschwind, R. M., Tilting the Balance: London Dispersion Systematically Enhances Enantioselectivities in Brønsted Acid Catalyzed Transfer Hydrogenation of Imines. *J. Am. Chem. Soc.* **2022,** *144* (43), 19861–19871.
10. Eschmann, C.; Song, L.; Schreiner, P. R., London Dispersion Interactions Rather than Steric Hindrance Determine the Enantioselectivity of the Corey–Bakshi–Shibata Reduction. *Angew. Chem. Int. Ed.* **2021,** *60* (9), 4823–4832.
11. Grimme, S.; Ehrlich, S.; Goerigk, L., Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011,** *32* (7), 1456–1465.

12. MacLeod, R.; Prosser, H.; Fikentscher, L.; Lanyi, J.; Mosher, H. S., Asymmetric Reductions. XII. Stereoselective Ketone Reductions by Fermenting Yeast*. *Biochemistry* **1964,** *3* (6), 838–846.
13. Corey, E. J.; Link, J. O.; Bakshi, R. K., A Mechanistic and Structural Analysis of the Basis for High Enantioselectivity in the Oxazaborolidine-Catalyzed Reduction of Trihalomethyl Ketones by Catecholborane. *Tetrahedron Lett.* **1992,** *33* (47), 7107–7110.
14. Corey, E. J.; Helal, C. J., Reduction of Carbonyl Compounds with Chiral Oxazaborolidine Catalysts: A New Paradigm for Enantioselective Catalysis and a Powerful New Synthetic Method. *Angew. Chem. Int. Ed.* **1998,** *37* (15), 1986–2012.
15. Cho, B. T., Recent Advances in the Synthetic Applications of the Oxazaborolidine-Mediated Asymmetric Reduction. *Tetrahedron* **2006,** *62* (33), 7621–7643.
16. Stemmler, R. T., CBS Oxazaborolidines - Versatile Catalysts for Asymmetric Synthesis. *Synlett* **2007,** *2007* (06), 0997–0998.
17. Natale, N. R.; Rider, K. C.; Burkhart, D. J.; Li, C.; McKenzie, A. R.; Nelson, J. K., Preparation of Chiral Isoxazole Carbinols *via* Catalytic Asymmetric Corey-Bakshi-Shibata Reduction. *ARKIVOC* **2010,** *2010* (8), 97–107.
18. Dai, S.; Li, G.; Zhang, W.; Zhang, C.; Song, X.; Huang, D., Efficient Synthesis of (R)-Phenylephrine Using a Polymer-supported Corey–Bakshi–Shibata Catalyst. *Chem. Lett.* **2017,** *46* (5), 740–743.
19. Ghosh, A. K.; Chen, Z.-H., Enantioselective Syntheses of FR901464 and Spliceostatin A: Potent Inhibitors of Spliceosome. *Org. Lett.* **2013,** *15* (19), 5088–5091.
20. Cannon, J. S., A Nitrone Dipolar Cycloaddition Strategy toward an Enantioselective Synthesis of Massadine. *Org. Lett.* **2018,** *20* (13), 3883–3887.
21. Corey, E. J.; Helal, C. J., Novel Electronic Effects of Remote Substituents on the Oxazaborolidine-Catalyzed Enantioselective Reduction of Ketones. *Tetrahedron Lett.* **1995,** *36* (50), 9153–9156.
22. Rinehart, N. I.; Zahrt, A. F.; Denmark, S., The Leveraging Machine Learning for Enantioselective Catalysis: From Dream to Reality. *CHIMIA* **2021,** *75* (7–8), 592.
23. Singh, S.; Pareek, M.; Changotra, A.; Banerjee, S.; Bhaskararao, B.; Balamurugan, P.; Sunoj, R. B., A Unified Machine-Learning Protocol for Asymmetric Catalysis as a Proof of Concept Demonstration Using Asymmetric Hydrogenation. *PNAS* **2020,** *117* (3), 1339–1345.
24. Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E., Prediction of Higher-Selectivity Catalysts by Computer-Driven Workflow and Machine Learning. *Science* **2019,** *363* (6424), eaau5631.
25. Rüscher, M.; Herzog, A.; Timoshenko, J.; Jeon, H. S.; Frandsen, W.; Kühl, S.; Roldan Cuenya, B., Tracking Heterogeneous Structural Motifs and the Redox Behaviour of Copper–Zinc Nanocatalysts for the Electrocatalytic CO_2 Reduction Using Operando Time Resolved Apectroscopy and Machine Learning. *Catal. Sci. Technol.* **2022,** *12* (9), 3028–3043.
26. Kitchin, J. R., Machine Learning in Catalysis. *Nat. Cat.* **2018,** *1* (4), 230–232.
27. O'Connor, N. J.; Jonayat, A. S. M.; Janik, M. J.; Senftle, T. P., Interaction Trends Between Single Metal Atoms and Oxide Supports Identified With Density Functional Theory and Statistical Learning. *Nat. Cat.* **2018,** *1* (7), 531–539.
28. Tran, K.; Ulissi, Z. W., Active Learning Across Intermetallics to Guide Discovery of Electrocatalysts for CO_2 Reduction and H_2 Evolution. *Nat. Cat.* **2018,** *1* (9), 696–703.
29. Ma, S.; Huang, S.-D.; Liu, Z.-P., Dynamic Coordination of Cations and Catalytic Selectivity on Zinc–Chromium Oxide Alloys During Syngas Conversion. *Nat. Cat.* **2019,** *2* (8), 671–677.
30. Lin, C.; Li, J.-L.; Li, X.; Yang, S.; Luo, W.; Zhang, Y.; Kim, S.-H.; Kim, D.-H.; Shinde, S. S.; Li, Y.-F.; Liu, Z.-P.; Jiang, Z.; Lee, J.-H., In-Situ Reconstructed Ru Atom Array on α-MnO_2 With Enhanced Performance for Acidic Water Oxidation. *Nat. Cat.* **2021,** *4* (12), 1012–1023.
31. Esterhuizen, J. A.; Goldsmith, B. R.; Linic, S., Interpretable Machine Learning for Knowledge Generation in Heterogeneous Catalysis. *Nat. Cat.* **2022,** *5* (3), 175–184.

32. Li, F.; Yuan, L.; Lu, H.; Li, G.; Chen, Y.; Engqvist, M. K. M.; Kerkhoven, E. J.; Nielsen, J., Deep Learning-Based k_{cat} Prediction Enables Improved Enzyme-Constrained Model Reconstruction. *Nat. Cat.* **2022,** *5* (8), 662–672.
33. Mou, T.; Pillai, H. S.; Wang, S.; Wan, M.; Han, X.; Schweitzer, N. M.; Che, F.; Xin, H., Bridging the Complexity Gap in Computational Heterogeneous Catalysis with Machine Learning. *Nat. Cat.* **2023,** *6* (2), 122–136.
34. Margraf, J. T.; Jung, H.; Scheurer, C.; Reuter, K., Exploring Catalytic Reaction Networks With Machine Learning. *Nat. Cat.* **2023,** *6* (2), 112–121.
35. Zhao, S.; Gensch, T.; Murray, B.; Niemeyer, Z. L.; Sigman, M. S.; Biscoe, M. R., Enantiodivergent Pd-catalyzed C–C Bond Formation Enabled Through Ligand Parameterization. *Science* **2018,** *362* (6415), 670–674.
36. Lau, S. H.; Borden, M. A.; Steiman, T. J.; Wang, L. S.; Parasram, M.; Doyle, A. G., Ni/Photoredox-Catalyzed Enantioselective Cross-Electrophile Coupling of Styrene Oxides with Aryl Iodides. *J. Am. Chem. Soc.* **2021,** *143* (38), 15873–15881.
37. Tsuji, N.; Sidorov, P.; Zhu, C.; Nagata, Y.; Gimadiev, T.; Varnek, A.; List, B., Predicting Highly Enantioselective Catalysts Using Tunable Fragment Descriptors. *Angew. Chem. Int. Ed.* **2023,** *62* (11), e202218659.
38. Miró, J.; Gensch, T.; Ellwart, M.; Han, S.-J.; Lin, H.-H.; Sigman, M. S.; Toste, F. D., Enantioselective Allenoate-Claisen Rearrangement Using Chiral Phosphate Catalysts. *J. Am. Chem. Soc.* **2020,** *142* (13), 6390–6399.
39. Xu, L.-C.; Frey, J.; Hou, X.; Zhang, S.-Q.; Li, Y.-Y.; Oliveira, J. C. A.; Li, S.-W.; Ackermann, L.; Hong, X., Enantioselectivity Prediction of Pallada-Electrocatalysed C–H Activation Using Transition State Knowledge in Machine Learning. *Nat. Synth.* **2023,** *2* (4), 321–330.
40. Milo, A.; Neel, A. J.; Toste, F. D.; Sigman, M. S., A Aata-Intensive Approach to Mechanistic Elucidation Applied to Chiral Anion Catalysis. *Science* **2015,** *347* (6223), 737–743.
41. Das, M.; Sharma, P.; Sunoj, R. B., Machine Learning Studies on Asymmetric Relay Heck Reaction—Potential Avenues for Reaction Development. *J. Chem. Phys.* **2022,** *156* (11).
42. Lexa, K. W.; Belyk, K. M.; Henle, J.; Xiang, B.; Sheridan, R. P.; Denmark, S. E.; Ruck, R. T.; Sherer, E. C., Application of Machine Learning and Reaction Optimization for the Iterative Improvement of Enantioselectivity of Cinchona-Derived Phase Transfer Catalysts. *Org. Process Res. Dev.* **2022,** *26* (3), 670–682.
43. LeCun, Y.; Boser, B.; Denker, J. S.; Henderson, D.; Howard, R. E.; Hubbard, W.; Jackel, L. D., Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput.* **1989,** *1* (4), 541–551.
44. Bausell, R. B.; Bausell, R. B., 15C1Publication Bias. In *The Problem with Science: The Reproducibility Crisis and What to do About It*, Oxford University Press: 2021; p 0.
45. Xu; Wei; Zhang, Effect of Temperature on the Enantioselectivity in the Oxazaborolidine-Catalyzed Asymmetric Reduction of Ketones. Noncatalytic Borane Reduction, a Nonneglectable Factor in the Reduction System. *J. Org. Chem.* **2003,** *68* (26), 10146–10151.
46. Stone, G. B., Oxazaborolidine Catalyzed Borane Reductions of Ketones: a Significant Effect of Temperature on Selectivity. *Tetrahedron: Asymmetry* **1994,** *5* (3), 465–472.
47. Landrum, G. RDKit: Open-Source Cheminformatics Software. https://www.rdkit.org. DOI: https://zenodo.org/record/6961488#.Y9znXezMKPS.
48. Ramsundar, B.; Eastman, P.; Walters, P.; Pande, V.; Leswing, K.; Wu, Z., *Deep Learning for the Life Sciences*. O'Reilly Media: 2019.
49. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steinerand, B.; Fang, L.; Bai, J.; Chintala, S., PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Adv. Neural Inf. Process. Syst.* **2019,** *32*.

50. Fey, M.; Lenssen, J. E., Fast Graph Representation Learning with PyTorch Geometric. In *International Conference on Learning Representations*, New Orleans, USA, 2019.
51. Brody, S.; Alon, U.; Yahav, E., How Attentive are Graph Attention Networks? *arXiv preprint arXiv* **2021**.
52. 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.* **2011,** *12* (85), 2825–2830.

TOC Figure



3. Acknowledgment – Danksagung

Die vorgestellten Arbeiten wären nicht ohne die Unterstützung (privat und wissenschaftlich) einiger wichtigen Schlüsselfiguren zustande gekommen. Mein Dank geht dabei besonders an:

Prof. Dr. Peter R. Schreiner, PhD, welcher mich in seine Arbeitsgruppe aufgenommen und gefördert hat. Von interessanten Anregungen und Hinweisen zu potenziell fachübergreifenden Anwendungen meiner Forschung, bis zu hin zur Beratung in persönlichen Belangen. Besonders dankbar bin ich für die Freiheiten in der Gestaltung meiner Arbeitszeit, meines Arbeitsorts, meines Arbeitsschwerpunkts und für das Lehren auf den Saiten der Forschung zu spielen.

Prof. Dr. Doreen Mollenhauer, für die Übernahem des Zweitgutachtens dieser Arbeit.

Dr. Dennis Gerbig, dafür, dass er immer ein Ohr für mich hatte und meine Motivation stets in einem angeregten Zustand hielt. Seine Hinweise bewahrten mich oftmals vor einem *unerlaubten Übergang* in ein niedrigeres „Motivationsniveau“. Ohne ihn wäre vieles nicht in der kurzen Zeit möglich gewesen. Durch ihn lernte ich es durch die Barriere zu tunneln. Er war die Aktivierungsfunktion meines neuralen Netzwerks.

Dr. Bastian Bernhardt, welcher mich schon früh während meines Masters an die Gepflogenheiten der Wissenschaft heranführte, für die zwei tollen Publikationen zusammen und den tollen „Bouldersessions“.

Dr. Elisa Franzmann-Don, ohne welche ich **niemals** mit dem Studium angefangen hätte.

Ephrath Solel, PhD, für die Möglichkeit etwas ML im Rahmen meiner damaligen HiWi-Stelle auszuprobieren.

Oliver Pereira, für die tolle Zusammenarbeit im CBS-Projekt und den netten Gesprächen.

Michaela Richter, für die unkomplizierte und schnelle Beratung in administrativen Angelegenheiten.

Dem **Verband** der **Chemischen Industrie** für die finanzielle Unterstützung in Form eines Kekulé Stipendiums während meiner Promotion.

Meine ehemaligen Kommilitonen: **Alexander Granichny**, **Lara Gronych**, **Paul Debes** und **Lysander Wagner** für das gelegentliche Bouldern.

Janek Mann, für die gemeinsame Obsession fürs Bouldern und das Testen von Kletterschuhen.

Florian Schilder, für die vielen tollen Stunden in Rust, welche sich als gutes Training zum Steigern des Durchhaltevermögens erwiesen.

André, Justin und **Werner Lukas**, für die vielen entspannenden Skiurlaube.

Meine Eltern, für die Möglichkeit studieren zu können und der emotionalen Unterstützung.