DISSERTATION

DATA-DRIVEN STRATEGIES FOR ORGANIC STRUCTURE-PROPERTY AND STRUCTURE-REACTIVITY RELATIONSHIPS

Submitted by

Shree Sowndarya Santhanalakkshmi Vejaykummar

Department of Chemistry

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2024

Doctoral Committee:

Advisor: Robert Paton

Ashok Prasad
Seonah Kim
Aaron Nielsen

ABSTRACT

DATA-DRIVEN STRATEGIES FOR ORGANIC STRUCTURE-PROPERTY AND STRUCTURE-REACTIVITY RELATIONSHIPS

The prediction of molecular properties plays a pivotal role in various domains, from drug discovery to materials science. With the advent of machine learning (ML) techniques, particularly in the field of cheminformatics, the prediction of properties for small organic molecules has witnessed significant advancements. This document delves into the diverse machine-learning strategies employed for the accurate prediction of properties crucial for understanding molecular behavior. In Chapter 1, I offer insights into the evolution of data-driven modeling through Quantitative Structure-Property Relationships (QSPR), highlighting promising advancements in utilizing chemical features to construct predictive models for molecular properties. In Chapter 2, I delve into the primary stage of modeling, focusing on data collection for predictive tasks. I illustrate how the integration of automation and computational tools' advancement can construct modular workflows for FAIR (Findable, Accessible, Interoperable, and Reusable) chemistry. This approach aims to enhance the usability and reproducibility of scientific data. In Chapter 3, I emphasize leveraging computational tools to access high-level data for small organic molecules. I showcase the creation of a novel metric for assessing organic radical stability, utilizing a comprehensive chemical database of radicals. This involves employing straightforward physical organic descriptors, namely fractional spin, and buried volume, computed through systematic computational workflows. In Chapter 4, I explore the progression of graph-based models designed to forecast molecular properties, specifically Bond

Dissociation Energy. Additionally, I conduct a thorough examination of two particular applications pertinent to pharmaceutical and atmospheric chemistry. I demonstrate that utilizing a minimal number of molecules from the relevant chemical space can notably enhance large-scale machine-learning models. Finally, in Chapter 5, I combine the developed tools from Chapters 3 and 4, to perform goal-directed molecular optimization in identifying novel radicals for aqueous redox flow batteries using graph neural networks (radical stability, redox potentials, and bond dissociation energy) and reinforcement learning. This *de novo* molecular optimization strategy has successfully identified 32 new radical candidates. By amalgamating insights from diverse studies, this dissertation endeavors to offer a comprehensive grasp of how machine-learning strategies are transforming the terrain of molecular property prediction.

ACKNOWLEDGMENTS

As I fly to Los Angeles with immense joy knowing that I can finally defend, setting out to write my acknowledgments I realize that multiple people have contributed to helping me grow both professionally and personally in my journey so far. With a long list of people that I have to thank, the first and foremost goes to my advisor, mentor, and guide, Prof. Robert Paton. He has constantly stood by my decisions throughout graduate school, motivated me when needed, and has offered immense support over the last five years. Rob has significantly impacted my decisions since my first meeting with him. While looking into prospective doctoral degree programs, I was fortunate to meet him at the University of Oxford, where I completed a master's degree. I fondly remember my chat with him, where we talked about possible project directions and how Colorado weather can be cold and dry! Knowing that I would be about continents away from home, doing a long-demanding Ph.D., the anxiety about who I would work with lessened after the talk with him. Over the years his mentorship, guidance, emotional support, encouragement, and his confidence in me have time and again proved that my choice to come to Colorado was one of the best decisions of my life and professional career.

I have also had the privilege to work with other distinguished scientists and professors during my time at Colorado State namely Dr. Peter St. John and Prof. Seonah Kim. I would like to take this as an opportunity to thank Peter for taking the time to educate me about topics I had seen for the first time. His patience and guidance throughout my initial years were instrumental in shaping my research career. I would like to thank Seonah for everything she has done for me as a mentor and for being someone I can turn to for a second opinion. I can confidently say with Seonah around, the theory suite would never lack parties, group outings, and retreats – which

have all been significant in creating a sense of belonging for international students like me. As a graduate student, I worked in the pharmaceutical industry at AbbVie in Chicago. I would like to thank my mentor Pritha, for the opportunity to gain insights into day-to-day life in the industry that was deterministic in my decision to pursue a career in the industry. I also believe reaching this milestone would have been incomplete without my undergraduate and master's advisors – Prof. Raghavan B Sunoj and Prof. Fernanda Duarte. While I was a young researcher trying to find my interests, they have shown me the path by providing insights to reach where I currently am.

As I submit this dissertation, I would like to thank my committee members Prof. Ashok Prasad, Prof. Delphine Farmer, Prof. Alan Van Orden, and Prof. Aaron Nielsen for their time in reading my thesis and providing feedback and suggestions. Over the years thank you for taking the time to evaluate my candidacy and being a critical judge of my work.

I would like to thank the most important people of my journey who have made it an unforgettable experience – my present and former colleagues within the Paton and Kim Lab. I am always going to remember our time together both inside and outside the office. Be it our scientific discussions, or our lazy Friday afternoon conversations on how we want it to be 5 pm already, I think we have grown into amazing scientists! I would like to thank Guilian Luchini, Heidi Klem, Liliana Gallegos, Dr. Yingzi Li, and Dr. Juan Alegre Requena for the warm welcome into the group. They became my academic family who have tolerated me for who I am and have witnessed my growth as a graduate student. I am thankful to have another member, Louis de Lescure, in my cohort. Thank you for answering my organic chemistry questions and all the support over the job-hunting process as I must say it was a rollercoaster of emotions. Other current members of both Paton and Kim Lab – Dr. Mihai Popsecu, Dr. Graham Haug, Dr. Zhitao Feng, Dr. Raúl Pérez-Soto Alex Platt, Niket Manoj, Abhijeet Singh Bhadauria, Jake King, Sabari

Kumar, Chris Stubbs, Collin Hansen, Hojin Jung for carefully listening to all my practice talks and reluctantly agreeing to go to Bawarchi Biryani for lunch on Fridays! I would also like to thank other past members of the Paton and Kim lab – Dr. Yanfei Guan, Dr. Yeonjoon Kim, Dr. Santeri Aikonen, Dr. Sreenithya Avadakkam, Lukas Sigmund, Wojtek Treyde, Susana Portela, Turki Alturaifi, Ricky Pena, Brandon Portela, and Jaehwan Lim for all the scientific discussions.

Over the years in Colorado, living so far away from home, my roommates, and friends – Likitha, Pavithra, Pranjali, Siddhi, Anshika, Poonam, Sanjana, and Swetam, allowed me to have the feeling of home. All the cooking sessions, movie nights, and hunting down different trains in Colorado are memories I will cherish. Thank you for living with me and listening to my rants about day-to-day activities.

Next, I would like to thank my friends from undergraduate – Smruti, Shivani, Nikhita, Mitali, Himani, Gauri, Trishala, and Aditi for being the support system so far away from home. Overcoming the time zone differences to have 5 hour long calls over the weekends to talk about life was something I looked forward to over the past years. My friends from masters – Natasha, Shivohum & Aditya, I thank you all for the conversations and the games we played over my lunchtime during the era of COVID-19. I have been fortunate to meet like-minded people who have truly become family!

Shreyam, I am at a loss for words to write about what role you play in my life. Over the past nine years that I have known you and been with you, I realize I am the luckiest girl in the world. We have stuck through thick and thin, overcome problems together, taken daring steps to go abroad together for higher studies, and traveled together – all my confidence stems from knowing that no matter what you'll always be there for me. The last 5 years as graduate students have been the toughest given the distance between us. But irrespective of that you have been just

a call away to offer emotional support and celebrate my small achievements over the phone. I thank you with all my heart for coming into my life and sticking by me for this long.

Finally, I would like to circle back to my origins and thank my family, Mom, Dad, Ishu, and my grandparents for always believing in me for the last 27 years, supporting every decision I have taken, and letting me branch out to follow my passion to do science. Thank you for being there when I was physically ill and making all the efforts to cure me. You are irreplaceable; everything that I do is to make you proud. I hope to make progress in that direction and bring you immense joy!

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 Introduction to data-driven computational modeling

Chemistry relies on a fundamental principle: similar molecules tend to exhibit similar properties. This principle implies that by altering molecular structures in the laboratory, we can fine-tune or optimize specific chemical properties. While chemists often grasp these structure-property relationships intuitively, translating them into quantitative terms requires a meticulous and sometimes mathematical description of the variations in molecular structures. By way of illustration, consider the realm of organic chemistry. Even minor modifications to molecular structures, such as the introduction or substitution of functional groups, can exert profound effects on properties like solubility, reactivity, and biological activity. Through systematic adjustments, chemists can strategically craft molecules to meet particular requirements, whether it involves enhancing the potency of a pharmaceutical compound, refining the efficiency of a catalyst, or optimizing the performance of a material.[1] The ability to understand and manipulate structure-property relationships not only facilitates the creation and synthesis of new compounds but also deepens our comprehension of the underlying mechanisms governing chemical behavior.[2-4]

The pursuit of quantitative structure-property relationships (QSPRs) in chemistry hinges on the formulation of quantitative molecular descriptors.[5, 6] These descriptors encapsulate differences in electronics, connectivity, shape, and other structural attributes, serving as the cornerstone for analyzing and predicting the correlations between molecular structures and properties. Ideally, these descriptors can be easily generated computationally for an arbitrary molecular structure.[7, 8] QSPR models have traditionally relied upon the use of easily

interpretable descriptors and simple statistical techniques, such as linear regressions, although with the rise of ML methods, increasingly complex computational representations of molecules are more common along with the application of more elaborate regression models, such as neural networks. [9, 10]

"Traditional" computational molecular modeling methods have been crafted to anticipate molecular properties, especially in scenarios where a well-established mechanism governs the process, and the pertinent structures are conducive to computational analysis. These methods encompass a spectrum of techniques grounded in fundamental principles of quantum mechanics, molecular mechanics, and statistical mechanics. In the realm of quantum mechanics-based methods, approaches such as density functional theory (DFT)[11] and ab initio methods offer precise insights into the electronic structure of molecules. By solving the Schrödinger equation, these methods yield accurate descriptions of molecular orbitals, energies, and properties. On the other hand, molecular mechanics methods provide a computationally efficient means to explore the structural and energetic aspects of molecular systems at a larger scale.[12] Utilizing simplified force fields, molecular mechanics simulations can elucidate conformational changes, and intermolecular interactions in molecular ensembles. Furthermore, statistical mechanics-based approaches, including molecular dynamics simulations[13] and Monte Carlo methods,[14] enable the investigation of molecular behavior over time and under various environmental conditions. By integrating Newton's equations of motion or employing stochastic sampling techniques, these methods provide dynamic insights into thermodynamic properties in molecular systems.

While traditional computational molecular modeling methods excel in scenarios where well-established principles and structural information are available, they also face challenges in dealing with time scales required to perform computational simulations hence making these an

inherently “low throughput” approach. In contrast, data-driven property prediction methods have emerged that are orders of magnitude faster than traditional methods.[15-18] In cases where ample, high-quality data is available, data-driven approaches have shown promise in outperforming complex computations in terms of accuracy.[19] These approaches harness ML and statistical modeling techniques to extract patterns directly from data, without relying on explicit mathematical models or detailed simulations. Data-driven methods excel in uncovering complex relationships between input features and output variables, particularly in scenarios where underlying mechanisms are not fully understood or difficult to model explicitly.[20, 21] They offer scalability advantages, being able to process large-scale datasets efficiently, making them well-suited for fields such as drug discovery. Moreover, data-driven approaches facilitate the integration of heterogeneous data sources and domain knowledge, providing comprehensive insights and predictions. They represent a shift in scientific research, offering powerful tools for driving innovation and accelerating discoveries of new molecular systems.

1.2 Quantitative structure-property relationships in large data regimes.

In the systematic development of data-driven QSPR modeling, the process unfolds through several key steps.[18] To elaborate 1) Identification of molecules from existing open-source molecular libraries such as PubChem, ChemBL, and CAS. 2) Calculation of the property of interest either using experimental or computational techniques. 3) Determination of quantitative representations of molecular structures that capture important information about the atoms, bonds, functional groups, etc., present within the molecule. 4) Building predictive models using simple methods such as linear regression or more complex tools such as neural networks. 5) Finally, employing the developed model to estimate chemical properties that can be used in

downstream applications (Fig. 1.1). While these steps are done sequentially, multiple considerations exist within each step where different methodologies can be used.

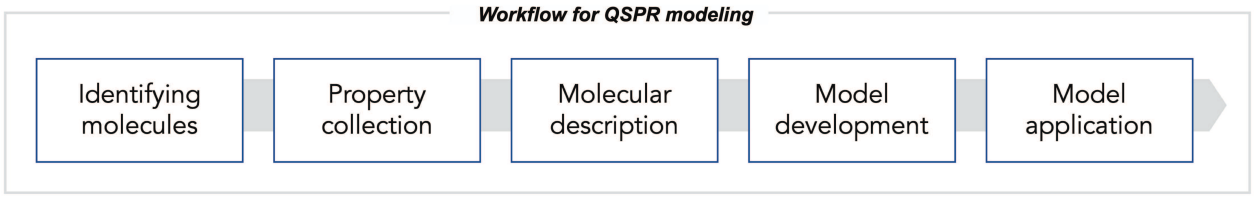


Figure 1.1. *Systematic workflow for modeling chemical properties*

1.2.1 Representations of molecules for computational modeling

In data-driven QSPR approaches to estimate molecular properties, we utilize various chemical aspects or 'features' of a molecule to determine the respective properties (Fig. 1.2A).[7, 22, 23] Featurization in the context of organic chemistry, involves capturing the various ways in which molecules differ. These features encompass a wide range of molecular characteristics that capture essential structural information. They may include descriptors such as molecular weight, size, shape, electronegativity, polarizability, bond lengths, bond angles, torsion angles, functional groups, aromaticity, and topological indices. Additionally, physicochemical properties such as solubility, partition coefficients, acidity, basicity, and surface area can also serve as features. More intensive features come from experimental or computational sources, such as atomic charges, NMR chemical shifts, or molecular dipole moment might be selected to quantify the electronic variation between a collection of compounds.[8] The suitability of a particular descriptor is generally context-dependent, and so a large variety of descriptors have been developed to date. There exist relatively simple representations, such as one-hot encoding or fingerprints, where 1's and 0's describe the presence or absence of predefined molecular substructures (Fig. 1.2A).[24] Extending beyond simple count-based descriptors, graph-based approaches can be used to encode every atom and bond present in a molecule. By incorporating a

diverse array of molecular features, data-driven approaches aim to capture the complex relationships between molecular structures and properties, enabling accurate predictions for a wide range of chemical compounds.

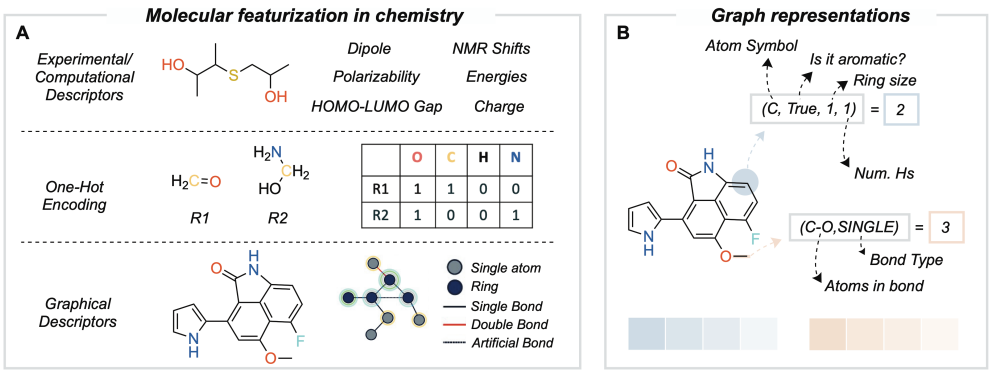


Figure 1.2. *(A) Highlighting different methods to featurized molecules. (B) Mathematical representation of molecules utilized in graphs for modeling*

Within this dissertation, graph-based approaches for chemical property prediction are primarily utilized.[25] This involves encoding every atom and bond in a molecule as nodes/vertices and edges in a graph, respectively. Mathematically, a graph is defined as G = (V, E) where V is a set of vertices, and E is a set of edges. Additional attributes such as atomic number, hybridization state, charge, and other relevant properties can be associated with each node to provide additional information about the atoms. Similarly, attributes such as bond type (single, double, triple), bond order, bond length, and stereochemistry can be associated with each edge to capture the characteristics of the bonds between atoms (Fig. 1.2B). In graph-based approaches, nodes representing atoms and edges representing bonds are tokenized based on feature vectors derived from their respective properties. For instance, a carbon atom (C) might be tokenized as "2" (blue), while a carbon-oxygen (C-O) bond could be tokenized as "3" (orange). This tokenization process is applied to all atoms and bonds in the molecule, generating a vector representation of

atoms and bonds. These encoded representations serve as numerical inputs for various modeling tasks, enabling the application of ML algorithms to predict molecular properties accurately.

1.2.2 Methodologies for modeling molecular properties

With a set of quantitative molecular features in hand, ML approaches to chemical property prediction often then rely on a given supervised or unsupervised learning approach.[26, 27] Clustering, dimensionality reduction, etc., relies on unsupervised learning that learns patterns from untagged data. Clustering algorithms group similar instances together based on their feature similarities, thereby revealing inherent structure within the dataset. Dimensionality reduction methods, such as principal component analysis (PCA)[28] or t-distributed stochastic neighbor embedding (t-SNE),[29] aim to reduce the dimensionality of the feature space while preserving relevant information, facilitating visualization and exploration of the data. On the other hand, regression and classification tasks achieved by mapping a specific input to a corresponding output fall under supervised learning (Fig. 1.3). Supervised learning algorithms, including linear regression,[30] decision trees,[31] support vector machines (SVM),[32] random forests,[33] and neural networks,[34] are commonly utilized to tackle such tasks by learning from labeled examples and making predictions on unseen data based on learned patterns.[35] Overall, both supervised and unsupervised learning approaches play crucial roles in ML-based chemical property prediction, each offering distinct capabilities for analyzing and extracting insights from molecular data.

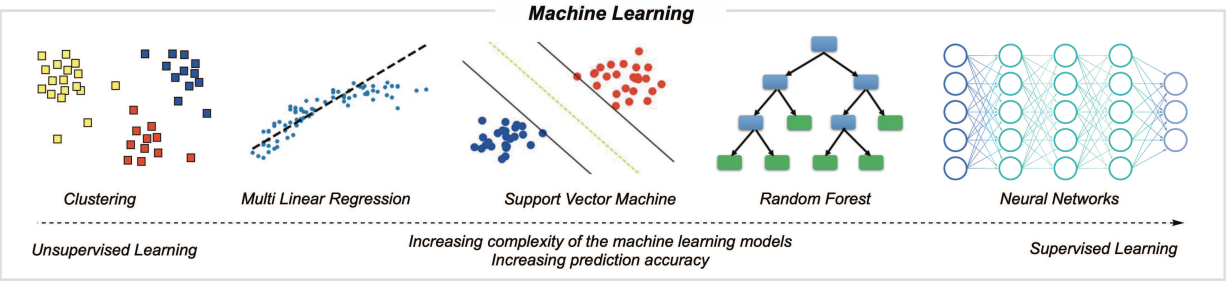


Figure 1.3. *Highlights different machine learning methods to perform unsupervised learning and supervised learning.*

Within is document, we deal with supervised learning methods such as neural networks.[19, 36-38] Neural networks, which mimic the human nervous system, are made of perceptons (circles on the neural network) that make decisions based on the input features. The working principle of a neural network lies in its ability to process and learn from data through interconnected layers of nodes, or neurons. Each neuron receives input signals, applies a transformation through an activation function, and passes the result to neurons in the next layer. This is called a Multi-Layer Perceptron (MLP) where information flows in one direction from the input layer (Fig. 1.3 dark blue circles) to hidden layers (Fig. 1.3 cyan circles) and finally to the output layer. There is no connection between neurons with a specific layer. Through a process called training, the network adjusts the weights (learnable parameters) associated with each connection between neurons to minimize the difference between its predictions and the actual outcomes. By iteratively fine-tuning these weights across multiple training examples, neural networks can effectively capture complex patterns and relationships within the data, enabling them to make accurate predictions and gain deeper insights into structure-property relationships in chemical compounds.

In recent years, neural network architectures have become the gold standard for machine/deep learning and have multiple uses in modeling molecular properties at atom, bond,

and molecule levels. Specifically, Graph Neural Networks (GNNs) that deal with graph structure as inputs for training neural networks have been shown to have high success.[39] These GNNs use the mechanism of message/information passing between atoms and bonds to iteratively learn about neighbors to a specific atom/ bond hence allowing the models to learn complex relationships present in molecules (Fig. 1.4 A).[40] In the message-passing mechanism, there are three important components: AGGREGATE, UPDATE, and READOUT. The AGGREGATE function performs the task of gaining and combining information from the neighboring atoms (e.g. connected by a bond) to compute new representations of the nodes. This can be operations such as sum, mean, or any other aggregation function (e.g. including edge information for aggregation). The UPDATE function takes the aggregated information along with the specific node to update a given node information. The UPDATE is typically a neural network layer such as with can linearly or non-linearly transform the node representations. Finally, the READOUT function collects all updated node representations of *n* message passing rounds to gain graph-level representations either *via* averaging or pooling operations. The above operation in message passing can be summarized using Equation (1) where for a node *u* in a graph for the *k+1*th round of message passing, to produce updated representations of h_u, information from the neighbors *N(u)* is used in the AGGEGATE function followed by utilization of UPDATE function to gain new representations for h_u. Various GNN architectures have been designed based on the types of AGGREGATE/UPDATE functions employed.[41]

$$h_u^{(k+1)} = UPDATE^{(k)}\left(h_u^{(k)}, AGGREGATE^{(k)}\left(\{h_v^{(k)}, \forall_v \in N(u)\}\right)\right) \quad (1)$$

Upon message passing to predict properties corresponding to atoms, updated/learned node embeddings are employed. For bond properties, edge information is utilized. At the molecular

level, aggregated node information after READOUT is used to predict molecular properties (Fig. 1.4 B).

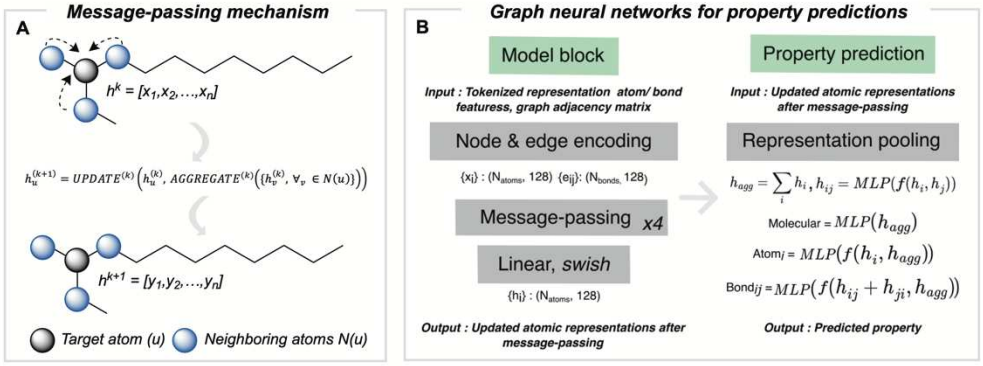


Figure 1.4. *(A) Methodology used in message passing in neural networks. Feature vectors of target atoms are updated based on the feature vectors of neighboring atoms. (B) An example model architecture of graph neural networks used in molecular property prediction.*

1.2.3 ML models for downstream goal-directed molecular optimization

Applying ML models for goal-directed optimization involves leveraging the predictive capabilities of ML algorithms to iteratively refine solutions toward achieving a specific objective.[42-45] For instance, concerning drug discovery, identifying a specific molecular candidate with the desired chemical properties.[46] However, knowing the vastness of chemical space, accounting for all molecular representations, and accurately predicting molecular properties is a challenging problem. 'Black-box' like prediction models developed using neural networks are trained to account for the vastness of chemical space. Strategies such as reinforcement learning are utilized for the exploration of chemical space.[47-49] This process typically entails formulating an optimization problem, where the model learns from historical data or interactions with the environment to identify optimal or near-optimal solutions. By training the model on relevant features and objective functions (e.g., predictions from QSPR models), and incorporating feedback from previous iterations, reinforcement learning-based optimization approaches can efficiently search through large solution spaces, adapt to changing

conditions, and converge towards desirable outcomes. Within this dissertation, the generation of new molecules with the help of reinforcement learning is highlighted.

1.3 Document overview

This work's primary focus is on the study of QSPRs in large data domains and their applications in goal-directed molecular optimization. In this dissertation, I will address challenges within QSPR modeling corresponding to the automation of large data collection, enumerating molecular structures as graph representations for utility in large neural networks, and the implementation of QSPR models for identifying new molecular candidates with desired chemical properties. This dissertation is broken down into three domains: 1) I will showcase how computational tools can be combined to make modular workflows for FAIR (Findable, Accessible, Interoperable, and Reusable) chemistry to enable the usability and reproducibility of scientific data, 2) I will delve into the diverse machine-learning strategies that I have employed for the accurate prediction of properties crucial for understanding molecular behavior. These properties are Radical Stability Score along with Reduction Oxidation potentials of organic radicals, and Bond Dissociation Energy for applications in the pharmaceutical/atmospheric relevant molecules, 3) Utilize the large-scale machine learning models in identifying novel radicals for aqueous redox flow batteries using graph neural networks and reinforcement learning. Overall, by synthesizing insights from various studies, this dissertation aims to provide a comprehensive understanding of how machine-learning strategies are reshaping the landscape of molecular property prediction and fostering innovation in molecular design and beyond.

1.3.1 Python toolkits for automation

In Chapter 2, I will highlight the developments of Automated Quantum Mechanical Environments (AQME) a free and open-source Python package for the rapid deployment of automated workflows using cheminformatics and quantum chemistry. AQME workflows integrate tasks performed across multiple computational chemistry packages and data formats, preserving all computational protocols, data, and metadata for machine and human users to access and reuse. AQME has a modular structure of independent modules that can be implemented in any sequence, allowing the users to use all or only the desired parts of the program. The code has been developed for researchers with basic familiarity with the Python programming language. The CSEARCH module interfaces to molecular mechanics and semi-empirical quantum mechanics (SQM) conformer generation tools (e.g., RDKit and Conformer–Rotamer Ensemble Sampling Tool, CREST) starting from various initial structure formats. The CMIN module enables geometry refinement with SQM and neural network potentials, such as ANI. The QPREP module interfaces with multiple quantum mechanics (QM) programs, such as Gaussian, ORCA, and PySCF. The QCORR module processes QM results, storing structural, energetic, and property data while also enabling automated error handling (i.e., convergence errors, wrong number of imaginary frequencies, isomerization, etc.) and job resubmission. The QDESCP module provides easy access to QM ensemble-averaged molecular descriptors and computed properties, such as NMR spectra. Overall, AQME provides automated, transparent, and reproducible workflows to produce, analyze, and archive computational chemistry results. SMILES inputs can be used, and many aspects of tedious human manipulation can be avoided. Installation and execution on Windows, macOS, and Linux platforms have been tested, and the code has been developed to support access through Jupyter Notebooks, the command line, and

job submission (e.g., Slurm) scripts. Examples of pre-configured workflows are available in various formats, and hands-on video tutorials illustrate their use.

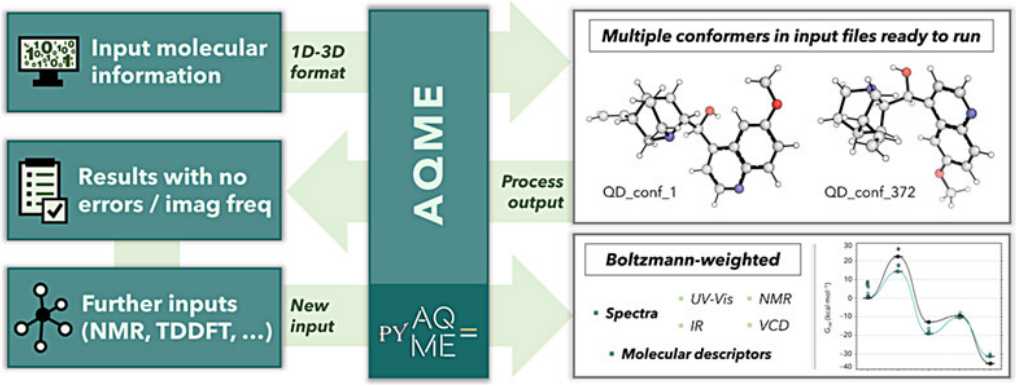


Figure 1.5. *Outline of work highlighted in Chapter 2 based on automation for computational tasks.*

1.3.2 Machine learning of molecular properties for small organic molecules

In Chapter 3, a specific class of molecules that are open-shell radicals is discussed. First I aim to develop a quantitative metric to understand radical stability. Next, I showcase the utility of graph neural networks to model large databases of radical stability and redox potentials. It is known that long-lived organic radicals are promising candidates for the development of high-performance energy solutions such as organic redox batteries, transistors, and light-emitting diodes. However, "stable" organic radicals that remain unreactive for an extended time and that can be stored and handled under ambient conditions are rare. A necessary but not sufficient condition for organic radical stability is the presence of thermodynamic stabilization, such as conjugation with an adjacent π-bond or lone-pair, or hyperconjugation with a σ-bond. However, thermodynamic factors alone do not result in radicals with extended lifetimes: many resonance-stabilized radicals are transient species that exist for less than a millisecond. Kinetic stabilization is also necessary for persistence, such as steric effects that inhibit radical dimerization or reaction with solvent molecules. I describe a quantitative approach to map organic radical stability, using

molecular descriptors intended to capture thermodynamic and kinetic considerations. The comparison of an extensive dataset of quantum chemical calculations of organic radicals with experimentally known stable radical species reveals a region of this feature space where long-lived radicals are located. These descriptors, based upon maximum spin density and buried volume, are combined into a single metric, the radical stability score, that outperforms thermodynamic scales based on bond dissociation enthalpies in identifying remarkably long-lived radicals. This provides an objective and accessible metric for use in future molecular design and optimization campaigns. Having defined the radical stability metric, in Chapter 5, I showcase the development of building graph neural network models for radical properties – radical stability and redox potentials.

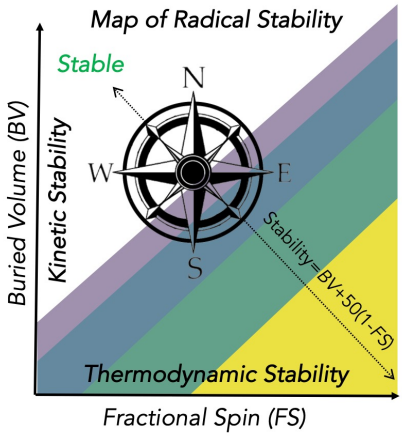


Figure 1.6. *Outline of work highlighted in Chapter 3 based on developing a metric for radical stability using physical organic descriptors.*

In Chapter 4, I aim to address another important bond property, Bond Dissociation Energy (BDE) to identify strategies to make radical species. Bond dissociation energetics underpin the thermodynamics of chemical transformations where bonds are broken or formed and can also be used to predict reaction rates and selectivities. Current machine learning (ML) models to predict bond dissociation energy (BDE) are largely limited in their elemental coverage to hydrogen and

the second-row elements. This has restricted the applicability of ML-derived BDE predictions, particularly for molecules of medicinal relevance, since the heteroatoms S, Cl, F, P, Br, and I are commonly found in approved pharmaceuticals. Atmospherically and environmentally relevant molecules containing multiple halogen atoms have been similarly inaccessible. In this study, we considerably expand the size, elemental composition, and bond types of an extensive BDE database and train a new ML BDE model that includes C, H, N, O, S, Cl, F, P, Br, and I. We curate a new quantum chemical dataset of 531 244 unique zero-point energy-inclusive homolytic dissociations of organic compounds. We investigate accuracy for out-of-sample molecules and implement iterative training and testing cycles during model development to improve the model accuracy. Improvements in predictive accuracy were achieved for datasets of pharmaceutically relevant molecules containing multiple C(sp2)–halogen bonds from 5.7 to 0.8 kcal mol^{-1} and Polyhaloalkyl compounds with multiple C(sp3)–halogen bonds from 2.7 to 1.2 kcal mol^{-1} through the targeted augmentation of training data by as little as eight additional molecules. Our updated and expanded model (ALFABET) achieves a mean absolute error of 0.6 kcal mol^{-1} for both enthalpies and free energies compared to the quantum chemical ground truth. The graph-based representations utilized here outperform traditional cheminformatics features such as radial fingerprints, and there is no discernible improvement in accuracy by including more expensive QM-derived parameters, such as optimized bond lengths. Finally, we illustrate high accuracy in external prediction tasks for large halogenated natural products, pharmaceutically relevant halogenated molecules, atmospherically important halocarbons, and poly-fluoroalkyl substances related to environmental toxicity.

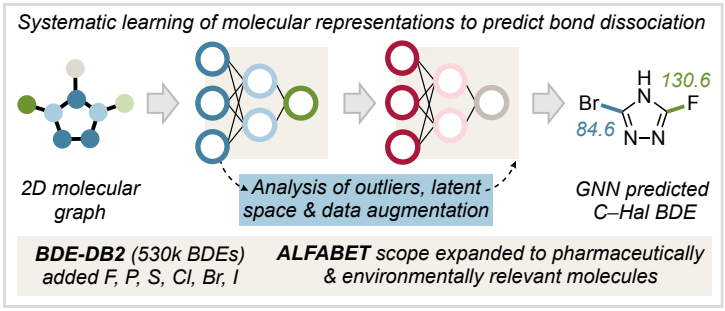


Figure 1.7. *Outline of work highlighted in Chapter 4 to build robust machine learning models for bond dissociation energy predictions.*

1.3.1 Quantitative approaches to chemical properties for goal-directed molecular optimization

In Chapter 5, I aim to combine the methodologies defined above to advance the field of goal-directed molecular optimization with the promise of finding feasible candidates for even the most challenging molecular design applications. One example of a fundamental design challenge is the search for novel stable radical scaffolds for an aqueous redox flow battery that simultaneously satisfies redox requirements at the anode and cathode, as relatively few stable organic radicals are known to exist. To meet this challenge, we develop a new open-source molecular optimization framework based on AlphaZero coupled with a fast, machine-learning-derived surrogate objective trained with nearly 100,000 quantum chemistry simulations. The objective function comprises three graph neural networks: one that predicts adiabatic oxidation and reduction potentials, a second that predicts electron density and local three-dimensional environment, previously shown to be correlated with radical persistence and stability, and a third to predict bond dissociation energy to determine the formation of radical species. With no hard-coded knowledge of organic chemistry, the reinforcement learning agent finds molecule candidates that satisfy a precise combination of redox, stability, and synthesizability requirements defined at the quantum chemistry level, many of which have reasonably predicted retrosynthetic pathways. The optimized molecules show that alternative stable radical scaffolds

may offer a unique profile of stability and redox potentials to enable low-cost symmetric aqueous redox flow batteries.

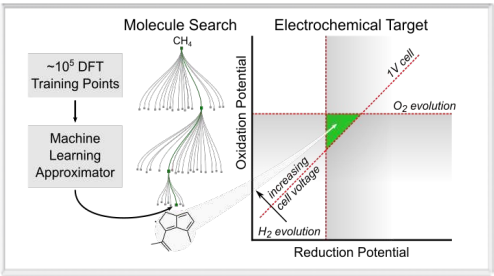


Figure 1.8. *Outline of work highlighted in Chapter 5 to identify new radical candidates for aqueous redox flow batteries.*

REFERENCES

1. Sliwoski, G.; Kothiwale, S.; Meiler, J.; Edward W. Lowe, J., Computational Methods in Drug Discovery. *Pharmacol. Rev.* **2014,** *66*, 334-395.

2. Xu, J.; Hagler, A., Chemoinformatics and Drug Discovery. *Molecules* **2002,** *7*, 566-600.

3. Muratov, E. N.; Bajorath, J.; Sheridan, R. P.; Tetko, I. V.; Filimonov, D.; Poroikov, V.; Oprea, T. I.; Baskin, I. I.; Varnek, A.; Roitberg, A.; Isayev, O.; Curtalolo, S.; Fourches, D.; Cohen, Y.; Aspuru-Guzik, A.; Winkler, D. A.; Agrafiotis, D.; Cherkasov, A.; Tropsha, A., QSAR without borders. *Chem. Soc. Rev.* **2020,** *49*, 3525-3564.

4. Lill, M. A., Multi-dimensional QSAR in drug discovery. *Drug Discov. Today* **2007,** *12*, 1013-1017.

5. Wells, P. R., Linear Free Energy Relationships. *Chem. Rev.* **1963,** *63*, 171-219.

6. Katritzky, A. R.; Lobanov, V. S.; Karelson, M., QSPR: the correlation and quantitative prediction of chemical and physical properties from structure. *Chem. Soc. Rev.* **1995,** *24*, 279.

7. Gallegos, L. C.; Luchini, G.; St. John, P. C.; Kim, S.; Paton, R. S., Importance of Engineered and Learned Molecular Representations in Predicting Organic Reactivity, Selectivity, and Chemical Properties. *Acc. Chem. Res.* **2021,** *54*, 827-836.

8. Karelson, M.; Lobanov, V. S.; Katritzky, A. R., Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chem. Rev.* **1996,** *96*, 1027-1044.

9. Yao, X. J.; Panaye, A.; Doucet, J. P.; Zhang, R. S.; Chen, H. F.; Liu, M. C.; Hu, Z. D.; Fan, B. T., Comparative Study of QSAR/QSPR Correlations Using Support Vector Machines, Radial Basis Function Neural Networks, and Multiple Linear Regression. *J. Chem. Inf. Comp. Sci.* **2004,** *44*, 1257-1266.

10. Xue, C. X.; Zhang, R. S.; Liu, H. X.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T., An Accurate QSPR Study of O−H Bond Dissociation Energy in Substituted Phenols Based on Support Vector Machines. *J. Chem. Inf. Comp. Sci.* **2004,** *44*, 669-677.

11. Kohn, W.; Becke, A. D.; Parr, R. G., Density Functional Theory of Electronic Structure. *J. Phys. Chem.* **1996,** *100*, 12974-12980.

12. Halgren, T. A., Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996,** *17*, 490-519.

13. Karplus, M.; McCammon, J. A., Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **2002,** *9*, 646-652.

14. Rosenbluth, M. N.; Rosenbluth, A. W., Monte Carlo Calculation of the Average Extension of Molecular Chains. *J. Chem. Phys.* **1955,** *23*, 356-359.

15. Williams, W. L.; Zeng, L.; Gensch, T.; Sigman, M. S.; Doyle, A. G.; Anslyn, E. V., The Evolution of Data-Driven Modeling in Organic Chemistry. *ACS Cent. Sci.* **2021,** *7*, 1622-1637.

16. Tu, Z.; Stuyver, T.; Coley, C. W., Predictive chemistry: machine learning for reaction deployment, reaction development, and reaction discovery. *Chem. Sci.* **2023,** *14*, 226-244.

17. Sigman, M. S.; Harper, K. C.; Bess, E. N.; Milo, A., The Development of Multidimensional Analysis Tools for Asymmetric Catalysis and Beyond. *Acc. Chem. Res.* **2016,** *49*, 1292-1301.

18. Raghavan, P.; Haas, B. C.; Ruos, M. E.; Schleinitz, J.; Doyle, A. G.; Reisman, S. E.; Sigman, M. S.; Coley, C. W., Dataset Design for Building Models of Chemical Reactivity. *ACS Cent. Sci.* **2023,** *9*, 2196-2204.

19. Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V., MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **2018,** *9*, 513-530.

20. Gasteiger, J.; Zupan, J., Neural Networks in Chemistry. *Angew. Chem. Int. Ed.* **1993,** *32*, 503-527.

21. Mitchell, J. B. O., Machine learning methods in chemoinformatics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014,** *4*, 468-481.

22. David, L.; Thakkar, A.; Mercado, R.; Engkvist, O., Molecular representations in AI-driven drug discovery: a review and practical guide. *J. Cheminform.* **2020,** *12*, 56.

23. Wigh, D. S.; Goodman, J. M.; Lapkin, A. A., A review of molecular representation in the age of machine learning. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2022,** *12*, e1603.

24. Kammeraad, J. A.; Goetz, J.; Walker, E. A.; Tewari, A.; Zimmerman, P. M., What Does the Machine Learn? Knowledge Representations of Chemical Reactivity. *J. Chem. Inf. Model.* **2020,** *60*, 1290-1301.

25. Guo, Z.; Guo, K.; Nan, B.; Tian, Y.; Iyer, R. G.; Ma, Y.; Wiest, O.; Zhang, X.; Wang, W.; Zhang, C., Graph-based molecular representation learning. *arXiv preprint arXiv:2207.04869* **2022**.

26. Bender, A.; Schneider, N.; Segler, M.; Patrick Walters, W.; Engkvist, O.; Rodrigues, T., Evaluation guidelines for machine learning tools in the chemical sciences. *Nat. Rev. Chem.* **2022,** *6*, 428-442.

27. Keith, J. A.; Vassilev-Galindo, V.; Cheng, B.; Chmiela, S.; Gastegger, M.; Müller, K.-R.; Tkatchenko, A., Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems. *Chem. Rev.* **2021,** *121*, 9816-9872.

28. Jolliffe, I., Principal Component Analysis. In *International Encyclopedia of Statistical Science*, Lovric, M., Ed. Springer Berlin Heidelberg: Berlin, Heidelberg, 2011; pp 1094-1096.

29. Cai, T. T.; Ma, R., Theoretical foundations of t-sne for visualizing high-dimensional clustered data. *J. Mach. Learn. Res.* **2022,** *23*, 1-54.

30. Pearson, K., LIII. On lines and planes of closest fit to systems of points in space. *Philos. Mag.* **1901,** *2*, 559-572.

31. Rokach, L.; Maimon, O., Decision Trees. In *Data Mining and Knowledge Discovery Handbook*, Maimon, O.; Rokach, L., Eds. Springer US: Boston, MA, 2005; pp 165-192.

32. Cristianini, N.; Ricci, E., Support Vector Machines. In *Encyclopedia of Algorithms*, Kao, M.-Y., Ed. Springer US: Boston, MA, 2008; pp 928-932.

33. Breiman, L., Random Forests. *Mach. Learn.* **2001,** *45*, 5-32.

34. McCulloch, W. S.; Pitts, W., A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biol.* **1943,** *5*, 115-133.

35. Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P. S., A Comprehensive Survey on Graph Neural Networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2021,** *32*, 4-24.

36. Mayr, A.; Klambauer, G.; Unterthiner, T.; Steijaert, M.; Wegner, J. K.; Ceulemans, H.; Clevert, D.-A.; Hochreiter, S., Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chem. Sci.* **2018,** *9*, 5441-5451.

37. Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M., Analyzing learned molecular representations for property prediction. *J Chem. Inf. Model.* **2019,** *59*, 3370-3388.

38. Shindo, H.; Matsumoto, Y., Gated Graph Recursive Neural Networks for Molecular Property Prediction. *arXiv preprint arXiv:1909.0025* **2019**.

39. Zhou, J.; Cui, G.; Hu, S.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; Sun, M., Graph neural networks: A review of methods and applications. *AI open* **2020,** *1*, 57-81.

40. Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. In *Neural message passing for quantum chemistry*, International conference on machine learning, PMLR: 2017; pp 1263-1272.

41. Duval, A.; Mathis, S. V.; Joshi, C. K.; Schmidt, V.; Miret, S.; Malliaros, F. D.; Cohen, T.; Liò, P.; Bengio, Y.; Bronstein, M., A Hitchhiker's Guide to Geometric GNNs for 3D Atomic Systems. *arXiv preprint arXiv:2312.07511* **2023**.

42. Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H., Molecular de-novo design through deep reinforcement learning. *J. Cheminform.* **2017,** *9*, 1-14.

43. Brown, N.; Fiscato, M.; Segler, M. H.; Vaucher, A. C., GuacaMol: benchmarking models for de novo molecular design. *J. Chem. Inf. Model.* **2019,** *59*, 1096-1108.

44. Popova, M.; Isayev, O.; Tropsha, A., Deep reinforcement learning for de novo drug design. *Sci. Adv.* **2018,** *4*, eaap7885.

45. Winter, R.; Montanari, F.; Steffen, A.; Briem, H.; Noé, F.; Clevert, D.-A., Efficient multi-objective molecular optimization in a continuous latent space. *Chem. Sci.* **2019,** *10*, 8016-8024.

46. Jensen, J. H., A graph-based genetic algorithm and generative model/Monte Carlo tree search for the exploration of chemical space. *Chem. Sci.* **2019,** *10*, 3567-3572.

47. Horwood, J.; Noutahi, E., Molecular Design in Synthetically Accessible Chemical Space via Deep Reinforcement Learning. *ACS Omega* **2020,** *5*, 32984-32994.

48. Thiede, L. A.; Krenn, M.; Nigam, A.; Aspuru-Guzik, A., Curiosity in exploring chemical space: Intrinsic rewards for deep molecular reinforcement learning. *Mach. Learn.: Sci. Technol.* **2020***, 3,* 035008.

49. Von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A., Exploring chemical compound space with quantum-based machine learning. *Nat. Rev. Chem.* **2020,** *4*, 347-358.

CHAPTER 2: DESIGNING COMPUTATIONAL TOOLS FOR AUTOMATION

2.1 Chapter overview

In this chapter, we tackle the challenge of automation in computational chemistry to build a robust recorded protocol that can be fully reproducible. We aim to incorporate FAIR data principles, which include guidelines for improving the Findability, Accessibility, Interoperability, and Reuse of data, we publish a Python package *Automated Quantum Mechanical Environment (AQME)* as open-source software for scientists in academia and industry. This collaborative research was published in Wiley's WIRES Computational Molecular Science journal with equal lead contributions from Dr. Juan V. Alegre and myself, Dr. Raúl Pérez-Soto, Turki M. Alturaifi, and Dr. Robert S. Paton. In this study, I contributed to developing the code for the automation, testing, and deploying of CSEARCH, CMIN, and QDESCP modules. I also worked on the end-to-end workflows highlighted in this study and helped write the manuscript for publication.

Reprinted with permission from: Alegre-Requena, J. V.; † **Sowndarya S. V. S.;** † Pérez-Soto, R,.; Alturaifi, T. M.; Paton, R. S. AQME: Automated quantum mechanical environments for researchers and educators. *WIREs Comput. Mol. Sci*. **2023**. *e1663*. with permission from Wiley.

2.2 Introduction

Continued improvements to computer hardware and algorithms have meant that quantum chemical studies are increasingly applied to study ever-larger and conformationally more flexible molecules and molecular datasets of increasing size. Computational high-throughput screening, chemical space exploration and molecular optimization, and the construction of high-

qualitydatasets to train emerging machine learning (ML) models rely on the ability to execute, analyze and store the results of large quantum chemistry campaigns.[1-8] These tasks typically require more than one calculation per molecule. For example, a sequence of molecule building, conformational analysis and refinement, optimization and thermochemical analysis, property prediction, and ensemble averaging is often performed. Each step may involve a different model chemistry (e.g., molecular mechanics, MM, quantum mechanics, QM, semi-empirical QM, SQM) executed by a distinct package. Further, these efforts are multiplied by the number of molecules. Automating these multi-step workflows minimizes manual effort and human error and enables the complex protocols and their associated data and metadata to be fully captured and reused.[9-13] Automated workflows for QM calculations, including those that address the important challenge of transition state (TS) location and conformational analysis (e.g., Wheeler's QChASM[14] and Duarte's autodE[15]), have emerged as powerful software tools, such as AARON,[16] ACE,[17] Aiida,[18] Auto-QChem,[19] CatVS,[20] Chemistream,[21] ChemShell,[22] FireWorks,[23] molSimplify,[24] PyADF,[25] and QMflows,[26] among others. In this work, we focus on the development of an automated end-to-end workflow software, AQME, to perform multi-step computational tasks spanning multiple programs and theoretical methods.

The modular design of AQME provides opportunities for workflow customization, use in Jupyter notebooks, and integration into other Python projects.[27] Ready-to-use examples with different degrees of complexity are provided via GitHub,[28] supplemented by hands-on video tutorials.[29] These examples can be trivially modified to create workflows that implement different software and levels of theory, or for application to different prediction tasks. For example, a researcher can create a workflow to calculate a reaction energy profile and, afterward,

tune the module combination to generate QM molecular descriptors to use in machine learning models.

Currently, the program contains five modules designed for different tasks (Fig. 2.1) that can be executed in any order; individual modules can be skipped if required. The input format can be a SMILES representation or many types of structure formats (SDF, PDB, XYZ, among others). The first module, CSEARCH, automates conformational analysis. This module is interfaced with molecular mechanics potentials through RDKit[30] and semi-empirical potentials through xTB. Searches can be performed externally using RDKit or CREST,[31] or using internally-coded systematic or Monte Carlo torsion sampling protocols. Then, CMIN refines these geometries and relative energies obtained from the initial conformer generation with semi-empirical methods (xTB) [32] or ML potentials (ANI).[33] However, this module can also be used to independently process 3D input formats. The next module, QPREP, converts a wide variety of 3D formats into input files for QM calculations with several packages, such as Gaussian,[34] ORCA,[35] and PySCF.[36] Tedious tasks such as (in Gaussian) creating Gen(ECP) sections or including final lines (i.e., NBO extra keywords) in the input files are handled automatically. QCORR is a cclib-based[37] module that detects issues and errors in QM output files, structures all output data, and creates ready-to-submit input files to correct those issues. User-specified criteria (i.e., spin contamination, isomerization, etc.) can be defined to filter output data. The last module, QDESCP, is designed to generate Boltzmann ensemble-averaged molecular QM properties or descriptors, which can be readily used in ML models. Commonly used descriptors such as atomic charges, bond orders, dipole moment, and solvation energy are included.

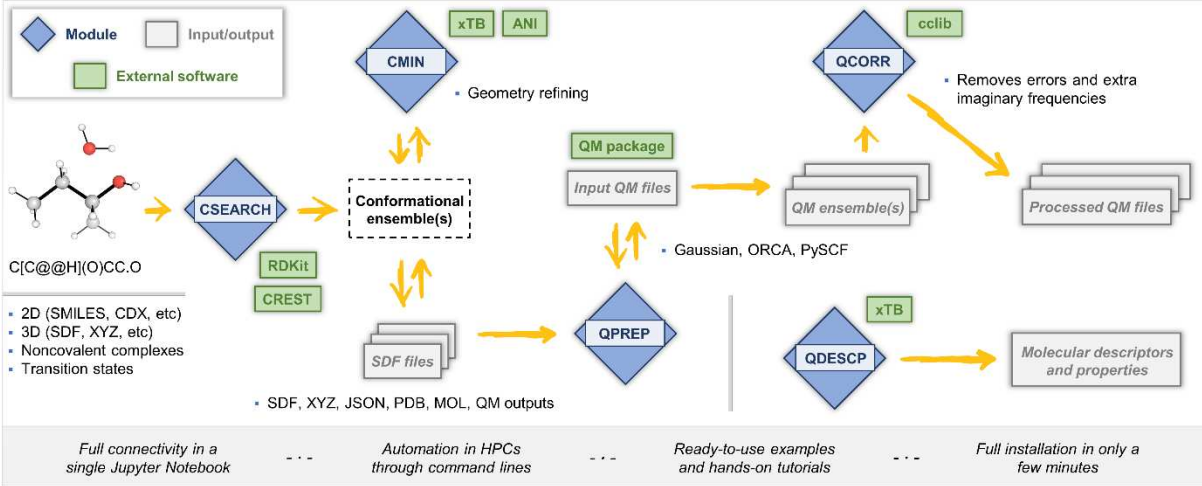


Figure 2.1. *General workflow including the modules available in AQME and their connectivity with external programs.*

2.3 Installation and Technical Details

The program presented is a free, open-source Python-based software installed via conda-forge (*conda install -c conda-forge aqme*) or Python Package Index (*pip install aqme*). All the dependencies required to run AQME are installed automatically, except for RDKit and Openbabel when using pip install. Along with the software, we set up tests through Pytest, Circle CI, and Codecov, which allows us to ensure a correct functioning of a significant proportion of the code. Also, multiple code analyzers (CodeFactor, Codacy, and LGTM) were employed to improve the quality standards and readability of the deployed code. A detailed documentation page is available at Read the Docs.[29]

2.4 Modules of AQME

AQME is divided into modules that can be called as part of a workflow or separately. Four main applications enclose these modules: i) conformer generation and geometry refinement (CSEARCH and CMIN), ii) generation of QM input files (QPREP), iii) post-processing of

output files (QCORR), and iv) generation of Boltzmann weighted descriptors (QDESCP). In this section, the technical details of the modules are disclosed in more detail.

2.4.1 Conformer Generation and Geometry Refinement: The CSEARCH and CMIN Modules

The availability of numerous conformer generators for small molecules (i.e., ConfGen,[38] OMEGA,[39] Frog2,[40] etc.) highlights the central importance of this task. In the CSEARCH module, AQME gathers multiple types of conformational search tools. It can be used simply as an interface to the external conformer generation protocols available in RDKit or CREST, or to perform torsion-based sampling internally while making use of the MM or SQM potentials in those packages (Fig. 2.2A). When starting from SMILES strings, CSEARCH attempts to derive molecular charge and multiplicity, although this can be manually overwritten (charge and mult options). The number of simultaneous processes is controlled by the max_workers option; the number of processors used by each process with the nprocs option.

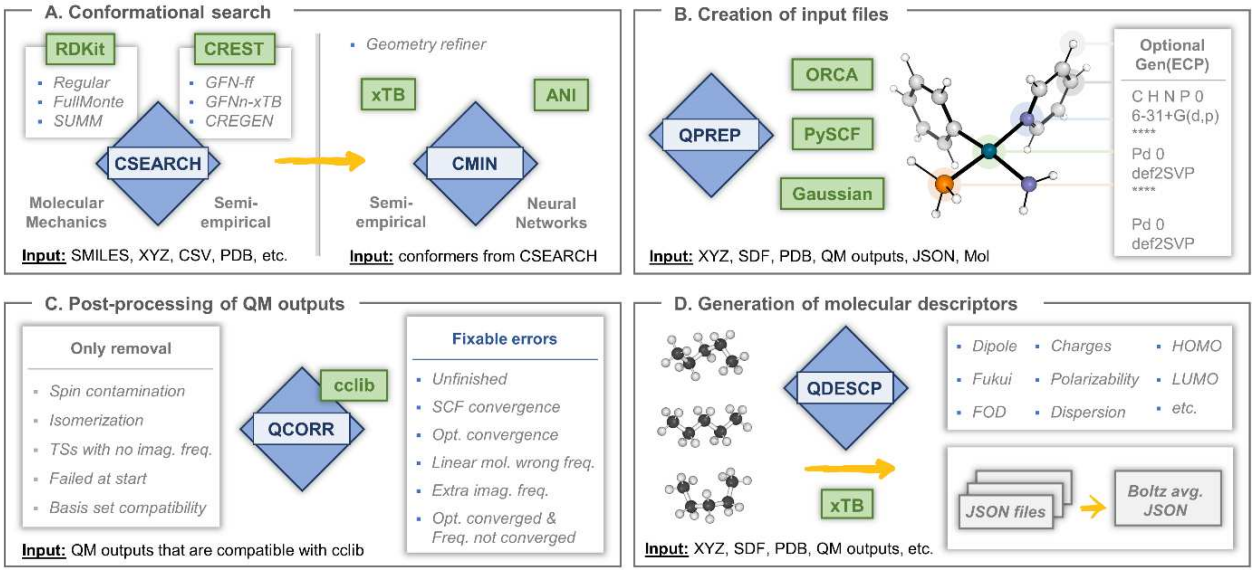


Figure 2.2. *Modules in AQME. (A) Methods and parameter options in CSEARCH and CMIN modules. (B) Recognition of atoms to include in the GenECP section of an example organometallic complex when using gen_atoms=["Pd"]. (C) Outline of error fixed in the QCORR module. Additional creation of input files for calculations for properties such as NMR,*

NBO, or higher-level single point energy evaluation. (D) Generation of Boltzmann averaged molecular properties using the QDESCP module for conformer ensembles.

The Grimme group's CREST generates conformers by extensive metadynamics sampling using semi-empirical methods (GFNn-xTB) or force fields (GFN-FF), with an additional genetic Z-matrix crossing step at the end. User-defined constraints for atom positions, bond distances, angles and dihedral angles enable approximate TS structures to be sampled. When starting from 1D and 2D structures, RDKit is used to generate the necessary 3D input for CREST. Before the sampling, two initial xTB optimizations are performed to avoid errors. During the first optimization, the calculations are performed with all the bonds frozen in addition to any user-defined constraints. This preparatory calculation avoids problems related to the superimposition of molecules when the input molecules are generated from 1D or 2D inputs. In the second optimization, only the user-defined constraints are included. Afterward, the CREST search is carried out, including any additional keywords specified in the *crest_keywords* option (i.e., crest_keywords="--nci --cbonds 0.5").

MM-based searches are faster and may be necessary for large systems or large numbers of molecules. The first method (using program="rdkit") performs RDKit-based conformer optimization and filters duplicates based on energy and geometry (with root mean square distance, RMSD). It starts with an energy window to remove high energy conformers (*ewin_csearch* option, default 5 kcal/mol above lowest), then removes conformers with similar energies (*initial_energy_threshold* option, default 0.0001 kcal/mol), and finally removes conformers with similar energy and RMSD (*energy_threshold*, default 0.25 kcal/mol and *rms_threshold*, default 0.25). The default values were chosen based on a benchmark study of flexible druglike compounds and natural products to yield significant conformers while reducing duplicates (see the *Benchmarking of CSEARCH-RDKit and CMIN* section in APPENDIX A.1).

When the initial conformational sampling fails, the program automatically tries to address the problems through a series of changes to its initial protocol (i.e., changing MMFF for UFF, using random coordinates for molecular embedding, etc.), making the protocol more robust. CSEARCH also tries to overcome other severe limitations in the conformational sampling of molecules that contain transition metals or atoms with uncommon hybridization (i.e., pentacoordinate P atoms). Additionally, common templates for organometallic compounds, such as linear, trigonal planar, and square-planar geometries, can be used. These geometries are not usually obtained with standard RDKit protocols (i.e., square-planar metal complexes lead to tetrahedral structures), which may then neglect an essential aspect of conformational behavior (see the *Highlighting the Importance of Specifying the Metal Type: ABEVUZ as an Example* section in APPENDIX A.2).

Internal torsional sampling approaches, such as the systematic unbounded multiple minimum (SUMM)[42] and Monte Carlo Multiple Minimum (MCMM) algorithms, are also implemented.[43, 44] The SUMM approach surveys dihedral angles that are progressively varied by a user-specified increment, while MCMM applies random values to a random subset of the rotatable torsions. These two methods require more time than the standard RDKit sampling, but they might render more accurate results in cases with complex conformational spaces. Finally, the CMIN module refines the energies and geometries of the structures obtained with RDKit or other low-level methods before optimizing with more demanding levels of theory, such as density functional theory (DFT). xTB or ANI methods typically result in a reordering of relative conformational stabilities closer to QM results and the removal of duplicate conformations.

2.4.2 QM Input File Preparation: The QPREP Module

The QPREP module is designed to convert multiple formats (SDF, XYZ, PDB, JSON, LOG/OUT) into input files for QM programs ready to be submitted without further modification. When using SDF and XYZ files, a QM input is generated for each structure in the files. For LOG/OUT calculations, only the final geometry is employed to create inputs for post-optimization single-point calculations (i.e., energy corrections at a higher theory level, TD-DFT calculations, etc.). QPREP currently generates inputs for Gaussian, ORCA, and PySCF. One of the most convenient features of this module is that the Gen and GenECP specifications from Gaussian input files are automatically written. When the user specifies atom types to use in gen_atoms, QPREP detects the atom types of the molecule and separates them into two groups for the input GenECP section. For example, the users can set the 6-31+G(d,p) basis set for C, H, N, and P atoms while using def2-SVP for Pd atoms (Fig. 2.2B). This automated protocol avoids the tedious manual setup of all the types of atoms in the GenECP part, which is especially helpful when working with different families of compounds or with big molecular datasets. The input keywords for the generated input files are specified through the *qm_input* option, and other parameters can be edited as preferred, such as charge, multiplicity, generation of CHK files, number of processors, and memory. Also, the user can include final lines after the molecular coordinates (i.e., NBO keywords).

2.4.3 Post-processing of Output Files: The QCORR Module

Typically, a tedious manual search and correction for error terminations, convergence issues, and extra imaginary frequencies is necessary after running QM calculations. Based on our experience with structure optimizations and frequency calculations for large databases (i.e., many

thousands) of organic compounds, such occurrences are relatively common. QCORR structures output data and automatically detects issues or errors, creating new input files that try to correct those issues, a cycle that can be repeated several times (Fig. 2.2C).

We conducted a study of 2709 calculations with organic molecules to find the optimal number of QCORR cycles for ground state optimization and frequency calculations in Gaussian 16 (QCORR_benchmarking.zip file in FigShare[41]). The initial geometries were obtained with a CSEARCH-RDKit standard conformer sampling. In the first round, 24% of the RDKit-generated conformers converged to duplicated QM conformers after optimization. From the unique structures, many outputs had problems: 1.5% showed imaginary frequencies, and 35% failed to converge to a stationary point in frequency calculation (albeit they converged during optimization). QCORR then automatically generated new inputs to fix the issues and these inputs were run. After the second round of QM calculations, 98% of the outputs had no issues, indicating that two QCORR rounds might be sufficient for most organic molecules. In our experience, an additional cycle is normally needed for complex systems with metals or supramolecular aggregates. The *freq_conv* keyword, which checks if a stationary point is found during optimization but not during frequency calculations, may be best avoided for flat or complex energy surfaces.

Structural isomerizations can be automatically detected and filtered. Also, calculations with spin contamination will be removed if $\langle S^2 \rangle$ differs from s(s+1) by more than 10%, as previously suggested.[45] The filter can be disabled or adjusted to other thresholds with the *s2_threshold* option. QCORR uses the *cclib* Python library and stores all parsed data as JSON files since this format allows other Python tools to retrieve the information easily. By default,

QCORR uses this information to detect all calculations for consistency in terms of calculation type and software version.

2.4.4 Generation of Boltzmann Weighted Properties and Descriptors: The QDESCP Module

When comparing computed results with experimental observables, Boltzmann-weighted values are generally advisable for molecules with multiple conformers. This also applies to molecular descriptors, where the utility of approaches that derive an ensemble average for a particular descriptor, or directly use minimum/maximum values, has been demonstrated.[46, 47] AQME is integrated with xTB to compute and curate computed descriptors for large compound databases. Starting from 3D geometries, atomic properties such as charges, fractional occupation densities, Fukui indices, D3-dispersion coefficients, and molecular properties including dipole moment, HOMO-LUMO gap, polarizability, energy are determined for every conformation of a molecule (Fig. 2.2D). Then, the Boltzmann averaged values of each descriptor are calculated and stored separately. This protocol enables the generation of SQM molecular descriptors as starting points for ML models. Additionally, QDESCP can be used to obtain Boltzmann averaged nuclear magnetic resonance (NMR) chemical shifts from DFT calculations. The user can specify slope and intercept to scale the results to the tetramethylsilane (TMS) scale using tools such as The Tantillo group's CHESHIRE repository,[48] rendering simulated spectra that can be compared directly with experimental spectra.

2.5 End-to-End workflows

In this section, we detail three illustrative workflows that have been adapted for different applications regularly carried out in our group, such as calculating energy profiles and generating

molecular descriptors for ML models. These workflows are available on Figshare[41] along with associated data and metadata, and three formats are available in the code and the Read the Docs webpage: Jupyter Notebook, SLURM script, and command-line script. Hands-on tutorials have been uploaded to YouTube.[29] For large systems or datasets, we typically execute end-to-end AQME workflows on a cluster using SLURM commands: the overall time taken is dominated by the QM calculation steps.

2.5.1 The Conformational Distribution and ^{1}H Chemical Shifts of Strychnine from SMILES Input

Strychnine is a natural alkaloid produced by different plants of the genus *Strychnos*, whose complexity and pharmacological properties have attracted many organic synthetic groups over time.[49] Recently, John, Reinscheid, and coworkers reported two different conformers in a 97:3 ratio observed in NMR studies.[50] The following AQME workflow aims to identify these two structures and simulate an averaged NMR spectra starting from a SMILES string, using a combination of i) RDKit conformer sampling, ii) Gaussian geometry optimization with B3LYP/6-31+G(d,p), iii) fixing errors and imaginary frequencies of the output files, iv) GoodVibes[51] calculation of Boltzmann distributions using Gibbs free energies at 298.15 K, and (v) Boltzmann averaged shielding tensor calculations (empirically-scaled to obtain chemical shifts) with B3LYP/6-311+G(2d,p), SMD = $CHCl_3$ (Fig. 2.3). Using the CSEARCH module with RDkit yields two conformers which are utilized further for DFT optimization. The calculated Boltzmann distribution for the two conformers is 99:1, which correlates well with the experimental observation of 97:3. Furthermore, the predicted ^{1}H chemical shifts present a low mean average error (MAE, 0.14 ppm for nine known ^{1}H signals) compared to the experimental values.[52] This workflow did not require manual intervention and suggests that further

applicability of AQME to automate NMR prediction and organic structure elucidation merits investigation.

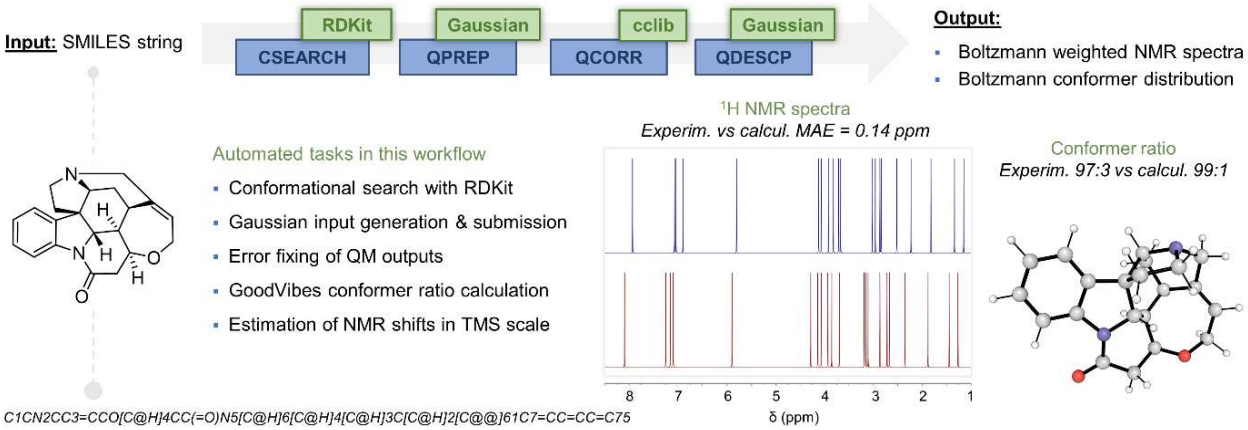


Figure 2.3. *End-to-end workflow to calculate the conformer distribution and scaled (TMS) ^{1}H NMR chemical shifts of strychnine in $CHCl_3$.*

2.5.2 Comparing Diels-Alder Activation Barriers from Multiple SMILES Inputs

A common computational task involves comparing the reactivity of several different substrates or reagents, for which the same elementary steps are studied separately for each of the related systems. Where transition states are involved, a common approach to conformational analysis involves constraining forming/breaking bonds while flexible regions are explored in much the same way as for a ground state structure, followed by saddle point optimizations. AQME can be employed to generate and compare energy profiles by automating this sequence of steps that is often performed manually. Fig. 2.4 summarizes a workflow where a CSV input containing a list of SMILES is used to generate the reaction energy profiles for the Diels-Alder cycloadditions of multiple cyclic dienophiles, each reacting with cyclopentadiene.[53] A Jupyter notebook was used to define SMILES strings and to identify the relevant atom numbers to define constraints that are used for the conformational analysis of TSs. Performed manually, these tasks would typically each structure to be built and visualized by the user. This approach is limited to reactions where

the TS structures are intuitively known; for automated TS location and energy profile generation without requiring prior knowledge, and mechanistic discovery, tools such as autodE are highly recommended.[15]

The workflow presented illustrates a typical multi-step combination of SQM conformer sampling, geometry optimizations and single point energy corrections with different levels of theory, and the generation of a potential energy surface diagram. AQME links together the following tasks: i) CREST conformer sampling, ii) Gaussian geometry optimization (B3LYP/def2-TZVP), iii) fixing errors and imaginary frequencies of the output files, iv) ORCA single point energy corrections using DLPNO-CCSD(T)/def2-TZVPP, and v) Boltzmann weighted thermochemistry calculation and PES generation with GoodVibes at 298.15 K. There is minimal manual intervention and the use of separate spreadsheets to create the PES is avoided.

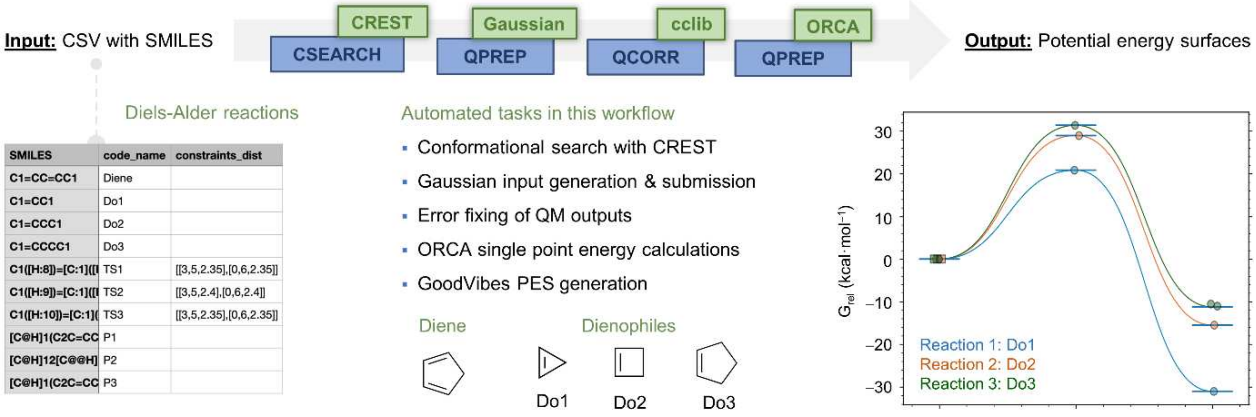


Figure 2.4. *End-to-end workflow to generate the energy profile of multiple Diels-Alder reactions using a CSV as the input file.*

2.5.3 Generating QM or SQM Molecular Descriptors for a Large Dataset

Statistical and ML applications in chemistry are often enhanced by using feature vectors or parameters derived from QM calculations, as opposed to features obtained solely from the 2D-molecular graph.[54] For example, QM-derived atomic charges or populations are often used in

multivariate linear regression or neural network models.[55] Fig. 2.5 shows a workflow performed on the SMILES-containing ESOL database,[56] which contains measured aqueous solubilities, that uses a message passing graph neural network (GNN) model. There are 1126 structures with experimental values available, which are split into training (901), validation (175) and test (50) sets. The protocol includes i) RDKit conformer sampling, ii) xTB descriptor generation (Boltzmann weighted), and iii) neural fingerprint (nfp)[57] based GNN model creation. In the first step, the CSEARCH module creates 3D conformations using the SMILES strings of the database with RDKit. Then, QDESCP uses xTB to generate molecular and atomic properties such as dipole moments, charges, Fukui indexes, dispersion parameters, polarizability, and HOMO-LUMO gaps, among others. These properties are provided individually for each conformer and as Boltzmann averaged values. In addition to properties generated from xTB, features in the Lipinski/Descriptors modules of RDKit are also included in the QDESCP analysis.

Boltzmann-weighted xTB parameters are then used as descriptors in a GNN model. The GFN2-xTB atomic properties are encoded as node features, and the molecular properties are passed as global features in the input graph structure. This graphical representation of the molecule with embedded atomic and molecular properties is used to build a message passing GNN. The GNN model utilized the AdamW optimizer, and model performance was assessed by measuring the mean absolute error during training for 500 epochs. The model showed an R^2 and MAE of 0.9 in the held-out test set of 50 molecules. Hyperparameter tuning can be performed along with feature selection to further improve this accuracy. Other ML models, such as random forests can also be employed in this workflow instead of the GNN shown.

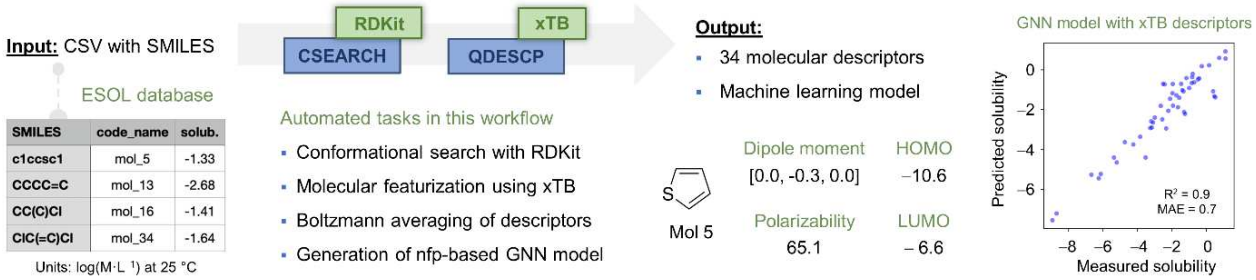


Figure 2.5. *End-to-end workflow to create and use molecular descriptors in a GNN model starting from a database of SMILES strings.*

2.6 Conclusion

AQME is an open-source Python package for building computational workflows to perform multi-step protocols efficiently that combine different packages and model chemistries. The approach is well-suited to general tasks incorporating conformational analysis, geometry refinement, QM optimizations, and ensemble-averaged property predictions. Representative examples of chemical shift prediction, reaction energy profile calculation, and dataset featurization are shown here, in each case operating as a fully automated ('end-to-end') workflow from the supplied inputs to desired Boltzmann-averaged outputs. Inputs can be supplied in SMILES or several 3D structure formats. This approach captures and preserves protocols used at every stage of the research process: there are no extraneous Spreadsheets involved and all steps can be reproduced by other researchers.

The software consists of independent modules that can be combined in any order. The CSEARCH module performs conformational sampling or interfaces to external conformational analysis tools using MM (RDKit) or SQM (xTB, CREST) levels of theory. CMIN allows the refinement of conformer ensembles with xTB or ANI methods. The QPREP module converts files with different 3D input formats into input files for external QM programs, such as Gaussian, ORCA, and PySCF. The QCORR module analyzes QM output data by systematically processing

output files. Filters for e.g., convergence errors, imaginary frequencies, and undesired structural isomerization, can be implemented to create new inputs automatically. The QDESCP module produces Boltzmann-weighted descriptors and properties. The modular structure means AQME can be used in Jupyter notebook environments or reused and imported by other Python projects.

REFERENCES

1. Burai Patrascu, M.; Pottel, J.; Pinus, S.; Bezanson, M.; Norrby, P.-O.; Moitessier, N., From desktop to benchtop with automated computational workflows for computer-aided design in asymmetric catalysis. *Nat. Catal.* **2020,** *3*, 574-584.

2. Sanchez-Lengeling, B.; Aspuru-Guzik, A., Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018,** *361*, 360-365.

3. Ahn, S.; Hong, M.; Sundararajan, M.; Ess, D. H.; Baik, M. H., Design and Optimization of Catalysts Based on Mechanistic Insights Derived from Quantum Chemical Reaction Modeling. *Chem Rev* **2019,** *119*, 6509-6560.

4. Durand, D. J.; Fey, N., Computational Ligand Descriptors for Catalyst Design. *Chem. Rev.* **2019,** *119*, 6561-6594.

5. Freeze, J. G.; Kelly, H. R.; Batista, V. S., Search for Catalysts by Inverse Design: Artificial Intelligence, Mountain Climbers, and Alchemists. *Chem. Rev.* **2019,** *119*, 6595-6612.

6. Vaissier Welborn, V.; Head-Gordon, T., Computational Design of Synthetic Enzymes. *Chem. Rev.* **2019,** *119*, 6613-6630.

7. Wagner, J. R.; Lee, C. T.; Durrant, J. D.; Malmstrom, R. D.; Feher, V. A.; Amaro, R. E., Emerging Computational Methods for the Rational Discovery of Allosteric Drugs. *Chem. Rev.* **2016,** *116*, 6370-90.

8. Colón, Y. J.; Snurr, R. Q., High-throughput computational screening of metal–organic frameworks. *Chem. Soc. Rev.* **2014,** *43*, 5735-5749.

9. Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A.; Chae, H. S.; Einzinger, M.; Ha, D. G.; Wu, T.; Markopoulos, G.; Jeon, S.; Kang, H.; Miyazaki, H.; Numata, M.; Kim, S.; Huang, W.; Hong, S. I.; Baldo, M.; Adams, R. P.; Aspuru-Guzik, A., Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater* **2016,** *15*, 1120-7.

10. Dunstan, M. T.; Jain, A.; Liu, W.; Ong, S. P.; Liu, T.; Lee, J.; Persson, K. A.; Scott, S. A.; Dennis, J. S.; Grey, C. P., Large scale computational screening and experimental discovery of novel materials for high temperature CO2 capture. *Energy Environ. Sci.* **2016,** *9*, 1346-1360.

11. Chakraborty, S.; Xie, W.; Mathews, N.; Sherburne, M.; Ahuja, R.; Asta, M.; Mhaisalkar, S. G., Rational Design: A High-Throughput Computational Screening and Experimental Validation Methodology for Lead-Free and Emergent Hybrid Perovskites. *ACS Energy Letters* **2017,** *2*, 837-845.

12. Hayes Robert, J.; Bentzien, J.; Ary Marie, L.; Hwang Marian, Y.; Jacinto Jonathan, M.; Vielmetter, J.; Kundu, A.; Dahiyat Bassil, I., Combining computational and experimental screening for rapid optimization of protein properties. *Proc. Natl. Acad. Sci. U.S.A.* **2002,** *99*, 15926-15931.

13. Sokolov, A. N.; Atahan-Evrenk, S.; Mondal, R.; Akkerman, H. B.; Sánchez-Carrera, R. S.; Granados-Focil, S.; Schrier, J.; Mannsfeld, S. C.; Zoombelt, A. P.; Bao, Z.; Aspuru-Guzik, A., From computational discovery to experimental characterization of a high hole mobility organic crystal. *Nat. Commun.* **2011,** *2*, 437.

14. Ingman, V. M.; Schaefer, A. J.; Andreola, L. R.; Wheeler, S. E., QChASM: Quantum chemistry automation and structure manipulation. *WIREs Comput. Mol. Sci.* **2021,** *11*, e1510.

15. Young, T. A.; Silcock, J. J.; Sterling, A. J.; Duarte, F., autodE: Automated Calculation of Reaction Energy Profiles— Application to Organic and Organometallic Reactions. *Angew. Chem. In. Ed.* **2021,** *60*, 4266-4274.

16. Guan, Y.; Ingman, V. M.; Rooks, B. J.; Wheeler, S. E., AARON: An Automated Reaction Optimizer for New Catalysts. *J. Chem. Theory. Comput.* **2018,** *14*, 5249-5261.

17. Corbeil, C. R. T., S.; Schwartzentruber, J. A.; Moitessier, N., *Angew. Chem. Int. Ed* **2008,** *47*, 2635-26.

18. Pizzi, G.; Cepellotti, A.; Sabatini, R.; Marzari, N.; Kozinsky, B., AiiDA: automated interactive infrastructure and database for computational science. *Comput. Mat. Sci.* **2016,** *111*, 218-230.

19. Żurański, A. M.; Wang, J. Y.; Shields, B. J.; Doyle, A. G., Auto-QChem: an automated workflow for the generation and storage of DFT calculations for organic molecules. *React. Chem. Eng.* **2022,** *7*, 1276-1284.

20. Rosales, A. R.; Wahlers, J.; Limé, E.; Meadows, R. E.; Leslie, K. W.; Savin, R.; Bell, F.; Hansen, E.; Helquist, P.; Munday, R. H.; Wiest, O.; Norrby, P.-O., Rapid virtual screening of enantioselective catalysts using CatVS. *Nat. Catal.* **2019,** *2*, 41-45.

21. Chemistream. https://chemistream.io

22. Metz, S.; Kästner, J.; Sokol, A. A.; Keal, T. W.; Sherwood, P., ChemShell—a modular software package for QM/MM simulations. *WIREs Comput. Mol. Sci.* **2014,** *4*, 101-110.

23. Jain, A.; Ong, S. P.; Chen, W.; Medasani, B.; Qu, X.; Kocher, M.; Brafman, M.; Petretto, G.; Rignanese, G.-M.; Hautier, G.; Gunter, D.; Persson, K. A., FireWorks: a dynamic workflow system designed for high-throughput applications. *Concurrency and Computation: Practice and Experience* **2015,** *27*, 5037-5059.

24. Ioannidis, E. I.; Gani, T. Z.; Kulik, H. J., molSimplify: A toolkit for automating discovery in inorganic chemistry. *J. Comput. Chem.* **2016,** *37*, 2106-17.

25. Jacob, C. R.; Beyhan, S. M.; Bulo, R. E.; Gomes, A. S. P.; Götz, A. W.; Kiewisch, K.; Sikkema, J.; Visscher, L., PyADF — A scripting framework for multiscale quantum chemistry. *J. Comput. Chem.* **2011,** *32*, 2328-2338.

26. Zapata, F.; Ridder, L.; Hidding, J.; Jacob, C. R.; Infante, I.; Visscher, L., QMflows: A Tool Kit for Interoperable Parallel Workflows in Quantum Chemistry. *J. Chem. Inf. Model.* **2019,** *59*, 3191-3197.

27. Van Rossum, G. D., F. L. , *Python 3 Reference Manual.* CreateSpace, Scotts Valley, CA.

28. AQME. https://github.com/jvalegre/aqme.

29. https://www.youtube.com/channel/UCHRqI8N61bYxWV9BjbUI4Xw.

30. RDKit: Open-source cheminformatics. http://www.rdkit.org.

31. 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.

32. Grimme, S.; Bannwarth, C.; Shushkov, P., A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1–86). *J. Chem. Theory Comput.* **2017,** *13*, 1989-2009.

33. 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.

34. 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 Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16 Rev. C.01*, Wallingford, CT, 2016.

35. Neese, F., The ORCA program system. *WIREs Comput. Mol. Sci.* **2012,** *2*, 73-78.

36. Sun, Q.; Berkelbach, T. C.; Blunt, N. S.; Booth, G. H.; Guo, S.; Li, Z.; Liu, J.; McClain, J. D.; Sayfutyarova, E. R.; Sharma, S.; Wouters, S.; Chan, G. K.-L., PySCF: the Python-based simulations of chemistry framework. *WIREs Comput. Mol. Sci.* **2018,** *8*, e1340.

37. O'Boyle, N. M.; Tenderholt, A. L.; Langner, K. M., cclib: A library for package-independent computational chemistry algorithms. *J. Comput. Chem.* **2008,** *29*, 839-845.

38. Watts, K. S.; Dalal, P.; Murphy, R. B.; Sherman, W.; Friesner, R. A.; Shelley, J. C., ConfGen: a conformational search method for efficient generation of bioactive conformers. *J. Chem. Inf. Model.* **2010,** *50*, 534-46.

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

40. Miteva, M. A.; Guyon, F.; Tufféry, P., Frog2: Efficient 3D conformation ensemble generator for small compounds. *Nucleic Acids Res.* **2010,** *38*, W622-7.

41. Alegre-Requena, J. V. S. V., S. S.; Paton, R.;, AQME paper examples. *Figshare Dataset* **2022**.

42. Goodman, J. M.; Still, W. C., An unbounded systematic search of conformational space. *J. Comput. Chem.* **1991,** *12*, 1110-1117.

43. Chang, G.; Guida, W. C.; Still, W. C., An internal-coordinate Monte Carlo method for searching conformational space. *J. Am. Chem. Soc.* **1989,** *111*, 4379-4386.

44. Robert S. Paton https://github.com/bobbypaton/FullMonte.

45. Young, D., Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems. *Wiley* **2001**, 228.

46. Brethomé, A. V.; Fletcher, S. P.; Paton, R. S., Conformational Effects on Physical-Organic Descriptors: The Case of Sterimol Steric Parameters. *ACS Catal.* **2019,** *9*, 2313-2323.

47. Newman-Stonebraker, S. H.; Smith, S. R.; Borowski, J. E.; Peters, E.; Gensch, T.; Johnson, H. C.; Sigman, M. S.; Doyle, A. G., Univariate classification of phosphine ligation state and reactivity in cross-coupling catalysis. *Science* **2021,** *374*, 301-308.

48. Lodewyk, M. W. S., M. R.; Tantillo, D. J. , CHESHIRE. *Chem. Rev.* **2012,** *112*, 1839-1862.

49. Bonjoch, J.; Solé, D., Synthesis of Strychnine. *Chem. Rev.* **2000,** *100*, 3455-3482.

50. Schmidt, M.; Reinscheid, F.; Sun, H.; Abromeit, H.; Scriba, G. K. E.; Sönnichsen, F. D.; John, M.; Reinscheid, U. M., Hidden Flexibility of Strychnine. *Eur. J. Org. Chem.* **2014,** *2014*, 1147-1150.

51. Luchini G, A.-R. J. V., ; Funes-Ardoiz, I.; Paton R.S.;, GoodVibes: automated thermochemistry for heterogeneous computational chemistry data *F1000Research* **2020,** *9*, 291.

52. SDBSWeb (National Institute of Advanced Industrial Science and Technology, date of access). Compound name: Strychnine, solvent: CDCl3 SDBS No.: 7596. https://sdbs.db.aist.go.jp/sdbs/cgi-bin/landingpage?sdbsno=7596

53. Liu, F.; Paton, R. S.; Kim, S.; Liang, Y.; Houk, K. N., Diels–Alder Reactivities of Strained and Unstrained Cycloalkenes with Normal and Inverse-Electron-Demand Dienes: Activation Barriers and Distortion/Interaction Analysis. *J. Am. Chem. Soc.* **2013,** *135*, 15642-15649.

54. Gallegos, L. C.; Luchini, G.; St. John, P. C.; Kim, S.; Paton, R. S., Importance of Engineered and Learned Molecular Representations in Predicting Organic Reactivity, Selectivity, and Chemical Properties. *Acc. Chem. Res.* **2021,** *54*, 827-836.

55. Stuyver, T.; Coley, C. W., Quantum chemistry-augmented neural networks for reactivity prediction: Performance, generalizability, and explainability. *J. Chem. Phys.* **2022,** *156*, 084104.

56. Delaney, J. S., ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *J. Chem. Infor. Comp. Sci.* **2004,** *44*, 1000-1005.

57. St. John, P. W., W.; S. V, S. S.; , NREL/nfp: patch for install without rdkit (0.3.12). *Zenodo* **2022**.

CHAPTER 3: A QUANTITATIVE METRIC FOR ORGANIC RADICAL STABILITY AND PERSISTENCE USING THERMODYNAMIC AND KINETIC FEATURES

3.1 Chapter overview

In this chapter, we develop a new metric to quantify organic radical stability utilizing both thermodynamic and kinetic features of a radical. We aimed to incorporate physical organic descriptors to quantify radical stability based on fractional spin and buried volume of the radical atom. To do so we curate a database of ~80,000 organic radicals and featurized the fractional spin and buried volume of the radical atom. These features are then combined to obtain a Radical Stability Score (RSS). The developed metric can be utilized to determine a Pareto front of stable organic radicals. This collaborative research was published in the Royal Society of Chemistry's Chemical Science with contributions from myself, Dr. Peter St. John, and Dr. Robert S. Paton. In this study, I contributed to defining the radical stability score metric, building the database of fractional spin and buried volume, analysis of the Pareto front, comparison of radical stabilization energy, and application to cascade reactions. This was done with guidance from my advisors Dr. Peter St. John and Dr. Robert Paton who also helped write the manuscript for publication.

Reprinted with permission from: **Sowndarya S. V. S.;** St. John, P.; Paton, R. S. A Quantitative Metric for Organic Radical Stability and Persistence Using Thermodynamic and Kinetic Features. *Chem. Sci*. **2021**. *12(39*), 13158–13166 with permission from the Royal Society of Chemistry.

3.2 Introduction

From the initial discovery of free radicals to their becoming textbook chemistry, it has been emphasized that a molecule containing an unpaired electron (i.e., a free radical) is likely very reactive. Over the years, however, organic chemists have addressed the importance of studying organic radicals' stability.[1] Since Gomberg's discovery that the triphenylmethyl radical does not dimerize to form hexaphenylethylene, the search for stable radicals has intensified.[2] In this field, the term "stable radical" is proposed to refer to a compound that is unreactive enough so that "the pure radical can be handled and stored in the lab with no more precautions than would be used for the majority of commercially available organic chemicals".[3] Ingold also classified radicals according to their lifetime; transient radicals are those with a half-life less than a millisecond, while persistent radicals are those with a longer half-life.[3] In this context, a stable radical able to be stored in air can be seen as an example of extreme persistence.

Most radicals are highly reactive and transient. In contrast, stable radical species possess unique electronic and reduction-oxidation (redox) properties that have spurred interest as potential materials in energy storage and energy conversion devices.[4-7] The electronic and steric features influencing radical stability have been well studied; however, predicting a radical's lifetime remains challenging, and only a handful of experimentally stable organic radical species have been reported (Fig. 3.1).[8, 9] While it may be possible to extend the lifetime of an already persistent radical through structural fine-tuning, the discovery of new stable radical functionalities has been restricted by the absence of a quantitative description of radical stability that extends beyond simple thermodynamic considerations, and that can be used predictively.

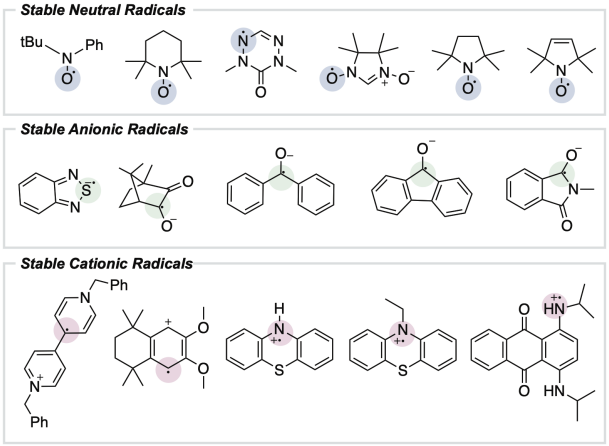


Figure 3.1. *Experimentally confirmed stable organic radicals.*

Comparative studies of radical stability have focused mainly on aspects of thermodynamic stabilization. For example, these effects in organic radicals have been quantified by defining a radical stabilization energy (RSE) scale.[10, 11] The RSE for a carbon-centered radical R· is defined by the difference between C–H bond dissociation energies (BDEs) of methane and R–H (an isodesmic H-atom transfer reaction).[12] Weaker R–H bonds give rise to increasingly positive RSE values that reflect increasing thermodynamic stabilization of the R· radical. Alternative RSE schemes have been proposed to minimize the differential contributions of C–H bond polarity in these comparisons, such as by comparing the C–C BDEs of CH_3–CH_3 vs. R–R.[13, 14] Additionally, for radicals centered on oxygen, nitrogen, or sulfur atoms, molecules such as H_2O, NH_3, and H_2S define separate RSE scales.[15-17]

RSE schemes have been instrumental in enabling quantitative comparisons of radical stabilization that can be performed computationally, using composite *ab initio* methods or lower-cost density functional theory (DFT) calculations. However, several fundamental issues limit the usefulness of RSE values for discovering new stable radicals. Firstly, persistence is a kinetic phenomenon relating to rates of reactivity and influenced by steric stabilization, while RSEs are

thermodynamic in origin. For example, although electron delocalization provides thermodynamic stabilization and an RSE value of 14.6 kcal/mol, the benzyl radical is a transient species with a lifetime of less than a millisecond.[18] Secondly, the referencing of RSE values with a specific bond-type (e.g., C–H) does not allow for a universal comparison of different radical types, such as carbon-centered vs. oxygen-centered radicals.[19]At the same time, quantitative metrics to describe kinetic persistence are still in their infancy, and no general method is yet available. To address these limitations, we propose a new metric for quantifying radical persistence, able to identify and predict stable organic radical structures. Its development incorporates kinetic and thermodynamic considerations, and the metric is generally applicable to carbon- or heteroatom (N, O, S) centered radicals. We assembled a sizeable computational database of organic radicals and have identified two DFT-derived features derived from the quantum mechanical spin density distribution and the molecular geometry that can be used to chart thermodynamic and kinetic variability. With this approach, we can cluster radicals into distinct regions and identify the region of this feature space where experimentally validated stable radicals reside.[20] The resulting quantitative metric for stable radicals provides a route to high-throughput searches and generative design efforts for stable organic radicals precisely tuned for emerging energy material applications.

3.3 Results and Discussion

We first consider how radical structures differ in the two principal characteristics introduced above, kinetic lifetime and thermodynamic stabilization. We can visualize these variations on a map of relative radical stability, which is divided into four quadrants (Fig. 3.2): (i) thermodynamically destabilized, kinetically transient radicals such as $CH_3\cdot$ in the SE quadrant;

(ii) thermodynamically stabilized, kinetically transient structures such as the benzyl radical in the SW quadrant; (iii) thermodynamically destabilized, kinetically persistent radicals such as a sterically protected 1,5-disubstituted phenyl radical in the NE quadrant; (iv) thermodynamically stabilized, kinetically persistent radicals in the NW quadrant. All stable radicals exist in this final quadrant.

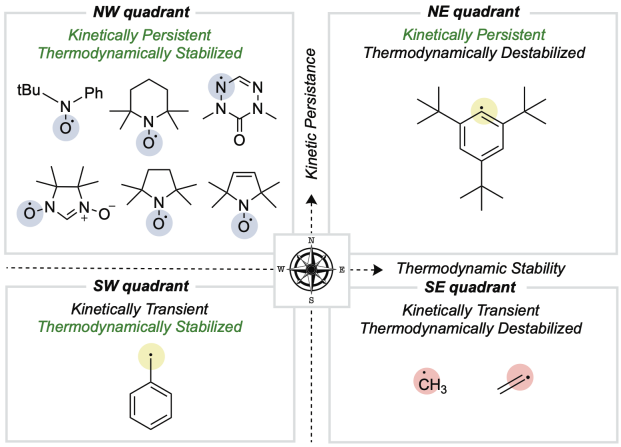


Figure 3.2. *Mapping organic radicals according to two criteria, kinetic persistence and thermodynamic stabilization. The most "stable" organic radicals reside in the NW quadrant.*

The thermodynamic stabilization of radicals can arise from resonance stabilization, such as via conjugation with an adjacent π-system or lone-pair, or hyperconjugation with a neighboring σ-bond.[21-25] Extended conjugation involving several π-bonds allows extensive spin delocalization. For example, the triphenylmethyl radical has an RSE of –24.7 kcal mol^{-1} due to extended conjugation, in contrast to benzyl radical, which is –14.6 kcal mol^{-1}.[26] Another factor influencing radical stability is the element(s) on which the spin density is located. For example, N-, O- and S-centered radicals are disproportionately represented among known stable radical structures.[27] This observation may appear counterintuitive from a thermodynamic standpoint since the BDE values of X–H bonds are generally larger than C–H bonds. However, these elements possess lone pairs and are highly electronegative, slowing the kinetics of radical

dimerization and reaction with molecular O_2, respectively.[27] Substituent effects also help stabilize radicals, where the presence of polar groups provides conjugative/inductive effects, which help in charge distribution and spin delocalization. The presence of both donor and acceptor substituents results in appreciable resonance, known as the captodative effect.[28] Alongside intrinsic structural effects, extrinsic factors such as changes in pH can result in radical stabilization due to changes in orbital configuration that result in the singly occupied molecular orbital (SOMO) no longer being highest in energy.[29]

While thermodynamic stabilization is solely electronic in origin, the kinetics of radical reactions can be profoundly influenced by steric effects. A simple example of an increase in steric bulk resulting in lower reactivity involves replacing H atoms with heavier elements such as Cl.[30] The incorporation of branched substituents such as *t*-butyl groups close to the radical center have an even more significant effect.[31] This phenomenon is further exemplified by stable radicals such as 2,2,6,6-tetramethylpiperidin-1-yl)oxyl (TEMPO), where two gem-dimethyl groups provide steric protection for the long-lived nitroxyl radical. Sigman and Sanford have illustrated the effects of N-alkyl group length in extending the half-life of pyridinium radicals.[32] Bulky substituents around a radical center result in steric hindrance towards bimolecular reactions, e.g., with other radicals, solvent molecules, and molecular O_2.

3.2.1 Towards a Quantitative Metric for Radical Stability

We sought to develop numerical descriptors for the two independent dimensions of radical stability: (i) thermodynamic stabilization via electron delocalization and (ii) kinetic persistence due to steric hindrance.[33] In contrast to RSE scales, we were keen to avoid an atom or bond-

specific scheme and instead focus on describing the extent of spin-delocalization directly. This led us to consider the largest atomic (Mulliken) spin density as a descriptor for the effect of thermodynamic stabilization. Steric effects on pyridinium radicals' lifetimes have been described previously using multidimensional Sterimol parameters[34, 35] to quantify the out-of-plane distance of N-alkyl substituents close to the radical center. To generalize this concept for radicals of arbitrary 3D-geometry, we focused on describing the extent to which adjacent functional groups occupy the space around a radical center. In our work, this is quantified by the percent buried volume, defined as the occupied percent of the total volume of a sphere with a defined radius centered around the radical. Cavallo and Nolan developed the buried volume parameter to capture a coordinated ligand's steric demands around a central metal.[36, 37] To our knowledge, this has not been used previously to describe radicals' kinetic persistence. In our buried volume calculations, we define the radical center as the atom with the maximum fractional spin, which we expect to exert a strong influence on radical reactivity.

Most radicals generated in organic chemistry are short-lived and unstable. Therefore, by comparing the spin and buried volumes scores of radicals known experimentally to be stable to a large sample of more typical organic radicals, we can validate that our metric can explain radical stability by separating known-stable from likely unstable radical species. A recently generated database of 200,000 quantum chemical calculations performed at the M06-2X/def2-TZVP[38] level for small, organic open-shell molecules generated by breaking single, non-cyclic bonds in molecules taken from the PubChem Compound database provides such a sample of expected organic radical configurations.[20, 39] These radicals were obtained after conformer sampling using the MMFF94 force field within RDKit,[40, 41] following which the lowest-energy conformer was

utilized for DFT optimizations. Molecules with 10 heavy atoms or fewer containing only C, N, S, O, and H atoms were re-optimized using water as an implicit solvent using Gaussian 16.[42] All calculations reported in the main text were optimized in water using the SMD solvation model,[43] while gas-phase values are reported in the APPENDIX (Fig. B.S1). Water was used due to its relevance in the performance of redox batteries. For all the radicals in this dataset, Mulliken spin densities were obtained for each atom, and the corresponding buried volume around each atom was computed from the optimized molecular geometry. For each molecule, the computed spin density values were normalized: the absolute magnitudes of heavy atom spins were summed, neglecting the small spin values on H atoms, and then converted to fractional spins that sum to one. The center with the highest fractional spin was assigned as the location of the radical center. Negligible basis set dependence of spin densities was confirmed by comparing with def2QZVP for a small set of radicals (APPENDIX Table B.S1). A Python package, DBSTEP, was developed to aid the high-throughput evaluation of buried volumes for almost 90,000 compounds, using numerical integration on a Cartesian grid with a spacing of 0.05 Å. Voxel occupancies were determined based on the unscaled atomic Bondi radii for all atoms. A sphere radius of 3.5 Å was used throughout (Fig. 3.3A).[44] The final breakdown of organic radicals in the dataset based on their location of maximum spin density are as follows: 73,080 carbon, 5,097 oxygen, 9,693 nitrogen, and 1,447 sulfur (APPENDIX, Fig. B.S2).

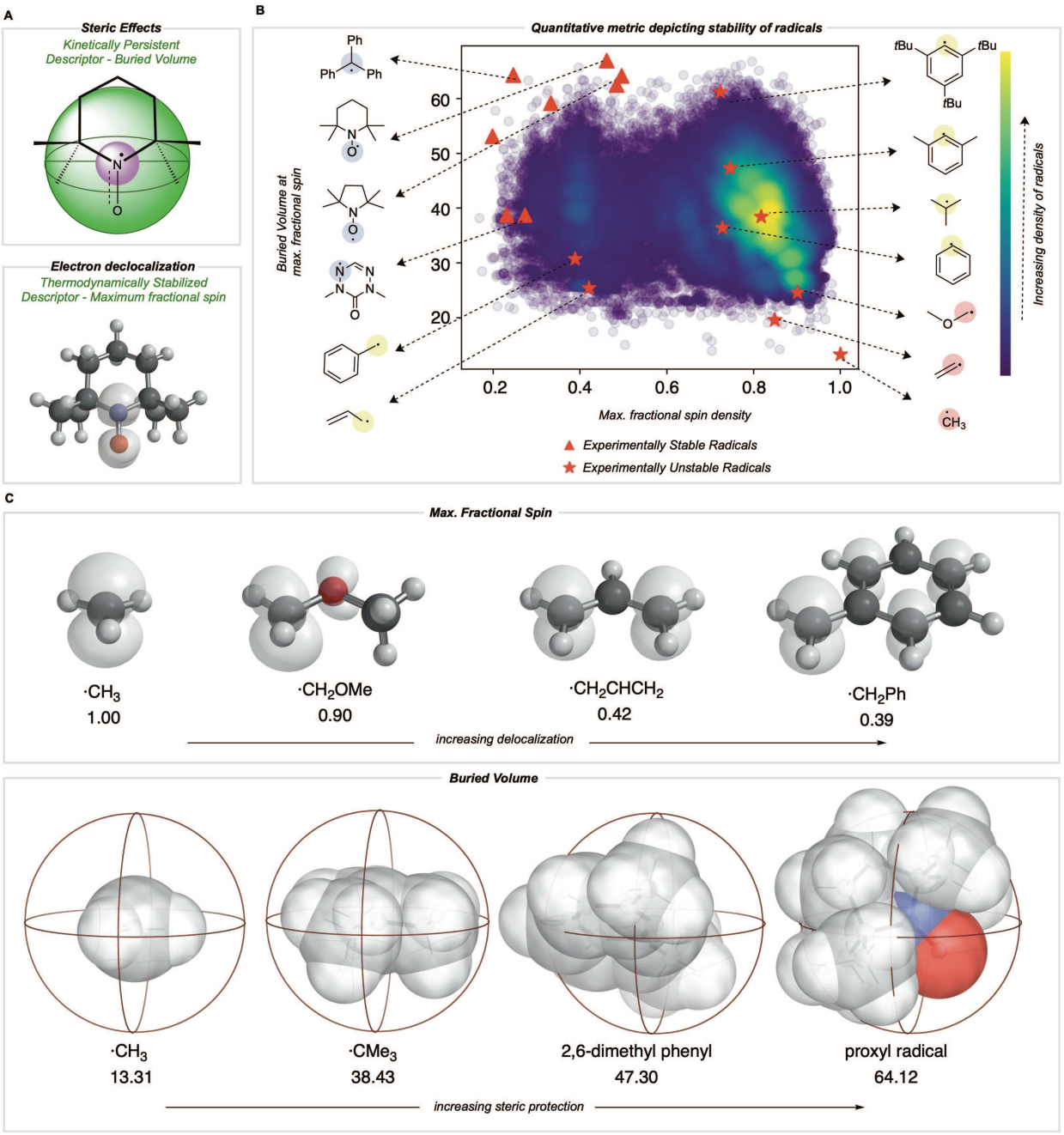


Figure 3.3. *(A) Parameters used to build a stability metric for organic radicals in two dimensions corresponding to kinetic and thermodynamic stabilization. (B) Quantitative evaluation of these metrics produces a map in which experimentally validated stable radicals (red triangle) are separated from common radical structures. (C) Visualization of electronic and steric descriptors.*

The distribution of spin density and steric descriptor values obtained for the entire dataset of 73,080 organic radicals is shown in Fig. 3.3B. We also computed descriptor values for specific

radicals whose stability and persistence are experimentally known (red points in Fig. 3.3B). The appearance of validated stable radicals, such as TEMPO, proxyl, and trityl radicals at the frontier of this distribution (top left corner), was encouraging since these structures score amongst the highest in terms of both thermodynamic stabilization (low spin density) and kinetic persistence (high buried volume). Structures for all plotted stable radical species are depicted in the APPENDIX (Fig. B.S2). The benzyl radical appears in the lower-left quadrant, indicative of favorable delocalization but low buried volume around the radical center, consistent with its transience. Resonantly stabilized radicals such as these play an essential role in soot precursor formation, as their relatively long lifetimes and multiple reactive sites lead to further ring growth during combustion. Conversely, the 2,4,6-tri-tert-butyl phenyl radical appears in the upper right quadrant due to high buried volume but a highly localized unpaired electron. Finally, small and unstable structures such as the methyl radical appear in the lower right, indicating an absence of either thermodynamic stabilization or kinetic protection. Additionally, a higher density of radicals is obtained in the lower right corner of this map, which signifies that an organic radical chosen at random is likely to be unstable and short-lived.

Both electron delocalization and steric protection play an important role, and some examples of spin density and buried volume descriptors are illustrated in Fig. 3.3C. The spin density in a methyl radical is fully localized to the carbon atom (our scheme only considers non-hydrogen atoms), giving a value of 1. However, spin delocalization in structures such as the benzyl radical results in a maximum spin density value less than 0.5 on the CH_2. In the most highly delocalized structures, we obtained maximum spin densities close to 0.2. Further quantitative justification for the use of max. spin density as a descriptor of thermodynamic

stability comes from the fact that it is highly correlated with RSE values defined by the cleavage of a particular bond type (e.g., C(sp^3)-H, R^2 = 0.94) and with the BDE values for over 100,000 C-H bonds (R^2 = 0.66) (APPENDIX, Fig. B.S5). Similarly, the buried volume values ranged from 13% for the smallest radicals such as ·CH_3 to around 65% for the most sterically protected structures, such as the proxyl radical or TEMPO. These steric descriptor values are most strongly influenced by the degree of substitution at the site of the maximum spin density and of each of the neighboring atoms, as illustrated by the contents of the 3.5 Å radius spheres shown in Fig. 3.3C. We found the results from using different sphere sizes to generate these buried volumes (2.5, 3.0, 3.5, 4.0Å) to be very highly correlated (R^2=0.96-0.99) and so we have retained the familiar value of 3.5Å throughout (APPENDIX, Fig. B.S9).

Having seen that empirically validated stable radicals occupy a distinct region of our parameter space, we next focused on identifying other molecules predicted to be similarly stable. This analysis relies on identifying the Pareto optimal set of radicals – those structures for which there are no other examples superior both in terms of buried volume and delocalization (Fig. 3.4)[45, 46]. We added several charged structures known to be stable radicals (e.g., phenyl-viologen) to our dataset (APPENDIX Fig. B.S2) at this point. The Pareto frontier set of radicals was identified from our large dataset, separated according to the atom (C, N, O, or S) with the largest spin density. The proximity of some of these computationally derived structures to known stable radicals (red triangles) encouraged us to explore these as potential new stable radical candidates.

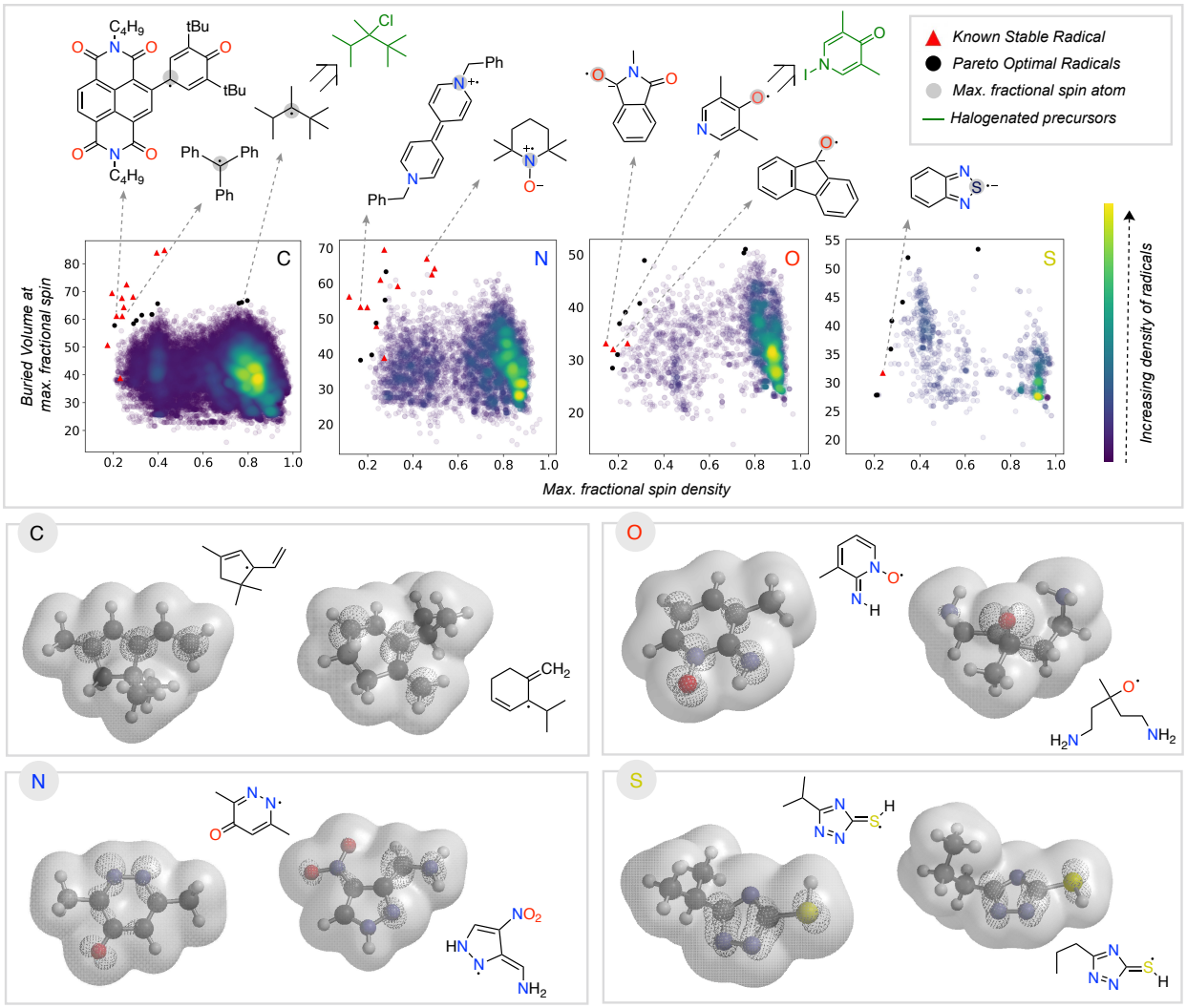


Figure 3.4. *Superimposition of known stable radicals with optimal predicted organic radicals according to fractional spin density and buried volume parameters. Analysis separated into C, N, O, and S- radical centers. Below, spin densities and molecular isosurfaces are show for selected molecules from the Pareto front.*

Some of the computationally optimized molecules on the Pareto frontier are shown in Fig. 3.4. They are divided into the type of atom the radical center lies on upon DFT optimization. All structures contain high delocalization, steric protection, or both. Based on their shared features with currently known stable radicals, we propose that compounds on the Pareto frontier are potential candidates for long-lived radicals. As an illustration, two such radicals computationally predicted to be Pareto optimal could be synthesized in one step from

commercially available halogenated precursors (shown in green, Fig. 3.4).[47, 48,49] Our newly developed metric is, therefore, able to aid in the identification of stable radical structures. On a general note, stable radicals lie on the top left region of each graph based on atom type. However, we also observe that stable heteroatom radicals tolerate lower buried volumes (as low as 40%) than their carbon-centered counterparts. As discussed above, we attribute this to their electronegativity and lone-pairs, which retard reactions at the radical center.

3.2.2 Comparison to radical stabilization energies

RSE values are commonly invoked to quantify radical stability.[10, 11] In this section, we compare our newly developed stability metric with traditional RSE values of carbon-centered radicals, derived from the difference in C–H BDE values relative to methane. To do this, we formulated our two parameters into a singular stability metric (equation 2).

$$Radical\ Stability\ Score = V_{bur} + 50 \times (1 - Max.\ Spin) \qquad (2)$$

The factor of 50 in this equation was chosen to give approximately equal weight to buried volume and spin terms, reflecting kinetic and thermodynamic contributions to radical stability. In contrast, RSE or BDE values are purely thermodynamic. The M06-2X/def2-TZVP gas-phase BDE values for 107,717 C–H bonds were collected for our dataset of 200,000 open-shell molecules. Comparing the number of C–H bonds to the C-centered radicals (73,080), the increase is due to the fragmentation scheme from a parent molecule. Multiple parent molecules can generate the same C-centered radical, corresponding to different C–H bonds breaking reactions. We compare these BDE values with our Radical Stability Score (RSS) in Fig. 3.5.

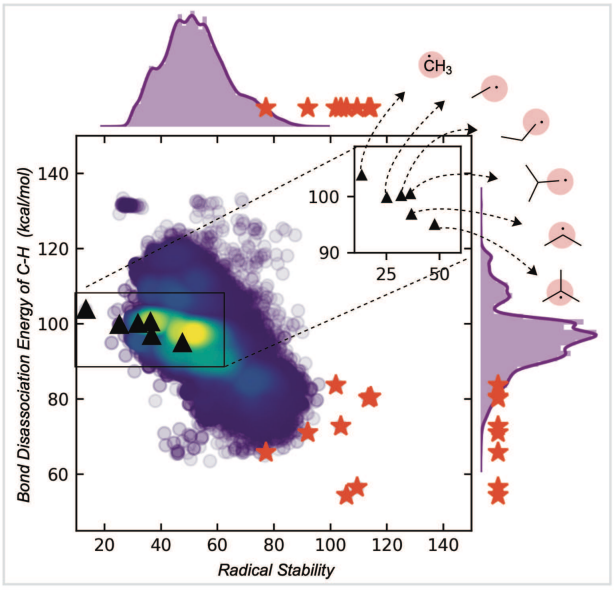


Figure 3.5. *Correlation of radical stability score with C–H BDE. The red stars are known stable radicals.*

While RSS values are inversely correlated with BDEs, the correlation is relatively modest (R^2 = 0.6). The RSS values provide different information, as illustrated for a family of alkyl radicals (Fig. 3.5). Relative C–H BDE values demonstrate the sequence of thermodynamic stabilities: methyl < primary < secondary < tertiary radicals. However, on solely thermodynamic grounds, there is little to distinguish between the stabilities of primary ethyl, *n*-propyl and *i*-propyl radicals, and the *t*-butyl radical is only marginally more stable than the secondary *i*-propyl radical. RSS values preserve the overall order of stabilities of primary, secondary, and tertiary radicals. However, they also provide additional stratification: the primary radicals are separated with longer or bulkier alkyl chains contributing to a greater stability score. The tertiary radical is more clearly differentiated from the rest of the structures due to steric effects. We suggest that this scale offers greater resolution to separate stable from unstable candidates. Furthermore, The RSS score exhibits a good correlation (R^2 = 0.85) with the relative rates of decomposition of 18

radicals, which exhibit second order kinetics consistent with radical homocoupling (APPENDIX, Fig. B.S9).

RSS values are also more helpful than BDE values for classification tasks. For example, the extreme instability of the methyl radical can be inferred from the fact that it is a statistical outlier in the overall distribution of RSS values. This is not the case when looking at the BDE distribution (APPENDIX, Fig. B.S6). At the other and potentially more helpful end of the spectrum, empirically stable radicals are also clustered towards the top end of the RSS distribution (red stars, Fig. 3.5), more so than for the distribution of BDE values, which were obtained using a machine learning BDE prediction tool, ALFABET.[39] All known stable radicals considered in this work are found in or above the 97th percentile of RSS values, which drops slightly to the 93rd percentile for BDE value. Therefore, we anticipate that the use of RSS values in predictive screening for new stable radicals will result in greater enrichment compared to more traditional metrics.

3.2.3 Utility in studies of radical cascade reactions

The RSS metric and the underlying two descriptions of thermodynamic stabilization and kinetic persistence enable radicals' stabilities to be compared quantitatively. We assessed the utility of this approach in comparing radical intermediates occurring sequentially along a reaction pathway as a collective variable for the reaction coordinate. We studied three cascade reactions involving sequential radical *exo*-trig/dig cyclization steps,[50-53] computing the thermochemistry at the M062X/def2TZVP level of theory (Fig. 3.6). Each successive intermediate is more stable than the last, and while this is influenced by the nature of bonds formed and broken, the evolution of

the fractional spin descriptor illustrates the contribution from greater radical delocalization being particularly noticeable in the second step, while the main increase in kinetic protection occurs in the first step. The final values of two of these structures suggest that these would be somewhat stable radicals with appreciable lifetimes. While the RSS metric describes structural and electronic changes around the site of an unpaired electron, other contributions to a reaction's driving force, such as strain-release from ring-opening, may also play a key role and would need to be separately accounted for.[54]

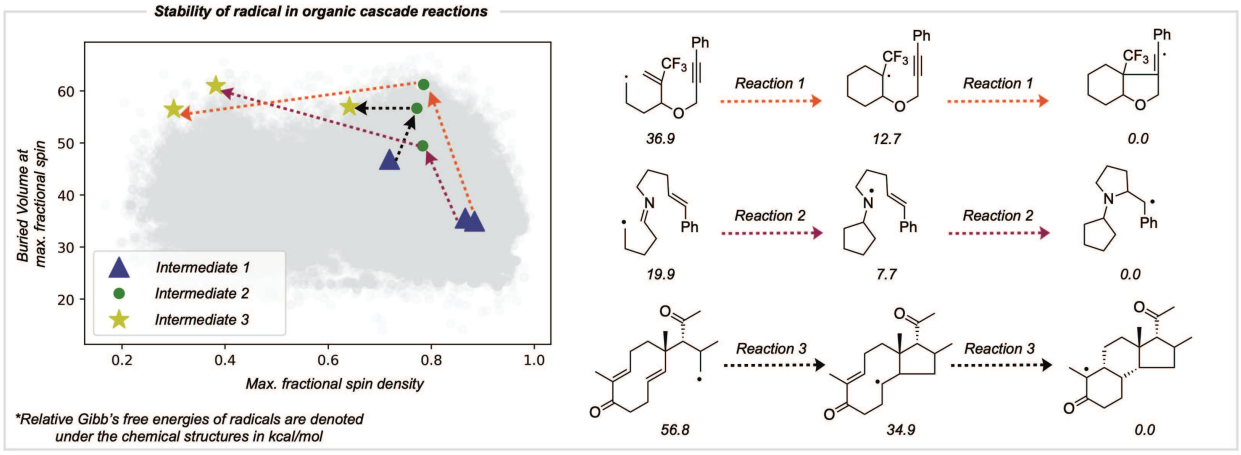


Figure 3.6. *Improvement in radical stability in consecutive intermediates in organic radical cascade reactions. The type of arrows on the reaction depicts the respective reaction on the stability graph. Computationally optimized molecules are depicted in grey in the background.*

3.3 Conclusion

Inspired by thermodynamic and kinetic considerations, we propose two key computational descriptors for quantifying radical stability, the maximum spin density and the buried volume around the atom where this spin is located. Two-dimensional plots of these descriptors were generated for thousands of common radicals and for experimentally verified stable, highly persistent radicals. Stable radicals appear in a distinct region of this feature space associated with thermodynamic stability and kinetic persistence. This analysis gives rise to the notion that stable

organic radicals can be computationally predicted and designed in a quantitative fashion. The two dimensions of radical stability can be considered independently or combined into a Radical Stability Score (RSS). This metric preserves the thermodynamic information contained in traditional metrics derived from BDE values while also allowing for differentiation between sterically distinct radical centers. Overall, this enables more straightforward distinctions between the small population of stable radicals and the vast majority of shorter-lived structures. We also demonstrated the use of these descriptors to compare successive radical intermediates occurring along a reaction pathway. This work provides a unified evaluation of radical stability that could be incorporated as an objective function in molecular design efforts using high-throughput screening or generative machine learning. The discovery of new, more stable radicals for use in electronic devices such as redox batteries, transistors, and light-emitting diodes is our overarching goal.

REFERENCES

1. Tang, B.; Zhao, J.; Xu, J.-F.; Zhang, X., Tuning the stability of organic radicals: from covalent approaches to non-covalent approaches. *Chem. Sci.* **2020,** *11* (5), 1192-1204.

2. Gomberg, M., AN INSTANCE OF TRIVALENT CARBON: TRIPHENYLMETHYL. *J. Am. Chem. Soc.* **1900,** *22* (11), 757-771.

3. Griller, D.; Ingold, K. U., Persistent carbon-centered radicals. *Acc. Chem. Res.* **1976,** *9* (1), 13-19.

4. Wilcox, D. A.; Agarkar, V.; Mukherjee, S.; Boudouris, B. W., Stable Radical Materials for Energy Applications. *Annu. Rev. Chem. Biomol. Eng.* **2018,** *9* (1), 83-103.

5. Gracia, R.; Mecerreyes, D., Polymers with redox properties: materials for batteries, biosensors and more. *Polym. Chem.* **2013,** *4* (7), 2206.

6. Root, S. E.; Savagatrup, S.; Printz, A. D.; Rodriquez, D.; Lipomi, D. J., Mechanical Properties of Organic Semiconductors for Stretchable, Highly Flexible, and Mechanically Robust Electronics. *Chem. Rev.* **2017,** *117* (9), 6467-6499.

7. Boudouris, B., Engineering optoelectronically active macromolecules for polymer-based photovoltaic and thermoelectric devices. *Curr. Opin. Chem. Eng.* **2013,** *2*, 294–301.

8. Ji, L.; Shi, J.; Wei, J.; Yu, T.; Huang, W., Air-Stable Organic Radicals: New-Generation Materials for Flexible Electronics? *Adv. Mater.* **2020,** *32* (32), 1908015.

9. Berville, M.; Richard, J.; Stolar, M.; Choua, S.; Le Breton, N.; Gourlaouen, C.; Boudon, C.; Ruhlmann, L.; Baumgartner, T.; Wytko, J. A.; Weiss, J., A Highly Stable Organic Radical Cation. *Org. Lett.* **2018,** *20* (24), 8004-8008.

10. Wood, G. P. F.; Moran, D.; Jacob, R.; Radom, L., Bond Dissociation Energies and Radical Stabilization Energies Associated with Model Peptide-Backbone Radicals. *J. Phys. Chem. A* **2005,** *109* (28), 6318-6325.

11. Henry, D. J.; Parkinson, C. J.; Mayer, P. M.; Radom, L., Bond Dissociation Energies and Radical Stabilization Energies Associated with Substituted Methyl Radicals. *J. Phys. Chem. A* **2001,** *105* (27), 6750-6756.

12. Luo, Y.-R., *Comprehensive Handbook of Chemical Bond Energies*. **2007**; pp. 1-1656.

13. Wodrich, M. D.; Mckee, W. C.; Schleyer, P. V. R., On the Advantages of Hydrocarbon Radical Stabilization Energies Based on R−H Bond Dissociation Energies. *J. Org. Chem.* **2011,** *76* (8), 2439-2447.

14. Coote, M. L.; Lin, C. Y.; Zavitsas, A. A., Inherent and transferable stabilization energies of carbon- and heteroatom-centred radicals on the same relative scale and their applications. *Phys. Chem. Chem. Phys.* **2014,** *16* (18), 8686-8696.

15. Hioe, J.; Šakić, D.; Vrček, V.; Zipse, H., The stability of nitrogen-centered radicals. *Org. Biomol. Chem.* **2015,** *13* (1), 157-169.

16. Zipse, H., Radical Stability—A Theoretical Perspective. In *Radicals in Synthesis I*, Springer-Verlag: **2006**; pp. 163-189.

17. Hioe, J.; Zipse. H., *Encyclopedia of Radical in Chemistry, Biology and Materials ,*. John Wiley & Sons Ltd: Chichester, UK, **2012**;pp. 449-477.

18. Coote, M.L.; Lin C. Y.; Zipse, H., ***Carbon-Centered Free Radicals and Radicals Cations***. John Wiley & Sons: **2010**; pp. 83 - 10.

19. Zavitsas, A. A., A Single Universal Scale of Radical Stabilization Energies Does Not Exist: Global Bond Dissociation Energies and Radical Thermochemistries Are Described by Combining Two Universal Scales. *J. Org. Chem.* **2008,** *73* (22), 9022-9026.

20. St. 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* (1).

21. Gobbi, A.; Frenking, G., Resonance Stabilization in Allyl Cation, Radical, and Anion. *J. Am. Chem. Soc.* **1994,** *116* (20), 9275-9286.

22. Kossiakoff, A.; Rice, F. O., Thermal Decomposition of Hydrocarbons, Resonance Stabilization and Isomerization of Free Radicals. *J. Am. Chem. Soc.* **1943,** *65* (4), 590-595.

23. Bader, R. F. W.; Slee, T. S.; Cremer, D.; Kraka, E., Description of conjugation and hyperconjugation in terms of electron distributions. *J. Am. Chem. Soc.* **1983,** *105* (15), 5061-5068.

24. Alabugin, I. V.; Dos Passos Gomes, G.; Abdo, M. A., Hyperconjugation. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2019,** *9* (2), e1389.

25. Mulliken, R. S.; Rieke, C. A.; Brown, W. G., Hyperconjugation*. *J. Am. Chem. Soc.* **1941,** *63* (1), 41-56.

26. Kubo, T., Synthesis, Physical Properties, and Reactivity of Stable, π-Conjugated, Carbon-Centered Radicals. *Molecules* **2019,** *24* (4), 665.

27. Hicks, R. G., What's new in stable radical chemistry? *Org. Biomol. Chem.* **2007,** *5* (9), 1321.

28. Viehe, H. G.; Janousek, Z.; Merenyi, R.; Stella, L., The captodative effect. *Acc. Chem. Res.* **1985,** *18* (5), 148-154.

29. Gryn'Ova, G.; Marshall, D. L.; Blanksby, S. J.; Coote, M. L., Switching radical stability by pH-induced orbital conversion. *Nat. Chem.* **2013,** *5* (6), 474-481.

30. Veciana, J.; Carilla, J.; Miravitlles, C.; Molins, E., Free radicals as clathrate hosts: crystal and molecular structure of 1: 1 perchlorotriphenylmethyl radical–benzene. *J. Chem. Soc., Chem. Commun.* **1987,** (11), 812-814.

31. Degirmenci, I.; Coote, M. L., Effect of Substituents on the Stability of Sulfur-Centered Radicals. *J. Phys. Chem. A* **2016,** *120* (37), 7398-7403.

32. Sevov, C. S.; Hickey, D. P.; Cook, M. E.; Robinson, S. G.; Barnett, S.; Minteer, S. D.; Sigman, M. S.; Sanford, M. S., Physical Organic Approach to Persistent, Cyclable, Low-Potential Electrolytes for Flow Battery Applications. *J. Am. Chem. Soc.* **2017,** *139* (8), 2924-2927.

33. Gallegos, L. C.; Luchini, G.; St. John, P. C.; Kim, S.; Paton, R. S., Importance of Engineered and Learned Molecular Representations in Predicting Organic Reactivity, Selectivity, and Chemical Properties. *Acc. Chem. Res.* **2021,** *54* (4), 827-836.

34. Verloop, A., *Drug Design, Vol.III*. Academic Press:New York, **1976**; Vol .III.

35. Brethomé, A. V.; Fletcher, S. P.; Paton, R. S., Conformational Effects on Physical-Organic Descriptors: The Case of Sterimol Steric Parameters. *ACS Catal.* **2019,** *9* (3), 2313-2323.

36. Hillier, A. C.; Sommer, W. J.; Yong, B. S.; Petersen, J. L.; Cavallo, L.; Nolan, S. P., A Combined Experimental and Theoretical Study Examining the Binding ofN-Heterocyclic Carbenes (NHC) to the Cp*RuCl (Cp* = η5-C5Me5) Moiety: Insight into Stereoelectronic Differences between Unsaturated and Saturated NHC Ligands. *Organometallics* **2003,** *22* (21), 4322-4326.

37. Falivene, L.; Credendino, R.; Poater, A.; Petta, A.; Serra, L.; Oliva, R.; Scarano, V.; Cavallo, L., SambVca 2. A Web Tool for Analyzing Catalytic Pockets with Topographic Steric Maps. *Organometallics* **2016,** *35* (13), 2286-2293.

38. 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 function. *Theor. Chem. Acc.* **2008,** *120* (1-3), 215-241.

39. St. John, P. C.; Guan, Y.; Kim, Y.; Kim, S.; Paton, R. S., Prediction of organic homolytic bond dissociation enthalpies at near chemical accuracy with sub-second computational cost. *Nat. Commun.* **2020,** *11* (1).

40. Landrum, G., RDKit: Open-Source Cheminformatics Software, **2016**. *URL http://www. rdkit. org/, https://github. com/rdkit/rdkit*.

41. Riniker, S.; Landrum, G. A., Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Inf. Model.* **2015,** *55* (12), 2562-2574.

42. 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 Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16 Rev. C.01*, Wallingford, CT, 2016.

43. Marenich, A. V.; Cramer, C. J.; Truhlar, D. G., Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **2009,** *113* (18), 6378-6396.

44. Luchini, G.; Paton, R. S. *DBSTEP: DFT Based Steric Parameters,* 2021, Zenodo DOI: 10.5281/zenodo.4702097 (*https://www.github.com/patonlab/DBSTEP)*

45. Brethomé, A. V.; Paton, R. S.; Fletcher, S. P., Retooling Asymmetric Conjugate Additions for Sterically Demanding Substrates with an Iterative Data-Driven Approach. *ACS Catal.* **2019,** *9* (8), 7179-7187.

46. Janet, J. P.; Ramesh, S.; Duan, C.; Kulik, H. J., Accurate Multiobjective Design in a Space of Millions of Transition Metal Complexes with Neural-Network-Driven Efficient Global Optimization. *ACS Cent. Sci.* **2020,** *6* (4), 513-524.

47. Baxter, A.; Brown, J. A.; Hirs, D.; Humphreys, P.; Jones, K. L.; Patel, V. K. Imidazole derivatives and their use in the treatment of autoimmune or inflammatory diseases or cancers, AU-2017317724-A1. 2019.

48. Yamauchi, N.; Kubota, T.; Nyugaku, T. Alkoxysilane composition, JP-2015212338-A. 2015.

49. Lalevee, J.; Fouassier, J. P., *Encyclopedia of Radical in Chemistry, Biology and Materials*. John Wiley & Sons, Weinheim: **2012**; Vol. 1, pp. 37-57.

50. McCarroll, A. J.; Walton, J. C., Programming Organic Molecules: Design and Management of Organic Syntheses through Free-Radical Cascade Processes. *Angew. Chem*. **2001**, *40*, 2224-2248.

51. Tomida, S.; Doi, T.; Takahashi, T., New approach to the progesterone BCD-ring system by utilizing a tandem transannular radical cyclization. *Tetrahedron Lett.* **1999,** *40* (12), 2363-2366.

52. Morikawa, T.; Nishiwaki, T.; Kobayashi, Y., Radical cyclization to the trifluoromethyl-substituted double bond: Regioselectivity and tandem cyclization. *Tetrahedron Lett.* **1989,** *30* (18), 2407-2410.

53. Bowman, W. R.; Stephenson, P. T.; Young, A. R., Synthesis of nitrogen heterocycles using tandem radical cyclisation of imines. *Tetrahedron Lett.* **1995,** *36* (31), 5623-5626.

54. Beckwith, A. L. J.; Bowry, V. W., Kinetics and regioselectivity of ring opening of substituted cyclopropylmethyl radicals. *J. Org. Chem.* **1989,** *54* (11), 2681-2688.

CHAPTER 4: EXPANSION OF BOND DISSOCIATION PREDICTION WITH MACHINE LEARNING TO MEDICINALLY AND ENVIRONMENTALLY RELEVANT CHEMICAL SPACE

4.1 Chapter overview

In this chapter, we address the limitations of existing machine learning models in the prediction of a bond property, Bond Dissociation Energy (BDE). Existing models are limited to molecules containing element types C, N, O, and H, hence restricting real-world application in the pharmaceutical industry and environmentally relevant molecules. To this end, we aim to develop a robust graph neural network incorporating new element types and varying chemical representations. In the process, we have assembled the largest BDE database known to date with ~500,000 unique bond dissociation reactions. This new model was interactively improved by analyzing chemical space and finally tested on multiple applications corresponding to natural products, perfluoroalkyl substances and regressed against experimentally obtained BDEs. This collaborative research was published in the Royal Society of Chemistry's Chemical Science with contributions from myself, Dr. Yeonjoon Kim, Dr. Seonah Kim, Dr. Peter St. John, and Dr. Robert S. Paton. In this study, I contributed to expanding the BDE database, training the model, applying the newly developed model to pharmaceutical/environmentally relevant compounds, analyzing the chemical space of applicability of the developed models, comparing simple random forest models vs complex neural networks, and finally estimating the BDE of real-world applications. This was done with guidance from my colleague Dr. Yeonjoon Kim and advisors Dr. Seonah Kim, Dr. Peter St. John, and Dr. Robert Paton who also helped write the manuscript for publication.

Reprinted with permission from: **Sowndarya S. V. S.;** Kim, Y.; Kim, S.; St. John, P.; Paton, R. S. Expansion of bond dissociation prediction with machine learning to medicinally and environmentally relevant chemical space. *Digit. Discov.* **2023**. *2*, 1900–1910 with permission from the Royal Society of Chemistry.

4.2 Introduction

The homolytic bond dissociation enthalpy (BDE) quantifies the thermodynamic stabilities of product radical fragments formed by the homolysis of a covalent bond in a reactant molecule. The quantitative evaluation of BDE values provides detailed insight into the thermodynamics of bond-breaking and forming reactions, and can also be related to kinetics and selectivity by using linear free-energy relationships or empirical scaling relations. This fundamental importance has led to the use of BDE (and its free energy counterpart) values across multiple domains in chemistry, ranging from the determination of possible reaction mechanisms to computational mass spectroscopy.[1] For example, BDE values have been used to assess the difference in bond strengths of primary, secondary, and tertiary C-H bonds,[2] to quantify the geometric deformation necessary to reach the transition structures for Pd-catalyzed carbon-halogen insertion,[3] for the prediction of likely molecular fragments observed in the mass spectra of short peptides[4], breakdown of large biomass such as lignin[5], comparing the stability of organic radicals[6] and design of *de novo* radicals for organic redox flow batteries.[7] Due to the broad utility and applicability of BDE values, significant effort has been invested in obtaining quantitative measurements. Experimental techniques such as pyrolysis[8], radical kinetics photoionization mass spectrometry, acidity/electron affinity cycles,[2] and electrochemistry have been used to obtain BDE measurements.[9]

Besides experimental measurements, quantum-mechanical (QM) calculations have become pivotal in assessing and predicting BDE values. Emerging computational methods for the automated enumeration and exploration of reaction mechanisms use estimated BDE values to identify energetically favorable paths among the numerous possibilities.[10] High levels of accuracy are attainable with composite *ab initio* computations of BDEs at 0 K (D_0). For example, the CBS-QB3 method yields mean absolute errors (MAEs) of 0.58 kcal mol^{-1} relative to experimental values for small molecules such as diatomics, hydrocarbons, and hydrides of N, S, Be, Li, and Si.[11,12] However, density functional theory (DFT) calculations are often more practical for larger, conformationally flexible compounds and have been increasingly used to compute BDEs:[13] the M06-2X hybrid meta-GGA functional gives an MAE of 2.1 kcal mol^{-1} relative to experimental hydrocarbon BDE measurements.[14,15] Compared to experimental measurements, which typically give information about the weakest bond(s) in a molecule, computations can be used to study all possible homolytic dissociations. However, due to the exponential scaling of electronic structure calculations with the number of basis functions (and hence molecular size), performing quantum chemical predictions for larger molecules or sizable datasets is challenging, if not intractable. Additionally, a detailed analysis of conformational space may be necessary to identify the most important stationary points on the potential energy surface of both the parent molecule and the two radicals formed upon homolysis. These practical limitations have led to the development of alternative approaches for BDE estimation, such as Quantitative Structure-Property Relationship (QSPR) and Machine Learning (ML) models.[16-18] A machine Learning derived, Fast, Accurate Bond dissociation Enthalpy Tool (ALFABET) achieved an MAE of 0.58 kcal mol^{-1} (vs. an M06-2X/def2-TZVPP oracle) for BDEs of unseen molecules containing C, H, N, O atoms using a message-passing Graph-convolutional Neural

Network (GNN), and has found utility in multiple applications.[14, 19] Rapid and accurate predictions of BDEs have enabled the application of ML to various domains of chemistry including biological metabolism and combustion chemistry. However, these models have been predominantly limited in element scope to second-row elements. This has restricted the application of BDE predictions in medicinal, atmospheric, and environmental domains, where molecules containing multiple larger heteroatoms or halogen atoms are frequently encountered.

Herein, we present an expanded and updated GNN model for BDE prediction. We also consider BDFE (bond dissociation free energy) values, the standard free energy change associated with the dissociation. The inclusion of S, Cl, F, P, Br, and I was motivated by the frequency of these elements alongside C, H, N, and O in approved drugs and enables ML-based BDE predictions to be applied routinely to medicinally and pharmaceutically relevant molecules. We describe the development of a large dataset (*BDE-db2*) containing over 530,000 unique M06-2X/def2-TZVP computed BDE and BDFE values, which underpins this effort. We explore the performance of the model on focused datasets of $C(sp^2)$ and $C(sp^3)$ halogenated molecules relevant to medicinal (polyhalogenated building blocks), atmospheric (halocarbons), and environmental chemistry (per- and polyfluoroalkyl substances, PFAS). Improvements in predictive performance are obtained by analyzing the latent space covered by these datasets and by the addition of a small number of new training samples. The expanded model retains the same levels of accuracy for C, H, N, and O as in previous work while significantly expanding the applicability to new bond types, which are predicted with similarly high levels of accuracy.

4.3 Results and Discussion

4.3.1 Computational BDE Dataset Curation

Efforts to construct generalizable models for BDE prediction of multiple bond types have been greatly enabled by the development of large computational datasets. Aires-de-Sousa and coworkers developed a dataset[17] of computed BDE values for 1000 neutral molecules from the fragment-like subset of the ZINC database.[20,21] Reference bond dissociation energies exclusive of zero-point vibrational energy (ZPE) were generated for 12,834 unique bonds (single and double) between C, H, N, O, and S at the B3LYP/6-311++G(d,p) level of theory. All geometries were optimized with the semi-empirical DFTB3 Hamiltonian. Subsequently, St. John and coworkers developed the BDE-db dataset,[22] taking 42,557 neutral $C_xH_yO_zN_m$ molecules from the PubChem Compound database. This study used an automated fragmentation, conformer generation, and DFT computation workflow to obtain 290,664 unique ZPE-inclusive bond dissociation enthalpies at the M06-2X/def2-TZVP level of theory,[23] which gave the best empirical performance when compared with values from the experimental *iBond* database.[14, 24] Motivated by energy storage and electrolyte applications, Persson and coworkers constructed the BDNCM dataset of 64,312 homolytic and heterolytic bond dissociations for 8,518 neutral and charged molecules containing C, H, O, F, and Li. BDFE values were obtained at the SMD-ωB97X-V/def2-TZVPPD level of theory. Additionally, DiLabio and coworkers have curated a high-quality benchmark dataset including 4502 datapoints of bond separation energies including H, B, C, N, O, F, Si, P, S, and Cl atoms at (RO)CBS-QB3 level of theory.[25]

In this work, we describe one of the most comprehensive quantum chemical bond dissociation datasets, *BDE-db2*, containing 531,244 unique homolytic BDE and BDFE values at

the M06-2X/def2-TZVP level of theory (Fig. 4.1A). 332,035 unique dissociations absent from other datasets have been newly added. We included the ten most common elements in approved pharmaceuticals: C, H, O, N, S, Cl, F, P, Br, and I atoms (in order of their abundance).[26] In addition to compounds originally sourced from PubChem present in BDE-db, we sourced 38,277 additional small molecules (10 heavy atoms or fewer) containing the above heteroatoms from the ZINC15 and PubChem compound libraries. M06-2X/def2-TZVP enthalpies (including the unscaled ZPE) and RRHO Gibbs energies (1 atm, 298K) were computed: the accuracy of this level of theory for halogenated molecules has been benchmarked, showing that hybrid functionals with a high proportion of exact exchange or long-range corrections are more accurate.[27-30] An automated workflow generated the structures of parent and radical fragments from SMILES inputs by enumerating all possible exocyclic single-bond dissociations. Following conformational analysis with *RDKit*, the most stable conformers were optimized with DFT (further details in APPENDIX Section C.1). Structures with imaginary frequencies or having undergone structural rearrangements or fragmentations were removed. Further, recent studies highlighting unphysical and anomalous harmonic vibrations computed for open-shell species (with double-hybrid density functionals)[31] led us to implement an additional filter for dissociations with abnormally large contributions from ΔZPE: 373 further dissociations with statistically significant deviations were removed in this way (APPENDIX Section C.1).

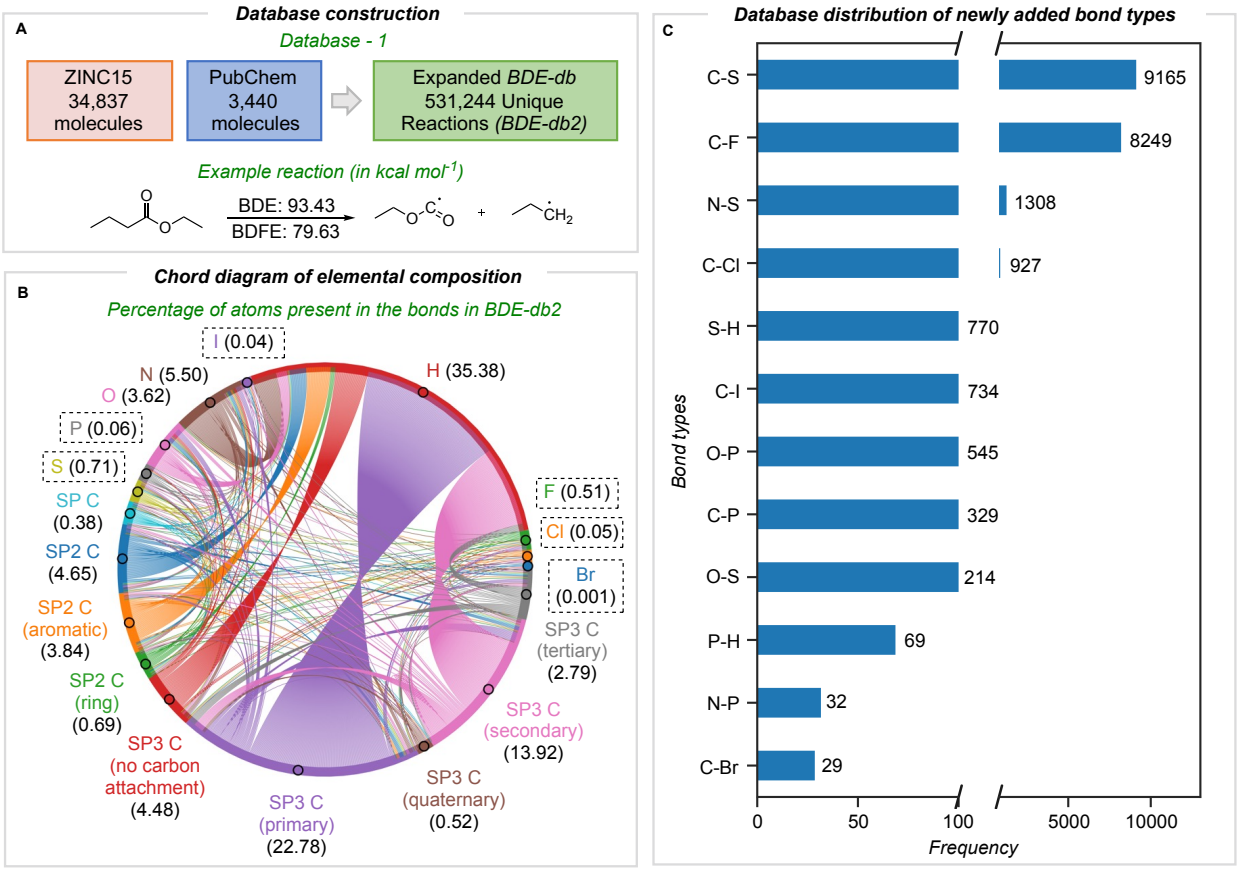


Figure 4.1. *(A) Composition of the BDE-db2 database. (B) Chord diagram representing the bond types present in the database. Link thickness between different segments of the circle reflects the number of bonds between those two atom types. Elemental composition (%)*

The elemental and bond composition of the *BDE-db2* dataset is shown in Fig. 4.1B, which contains 806,433 bond-breaking reactions. After sampling chemical compounds in PubChem and ZINC15 randomly, the elements S, F, P, Cl, I, and Br are present in around 1.4% (22,385) of all bond dissociations alongside C, H, N, and O. The majority (54%) of bonds broken involve at least one carbon atom, with bonds to H the next most populous (35%). Bonds to carbon in all formal hybridization states and degree of substitution are well sampled (Fig. 4.1B). The frequency density of newly added bond types is shown in Fig. 4.1C, with the highest number corresponding to C-S bonds (9165). Following this is C-F bonds with 8249. C-Cl, C-Br, and C-P bonds are around an order of magnitude less frequent than C-F bonds, while C-Br bonds are the

most scarce, with 29. The counts of all bond types in the *BDE-db2* are present in the APPENDIX (Section C.2).

A message passing GNN was then trained to predict the M06-2X/def2-TZVP values of homolytic BDE and BDFE directly from SMILES line notation (Fig. 4.2A).[32] In this approach, molecules input as SMILES are embedded as 2D-molecular graphs using *rdkit*.[33] Nodes (atoms) and edges (bonds) are then assigned to independent classes depending on several features easily obtained from *rdkit*: for atoms, the element, atomic number, formal charge, chiral tag, aromatic state, ring state, degree, and the number of attached H atoms, while for bonds, the pair of bonded elements, formal bond order, and ring state. The embedded vector representations (of length 128) used for atoms and bonds are updated in every message-passing layer of the GNN, utilizing the representations of neighboring bonds and atoms. Benchmarking the number of message-passing layers reveals no gain in accuracy beyond six.[14] During message passing, bond states are updated based on adjacent atoms first, after which atom states are updated systematically (grey box in Fig. 4.2A). Having performed this process six times, atom and bond representations have been encoded with structural information from up to 5/6 bonds away.[14] Bond states from the final message passing layer are reduced to BDE and BDFE predictions by passing them through a linear output layer. Model learning performance was enhanced by utilizing the AdamW optimizer with an inverse time decay schedule for both learning rate (10^{-3}) and weight decay (10^{-5}), and model performance was assessed by measuring the mean absolute error during training for 500 epochs for a batch size of 128 molecules.

Our training and validation set consisted of 514,942 and 8128 unique BDEs, and the performance of the final model was tested on a held-out test set of 1000 molecules comprising

8084 unique dissociations. The MAE on this test set (vs. DFT) is 0.61 and 0.60 kcal mol^{-1} for BDE and BDFE, respectively (Fig. 4.2B). This significantly outperforms chemical descriptor-based approaches, such as that based on an Associative Neural Network (ASNN), giving an MAE of 3.35 kcal mol^{-1} for 887 BDE values involving C, H, O, N, or S. The predictive accuracy is comparable with previous GNN models, ALFABET and BonDNet, with MAE values of 0.58 and 0.50 kcal mol^{-1}, respectively, while encompassing many more bond types. Analysis of the 20 most populous bond types in the held-out test set (APPENDIX Section C.3) shows that C–C and C–H bonds, which are the most frequently encountered, are well predicted (with MAEs of 0.77 kcal mol^{-1} and 0.74 kcal mol^{-1}). Encouragingly, newly added bond types that are less frequently encountered are predicted with only slightly (0.5 – 0.7 kcal mol^{-1}) higher MAE values and all errors fall under 1.7 kcal mol^{-1}. This includes bond types rarely sampled, such as C–Br, P–H, and O–S, where there are tens to hundreds of values in the dataset, in contrast to hundreds of thousands of C–C and C–H bonds. Comparable predictive accuracy is obtained for BDE and BDFE values, which is perhaps unsurprising since these ground truth values are highly correlated.

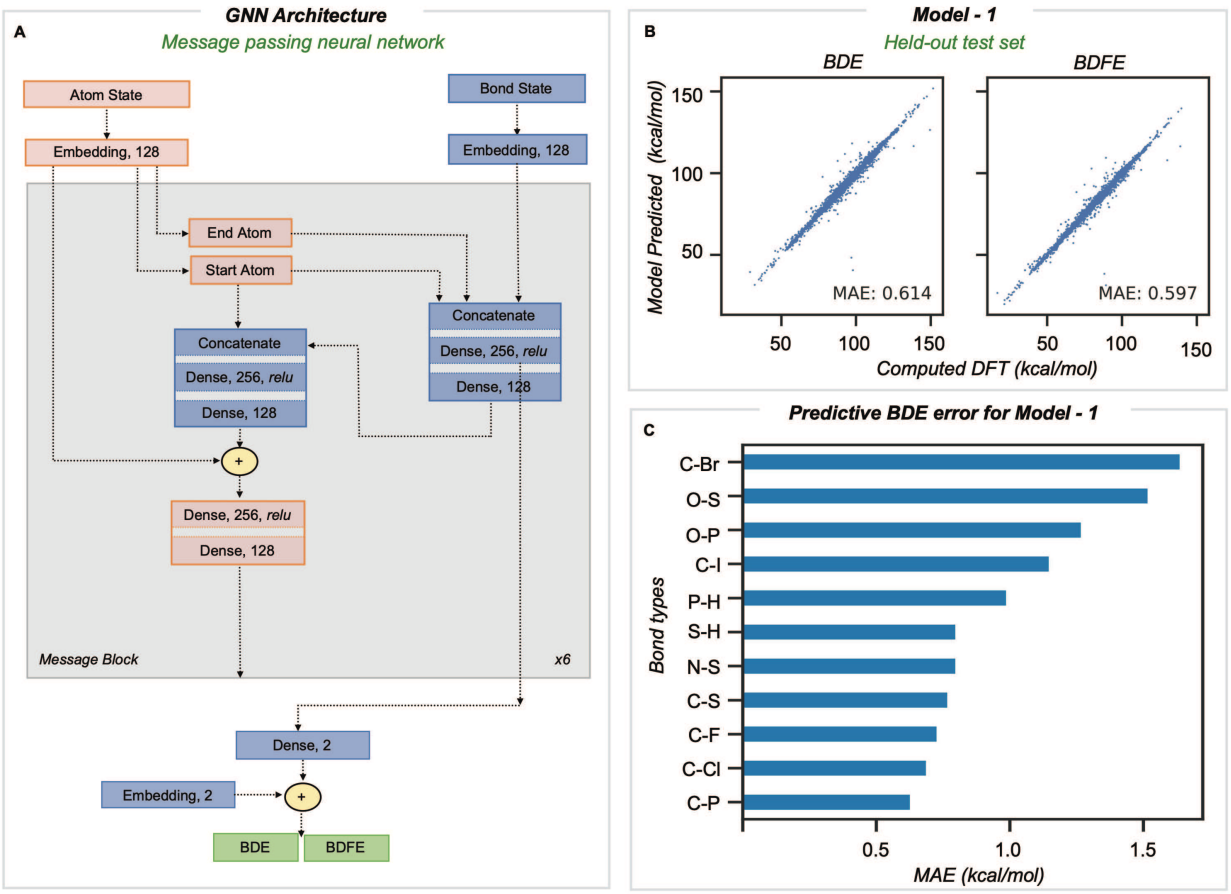


Figure 4.2. *(A) The GNN architecture utilized for the prediction of BDE and BDFE. (B) Predictions for held-out test set (test set – 1). (C) BDE prediction error for a held-out test set based on bond types.*

4.3.2 Application to aryl halide building block compounds

In medicinal chemistry, the modular synthesis of novel drug candidates can be carried out using building blocks, functionalized chemical reagents typically selected for their drug-like properties.[34-37] Aromatic and heteroaromatic fragments are typically functionalized by one or multiple halogen atoms, enabling an array of cross-coupling reactions to be performed. Computed carbon–halogen BDE values have been used to predict the relative rates of oxidative addition by a Pd-catalyst, enabling site-selectivities in the Suzuki cross-couplings of polyhalogenated aromatics to be predicted.[3] The general ability to accurately predict C–X BDE

values for singly- and multiply-halogenated (hetero)aromatics with ML is thus desirable from the perspective of synthesis planning and reaction prediction. We thus focused on commercially available building block libraries for medicinal chemistry developed by Enamine.[38] We tested our newly developed model (herein *model 1*) on halogenated compounds from the Enamine database:[39, 40] a randomly sampled subset of 624 aryl and alkyl halogenated compounds, each with at least one C–F, C–Cl, C–Br or C–I bond, were selected. 64 molecules were common to the original training set and were removed, leaving 560 molecules. These halogenated molecules were fragmented and optimized to generate DFT values for all exocyclic bond dissociations. A total of 6295 BDEs (with 4078 unique BDEs) were collected (*test set 2*) containing 792 C–X bonds (with 696 unique C–X bonds), with a breakdown of 213 C–F, 265 C–Cl, 276 C–Br, and 38 C–I bonds (Fig. 4.3A).

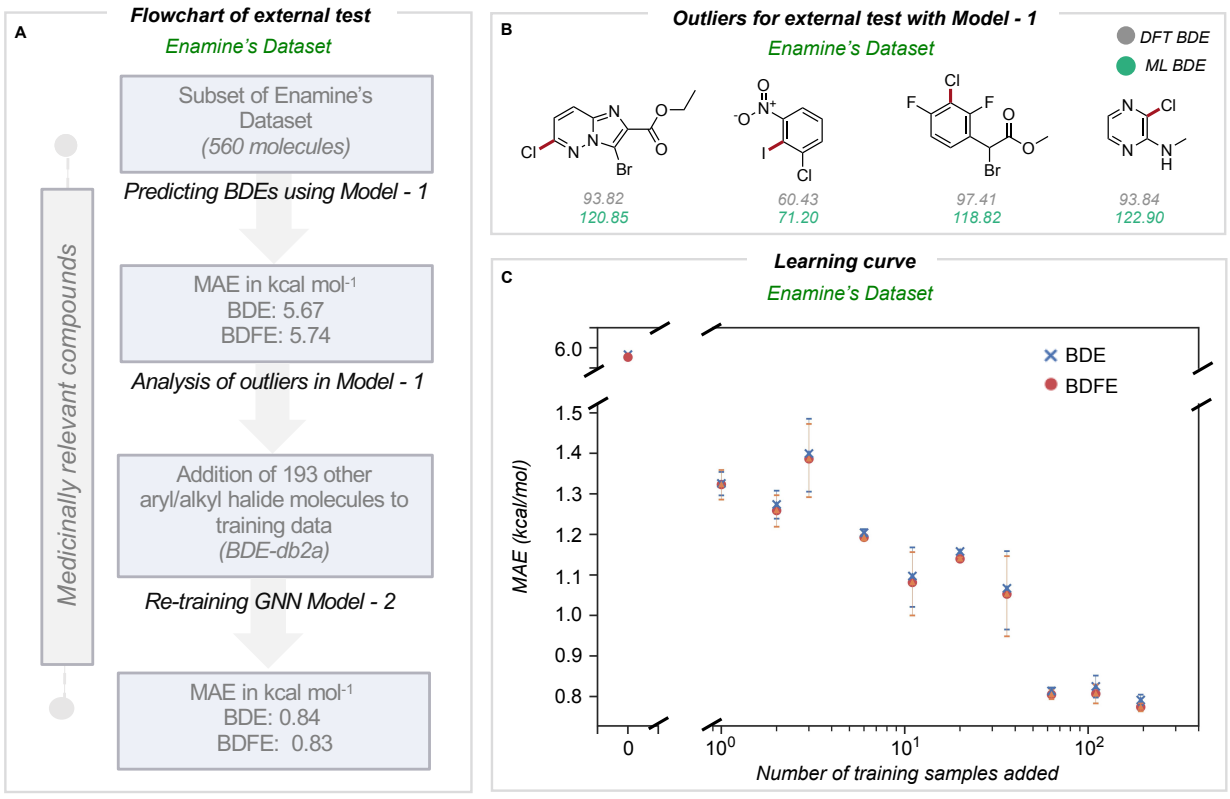


Figure 4.3. *(A) Workflow for model optimization for C(sp^2)–Halogen BDE prediction. (B) Outliers when using model 1 for BDE prediction of Enamine's dataset. (C) GNN learning curve showing systematic improvements in the MAE upon the addition of halogenated molecules to BDE-db2 to obtain BDE-db2a. The error bar corresponds to three different runs.*

BDE and BDFE predictions for these halogenated molecules initially gave MAEs of 5.67 and 5.74 kcal mol^{-1} relative to the DFT oracle. These errors significantly exceed those obtained for the original test set (Fig. 4.2C), primarily due to poor performance for C–Cl and C–I bonds with MAEs of 12 and 8 kcal mol^{-1} (APPENDIX Section C.4). Comparing *test set 2* against molecules in the training database reveals differences in the total number of atoms and the number of halogen and nitrogen atoms. The most pronounced outlier molecules (Fig. 4.3B) contain structural motifs absent from the original training set, such as multiple halogen atoms, and can have errors >20 kcal mol^{-1} (APPENDIX Section C.5). To improve model performance, molecules containing multiple halogens were randomly sampled and added to the training dataset (*BDE-db2a*). These additional molecules correspond to a distinct subset from the total enamine database: we used 193 molecules with 1634 unique BDEs, 413 of which correspond to C–X bonds (139 C–F, 141 C–Cl, 119 C–Br and 14 C–I). This corresponds to an increase in training set size by a modest 0.3%. This expanded dataset was used to train a new GNN (*model* 2) whose architecture is the same as Fig. 4.2A. Upon testing on the Enamine dataset, model performance is considerably improved, giving MAEs of 0.84 and 0.83 kcal mol^{-1} for BDE and BDFE values, respectively without degraded performance on the original *test set 1* (0.64 and 0.62 kcal mol^{-1}, APPENDIX Section C.6). To determine how adding new structures to the training data enhances model performance, we experimented by adding different numbers of randomly sampled halogenated structures (Fig. 4.3C). These runs were performed in triplicate. Surprisingly, the addition of fewer than ten additional structures reduces MAE values to around

~ 1kcal mol $^{-1}$, while continuous improvement is observed as more structures are added to reach a limiting accuracy at around 100 additional structures. This behavior suggests that model performance for fairly broad areas of chemical space, such as poly-halogenated heterocycles, can be improved by adding a relatively small but targeted number of compounds to the training process.

4.3.3 Application to environmentally relevant compounds

We next assessed the model's predictive accuracy for polyhaloalkyl compounds, such as environmentally relevant chlorofluorocarbons containing several C(sp^3)–X bonds. A dataset of 40 molecules (*test set 3*) containing 274 bond dissociations (155 unique bond dissociations) was curated from PubChem, following which systematic fragmentation and DFT optimizations were performed. One molecule was common to the original training set and was removed. The total number of C–X bonds is 212 (104 unique C–X bonds), with a breakdown of 123 C–F, 85 C–Cl, and 4 C–Br bonds (APPENDIX Section C.7). Applying the improved *model 2* on this dataset led to MAEs of 2.69 and 2.86 kcal mol^{-1} for BDE and BDFE values (parity plots are shown in APPENDIX Section C.8).

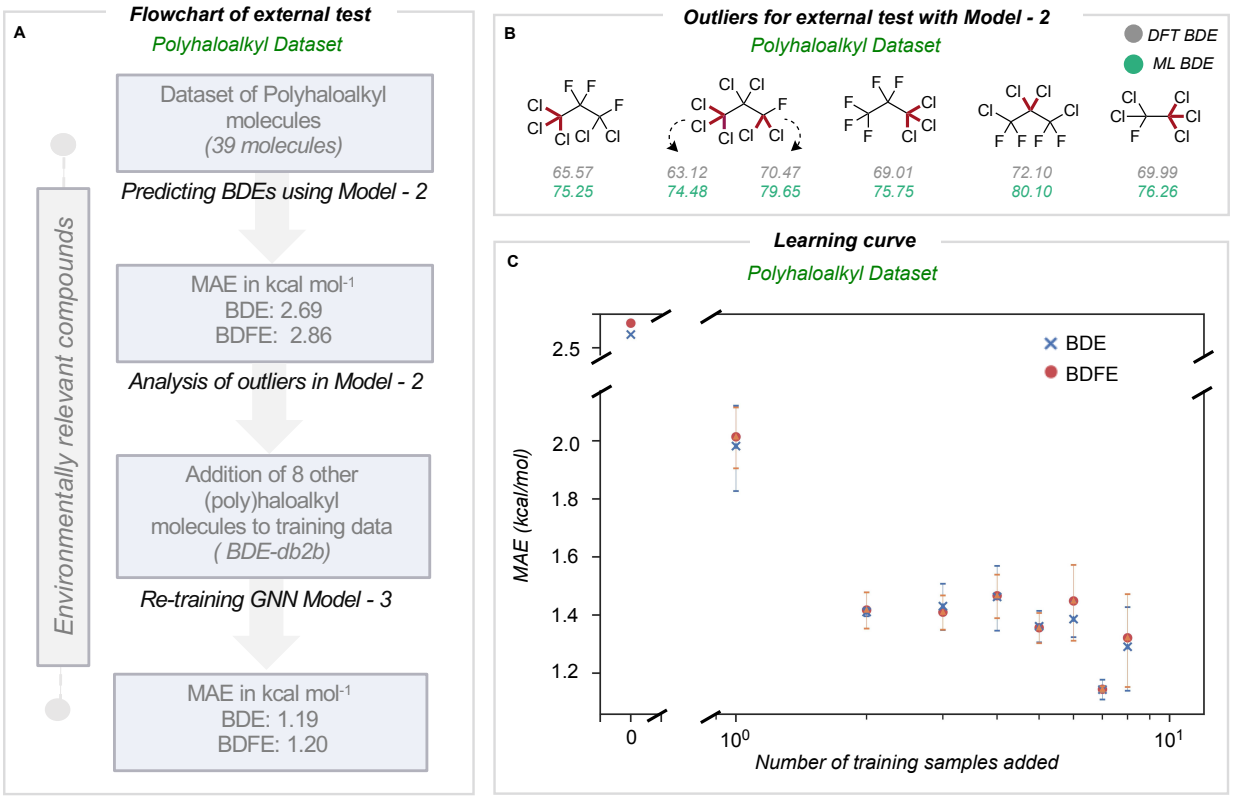


Figure 4.4. *(A) Workflow for model optimization for C(sp^3)–Halogen BDE prediction. (B) Outliers when using model 2 for BDE prediction of polyhaloalkyl compounds. (C) GNN learning curve showing systematic improvements in the MAE upon the addition of halogenated molecules to BDE-db2a to obtain BDE-db2b. The error bar corresponds to three different runs.*

For this dataset, we found relatively large errors for the weakest bonds (BDE values under 80 kcal mol^{-1}). These correspond to dissociation at multiply halogenated carbons, which yield resonance-stabilized radicals and hence smaller BDE values (Fig. 4.4B). Looking to see if similar molecules existed in our training database from *BDE-db2a*, we found that only 3.4% of molecules (2,102 of 61,630 molecules) have multiple halogens on the same atom. In comparison, 0.2% (1,407 of 516,570 unique bonds) of the bond-breaking reactions had at least one fragment with >1 halogen atom on the radical atom. Based on our earlier observations, we hypothesized that adding a relatively small number of compounds bearing multiply-halogenated carbon atoms to training data could considerably improve predictive accuracy for this dataset.

Eight molecules were constructed and added to the training dataset to build *BDE-db2b* (APPENDIX Section C.9) and develop a new GNN *model 3*. This new model showed a notable reduction in MAE values to 1.19 and 1.20 kcal mol^{-1} (Table 4.1). Additionally, the learning curve depicts how the mean MAE values over three different types of additions vary when each molecule is added (Fig. 4.4C). Prediction accuracies with this model are maintained for prior test sets (APPENDIX Section C.10). The results of the successive improvements made to the BDE prediction model are summarized in Table 4.1.

Table 4.1. *BDE and BDFE prediction accuracy (MAE in kcal mol^{-1}) obtained from GNN following test-train cycles.*

	Model 1		*Model 2*		*Model 3*	
	BDE	BDFE	BDE	BDFE	BDE	BDFE
General test set (n = 1000)	0.61	0.60	0.64	0.62	0.64	0.61
Haloheterocycle set (n = 560)	5.67	5.74	0.84	0.83	0.74	0.70
Polyhaloalkyl set (n = 39)	2.78	2.74	2.69	2.86	1.19	1.20

4.3.4 Chemical space and neighbor analysis

To understand the relationship between prediction accuracy and training set composition, we visualized the representations of bonds learned by the final model (Fig. 4.5A). Since the bond states have a dimension of 128, we performed dimensionality reduction with the t-SNE method to project 8924 different C–X bonds in two dimensions. The newly added C–X bonds used to build *BDE-db2a* and *BDE-db2b* are spread in chemical space and cover regions not present in the original dataset. Further, we studied specific examples of outlier predictions that were improved by successive generations of our model (Fig. 4.5B). For these bonds, we found the ten nearest neighbors (in the model's 128-dimensional latent space) from the training dataset and computed the mean distance of these nearest neighbors. In each case, we found that poor

predictions result where the closest training bonds are highly chemically dissimilar to the query bond. Previous work has also shown that determining the distance in latent space enables the identification of high and low confidence points.[41] Overall, the systematic improvement in performance can be attributed to the incorporation of new regions of chemical space in the training set containing more diverse structural features.

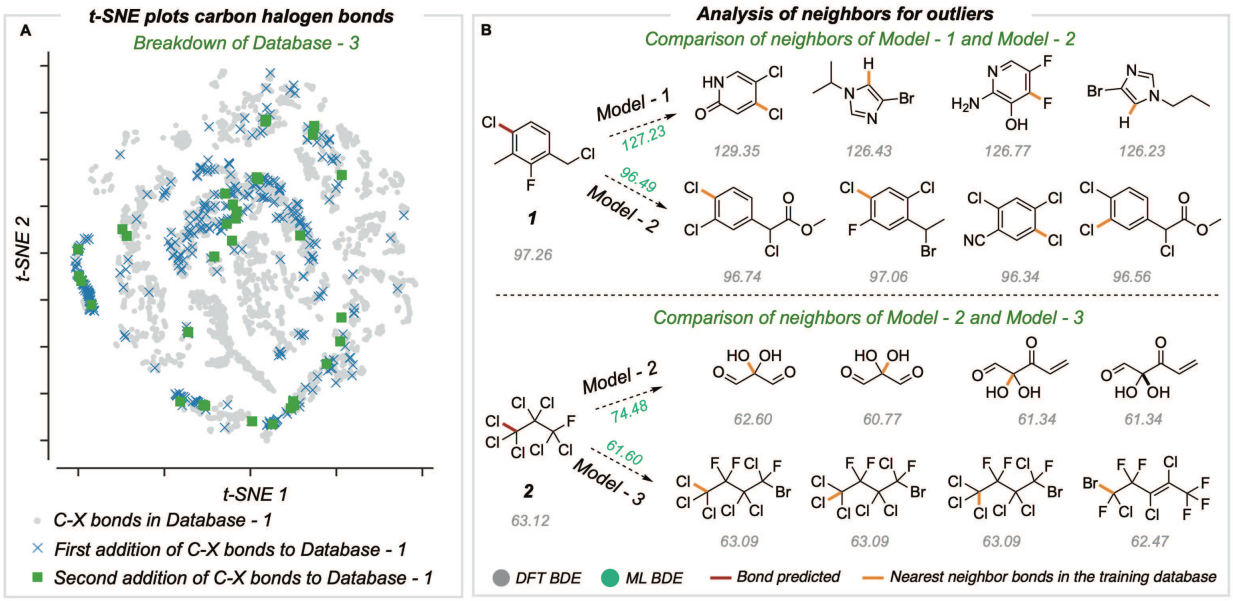


Figure 4.5. *(A) t-SNE plot showing a reduced dimensionality projection of embeddings representing the final bond states used for BDE prediction for C-F, C-Cl, C-Br, and C-I bonds. (B) Analysis of neighbors of one outlier in Enamine's dataset and polyhaloalkyl datasets.*

4.3.5 Comparison of current GNN models with traditional cheminformatics features and QM features

We compared the performance of end-to-end learned representations, as in our GNN model, against cheminformatics-based features such as circular atomic fingerprints. We also evaluated whether our GNN could be improved by including additional features, such as DFT-optimized bond lengths (Fig. 4.6).[42] We studied these models' learning by systematically increasing the number of training samples: while the number of samples was randomly selected for three

different runs, they were kept consistent across the different models. For GNN Model 3, the MAE decreases with training dataset size, achieving kcal/mol accuracy at around ~5000 samples. In contrast, a Random Forest (RF) based model using Morgan fingerprints generated for the bonded pair of atoms (radius of 3 encoded as 512 bits) demonstrates slower learning, with a BDE error of 2.2 kcal mol^{-1} for the same training set size. Ultimately, achieving chemical accuracy with this model would require training on a dataset more than an order of magnitude larger than that used to train the GNN. Including optimized bond lengths as part of the initial embedding before GNN model training led to an improvement in performance only for very small dataset sizes (<100), while in the limit of larger datasets, there was a negligible reduction in MAE values. This result suggests that representation learning occurs efficiently for this problem, with datasets on the scale of BDE-db2 containing tens of thousands of training examples.

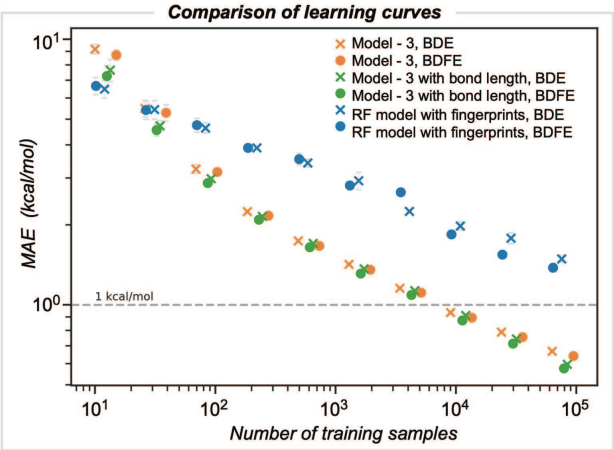


Figure 4.6. *Model learning behavior, comparing the original GNN model (orange), a GNN model with features augmented by DFT-optimized bond lengths (green), and a Random Forest regression using Morgan fingerprints (blue).*

4.3.6 Validation against computed and experimental BDE values for diverse halogenated compounds

The newly developed model was validated using three external datasets: halogenated natural products and environmentally relevant polyfluoroalkyl substances (PFAS), for which we computed reference BDE values at the M06-2X/def2-TZVP level of theory,[43] and an experimental dataset of BDE values for aliphatic chlorofluorocarbons.

Four natural products containing F, Cl, Br, and I were identified (Fig. 4.7A): Nucleocidin,[44] Salinosporamide A,[45] Tyrian purple,[46] and Levothyroxine.[47] All of these natural products were fragmented and optimized with a similar method as the training database. The average number of heavy atoms is 23, more than twice the size of those in the training database. Across all four molecules, an MAE of 1.76 and 1.75 kcal mol^{-1} is obtained for BDE and BDFE for 65 unique bonds. Using Model 3 for predictions is orders of magnitude faster (~ seconds) than the quantum chemical reference values (~ 2 days/molecule of CPU time). The predicted BDE values for each molecule have an R^2 equal to or higher than 0.9 and a mean absolute error under 2 kcal mol^{-1} (Fig. 4.7A).

The second validation set comprises experimentally reported bond dissociation enthalpies of 40 fluorinated and chlorinated alkanes, including chlorofluorocarbons (CFCs) and hydrochlorofluorocarbons (HFCs), atmospheric trace gases that influence stratospheric ozone, climate, and air quality.[43, 48] Primary mechanisms of atmospheric degradation, such as photolysis, are influenced by the C–halogen bond strength, while the reactions of HFCs with hydroxyl radicals are influenced by C–H bond strengths, and so their prediction is of practical importance.

We removed 7 molecules from this dataset present in our original training database to avoid data leakage. ML-predicted bond dissociation enthalpies for the remaining compounds compare well with experimental values, with a correlation coefficient of 0.96 and an MAE of 2.12 kcal mol^{-1} for 69 unique bonds (23 C–C, 20 C–H, 10 C–F, and 16 C–Cl bonds) (Fig. 4.7B). This is encouraging since no experimental data was used to train the model, and the largest error obtained for this test set is ~ 3 kcal mol^{-1}.

Finally, we test the model's predictive power in determining the weakest homolytic bond dissociation in per- and polyfluoroalkyl substances (PFAS) containing ether, alcohol, amide, and sulfonamide functional groups. Thermodynamic BDE values have been related to breakdown pathways during combustion,[14] while the mechanisms of PFAS degradation have been linked to BDE values.[49, 50] We studied 57 molecules, with 557 C–F bonds and a total of 697 unique bonds overall. On comparing different bond types, C–S and C–C bonds have lower BDE values: we would expect these bonds to undergo homolysis first during pyrolytic decomposition (Fig. 4.7C). The accuracy of prediction against DFT is 1.15 and 1.51 kcal mol^{-1} for BDE and BDFE respectively. Previous data-driven models have focused on the prediction of C–F BDE values in PFAS molecules;[49] however, with the ability to predict across the breadth of bond types present in these compounds, we observe that C–S and C–C bonds (rather than C–F) are thermodynamically much more likely candidates for the primary site of homolysis. For 60% of the PFAS considered, the ML-predicted weakest bond matches DFT, while for the remaining 40% of cases, the weakest bond (from DFT) lies within 4 kcal mol^{-1} from the ML-minimum energy. This suggests one possible use of ML could be to quickly survey and rank possible homolytic cleavages, returning a focused set of candidate bonds to be investigated in greater depth with QM calculations.

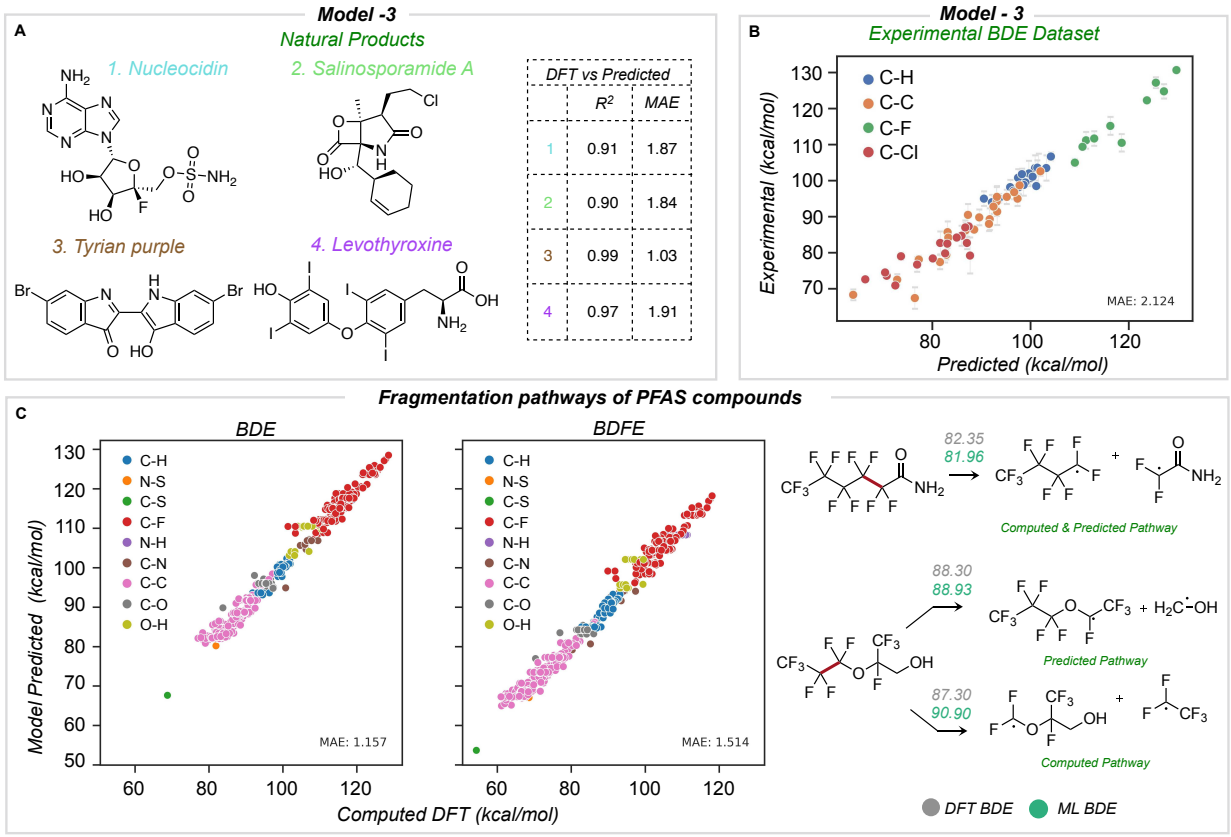


Figure 4.7. *(A) BDE prediction of halogenated natural products: R^2 and MAE relative to DFT ground truth for each molecule shown. (B) Comparison of predicted and experimental BDE values for chlorofluorocarbons (CFCs) and hydrochlorofluorocarbons (HFCs). (C) Predicted BDE and BDFE values for PFAS relative to DFT ground truth. Homolysis of the weakest bond as a potential mechanism of PFAS thermal decomposition. All values in kcal/mol.*

4.4 Conclusion

Bond dissociation enthalpies and free energies are fundamental quantities used to assess reaction thermodynamics. BDE values also influence reaction kinetics and are often used as essential ingredients to understand mechanism and selectivity. We have developed a broadly applicable BDE prediction tool based on a graph neural network that yields quantitative predictions close to DFT values across a range of organic molecules containing heteroatoms. This tool enables a broader range of chemical space to be studied by this approach than was previously possible, which now includes aromatic and aliphatic compounds with multiple halogen atoms relevant to

medicinal and atmospheric applications. For multiply halogenated chlorofluorocarbons, this approach yields results within 2 kcal/mol of experimental BDE values. For training dataset sizes on the order of thousands or tens of thousands of compounds, we observe that the learned embeddings of the GNN are not improved by the addition of additional QM descriptors and that the model learning performance exceeds a more traditional cheminformatics approach using fixed circular fingerprints by more than an order of magnitude. While the requirement for training datasets containing thousands of compounds qualifies as a data-hungry approach, we observed that successive expansion of the model's domain of applicability to encompass new bond types was possible through the addition of relatively small (i.e., fewer than hundreds) targeted compound libraries to the training data. We suggest that this may indicate some level of model generalization according to molecular substitution patterns around the site of dissociation, such that only a few examples of new bond types are required. This suggests that relatively small, focused datasets can be used to continually expand the scope of this, and other GNN-based models for property predictions.

REFERENCES

1. Hill, A. W.; Mortishire-Smith, R. J., Automated assignment of high-resolution collisionally activated dissociation mass spectra using a systematic bond disconnection approach. *Rapid Commun. Mass Spectrom.* **2005,** *19*, 3111-3118.

2. Blanksby, S. J.; Ellison, G. B., Bond Dissociation Energies of Organic Molecules. *Acc. Chem. Res.* **2003,** *36*, 255-263.

3. Garcia, Y.; Schoenebeck, F.; Legault, C. Y.; Merlic, C. A.; Houk, K. N., Theoretical Bond Dissociation Energies of Halo-Heterocycles: Trends and Relationships to Regioselectivity in Palladium-Catalyzed Cross-Coupling Reactions. *J. Am. Chem. Soc.* **2009,** *131*, 6632-6639.

4. Obolensky, O. I.; Wu, W. W.; Shen, R.-F.; Yu, Y.-K., Using dissociation energies to predict observability of b- and y-peaks in mass spectra of short peptides. *Rapid Commun. Mass Spectrom.* **2012,** *26*, 915-920.

5. Kim, S.; Chmely, S. C.; Nimlos, M. R.; Bomble, Y. J.; Foust, T. D.; Paton, R. S.; Beckham, G. T., Computational Study of Bond Dissociation Enthalpies for a Large Range of Native and Modified Lignins. *J. Phys. Chem. Lett.* **2011,** *2*, 2846-2852.

6. Sowndarya S. V, S.; St. John, P. C.; Paton, R. S., A quantitative metric for organic radical stability and persistence using thermodynamic and kinetic features. *Chem. Sci.* **2021,** *12*, 13158-13166.

7. S. V, S. S.; Law, J. N.; Tripp, C. E.; Duplyakin, D.; Skordilis, E.; Biagioni, D.; Paton, R. S.; St. John, P. C., Multi-objective goal-directed optimization of de novo stable organic radicals for aqueous redox flow batteries. *Nat. Mach. Intell.* **2022,** *4*, 720-730.

8. Szwarc, M., The Determination of Bond Dissociation Energies by Pyrolytic Methods. *Chem. Rev.* **1950,** *47*, 75-173.

9. Fu, Y.; Liu, L.; Wang, Y.-M.; Li, J.-N.; Yu, T.-Q.; Guo, Q.-X., Quantum-Chemical Predictions of Redox Potentials of Organic Anions in Dimethyl Sulfoxide and Reevaluation of Bond Dissociation Enthalpies Measured by the Electrochemical Methods. *J. Phys. Chem. A* **2006,** *110*, 5874-5886.

10. Koerstz, M.; Rasmussen, M. H.; Jensen, J. H., Fast and automated identification of reactions with low barriers: the decomposition of 3-hydroperoxypropanal. *SciPost Chem.* **2021,** *1*, 003.

11. Zhao, Y.; Truhlar, D. G., How Well Can New-Generation Density Functionals Describe the Energetics of Bond-Dissociation Reactions Producing Radicals? *J. Phys. Chem. A* **2008,** *112*, 1095-1099.

12. Montgomery, J. A.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A., A complete basis set model chemistry. VI. Use of density functional geometries and frequencies. *J. Chem. Phys.* **1999,** *110*, 2822-2827.

13. Mardirossian, N.; Head-Gordon, M., Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. *Mol. Phys.* **2017,** *115*, 2315-2372.

14. St. John, P. C.; Guan, Y.; Kim, Y.; Kim, S.; Paton, R. S., Prediction of organic homolytic bond dissociation enthalpies at near chemical accuracy with sub-second computational cost. *Nat. Commun.* **2020,** *11*.

15. Trung, N. Q.; Mechler, A.; Hoa, N. T.; Vo, Q. V., Calculating bond dissociation energies of X−H (X=C, N, O, S) bonds of aromatic systems via density functional theory: a detailed comparison of methods. *Royal Soc. Open Sci*. **2022,** *9*, 220177.

16. Yu, H.; Wang, Y.; Wang, X.; Zhang, J.; Ye, S.; Huang, Y.; Luo, Y.; Sharman, E.; Chen, S.; Jiang, J., Using Machine Learning to Predict the Dissociation Energy of Organic Carbonyls. *J. Phys. Chem. A* **2020,** *124*, 3844-3850.

17. Qu, X.; Latino, D. A.; Aires-De-Sousa, J., A big data approach to the ultra-fast prediction of DFT-calculated bond energies. *J. Cheminform.* **2013,** *5*, 34.

18. Xue, C. X.; Zhang, R. S.; Liu, H. X.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T., An Accurate QSPR Study of O−H Bond Dissociation Energy in Substituted Phenols Based on Support Vector Machines. *J. Chem. Inf. Comput. Sci.* **2004,** *44*, 669-677.

19. Zulueta, B.; Tulyani, S. V.; Westmoreland, P. R.; Frisch, M. J.; Petersson, E. J.; Petersson, G. A.; Keith, J. A., A Bond-Energy/Bond-Order and Populations Relationship. *J. Chem. Theory Comput.* **2022,** *18*, 4774-4794.

20. Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G., ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model* **2012,** *52*, 1757-1768.

21. Carr, R. A. E.; Congreve, M.; Murray, C. W.; Rees, D. C., Fragment-based lead discovery: leads by design. *Drug Discov. Today* **2005,** *10*, 987-992.

22. St. John, P.; Guan, Y.; Kim, Y.; Kim, S.; Paton, R. S., BDE-db: A collection of 290,664 Homolytic Bond Dissociation Enthalpies for Small Organic Molecules. figshare. *figshare* , *https://doi.org/10.6084/m9.figshare.10248932.v1* **2019**.

23. St. 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*.

24. Internet Bond-energy Databank (pKa and BDE)—iBonD Home Page. http://ibond.nankai.edu.cn/ 2022.

25. Prasad, V. K.; Khalilian, M. H.; Otero-de-la-Roza, A.; DiLabio, G. A., BSE49, a diverse, high-quality benchmark dataset of separation energies of chemical bonds. *Sci. Data* **2021,** *8*, 300.

26. Smith, B. R.; Eastman, C. M.; Njardarson, J. T., Beyond C, H, O, and N! Analysis of the Elemental Composition of U.S. FDA Approved Drug Architectures. *J. Med. Chem.* **2014,** *57*, 9764-9773.

27. Kozuch, S.; Martin, J. M. L., Halogen Bonds: Benchmarks and Theoretical Analysis. *J. Chem. Theory Comput.* **2013,** *9*, 1918-1931.

28. Forni, A.; Pieraccini, S.; Rendine, S.; Sironi, M., Halogen bonds with benzene: An assessment of DFT functionals. *J. Comput. Chem.* **2014,** *35*, 386-394.

29. Siiskonen, A.; Priimagi, A., Benchmarking DFT methods with small basis sets for the calculation of halogen-bond strengths. *J. Mol. Model.* **2017,** *23*.

30. Xu, S.; Wang, Q.-D.; Sun, M.-M.; Yin, G.; Liang, J., Benchmark calculations for bond dissociation energies and enthalpy of formation of chlorinated and brominated polycyclic aromatic hydrocarbons. *RSC Adv.* **2021,** *11*, 29690-29701.

31. Simmie, J. M.; Somers, K. P., Snakes on the Rungs of Jacob's Ladder: Anomalous Vibrational Spectra from Double-Hybrid DFT Methods. *J. Phys. Chem. A* **2020,** *124*, 6899-6902.

32. St. John, P. C.; Guan, Y.; Kim, Y.; Kim, S.; Paton, R. S. Prediction of Organic Homolytic Bond Dissociation Enthalpies at near Chemical Accuracy with Sub-Second Computational Cost. *Nat. Commun.* **2020**, *11 (1)*, 1–12.

33. Riniker, S.; Landrum, G. A., Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Inf. Model* **2015,** *55*, 2562-2574.

34. Zabolotna, Y.; Volochnyuk, D. M.; Ryabukhin, S. V.; Horvath, D.; Gavrilenko, K. S.; Marcou, G.; Moroz, Y. S.; Oksiuta, O.; Varnek, A., A Close-up Look at the Chemical Space of Commercially Available Building Blocks for Medicinal Chemistry. *J. Chem. Inf. Model* **2022,** *62*, 2171-2185.

35. Grygorenko, O. O.; Volochnyuk, D. M.; Vashchenko, B. V., Emerging Building Blocks for Medicinal Chemistry: Recent Synthetic Advances. *Eur. J. Org. Chem.* **2021,** *2021*, 6478-6510.

36. Kalliokoski, T., Price-Focused Analysis of Commercially Available Building Blocks for Combinatorial Library Synthesis. *ACS Comb. Sci.* **2015,** *17*, 600-607.

37. Helal, C. J.; Bundesmann, M.; Hammond, S.; Holmstrom, M.; Klug-McLeod, J.; Lefker, B. A.; McLeod, D.; Subramanyam, C.; Zakaryants, O.; Sakata, S., Quick Building Blocks (QBB): An Innovative and Efficient Business Model To Speed Medicinal Chemistry Analog Synthesis. *ACS Med. Chem. Lett.* **2019,** *10*, 1104-1109.

38. Zabolotna, Y.; Volochnyuk, D. M.; Ryabukhin, S. V.; Horvath, D.; Gavrilenko, K. S.; Marcou, G.; Moroz, Y. S.; Oksiuta, O.; Varnek, A., A Close-up Look at the Chemical Space of Commercially Available Building Blocks for Medicinal Chemistry. *J. Chem. Inf. Model* **2021**.

39. Enamine Functional Classes: Alkyl Halides. Available at: https://enamine.net/building-blocks/functional-classes/alkyl-halides. **2020**.

40. Enamine Functional Classes: Aryl Halides. Available at: https://enamine.net/building-blocks/functional-classes/aryl-halides. **2020**.

41. Janet, J. P.; Duan, C.; Yang, T.; Nandy, A.; Kulik, H. J., A quantitative uncertainty metric controls error in neural network-driven chemical discovery. *Chem. Sci.* **2019,** *10*, 7913-7922.

42. (a) Gallegos, L. C.; Luchini, G.; St. John, P. C.; Kim, S.; Paton, R. S., Importance of Engineered and Learned Molecular Representations in Predicting Organic Reactivity, Selectivity, and Chemical Properties. *Acc. Chem. Res.* **2021,** *54*, 827-836; (b) This would of course not be a practical approach for rapid BDE prediction; rather, we performed this analysis to see whether the learned embeddings were optimal or could be further improved through the addition of expensive DFT-level features.

43. Shi, J.; He, J.; Wang, H.-J., A computational study of C–X (X = H, C, F, Cl) bond dissociation enthalpies (BDEs) in polyhalogenated methanes and ethanes. *J. Phys. Org. Chem.* **2011,** *24*, 65-73.

44. Carvalho, M. F.; Oliveira, R. S., Natural production of fluorinated compounds and biotechnological prospects of the fluorinase enzyme. *Crit. Rev. Biotechnol.* **2017,** *37*, 880-897.

45. Zeng, J.; Zhan, J., Chlorinated Natural Products and Related Halogenases. *Isr. J. Chem.* **2019,** *59*, 387-402.

46. Cooksey, C., Tyrian Purple: 6,6'-Dibromoindigo and Related Compounds. *Molecules* **2001,** *6*, 736-769.

47. Bianco, A. C.; Salvatore, D.; Gereben, B. Z.; Berry, M. J.; Larsen, P. R., Biochemistry, Cellular and Molecular Biology, and Physiological Roles of the Iodothyronine Selenodeiodinases. *Endocr. Rev.* **2002,** *23*, 38-89.

48. Luo, Y. R., *Bond disassociation energies, CRC handbook of Chemistry and Physics*. CRC press, Boca Raton: **2002**.

49. Raza, A.; Bardhan, S.; Xu, L.; Yamijala, S. S. R. K. C.; Lian, C.; Kwon, H.; Wong, B. M., A Machine Learning Approach for Predicting Defluorination of Per- and Polyfluoroalkyl Substances (PFAS) for Their Efficient Treatment and Removal. *Environ. Sci. Technol. Lett.* **2019,** *6*, 624-629.

50. Kurniawan, D.; Arai, H.; Morita, S.; Kitagawa, K., Chemical degradation of Nafion ionomer at a catalyst interface of polymer electrolyte fuel cell by hydrogen and oxygen feeding in the anode. *Microchem. J.* **2013,** *106*, 384-388.

CHAPTER 5: MULTI-OBJECTIVE GOAL-DIRECTED OPTIMIZATION OF DE NOVO STABLE ORGANIC RADICALS FOR AQUEOUS REDOX FLOW BATTERIES

5.1 Chapter overview

In this chapter, we combine the newly developed machine learning models to perform goal-directed molecular optimization of organic radicals for utility in aqueous redox flow batteries. We utilize Google Deep Mind's AlphaZero algorithm to search the vast chemical space. Upon successfully identifying 32 new radical candidates de novo, a systematic analysis of them was done for stability, redox potentials, and synthesizability. This collaborative research was published in the Nature Portfolio under Nature Machine Intelligence with contributions from myself, Jeffrey N. Law, Charles E. Tripp, Dmitry Duplyakin, Erotokritos Skordilis, David Biagioni, Robert S. Paton, and Peter C. St. John. In this study, I contributed to performing the benchmarks of redox potentials against experimental data, building databases of ~80,000 for organic radical's reduction and oxidation potentials, modeling the radical properties using graph neural networks, and finally analyzing the generated radicals from the AlphaZero algorithm in terms of successful/failed candidates, synthesizability and identifying chemical properties learned by the algorithm itself. I also helped write the manuscript for publication. Our work on goal-directed molecular optimization was highlighted in *Nature Chemistry Reviews*, *Nature Machine Intelligence News*, and featured in *Tech Xplore*.

Reprinted with permission from: **Sowndarya S. V. S.;** Law, J.N., Tripp, C.E. *et al.,* Multi-objective goal-directed optimization of de novo stable organic radicals for aqueous redox flow batteries. *Nat. Mach. Intell*., **2022**, **4**, 720–730.

5.2 Introduction

The development of materials with precisely tuned electrochemical and physical properties is critical in enabling next-generation energy technologies. One example appears in redox flow batteries (RFBs), which offer the potential to deliver low-cost and reliable energy storage at the grid-scale.[1] While vanadium-based RFB chemistries are currently the most well-studied, battery formulations using organic molecules as the active species are a promising alternative as they are domestically manufacturable, decoupled from markets for transition metals, and have a lesser ecological footprint.[2–4] The stability window of the solvent determines the desired electrochemical potential for both anode and cathode half-reactions: for water-based batteries, this dictates a maximum thermodynamically stable voltage of 1.23 V at 25°C, although in practice voltages of approximately 2.0 V are possible due to the slow kinetics of hydrogen evolution on carbon electrodes.[5]

A wide range of organic redox couples exist and have been explored as charge carriers in flow battery applications.[6] Among these, persistent organic radicals are a promising class of active species with highly reversible redox processes.[7] These molecules have an unpaired valence electron that can either be donated or paired with an accepted electron to form a closed-shell species. However, due partly to their unique and complex chemistry, relatively few stable radical-containing materials are known to exist.[7,8] As a result, most studies have focused on chemical modifications of a handful of well-known stable radical scaffolds,[9] primarily via mechanism-based approaches that identify optimal side-chains to improve performance: i.e., increasing solubility, limiting possible decomposition reactions, or tuning redox potential.[10–18] The scarcity of radical scaffolds complicates the tuning of their physical and electrochemical properties to meet the strict demands of high-performance, low-cost RFBs.[2,3] For example,

TEMPO (2,2,6,6-tetramethylpiperidine-N-oxyl) is currently a leading organic catholyte candidate due to its persistence and ability to undergo reversible one-election oxidation (Fig. 5.1). However, water-soluble TEMPO derivatives remain uneconomical, with a low oxidation potential of +0.8 V vs. Standard Hydrogen Electrode (SHE) compared to the thermodynamic limit of +1.23 V in water.[19,20] Viologen derivatives have similarly been explored as anolyte materials due to their highly reversible +1/+2 redox couple with a standard reduction potential of -0.45 V vs. SHE.[15,21] However, with a molecular weight of 257 g/mol (vs. 156 g/mol for TEMPO), batteries based on methyl viologen have a high equivalent weight (molecular weight per mol of electrons transferred).[3,18] Additionally, the use of separate electrolytes for the anode and cathode can result in capacity fade with chemical cross-over driven by concentration gradients.[22] The discovery of new stable organic radical scaffolds may therefore unlock performance and cost targets unachievable with current materials. Recent work has demonstrated that the stability of organic radicals, viewed in terms of thermodynamic stabilization and kinetic persistence, can be estimated using density functional theory (DFT).[23] In addition to stability, the single-electron half-reaction potentials of organic radicals can be reliably estimated via DFT from their adiabatic electron affinity and ionization energy.[24] Computational screening of many requirements for new redox-active moieties is therefore feasible, enabling a high-throughput search for potential candidates.

The field of goal-directed molecular optimization has evolved rapidly in recent years, boosted in part by improved machine learning (ML) tools and generative algorithms.[25] Computational lead generation has been predominantly studied in pharmaceutical research, often through generating serialized molecular structures as SMILES strings that resemble a given training database of compounds.[26–29] While algorithmic approaches like Bayesian optimization

can help to efficiently navigate predefined chemical spaces in combination with surrogate models, the number of pre-enumerated chemical species is constrained. In contrast, performing on-the-fly molecular generation during exploration enables searching even larger chemical spaces.[25] Techniques from reinforcement learning (RL) have shown an excellent ability to generate valid molecules with desired properties without relying on an existing database of molecular structures to learn valid structural motifs.[30,31] In particular, methods based on a direct tree search of molecular structures using techniques such as Monte Carlo Tree Search (MCTS) offer the ability to precisely control the search space of candidate molecules.[32–36] Most molecular optimization work has been benchmarked on fairly simple and fast-to-compute functions, such as Quantitative Estimate of Druglikeness (QED)[37] or penalized octanol-water partition coefficient (Penalized Log-P)[38] that do not provide a realistic picture of the challenges of molecular design,[39] or optimize just a single objective.[40] Additionally, generative models explore millions of potential candidates during a typical search – precluding the incorporation of more detailed and computationally intensive molecular evaluation criteria using DFT or other high-fidelity simulations. Machine learning (ML) surrogate models, given sufficient training data, have been shown to reproduce quantum chemical calculations at a fraction of the computational cost.[41–44]

In this study, we develop a complex and multi-factored objective function for organic radical charge carriers that includes radical stability, redox potential, and synthesizability considerations backed by $O(10^5)$ DFT calculations. We next implement a scalable RL approach based on single-player AlphaZero[45,46] that guarantees validity and low synthetic accessibility scores[37] for optimized molecules. As an even more ambitious goal than simply improving on the single redox couple performance of either TEMPO or methyl viologen, we sought to find stable radical candidates that simultaneously satisfy both the oxidation and reduction requirements for a

symmetric aqueous redox flow battery. Finding a single redox-active species that can perform both the oxidation and reduction reactions simplifies the battery design and reduces capacity fade through membrane cross-over.[47] Compared to the optimization of an asymmetric battery candidate, this requirement imposes a more complex multi-objective optimization challenge, as the quantum chemical energies of two one-electron redox processes must be balanced within a single small radical scaffold. Further, we performed a "synthetically aware" exploration of chemical space by evaluating synthetic accessibility on the fly and pruning the search tree where it enters synthetically or topologically impractical regions. The generative model yielded a large distribution of molecules predicted to meet the desired stability criteria while simultaneously having suitable oxidation and reduction potentials. The accuracy of these ML surrogate predictions was then validated against DFT calculations, with many radical candidates passing all criteria at the DFT level. Furthermore, we performed a post hoc analysis of the predicted retrosynthetic routes for the optimized molecules, finding many molecules with reasonable synthetic pathways.[48] This study demonstrates that goal-directed molecular optimization, coupled with a highly detailed ML surrogate model, can produce realistic candidates for demanding applications. Additionally, this study suggests that stable radical scaffolds for RFBs are likely more abundant than the limited but well-known set of experimentally characterized motifs.

5.3 Results and Discussion

An overview of the computations performed in this work is shown in Fig. 5.1. We first define our optimization criteria and benchmark a DFT workflow against experimental data. We then construct a database of radicals using this workflow, and subsequently train and validate ML

surrogate models. Reinforcement learning optimizes these surrogate objectives, yielding a set of candidate radicals. We perform both DFT confirmation and a *post hoc* synthesizability analysis on these radicals, yielding a final set of candidate results.

5.3.1 Computational screening of required features for organic active species

We begin by defining the features required for organic stable radical active species to be viable candidates for redox flow batteries (Fig. 5.1). For commercial viability, RFBs need to achieve a high charge density and high reversibility (i.e., longevity) at low cost. Active species must therefore have a precisely tuned redox potential to take full advantage of the solvent's electrochemical stability window, and a highly stable, long-lived radical center to avoid reactions that might reduce the battery's capacity over time.[49] We estimate standard redox potentials with adiabatic (i.e., geometry-optimized) ionization potentials and electron affinities obtained from implicitly solvated DFT thermochemistry including vibrational zero-point energy. Further, we estimate radical stability using a recently developed metric that incorporates both thermodynamic and kinetic stabilization of the radical center using 3D structural features and electron spin density obtained via DFT.[23] Highly-delocalized and sterically protected radicals are prioritized by this approach.

Radical groups must also be synthetically accessible. Synthesizability is considered by constraining the synthetic accessibility score (SAscore) of the closed-shell R-H molecule to be less than 4.0 [37,50] and by ensuring that the R-H bond is relatively weak, with a homolytic bond dissociation enthalpy (BDE) of 60-80 kcal/mol.[51,52] While many thermal and photochemical synthetic protocols exist to form radicals from a closed-shell parent organic compound (e.g.,

deoxygenation, dehalogenation, etc.), this BDE constraint limits our candidates to those that could be generated by a facile and selective late-stage H-atom abstraction.

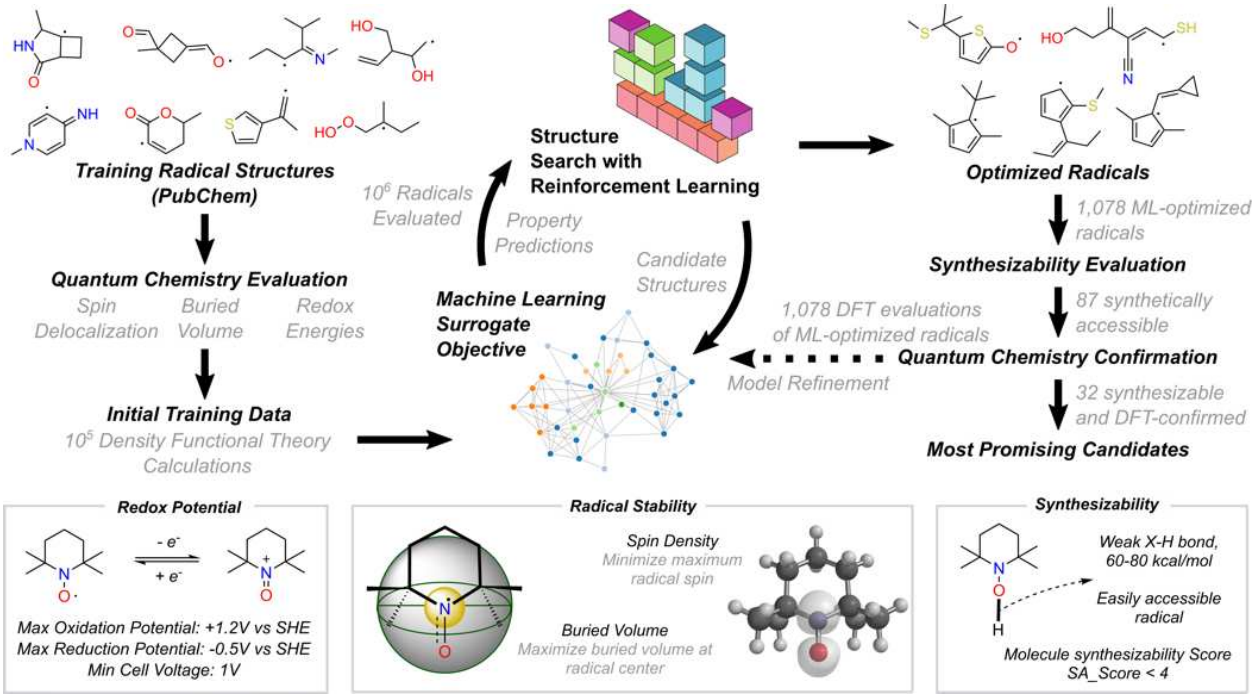


Figure 5.1*. Overview of the computational strategy and molecular design criteria used for goal-directed optimization of redox-active stable radical moieties.*

5.3.2 Development of a fast surrogate multi-objective function

To ensure the accuracy of aqueous redox calculations, we first benchmarked a wide number of different density functional, basis set, and solvation model combinations on an experimental dataset of 174 redox potentials in acetonitrile (Fig. 5.2A, APPENDIX Fig. D.S1).[53] The lowest mean absolute error (MAE) was achieved using M06-2X/def2-TZVP[54] and the SMD solvation model.[55] An additional benchmark comparing calculated and experimental redox potentials in water is outlined in APPENDIX Fig. D.S6.[56,57] We obtain an MAE of 0.25 V for 46 molecules using M06-2X/def2-TZVP with SMD solvation, with M06-2X similarly yielding the lowest error among the functionals considered.

To enable goal-directed molecular optimization, we constructed a database of 50,547 oxidation potential (OP) and 81,854 reduction potential (RP) calculations by re-optimizing radical and charged structures from an existing database of organic radicals in an implicit water solvent.[58] We impose several quality checks to ensure convergence of the DFT optimization and validity of the resulting energy calculations, including checking for normal termination of the DFT method, ensuring that bonds were not broken or formed during optimization and that the optimized open-shell molecules have minimal spin contamination (see APPENDIX D, Methods).

We next trained a graph neural network (GNN) model to predict both OP and RP directly from a radical's chemical connectivity, i.e., only based on atoms and bonds without considering a specific 3D conformation (Fig. 5.2B).[41] A test set of 2000 radicals was withheld for validation, consisting of 1773 and 1052 converged RP and OP calculations, respectively. Learning curves plot the models' prediction error as a function of database size (Fig 5.2C) and demonstrate the models continue to benefit from additional data even at the full database limit. Distributions of prediction errors (in volts) for test-set compounds using the entire training dataset are shown in Fig. 5.2D, with an MAE of 47.4 and 37.4 mV (1.1 and 0.9 kcal/mol) for OP and RP, respectively, close to the 'chemical accuracy' target of 1 kcal/mol.

Using the same chemical connectivity inputs, we trained a second surrogate GNN model on a recently published database of radical stability scores.[23,59] In this dataset, radical stability is correlated with two quantum chemical descriptors: the delocalization of the radical electron's spin, and the buried volume at the location of maximum spin (calculated as the fractional occupancy of a 3.5 Å radius sphere surrounding a target atom).[60] This GNN is trained to predict local aspects of the optimized 3D-geometry along with the quantum mechanical electron density (more precisely, the density difference between a- and b-spin electrons) at each atomic position.

Buried volume and spin density are fractional quantities bounded between 0 and 100 percent. The model achieves an MAE of 1% in buried volume prediction and 0.7% in predicting QM spin densities on each heavy (i.e., non-hydrogen) atom on 5000 radicals withheld for validation.

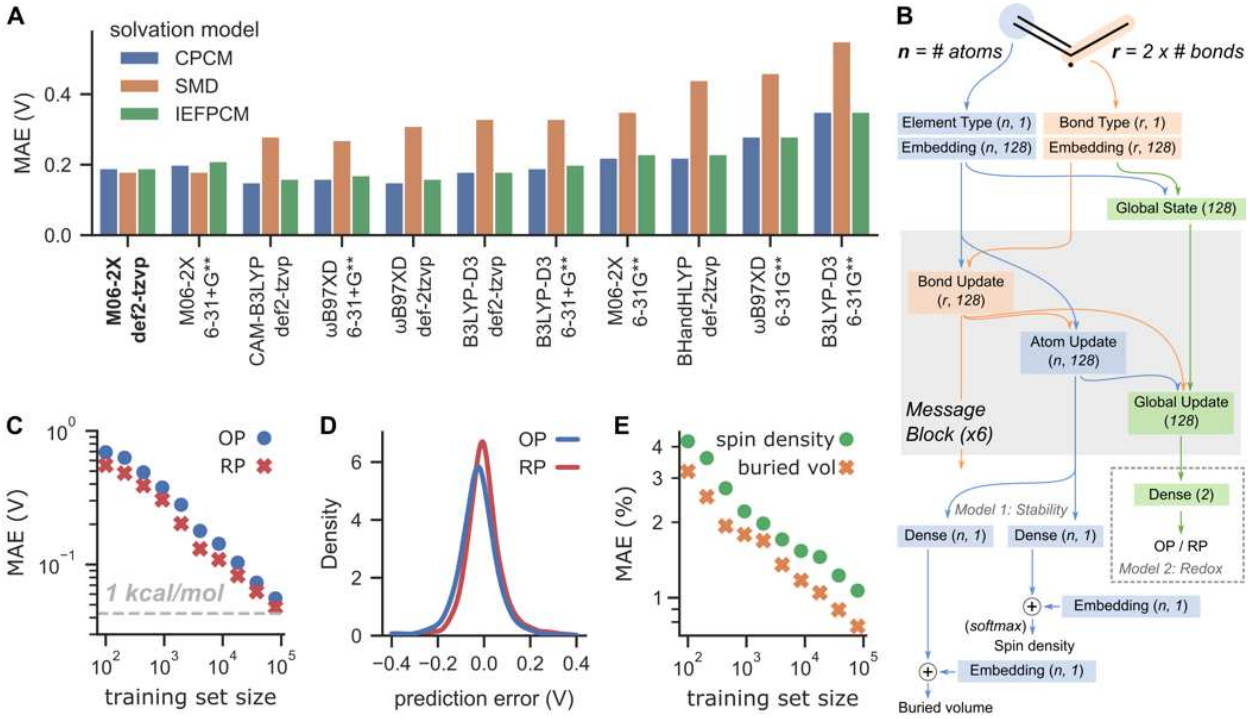


Figure 5.2. *Development of a fast surrogate objective function. (A) Prediction accuracy as a function of density functional, basis set, and solvation model on an experimental database of 174 redox potentials.*[53] *(B) Graph neural network topology for predicting stability and redox potential. Two separate models are trained. The first predicts spin density and buried volume for an input molecule at each atom. The second predicts the oxidation (OP) and reduction (RP) potentials for an entire molecule. (C) Learning curve for redox potentials showing prediction accuracy vs. number of training molecules. (D) Distribution of redox prediction errors for the final trained model. (E) Learning curve of the prediction of parameters governing radical stability.*

Stabilized radicals tend to have highly delocalized electronic structures, where substituents can potentially have a long-range influence on stability and redox potential. As demonstrated by the learning curves in Figs. 5.2C and 5.2E, the trained GNN models continue to benefit from additional training data even with nearly 100,000 training examples. The GNNs employed in this study use six message-passing layers and are therefore able to exchange

localized chemical information within a radius of 6 bonds. The ability of GNN models to learn long-distance and nonlinear interactions between functional moieties given sufficient training data is an advantage over traditional fingerprint or descriptor-based methods.

These two trained ML models, one for redox potential and one for radical stability, quickly and accurately predict many of the relevant parameters for organic radical viability in RFB applications, thus fulfilling the role of a viable surrogate for DFT calculations. Since RL frameworks typically operate with scalar reward functions, we converted the outputs of these two models into a single reward value as follows. First, we computed radical stability scores by combining the maximum predicted spin and the buried volume at the location of maximum spin. Stability scores range from near zero for highly unstable radicals (i.e., the methyl radical) to 75 or higher for radicals known to be stable experimentally.[23] Second, for the redox potential score, a maximum of 100 extra points were awarded for meeting each of four separate criteria (25 points each): (i) a reduction potential between -0.5 V and +0.2 V, (ii) an oxidation potential between +0.5 V and +1.2 V, (iii) a total voltage difference of at least 1 V, and (iv) an R-H BDE between 60-80 kcal/mol. BDEs were predicted for the hydrogen-terminated radical using a previously published ML model.[41] We added these two scores together to obtain a single reward value. Further details on the exact structure of the reward function are provided in the Methods section in APPENDIX D.

After constructing an efficient surrogate objective function, we next sought to find radicals that maximize this function. Molecule optimization was posed as a search over a directed acyclic graph (DAG), beginning the search at an initial state of a lone carbon atom. In a similar fashion to previous studies, we next considered possible actions to transition between states.[30,31] In this study, each action adds a new bond to the molecule, either between two atoms

with free valence in the original molecule (forming a ring) or between an atom in the original molecule and one of a set of possible atom additions. We considered only C, N, O, or S atoms as common elements found organic electronic materials (Fig. 5.3A). To ensure the molecules we generated were realistic, we refined the set of possible successor states from a given starting structure by (i) enumerating possible stereoisomers, (ii) canonicalizing molecules to tautomer forms,[61] and (iii) removing of molecules with high SAscore values or highly constrained ring systems. Additionally, we removed molecules containing moieties that differed substantially from the redox and stability training database. Hydrogen atoms were handled implicitly and filled free valence positions in each final molecule. To generate radical structures, additional terminal successor states were created from intermediate molecules where one atom has a hydrogen atom replaced with an unpaired electron. A more complete description of the action space, including a comparison against previous approaches, is given in APPENDIX D, Methods section.

5.3.3 Candidate optimization through reinforcement learning

In this study, we limited constructed molecules to a maximum of 12 heavy atoms (approximately the size of TEMPO), as lower molecular weight redox-active moieties are preferred for a lower equivalent weight. Including selecting of the location of the radical electron, this yields a search space of approximately 10^9 possible valid radicals, estimated via extrapolating from smaller maximum sizes and consistent with previous results.[62] The computational cost of enumerating this space grows exponentially with the maximum molecule size, motivating a more efficient strategy for finding top-performing molecules (APPENDIX Fig. D.S2).

A framework for MCTS optimization over the defined DAGs was implemented that allows for transpositions, where the same molecule can be reached through multiple paths.[63] Following the approach of AlphaZero[45], this framework was augmented with a policy model that replaces the simulation phase (using a random policy) of MCTS with a value score predicted from a GNN, which also initializes the prior scores for successor states from the given molecule. This policy model is trained in a concurrent process by maintaining a buffer of recent MCTS rollouts, sampling in-progress molecules, and minimizing a multi-objective loss function. The loss function contains both the difference between the predicted value score and the final rollout reward and the difference between predicted prior probabilities and the actual search probabilities for each of the molecule's successor nodes (see APPENDIX D, Methods). As MCTS and the AlphaZero framework were originally designed for competitive games, the ranked reward strategy was used to enable *tabula rasa* self-play for the single-player combinatorial optimization problem.[46] In this strategy, the final reward of a rollout is rescaled to {0, 1} depending on whether the reward is greater than the 75th percentile of the last 250 results. An overview of the connectivity between the rollouts, the data buffer, and the policy model is shown in Fig. 5.3B. In this fashion, the policy-guided rollouts evolve from an initial random walk over molecular space to a highly targeted exploration of regions likely to contain high-reward molecules.

To search for potential candidate radicals, 200 rollout workers were split across 50 compute nodes for four hours, with a single node equipped with dual Tesla V100 GPUs handling the continual training of the policy model. This approach resulted in a total of 34,626 rollouts and over 3.8 million terminal state radicals evaluated with the surrogate objective function. Fig. 5.3C plots the final reward from each molecule rollout as a function of time, along with the loss

values for training the policy model to predict the final value and prior probabilities for intermediate molecule states. Using ranked rewards to rescale the final reward as a function of recent rollouts means the policy model is forced to continually adapt to predict which intermediate states are the most likely to lead to higher-performing radicals.

By maintaining a cache of all surrogate reward calculations performed during the search, we can easily query the data buffer for the top radical candidates found during the optimization. Of the 3.8 million radicals evaluated, 1,078 had a total surrogate reward greater than 195, corresponding to a minimum stability score of 95. The radical stability metric rewards molecules with highly delocalized electrons and bulky groups offering steric protection of the radical center. As such, the stability metric tends to have a higher maximum value for larger molecules. Known stable radicals in this size range include TEMPO (with a stability score of 93.9) and the phenoxy radical (77.2). The reward function includes a maximum of 100 points for meeting all redox and bond strength criteria in addition to the radical stability score. From the radical training database, no radicals were found that had a stability score greater than 90 while satisfying the redox criteria.

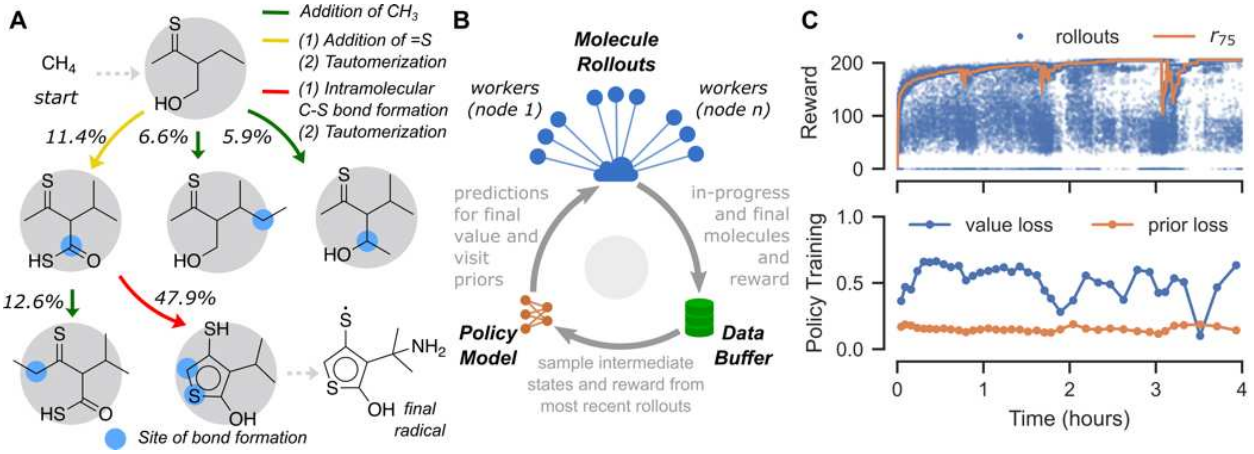


Figure 5.3. *Overview of the Reinforcement Learning (RL) structure optimization strategy. (A) An example of how molecules are constructed through the iterative addition of bonds and atoms. Percentages indicate transition probabilities between states near the end of the RL optimization (not all possible states are shown). (B) Schematic of the architecture of the computational*

search. Molecule rollouts and policy training are performed asynchronously and coordinated by a data buffer. (C) Evolution of the rewards for individual molecule rollouts (top) and losses for the policy model (bottom) versus time as the optimization proceeds. The 75th percentile of the final reward from the most recent 200 rollouts is denoted as r75 (top), and is used to reshape the reward through the ranked rewards strategy.

5.3.4 Confirmation of RL-optimized candidates with DFT

All 1078 molecules predicted to have the desired properties were subsequently analyzed with DFT to verify the accuracy of the ML models. Most top-performing candidates had close to the maximum molecule size, with 960 molecules having 12 heavy atoms, 110 molecules having 11 heavy atoms, 6 atoms having 10 heavy atoms, and just 2 molecules with 9 heavy atoms.

In Fig. 5.4A, we plot the ML-predicted redox voltages for the chosen subset, which all lie within the target triangle to permit a single radical to function as both the electron donor and acceptor in an aqueous redox flow battery with a total voltage of at least 1V. For radicals for which the DFT calculations converged, 80.5% fell within the desired target region (Fig. 5.4B). The stability scores of the radicals predicted via machine learning were then checked against those obtained via DFT. Fig. 5.4C shows the distribution of stability scores for both approaches and that stability scores obtained via DFT tended to be lower than those predicted with the surrogate objective function. Using a cutoff score of 90, well within the stability scores observed for experimentally known stable species, 41.9% of radicals were still classified as stable. As shown in APPENDIX Fig. D.S3, while buried volume predictions for optimized radicals were highly consistent with those obtained from DFT, accurate prediction of spin density was more difficult for these highly delocalized radicals. Additional training data in this region of molecular space may improve accuracy in subsequent experiments, as the generated radicals tended to be much more stable than those found in the training data (Fig. 5.4C).

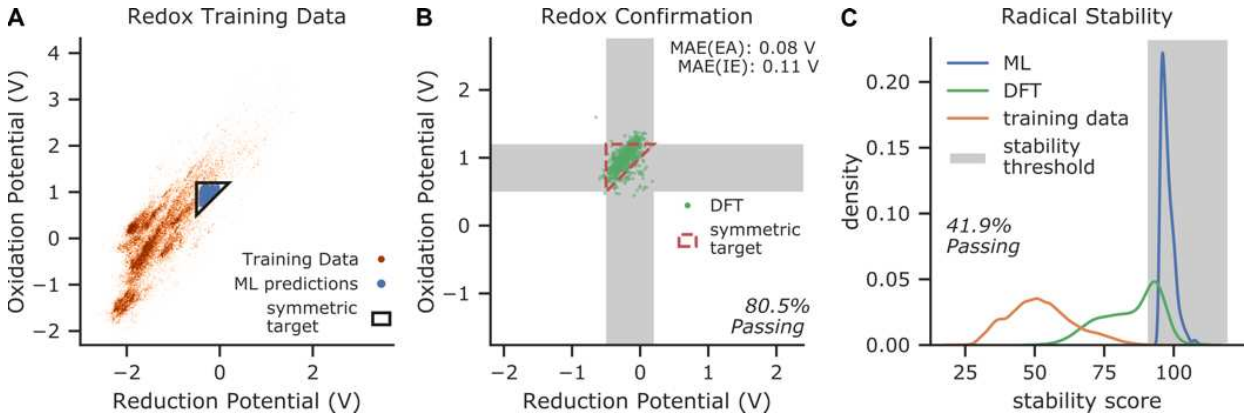


Figure 5.4. *Confirmation of top-scoring RL-generated radicals with DFT calculations. (A) Distribution of redox potentials for radicals in the training database, highlighting the target zone for an aqueous, symmetric redox battery. Predicted redox potentials for optimized radicals are shown in blue. (B) DFT computed redox properties for RL-optimized radicals that were predicted to fall within the target zone. (C) A comparison of ML (blue) and DFT-derived (green) stability estimates for optimized radicals, compared with the distribution of stability scores for radicals in the training database (orange). A stability threshold of 90 (gray) was used as a lower bound for determining whether a radical could be classified as stable.*

5.3.5 Evaluation of the synthesizability of generated molecules

The synthesizability of molecules proposed by generative algorithms has been identified as an area of concern, as theoretically optimized molecules that cannot be experimentally tested are of limited practical value.[50] To address this concern, the ASKCOS retrosynthesis prediction web service was applied post hoc to evaluate the 1078 top-ranked candidates in addition to on-the-fly constraints imposed during optimization to prioritize the search of synthetically tractable space (see APPENDIX D, Methods).[48,64] Of these, 87 returned putative synthetic routes with a median of 5 synthetic routes per candidate and an average depth of 7.9 steps. Following DFT validation, a total of 32 molecules were confirmed to satisfy the redox requirements while having high stability (>90). Chemical structures for a representative subset of these molecules are depicted in Fig. 5.5A. The RL-optimized molecules show structural variability through the varied inclusion of N, S, and O heteroatoms and extended delocalized structures, frequently with unsaturated

carbo- and heterocyclic (e.g., cyclopentadienyl, pyrrole, furan, thiophene) cores. We notice a trend of alkoxy thiophene subunits among the chemically synthesizable RL candidates. This is in correlation with the increasing use of similar molecules in organic bioelectronic devices. Previous work has shown that the electrical properties of thiophenes can be tuned by introducing alkoxy substituents.[65,66] Their enrichment among the chemically synthesizable radicals may also be due to the availability of reaction templates for this chemistry, as only 104 of the 232 DFT-confirmed RL candidates possess this functional group.

As required by the objective function, all radicals demonstrate high spin delocalization and high steric protection of the site of highest spin density. In Fig. 5.5B, we visually compare spin delocalization and buried volumes for both experimentally known and RL-optimized radicals. As expected, a high predicted stability is achieved by delocalizing the radical electron density across multiple atoms and centering the location of highest spin density on an atom with a high buried volume. We note that in Fig. 5.5A-B, the surrogate model correctly predicts that the spin is predominantly focused on the location of highest buried volume, matching DFT results, even though the radical center is formally specified at a different atom in the SMILES string.

We next investigated the predicted retrosynthetic pathways by which the wide variety of top-performing radicals might be experimentally prepared. In Fig. 5.5C, we show a putative pathway from ASKCOS for the hydrogenated form of a thiophene-based radical. Thiophenes are well-known fragments in organic electronics, where their semi-conducting properties are exploited for high efficiency.[67] The retrosynthetic route consists of a minimum of two well-established transformations involving a Friedel-Crafts alkylation and an acidic methyl ether cleavage, starting from commercially available 2-methoxythiophene and *tert*-Butyl chloride.[68–70]

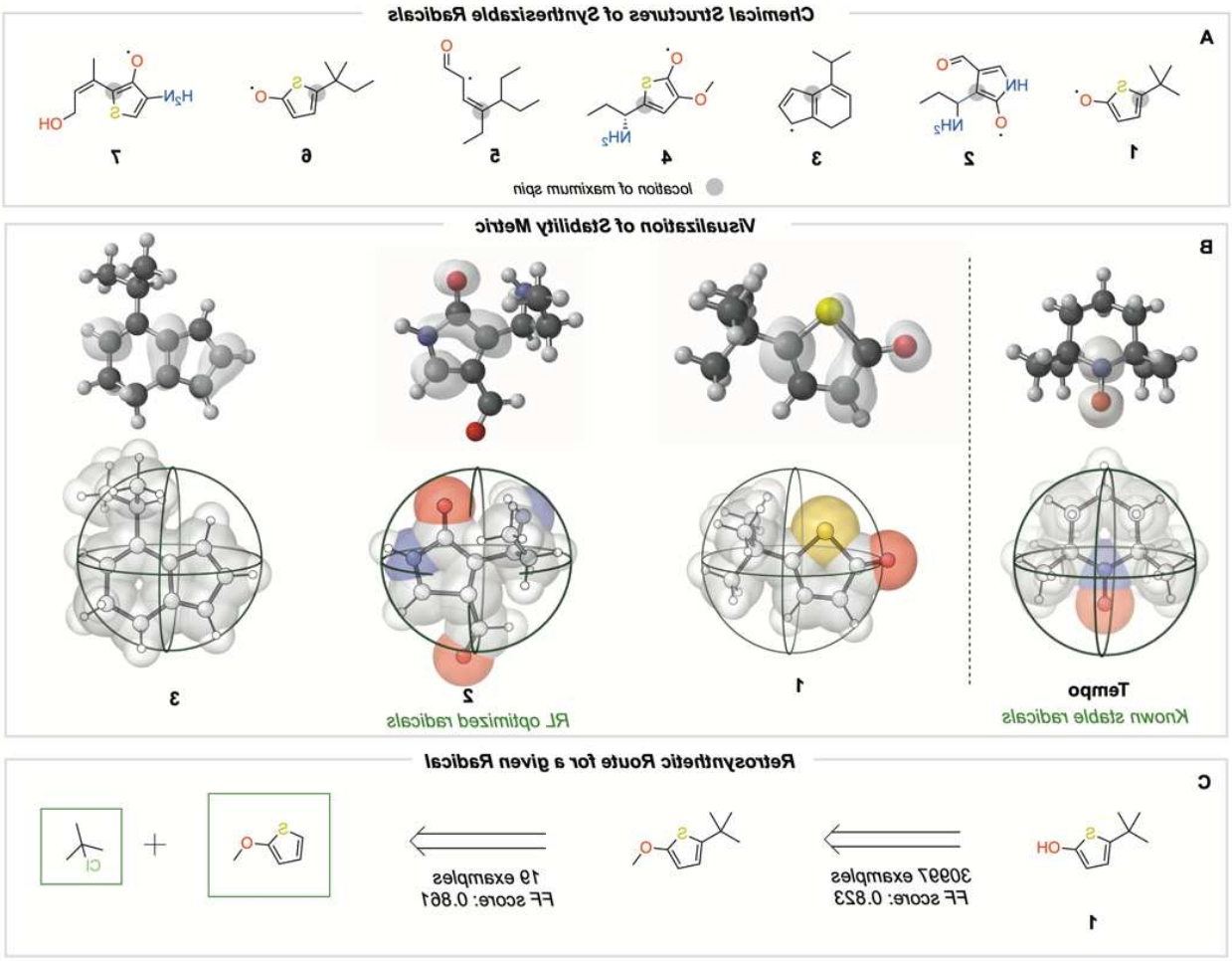


Figure 5.5. *De novo structures generated by RL. (A) Radical structures passing the design criteria subsequently validated with DFT. (B) Visualization of the radical stability metric, consisting of both spin delocalization (top row) and buried volume at the center of maximum spin (bottom row) for known stable radicals (left) and those generated via the RL algorithm (right). (C) A possible retrosynthetic pathway for the hydrogenated form of radical* ***1*** *generated by ASKCOS. FF score is the estimated plausibility value for each reaction.*

5.3.6 Error analysis of the surrogate objective function

The surrogate objective function successfully guided molecule optimization towards regions meeting the desired criteria at the DFT level. However, approximately half of the radicals predicted to meet the desired criteria ultimately fell short upon DFT confirmation. Understanding the primary modes by which the surrogate objective fails will help better understand the

limitations of machine-learning guided molecular design and further improve predictive methods for subsequent rounds of optimization.

In Fig. 5.6A, we show optimized radicals with a substantially lower DFT-calculated stability than that predicted with the surrogate model. One reason for such failure is the incorrect prediction of the maximum spin location, with a higher fraction of spin residing on an atom that is not highly shielded by bulky substituents. This failure represents in part a weakness of the chosen stability metric, as minor differences in predicted resonance can lead to large swings in the combined score. However, erroneously predicted loci of maximum spin were chemically reasonable, generally corresponding to the location of the second highest DFT spin density. Extended conjugated thiocarbonyl-based radicals and five membered cyclic alkoxy thiophenes are encountered frequently in these outliers. With DFT relaxation, the maximum spin typically locates on the terminal S atom, while the surrogate objective model predicts greater spin at a- or g- positions, in accordance with the general principle of vinylogy.[71] Retraining the surrogate objective with additional examples of these systems may improve predictions in subsequent generation rounds.

Errors in redox predictions also tended to occur for functional groups absent from the training data. In Fig. 5.6B we show the structure of one such outlier. Using the embeddings assigned by the surrogate model's penultimate prediction layer, we can explore which training set molecules are closest in structure to the target prediction. A nearest-neighbors search on this latent space reveals several cyclopentadienyl radicals with calculated redox potentials close to the erroneous prediction. The thioether substituent on a cyclopentadienyl, which is not found in any of the molecules in the redox potentials training database, strongly influences redox behavior in a way not captured by the surrogate model. The sulfur atom provides additional stabilization

of the oxidized form through resonance and of the reduced form through inductive effects (i.e., stabilization of the a-anion resonance structure). These types of prediction outliers could be remedied by augmenting the redox potential database with additional structural diversity. Learned strategies for precisely tuning redox potential

Searching for a symmetric electrolyte candidate places a challenging constraint on the electronic properties of optimized molecules, as both the oxidation and reduction potentials must be precisely and independently tuned. To explore the strategies used by the RL algorithm, we plot the relationship between oxidation potential (derived from the ionization energy) and the reduction potential (derived from the electron affinity) in Fig. 5.6C. Electron-rich, more readily oxidized (e.g., planar aminal) radicals are found in the lower-left corner, while electron-deficient, more readily reduced (e.g., heterocyclic sp^2) radicals are found in the top right corner of this plot (APPENDIX Fig. D.S4).

For open-shell molecules studied with spin-unrestricted Kohn–Sham DFT, the analog to Koopmans' theorem relates the energy of the highest singly-occupied molecular orbital (SOMO) with the vertical ionization energy.[72] Similarly, the lowest unoccupied molecular orbital (LUMO) is linked closely to vertical electron affinity, mainly when using long-range corrected density functionals.[73] From our computations we observe a correlation between a radical's SOMO energy and both redox potentials (Fig. 5.6C). This interdependence of the SOMO energy and both redox potentials illustrates the challenge of independently tuning the anode and cathode half-reactions. Qualitatively, electron-poor (low SOMO energy, electrophilic) radicals tend to be easily reduced and difficult to oxidize, while electron-rich (high SOMO energy, nucleophilic) radicals are easily oxidized and hard to reduce. However, the RL algorithm still manages to find candidates that meet both redox criteria. Radicals with the required redox properties for aqueous

batteries have a SOMO energy in the range of -6.5 to -7 eV (gray region in Fig. 5.6C). To independently optimize oxidation and reduction potentials at a fixed SOMO energy, AlphaZero learns to harness captodative stabilization of the radical center.[74] Captodative stabilization involves the incorporation of conjugated electron-donating and electron-withdrawing groups, and provides enhanced stability to all three important redox states: the radical, oxidized, and reduced states. Interestingly, this strategy mirrors the use of bipolar redox-active molecules (BRMs), an emerging strategy in the development of non-aqueous RFBs such as 2-phenyl-4,4,5,5-tetramethylimidazoline-1-oxyl-3-oxide (PTIO).[75] The algorithm thus rediscovers a fundamental concept in radical chemistry that has shown promise in the development of symmetric RFBs. However, unlike existing BRMs with relatively bulky functional groups, those discovered by RL more efficiently blend all required functionality into a much lower molecular weight moiety.

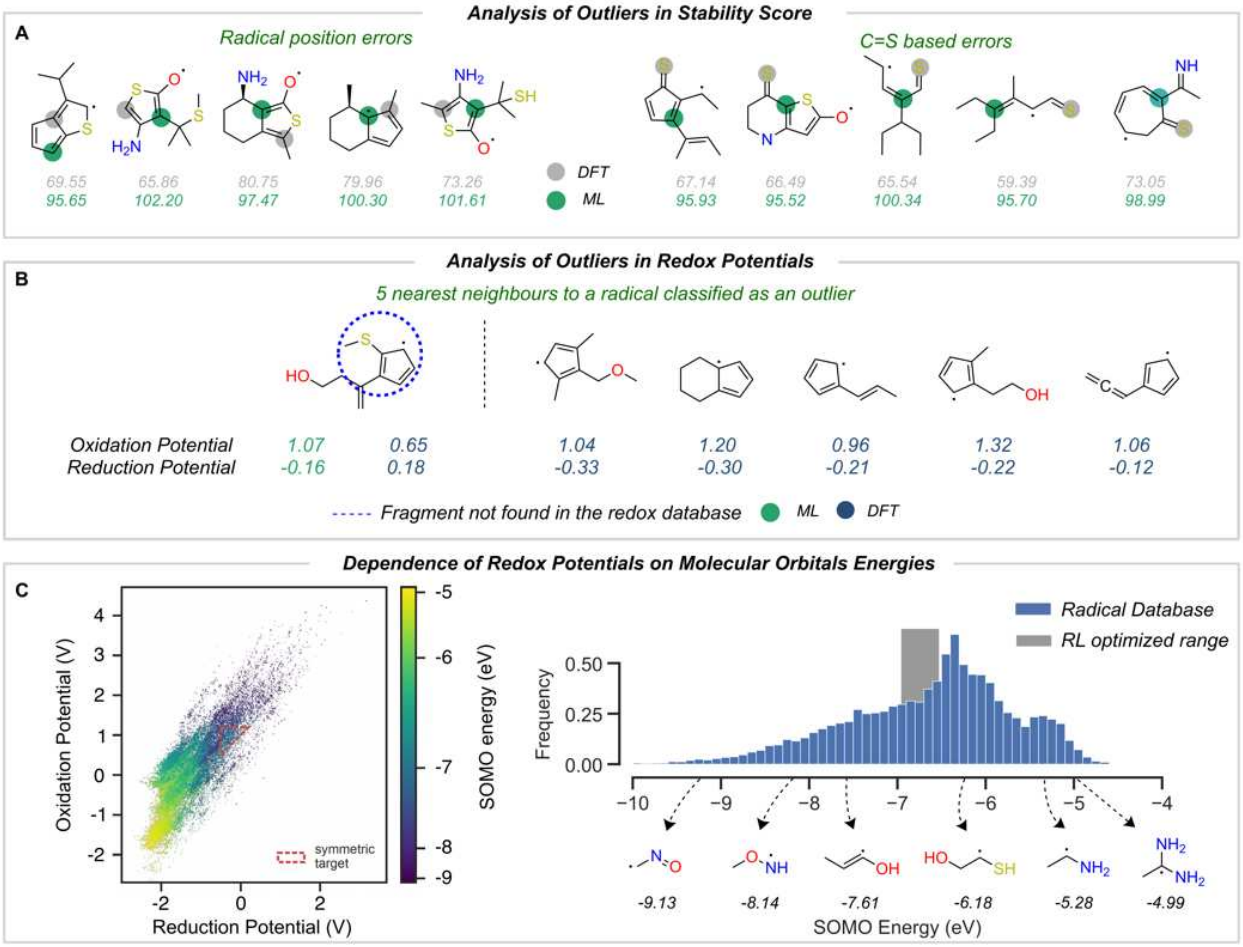


Figure 5.6. *Sources of error in the ML surrogate model and strategies for tuning redox potential. (A) Outliers in ML-predicted stability scores relative to DFT values. The predicted location of maximum spin is highlighted for both methods, and the resulting stability score is shown in gray for DFT and green for ML. (B) An example prediction error for redox potential due to lack of similar molecules present in the training database. The input radical is shown on the left, with the five closest training set radicals shown on the right. (C) On the left, the distribution in redox potentials for all training set molecules is shaded by each radical's SOMO energy. On the right, the distribution of SOMO energies of radicals in the training database is shown, with the gray region extending from the 5th to 95th percentiles of the range observed in RL-optimized candidates. Example structures from the radical database are shown for various SOMO energies.*

5.4 Conclusion

In this study, we have performed a search for molecular structures that simultaneously satisfy several complex quantum chemical phenomena important in advanced energy applications. We

have demonstrated that combining high-fidelity quantum chemistry simulations, machine learning predictive models, and state-of-the-art reinforcement learning strategies is an effective tool in efficiently exploring molecular space. Without being explicitly programmed on how to construct resonantly stabilized radicals with appropriate orbital energies, the RL algorithm learns a range of strategies that lead to high-performance candidates. We expect future work will further expand on the generative methods employed here, as new methods for goal-directed molecular design continue to emerge.[76] We provide our full training data as well as pretrained surrogate models to enable direct comparison with our results: all software, data, and models generated in this study have been made available as open-source resources (see APPENDIX D, Methods).

The construction and optimization of a surrogate model for an otherwise costly simulation or experiment is widely used in many fields, including molecular design. While in this study, candidates were found with reasonable efficiency (50% of optimized radicals), iterative refinement of the surrogate model with respect to the ground-truth calculations would improve the model's accuracy. Additionally, while molecules with putative synthetic routes were found from among the top-performing candidates, more accurate and faster methods of searching synthetically accessible space are required.

We demonstrate that organic radicals can be found that simultaneously possess the correct oxidation potential and reduction potential for symmetric aqueous redox flow batteries with a high equivalent weight. Additional refinement of the top-performing candidates is also required before they are likely to be applicable in aqueous organic redox flow batteries. Optimizing solubility with predictive models[77,78] and including charged moieties in both the training data and action space will be particularly important in achieving a high charge density. The stability metric employed also may have limitations that will need to be refined following

experimental investigation. The radical stability metric was developed to capture processes sensitive to steric effects such as bimolecular recombination or disproportionation. Other fates, such as oxidative aromatization may not be adequately predicted, requiring further refinement. Addressing these limitations to achieve holistic prediction of improved bipolar redox active candidates remains a future goal.

REFERENCES

1. Ding, Y., Zhang, C., Zhang, L., Zhou, Y. & Yu, G. Molecular engineering of organic electroactive materials for redox flow batteries. *Chem. Soc. Rev.* **2018**, *47*, 69–103.

2. Ha, S. & Gallagher, K. G. Estimating the system price of redox flow batteries for grid storage. *J. Power Sources*, **2015**, *296*, 122–132.

3. Darling, R. M., Gallagher, K. G., Kowalski, J. A., Ha, S. & Brushett, F. R. Pathways to low-cost electrochemical energy storage: a comparison of aqueous and nonaqueous flow batteries. *Energy Environ. Sci.*, **2014**, *7*, 3459–3477.

4. Hu, B., Debruler, C., Rhodes, Z. & Liu, T. L. Long-Cycling aqueous organic Redox flow battery (AORFB) toward sustainable and safe energy storage. *J. Am. Chem. Soc.* **2017**, 139, 1207–1214.

5. Kühnel, R.-S., Reber, D. & Battaglia, C. Perspective—Electrochemical Stability of Water-in-Salt Electrolytes. *J. Electrochem. Soc.* **2020**, *167*, 070544.

6. Kwabi, D. G., Ji, Y. & Aziz, M. J. Electrolyte Lifetime in Aqueous Organic Redox Flow Batteries: A Critical Review. *Chem. Rev.* **2020**, *120*, 6467–6489.

7. Wilcox, D. A., Agarkar, V., Mukherjee, S. & Boudouris, B. W. Stable Radical Materials for Energy Applications. *Annu. Rev. Chem. Biomol. Eng.* **2018**, *9*, 83–103.

8. Muench, S. *et al.* Polymer-Based Organic Batteries. *Chem. Rev.* **2016**, *116*, 9438–9484.

9. Liu, B., Tang, C. W., Jiang, H., Jia, G. & Zhao, T. Carboxyl-Functionalized TEMPO Catholyte Enabling High-Cycling-Stability and High-Energy-Density Aqueous Organic Redox Flow Batteries. *ACS Sustain. Chem. Eng.* **2021**, *9*, 6258–6265.

10. Wei, X. *et al.* Materials and Systems for Organic Redox Flow Batteries: Status and Challenges. *ACS Energy Lett.* **2017**, *2*, 2187–2204.

11. Dai, G. *et al.* The Design of Quaternary Nitrogen Redox Center for High-Performance Organic Battery Materials. *Matter* **2019**, *1*, 945–958.

12. Zhang, C. *et al.* Phenothiazine-Based Organic Catholyte for High-Capacity and Long-Life Aqueous Redox Flow Batteries. *Adv. Mater.* **2019**, ***31***, 1–8.

13. Yan, Y., Robinson, S. G., Vaid, T. P., Sigman, M. S. & Sanford, M. S. Simultaneously Enhancing the Redox Potential and Stability of Multi-Redox Organic Catholytes by Incorporating Cyclopropenium Substituents. *J. Am. Chem. Soc.* **2021**, *143*, 13450–13459.

14. Hu, B. *et al.* Improved radical stability of viologen anolytes in aqueous organic redox flow batteries. *Chem. Commun.* **2018**, *54*, 6871–6874.

15. Liu, T., Wei, X., Nie, Z., Sprenkle, V. & Wang, W. A Total Organic Aqueous Redox Flow Battery Employing a Low Cost and Sustainable Methyl Viologen Anolyte and 4-HO-TEMPO Catholyte. *Adv. Energy Mater.* **2016**, *6*.

16. Wang, W. *et al.* Recent progress in redox flow battery research and development. *Adv. Funct. Mater.* **2013**, ***23***, 970–986.

17. Shrestha, A., Hendriks, K. H., Sigman, M. S., Minteer, S. D. & Sanford, M. S. Realization of an Asymmetric Non-Aqueous Redox Flow Battery through Molecular Design to Minimize Active Species Crossover and Decomposition. *Chem. - A Eur. J.* **2020**, ***26***, 5369–5373.

18. Perry, M. L., Rodby, K. E. & Brushett, F. R. Untapped Potential: The Need and Opportunity for High-Voltage Aqueous Redox Flow Batteries. *ACS Energy Lett.* **2022**, *7*, 659–667.

19. Tian, Y. *et al.* Unlocking high-potential non-persistent radical chemistry for semi-aqueous redox batteries. *Chem. Commun.* **2019**, ***55***, 2154–2157.

20. Suo, L. *et al.* 'Water-in-salt' electrolyte enables high-voltage aqueous lithium-ion chemistries. *Science (80-.).* **2015**, *350*, 938–943.

21. Janoschka, T. *et al.* An aqueous, polymer-based redox-flow battery using non-corrosive, safe, and low-cost materials. *Nature* **2015**, ***527***, 78–81.

22. Potash, R. A., McKone, J. R., Conte, S. & Abruña, H. D. On the Benefits of a Symmetric Redox Flow Battery. *J. Electrochem. Soc.* **2016**, *163*, A338–A344.

23. Sowndarya, S. S. V., St. John, P. C. & Paton, R. S. A Quantitative Metric for Organic Radical Stability and Persistence Using Thermodynamic and Kinetic Features. *Chem. Sci.* **2021**, *12*, 13158–13166.

24. Lin, K. *et al.* A redox-flow battery with an alloxazine-based organic electrolyte. *Nat. Energy* **2016**, *1*.

25. Coley, C. W. Defining and Exploring Chemical Spaces. *Trends Chem.* **2021**, *3*, 133–145.

26. Segler, M. H. S. S., Kogej, T., Tyrchan, C. & Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* **2018**, *4*, 120–131.

27. Yang, Y. *et al.* Discovery of Highly Potent, Selective, and Orally Efficacious p300/CBP Histone Acetyltransferases Inhibitors. *J. Med. Chem.* **2020**, *63*, 1337–1360.

28. Moret, M., Helmstädter, M., Grisoni, F., Schneider, G. & Merk, D. Beam search for automated design and scoring of novel ROR ligands with machine intelligence. *Angew. Chemie Int. Ed.* **2021**, 2–8 doi:10.1002/anie.202104405.

29. Krenn, M., Häse, F., Nigam, A., Friederich, P. & Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Mach. Learn. Sci. Technol.* **2020**, *1*, 045024.

30. You, J., Liu, B., Ying, R., Pande, V. & Leskovec, J. Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation. in *NeurIPS* (Cornell University Library, 2018).

31. Zhou, Z., Kearnes, S., Li, L., Zare, R. N. & Riley, P. Optimization of Molecules via Deep Reinforcement Learning. *Sci. Rep.* **2019**, *9*, 10752.

32. Yang, X., Aasawat, T. K. & Yoshizoe, K. Practical Massively Parallel Monte-Carlo Tree Search Applied to Molecular Design. in *International Conference on Learning Representations*, **2021**.

33. Jensen, J. H. A graph-based genetic algorithm and generative model/Monte Carlo tree search for the exploration of chemical space. *Chem. Sci.* **2019**, *10*, 3567–3572.

34. Yang, X., Zhang, J., Yoshizoe, K., Terayama, K. & Tsuda, K. ChemTS: an efficient python library for

de novo molecular generation. *Sci. Technol. Adv. Mater.* **2017**, *18*, 972–976.

35. Rajasekar, A. A., Raman, K. & Ravindran, B. Goal directed molecule generation using Monte Carlo Tree Search. a*rXiv:2010.16399,* **2020**.

36. Kajita, S., Kinjo, T. & Nishi, T. Autonomous molecular design by Monte-Carlo tree search and rapid evaluations using molecular dynamics simulations. *Commun. Phys.* **2020**, *3*, 1–11.

37. Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **2009**, *1*, 8.

38. Kusner, M. J., Paige, B. & Hernández-Lobato, J. M. Grammar Variational Autoencoder. *34th Int. Conf. Mach. Learn. ICML* **2017,** *4*, 3072–3084.

39. Brown, N., Fiscato, M., Segler, M. H. S. & Vaucher, A. C. GuacaMol: Benchmarking Models for de Novo Molecular Design. *J. Chem. Inf. Model.* **2019**, *59*, 1096–1108.

40. Sumita, M., Yang, X., Ishihara, S., Tamura, R. & Tsuda, K. Hunting for Organic Molecules with Artificial Intelligence: Molecules Optimized for Desired Excitation Energies. *ACS Cent. Sci.* **2018**, *4*, 1126–1133.

41. St. John, P. C., Guan, Y., Kim, Y., Kim, S. & Paton, R. S. Prediction of organic homolytic bond dissociation enthalpies at near chemical accuracy with sub-second computational cost. *Nat. Commun.* **2020**, *11*, 1–12.

42. Guan, Y., Shree Sowndarya, S. V., Gallegos, L. C., St. John, P. C. & Paton, R. S. Real-time prediction of 1 H and 13 C chemical shifts with DFT accuracy using a 3D graph neural network. *Chem. Sci.* **2021**, *12*, 12012–12026.

43. Tabor, D. P. *et al.* Mapping the frontiers of quinone stability in aqueous media: Implications for organic aqueous redox flow batteries. *J. Mater. Chem. A* **2019**, *7*, 12833–12841.

44. Jinich, A., Sanchez-Lengeling, B., Ren, H., Harman, R. & Aspuru-Guzik, A. A Mixed Quantum Chemistry/Machine Learning Approach for the Fast and Accurate Prediction of Biochemical Redox Potentials and Its Large-Scale Application to 315 »000 Redox Reactions. *ACS Cent. Sci.* **2019**, *5*, 1199–1210.

45. Silver, D. *et al.* A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science,* **2018**, *362*, 1140–1144.

46. Laterre, A. *et al.* Ranked Reward: Enabling Self-Play Reinforcement Learning for Combinatorial Optimization. *arXiv:1807.01672*, **2018**.

47. Tong, L., Jing, Y., Gordon, R. G. & Aziz, M. J. Symmetric All-Quinone Aqueous Battery. *ACS Appl. Energy Mater.* **2019**, *2*, 4016–4021.

48. Coley, C. W., Rogers, L., Green, W. H. & Jensen, K. F. Computer-Assisted Retrosynthesis Based on Molecular Similarity. *ACS Cent. Sci.* **2017**, *3*, 1237–1245.

49. Sevov, C. S. *et al.* Physical Organic Approach to Persistent, Cyclable, Low-Potential Electrolytes for Flow Battery Applications. *J. Am. Chem. Soc.* **2017**, *139*, 2924–2927.

50. Gao, W. & Coley, C. W. The Synthesizability of Molecules Proposed by Generative Models. *J. Chem.*

Inf. Model. **2020**, *60*.

51. Henry, D. J., Parkinson, C. J., Mayer, P. M. & Radom, L. Bond Dissociation Energies and Radical Stabilization Energies Associated with Substituted Methyl Radicals. *J. Phys. Chem. A* **2001**, *105*, 6750–6756.

52. Galli, C. Nitroxyl Radicals. in *The Chemistry of Hydroxylamines, Oximes and Hydroxamic Acids* 705–750 (John Wiley & Sons, Ltd, **2008**, doi:https://doi.org/10.1002/9780470741962.ch15.

53. Roth, H. G., Romero, N. A. & Nicewicz, D. A. Experimental and Calculated Electrochemical Potentials of Common Organic Molecules for Applications to Single-Electron Redox Chemistry. *Synlett* **2016**, *27*, 714–723.

54. 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 function. *Theor. Chem. Acc.* **2007**, *120*, 215–241.

55. Marenich, A. V., Cramer, C. J. & Truhlar, D. G. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *J. Phys. Chem. B* **2009**, *113*, 6378–6396.

56. Ortiz-Rodríguez, J. C., Santana, J. A. & Méndez-Hernández, D. D. Linear correlation models for the redox potential of organic molecules in aqueous solutions. *J. Mol. Model.* **2020**, *26*, 70.

57. Isegawa, M., Neese, F. & Pantazis, D. A. Ionization Energies and Aqueous Redox Potentials of Organic Molecules: Comparison of DFT, Correlated ab Initio Theory and Pair Natural Orbital Approaches. *J. Chem. Theory Comput.* **2016**, *12*, 2272–2284.

58. St. John, P. C. *et al.* Quantum chemical calculations for over 200,000 organic radical species and 40,000 associated closed-shell molecules. *Sci. Data* **2020**, *7*, 244.

59. Sowndarya S. V., S., St. John, P. & Paton, R. Radical stability and redox potential calculations for 89,320 organic radicals in the water phase, **2021** doi:10.6084/m9.figshare.14597556.v4.

60. Cavallo, L., Correa, A., Costabile, C. & Jacobsen, H. Steric and electronic effects in the bonding of N-heterocyclic ligands to transition metals. *J. Organomet. Chem.* **2005**, *690*, 5407–5413.

61. Sitzmann, M., Ihlenfeldt, W. D. & Nicklaus, M. C. Tautomerism in large databases. *J. Comput. Aided. Mol. Des.* **2010**, *24*, 521–551.

62. Blum, L. C. & Reymond, J. L. 970 Million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732–8733.

63. Saffidine, A., Cazenave, T. & Méhat, J. UCD: Upper confidence bound for rooted directed acyclic graphs. *Knowledge-Based Syst.* **2012**, *34*, 26–33 .

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

65. Giovannitti, A. *et al.* Redox-Stability of Alkoxy-BDT Copolymers and their Use for Organic Bioelectronic Devices. *Adv. Funct. Mater.* **2018**, *28*.

66. Moreira, T. *et al.* Processable Thiophene-Based Polymers with Tailored Electronic Properties and their Application in Solid-State Electrochromic Devices Using Nanoparticle Films. *Adv. Electron. Mater.* **2021**, *7*, 1–12.

67. Dou, L., Liu, Y., Hong, Z., Li, G. & Yang, Y. Low-Bandgap Near-IR Conjugated Polymers/Molecules for Organic Electronics. *Chem. Rev.* **2015**, *115*, 12633–12665.

68. Belen'Kii, L. I. & Yakubov, A. P. Stable heteroareniumions - VIII some transformations of alkylthiophenium ions and new synthesis of 2-t-butylthiophene. *Tetrahedron* **1984**, *40***,** 2471–2477.

69. 2-Methoxythiophene | Sigma-Aldrich. https://www.sigmaaldrich.com/US/en/product/aldrich/331597.

70. 2-Chloro-2-methylpropane | Sigma-Aldrich.

71. Curti, C., Battistini, L., Sartori, A. & Zanardi, F. New Developments of the Principle of Vinylogy as Applied to π-Extended Enolate-Type Donor Systems. *Chem. Rev.* **2020**, *120*, 2448–2612.

72. Gritsenko, O. V. & Baerends, E. J. The spin-unrestricted molecular Kohn-Sham solution and the analogue of Koopmans's theorem for open-shell molecules. *J. Chem. Phys.* **2004**, *120*, 8364–8372.

73. Tsuneda, T., Song, J. W., Suzuki, S. & Hirao, K. On Koopmans' theorem in density functional theory. *J. Chem. Phys.* **2010**, *133*.

74. Bordwell, F. G. & Lynch, T. Y. Radical stabilization energies and synergistic (captodative) effects. *J. Am. Chem. Soc.* **1989**, *111*, 7558–7562.

75. Li, M., Case, J. & Minteer, S. D. Bipolar Redox-Active Molecules in Non-Aqueous Organic Redox Flow Batteries: Status and Challenges. *ChemElectroChem* **2021**, *8*, 1215–1232.

76. Nigam, A., Pollice, R., Krenn, M., Gomes, G. D. P. & Aspuru-Guzik, A. Beyond generative models: Superfast traversal, optimization, novelty, exploration and discovery (STONED) algorithm for molecules using SELFIES. *Chem. Sci.* **2021**, *12*, 7079–7090.

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

78. Alibakhshi, A. & Hartke, B. Improved prediction of solvation free energies by machine-learning polarizable continuum solvation model. *Nat. Commun.* **2021**, *12*, 1–7.

79. Hammerich, O. & Speiser, B. Techniques For Studies Of Electrochemical Reactions In Solution. in *Organic Electrochemistry* 117–188 (CRC Press, **2015**). doi:10.1201/b19122-8.

80. Frisch, M. J. *et al.* Gaussian 16 Rev. C.01. *Gaussian 16*, **2016**.

81. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **1988**, *28*, 31–36.

82. St. John, P. C. NFP: keras layers for end-to-end learning on molecular structure, **2019**. doi:10.5281/zenodo.5899629.

83. Landrum, G. A. RDKit: Open-source cheminformatics. http://www.rdkit.org, **2020**.

84. Gallegos, L. C., Luchini, G., St. John, P. C., Kim, S. & Paton, R. S. Importance of Engineered and

Learned Molecular Representations in Predicting Organic Reactivity, Selectivity, and Chemical Properties. *Acc. Chem. Res.* **2021**, *54 (4)*, 827–836.

85. Ruddigkeit, L., Van Deursen, R., Blum, L. C. & Reymond, J. L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875.

86. Biagioni, D., Skordilis, E., Tripp, C., Duplyakin, D. & St. John, P. rlmolecule: A library for general-purpose material and molecular optimization using AlphaZero-style reinforcement learning. **2020**. doi:10.5281/zenodo.5899577.

87. Sowndarya, S. S. V. & St. John, P. C. Data and trained models for predicting the stability and redox potentials of organic radicals. **2022**. doi:10.5281/zenodo.5902549.

CONCLUDING REMARKS

Over the years, Quantitative Structure-Property Relationship (QSPR) has emerged as a cornerstone in computer-aided chemical modeling, facilitating the understanding of crucial molecular behavior and properties. With the advent of machine learning, this field has witnessed substantial advancements through the exploration of new molecular representations and the utilization of complex model architectures. By harnessing the power of machine learning, we can now leverage vast amounts of data to uncover intricate relationships between molecular structures and properties, paving the way for more targeted and effective computational design strategies. This integration of QSPR and machine learning not only accelerates molecular discovery but also holds the promise of unlocking new avenues and addressing previously unmet needs. Throughout this dissertation, I have illuminated multiple facets of structure-property relationship modeling, embarking from the imperative for automation, scaling up to construct large-scale predictive models, and culminating in the application of these tools within molecular optimization strategies. This journey underscores the importance of harnessing computational methodologies to unravel the intricate connections between molecular structures and their ensuing properties, ultimately fostering advancements across diverse domains such as drug discovery, materials science, and beyond.

LIST OF PUBLICATIONS

[*]*Denotes equal contribution*.

Haas, B.;[^] Hardy, M.;[^] **S. V, S. S.;**[*] Adams, K.;***** Coley, C; Paton, R. S. Sigman, M., 'Predicting DFT-Level Descriptors for Amide Coupling Reactions' *ChemRxiv*, **2024**.

Sigmund, L. M.; **S. V, S. S.;** Albersa, A.; Erdmanna, P; Paton, R. S. Greb, L., 'Predicting Lewis Acidity: 2024 Machine Learning the Fluoride Ion Affinity of p-Block Atom-based Molecules' *Angew. Chem. Int. Ed*., e202401084.

S. V, S. S.; Kim, Y.; Kim, S.; St. John, P. C; Paton, R. S., 'Extending Bond Dissociation Prediction with Machine Learning to Medicinally and Environmentally Relevant Chemical Space' *Digit. Discovery*, **2023**, *2*, 1900-1910.

Alegre-Requena, J. V.;[*] **S. V, S. S.;**[*] Alturaifi, T.; Pérez-Soto, R.; Paton, R. S., 'AQME: Automated Quantum Mechanical Environments for Researchers and Educators' *WIREs Comput Mol Sci.* **2023**, e1663

S. V, S. S.; Law, J.; Tripp, C.; Duplyakin, D.; Skordilis, E.; Biagioni, D.; Paton, R. S.; St. John, P., 'Multi-objective goal-directed optimization of de-novo stable organic radicals for aqueous redox flow batteries' *Nat. Mach. Intell.,* **2022**, *4*, 720–730

S. V, S. S.; St. John, P.; Paton, R. S., 'A quantitative metric for organic radical stability and persistence using thermodynamic and kinetic features' *Chem. Sci.*, **2021**, *12*, 13158-13166

Guan, Y.; **S. V, S. S.;** Gallegos, L. C.; St. John, P.; Paton, R. S., 'Real-time prediction of H and C chemical shifts with DFT accuracy using a 3D graph neural network' *Chem. Sci.,* **2021**, *12*, 12012-12026

APPENDIX A: SUPPLEMENTARY MATERIALS FOR CHAPTER 2

A.1. Benchmarking of CSEARCH-RDKit and CMIN

A.1.1 Software

AQME was executed on a local machine containing an Intel Xeon Platinum 8260 Processor (48 (2 x 24) CPUs, 35.75 M cache, 2.40 GHz, 256 GB RAM) with Linux (CentOS 7). For the external and internal software employed, the following versions were used:

- Gaussian 16 vB.01[1]
- RDKit v2020.03.2[2]
- xTB v6.3.1[3]
- ANI1ccx v2.2[4]
- GoodVibes v3.0.1[5,6]
- AQME v1.0[7]

A.1.2 Benchmarking Set-1: Aromatic Vicinal Diols

A simple set of diols was used to optimize the parameters of the energy (E) and RMSD geometrical filters used by AQME during conformational search. The molecules selected have different number of atoms, and they only lead to three unique conformers, which allows us to detect when the parameters are too tight (missing conformers) or loose (excess of duplicates) (Fig. A.S1).

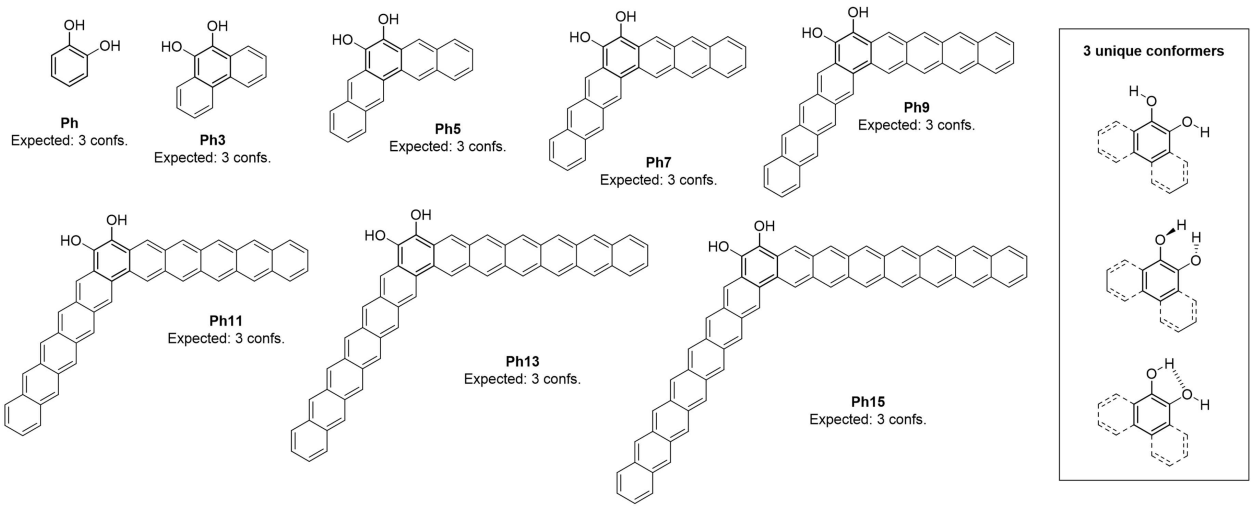


Figure A.S1*. Molecules of the initial benchmarking set along with the expected unique conformers.*

Two types of duplicates were found after the RDKit conformational samplings. In the first type, the molecules are identical (type 1) and, in the other type, the differences between conformers were based on small bond rotations (type 2). For both types of duplicates found with RDKit, the conformers converge to the same geometry after DFT optimization and, therefore, AQME should be able to remove as many of these duplicates as possible. For type 1 duplicates, we found E differences between duplicates from $4.06 \cdot 10^{-5}$ to $4.80 \cdot 10^{-11}$ kcal·mol^{-1} along the dataset for RDKit, whereas for type 2 duplicates the E differences ranged from 0.17 to 0.26 kcal·mol^{-1} for RDKit and from 0.14 to 0.21 kcal·mol^{-1} for xTB (Table A.S1). The RMSD differences varied from 0.09 to 0.19 for RDKit and from 0.09 to 0.10 for xTB for type 2 duplicates. Based on these results, we decided to include two different types of filters targeting both types of duplicates. The first filter is called E pre-filter within the program code and is designed to exclude type 1 duplicates, filtering off all the duplicates with E differences lower than 10^{-4} kcal·mol^{-1}. This E pre-filter saves a tremendous amount of computing time when running AQME since filtering based on E is much faster than filtering based on RMSD (avoiding unnecessary RMSD calculations). The second filter eliminates type 2 duplicates using E and RMSD differences. With this two-step filtering protocol, no unique conformers were discarded in RDKit and xTB. These standard internal parameters can be edited with the *energy_threshold* and *rms_threshold* options. Also, the energy window for conformers can be adapted with the *ewin_csearch* option. By default, conformers that are more than 5 kcal·mol^{-1} than the most stable conformer are discarded.

Table A.S1. *E (in kcal·mol^{-1}) and RMSD differences for type 1 and 2 duplicates obtained for benchmarking set-1. The number of initial conformers was 30 in all the cases.*

	Type 1 duplicates	Type 2 duplicates	
Molecule	*RDKit* (CSEARCH)	*RDKit* (CSEARCH)	*xTB* (CMIN)
Ph	E = $4.67 \cdot 10^{-8}$ to $4.91 \cdot 10^{-11}$	-	-

Ph3	$E = 2.14 \cdot 10^{-6}$ to $4.80 \cdot 10^{-11}$	RMSD = 0.10, E = 0.17	RMSD = 0.10, E = 0.20
Ph5	$E = 4.41 \cdot 10^{-7}$ to $1.49 \cdot 10^{-8}$	RMSD = 0.10, E = 0.25	RMSD = 0.10, E = 0.14
Ph7	$E = 3.76e \cdot 10^{-6}$ to $6.52 \cdot 10^{-9}$	RMSD = 0.19, E = 0.26	RMSD = 0.09, E = 0.20
Ph9	$E = 1.31 \cdot 10^{-5}$ to $3.90 \cdot 10^{-9}$	RMSD = 0.09, E = 0.26	RMSD = 0.09, E = 0.21
Ph11	$E = 2.05 \cdot 10^{-5}$ to $8.11 \cdot 10^{-9}$	-	-
Ph13	$E = 2.05 \cdot 10^{-5}$ to $1.24 \cdot 10^{-8}$	-	-
Ph15	$E = 4.06 \cdot 10^{-5}$ to $5.08 \cdot 10^{-9}$	-	-

A.1.3 Benchmarking Set-2: More Complex Molecules

The parameters previously set for AQME were tested using a new dataset of molecules that included wider structural diversity than the initial vicinal diol dataset (Fig. A.S2). The number of conformers tested was automatically calculated by AQME using the number of rotatable bonds and aliphatic rings, including a multiplying factor of 20 to ensure the creation of a significant number of initial samples.

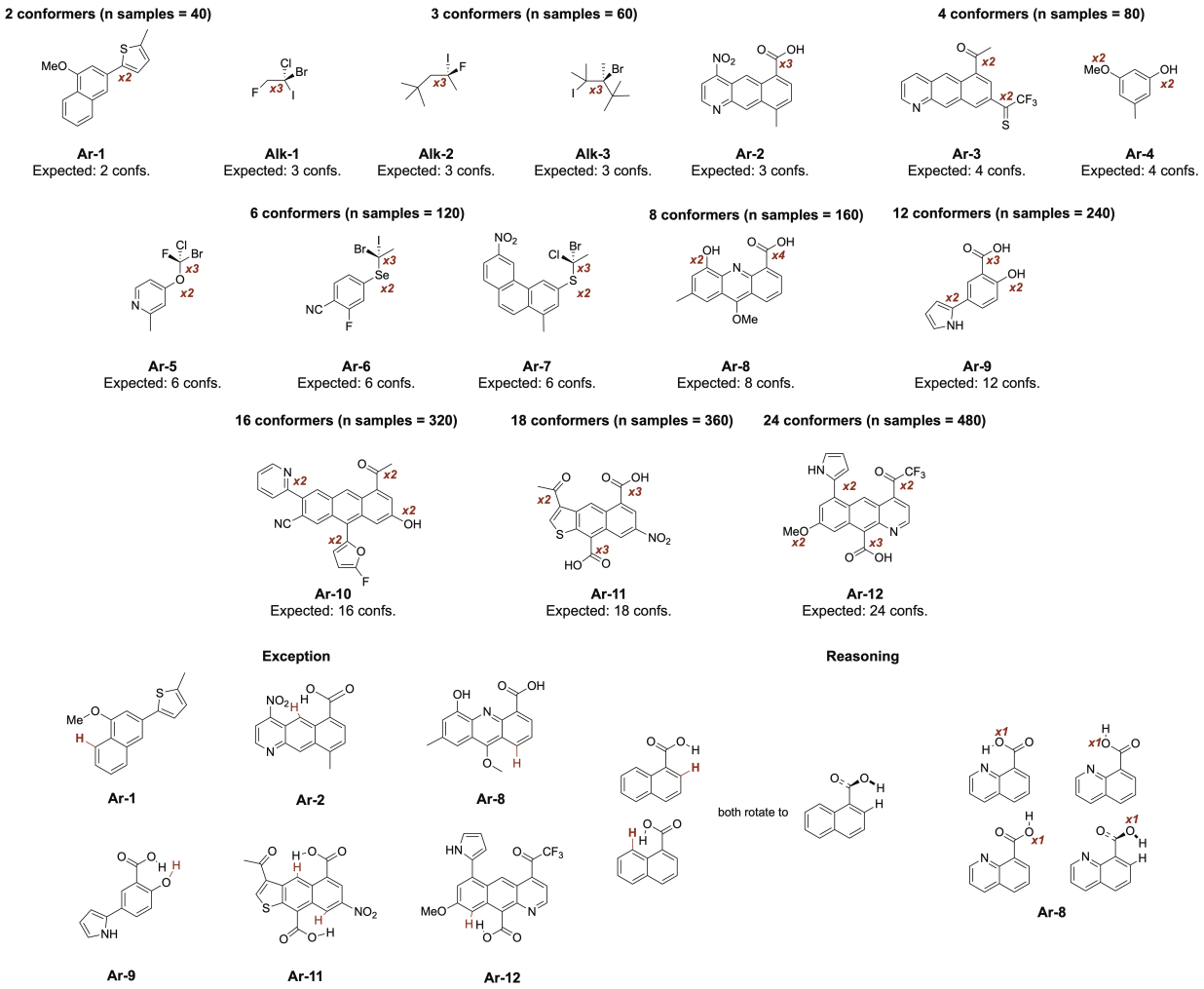
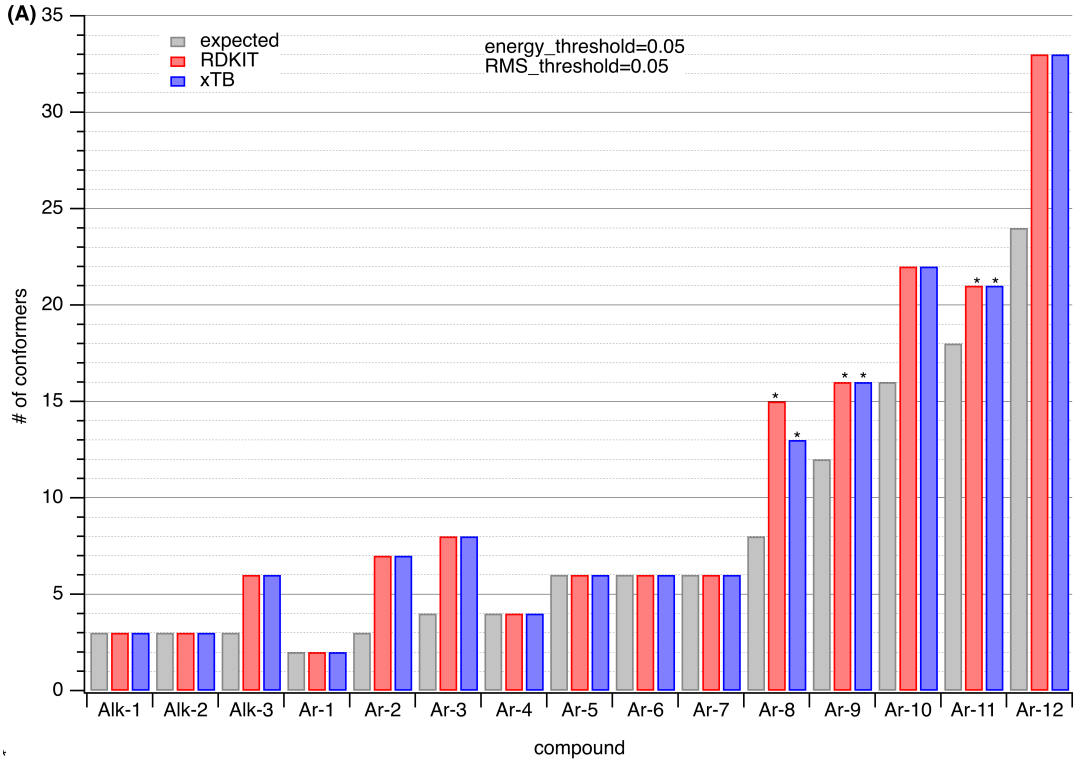
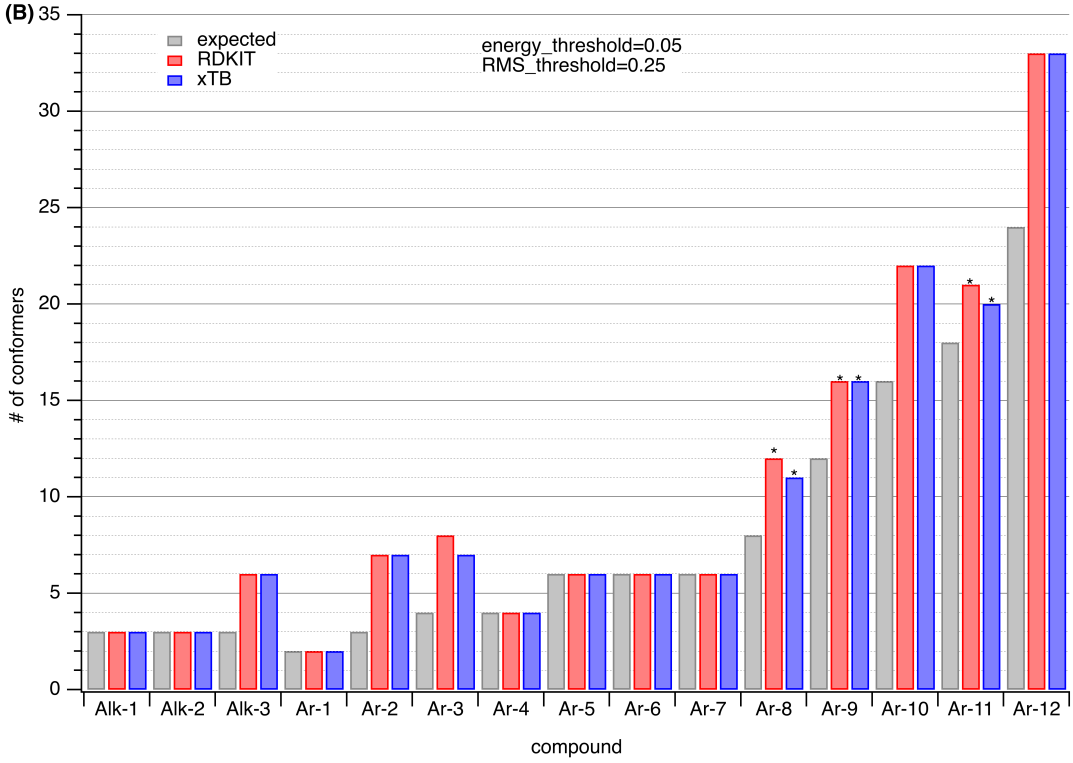


Figure A.S2. *Molecules from the second benchmarking set including the expected conformers for each case as well as the initial number of sampled conformers. These expected conformers were calculated using the different x2 and x3 rotatable groups. The exceptions to these general rules are displayed on the bottom of the figure.*

There are two exceptions to the general rules that are displayed in Fig. A.S2 to estimate the number of possible conformers. First, methoxy groups in the 1-naphthyl position (**Ar-1**) cannot be placed in the left side of the ring due to steric repulsion and, therefore, only one conformer is expected from methoxy rotation (Fig. A.S2, bottom). Also, aromatic CO_2H groups that are surrounded by two CH groups (**Ar-2**, **Ar-9**, **Ar-11** and **Ar-12**) count as x3, rather than x4 as seen in **Ar-8**, due to the steric repulsion caused by the aromatic hydrogens (Fig. A.S2, bottom).

Fig. A.S3-A.S6 show the results of 16 different energy threshold and RMS threshold combinations. The most useful combination is defined as the combination that generates all the expected conformers with least number of duplicated conformers. When the energy threshold is lowered from 0.25 kcal·mol^{-1} to 0.05 kcal·mol^{-1} while keeping a constant RMS of 0.05, no change is observed (Fig. A.S3A vs A.S4A). However, when the energy threshold is lowered from 0.25 kcal·mol^{-1} to 0.05 kcal·mol^{-1} while keeping a constant RMS of 0.25, an increase in the number of the duplicated conformers is observed (i.e. xTB in **Ar-3** and **Ar-8**, Fig. A.S3B vs A.S4B). Raising the RMSD threshold to 0.5 caused **Ar-11** to lose two of the expected conformers compared to 0.25 (Fig. A.S4B vs A.S4C). After a thorough analysis of the results for all the combinations in which we considered the number of duplicates and unique conformers, we suggest to use the standard thresholds of 0.25 kcal·mol^{-1} for E and 0.25 for RMSD filters.




(B)
expected
RDKIT
xTB
energy_threshold=0.05
RMS_threshold=0.25
# of conformers
compound
Alk-1
Alk-2
Alk-3
Ar-1
Ar-2
Ar-3
Ar-4
Ar-5
Ar-6
Ar-7
Ar-8
Ar-9
Ar-10
Ar-11
Ar-12

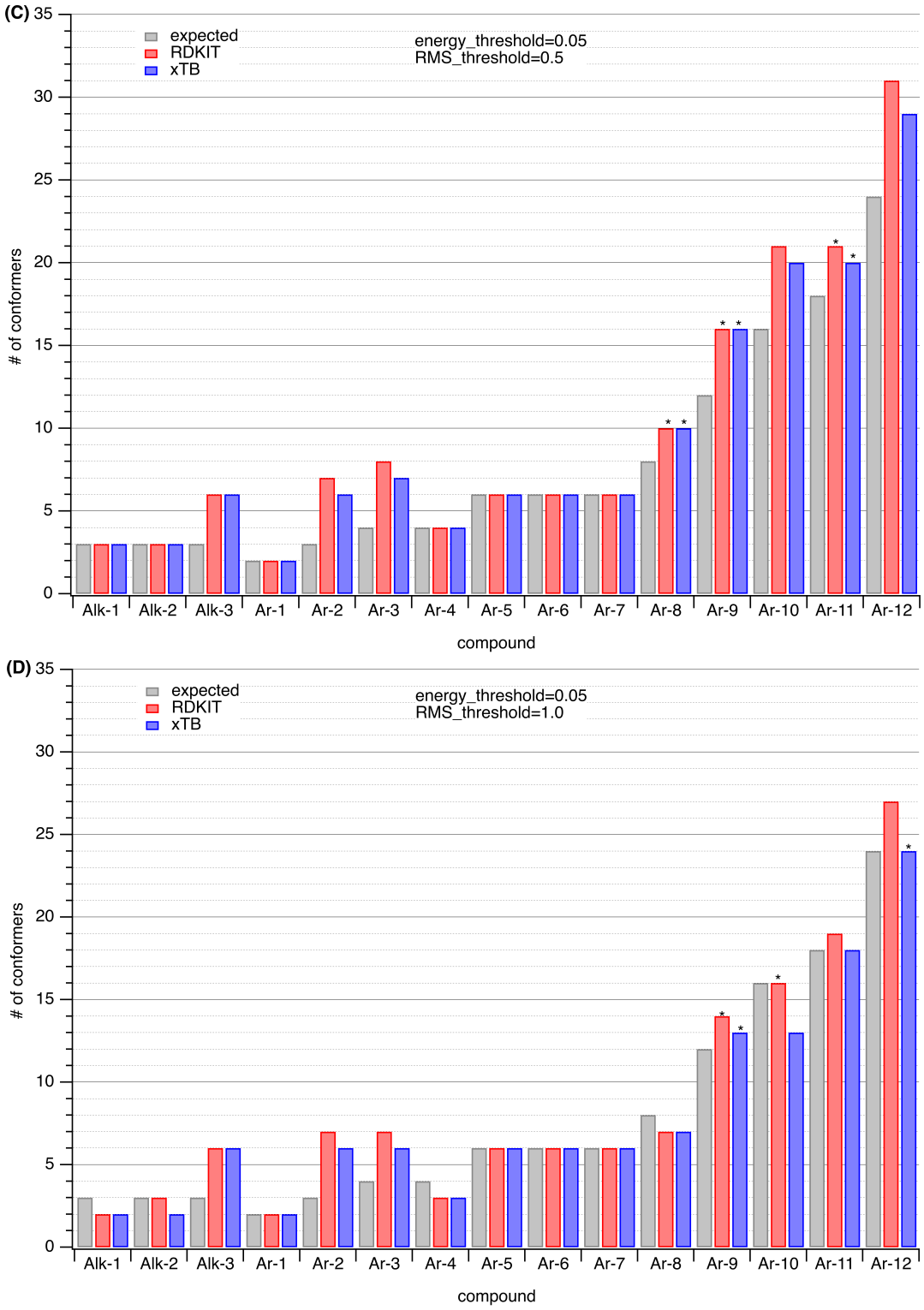
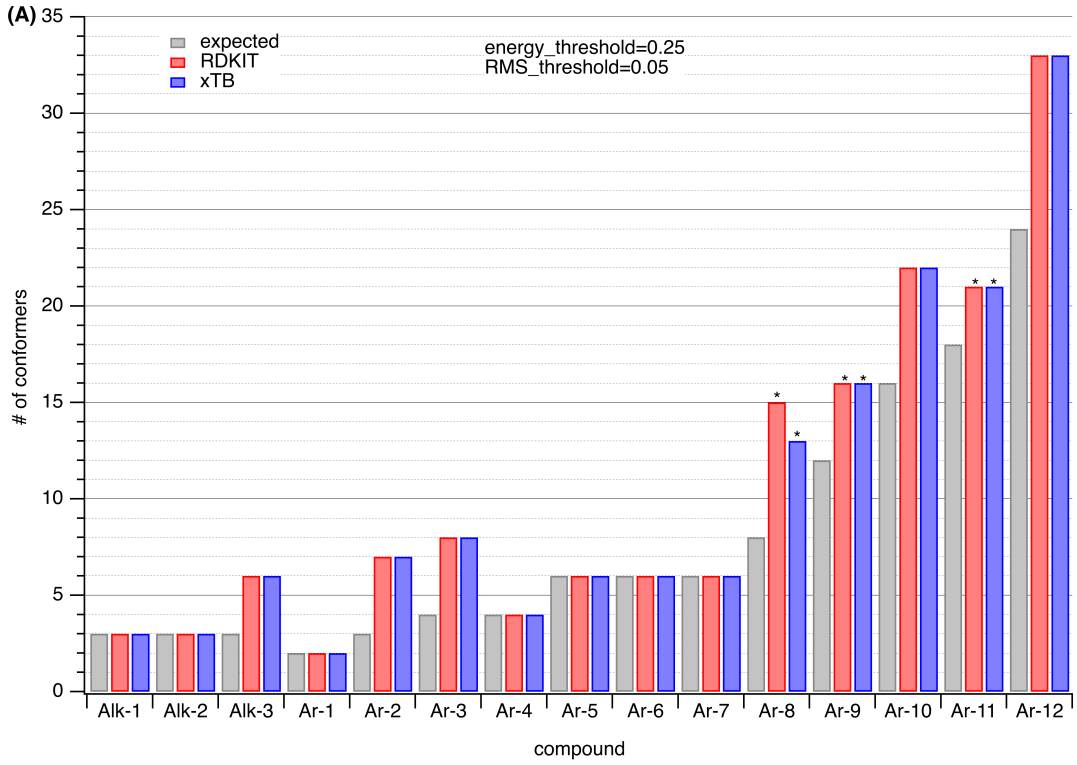
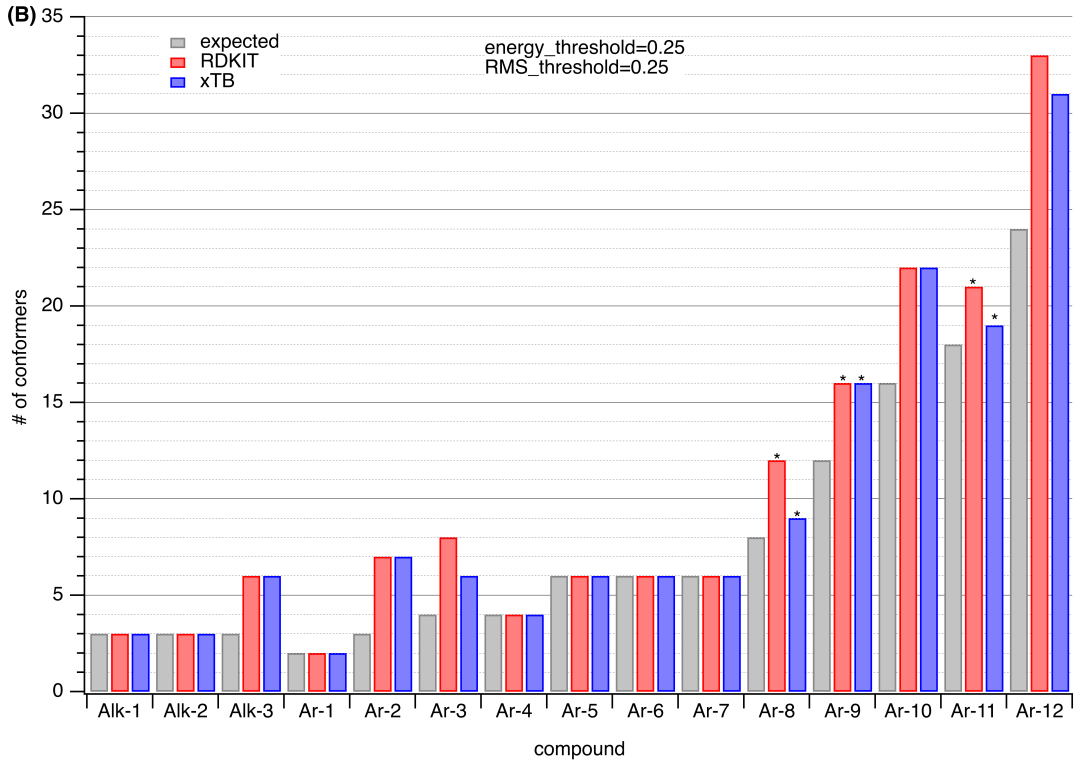


Figure A.S3. *(A-D) Number of generated conformers and the expected conformers for a constant energy threshold of 0.05 kcal·mol^{-1} and different values for the RMSD threshold ranging between 0.05 – 1.0. The asterisk corresponds to one or more missing expected conformers after doing manual comparison between the expected and the generated conformers.*


(A)
expected
RDKIT
xTB
energy_threshold=0.25
RMS_threshold=0.05
# of conformers
compound
Alk-1
Alk-2
Alk-3
Ar-1
Ar-2
Ar-3
Ar-4
Ar-5
Ar-6
Ar-7
Ar-8
Ar-9
Ar-10
Ar-11
Ar-12


(B)
expected
RDKIT
xTB
energy_threshold=0.25
RMS_threshold=0.25
# of conformers
compound
Alk-1
Alk-2
Alk-3
Ar-1
Ar-2
Ar-3
Ar-4
Ar-5
Ar-6
Ar-7
Ar-8
Ar-9
Ar-10
Ar-11
Ar-12

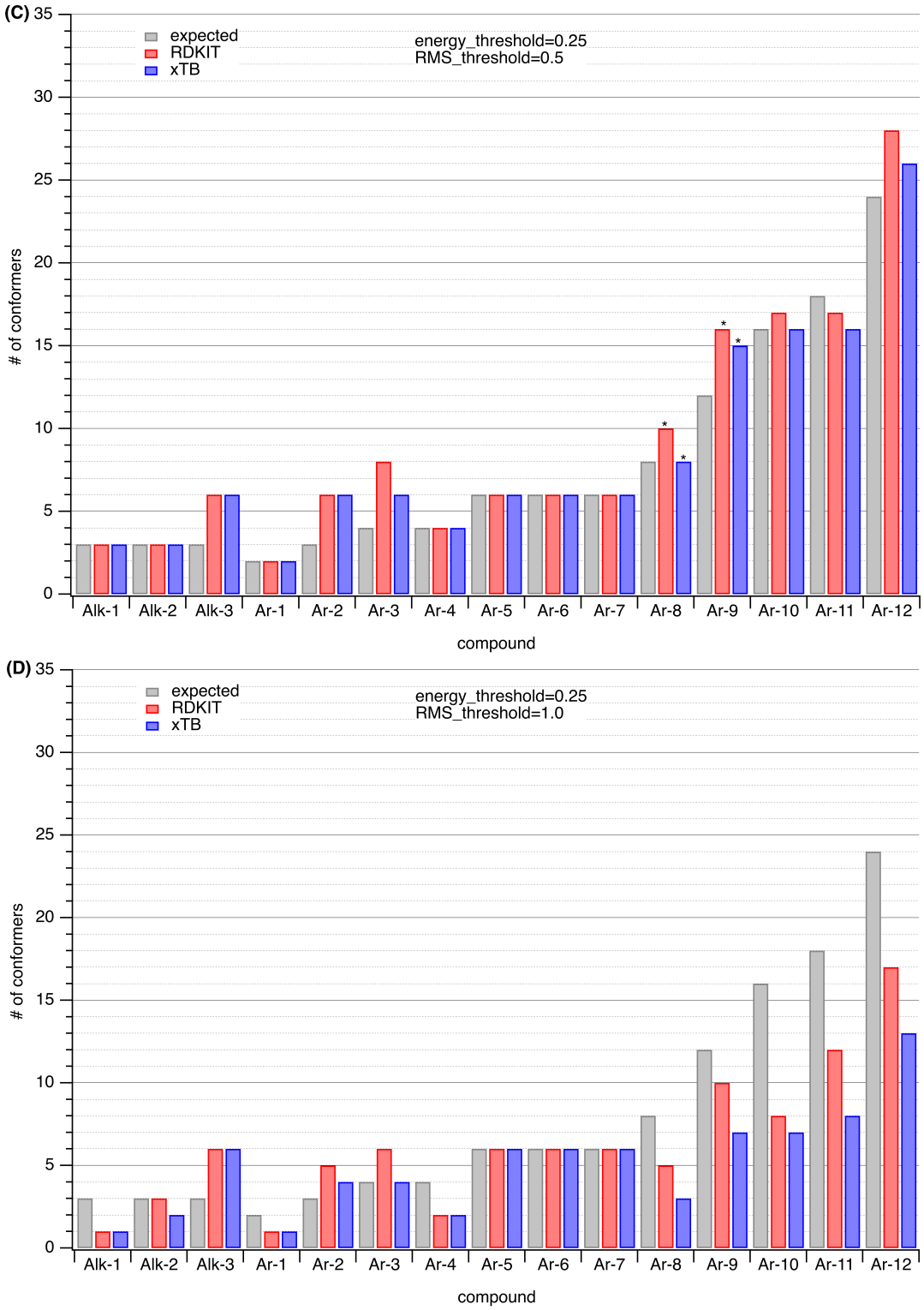
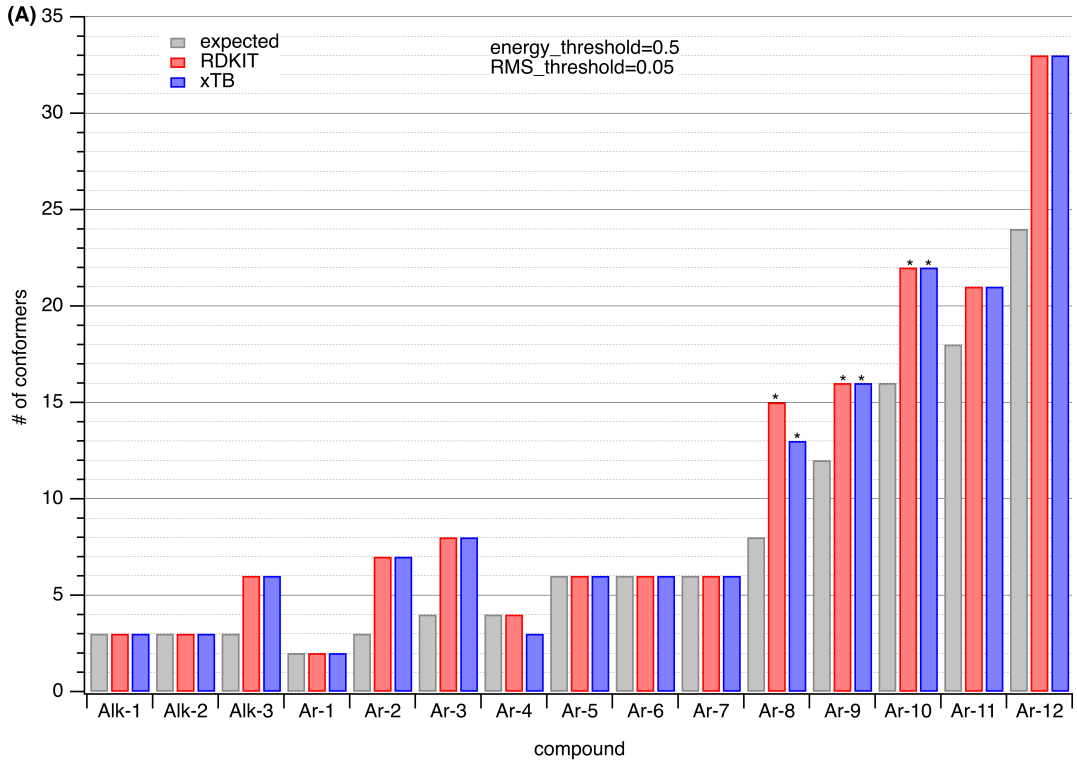
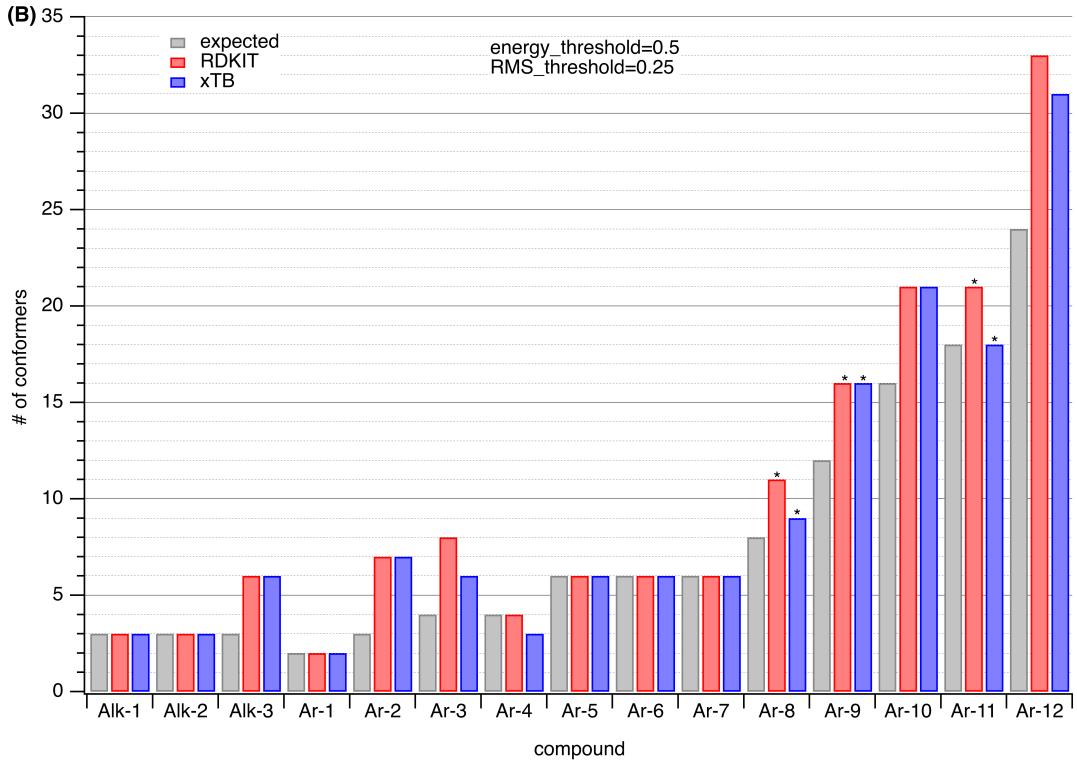


Figure A.S4. *(A-D) Number of generated conformers and the expected conformers for a constant energy threshold of 0.25 kcal·mol^{-1} and different values for the RMSD threshold ranging between 0.05 – 1.0. The asterisk corresponds to one or more missing expected conformers after doing manual comparison between the expected and the generated conformers.*


(A)
expected
RDKIT
xTB
energy_threshold=0.5
RMS_threshold=0.05
# of conformers
compound
Alk-1
Alk-2
Alk-3
Ar-1
Ar-2
Ar-3
Ar-4
Ar-5
Ar-6
Ar-7
Ar-8
Ar-9
Ar-10
Ar-11
Ar-12


(B)
expected
RDKIT
xTB
energy_threshold=0.5
RMS_threshold=0.25
# of conformers
compound
Alk-1
Alk-2
Alk-3
Ar-1
Ar-2
Ar-3
Ar-4
Ar-5
Ar-6
Ar-7
Ar-8
Ar-9
Ar-10
Ar-11
Ar-12

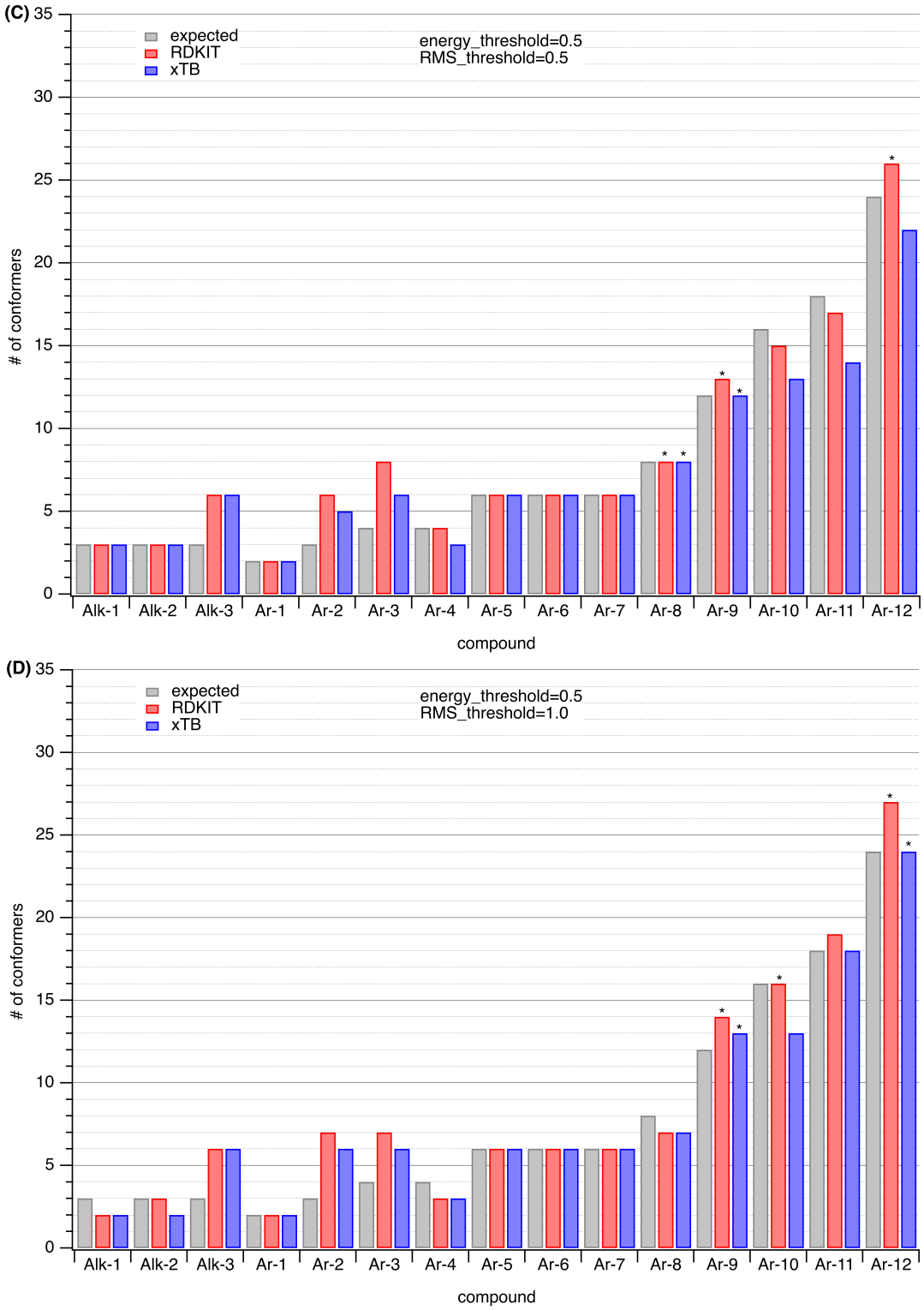
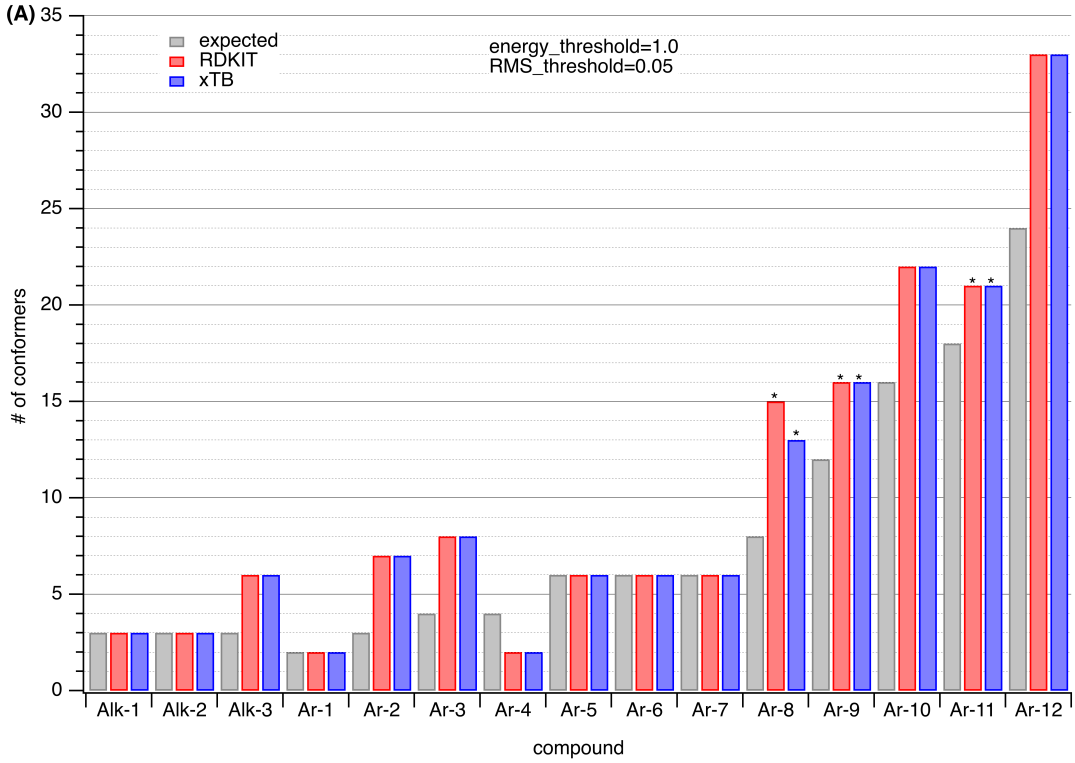
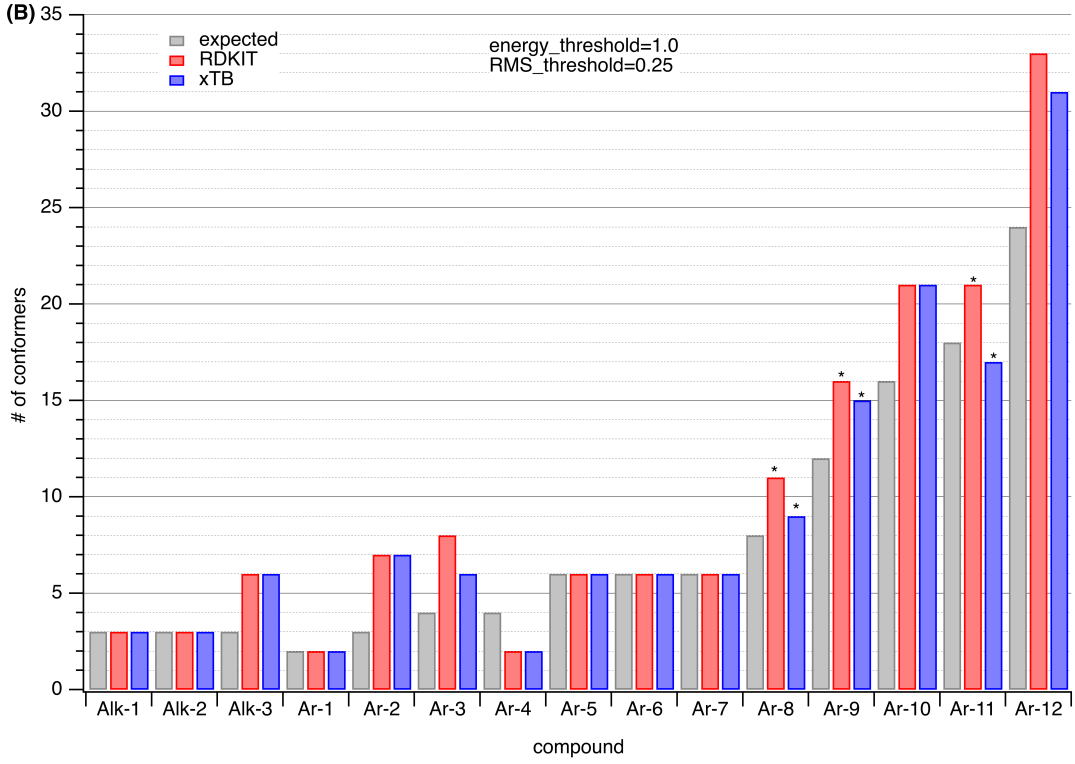


Figure A.S5. *(A-D) Number of generated conformers and the expected conformers for a constant energy threshold of 0.5 kcal·mol^{-1} and different values for the RMSD threshold ranging between 0.05 – 1.0. The asterisk corresponds to one or more missing expected conformers after doing manual comparison between the expected and the generated conformers.*


(A)
expected
RDKIT
xTB
energy_threshold=1.0
RMS_threshold=0.05
# of conformers
compound
Alk-1
Alk-2
Alk-3
Ar-1
Ar-2
Ar-3
Ar-4
Ar-5
Ar-6
Ar-7
Ar-8
Ar-9
Ar-10
Ar-11
Ar-12


(B)
expected
RDKIT
xTB
energy_threshold=1.0
RMS_threshold=0.25
# of conformers
compound
Alk-1
Alk-2
Alk-3
Ar-1
Ar-2
Ar-3
Ar-4
Ar-5
Ar-6
Ar-7
Ar-8
Ar-9
Ar-10
Ar-11
Ar-12

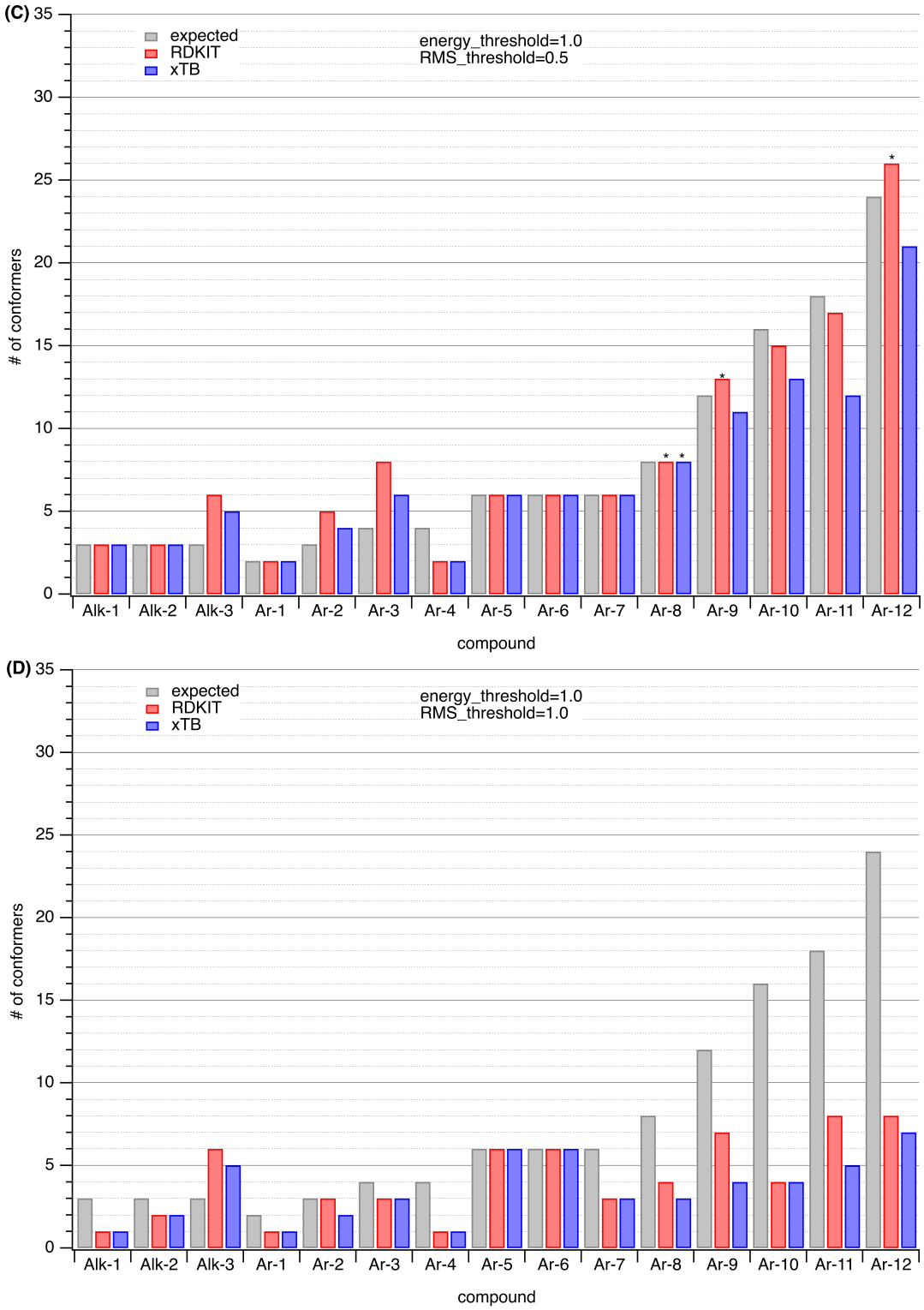


Figure A.S6. *(A-D) Number of generated conformers and the expected conformers for a constant energy threshold of 1.0 kcal·mol^{-1} and different values for the RMSD threshold ranging between 0.05 – 1.0. The asterisk corresponds to one or more missing expected conformers after doing manual comparison between the expected and the generated conformers.*

A.1.4 Benchmarking Set-3: Crystal Geometries

The validity of our optimized RDKit protocol was tested by using crystal structures of top-selling pharmaceuticals and important natural products from the Cambridge Structural Database (CSD) database[8] Q(Fig. A.S7 and Tables A.S2-A.S3). A summary of the energy difference and relative times with different methods for organic crystals is provided in Tables A.S2 and A.S3, respectively. The standard RDKit sampling yielded the most reliable results, finding the crystal geometry in eight out of nine cases. The highest difference in energy between the crystal structure and the lowest stable conformer is 3.64 kcal·mol^{-1} (**OMAJIX**). This result further strengthens the choice of a 5 kcal·mol^{-1} energy window that is set as default in the conformational sampling methods of AQME.

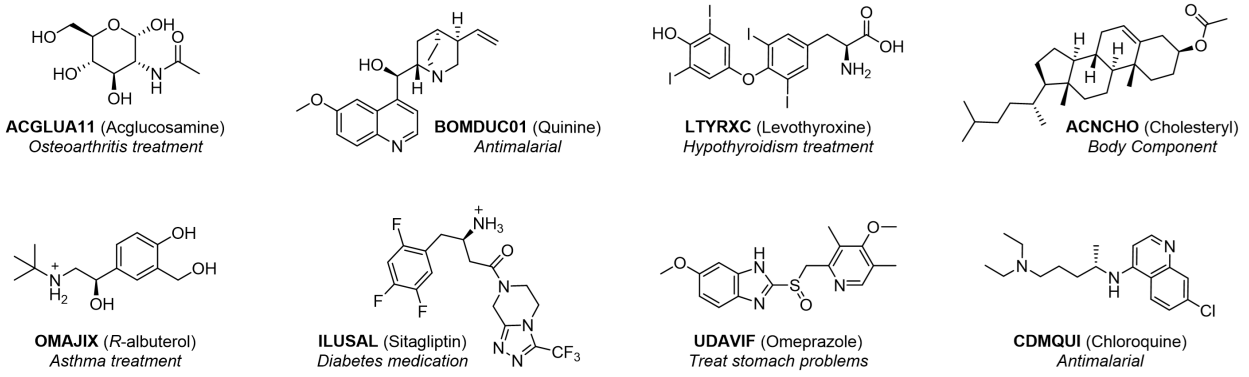


Figure A.S7. *Top-selling pharmaceuticals and important natural products from the CSD crystal database. CSD identifiers are included.*

Table A.S2. *Energy differences between the conformer that matches the crystal geometry and the most stable conformer found when using different methods for conformer generation starting from organic crystals. The number of times the method found the crystal structure in the sampling is included at the bottom.*

CCDC identifier	Energy difference (kcal·mol^{-1})			
	RDKit	xTB	ANI1ccx	ANI2x
CIDDEZ	N.F.	N.F.	N.F.	N.F.
ACGLUA11	1.43	2.00	0.00	0.69
LTYRXC	0.74	3.91	Incomp.	Incomp.

OMAJIX	3.64	N.F.	1.77	3.20
UDAVIF	1.15	3.34	Incomp.	2.60
ILUSAL	3.09	N.F.	Incomp.	2.72
CDMQUI	2.81	0.86	Incomp.	Incomp.
BOMDUC01	0.82	0.00	0.40	0.00
ACNCHO	0.59	0.35	0.29	0.26
Found crystal geometry	8 out of 9	6 out of 9	4 out of 9	6 out of 9

N.F. = conformational search did not find the conformer observed in the crystal structure. *Incomp*. = conformational search was incompatible with the method used.

Table A.S3. *Calculation time for each organic crystal using different methods for conformer generation.*

CCDC identifier	Time (seconds) RDKit	xTB	ANI1ccx	ANI2x
CIDDEZ	21.22	46.35	142.07	160.02
ACGLUA11	26.92	13.86	78.25	128.46
LTYRXC	27.25	163.48	Incomp.	Incomp.
OMAJIX	36.58	31.16	48.84	94.85
UDAVIF	42.95	57.15	Incomp.	201.86
ILUSAL	35.74	288.12	Incomp.	451.22
CDMQUI	57.82	233.96	Incomp.	Incomp.
BOMDUC01	39.99	53.8	166.93	241.74
ACNCHO	126.05	105.05	262.93	372.01
Time performance	1st (fastest)	2nd	3rd	4th

Incomp. = conformational search was incompatible with the method used.

AQME includes a modified RDKit protocol to improve the conformational samplings when using metals or atoms with unusual hybridizations, which is explained in the next section. We used multiple crystals of organometallic compounds (Fig. A.S8). The times required for the searches are normally longer when metals are included, but the precision is similar (six out of eight crystal geometries were found and the highest energy difference with the most stable conformer was 3.69 kcal·mol^{-1} for **BESREW**, Table A.S4).

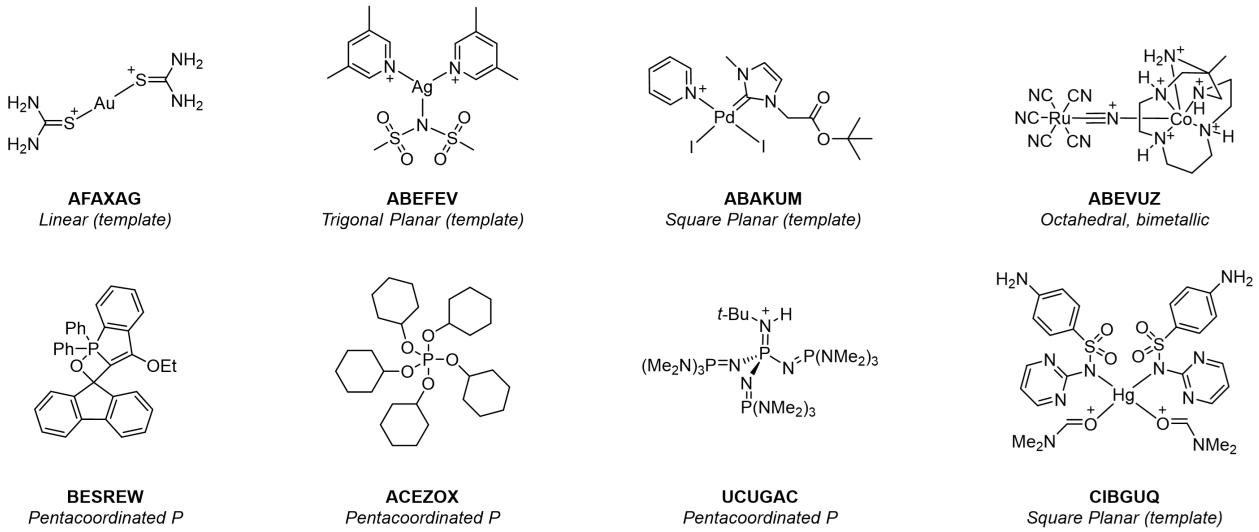


Figure A.S8. *Organometallic structures and molecules with atoms showing unusual hybridization states selected from the CSD crystal database. CSD identifiers are included.*

Table A.S4. *Energy differences between the conformer that matches the crystal geometry and the most stable conformer found when using CSEARCH-RDKit for conformer generation starting from metal crystals. The time required is also included.*

CCDC identifier	RDKit Energy difference (kcal·mol^{-1})	Time (seconds)
ABEFEV	0.50	168.76
ABEVUZ	N.F.	313.23
BESREW	3.69	107.85
ACEZOX	N.F.	5.59
AFAXAG	0.00	13.28
UCUGAC	0.85	22.53
CIBGUQ	1.26	1558.73
ABAKUM	0.39	39.89

N.F. = conformational search did not find the conformer observed in the crystal structure.

A.2 Highlighting the Importance of Specifying the Metal Type: ABEVUZ as an Example

When using RDKit conformational samplings in AQME, it is very important to include the atom types (*metal_atoms* option) and the oxidation numbers (*metal_oxi* option) of the metal centers. These two options modify the standard RDKit sampling method, making it compatible with

transition metals and atoms that show uncommon hybridization states (i.e. pentacoordinated phosphorus). As an example, we compare the most stable structures of **ABEVUZ** when using and omitting these options. Fig. A.S9A shows a valid molecular structure, which was obtained by including the two options for metals. However, a very distorted geometry is obtained when using the standard RDKit sampling protocol (Fig. A.S9B).

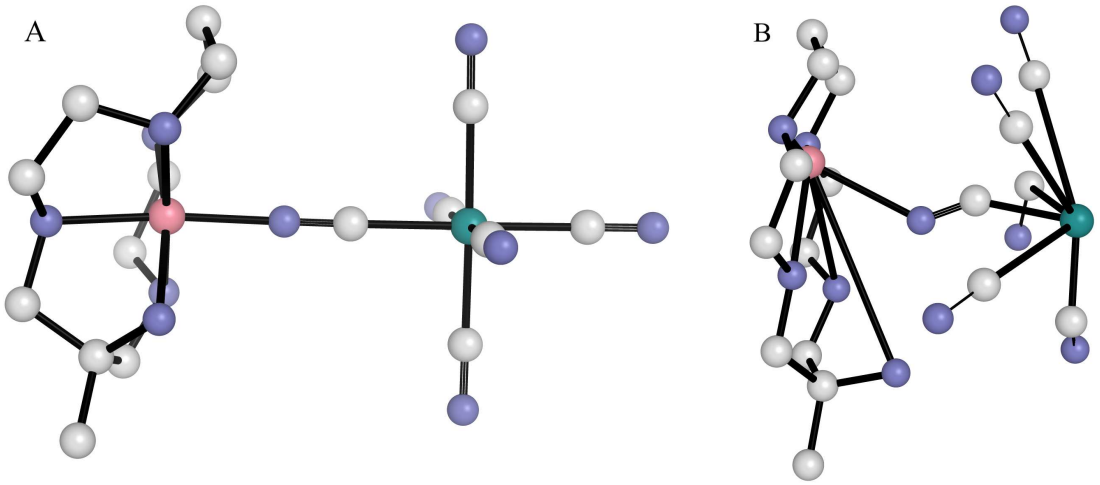


Figure A.S9. *(A) Most stable structure of ABEVUZ when specifying the metals and their oxidation numbers in the RDKit sampling. (B) Most stable structure when the metals and their oxidation numbers are not included in the RDKit sampling. Hydrogens are omitted for clarity. Carbon, gray; nitrogen, purple; cobalt, salmon; ruthenium, teal.*

REFERENCES

1. 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 Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. Gaussian 16 Rev. C.01, Wallingford, CT, **2016**.

2. RDKit: Open-source cheminformatics. http://www.rdkit.org.

3. Grimme, S.; Bannwarth, C.; Shushkov, P., A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1–86). *J. Chem. Theory Comput.* **2017**, *13*, 1989-2009.

4. 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.

5. Luchini, G.; Alegre-Requena, J. V.; Funes-Ardoiz, I.; Rodríguez-Guerra, J.; Chen, J.; Paton, R. S., bobbypaton/GoodVibes: GoodVibes v3.0.0 (v3.0.0). *Zenodo.* **2019**. https://doi.org/10.5281/zenodo.3346166.

6. Luchini G.; Alegre-Requena, J. V.; Funes-Ardoiz, I.; Paton R. S., GoodVibes: automated thermochemistry for heterogeneous computational chemistry data *F1000Research*. **2020**, *9*, 291.

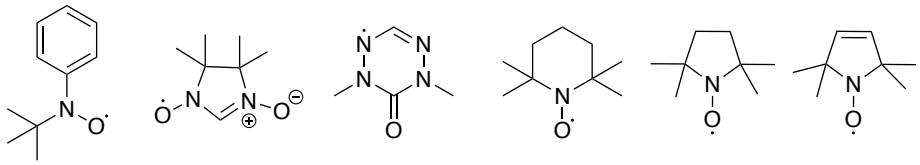
7. AQME v1.0, Alegre-Requena, J. V.; Sowndarya, S.; Pérez-Soto, R.; Alturaifi, T. M.; Paton, R. S., **2022**. https://github.com/jvalegre/aqme.

8. The Cambridge Structural Database, *Acta Crystallographica Section B*. **2016**, *72*, 171-179.

APPENDIX B: SUPPLEMENTARY MATERIALS FOR CHAPTER 3

B.1. Dependence of basis set for spin density calculations

Table B.S1. *Largest Fractional Spin Density with M06-2X using two basis sets: def2TZVP and def2QZVP. There is small basis set dependence (0.01-0.03) for the value of the normalized fractional spin.*



Name	def2-TZVP		def2-QZVP	
	spin density	fractional spin	spin density	fractional spin
nn	0.31	0.20	-0.37	0.21
proxyl	0.54	0.50	0.49	0.47
s-proxyl	0.53	0.48	0.49	0.46
TBAN	0.48	0.33	0.46	0.31
TEMPO	0.52	0.46	0.48	0.45
verdazyl	0.41	0.27	0.39	0.25

B.2 Distribution of fractional spin and buried volume of data set

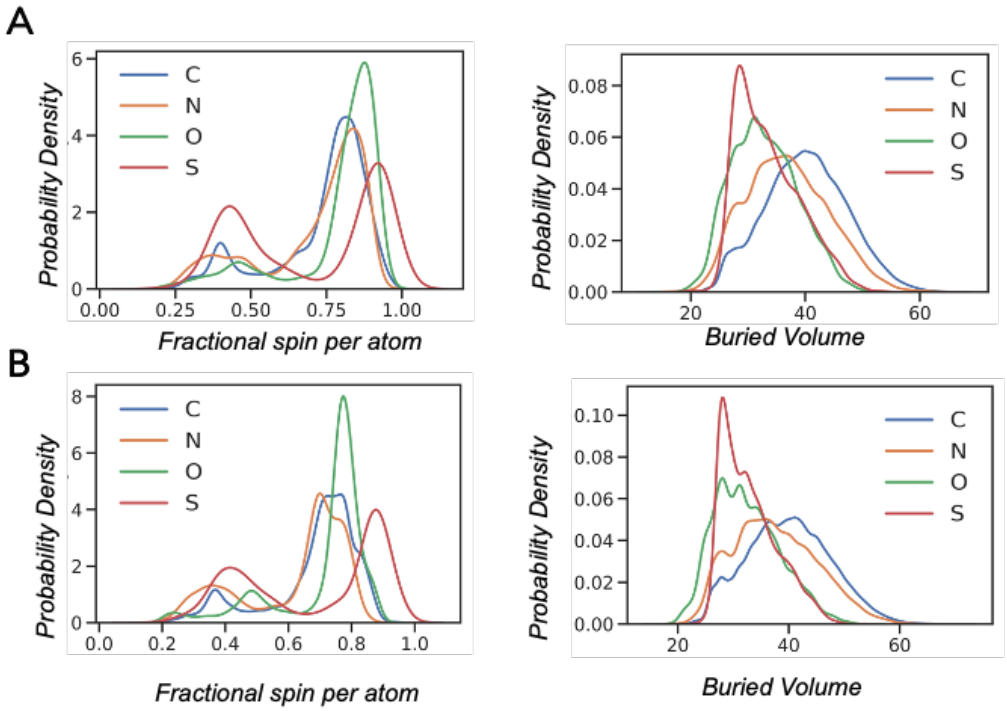


Figure B.S1. *The kernel density plots represent the spread of buried volume and fractional spin based on atom type in the data set. (A) in water phase (B) in gas phase*

B.3 Structures considered for evaluating the organic radical stability metric

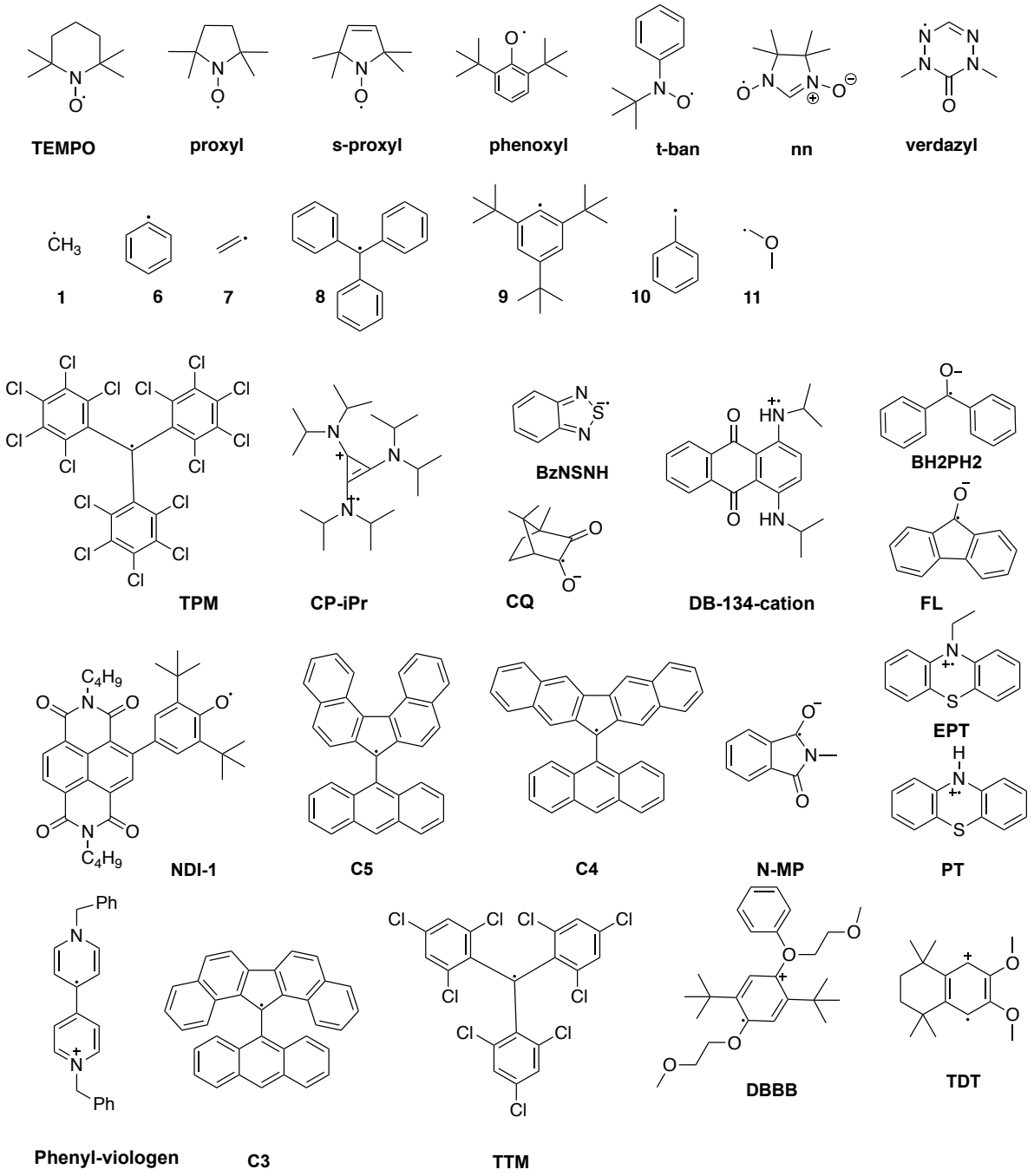


Figure B.S2. *Structures of radicals considered in the study of organic radical stability.*

B.4 Fractional spin and buried volumes for experimentally known radicals

Table B.S2. *M062X/def2-TZVP fractional Spin Density and V_{bur} (%) for selected stable radicals in gas and water phase.*

	Water			*Gas phase*		
Radical	**Atom**	**V_{bur} (%)**	**Max. spin**	**Atom**	**V_{bur} (%)**	**Max. spin**
TPM	C	84.83	0.4275	C	84.80	0.4271
CP-iPr	N	69.47	0.2731	N	69.41	0.2569
Proxyl	N	64.13	0.4954	O	43.95	0.5037
CQ	O	33.12	0.2407	O	32.98	0.2922
TEMPO	N	66.92	0.4611	O	46.85	0.4865
NDI-1	C	61.13	0.2139	C	60.93	0.1999
s-proxyl	N	62.45	0.4840	O	43.52	0.4898
car-C5	C	67.99	0.2872	C	67.97	0.2840
verdazyl	N	38.75	0.2730	N	38.75	0.2808
car-C4	C	67.62	0.2384	C	67.62	0.2366
Phenoxyl	C	38.76	0.2306	C	38.70	0.2149
phenyl-viologen	N	53.19	0.1680	N	53.11	0.1667
car-C3	C	72.45	0.2585	C	72.51	0.2622
Nn	N	53.12	0.1984	O	36.12	0.2254
TTM	C	83.93	0.3929	C	83.89	0.3959
DBBB	C	61.05	0.2391	C	60.89	0.2236
TDT	C	69.35	0.1941	C	69.30	0.2251
Tban	N	59.13	0.3330	O	41.54	0.3491
BP	C	50.61	0.1736	O	36.47	0.1819
Trityl (8)	C	64.27	0.2460	C	64.31	0.2457
N-MP	O	33.09	0.1459	C	52.62	0.1595
FL	O	31.99	0.1778	O	32.08	0.2438
DB-134-cation	N	56.14	0.1181	N	56.13	0.1347
EPT	N	60.93	0.2565	N	60.92	0.2413
PT	N	47.83	0.2388	S	43.85	0.2460
BzNSN	S	31.68	0.2364	S	31.69	0.2098
11	C	24.61	0.9008	C	24.58	0.8963
1	C	13.32	1.0000	C	13.30	1.0000
10	C	30.77	0.3886	C	30.75	0.3914
7	C	19.63	0.8487	C	19.61	0.8501
6	C	36.39	0.7280	C	36.36	0.7317

9	C	61.18	0.7230	C	61.30	0.7249

B.5 Plot of fractional spin *vs.* buried volume for radicals optimized in the gas phase

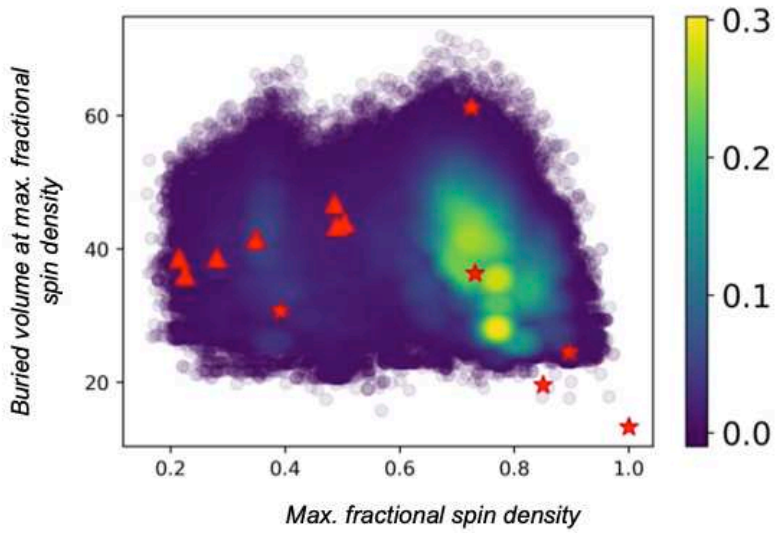


Figure B.S3. *Depiction of the stability metric to classify know radicals according to their stability. Experimentally stable radicals are shown as red triangles, located in the top left region of the graph.*

B.6 Plot of pareto-front for radicals optimized in gas phase

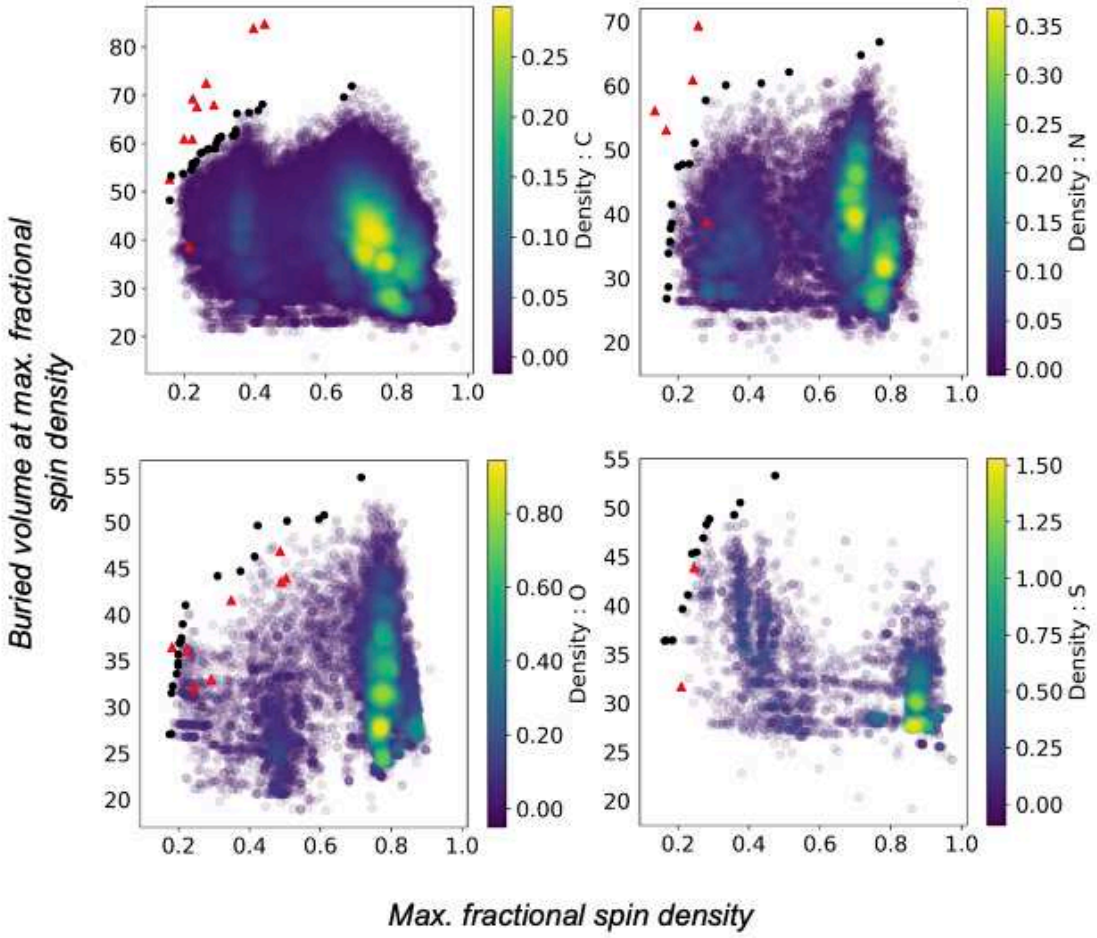


Figure B.S4. *Pareto front plots based on atom type (C, N, O, S) using fractional spin density and buried volume parameters.*

B.7 Comparison of Max fractional spin with thermodynamic quantities.

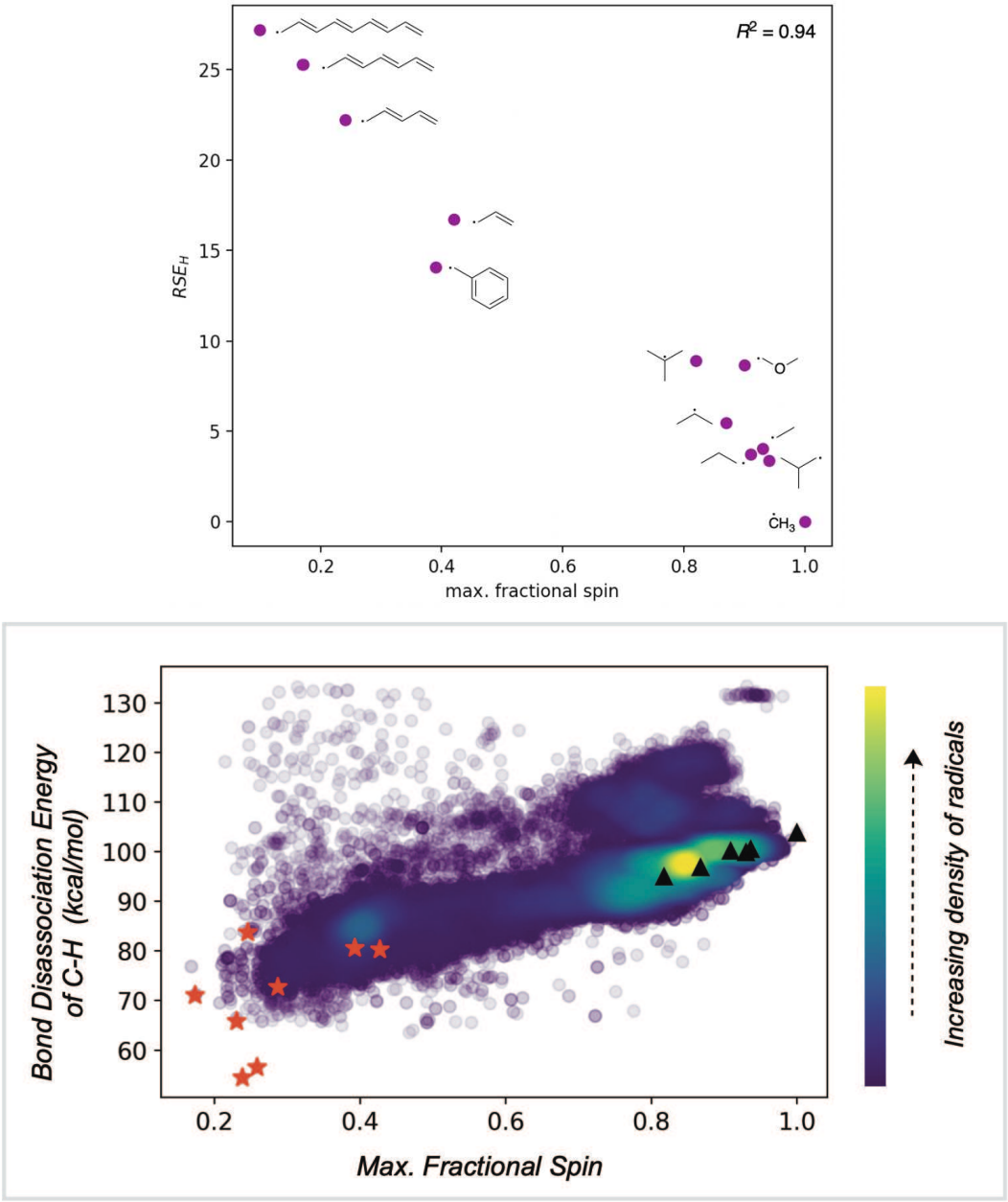


Figure B.S5. *Top: Correlation of max. fractional spin with RSE involving the cleavage of a C(sp^3)-H bond. Bottom: Correlation of max. fractional spin with C–H BDE. The red stars are known stable radicals. The black triangles are aliphatic hydrocarbons*

B.8 IQR plots for Radical Stability and Radical stabilization energies.

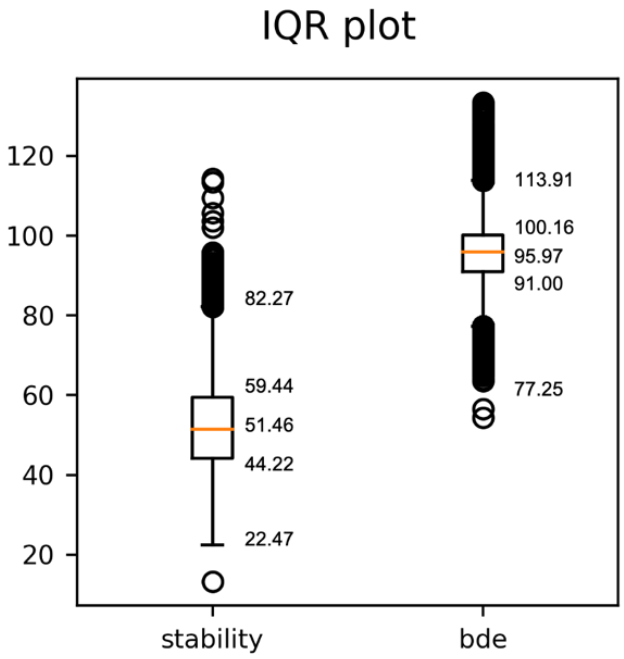


Figure B.S6. *Chemical molecular structures of stable radicals considered for comparison in C-H bond disassociation energies with the radical stability metric.*

B.9 Structures of stable radicals for the comparison of C-H BDE values.

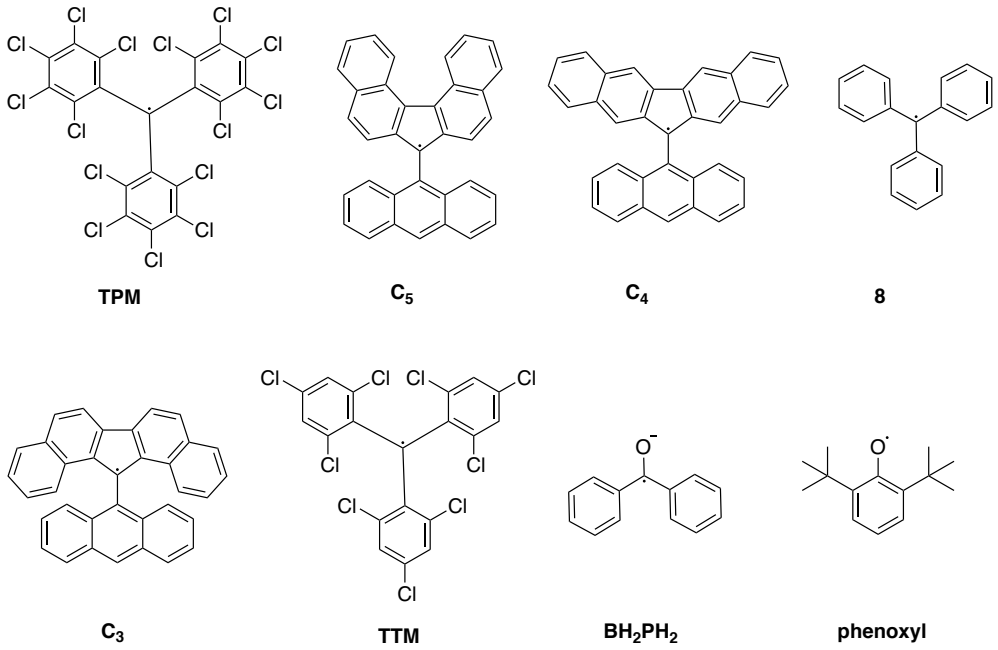


Figure B.S7*. Molecular structures of stable radicals for which C-H bond dissociation energies are compared with the radical stability (RSS) metric.*

B.10 Fractional spins, buried volumes and thermochemistry of radical cascade reactions.

Table B.S3. *Fractional Spin Density and Buried Volume for the Organic Reaction Cascades.*

Rxn	Int	Fractional Spin	Buried Volume	E (Hartree)	ZPE (Hartree)	H (Hartree)	G(T) (Hartree)	qh-G(T) (Hartree)
1	1	0.8655	35.57	-993.943240	0.274418	-993.647487	-993.722678	-993.717100
	2	0.7850	61.24	-993.988158	0.279212	-993.689260	-993.760967	-993.755715
	3	0.3001	56.50	-994.013187	0.281487	-993.713355	-993.779170	-993.775948
2	1	0.8835	35.02	-677.549588	0.332449	-677.197937	-677.265108	-677.262000
	2	0.7830	49.50	-677.572546	0.335977	-677.218424	-677.285767	-677.281396
	3	0.3817	60.99	-677.589464	0.338848	-677.233884	-677.297288	-677.293643
3	1	0.7190	46.87	-852.450664	0.400788	-852.026430	-852.099506	-852.097533
	2	0.7718	56.66	-852.490311	0.404021	-852.064150	-852.134397	-852.132488
	3	0.6411	56.99	-852.549563	0.407272	-852.121107	-852.190523	-852.188037

B.11 Correlation of RSS metric and relative bimolecular rates of radical decomposition.

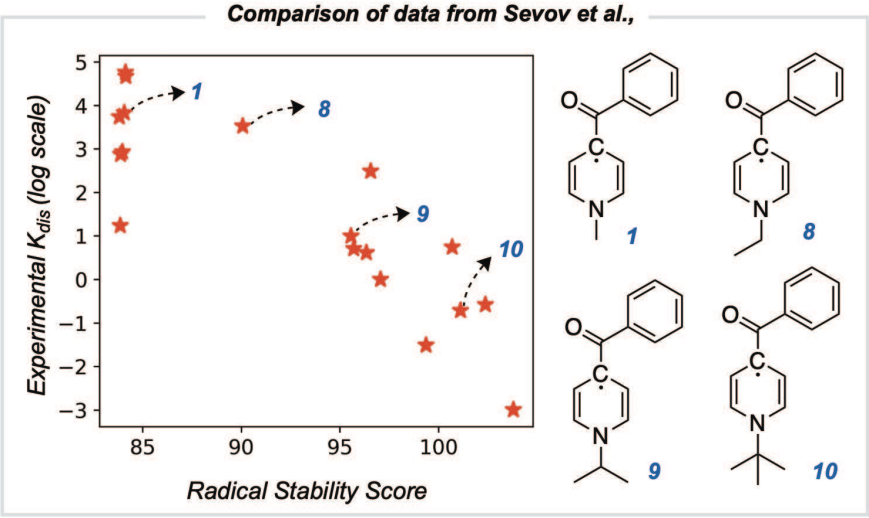


Figure B.S8. *Comparison of log(k_{rel}) and RSS for 18 radicals originally studied by Sevov (J. Am. Chem. Soc. 2017, 139, 8, 2924–2927). Buried volumes were generated at the N atoms.*

B.12 RSS metrics generated with different radii used for buried volume analysis.

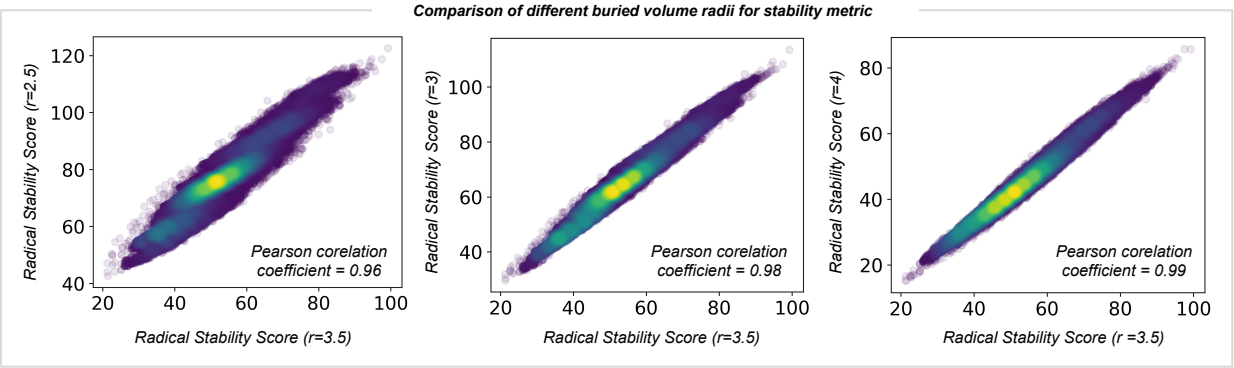


Figure B.S9. *Comparison of RSS metric generated with radii at 2.5, 3.0 and 4.0 Å against the more traditional value of 3.5 Å used for buried volume calculations.*

APPENDIX C: SUPPLEMENTARY MATERIALS FOR CHAPTER 4

C.1. Development of the *BDE-db2* dataset

A total of 244,850 Density Functional Theory (DFT) calculations were performed for both the (closed-shell) parent molecules and homolytically-dissociated (open-shell) radicals. All structures were fully optimized at the (U)M06-2X/def2-TZVP[1, 2] level of theory with Gaussian 16.[3] Starting from the SMILES representation (H-capped for open-shell species) we identify the lowest energy conformer from RDKit[4, 5] and use this geometry as the starting point for DFT optimization and to obtain the energetics and thermochemistry of open- and closed-shell species (Fig. S1A)). The forcefield utilized is MMFF94s, the number of embedded conformers is dependent on the number of rotatable bonds (n) where the mid of 100, 3^n, 1000 is used with the following setting for EmbedMultipleConfs (pruneRmsThresh=0.2, randomSeed=1, useExpTorsionAnglePrefs=True, useBasicKnowledge=True). Following this workflow, 219,834 calculations terminated normally while 25,016 encountered some form of automatically detected error – most frequently the presence of an imaginary frequency or a failure to determine a suitable initial 3D conformer from the SMILES embedding. Before additional data quality checks, these newly added calculations describe 38,277 new small molecules and 199,209 unique BDE values. Augmenting our original efforts[6, 7] with these values results in a total of 509,740 DFT computed molecular enthalpies and 531,244 unique homolytic BDE values.

A linear model was used to detect outliers in computed absolute enthalpy values by regressing against the element counts (including explicit Hs) as independent variables. The Inner Quartile Range (IQR) was tabulated for the residuals, and calculations that were more than three times the IQR from the upper quartile were removed. This technique finds molecules with

exceptionally high enthalpies for their given molecular composition, typically implying convergence to a particularly unstable conformation. Of the 509,740 enthalpy calculations, 4,309 were removed as enthalpy outliers with this method. The scatter plot for detection of outliers with abnormally large contributions from ΔZPE is shown in (Fig. S1B). The resulting B*DE-db2*, dataset is shared openly and hosted on GitHub at https://github.com/patonlab/BDE-db2. The dataset is available in the folder titled Dataset/bde-db2.

All additional data for studies involving test set 2 and test set 3 can be found in the following GitHub location https://github.com/patonlab/BDE-db2/Datasets

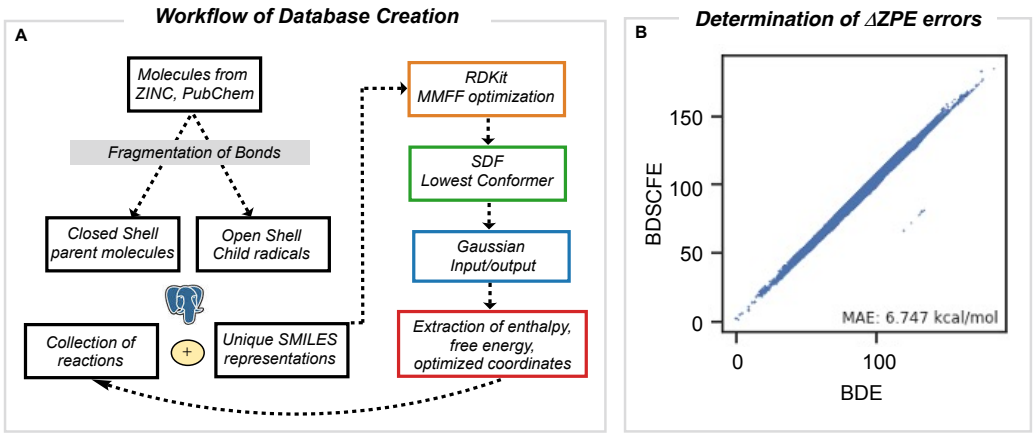


Figure C.S1. *(A) Workflow for the construction of an updated bond dissociation enthalpy database containing halogenated species. (B) Plot of dissociation energy (BDSCFE) vs enthalpy (BDE) exposes errors due to large and unphysical changes in ZPE for a given dissociation, for which the reaction data is removed.*

C.2 Breakdown of types of bond dissociation in the *BDE-db2* dataset

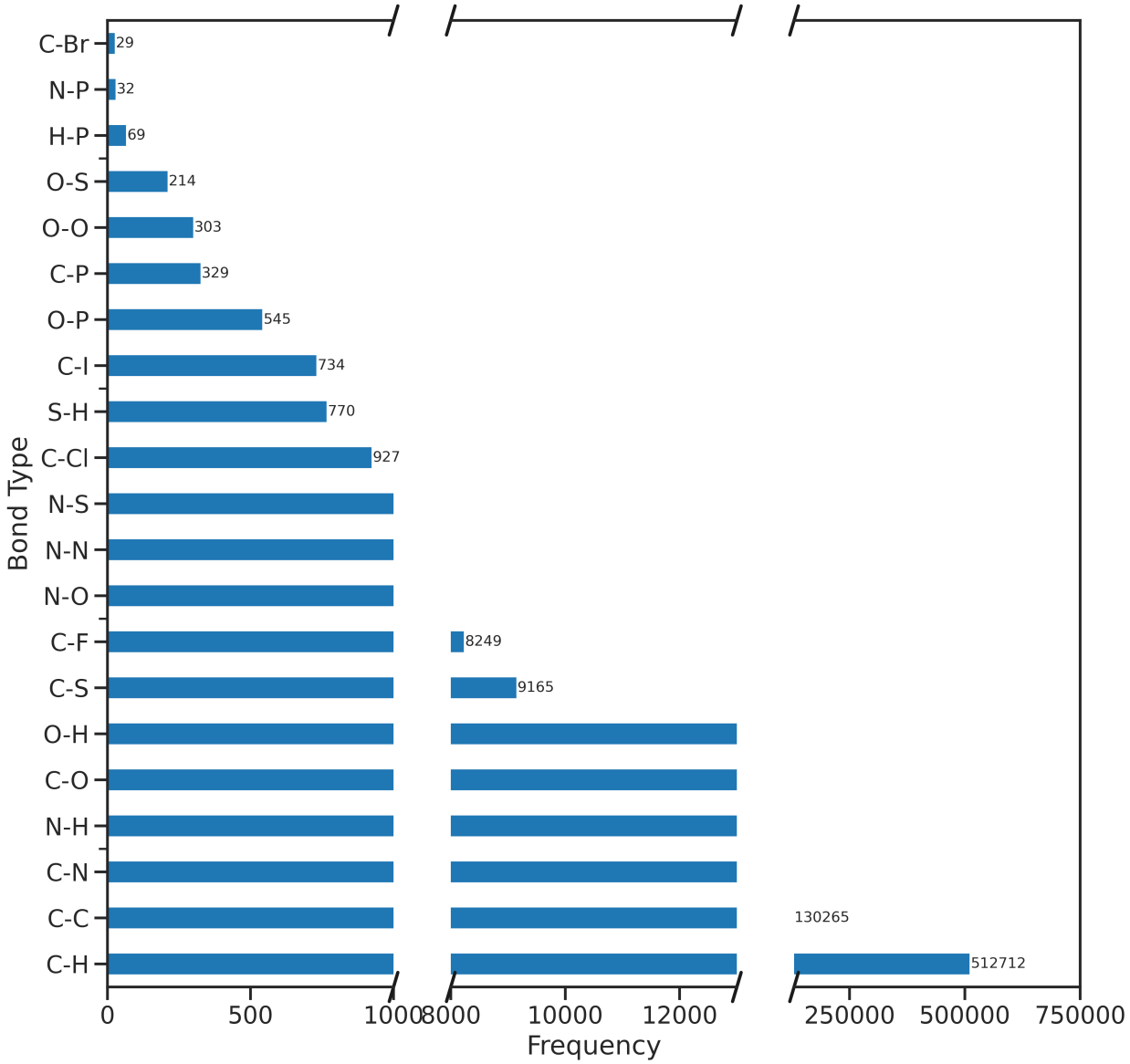


Figure C.S2. *Frequency counts of all dissociation bond-types greater than 25 present in BDE-db2.*

C.3 Prediction accuracy of newly added bond types in held-out test set

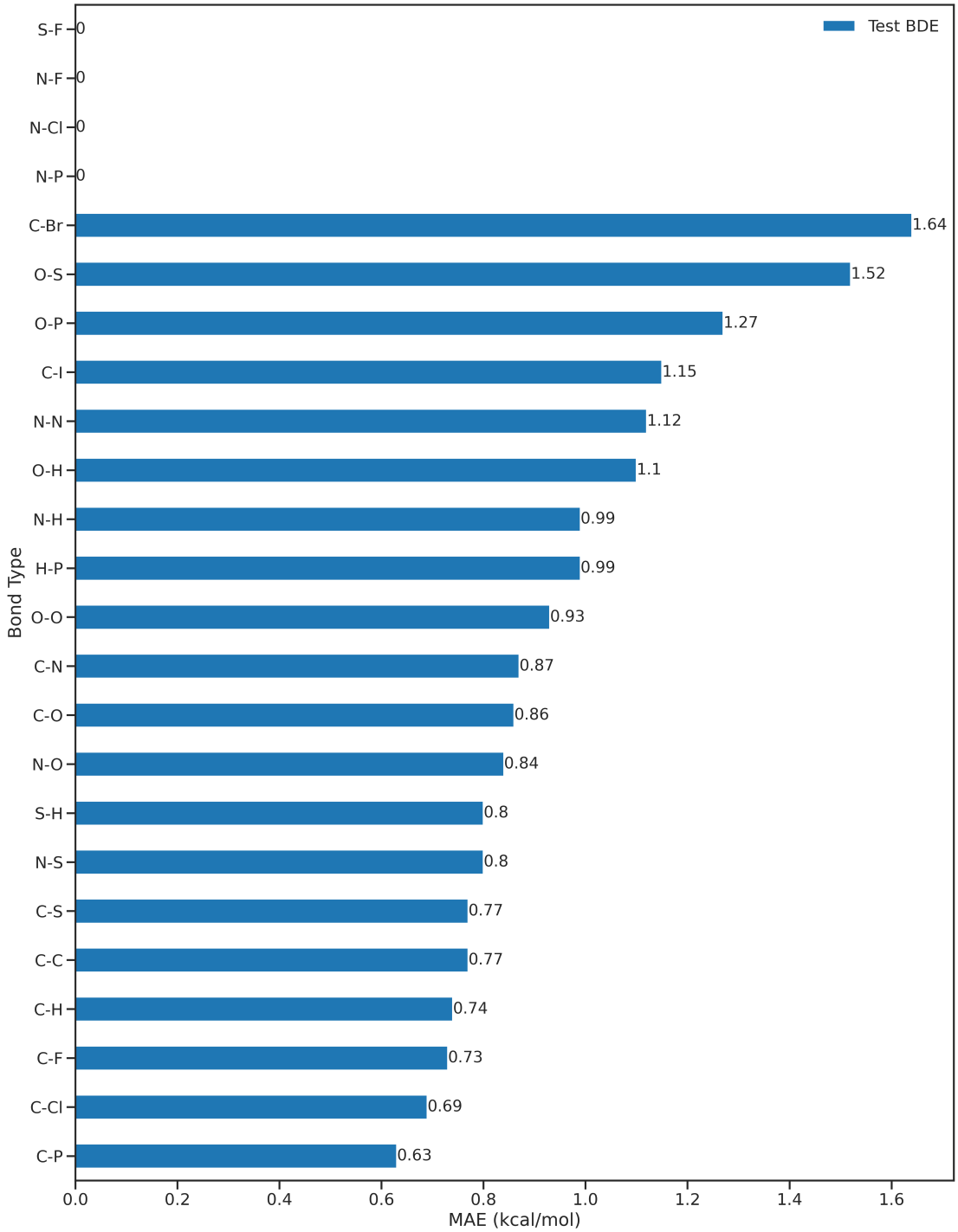


Figure C.S3. *Prediction accuracy relative to DFT oracle for each bond type in held-out set set with Model 1.*

C.4 Model 1: Performance on an external test set of halogenated heterocycles

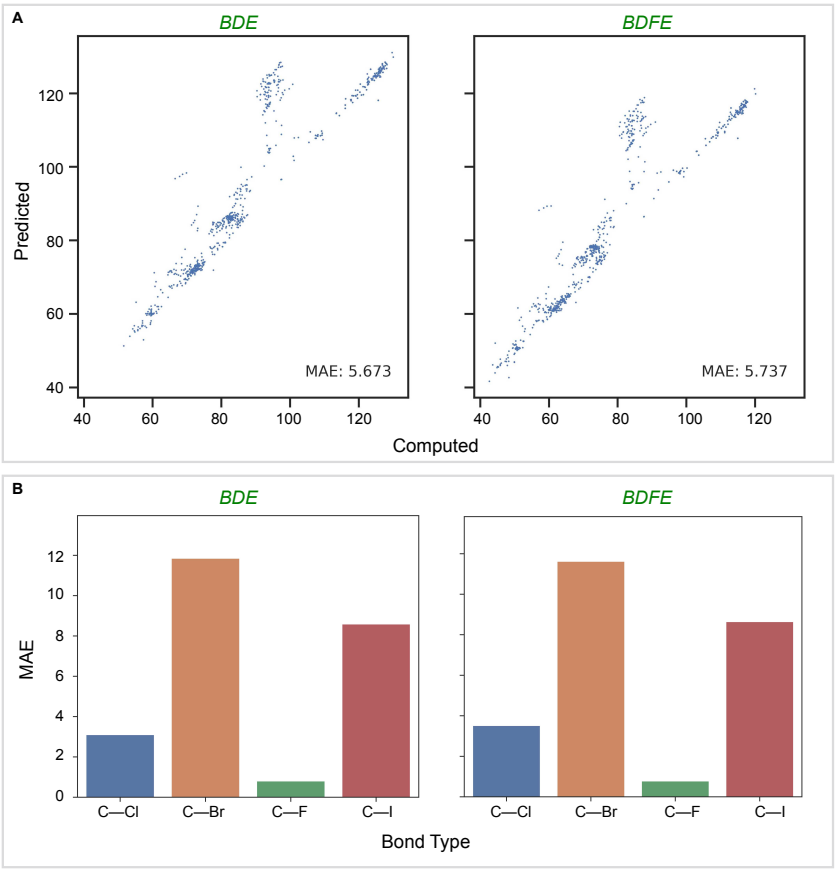


Figure C.S4. *(A) Parity plots for BDE and BDFE prediction (kcal/mol) for aryl halides with Model 1. (B) Spread of errors for aryl halides according to bond type with Model 1.*

C.5 Dataset composition: *BDE-db2* vs. halogenated heterocycle test set

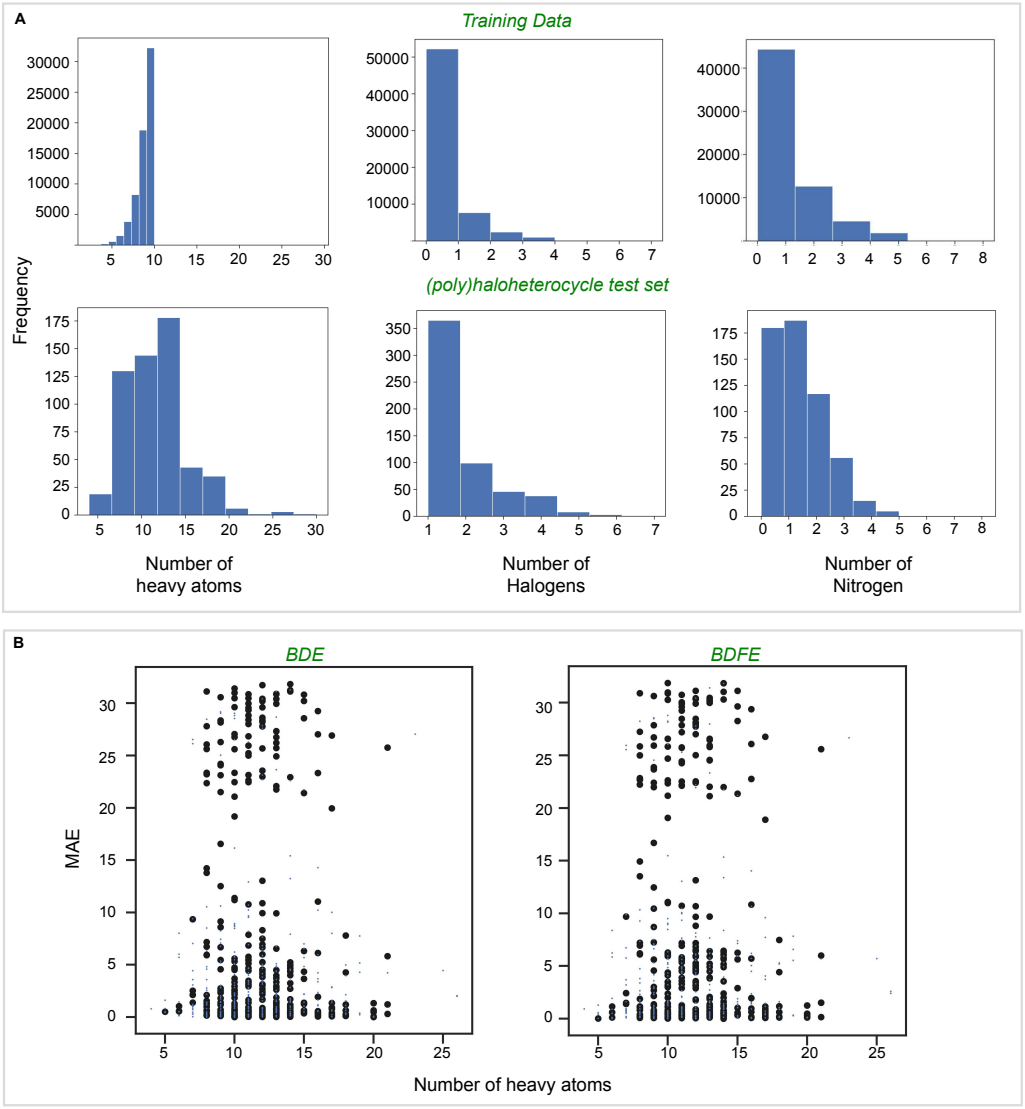


Figure C.S5. *(A) Comparison of the atomic composition of the model training set and aryl halide external test set. (B) The presence of multiple halogens in a molecule is associated with larger prediction errors (black dots represent molecules with more than one halogen).*

C. 6 Model 2: Performance on *BDE-bd2* and an external test set of halogenated heterocycles

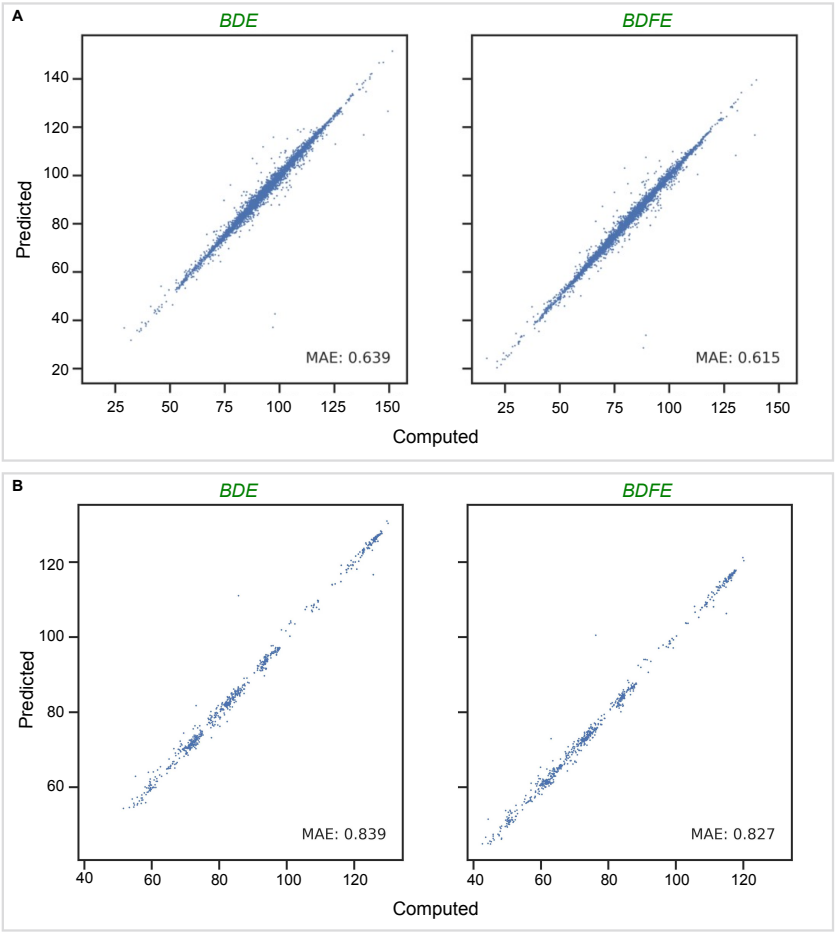


Figure C.S6. *Parity plots for prediction of BDE and BDFE (kcal/mol) with Model 2 for (A) held-out test set and (B) external test set of halogenated heterocycles.*

C.7 Composition of the polyhaloalkyl test set

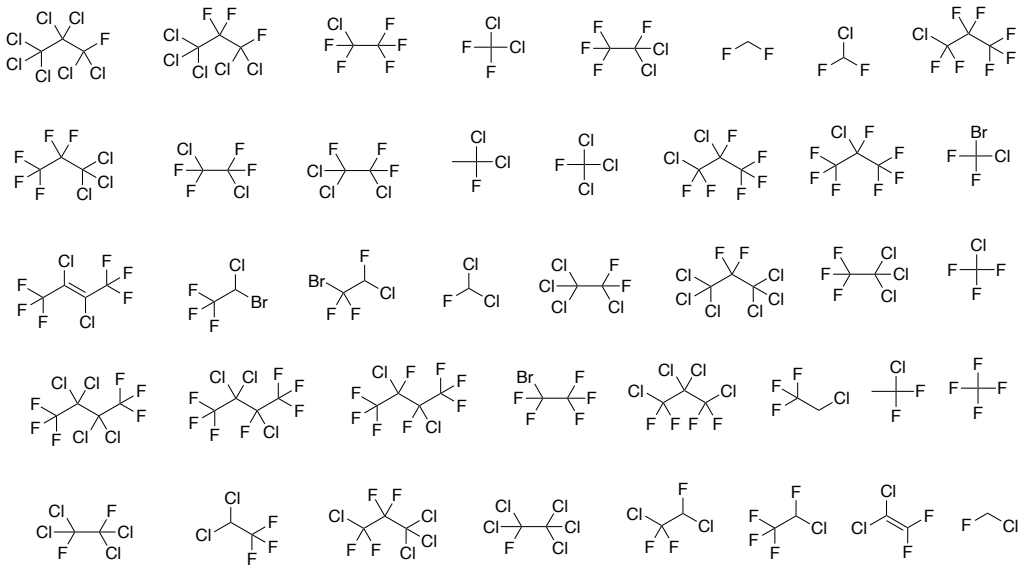


Figure C.S7. *Molecules in the polyhaloalkyl test set.*

C.8 Model 2: Performance on polyhaloalkyl test set

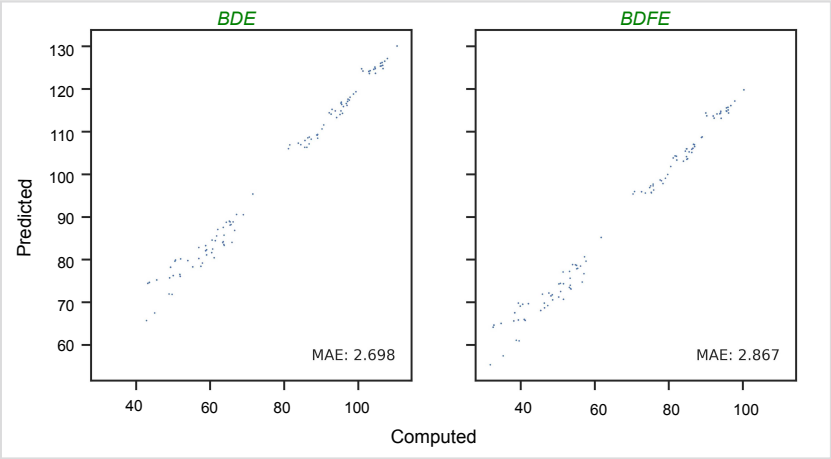


Figure C.S8. *Parity plots for prediction of BDE and BDFE (kcal/mol) for polyhaloalkyl test set with Model 2.*

C.9 Polyhaloalkyl molecules added to training data

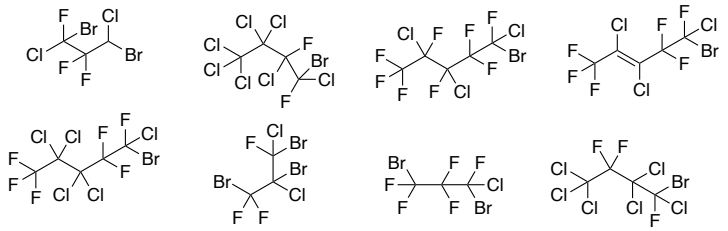


Figure C.S9. *Molecular structures added to train Model 3. N.B. The graph representation used does not distinguish between (R)- and (S)-stereogenic centers, and so configuration at these centers is not shown.*

C.10 Model 3: Performance on *BDE-db2*, halogenated heterocycle, and polyhaloalkyl test sets.

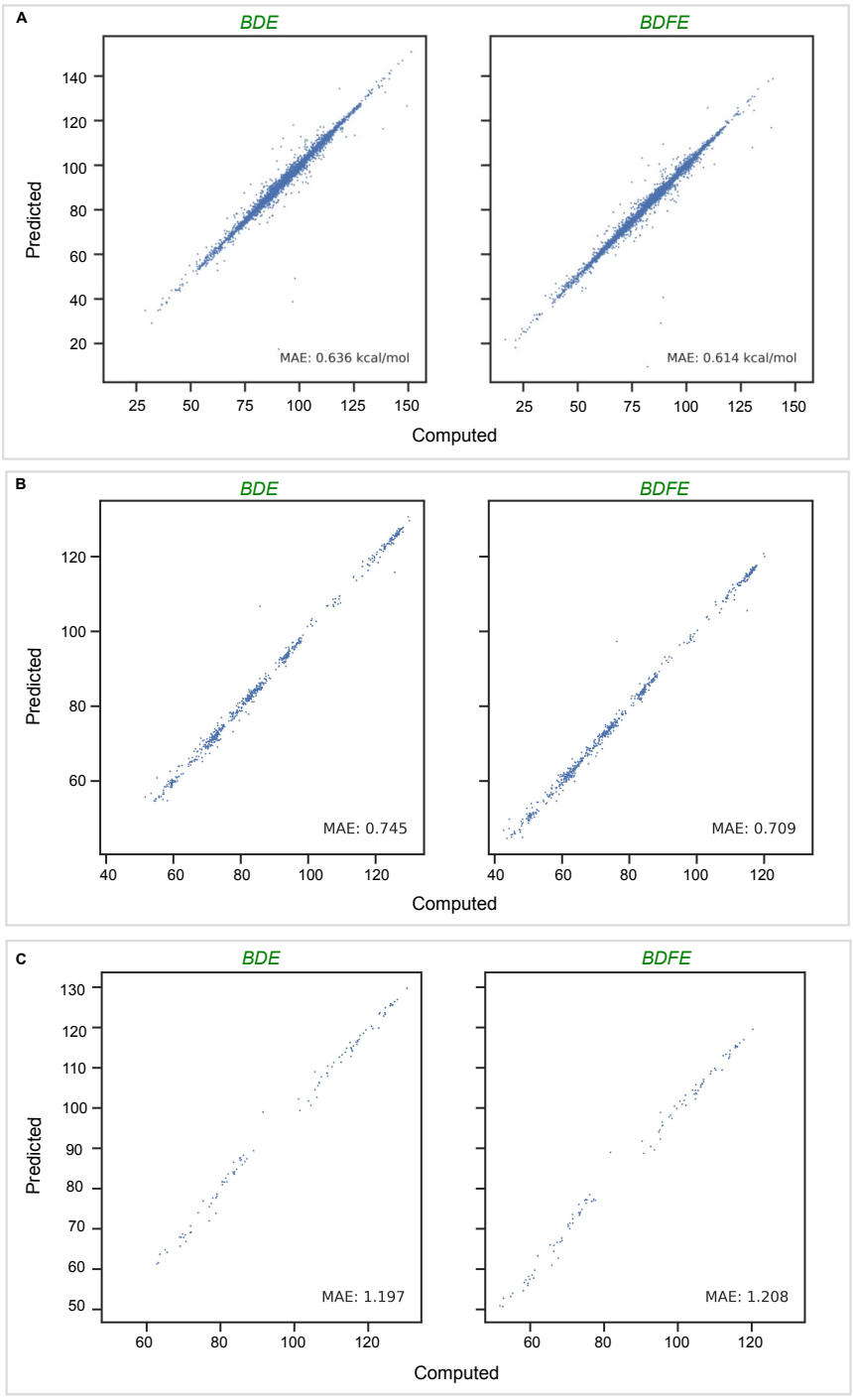


Figure C.S10. *BDE and BDFE predictions (kcal/mol) obtained with an improved model, Model 3: (A) held-out test set, (B) halogenated heterocycles and (C) polyhaloalkyl test set.*

C.11 Additional Details on comparison of traditional cheminformatics features and QM features.

For the models developed with Random Forest the following input and parameters were utilized. The inputs of fingerprints were created using Morgan Fingerprints with a radius of 3 and 512 bits defined around the bond of interest by specifying the atoms involved in the bond. The hyperparameters for modelling with random forest (RandomForestRegressor) was scanned using RandomizedSearchCV. The hyperparameter search include n_estimators = [100,200,300,400], max_features = [1,3,5,7, 'auto'], max_depth = [15, 10, 100, 1000]. Each model on the learning curve is optimized to get the respective hyperparameters for 10 different runs. The best parameters is chosen across the 10 runs to test on the held out test set 1. For graph neural network models with added QM description of bond lengths, the bond lengths were curated for each bond from the respective DFT calculation. The RBFExpansion bond length (dimension of 128) is concatenated with the tokenized embedding of the bond state built from RDKit features. This updated bond state is utilized in the message passing operation in the graph neural networks. The newly developed model with the QM features was test on the held-out test set 1.

REFERENCES

1. Zhao, Y.; Truhlar, D. G., How Well Can New-Generation Density Functionals Describe the Energetics of Bond-Dissociation Reactions Producing Radicals? *J. Phys. Chem. A* **2008,** *112*, 1095-1099.

2. 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.

3. 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 Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16 Rev. C.01*, Wallingford, CT, 2016.

4. Riniker, S.; Landrum, G. A., Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Info. Model.* **2015,** *55*, 2562-2574.

5. "RDKit: Open-source cheminformatics. https://www.rdkit.org".

6. St. John, P. C.; Guan, Y.; Kim, Y.; Kim, S.; Paton, R. S., Prediction of organic homolytic bond dissociation enthalpies at near chemical accuracy with sub-second computational cost. *Nat. Commun.* **2020,** *11*.

7. St. 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.

APPENDIX D: SUPPLEMENTARY MATERIALS FOR CHAPTER 5

D.1. Methods

D.1.1 Calculation and validity analysis of redox potentials

Full DFT estimation of the adiabatic (i.e., including the effects of geometric relaxation and vibrational zero point energy) ionization energy (IE) and electron affinity (EA) for a given radical takes hours per candidate, and requires three separate geometry optimizations to obtain standard-state Gibbs energies of the neutral radical and both anionic and cationic closed-shell species. We further obtain oxidation potential (OP) and reduction potential (RP) values in Volts by referencing the standard-state Gibbs energy changes to the absolute potential of the SHE.[79] Gaussian 16[80] was used for all DFT calculations with a default ultra-fine grid for all numerical integration. The primary database of redox potentials was built using the M06-2x/def2-TZVP level of theory by separately optimizing the neutral, oxidized, and reduced radical species. The calculations were performed using the SMD solvation model with a water solvent at 298K.[55] The same initial structures were used for all three calculations and were taken from previous calculations performed in the gas phase.[58] Where an initial relaxed structure is not available, a single lowest-energy conformer is found using the MMFF forcefield in RDKit, as described previously.[58] Iodine-based molecules in the experimental redox benchmark were optimized with the LAN2DZ basis set in combination with 6-31G(d,p) and 6-31g+(d,p).

An automated workflow was developed to check optimizations for convergence by ensuring the absence of imaginary vibrational frequencies and that all bond lengths remained within 0.4 Å of the sum of their covalent radii. Additionally, molecules were inspected to see whether new bonds were formed during optimization, as this often led to difficult-to-predict

redox potentials. Atom adjacency matrices were used to determine if any two atoms were closer than 1.3 times the sum of their covalent radii, and these molecules were removed from the training dataset. This primarily occurred during oxidation, as 8,566 oxidized molecules, 602 reduced molecules, and 177 neutral radicals were removed from the database in this fashion.

$$\text{Reduction Potential (V)} = \frac{G(R\cdot) - G(R^-)}{F} - E^{\ominus}\left(\frac{H^+}{H_2}\right)_{\text{abs}}$$

$$\text{Oxidation Potential (V)} = \frac{G(R^+) - G(R\cdot)}{F} - E^{\ominus}\left(\frac{H^+}{H_2}\right)_{\text{abs}}$$

Spin contamination was checked by looking at the expectation value of the total spin, $\langle S^2 \rangle$. Radicals were expected to have an $\langle S^2 \rangle = 0.75$, and a handful of optimizations were discarded where spin contamination resulted in an $\langle S^2 \rangle > 0.8$. Anions and cations were assumed to adopt a closed-shell singlet state, with an $\langle S^2 \rangle \sim 0$. To improve the consistency of the dataset, open-shell anions and cations with $\langle S^2 \rangle > 0.25$ were removed.

D.1.2 Training the surrogate objective models

Two separate machine learning models were developed to predict quantum mechanical properties as a function of a candidate radical's SMILES[81] notation. The first model predicts spin delocalization and buried volume on each heavy atom in the molecule. The second model predicts the radical's oxidation and reduction potential (in V relative to SHE). SMILES strings were first converted to a graph representation using the nfp[82] and RDKit[83] python libraries. Atoms and bonds were classified depending on features determined via RDKit. For atoms, this included their atomic type, chirality, presence in a ring, number of heavy atom neighbors (degree), aromaticity, number of neighboring hydrogens, and presence of a formal radical center.

For bonds, this included the atom types of the joined atoms, the bond type (single, double, aromatic), and presence in a ring, and Z/E stereochemistry (if present). The GNN edges are directional, and therefore two graph edges are created for each bond in the molecule, one pointing from atom A to atom B and another pointing from atom B to atom A. Each model consisted of a GNN with a similar core structure depicted in Fig. 2 and further detailed in Extended Data Figure 5. The GNN generates representative embeddings at the atom, bond, and global level by passing the initial features through a series of message blocks.[84] In the stability prediction GNN, the final atom feature vector is reduced to two output predictions for each atom's buried volume and spin density. In the redox GNN, the final global feature vector is reduced to two output predictions for the reduction and oxidation potentials. Both models are trained with a batch size of 128 molecules for 500 epochs over the training data, using the AdamW optimizer with an initial learning rate of 1E-4, decayed by 1E-5 each update step. The weight decay was set to an initial value of 1E-5 (1E-6 for the redox model) and was decayed by 1E-5 each update step.

Learning curves were generated by restricting the training set to a random subset of examples while keeping the same validation throughout the different models. Models were trained as described above using a fixed number of gradient updates (equivalent to 500 passes over the entire training dataset), and the performance on the held-out validation set was recorded. To improve predictive performance in the region of high-reward radicals, 3978 high-scoring molecules from previous RL runs were optimized with DFT and added to the training data for the surrogate objective functions in the final RL iteration. These molecules ranged in size between 9 and 15 heavy atoms with a maximum stability score of 109.7. DFT confirmation of

the final 1078 candidates could be used to augment this training set of high-scoring molecules in subsequent iterations.

D.1.3 Details of the reward function

To find radicals that meet all the desired criteria, predictions and desired ranges for multiple properties were synthesized into a scalar reward function. A continuous piecewise linear function was used to convert predictions to a score between 0 and 1, where a 1 was assigned if the predicted was inside the desired range, 0 outside, and a linear transition between the two scores if the prediction was near the boundary (with width equal to one sixth the width of the desired region).

The overall reward function was then composed by summing over individual scores from different properties

$$\begin{aligned} reward = {} & 50\,(1 - \max s_i) + 100\,BV_j + 25\,\mathrm{window}\big(R_{pred}, [-.5\,V, 0.2V]\big) \\ & + 25\,\mathrm{window}\big(O_{pred}, [0.5\,V, 1.2V]\big) + 25\,\mathrm{window}\big(R_{pred} - O_{pred}, [1\,V, 0.2V]\big) \\ & + 25\,\mathrm{window}\big(BDE_{pred}, [60, 80]\big) \end{aligned}$$

where s_i represents the predicted fractional spin on atom i, and BV is a vector of predicted buried volumes. The reward function was constructed to place approximately equal weight between the stability score (including spin and buried volume contributions, typically near 100 for highly stable radicals) and the remaining BDE (in kcal/mol) and redox requirements.

D.1.4 Description of the molecular action space

Beginning with the initial state of a single carbon atom (i.e., methane after adding implicit hydrogens), possible actions were enumerated following a series of expansion and filtering steps.

First, all possible tautomers of the given starting molecule were considered as possible starting states.[61] From each starting state, a new bond was added between an atom in the molecule and a second atom, either already in the molecule (forming a ring) for an unbonded C, N, O, or S atom. New molecules were generated for every possible atom pair and bond type (single, double, or triple) for which valency rules were satisfied. From this set of all possible next actions, molecules were filtered according to several ring, saturation, and synthetic accessibility criteria,[85] including restricting molecules to a maximum SAScore of 4.0. The action space was then further expanded by enumerating all possible stereochemical configurations of the starting molecule, followed by a reduction to canonical tautomer forms. Next actions were then de-duplicated by SMILES string.

Our action space differs from previously described molecular 'environments' in several ways. Unlike the environment proposed in MolDQN,[31] our approach results in a directed acyclic graph (DAG) over possible molecules by eliminating the possibility to remove atoms and bonds from molecules under construction. This DAG property prevents the search from searching cyclically and guarantees forward progress when building a radical. It also makes learning the value function easier by eliminating conflation and cross-contamination from cyclical paths in the search graph. Our approach is similar in that respect to the generation environment proposed by You *et al.,*[30] where atom and bond additions are guided by a policy network. Unlike in You *et al.,* we do not decompose the action into a node selection and link selection step, and instead only evaluate policy predictions once both the atom and bond type have been chosen. This allows us to easily filter the action space based on valency rules and dramatically reduces the number of invalid molecules constructed by the algorithm. It additionally allows us to easily consider additional modifications of the action space, including stereochemical enumeration

(e.g., at tetrahedral carbon atoms and of double bonds), tautomerization, and synthesizability considerations.

D.1.5 Details of the RL algorithm

The RL optimization was performed using the rlmolecule library[86] (github.com/nrel/rlmolecule), which implements the AlphaZero approach for molecule and material design. In this study, the RL agent learned to select from a parametric action space, where the molecular structures resulting from possible next actions were passed through a trainable policy GNN. The policy GNN had a similar structure to that used for redox prediction (Fig. 2B), using only 3 message passing layers and a feature dimension of 64. The policy model was trained with the ADAM optimizer with a learning rate of 1E-3 and a batch size of 32 positions. Within each batch, the policy model is presented with a molecule state and a list of potential next actions from a recently played MCTS rollout. The policy model is trained to simultaneously predict the actual visitation frequency from MCTS, as well as the outcome of the resulting molecule rollout (0 or 1 as scored via ranked rewards). Starting at the root methane state, molecule rollouts consisted of conducting 250 MCTS samples (or for a maximum of 30 seconds) and selecting the subsequent molecule state with probability proportional to the softmax of the visit counts. This procedure is repeated until a terminal state is selected.

The wall time limit was imposed to mitigate the effects of problematic regions of chemical space where the number of possible next actions per molecule, and therefore the time required to enumerate them, vastly outnumbered typical molecules. This was typically encountered with molecules with many possible tautomers and resulted in rollouts being added to the replay buffer that used an outdated version of the policy model.

Communication between the policy network training script and the MCTS rollout workers is handled through a shared filesystem and a PostgreSQL server. Policy checkpoints are previously written to a shared filesystem location, which is checked at the beginning of each rollout by the workers. Final statistics and molecule reward calculations are then written to the shared SQL database. The policy training script in turn selects the 256 most recent rollouts each training epoch, with each epoch consisting of 100 training steps.

D.1.6 Synthesizability prediction

Retrosynthetic routes are predicted using the ASKCOS web interface tool (https://askcos.mit.edu) using the tree builder module. Settings were chosen to match those used in a previous study evaluating the synthesizability of generative models.[50] Specifically, the maximum tree depth was limited to 9 steps, the maximum branching ratio is set to 25, a maximum wall time of each expansion is limited to 60 s, a maximum reagent cost of $100/g, 1000 max templates, and a maximum target probability of 0.999. Employing these settings along with no defined banned chemicals and reactions for the radical 1(Fig. 5A), we obtained a total of 43 routes, containing 90 chemicals and 970 reactions. Computation time for each parent molecule's retrosynthesis tree took approximately one hour.

D.2 Extended Data Figure Captions

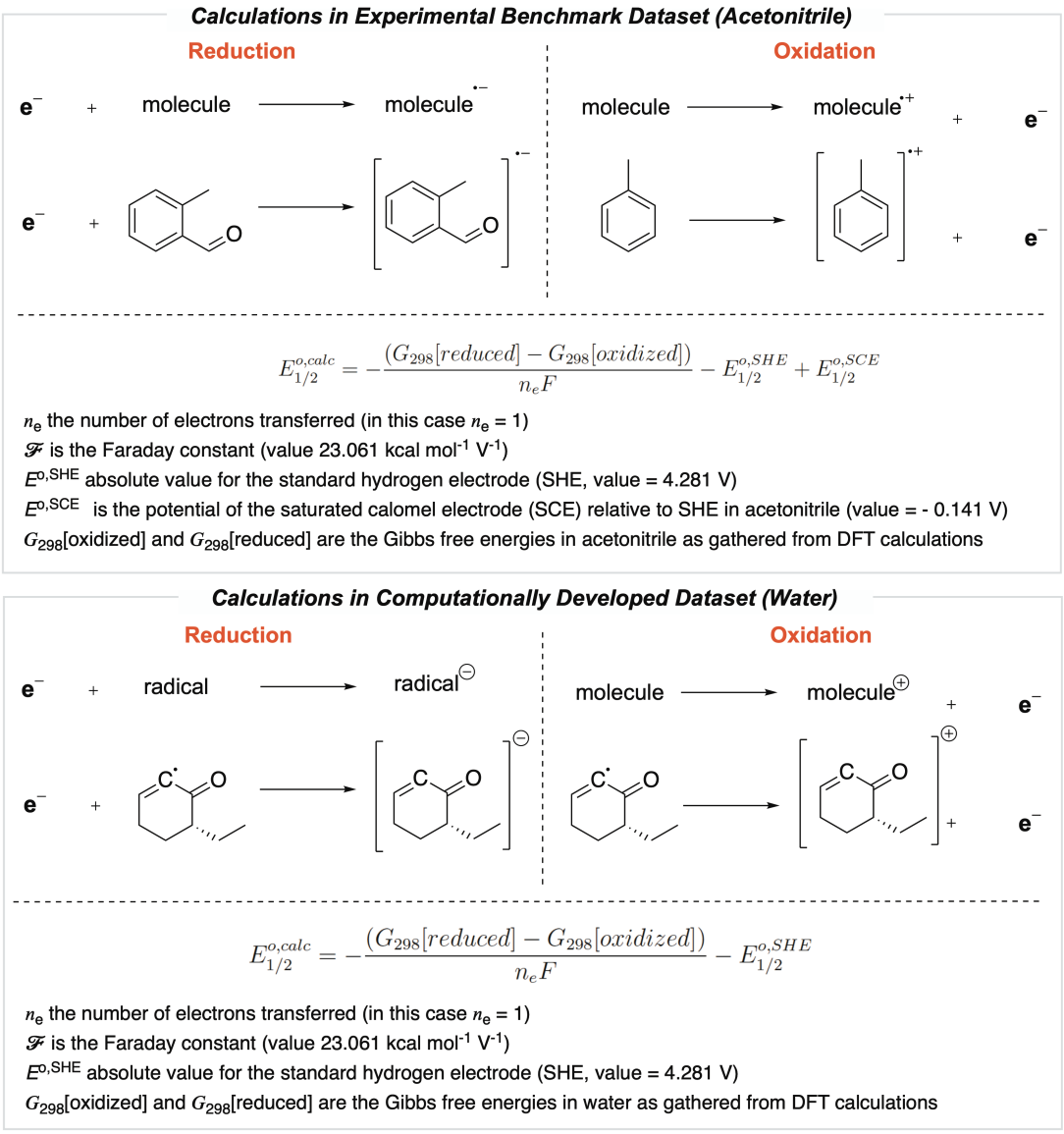


Figure D.S1. *Methods used in calculation of redox potentials of the reduction and oxidation reactions for the benchmark experimental dataset (top) and the computational dataset in water (bottom). The respective equations used to determine the redox potential are also shown.*

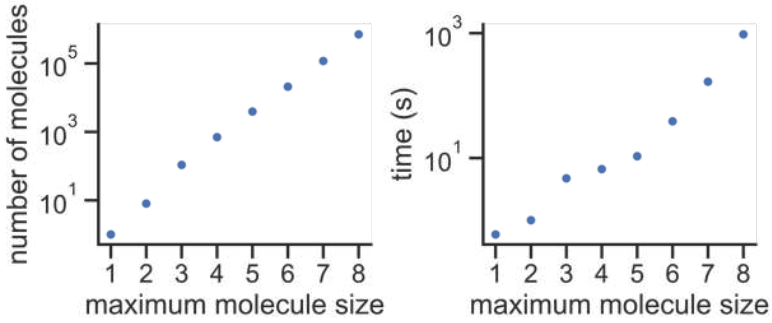


Figure D.S2. *(left) Maximum number of molecule states versus maximum molecule size for the search tree described in this study. Extrapolating from these results yields approximately* 1.9×10^9 *valid molecules with 12 or fewer heavy atoms. (right) Computational time required to enumerate the search space as a function of maximum molecule size. In addition to requiring more evaluations, larger molecules require additional computational time to check for a valid 3D embedding and to enumerate possible stereoisomers, and the time required for larger molecules may grow faster than simply exponential. An exponential fit to the last four datapoints indicates that a full enumeration of the 12 or fewer heavy atom search space would require 17.25 days. In addition to the time required to enumerate all candidates, a high-throughput screening would require evaluation of the final molecules with the reward function. This would require three separate neural network evaluations (which could be called in parallel) to estimate radical stability, redox potential, and X-H bond strength, and would likely add several days to the overall computational cost.*

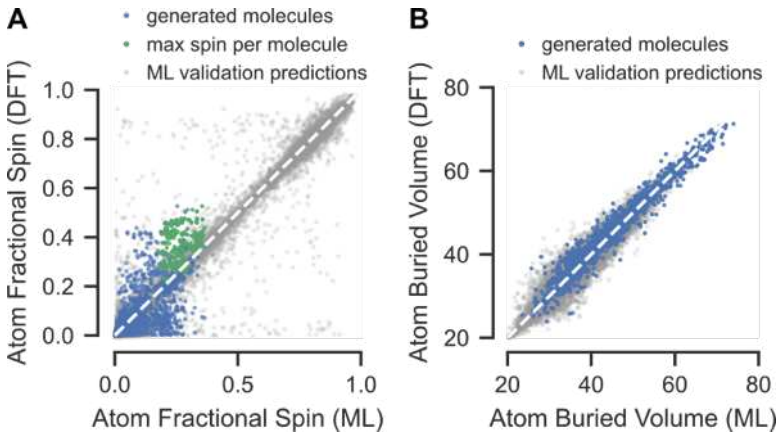


Figure D.S3. *Plot of ML versus DFT stability sub-scores for top-performing RL candidates. (A) RL-generated molecules tend to have highly distributed electrons, making the pre-diction of their spin difficult. (B) Buried volume predictions for these molecules are in line with expected errors from the validation set radicals.*

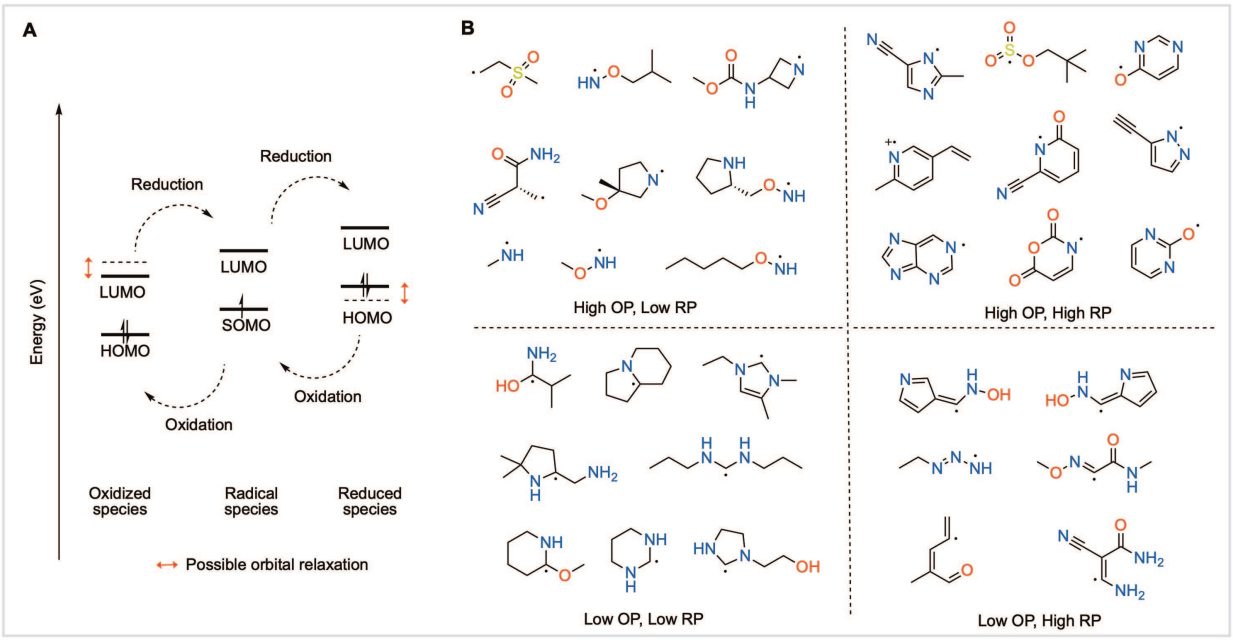


Figure D.S4. *(A) Molecular orbitals involved in the three redox states that are leveraged in a symmetric battery candidate. (B) Example structures of radicals with high and low oxidation potentials (OP) and reduction potentials (RP).*

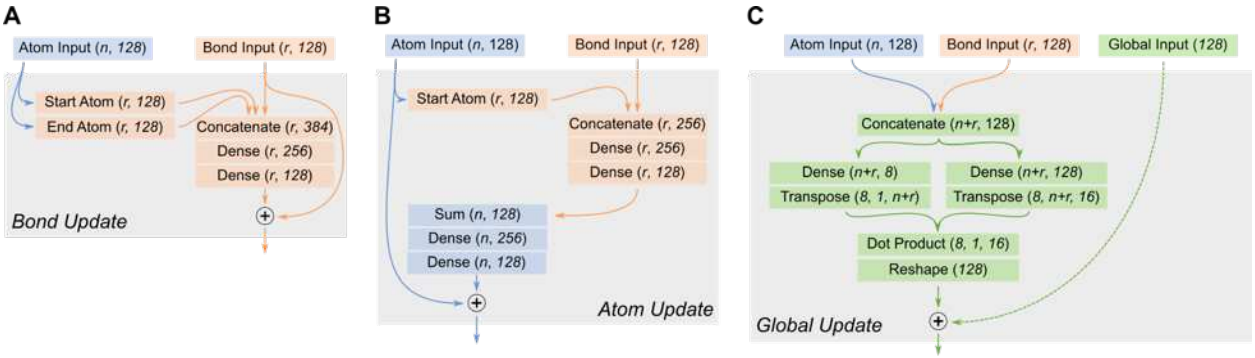


Figure D.S5. *Additional details for the GNN update blocks. (A) The bond update block concatenates the features of the source atom and the target atom together with bond's starting features and passes them through two dense layers. The output of this network is summed with the input bond features in a residual fashion. (B) The atom update block separately concatenates the atom's features with the features of each bond targeting the given atom and passes them through two dense layers. The output of these layers is summed over the incoming bonds and further passed through a series of dense layers to form an updated atom state. This updated state is summed with the previous atom feature vector in a residual fashion. (C) The global update block concatenates all bonds and atoms in the molecule and performs a multi-head dot product attention operation. This operation extracts a single feature vector for the entire molecule in a manner that is in variant to the ordering of atoms and bonds in the graph.*

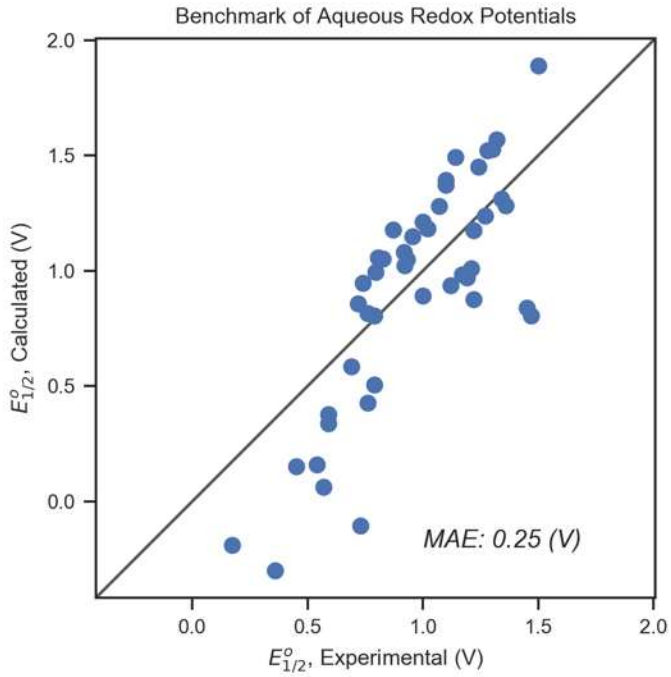


Figure D.S6. *Comparison of experimental redox potentials in water against computed redox potential using M06-2x/def2-TZVP methodology for a set of 46 molecules.*