

DISSERTATION

EMERGENT-SCALE PREDICTION OF MOLECULAR PROPERTIES WITH CHEMICALLY-
INFORMED NEURAL NETWORKS

Submitted by

Sabari Kumar

Department of Chemistry

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2026

Doctoral Committee:

Advisor: Seonah Kim

Justin Sambur

Peter St. John

Shrideep Pallickara

Copyright by Sabari Kumar 2026

All Rights Reserved

ABSTRACT

EMERGENT-SCALE PREDICTION OF MOLECULAR PROPERTIES WITH CHEMICALLY INFORMED NEURAL NETWORKS

Machine learning (ML) models have emerged as powerful tools for scientific discovery. Leveraging modern computing hardware allows for the generation of large data sets which can be used to train accurate, generalizable models to predict the physico-chemical behavior of molecular systems. However, the predictive power of such models is bounded by how chemical knowledge is encoded within the model. Unlike general-purpose neural networks, which treat molecular data as abstract numerical arrays, chemically-informed architectures incorporate physical and structural constraints directly into their mathematical form. This work examines the design principles behind such architectures and traces their application across different intra- and intermolecular scales of organization.

At the level of individual molecules, graph neural networks represent atoms as nodes and bonds as edges, propagating information through the molecular graph in a manner that mirrors molecular topology. We will examine the results of different parameterizations of this information in the context of predicting bond dissociation enthalpies and adiabatic singlet-triplet excitation energies, and present novel analytic tools which connect learned model behaviors with chemical intuition.

We will demonstrate how advances in ML algorithm design can enable the accurate prediction of properties which are not possible to simulate through conventional means. By adapting techniques from computational topology and geometric deep learning, we construct a novel neural network architecture capable of predicting protein solubility from predicted protein structures. We show that the model’s internal representations of protein structure align with those obtained through conventional molecular dynamics simulations.

We then extend the single molecule GNN approach to predict the interaction behaviors of molecular species with their surroundings. We show that learned mixing operators mirroring established chemical intuition can reliably predict nonlinear blending behavior, and introduce a new tool inspired by operator symmetrization to analyze this interaction.

Lastly, we present ongoing efforts to extend our study of interaction behavior to model the interactions of molecular conformer ensembles with their surroundings. We develop a preliminary foundation model for macrocyclic peptides, incorporating explicit learned reparameterization of conformational Boltzmann weights to accurately map conformational ensemble information to peptide properties.

Taken together, these results reflect a common organizing principle: the most transferable and physically meaningful models are those whose structures recapitulate the interactions that govern chemical behavior.

ACKNOWLEDGMENTS

As I come to the end of my PhD journey, my gratitude goes out to the incredible mentors I have been blessed with. First and foremost, my wholehearted thanks go out to my mentor, Dr. Seonah Kim. I have learned so much from working with Seonah; I am profoundly grateful for her mentorship and patience, and for giving me the leeway to learn to teach myself and to learn from my mistakes. Early on in my PhD, Seonah taught me a valuable lesson: she told me that if I knew everything already, then there would be no need for me to be in a PhD program. I think about those words regularly, and I'm sure I will remember them for the rest of my life. Seonah taught me how to be a scientist, and I am truly grateful for her guidance.

My thanks also go out to my other mentors: Dr. Yeonjoon Kim, Dr. Peter St. John, Dr. Rob Paton, and Dr. Tony Rappe, who have all given me invaluable advice over the course of my PhD. I learned how to do real science by modeling Yeonjoon; I still consider him a role model and the Platonic ideal of a mentor. Yeonjoon's skills at guiding others through rubber duck debugging are second to none, and I hope to one day be his equal. Every time I talk to Peter, he has given me excellent advice, from career progression to debugging neural networks. During my oral exams, Peter made a comment that running a simulation or training a model is an experiment, and that I should be treating my work as such. That advice completely changed the way I approach my work, and I will always be grateful to Peter for it. Over the course of my PhD, there have been 4 or 5 times where I have felt truly stuck. Every time, I have walked into Tony's office and left 15 minutes later with a solution. Once, as I was

struggling with understanding the group theory behind equivariant neural networks, I walked into Tony's office, at my wits' end. Tony listened to my word salad as I tried to explain my confusion, pulled a book, Shoon Kyung Kim's Group Theoretical Methods and Applications to Molecules and Crystals, from his shelf and told me that the answer to my question was probably contained within. Indeed, the first chapter cleared up my confusion. It remains one of the most impressive displays of knowledge I've ever seen, and I hope one day I can get to that level. Rob has been a patient, long suffering sounding board to my harebrained neural network schemes; I remain ever grateful for his patience in helping me think through ideas and helping me keep my work focused and directed. I'm also grateful for the support of members of the Theory Suite at CSU. Thank you all for the great discussions and the good times.

I also would like to thank my undergraduate advisor, Dr. Roberto Garza-Lopez. As a twenty-year-old, I wanted to be just like Roberto when I grew up - I still have that dream today. Working with Roberto sparked my love for interdisciplinary computational chemistry, and Roberto's volunteer and mentorship work are a continual source of inspiration. My profound thanks go out to Roberto for always believing in me, even when I didn't; I wouldn't be here today without him.

My thanks go out to my friends and family, whose patience and support through the late nights and long days have kept me going. Thanks to my parents, Amma, Acha, Anna's parents, Laura and Bruce, and my sister, Arathi, who have always supported me and fought for the best for me. My gratitude goes out to my friends- Simon, Heidi, Azissa, Daniel, Ian,

Shaya, Joe, Ian, Bradley, Emma, Carter, Brandon, and many more, for patiently supporting me throughout this process and coaxing me out of my computer cave from time to time.

Last, but certainly not least, my profound thanks and love go out to Anna and Sage, who have always been my biggest supporters. Thank you Anna for always being in my corner, and thanks for supporting me through all of this. I truly couldn't have done it without you. And of course, thanks to Sage, the most interesting dog in the world, for her usually unconditional, sometimes conditional love. I've closed every acknowledgement so far in my PhD with a picture of Sage, and I see no reason to stop now



TABLE OF CONTENTS

ABSTRACT	ii
ACKNOWLEDGMENTS	iv
CHAPTER 1: Chemistry as a Hierarchy of Emergent Scale Behavior	1
1.1: Chapter Overview	1
1.2: Neural Networks for Chemical Property Prediction	7
1.3: References	15
CHAPTER 2: Bond Properties from Atomic Interactions: Predicting Homolytic Ring Opening Bond Dissociation Enthalpies	18
2.1: Chapter Overview	18
2.2: Introduction.....	19
2.3: Results and Discussion	25
2.3.1: Expanding Dataset Size.....	25
2.3.2: Training the ALFABET3 Model and Analysis of Model Performance.....	30
2.3.3: Prediction of Ring Opening Polymerization Sites in Lactone Precursors	37
2.4: Conclusion	42
2.5: References	43
CHAPTER 3: Molecular Properties from Fragment Interactions: Predicting Adiabatic Singlet- Triplet Gaps	51
3.1: Chapter Overview	51
3.2: Introduction.....	52
3.2.1: Exciton Localization Problem in Multichromophore Systems:	54
3.3: Results and Discussion	57
3.3.1: Molecular Fragmentation Algorithm.....	57
3.3.2: Databases Overview and Curation.....	60
3.3.3: Database Analysis and Comparison	64
3.3.4: Architecture Selection and Model Extrapolation	69
3.3.5: Fragment-Based Models	75
3.4: Conclusions	83
3.5: References	85
CHAPTER 4: Protein-Environment Interactions from Constituent Residues: Predicting Protein Solubility.....	91

4.1: Chapter Overview	91
4.2: Introduction.....	93
4.3: Results and Discussion	99
4.3.1: Generation and Validation of Protein Structural Databases	99
4.3.2: Generation of Topological Barcode Embeddings (TopoCoder)	103
4.3.3: Equivariant Protein Solubility Prediction Model (TopoFormer) Training and Evaluation	108
4.3.4: Structural Mapping of Solubility in Carbonic Anhydrase Mutants	115
4.4: Conclusions	119
4.5: References	120
 CHAPTER 5: Solvent-Solute Interactions from Atom and Bond Structure: Predicting Small Molecule Solubility	 126
5.1: Chapter Overview	126
5.2: Introduction.....	127
5.3: Results and Discussion	132
5.3.1: Semi-supervised distillation scheme for developing solubility prediction models	132
5.3.2: Development of graph neural networks for solubility prediction	136
5.3.3: Semi-supervised distillation for self-evolving databases and graph neural networks	138
5.3.4: Analysis of solubility model performance.....	141
5.3.5: Error analysis of solubility prediction	151
5.4: Conclusions	160
5.5: References	161
 CHAPTER 6: Reactive Behavior through Intermolecular Interactions: Predicting Blend Cetane Number	 170
6.1: Chapter Overview	170
6.2: Introduction.....	172
6.3: Results and Discussion	176
6.3.1: Nonlinear Blending Behavior	176
6.3.2: Formulation For Modeling Nonlinear Blending Behavior	178
6.3.3: Baseline Analysis for the Predictive Model Development	180
6.3.4: Graph Neural Networks and Molecular Representations.....	183
6.3.5: Model Training and Evaluation of Model Accuracy	185
6.3.6: Interpretation of the W Matrix Trained for Modeling Nonlinear Blending Behavior	187

6.4: Conclusions	192
6.5: References	193
CHAPTER 7: Interactions of Molecular Ensembles with their Environments:	
Conformationally-Aware Foundation Models for Property Prediction	197
7.1: Chapter Overview	197
7.2: Introduction.....	198
7.3: Results, Future Work, and Discussion	204
7.3.1: Dataset Generation and Conformer Ensemble Processing	204
7.3.2: PECAN Model Design and Training	210
7.3.3: Preliminary Finetuning Studies	213
7.4: Conclusion	218
7.4: References	219
CHAPTER 8: Concluding Remarks and Future Directions.....	225
8.1: References	226
APPENDIX A: Supplementary Materials for Chapter 2.....	227
APPENDIX B: Supplementary Materials for Chapter 3.....	235
APPENDIX C: Supplementary Materials for Chapter 4	296
APPENDIX D: Supplementary Materials for Chapter 5	319

CHAPTER 1: Chemistry as a Hierarchy of Emergent Scale Behavior

1.1: Chapter Overview

Chemistry is a science of scales. The properties of a chemical bond are dictated by the atoms that form it and the molecular environment surrounding it. The properties of a molecule depend on how its constituent atoms, bonds, and fragments interact, fold, and polarize one another. The behavior of more complex chemical systems, such as a fuel mixture combusting, a drug dissolving in the fluid of the intestinal lumen, or a protein folding into a stable, active conformation in the cytosol of a cell, depends on how individual molecules interact with their neighbors. Each level of interaction scale is distinct, and each is incomplete without the cumulative sum of those below it.

This work uses this hierarchy of scale as a design framework for predictive models. The central argument of this framework is that a model trained to predict a molecular property should encode the same scale-based physicochemical relationships that *govern* that property. Bond dissociation energy, the energetic change associated with cleavage of a chemical bond, is determined by local electronic environment; the model should reason about bonds and their atomic contexts. Photochemical reactivity is determined by how sub-molecular chromophores interact; the model should thus reason about the nature of fragments and their interactions. Solubility is determined by solute-solvent interactions; the model should reason about the complete set of molecules, not isolated structures of the

solute or solvent alone. When the unit of representation matches the unit of physical causation, the model learns a representation of chemical behavior that generalizes by virtue of the underlying chemical interactions. When it does not, the model memorizes instead, behaving as a “glorified convex optimizer”.¹

It is this drive towards generalizability which motivates the model-building developments described in this work. A multitude of existing work in computational chemistry focuses on extrapolating the results of experimental or focused computational study to highly specific chemical systems.²⁻⁸ For instance, previous work by Sigman and Doyle used hand-in-hand experimental and theoretical mechanistic studies to identify stereoelectronic molecular structure attributes relevant in determining the yield and stereoselectivity of palladium catalyzed coupling reactions.⁹ A linear regression model was constructed, utilizing these features to accurately predict the yield and stereoselectivity of unseen reactions. The authors use careful parameterization of model input molecules to achieve high predictive accuracy and allow interpretation of model results. However, this model is specific to the Suzuki-Miyura catalytic system the authors focus their efforts on: its predictive behavior does not extrapolate to other types of coupling reactions, such as Heck or Stille couplings.

In this work, we focus on the converse strategy: we seek to build models that generalize across broad swathes of chemical space, sacrificing model simplicity for improved accuracy and expanded domains of applicability. We will make use of neural networks, machine learning models that offer increased parametric flexibility, trading the

need to use domain specific, handcrafted customized inputs for greatly increased input training data requirements and reduced interpretability. We will seek to mitigate the impact of both tradeoffs through creative incorporation of chemical structure into neural network algorithms and the subsequent use of this structure to devise new analytical tools. We will provide a brief introduction to the theory of such models in the latter portion of this chapter.

This work subsequently proceeds in ascending order of scale, from the intramolecular properties encoded by atoms, bonds, and fragments, to bulk-phase mixing behavior emerging from individual molecular properties. Each chapter identifies the physical scale at which the target property is governed, argues for the representation that matches that scale, and evaluates the resulting model against the alternatives.

Chapter 2 begins at the bond level with the prediction of ring-opening homolytic bond dissociation energy (BDE). BDE is among the most local of molecular properties: it is defined by the electronic environment within a few bonds of the breaking site, modulated by molecular strain and electronic conjugation behavior. The ALFABET3 model developed here extends BDE prediction to ring systems by introducing a message-passing scheme that treats molecular topology (and thus ring membership) as a structural primitive rather than an emergent pattern to be learned. On a dataset spanning over 700,000 bonds across diverse organic structures including strained carbocycles and heterocycles, ALFABET3 achieves state-of-the-art accuracy with substantially improved coverage compared to prior models. The trained model is used to rapidly screen proposed lactone monomers for use in ring opening polymerization reactions, and prediction outliers are analyzed using a

combination of techniques from traditional computational chemistry as well as techniques from the machine learning literature. The chapter establishes the core methodological vocabulary — message passing, physics-informed inductive bias, scale-matched representation — that the subsequent chapters build on.

Chapter 3 moves up to the fragment scale, using molecular fragments to parameterize photochemical reactivity. Photochemical reactivity cannot be assigned to individual bonds, nor does it emerge from the molecule as an undifferentiated whole. Rather, it is governed by chromophoric sub-fragments and by how those fragments interact through conjugation, energy transfer, and steric contact. The fragment-based approach developed in this chapter decomposes molecules into chemically meaningful sub-units, trains representations at the fragment level, and aggregates them to predict excited-state properties across a structurally diverse photochemical dataset. The key finding is that fragment-level representations transfer across molecular scaffolds more reliably than whole-molecule fingerprints, precisely because the governing physics operates at the fragment scale.¹⁰

Chapter 4 addresses molecular-level emergence through the lens of protein solubility, whole-molecule property which cannot be predicted from 2D connectivity alone. Solubility depends on the 3D geometry of the folded protein and how the surface interacts with the solvent. The TopoFormer model predicts protein solubility from predicted protein structures by encoding protein shape using persistent homology, a topological representation that captures cavity structure and surface porosity without requiring explicit

atom-by-atom geometry. By combining this shape-based representation with geometric representations, TopoFormer accurately predicts protein solubility from predicted protein structures.¹¹

Chapter 5 crosses into the intermolecular regime with solubility in single-solvent and multi-solvent systems. Solubility, a bulk-phase property, cannot be predicted from the properties of the solute and each solvent independently — it depends on the specific interactions between solute, co-solvents, and the mixture microstructure. Two related models are developed here, both framing solubility prediction as a problem over solute-solvent pairs rather than isolated molecules. First, a model for single solvent solubility is developed, using a semi-supervised knowledge distillation strategy to incorporate data from both experimental and computational studies: the self-evolving database strategy addresses the sparse coverage of solvent-solute chemical space by iteratively expanding the training set through active learning. Across a benchmark spanning diverse solute-solvent combinations, the pair-based GNN substantially outperforms models trained on single-molecule representations, confirming that the intermolecular unit of causation demands an intermolecular unit of representation.¹² This model is then extended to make predictions of solubility in multi-solvent systems.¹³

Chapter 6 completes the intermolecular scale in a physically and practically distinct setting: fuel combustion. Cetane number and sooting tendency are macroscopic outcomes of how fuel molecules interact under high-temperature, high-pressure conditions. The models developed here, physics-informed GNNs trained on carefully curated experimental

datasets, predict these properties for both single-component fuels and multicomponent fuel surrogate mixtures. The combustion setting differs from the solubility setting in the nature of the intermolecular interactions involved, but the representational logic is the same: mixture-level properties require mixture-level inputs. The chapter also demonstrates that physics-informed inductive biases (incorporating known combustion chemistry into the message-passing scheme) improve both accuracy and generalization to out-of-distribution fuel structures.^{14,15}

Lastly, in **Chapter 7**, we revisit the opening argument. Across four scales and six distinct prediction tasks, the models that perform best are consistently those whose architecture reflects the physical scale at which the target property is governed. That result has practical implications for how molecular ML models should be designed and evaluated, and theoretical implications for what it means to learn chemistry rather than fit data. The chapter closes with open problems: cross-scale models that propagate information across levels of the hierarchy simultaneously, uncertainty quantification that respects scale boundaries, and the prospect of autonomous chemical discovery in systems where the scale of governing interactions is itself unknown.

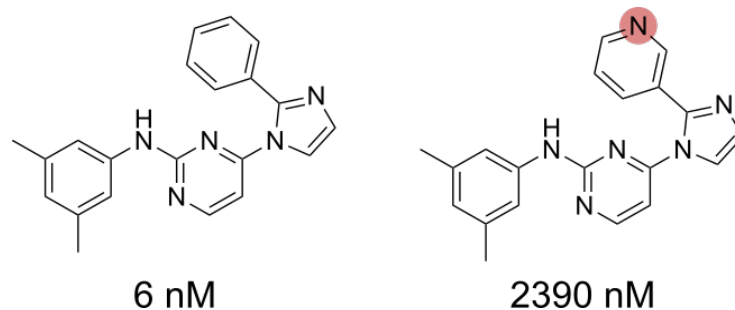


Figure 1.1: Proposed VEGFR-2 inhibitors. The substitution of a nitrogen for a carbon atom in the compound on the right results in a 388 fold decrease in potency.

1.2: Neural Networks for Chemical Property Prediction

On its surface, the breadth of chemical space constitutes a major challenge for the construction of predictive models for chemistry: the space of possible chemical compounds is vast and incredibly diverse. While the size of this space is difficult to bound, preliminary estimates are on the order of 10^{12} possible compounds; this is almost certainly a vast underestimate.¹⁶ Yet, this diversity belies another challenge – the mappings of chemical compounds to properties of interest is rough, leading to so called “activity cliffs”. To illustrate this, we provide an example from the field of drug discovery, where such relationships can be particularly pronounced. The vascular endothelial growth factor receptor-2 (VEGFR-2) is an example of a receptor tyrosine kinase extracellular signaling protein implicated in tumor growth and proliferation. As such, it is a common target in designing oncological therapeutics. Two such potential VEGFR-2 inhibitors are shown in **Figure 1.1**. The single substitution of a carbon atom to a nitrogen atom in one of the terminal aromatic rings reduces the inhibitor potency by a factor of nearly 400, showing that small changes in chemical structure can result in large changes in downstream molecular properties.¹⁷

While the effects of this roughness can be mitigated through careful descriptor selection in studies confined to narrow regions of chemical space, the construction of models that are accurate over disparate and large regions of chemical space require the use of model archetypes with high parametric flexibility to capture this roughness. Many models in the literature have shown promise in this regard, including support vector machines with learnable kernels, tree-based models such as random forests or gradient-boosted trees, and kernel ridge regression models.^{18–23} Yet frequently, deep neural networks have been the class of models showing the best generalizability and accuracies for prediction tasks on large datasets. We will focus on this class of models in this work, while using other types of models as baselines for comparison.

Neural networks are a class of machine learning models capable of accurately describing complex non-linear phenomena. Each neural network consists of a series of “layers”, which in turn consist of a matrix of trainable parameters and a subsequent nonlinearity. By sequentially transforming inputs using these layers, an accurate mapping from inputs to desired outputs may be created. Inputs to these models take the form of a vector; inputs – text, images, or in our case, molecules – must be converted to vectors in a preprocessing step before being fed into the model (*vide infra*). Consider the case of a molecular dataset for which we wish to obtain molecule-level predictions (eg., the enthalpy of formation). If the dataset consists of N molecules, we will represent this input vector as $[N, C]$, where C represents the “feature dimension”, the dimension of each molecule level vector. If instead, we were interested in predicting atom-level properties (eg., the NMR chemical shifts), our inputs may be of the form $[N, A, C]$, where A represents the maximum

number of atoms in any of the molecules in our dataset. (Representations for molecules with fewer atoms may be padded, say with zeros, such that the vector representation for each molecule is the same shape). These vectors are then multiplied by a matrix of weights, of the shape $[C, H]$, where H is the dimension of the given layer; this dimension is a choice made by the model creator (a “hyperparameter”) and must be carefully tuned to ensure optimal performance. Thus, the dimension of the data after the multiplication operation is $[N, H]$ or $[N, A, C]$ in the case of atom-level predictions. Subsequently, this resultant matrix is subject to a nonlinear activation function. Examples of these include the ReLU, tanh, GeLU, and SiLU functions; the choice of which function to use is also a model hyperparameter. The composition of these two functions – linear matrix multiplication and nonlinear activation – constitutes a “layer”. By stacking several of these layers in sequence, a neural network with sufficient depth (number of layers) or breadth (layer dimension) can approximate any function; this result, known as the Universal Approximation Theorem, was variously proved by Funahashi²⁴, Cybenko²⁵, Hornik²⁶, and Pinkus²⁷. The first layer is frequently referred to as the “input layer”, the last layer as the “output layer”, and the intermediate layers as “hidden layers”. The output layer’s dimension will necessarily be determined by the quantity we wish to predict. For instance, if we wish to predict only the enthalpy of formation, the output layer weights may be of the form $[H, 1]$, resulting in an output shape of $[N, 1]$ if inputs to the output layer are of the form $[N, H]$. Alternatively, if we want to simultaneously predict the energy of formation and the entropy of formation, the output layer weights may take the form $[H, 2]$, resulting in an output shape of $[N, 2]$ if inputs to the output layer are of the form $[N, H]$.

The constituent weights in the weight matrices (one matrix for each layer) referred to previously are updated during model training in a process termed backpropagation, in such a manner that updates minimize the model’s final prediction error. Initially, weights are initialized by random sampling from specific prior distributions. An initial prediction is made by propagating the inputs through all model layers, and the error of this prediction (the “loss”) is determined using a user-specified metric known as the “loss function”. The choice of this function is critical; the mean absolute error and root mean square error are common choices for regression tasks, involving the prediction of continuous values, while cross entropy or hinge losses are frequently used for classification tasks involving the prediction of discrete values. However, the use of additional terms in the loss function provides a handle to steer model training using known patterns in the training data.

Once a loss has been calculated, adjustments to the model weights are calculated by taking the gradient of the model weights with respect to the loss. The exact formulation of the update is governed by the choice of model optimizer, another hyperparameter. Common choices include stochastic gradient descent (SGD)^{28,29} and variants of adaptive moment estimation (Adam).³⁰ We will illustrate this process using the deterministic form of gradient descent. Formally, consider a neural network of n layers, with inputs of x and target outputs of y . Each layer consists of a weight matrix W and an activation function σ ; thus, we write the overall network as a composition of individual layers:

$$g(x) := \sigma^N(W^N \sigma^{N-1}(W^{N-1} \dots \sigma^1(W^1 x) \dots)) \quad 1.1$$

We will then denote the output of an individual layer as:

$$a^l = \sigma^l(W^l a^{l-1}) \quad 1.2$$

During model training, the weights are initially fixed and a prediction made by applying the composed neural network function to the inputs. Then, given a loss function C , we write the loss as:

$$L = C(y_i, g(x_i)) \quad 1.3$$

Thus, we can write the error in the output prediction of the network, δ^L , as a sum over the error of each output component, δ_j^L (assuming the general case of a network with multiple outputs. For convenience of notation, we denote the product of the weight matrix of the layer l with the output of layer $l-1$ as z .

$$\delta_j^L = \frac{\partial L}{\partial a_j^L} \sigma'(z_j^L) \quad 1.4$$

Here, the $\partial L / \partial a_j^L$ term measures the change in the loss with respect to the j^{th} output activation, while the $\sigma'(z_j^L)$ term represents how fast the activation function changes at z_j^L .

In matrix form, the above can be represented as:

$$\delta^L = \nabla_a L \odot \sigma'(z^L) \quad 1.5$$

where $\nabla_a C$ is the gradient of C with respect to each layer output. We can extend this formulation to a general layer l in the network:

$$\delta^L = ((w^{l+1})^T \delta^{l+1}) \odot \sigma'(z^l) \quad 1.6$$

where $(w^{l+1})^T$ is the transpose of the $l + 1$ layer weight matrix, the application of which can be thought of as moving the error backwards through the network. The application of these two equations then allows for the calculation of the error at any layer in the network: first, Equation 1.5 is used to calculate the error in the output layer. This error is then used in Equation 1.6 to calculate the error of the penultimate layer, and so forth. Then, we can write the rate of change of the cost with respect to any weight in the network:

$$\frac{\partial L}{\partial w_{jk}^l} = a_k^{l-1} \delta_j^l \quad 1.7$$

Then, to adjust the weights to minimize the prediction error, we “take a step” in the direction opposite to this gradient and subtract this from the value of the weight:

$$w_{jk}^{l \text{ updated}} = w_{jk}^l - \alpha \frac{\partial L}{\partial w_{jk}^l} \quad 1.8$$

Here, α is a hyperparameter (“learning rate”) which controls the rate at which weights are updated. A value of α which is too high can cause training instability due to “overshooting” optimal weight values during the training process, while a value that is too low can cause slow training.

Since inputs to neural network models take the form of vectors, molecular inputs must be converted to vectors before they can be used as inputs, converting a discrete point in chemical space to a continuous input. A variety of methods to do so exist in the literature; one such common technique to do so involves computing easy-to-calculate “descriptors” – physicochemical properties which describe the molecule of interest. Descriptors are

DIMENSION	DESCRIPTION	EXAMPLES
1D	Constitutional descriptors calculated from 1D structure	Molecular weight, number of rings
2D	Atomic connectivity: topological branching/graph index based	Morgan fingerprint, Balaban J index
3D	3D structure descriptors mapping atoms/bonds to points in space	Average steric occupancy, Sterimol
4D	Conformational ensemble descriptors	wSterimol, Boltzmann averaged values

Table 1.1: Molecular descriptors categorized by dimension. Examples of common descriptors for each type are also shown.

commonly categorized into 1D, 2D, 3D, and 4D descriptors, representing their spatial/topological dimension (**Table 1.1**). Appropriate choice of descriptors is vital to ensure molecular properties of interest are maximally captured; it is here where the judicious application of chemical intuition plays a key role. A combination of some or all descriptor types may be used to this effect. For instance, in the work by Sigman and Doyle cited earlier, mechanistic investigation revealed that ligand sterics are critical in determining reaction enantioselectivity; thus, Sterimol³¹ descriptors are used as model inputs.

The atom-bond connectivity structure of molecules offers a supplemental means to the descriptors discussed above to vectorize molecules. The structure of graphs is a natural means to parameterize this connectivity: a graph is a mathematical structure used to model pairwise relations between different entities. Each graph is made up of constituent nodes connected by edges, representing atoms and bonds respectively. Alternatively, if the

relationships between different molecules is of interest, each molecule may be represented as nodes, with intermolecular interactions then being represented by edges. This discrete connectivity structure thus neatly encodes intuitive chemical principles: atoms are most strongly influenced by their nearest neighbors, and the properties of a whole molecule can be thought of as some combination of atomic properties.

Graph neural networks (GNNs) are a class of neural networks specifically designed to exploit this connectivity as an “inductive bias”: an assumption about the structure of the underlying data a network makes predictions on. The most general form of these models is known as a message passing GNN (MPGNN) model; here, each node and edge is assigned a vector representing a unique state. Each layer of the MPGNN model updates this state vector using an operation known as a graph convolution. The graph convolution serves to incorporate information about neighboring nodes and edges, with each layer incorporating information from more distant nodes and edges using edge connectivity as a scaffold. In this way, each state vector captures a detailed, unique environment for each node and edge in the input graph. Formally, updates are given by the following equation; we consider node updates for convenience, but the same update rule holds for edge updates as well:

$$h_j^{l+1} = \sigma\left(h_j^l, \mathcal{N}(h_j^l)\right) \quad 1.9$$

Here, h_j^{l+1} represents a state vector for node j at layer $l + 1$. This state vector consists of an update function σ , which takes as inputs the node state at layer l and the result of a neighborhood aggregation function $\mathcal{N}$ on node j . This neighborhood aggregation function may be any permutation invariant function (eg: sum, average, maximum, etc.) and serves to

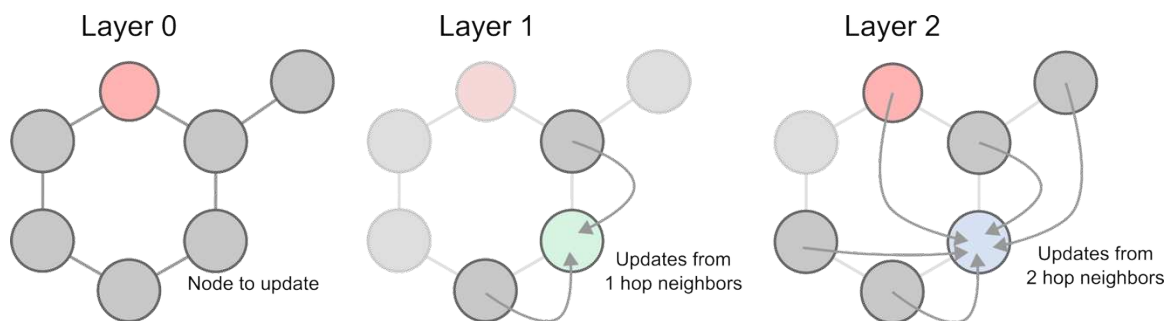


Figure 1.2: Node update scheme. Subsequent layers introduce updates from more distant nodes, building a representation of each node's unique environment.

combine the representations of connected nodes at l “hops” away from node j (**Figure 1.2**).

So, for example, in layer 2, a neighborhood aggregation function may average state vectors of both the first and second neighboring nodes of node j . On the other hand, the update function σ can be parameterized by a neural network, invoking the previously discussed Universal Approximation Theorem. Thus, h_j^{l+1} is given by the output of a neural network layer which takes as inputs the previous node state and an aggregation of the node's neighborhood state. By restricting the form of the aggregation and update functions, we arrive at alternative model formulations, such as graph attention networks, graph transformers, etc.

1.3: References

1. Denmark, S. Arthur C. Cope Award Lecture. (2023).
2. Cho, J. *et al.* Bioderived ether design for low soot emission and high reactivity transport fuels. *Sustainable Energy Fuels* **6**, 3975–3988 (2022).
3. Ito, Y., Taniai, T., Igarashi, R., Ushiku, Y. & Ono, K. CrystalFramer: Rethinking the Role of Frames for SE(3)-Invariant Crystal Structure Modeling. Preprint at <https://doi.org/10.48550/arXiv.2503.02209> (2025).
4. Kim, Y. *et al.* Designing green chemicals by predicting vaporization properties using explainable graph attention networks. *Green Chem.* **26**, 10247–10264 (2024).
5. Kim, Y. *et al.* Prediction of Hydroxymethylfurfural Yield in Glucose Conversion through Investigation of Lewis Acid and Organic Solvent Effects. *ACS Catal.* **10**, 14707–14721 (2020).

6. Ohta, A. *et al.* Validation of a New Methodology to Create Oral Drugs beyond the Rule of 5 for Intracellular Tough Targets. *J. Am. Chem. Soc.* **145**, 24035–24051 (2023).
7. Yang, J. H. *et al.* Room-Temperature Decomposition of the Ethaline Deep Eutectic Solvent. *J. Phys. Chem. Lett.* **16**, 3039–3046 (2025).
8. Tanada, M. *et al.* Development of Orally Bioavailable Peptides Targeting an Intracellular Protein: From a Hit to a Clinical KRAS Inhibitor. *J. Am. Chem. Soc.* **145**, 16610–16620 (2023).
9. Newman-Stonebraker, S. H. *et al.* Univariate classification of phosphine ligation state and reactivity in cross-coupling catalysis. *Science* **374**, 301–308 (2021).
10. Pérez-Soto, R. *et al.* A Fragment Based Approach Towards Curating, Comparing and Developing Machine Learning Models Applied in Photochemistry. *Chem. Sci.* <https://doi.org/10.1039/D5SC05615B> (2025) doi:10.1039/D5SC05615B.
11. Kumar, S. *et al.* Combining Geometry and Topology for Accurate Protein Solubility Prediction. Preprint at <https://doi.org/10.26434/chemrxiv-2025-jmzwn> (2025).
12. Kim, Y., Jung, H., Kumar, S., Paton, R. S. & Kim, S. Designing solvent systems using self-evolving solubility databases and graph neural networks. *Chem. Sci.* **15**, 923–939 (2024).
13. Jung, H. *et al.* Enhancing predictive models for solubility in multicomponent solvent systems using semi-supervised graph neural networks. *Digital Discovery* **4**, 1492–1504 (2025).
14. Kim, Y. *et al.* Physics-informed graph neural networks for predicting cetane number with systematic data quality analysis. *Proceedings of the Combustion Institute* **39**, 4969–4978 (2023).
15. Yeonjoon Kim *et al.* Physics-informed graph neural networks for predicting cetane number with systematic data quality analysis. *Proceedings of the Combustion Institute* **39**,.
16. Coley, C. W. Defining and Exploring Chemical Spaces. *TRECHEM* **3**, 133–145 (2021).
17. Stumpfe, D. & Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *J. Med. Chem.* **55**, 2932–2942 (2012).
18. Fehér, P. P., Madarász, Á. & Stirling, A. Prediction of Redox Power for Photocatalysts: Synergistic Combination of DFT and Machine Learning. *J. Chem. Theory Comput.* **19**, 4125–4135 (2023).

19. Kalita, B., Li, L., McCarty, R. J. & Burke, K. Learning to Approximate Density Functionals. *Acc. Chem. Res.* **54**, 818–826 (2021).
20. Kang, B., Seok, C. & Lee, J. Prediction of Molecular Electronic Transitions Using Random Forests. *J Chem Inf Model* **60**, 5984–5994 (2020).
21. Kim, B. *et al.* Enhancing Permeability Prediction of Heterobifunctional Degraders Using Machine Learning and Metadynamics-Informed 3D Molecular Descriptors. *J. Chem. Inf. Model.* <https://doi.org/10.1021/acs.jcim.5c01600> (2025) doi:10.1021/acs.jcim.5c01600.
22. Stubbs, C. D. *et al.* Predicting homopolymer and copolymer solubility through machine learning. *Digital Discovery* <https://doi.org/10.1039/D4DD00290C> (2024) doi:10.1039/D4DD00290C.
23. Pigeon, T. *et al.* Computing Surface Reaction Rates by Adaptive Multilevel Splitting Combined with Machine Learning and Ab Initio Molecular Dynamics. *J. Chem. Theory Comput.* **19**, 3538–3550 (2023).
24. Funahashi, K.-I. On the approximate realization of continuous mappings by neural networks. *Neural Networks* **2**, 183–192 (1989).
25. Cybenko, G. Approximation by superpositions of a sigmoidal function. *Math. Control Signal Systems* **2**, 303–314 (1989).
26. Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural Networks* **4**, 251–257 (1991).
27. Pinkus, A. Approximation theory of the MLP model in neural networks. *Acta Numerica* **8**, 143–195 (1999).
28. Hinton, G. E. & Salakhutdinov, R. R. Reducing the Dimensionality of Data with Neural Networks. *Science* **313**, 504–507 (2006).
29. Hinton, G. *et al.* Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine* **29**, 82–97 (2012).
30. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. Preprint at <https://doi.org/10.48550/arXiv.1412.6980> (2017).
31. Development and Application of New Steric Substituent Parameters in Drug Design. in *Drug Design* 165–207 (Academic Press, 1976). doi:10.1016/B978-0-12-060307-7.50010-9.

CHAPTER 2: Bond Properties from Atomic Interactions: Predicting Homolytic Ring Opening Bond Dissociation Enthalpies

2.1: Chapter Overview

In this chapter, we examine how the interactions of atom and bond-level properties across molecular environments can be combined to infer the properties of chemical bonds. We study the bond dissociation enthalpy, or BDE, the prediction of which is a critical task in understanding and manipulating the behavior of chemical reactions. While previous efforts have modeled this enthalpy change, they have not been extended to predict the BDE associated with homolytic bond cleavage in a ring structure, forming a diradical intermediate. In this work, we supplement our previously established ALFABET2 database of 531,244 DFT-calculated BDEs and bond dissociation free energies (BDFEs) curated from sources such as PubChem and ZINC15 to include ring-opening reactions, creating a database of 731,825 bond dissociation enthalpy and free energy values; this database was then used to retrain the ALFABET BDE prediction model, resulting in a unified model capable of predicting the BDE of compounds containing C, O, H, N, S, F, Cl, Br, I, P, Si, and ring structures. The retrained model predicts BDEs and BDFEs with mean absolute errors of 0.86 and 0.84 kcal/mol, respectively. We then assessed the model's predictive capacity and conducted an ablation analysis to identify the chemical properties that inform its predictions. Lastly, we present an application that highlights the use of the retrained model to predict the BDEs of lactones commonly used in ring-opening polymerization reactions.

This collaborative research was performed with lead contributions from myself and additional contributions from Dr. Shree Sowndarya S. V., Dr. Yeonjoon Kim, Dr. Peter St. John, Dr. Robert S. Paton, and Dr. Seonah Kim. I developed the high-throughput DFT pipeline, trained and evaluated the neural network architectures, performed analysis of the presented application to lactone polymerization, and wrote and edited the manuscript for publication.

2.2: Introduction

The enthalpy change accompanying bond breaking is critical to our understanding of the nature of chemical reactions. Lower enthalpy changes are correlated with weaker bond strengths and thus indicate likely sites for chemical reactions.^{1,2} Quantitative knowledge of this bond dissociation enthalpy (BDE) aids synthetic planning and reaction elucidation across domains as diverse as electrostatic catalysis,³ polyfluoroalkyl substance degradation,⁴ and directed functionalization of organic molecules.⁵

While experimental techniques such as pyrolysis and electrochemical redox potential measurements^{6,7} have been traditionally used to measure BDEs, they are subject to relatively large experimental errors on the order of several kcal/mol. *In-silico* methods for calculating BDEs have been used to circumvent this problem; however, these methods often rely on quantum-mechanical calculations, which are computationally intensive and require expert knowledge to be performed successfully. Previous studies using *ab initio* or density functional theory (DFT) methods have been limited in scope, generating calculated BDE values for only tens⁸ or hundreds of species,⁹ limiting their use in predictive models for large portions of chemical space. For example, the recently developed BSE49 dataset uses highly

accurate (RO)CBS-QB3 methods to calculate the bond separation energy accompanying homolytic cleavage of a diverse range of chemical bond types.¹⁰

However, much of the prior literature neglects ring opening reactions involving the formation of a single product species, due in part to the difficulty in computationally studying the diradical species formed by homolytic bond cleavage. Such difficulties include handling spin contamination, distinguishing between triplet and open-shell singlet electronic states, and achieving self-consistent-field convergence for diradical species (for details, *vide infra*). This presents a significant barrier to entry for the use of in silico BDE prediction tools in investigating reaction mechanisms where ring strains play a key role in determining product selectivity. Many novel strategies for organic synthesis exploit such mechanisms to engender selectivity, or make use of ring strain energy release for thermodynamic driving force. For instance, the ring-opening of iodoaziridine derivatives is a key step that enables [3+2] regioselective cycloaddition for the construction of cyclopentenyl scaffolds.¹¹ More complex scaffolds, such as substituted quinanes found in many natural products, can be accessed via the strain relief provided by ring-opening following a photochemical strain-installing step.^{12,13} This use of strain relief as a thermodynamic driver of reactivity is also often exploited in ring opening polymerization reactions for the production of polylactone chains.^{14,15}

Multiple other fields of chemistry also make use of BDEs to inform molecule design. In medicinal chemistry, the ring strain of strained heterocycles, including epoxides and β -lactams, governs the intrinsic reactivity of covalent drug warheads, with homolytic ring-opening BDEs providing a computationally accessible measure of the propensity for

bioactivation and reactive metabolite formation that can drive off-target toxicity.^{16,17} Accurate BDE predictions, therefore, directly support the design of covalent inhibitors with optimized potency-selectivity profiles, an increasingly important strategy in modern drug discovery.¹⁸ In the context of combustion and energy, homolytic ring-opening BDEs of cyclic hydrocarbons and oxygenated ring systems are essential thermochemical inputs to detailed chemical kinetic mechanisms: cycloalkanes comprise up to 40% of commercial fuels by mass, and their ring-opening BDEs directly influence predicted ignition delay times and laminar flame speeds.^{19,20} Additionally, in the emerging field of mechanochemistry, the BDE of a strained ring bond determines the force threshold for mechanochemical activation of mechanophore units, enabling the rational design of force-responsive, self-healing, and stress-reporting polymer materials.^{21,22} In each of these domains, the prohibitive cost of DFT calculations has previously limited high-throughput exploration of ring-containing chemical space.

Machine learning (ML) methods have gained widespread popularity over the past decade as a means to surmount this barrier, bolstered by advances in computing hardware and improved predictive algorithms. Several such models have been developed to predict BDEs, and can be broadly divided into two classes: those that directly predict BDEs from molecular structure and those that leverage a machine-learning interatomic potential (MLIP) to estimate them. Models that predict BDEs directly from input structures have historically focused on specific reaction types, thereby restricting the domain of applicability of their predictions to narrow regions of chemical space. For instance, Nakajima and Nemoto recently developed an ElasticNet²³-based model to predict BDEs for compounds containing

hypervalent iodine using molecular fingerprints derived from SMILES string representations, achieving a test set accuracy of 1.58 kcal/mol.²⁴ Aires-de-Sousa and coworkers developed a nearest-neighbors corrected random forest model to predict the BDEs of C-H bond cleavage for sp³ hybridized carbon atoms, using DFT calculated BDEs of 2263 unique bonds and 409 experimental BDEs from the iBond database,²⁵ achieving an accuracy of 2.22 kcal/mol on a held-out test set.²⁶

Recently, message-passing graph neural networks (GNNs) have emerged as an accurate ML methodology for simulating chemical properties.²⁷⁻²⁹ GNNs represent molecular structure by encoding atom and bond information via graph nodes and edges. By using the graph structure as an inductive bias, model predictions are less affected by descriptor selection than of those to other ML methods, removing a source of human bias that can influence model predictions.^{30,31} The message passing GNN framework developed by Gilmer et. al.³² has been used effectively to model chemical processes in areas as disparate as coarse graining molecular dynamics simulations,³³ optical property tuning,³⁴ and structural optimization,³⁵ illustrating its versatility and utilization in disparate fields of chemistry and materials science. Persson and coworkers developed the BondNet model, which can predict bond dissociation enthalpies for heterolytic bond cleavage in both neutral and charged species to within 0.55 kcal/mol.³⁶ While the accuracy of these models rivals that of quantum-mechanical calculations, their applicability remains limited due to the chemical space covered in the model training sets, particularly for portions relevant to pharmaceutical and materials synthesis.

On the other hand, models leveraging recently developed MLIPs for predicting BDEs have also been developed. Such models use ML models to predict accurate force-field terms, allowing molecular mechanics-based simulation techniques to approach the accuracy QM simulations, thus providing significant computational savings when compared to QM approaches. While several highly accurate MLIP models have been developed, their use for BDE prediction has been somewhat limited. Csányi and coworkers developed an interatomic potential prediction model³⁷ based on the MACE MLIP framework,^{38,39} achieving accuracies of up to 1.37 kcal/mol when predicting the BDEs for C-H bond cleavage for compounds in the CYP dataset.⁴⁰ Similarly, the ANI-1xBB model developed by Isayev and coworkers uses a reactive MLIP trained on nonequilibrium conformer geometries.⁴¹ With no explicit radicals included in the training set, the model achieves a predictive accuracy of 7.74 kcal/mol. Such MLIP-based models offer strong generalizability across chemical space. However, predictive accuracies frequently lag behind those of bespoke direct-prediction models, and both training and inference incur large computational costs due to model size and architecture choices.^{42–44}

Previous work by St. John et al. used the message-passing GNN framework to develop the ALFABET1 ML-based BDE prediction tool, capable of predicting homolytic BDEs for

DATABASE/ MODEL	ELEMENTS	DB SIZE	RING OPENING INCLUDED?	TEST SET ACCURACY (BDE, KCAL/MOL)
ALFABET1	C, H, O, N	290,644	No	0.58
ALFABET2	C, H, O, N, S, Cl, F, I, P	531,244	No	0.60
ALFABET3	C, H, O, N, S, Cl, F, I, P, Si	731,825	Yes	0.86 (1.54: Ring, 0.69: Non-ring)

Table 2.1: Comparison of ALFABET model phases. ALFABET3 test set accuracy is shown for the entire test set and for ring and non-ring bonds within the test set.

compounds containing C, H, O, and N atoms from molecular structure, with accuracies approaching those of quantum-mechanical calculations.⁴⁵ The subsequent ALFABET2 model developed by Shree Sowndarya et al. further expanded the scope of model predictions to include compounds containing S, P, F, Cl, Br, and I by incorporating molecules from PubChem⁴⁶ and ZINC15,⁴⁷ and enabled the calculation of free energies of homolytic bond cleavage (**Table 2.1**).⁴⁸ However, the ability to predict the enthalpy associated with the homolytic cleavage of a bond in a ring structure has, so far, been omitted. In this work, the ALFABET2 model is retrained to expand the domain of predictions to include these reactions using a newly calculated dataset of homolytic ring opening BDEs.

In this work, we retrain the ALFABET2 BDE prediction tool to extend its scope to homolytic ring-opening reactions, a reaction class forming diradical intermediates that has been systematically excluded from all prior BDE prediction databases (**Figure 2.1**). To this end, we construct a large-scale DFT-calculated dataset of ring-opening BDEs by iteratively breaking ring bonds in compounds from the ALFABET2 database, with extensive benchmarking of the computational methodology required to treat the resulting diradical species. To our knowledge, ALFABET3 is the first ML model trained on a dedicated dataset of

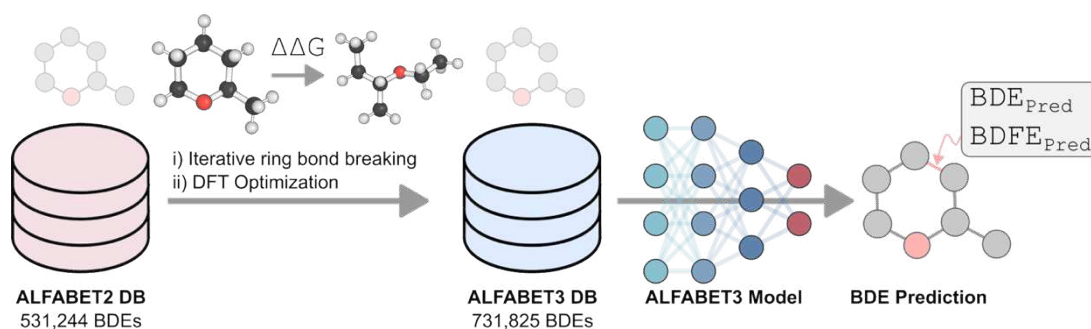


Figure 2.1: Overall scheme for the ALFABET3 model. Molecules from the existing ALFABET2 database were subject to iterative ring bond breaking, and the BDE and BDFE of homolytic bond cleavage calculated. The compiled database of BDE/BDFE values was used to train the ALFABET3 graph neural network model to predict the BDE/BDFE of all bonds in a given structure.

homolytic ring-opening BDEs, achieving predictive accuracy well within the error of the underlying DFT methodology. We then critically analyze model performance through outlier characterization and an ablation study of ring-specific molecular features. Finally, we demonstrate the practical utility of ALFABET3 by applying it to screen ring-opening sites in lactone monomers relevant to ring-opening polymerization, illustrating how the model enables rapid, low-cost prediction of ring-opening reactivity across diverse chemical space.

2.3: Results and Discussion

We aim to expand the utility of the ALFABET2 BDE/BDFE prediction tool to include homolytic ring opening BDEs; this was performed by expanding the size of the dataset used for ML training and retraining the model. By utilizing compounds from ALFABET2 for this purpose as starting structures, the need for computationally expensive DFT calculations was minimized, as DFT ground truth values for the non-ring opening steps are reused for training.

2.3.1: Expanding Dataset Size

Compounds in the ALFABET2 BDE database were used to generate training data for the current work (**Figure 2.1**, left). In summary, parent compounds with ring structures were identified, and ring bonds were iteratively broken to generate diradical species, yielding a diverse array of ring opened species. DFT calculations using the M06-2X functional⁴⁹ and the def2-TZVP valence triple-zeta polarizable basis set⁵⁰ were performed to calculate the enthalpy change accompanying homolytic ring opening bond cleavage. This enthalpy

change was calculated by subtracting the formation enthalpy of the diradical species from that of the parent species.

Extensive benchmarking studies were performed to validate the computational methodology used for dataset generation owing to the difficulty in modeling diradical species, as opposed to the single radical species considered in the construction of the ALFABET1 and ALFABET2 models. The relatively large contributions of static correlation to the electronic energy of such structures are addressed in wavefunction-based theories via the use of multiple initial references. This contribution is implicitly addressed through the exchange and correlation functional approximations in Kohn-Sham DFT,⁵¹ yielding good approximations for systems for which wavefunction theories struggle, such as for the ozone molecule.⁵² Since the exchange-correlation approximation is critical to the accurate representation of electronic structure, DFT functionals and basis set combinations were benchmarked. The M06-2X/def2-TZVP level of theory-basis set combination was chosen after performing benchmarking calculations utilizing 40 experimentally determined diradical formation enthalpy values found in the existing body of literature.^{53–56} In total, the results of calculations using 9 different levels of theory and basis set combinations were compared to

experimental results and assessed on both accuracy (**Figure 2.2a**) and computational cost (**Figure 2.2b**). Further details of this benchmarking process can be found in Appendix A, Section S1, “Benchmarking Computational Methods for Dataset Generation.”

Among the benchmarked methods, M062X/def2-TZVP offers the best balance of computational accuracy and cost. This level of theory offers the best accuracy with a moderately high computational cost. However, since the previous ALFABET2 database was generated using this level of theory-based set combination, selecting this combination enabled seamless integration of ring-opening BDE results with the previous database and avoids expensive re-calculation.

All DFT calculations for diradical species were performed in the triplet state. The use of a triplet state reduces the multireference character of the system studied and greatly

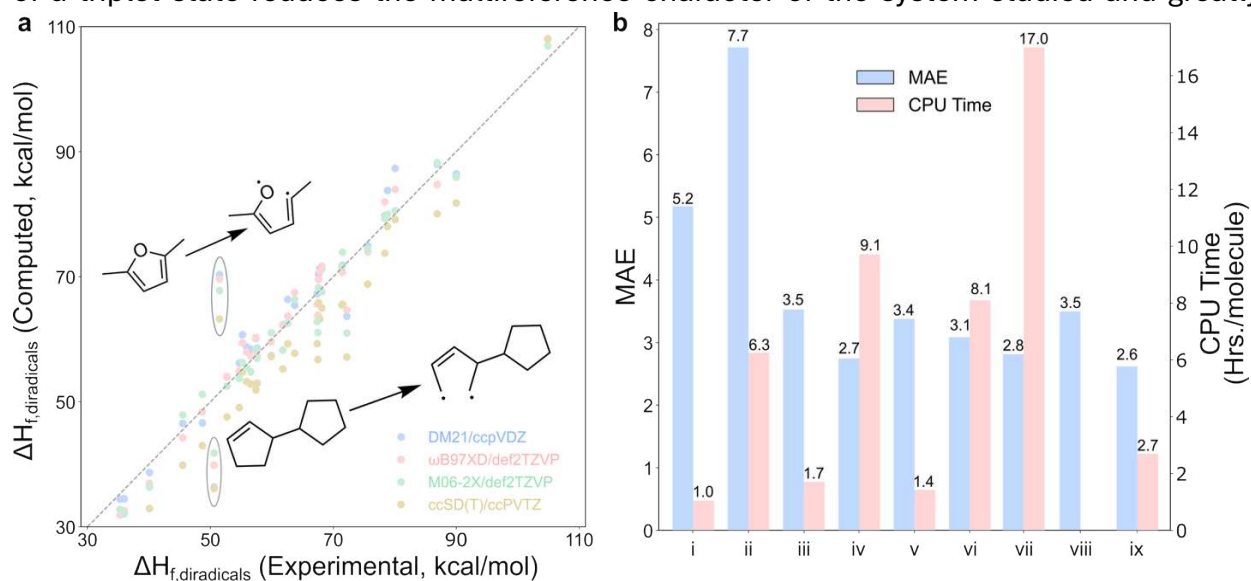


Figure 2.2: Benchmarking results for homolytic ring opening BDEs calculated with different levels of theory. a) Parity plot of calculated BDEs vs. experimentally measured BDEs curated from the literature. Two outlier molecules are observed, in which the diradical centers are separated by a conjugated π -system. b) Average calculation error and CPU time cost for benchmarked levels of theory: i: B3LYPD3/6-31G(d), ii: B3LYPD3/def2TZVP, iii: M06-2X/6-31G(d), iv: M06-2X/def2-TZVP, v: ω B97XD/6-31G(d), vi: ω B97XD/def2-TZVP, vii: CBS-QB3, viii: DM21/ccpVDZ, ix: DLPNO-ccSD(T)/ccpVDZ. Accurate wall times could not be obtained for the DM21 functional and are omitted here (details in SI Section S1, Benchmarking Computational Methods for Dataset Generation). We also note that the CPU time shown for the DLPNO-ccSD(T)/ccpVDZ calculation reflects only the time taken for SCF convergence; an optimized starting geometry was obtained from the M062X/def2TZVP optimized structure.

simplifies the computational complexity of the large scale high-throughput calculations required to build a training dataset for ML models.⁵⁷ Prior benchmarking found that the energies of open-shell singlet states were higher than those of triplet states, indicating that triplet-state energies are more representative of the system state for reactivity predictions. In addition, to validate the correct spin localization of the radical spin density on the correct atoms, CBS-QB3⁵⁸ and DLPNO-CCSD(T)⁵⁹ calculations were performed on the benchmark set. Yamaguchi's broken symmetry corrections were also performed on these systems to investigate the influence of triplet state spin contamination on a single-reference singlet calculation (details in Appendix A, Section S1).^{60,61} It was found that the Yamaguchi broken-symmetry corrections provided only marginal improvements in prediction accuracy compared to DFT calculations. Lastly, calculations with the DM21 neural functional were also performed; the DM21 functional has shown good performance in localizing spin density on radical species due to its ability to treat "fractional" electron spins.⁶² It was found that the selected M06-2X/def2-TZVP functional-basis set combination performed similarly to the tested high-level methods; details of these calculations are provided in Section S1 of Appendix A. Further, it was found that when the diradical centers were separated by a conjugation path, none of the tested methods correctly localized the spin density. This result was used to filter the final generated ALFABET3 dataset (*vide infra*).

An automated, high-throughput workflow was subsequently used to iteratively break ring bonds in compounds from the ALFABET2 database and to compute their BDEs. The rdkit Python package⁶³ was first used to identify ring bonds. The identified bonds were deleted from the molecular structure, and the conj_path Python

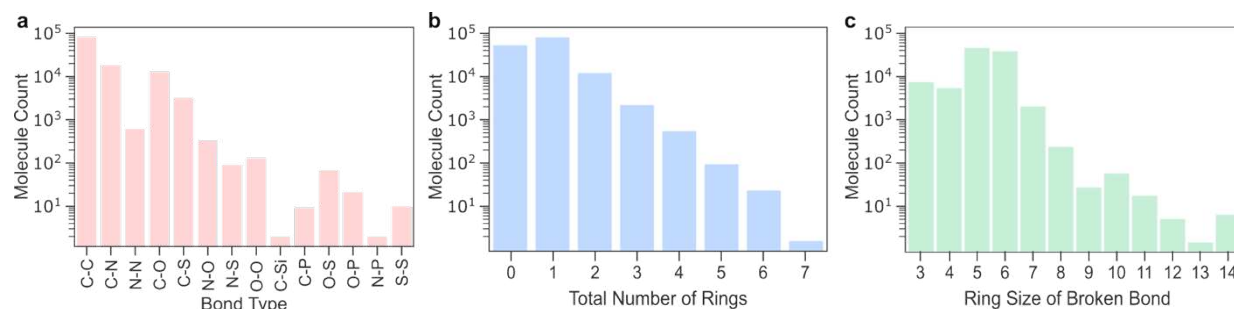


Figure 2.3: Database statistics for the ALFABET3-DB. a) Distribution of bond types for the newly added ring-opening reactions. b) Distribution of ring counts in each molecular structure for all molecules. c) Distribution of sizes of rings from which the bond was broken

package⁶⁴ was used to identify and remove molecular structures in which diradical centers are separated by a conjugation path, resulting in the removal of 1,310 molecular species. The MMFF94 forcefield was then used to generate initial conformers. Since the MMFF94 forcefield is not parameterized for diradical species, generated diradicals were temporarily capped with hydrogens, conformer search was performed, and hydrogens were removed to regenerate the diradical species. Stable conformers were subjected to optimization and single point energy calculation with the selected M06-2X/def2-TZVP functional/basis set combination. Screening of calculations was performed to filter out abnormal terminations (including those with imaginary frequencies), fragmentation, or isomerization using the cclib Python package⁶⁵ in conjunction with the rdkit.⁶³ The Goodvibes Python package⁶⁶ was then used to obtain the BDFE for each calculated BDE. A dataset of 1,130,903 BDE/BDFE values was generated using this pipeline, comprising 731,825 unique values corresponding to 76,790 molecular species. These unique values make up the ALFABET3 database, which adds 134,270 datapoints to the ALFABET2 database.

The final database contains a diverse array of BDEs and BDFEs corresponding to homolytic ring opening. While the most common bond in the ALFABET3 dataset is a C-C bond, over 100 unique molecules containing at least one C-N, C-O, C-S, N-N, N-O, and O-O bonds are also present (**Figure 2.3a**). In addition, a wide range of ring structures are also present; while 27,415 compounds in the ALFABET3-DB have no ring structures, 49,375 compounds have at least one ring structure (**Figure 2.3b**). Since the size of ring structures is known to impact their BDE/BDFE, the distribution of ring sizes was also obtained (**Figure 2.3c**). It is found that 5 and 6-membered rings are the most common, with 45,915 and 40,373 unique BDE/BDFE values, respectively.

2.3.2: Training the ALFABET3 Model and Analysis of Model Performance

The newly established ALFABET3 database was used to retrain the ALFABET2 model to expand the model predictive scope to homolytic ring opening reactions, resulting in the ALFABET3 model. To maintain compatibility with previous versions, the message-passing model architecture used for the ALFABET2 model was reused for ALFABET3.

The ALFABET3 model represents a molecule as a directed graph: atoms are nodes, and bonds are edges. An adjacency matrix representing the bond connectivity of atoms in the molecule is then created. The adjacency matrix is then used to propagate information describing the chemical identity of each atom and bond, ie, node and edge features. This information is conveyed through messages that update the network's representations of atom and bond environments. After several

iterations of this message passing, precise information about atom and bond properties can be inferred.

A minimal set of features, obtained from analysis of the molecular graph topology and chemical identity using the rdkit Python package, is used to categorize each atom and bond type; this information is fed into the neural network as node and edge features, respectively. Compared to ALFABET2, the key new features introduced in ALFABET3 are ring-related: ring size was added as both an atom feature and a bond feature (set to 0 for non-ring atoms/bonds), reflecting the importance of ring membership and ring strain in determining BDE values for ring-opening reactions. The element identity, number of radical electrons, formal charge, chiral tag, aromaticity, ring size (0 if not part of a ring), degree, and the total number of hydrogen bonds were used as atom features, and start/end atom identity, rdkit bond type, binary conjugation status, and ring size (0 if not part of a ring) were used as bond features.

The final ALFABET3 model consists of 6 total message passing layers (**Figure 2.4a**). Initial atom and bond features from rdkit are fed into a linear embedding to project into 128-dimensional vectors. These vectors are used for subsequent message passing operations. The final bond state representations are passed through an output layer to yield a 2D vector for each molecular input, representing BDE and BDFE, respectively. The AdamW optimizer with learning rate 10^{-3} and weight decay of 10^{-5} was used for model training for 200 epochs, using a mean absolute error (MAE) loss function (calculated as the sum of BDE and BDFE losses). Final model cross validation was performed using five random splits of the training dataset. Model training losses are shown in **Figure 2.4b**. Further details on model architecture and training can be found in Appendix A, Section S2, ALFABET3 Model Training.

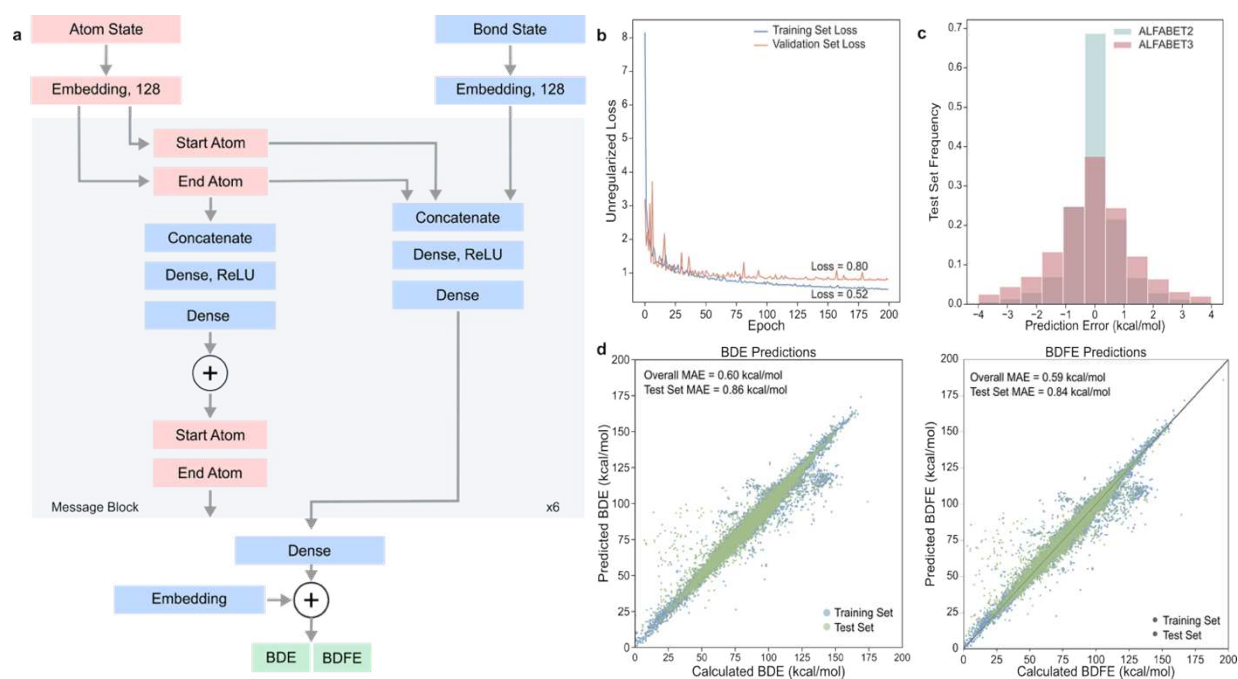


Figure 2.4: ALFABET3 model training details. a) Architecture of the ALFABET3 model. b) Training and validation losses for network training. c) Distribution of BDE prediction errors between test set compounds in the ALFABET2 (blue) and ALFABET3 (red) datasets. d) Parity plots of BDE and BDFE predictions for train and test sets.

The final training dataset comprises 711,169 unique BDE values, spread over 74,570 unique molecules. A held-out test set of 80,866 unique BDE values from 8,425 distinct molecules was used to evaluate model performance. We note that although a random split is used to generate the test and training sets, the percentage of ring bonds in each set is approximately the same (12.2% in the training set, 11.9% in the test set). The retrained model achieves an MAE on this test set of 0.86 kcal/mol and 0.84 kcal/mol for predicted BDE and BDFE, respectively, with 95% of predictions within 0.60 kcal/mol and 0.58 kcal/mol, respectively (**Figure 2.4d**). On just ring compounds in the held-out test set, the model achieves predictive accuracies of 1.54 and 1.50 kcal/mol for BDE and BDFE, respectively (**Figure 2.4c**, diradicals, red), compared to 0.69 and 0.66 kcal/mol for non-ring compounds (**Figure 2.4c**, single radicals, blue). These errors are well within the reported error ranges for single-point energy calculations utilizing the M06-2X functional⁶⁷ and illustrate the predictive utility of the ALFABET model. While these values are slightly higher than the BDE and BDFE MAEs of 0.72 and 0.68 kcal/mol, respectively (achieved by the ALFABET2 model on non-ring bonds only) achieved by the ALFABET2 model, the inclusion of ring bonds greatly enhances the utility of the ALFABET3 model, enabling more complete coverage of the synthetic chemical space.

Model test set outliers were identified by filtering test set predictions for bonds with errors exceeding 10% relative to DFT-calculated values. 676 such test set outliers were found, corresponding to 434 distinct molecules. 372 of these outliers correspond to homolytic ring opening reactions, constituting 55.0% of the total

outliers. The mean and standard deviation of the BDE prediction errors for all outlier molecules were found to be 5.56 kcal/mol and 16.5 kcal/mol, while the mean and standard deviation of the BDFE prediction errors were found to be 5.57 kcal/mol and 16.5 kcal/mol, respectively. A plot of outlier error distributions is available in Appendix A, Section S2, ALFABET3 Model Training.

Additionally, several molecules contain multiple outlier bonds; a plot of outlier counts per molecule is provided in the Appendix A, Section S2, ALFABET3 Model Training. To investigate the distribution of outliers, we use the rdkit-computed synthetic accessibility (SA) score as a measure of chemical structural complexity.⁶⁸ We find that on average, the outlier compounds have an SA score of 3.82 ± 0.98 , as compared to 3.06 ± 0.67 for the full test dataset and 3.01 ± 0.66 for the non-outlier test set compounds, indicating that the outlier molecules are, on average, more chemically complex than the rest of the test set. To quantify the difference in SAs between the test set and the test set outliers, a Cohen's d test was performed: Cohen's d quantifies the standardized difference between two group means (effect size); a value of 0.00 indicates identical distributions, values from 0.0 – 0.2 indicate small dissimilarity, values from 0.2 – 0.8 indicate medium to large dissimilarity, and values above 0.8 indicate high dissimilarity.⁶⁹ The Cohen's d test yields a value of 1.5, indicating a large and significant difference between the SA scores of the two sets.

To qualitatively evaluate the model's ability to learn differences in chemical environments across bonds, the latent-space representations of atoms and bonds are visualized as information passes through the model. Since each molecule is

represented by a 128-dimensional vector, assessing differences in this space is complex; hence, we employ principal component analysis followed by t-Distributed Stochastic Neighbor Embedding (t-SNE) to reduce these representations to 2 dimensions, enabling visualization.⁷⁰ It is observed that the bond and atom representations become increasingly spread out as information propagates through the message-passing layers, as evidenced by the dispersal of clustered representations (**Figures 2.5a** and **2.5b**, Layer 1) from layers 1 to 6. This represents increased resolution of unique atom and bond environments in the neural network's latent space, as information from atoms and bonds is increasingly incorporated. In Layer 1, where only information about the nearest neighboring atoms and bonds is considered, several groups of similar atoms are observed. As information about more distal neighbor environments is incorporated through additional message passing layers, the groupings of similar atoms and bonds become more dispersed, with ring and non-ring variants of the same bond type (e.g., C–N vs. C–N ring, C–O vs. C–O ring)

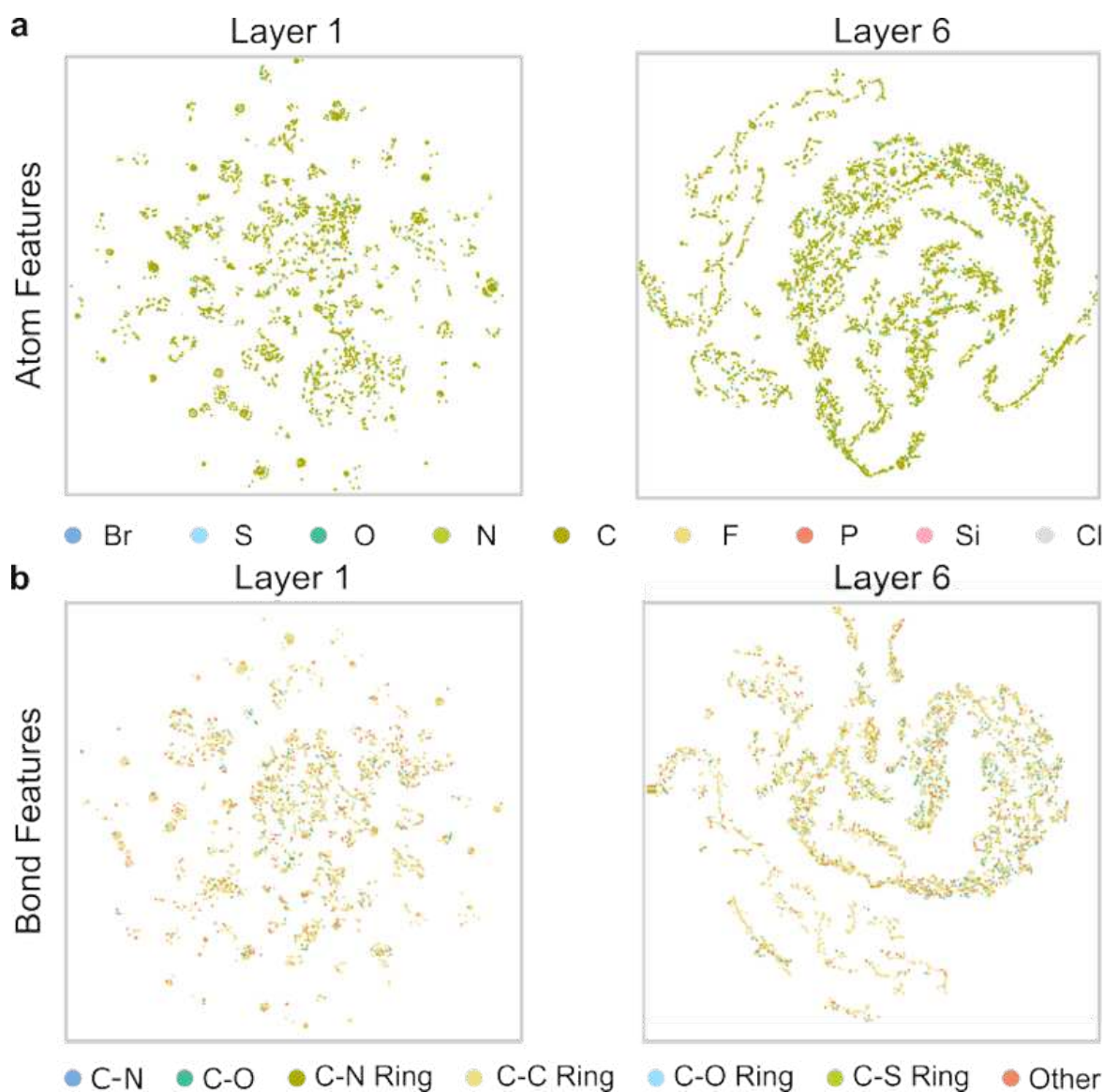


Figure 2.5: t-SNE visualization of a) atom state and b) bond state feature vectors. As the number of layers increases, information from atoms and bonds further from the reactive centers is included, allowing the chemical environments of the broken bond to be increasingly well resolved and dispersing embedding points.

becoming increasingly resolved into distinct clusters by Layer 6 (**Figures 2.5a** and **2.5b**, Layer 6).

Lastly, to determine the effect of ring information on the ALFABET3 model, we perform an ablation analysis. Ablation analysis involves selectively removing neural network components to identify which parts of the network contribute to its predictive ability.⁷¹ Here, we remove ring-related input features from the model's

inputs (ie, ring size features were removed from both atom and bond features). It is observed that the MAE on the full held-out test set increases to 3.64 kcal/mol and 3.34 kcal/mol for the BDE and BDFE, respectively, and that of the ring compounds in the test set increases to 6.11 and 6.14 kcal/mol, respectively. Due to the increase in the magnitude of prediction errors, it is clear that the ring information provided as model inputs is critical to ensuring predictive accuracy.

2.3.3: Prediction of Ring Opening Polymerization Sites in Lactone Precursors

Much recent work has highlighted the utility of lactone species as precursor building blocks for the synthesis of environmentally friendly biodegradable polymers.⁷²⁻⁷⁴ Many such synthetic strategies for block, star, and graft copolymers utilize the thermodynamic driving force of ring opening via ring-strain release to complete the chain-propagation catalytic cycle. The site of bond cleavage is critical to the design of this process; to continue the catalytic cycle, the ring-opening bond cleavage must regenerate an ion or radical. The dispersity of the resultant polymer chain is thus greatly affected by side reactions, leading to shuffling of the catalytically active chain ends and chain termination, as seen when the site of ring-opening differs from that targeted during chain growth. This fluctuation in the number of monomer units in the polymer chains, in turn, affects the polymer's physical properties, such as tensile strength, heat resistance, etc.⁷⁵⁻⁷⁷ Thus, foreknowledge of the site of lactone ring opening is of great importance in the selection of monomer units for the creation of polymers with tailored physical properties.

Previous methodologies for predicting the site of ring opening in lactone monomers have relied on expensive quantum-mechanical calculations. In one

strategy, isodesmic bond breaking reactions are used to estimate the ring strain energy: the ring-opened lactone monomer is capped with suitable end groups in order to conserve the total number and types of bonds in the cyclic form, and quantum mechanics calculations are used to obtain the energy of the cyclic monomer and the ring-opened and capped monomer. The difference in energies is then taken as the ring-strain energy.^{78,79} Another recent approach used detailed conformational sampling of both monomer and polymer geometries using *ab-initio* molecular dynamics (AIMD) approaches to calculate ring opening polymerization enthalpies ($\Delta_{\text{ROP}}H$), which account for the complex spatial configurations of polymer morphologies.⁸⁰ Despite the high accuracy of this method when benchmarked against experimentally determined $\Delta_{\text{ROP}}H$ values, the computational cost of running AIMD studies precludes its large-scale adoption.

To balance the competing demands of low computational cost and high accuracy in screening potential monomer units, we use the trained ALFABET3 model to predict bond BDEs for 192 industrially significant lactone monomer units collected from the literature. In particular, the lactone monomers considered here have been identified as candidates for the synthesis of complex hydrosoluble macromolecules via precise control of the number of oligomer units per chain (ie, limiting total chain mass to below 10,000 g/mol) and thus are of great interest in the manufacture of easily biodegradable novel green plastics. Monomer units are considered suitable for feedstock use if bond cleavage occurs at the C-O bond, enabling chain propagation.

The predicted BDEs are then ranked for each molecule to screen potentially useful monomer units. For 23 out of 192 monomers, the lowest BDE is the C-O ring bond, indicating the potential for use as a polymer feedstock. In 152 lactones, the lowest BDE was found to be a C-C ring bond; the remaining 17 lactones had C-S or O-O bonds as the lowest energy bond. The MAE of BDE prediction for all bonds was found to be 1.66 kcal/mol, while that of just the C-O ring bonds was found to be 1.85 kcal/mol (**Figure 2.6a**). Bonds with MAEs over 10 kcal/mol were classified as outliers; analysis of these outliers shows that two molecules contain all outlier bonds.

On examination of the first outlier molecule (termed Mol1 for brevity), it is clear that Mol1 consists of 27 heavy atoms, almost double that of the largest structures in ALFABET3 DB (a maximum of 14 heavy atoms). It is also noted that the structures generated by homolytic cleavage of outlier bonds 1 and 2 situate the diradical

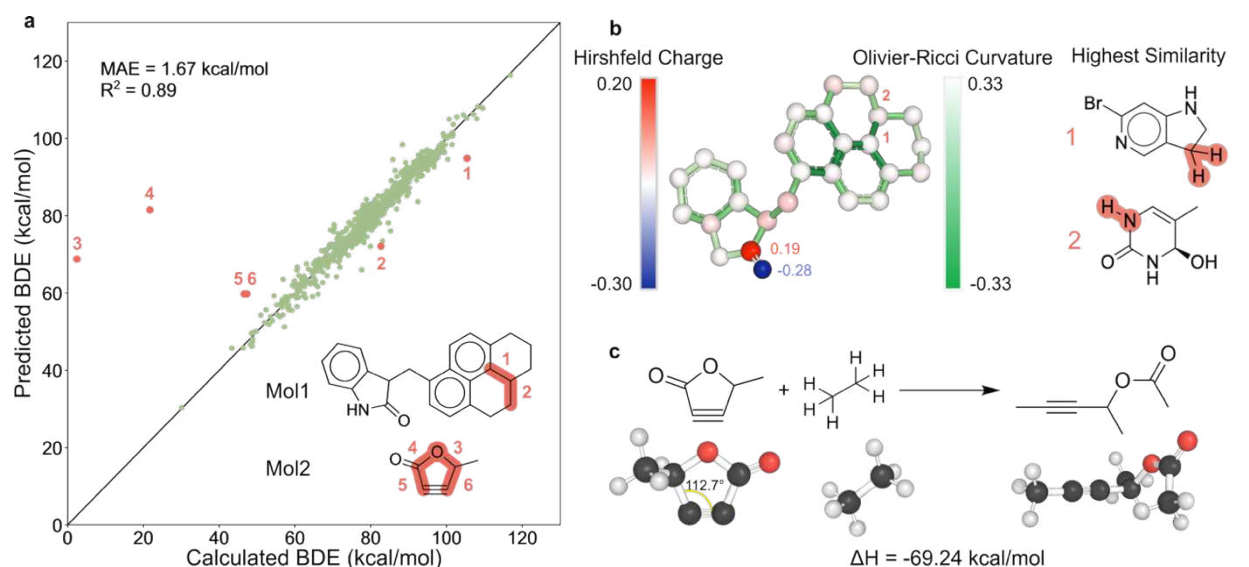


Figure 2.6: Model predictions for the homolytic ring opening bond dissociation enthalpy of individual lactone monomer bonds. a) Plot of model-predicted BDE vs DFT-calculated BDEs for 193 lactone monomer units. Outlier BDE prediction bonds (error > 10 kcal/mol) are shown in red with outlier bonds numbered. b) Left: Hirshfeld charges and Olivier-Ricci curvature for Mol1. Outlier bonds are shown in red numbers, corresponding to a). Computed Hirshfeld charges for the carbonyl group are shown in red and blue texts; other computed Hirshfeld charges are omitted for clarity. Right: Highest similarity test set bond state vectors to corresponding outlier bonds. c) Hyperhomodesmotic calculation results for the ring opening reaction of Mol2.

electrons in close proximity, leading to possible energetic destabilization in the electronic structure calculations. Further, two heteroatom motifs are found (a secondary amine and a carbonyl group), in a ring structure separated from the ring structure containing the outlier bonds by an alkyl chain. To quantify the effects of the heteroatom motifs, Hirshfeld charges⁸¹ were computed for the optimized structure using the M062X/def2TZVP level of theory. It was found that Hirshfeld charges are localized on the carbonyl group with charges of -0.28 and 0.19 on the C and O atoms, respectively (**Figure 2.6b**), thus indicating that electronic effects influencing bond strength are localized to the carbonyl constituent atoms.

To diagnose the mechanism of failure, we directly interrogated the model's internal representations by computing the cosine similarity of bond embedding vectors for outlier bonds 1 and 2 against all bonds in the ALFABET3 test set (**Figure 2.6b**, right). The most similar test-set bonds correspond to C-H and N-H bonds, respectively, indicating that the model fails to encode the unusual long-range electrostatic environment of these bonds, as the influence of the charged carbonyl motif is not reflected in what the GNN 'sees' at the outlier bond sites. The Ollivier-Ricci curvature of Mol1 (**Figure 2.6b**, left) provides a geometric interpretation consistent with this finding: all bonds along paths connecting the charged carbonyl to the outlier bonds carry negative curvature, a topological signature of over-squashing bottlenecks in message-passing GNNs that prevent distant atomic environments from being propagated to the prediction site. Together, these analyses indicate that the error in Mol1 arises from a combination of out-of-distribution long-

range electrostatic structure and insufficient message-passing depth to bridge the structural distance between the carbonyl and the predicted bond. Future efforts in database expansion, specifically targeting molecules with remote electron-withdrawing groups, and exploration of graph rewiring or higher-order GNN architectures to alleviate over-squashing, may help resolve this class of errors.

We hypothesize that errors for bond prediction in the second outlier molecule (Mol2) may be due to the large amount of ring strain energy stemming from the C_{sp} - C_{sp} triple bond present in the ring structure. To quantify this, hyperhomodesmotic calculations were performed using the scheme introduced by Wheeler and coworkers (**Figure 2.6c**).⁷⁹ The formation enthalpy of Mol2 and ethane was first calculated at the M062X/def2TZVP level of theory. The optimized Mol2 structure was then ring opened and capped with two methyl groups, thus preserving the total number and hybridization of carbon-carbon bonds. The resultant structure was then optimized and its formation enthalpy calculated. The ring strain energy was calculated to be 69.24 kcal/mol for the “lactyne” Mol2, much higher than the 8.6 – 9.5 kcal/mol values reported for the saturated γ -lactone analog.^{82,83} Since the ALFABET3 model does not incorporate molecular geometries as inputs, the geometric distortion induced by the presence of the triple bond is not directly incorporated into model predictions. In addition, molecules with triple bonds in ring systems are comparatively rare; only 22 such bonds exist in ALFABET3DB training set, representing only 0.003% of the training set. We note that the DFT-calculated values for outlier bonds 3, 4, 5, and 6 are extremely low: all bonds are below the 1st

percentile of DFT- calculated BDEs in the ALFABET3 DB (0.007, 0.05, 0.61, 0.65 percentiles, respectively). While such outlier bonds are uncommon, future work in improving model generalizability may build on such examples to further augment the model's training set. Useful outliers may be curated from specialized datasets such as the Mindless Benchmarking set curated by Korth and Grimme,⁸⁴ or identified through the use of targeted reinforcement learning strategies that use ML methods to identify optimal data.⁸⁵

2.4: Conclusion

Knowledge of the enthalpy change associated with ring opening is crucial for planning synthetic strategies and elucidating chemical mechanisms. Traditional computational approaches rely on using quantum chemical software to calculate these values, incurring significant computational cost and making these calculations inaccessible to chemists unfamiliar with these software packages. To rectify this, we developed the ALFABET3 model and ALFABET3DB dataset, building on our previously introduced ALFABET2 model to incorporate the ability to predict the bond dissociation enthalpy and bond dissociation free enthalpy for homolytic ring-opening reactions.

A dataset of DFT enthalpies of homolytic ring opening reactions was generated using the compounds in the ALFABET2 dataset, and the ALFABET2 model was retrained to incorporate this new data. The resultant ALFABET3 model achieves predictive accuracies of 0.86 kcal/mol and 0.84 kcal/mol for predicted BDE and BDFE, respectively, on test set BDEs ALFABET3DB. The accuracy for homolytic ring

opening reactions alone was found to be 1.54 kcal/mol, while that of non-ring homolytic bond dissociation was found to be 0.69 kcal/mol, indicating that additions of homolytic ring opening reactions to the ALFABET2 model do not significantly degrade predictive accuracy on the existing heterolytic bond cleavage target domain. A qualitative assessment of the model's ability to resolve intramolecular atomic environments is performed via t-SNE analysis, which shows that the model learns the roles of distinct chemical environments during training. An ablation study is conducted to assess the role of the ring-specific model inputs in influencing model predictions.

Finally, the predictive power of the new model is demonstrated through applications in ring opening polymerization reactions, highlighting its utility in predicting the viability of synthetic strategies. A thorough outlier analysis showed that model predictions with large errors are localized to molecules with atom and bond environments not represented in the ALFABET3DB training set. Future work will seek to rectify this gap by expanding the ALFABET3 dataset to encompass larger regions of chemical space and increasing the expressivity of the ALFABET3 neural network model to better incorporate subtle variations in chemical environments.

2.5: References

1. Lam, A. Y. S. & Li, V. O. K. Chemical Reaction Optimization: a tutorial. *Memetic Comp.* **4**, 3–17 (2012).
2. Tu, Z., Stuyver, T. & Coley, C. W. Predictive chemistry: machine learning for reaction deployment, reaction development, and reaction discovery. *Chem. Sci.* **14**, 226–244 (2023).

3. Gundelach, L., Fox, T., Tautermann, C. S. & Skylaris, C.-K. Protein–ligand free energies of binding from full-protein DFT calculations: convergence and choice of exchange–correlation functional. *Phys. Chem. Chem. Phys.* **23**, 9381–9393 (2021).
4. Merino, N. *et al.* Degradation and Removal Methods for Perfluoroalkyl and Polyfluoroalkyl Substances in Water. *Environmental Engineering Science* **33**, 615–649 (2016).
5. Li, J. *et al.* Site-specific allylic C–H bond functionalization with a copper-bound N-centred radical. *Nature* **574**, 516–521 (2019).
6. Szwarc, M. & Ghosh, B. N. Determination of the Dissociation Energy of the C–Br Bond by Pyrolysis. *J. Chem. Phys.* **17**, 744–745 (1949).
7. Fu, Y. *et al.* 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* **110**, 5874–5886 (2006).
8. Lazarou, Y. G., Prosmittis, A. V., Papadimitriou, V. C. & Papagiannakopoulos, P. Theoretical Calculation of Bond Dissociation Energies and Enthalpies of Formation for Halogenated Molecules. *J. Phys. Chem. A* **105**, 6729–6742 (2001).
9. Chan, B. & Radom, L. BDE261: A Comprehensive Set of High-Level Theoretical Bond Dissociation Enthalpies. *J. Phys. Chem. A* **116**, 4975–4986 (2012).
10. 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* **8**, 300 (2021).
11. Kitagawa, O., Miyaji, S., Yamada, Y., Fujiwara, H. & Taguchi, T. Iodine Atom Transfer [3 + 2] Cycloaddition Reaction with Electron-Rich Alkenes Using N-Tosyliodoaziridine Derivatives as Novel Azahomoallyl Radical Precursors. *J. Org. Chem.* **68**, 3184–3189 (2003).
12. Rawal, V. H. & Dufour, C. A General Strategy for Increasing Molecular Complexity: Photocycloaddition-Fragmentation Route to Functionalized Di- and Triquinanes. *J. Am. Chem. Soc.* **116**, 2613–2614 (1994).
13. Dvorak, C. A., Dufour, C., Iwasa, S. & Rawal, V. H. Rapid Synthesis of Di- and Triquinanes by Direct Reductive Fragmentation of Paterno–Büchi-Derived Oxetanes. *J. Org. Chem.* **63**, 5302–5303 (1998).
14. Albertsson, A.-C. & Varma, I. K. Recent Developments in Ring Opening Polymerization of Lactones for Biomedical Applications. *Biomacromolecules* **4**, 1466–1486 (2003).

15. Engel, J., Cordellier, A., Huang, L. & Kara, S. Enzymatic Ring-Opening Polymerization of Lactones: Traditional Approaches and Alternative Strategies. *ChemCatChem* **11**, 4983–4997 (2019).
16. Kalgutkar, A. S. Designing around Structural Alerts in Drug Discovery. *J. Med. Chem.* **63**, 6276–6302 (2020).
17. Hughes, T. B., Miller, G. P. & Swamidass, S. J. Modeling Epoxidation of Drug-like Molecules with a Deep Machine Learning Network. *ACS Cent. Sci.* **1**, 168–180 (2015).
18. Gehringer, M. & Laufer, S. A. Emerging and Re-Emerging Warheads for Targeted Covalent Inhibitors: Applications in Medicinal Chemistry and Chemical Biology. *J. Med. Chem.* **62**, 5673–5724 (2019).
19. Tran, L.-S., Herbinet, O., Carstensen, H.-H. & Battin-Leclerc, F. Chemical kinetics of cyclic ethers in combustion. *Progress in Energy and Combustion Science* **92**, 101019 (2022).
20. Sirjean, B., Glaude, P. A., Ruiz-Lopez, M. F. & Fournet, R. Detailed Kinetic Study of the Ring Opening of Cycloalkanes by CBS-QB3 Calculations. *J. Phys. Chem. A* **110**, 12693–12704 (2006).
21. Ramirez, A. L. B. *et al.* Mechanochemical strengthening of a synthetic polymer in response to typically destructive shear forces. *Nature Chem* **5**, 757–761 (2013).
22. Janissen, R. & Filonenko, G. A. Mechanochemistry of Spiropyran under Internal Stresses of a Glassy Polymer. *J. Am. Chem. Soc.* **144**, 23198–23204 (2022).
23. Zou, H. & Hastie, T. Regularization and Variable Selection Via the Elastic Net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67**, 301–320 (2005).
24. Nakajima, M. & Nemoto, T. Machine learning enabling prediction of the bond dissociation enthalpy of hypervalent iodine from SMILES. *Sci Rep* **11**, 20207 (2021).
25. iBonD. *Internet Bond-energy Databank (pKa and BDE)--iBonD Home Page* <http://ibond.nankai.edu.cn/>.
26. Li, W., Luan, Y., Zhang, Q. & Aires-de-Sousa, J. Machine Learning to Predict Homolytic Dissociation Energies of C–H Bonds: Calibration of DFT-based Models with Experimental Data. *Molecular Informatics* **42**, 2200193 (2023).
27. Guan, Y., Sowndarya, S. V. S., C. Gallegos, L., John, P. C. S. & S. Paton, R. Real-time prediction of ¹H and ¹³C chemical shifts with DFT accuracy using a 3D graph neural network. *Chemical Science* **12**, 12012–12026 (2021).

28. Kim, Y., Jung, H., Kumar, S., Paton, R. S. & Kim, S. Designing solvent systems using self-evolving solubility databases and graph neural networks. *Chem. Sci.* **15**, 923–939 (2024).
29. Kim, Y. *et al.* Physics-informed graph neural networks for predicting cetane number with systematic data quality analysis. *Proceedings of the Combustion Institute* **39**, 4969–4978 (2023).
30. Garg, V., Jegelka, S. & Jaakkola, T. Generalization and Representational Limits of Graph Neural Networks. in *Proceedings of the 37th International Conference on Machine Learning* 3419–3430 (PMLR, 2020).
31. Dwivedi, V. P. *et al.* Benchmarking Graph Neural Networks. Preprint at <https://doi.org/10.48550/arXiv.2003.00982> (2022).
32. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural Message Passing for Quantum Chemistry. Preprint at <https://doi.org/10.48550/arXiv.1704.01212> (2017).
33. Husic, B. E. *et al.* Coarse graining molecular dynamics with graph neural networks. *J. Chem. Phys.* **153**, 194101 (2020).
34. Joung, J. F. *et al.* Deep Learning Optical Spectroscopy Based on Experimental Database: Potential Applications to Molecular Design. *JACS Au* **1**, 427–438 (2021).
35. Mansimov, E., Mahmood, O., Kang, S. & Cho, K. Molecular Geometry Prediction using a Deep Generative Graph Neural Network. *Sci Rep* **9**, 20381 (2019).
36. Wen, M., Blau, S. M., Spotte-Smith, E. W. C., Dwaraknath, S. & Persson, K. A. BondNet: a graph neural network for the prediction of bond dissociation energies for charged molecules. *Chem. Sci.* **12**, 1858–1868 (2021).
37. Gelžinytė, E., Öeren, M., Segall, M. D. & Csányi, G. Transferable Machine Learning Interatomic Potential for Bond Dissociation Energy Prediction of Drug-like Molecules. *J. Chem. Theory Comput.* **20**, 164–177 (2024).
38. Batatia, I., Kovács, D. P., Simm, G. N. C., Ortner, C. & Csányi, G. MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields. Preprint at <https://doi.org/10.48550/arXiv.2206.07697> (2023).
39. Batatia, I. *et al.* A foundation model for atomistic materials chemistry. Preprint at <https://doi.org/10.48550/arXiv.2401.00096> (2024).

40. Tyzack, J. D., Hunt, P. A. & Segall, M. D. Predicting Regioselectivity and Lability of Cytochrome P450 Metabolism Using Quantum Mechanical Simulations. *J. Chem. Inf. Model.* **56**, 2180–2193 (2016).
41. Zhang, S., Zubatyuk, R., Yang, Y., Roitberg, A. & Isayev, O. ANI-1xBBB: An ANI-Based Reactive Potential for Small Organic Molecules. *J. Chem. Theory Comput.* **21**, 4365–4374 (2025).
42. Batzner, S. et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat Commun* **13**, 2453 (2022).
43. Batatia, I. et al. The design space of E(3)-equivariant atom-centred interatomic potentials. *Nat Mach Intell* **7**, 56–67 (2025).
44. Joshi, C. K. et al. All-atom Diffusion Transformers: Unified generative modelling of molecules and materials. Preprint at <https://doi.org/10.48550/arXiv.2503.03965> (2025).
45. 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* **11**, 2328 (2020).
46. Kim, S. et al. PubChem 2025 update. *Nucleic Acids Res* **53**, D1516–D1525 (2025).
47. Sterling, T. & Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **55**, 2324–2337 (2015).
48. V, S. S. S., Kim, Y., Kim, S., John, P. C. S. & Paton, R. S. Expansion of bond dissociation prediction with machine learning to medically and environmentally relevant chemical space. *Digital Discovery* <https://doi.org/10.1039/D3DD00169E> (2023) doi:10.1039/D3DD00169E.
49. Zhao, Y. & Truhlar, D. G. A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions. *J. Chem. Phys.* **125**, 194101 (2006).
50. Weigend, F. Accurate Coulomb-fitting basis sets for H to Rn. *Phys. Chem. Chem. Phys.* **8**, 1057–1065 (2006).
51. Cremer, D. Density functional theory: coverage of dynamic and non-dynamic electron correlation effects. *Molecular Physics* **99**, 1899–1940 (2001).
52. Wang, S. G. & Schwarz, W. H. E. Simulation of nondynamical correlation in density functional calculations by the optimized fractional orbital occupation approach: Application to the potential energy surfaces of O₃ and SO₂. *J. Chem. Phys.* **105**, 4641–4648 (1996).

53. *Handbook of Bond Dissociation Energies in Organic Compounds*. (CRC Press, 2002). doi:10.1201/9781420039863.
54. Sirjean, B., Fournet, R., Glaude, P.-A. & Ruiz-López, M. F. Extension of the composite CBS-QB3 method to singlet diradical calculations. *Chemical Physics Letters* **435**, 152–156 (2007).
55. Verdicchio, M., Sirjean, B., Tran, L. S., Glaude, P.-A. & Battin-Leclerc, F. Unimolecular decomposition of tetrahydrofuran: Carbene vs. diradical pathways. *Proceedings of the Combustion Institute* **35**, 533–541 (2015).
56. Herbinet, O. *et al.* Primary Mechanism of the Thermal Decomposition of Tricyclodecane. *J. Phys. Chem. A* **110**, 11298–11314 (2006).
57. Stuyver, T. *et al.* Do Diradicals Behave Like Radicals? *Chem. Rev.* **119**, 11291–11351 (2019).
58. Montgomery, J. A., Frisch, M. J., Ochterski, J. W. & Petersson, G. A. A complete basis set model chemistry. VII. Use of the minimum population localization method. *J. Chem. Phys.* **112**, 6532–6542 (2000).
59. Riplinger, C. & Neese, F. An efficient and near linear scaling pair natural orbital based local coupled cluster method. *J. Chem. Phys.* **138**, 034106 (2013).
60. Yamaguchi, K., Jensen, F., Dorigo, A. & Houk, K. N. A spin correction procedure for unrestricted Hartree-Fock and Møller-Plesset wavefunctions for singlet diradicals and polyradicals. *Chemical Physics Letters* **149**, 537–542 (1988).
61. Bernardi, Fernando. *et al.* MC-SCF study of the Diels-Alder reaction between ethylene and butadiene. *J. Am. Chem. Soc.* **110**, 3050–3055 (1988).
62. Kirkpatrick, J. *et al.* Pushing the frontiers of density functionals by solving the fractional electron problem. *Science* **374**, 1385–1389 (2021).
63. RDKit: Open source cheminformatics. <https://doi.org/10.5281/zenodo.591637> doi:10.5281/zenodo.591637.
64. Kim, Y. Yeonjoon-Kim/conj_path. (2025).
65. Berquist, E. *et al.* cclib 2.0: An updated architecture for interoperable computational chemistry. *J. Chem. Phys.* **161**, 042501 (2024).
66. GoodVibes: automated thermochemistry for... | F1000Research. <https://f1000research.com/articles/9-291/v1>.

67. 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 Account* **120**, 215–241 (2008).
68. Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics* **1**, 8 (2009).
69. Lakens, D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Front. Psychol.* **4**, (2013).
70. Maaten, L. van der & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**, 2579–2605 (2008).
71. Meyes, R., Lu, M., de Puiseau, C. W. & Meisen, T. Ablation Studies in Artificial Neural Networks. Preprint at <https://doi.org/10.48550/arXiv.1901.08644> (2019).
72. Tan, C., Xiong, S. & Chen, C. Fast and Controlled Ring-Opening Polymerization of Cyclic Esters by Alkoxides and Cyclic Amides. *Macromolecules* **51**, 2048–2053 (2018).
73. Jianming, R., Anguo, X., Hongwei, W. & Hailin, Y. Review – recent development of ring-opening polymerization of cyclic esters using aluminum complexes. *Designed Monomers and Polymers* **17**, 345–355 (2014).
74. P. Dove, A. Controlled ring-opening polymerisation of cyclic esters : polymer blocks in self-assembled nanostructures. *Chemical Communications* **0**, 6446–6470 (2008).
75. Romio, M. *et al.* Dispersity within Brushes Plays a Major Role in Determining Their Interfacial Properties: The Case of Oligoxazoline-Based Graft Polymers. *J. Am. Chem. Soc.* **143**, 19067–19077 (2021).
76. B. Doncom, K. E., D. Blackman, L., B. Wright, D., I. Gibson, M. & K. O'Reilly, R. Dispersity effects in polymer self-assemblies: a matter of hierarchical control. <https://doi.org/10.1039/C6CS00818F> (2017) doi:10.1039/C6CS00818F.
77. Perego, G., Cella, G. D. & Bastioli, C. Effect of molecular weight and crystallinity on poly(lactic acid) mechanical properties. *Journal of Applied Polymer Science* **59**, 37–43 (1996).
78. Dudev, T. & Lim, C. Ring Strain Energies from ab Initio Calculations. *J. Am. Chem. Soc.* **120**, 4450–4458 (1998).

79. Wheeler, S. E., Houk, K. N., Schleyer, P. v. R. & Allen, W. D. A Hierarchy of Homodesmotic Reactions for Thermochemistry. *J. Am. Chem. Soc.* **131**, 2547–2560 (2009).
80. Tran, H. *et al.* Toward Recyclable Polymers: Ring-Opening Polymerization Enthalpy from First-Principles. *J. Phys. Chem. Lett.* **13**, 4778–4785 (2022).
81. Hirshfeld, F. L. Bonded-atom fragments for describing molecular charge densities. *Theoret. Chim. Acta* **44**, 129–138 (1977).
82. Bond, J. Q., Martin Alonso, D., West, R. M. & Dumesic, J. A. γ -Valerolactone Ring-Opening and Decarboxylation over SiO₂/Al₂O₃ in the Presence of Water. *Langmuir* **26**, 16291–16298 (2010).
83. Emel'yanenko, V. N., Verevkin, S. P., Burakova, E. N., Roganov, G. N. & Georgieva, M. K. The thermodynamic properties of 4-pentenoic acid. *Russ. J. Phys. Chem.* **82**, 1521–1526 (2008).
84. Korth, M. & Grimme, S. “Mindless” DFT Benchmarking. *J. Chem. Theory Comput.* **5**, 993–1003 (2009).
85. Sridharan, B. *et al.* Deep reinforcement learning in chemistry: A review. *Journal of Computational Chemistry* **45**, 1886–1898 (2024).

CHAPTER 3: Molecular Properties from Fragment Interactions: Predicting Adiabatic Singlet-Triplet Gaps

3.1: Chapter Overview

In this chapter, we examine how the interactions of intra-molecular fragments across can be combined to model the adiabatic triplet energy gap (the thermodynamic driving force between the S_0 and T_1 electronic states of a given molecule, referred to as the S_0 - T_1 gap). Knowledge of this property enables design of efficient photocatalysts amenable to accelerating charge transfer or Dexter energy transfer based synthetic strategies. However, calculating the adiabatic S_0 - T_1 gap for a given molecules requires careful quantum mechanics calculations, making it both computationally expensive and challenging to perform for inexperienced users. In this work, we introduce a novel methodology for the prediction of adiabatic S_0 - T_1 gap based on chemical intuition around chromophores, the fragments of a molecule responsible for absorbing light. We construct a large-scale dataset of adiabatic S_0 - T_1 gaps for both whole molecule species and constituent chromophore fragments. This database is then used to build a range of predictive models for whole-molecule adiabatic S_0 - T_1 gaps; we find that a Δ -learning strategy combining individual predictions for both fragments and whole molecules achieves the most accurate and generalizable predictive performance. This model is applied to predict the site of ring closure in light driven [2+2] cyclization reactions.

This collaborative research was published in the RSC journal Chemical Science with equal lead contributions from Dr. Raúl Pérez-Soto, Dr. Mihai Popsecu, and myself, and additional contributions from Leticia Adao Gomes, Changyeob Lee, Elijah Shore, Dr. Steven A. Lopez, Dr. Robert S. Paton, and Dr. Seonah Kim. I helped in the development and analysis of neural network architectures and the high-throughput DFT pipeline, and helped to write and edit the manuscript for publication.

3.2: Introduction

With the re-emergence of photochemistry and photocatalysis as a major driving force of organic chemical synthesis at the start of the 21st century,^{1–6} the development of highly accurate computational methods for the prediction of photophysical properties has been deemed an essential driving force for the acceleration of the discovery of new chemical reactivity and development of new materials.^{7–11} More recently, however, the development of machine learning-based approaches has emerged as a promising alternative to highly computationally expensive quantum mechanical methods for predicting chemical properties, which are then used as an alternative to experimental data.^{12–15} One reason for the rise of machine learning in the last two decades has been an increase in the amount of data available, along with easier access to the most novel model architectures. However, the reliability of any model is highly dependent on the quality of the data used to formulate the model. Thus, the time-consuming step of data curation and validation becomes the rate-determining step in model development.

Due to the difficulties associated with collecting experimental data, models have relied on the construction of large datasets from computational simulations, based on methods such as Density Functional Theory (DFT).¹⁶ This approach sidesteps several of the limitations associated with data collection, quality assurance, and data interpretation. A plethora of models trained on such computational datasets have been successfully developed, with applications in various fields, including solubility,^{17,18} bond dissociation energy,^{19–21} synthesizability,²² and prediction of catalytic efficiency.^{23,24} Thanks to the relative ease of constructing these virtual datasets, many research groups have constructed independent datasets for model development, frequently competing to simulate the same properties. This in turn raises a key question: how can we compare the diversity and quality of competing datasets beyond just the numerical range of the property in question? How can we enhance new datasets by leveraging existing ones to cover a broader range of the chemical space?

One such area of high interest to both the experimental and theoretical communities involves the development and curation of large datasets of photophysical properties. Specifically, within the context of both photoredox and energy transfer catalysis, knowledge of the catalyst and substrate triplet energies is essential for calculating excited-state redox potentials. Also, it facilitates the prediction of reactions amenable to Dexter energy transfer. Within this regard, DFT computation of the adiabatic triplet energy gap (i.e., the thermodynamic driving force between S_0 and T_1 state) has become a central method for determining molecular

triplet energies, largely due to the experimental challenges involved. To this extent, the adiabatic triplet energy has become a key photophysical property, featured in numerous databases, such as the Verde Materials DB.²⁵ Similarly, Glorius and coworkers have recently published the first machine learning model, EnT-Decker, for predicting the adiabatic S_0 - T_1 value, trained on an in-house-generated DFT dataset of 34,848 molecules.²⁶ In this work, not only do they demonstrate that Graph Neural Network models outperform other simpler models, but also showcase the practical application of the developed models and build a user-friendly platform for chemists to use their models.

However, as demonstrated in this work, high-throughput DFT calculations of adiabatic S_0 - T_1 gaps pose challenges when applied to molecules with multiple functional groups. In such cases, exciton localization can vary semi-randomly across different regions of the molecule, leading to significant energy differences and introducing substantial error in model construction. In this work, we present a new fragmentation algorithm designed to address this issue. This algorithm enables the training of a Δ -learning model on a new dataset of 46,415 molecules (ALFAST-DB), following a similar protocol to EnT-Decker. Additionally, the algorithm offers a novel approach to comparing the chemical compositions of different databases and evaluating the transferability of ML models trained on them.

3.2.1: Exciton Localization Problem in Multichromophore Systems:

As highlighted earlier, DFT optimizations on excited states are particularly challenging for molecules with multiple chromophores due to the issue of exciton

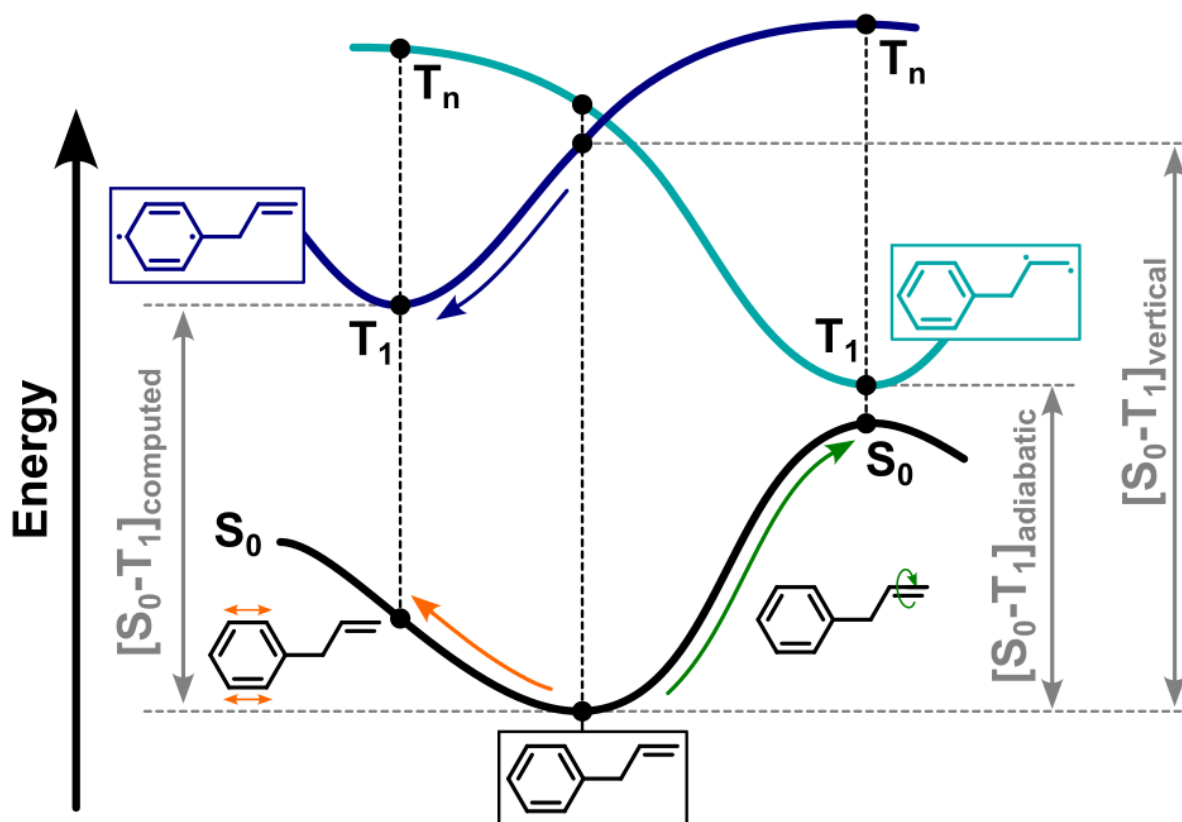


Figure 3.1: Diagram describing the computational structural optimization on the T1 surface starting from a ground state geometry

localization. Taking allylbenzene as an example, the standard approach begins by using the optimized geometry of the molecule in its ground state (S_0) as the starting point for optimizing the corresponding triplet state (T_1). The geometry optimization on the T_1 surface proceeds from this initial geometry, and the algorithm directs the structure toward the nearest local minimum based on the gradient of the surface curvature.

In the case of allylbenzene, this initial optimization leads to a structure where the triplet diradical localizes on the phenyl group, as illustrated in **Figure 3.1**. However, this localization isn't always unique. If the molecule is initially excited to a higher triplet state (T_n) instead, the system may follow a different potential energy

surface. This new surface can intersect the initial T_1 surface and, in some cases, drop below it in energy, effectively becoming the new T_1 state. Such a scenario results in the localization of the spin density on a different region of the molecule.

For allylbenzene, the system can alternatively find a T_1 minimum where the spin density localizes on the alkene group. While the optimization process preferentially leads to the excitation of the aryl ring, this alkene-localized minimum is also a valid solution. It is energetically more favorable, resulting in a lower adiabatic triplet energy. In this case, when computed at the M06-2X/def2-TZVP level using Gaussian 16,²⁷ the adiabatic triplet energy (62.4 kcal/mol) is lower by an impressive 23.3 kcal/mol compared to the phenyl-localized minimum (85.7 kcal/mol).

This difference in localization presents a significant challenge for automated excited-state optimization in database construction. In systems with multiple chromophores, it becomes nearly impossible to predict a priori where the triplet state will localize. As a result, this uncertainty complicates the construction of a coherent dataset, often leaving the data effectively "unlabeled" in terms of which chromophore is involved in the triplet state localization. This type of unstructured data introduces confusion into machine learning models, as the models may struggle to learn from or differentiate between cases involving different chromophores. Furthermore, it raises an important question regarding the goal of adiabatic triplet energy predictions: is the objective to predict the global minimum energy state, or should the models aim to predict all possible local minima? Defining this objective becomes crucial in ensuring the predictive power and applicability of the machine learning models.

One potential solution, as implemented by the EnT-Decker authors, is to provide a prediction of spin density in conjunction with the adiabatic triplet energy value. This strategy adds valuable information to the user, offering insight into where the triplet localization occurs within the molecule. However, while this represents an important step forward, it does not fully resolve the underlying issue of training models on unlabeled data. Since the spin density localization is not systematically assigned to a specific chromophore or fragment, the model is still learning from data that may not be explicitly categorized.

Moreover, since these spin density models were trained independently of the energy predictions, there is the potential for inconsistencies. In some cases, the energy and spin density predictions may not align, potentially leading to misleading conclusions for the user.

3.3: Results and Discussion

3.3.1: Molecular Fragmentation Algorithm

To address the challenges associated with exciton localization and unlabeled data in multi-chromophore systems, a first step is to develop a workflow for identifying potentially relevant sites where the exciton can be localized. Here we developed a fragmentation algorithm that generates chemically meaningful fragments from complex molecules, akin to the one described by Ertl in 2017.²⁸ This approach was inspired by the concept of functional groups in organic chemistry, which serve as the building blocks of molecules and represent key sites of chemical

reactivity. Functional groups, often composed of π -bonds and heteroatoms linked by resonance, are rich in π -electrons and lone pairs. These electron-rich regions are frequently the most likely sites for excitations due to their lower energy gaps between π - π^* and n - π^* orbital levels.²⁹ By leveraging this understanding of functional groups, this algorithm systematically breaks down molecules into these reactive components based on extended conjugation paths, leading to more structured data that can be effectively used for machine learning. The steps of the algorithm are illustrated below (**Figure 3.2**), and include the following key processes:

1. Identify all relevant heteroatoms based on a predefined list, as well as all carbon atoms containing a double, triple, or aromatic bond. Specifically for this work, we considered the following candidate heteroatoms: N, O, S, P, Se, F, Cl, and Br.
2. For each candidate atom, a list is constructed that contains these atoms. Then, the atoms directly bonded to each candidate atom are recursively added. If the connecting atoms are also candidate atoms, the process continues to add their connected neighbors until no candidate atoms are found. This recursive process is stopped when either there are no more atoms or the last atom added was a non-candidate atom.
3. Once no candidate atoms remain in reach of the molecule, duplicated fragments are removed.
4. Finally, for database storage analysis and curation, each fragment is transformed into its SMILES representation. During this step, the end-atoms of

the fragments are implicitly capped with H to match their valence. As all valid SMILES are valid SMARTS, H-capping allows the generation of smaller, valid, typically closed-shell molecules containing the key functionality of the fragment.

Following the example highlighted above (**Figure 3.2**), the outcome of the fragmentation scheme as applied to 3-(2-(4,5-dichlorocyclohex-2-en-1-yl)ethyl)-1,1-

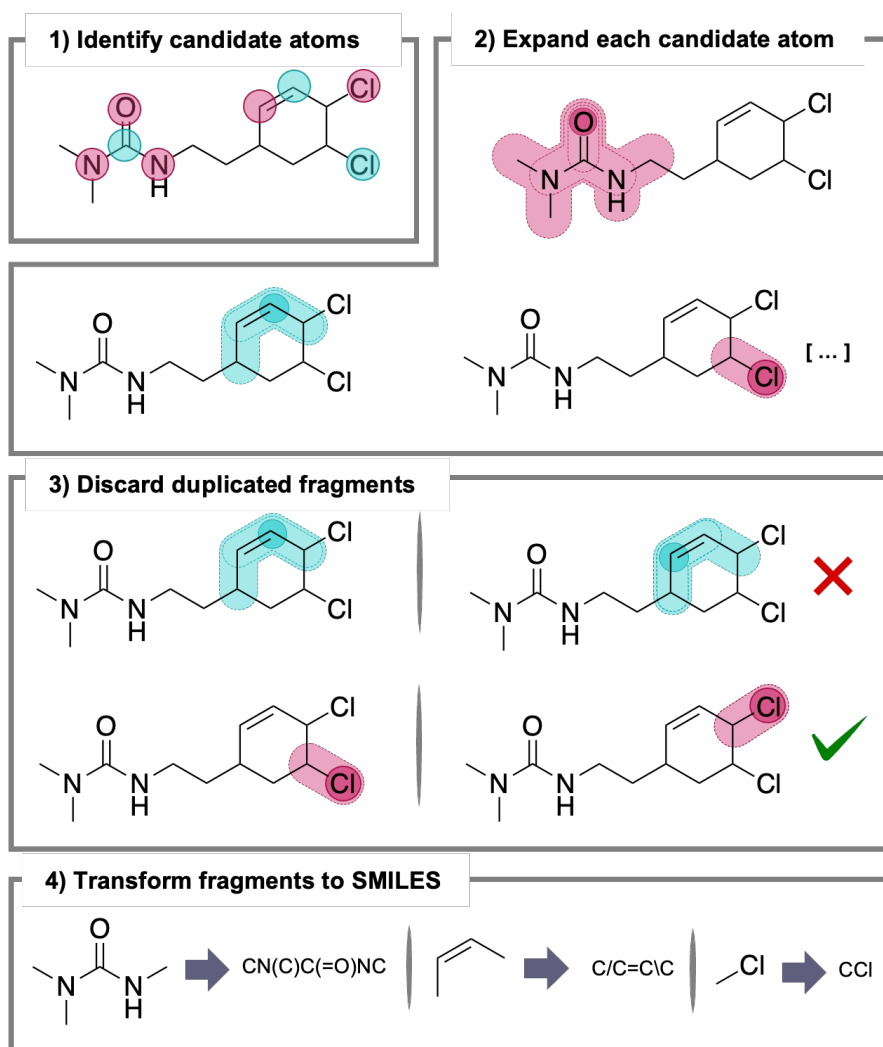


Figure 3.2: Scheme and example of the application of the fragmentation algorithm used in the present work.

dimethylurea will lead to the identification of four distinct functional groups, namely: a trisubstituted urea, a disubstituted (*Z*)-alkene, and two alkyl chlorides.

3.3.2: Databases Overview and Curation

The present work utilizes three separate datasets comprising adiabatic S_0 - T_1 energy gaps. One such dataset is the EnT-DB database, developed by Glorius and coworkers, which was used to develop the EnT-Decker predictive model. The EnT-DB dataset consists of DFT-computed adiabatic S_0 - T_1 energy gaps for a total of 34,848 molecular species. Energies are computed at the ω B97x-D3/def2-TZVPP:PCM(CH_3CN)/B3LYP-D3/def2-SVP:PCM(CH_3CN)^{30–35} level of theory, using the ORCA software package.³⁶ Molecules were selected to maximize chemical space coverage of medically relevant structures, and were selected from the ZINC dataset³⁷ clade of commercially available species with molecular weights under 325 Da. The size distribution of molecules in the dataset is roughly symmetric, with a median of 30 atoms. Calculated adiabatic S_0 - T_1 gaps span a broad range, with a median of 69.4 kcal/mol (**Figure 3.3**, EnT-DB).

The second database used in this work is the Verde Materials Database previously developed by Lopez and coworkers (Verde-DB). The version of the dataset used here is more comprehensive than the one previously published; this version of the Verde-DB dataset contains 3,286 molecules, compared to the 1,500 molecules in the original work. The S_0 - T_1 adiabatic gaps in this database are also obtained using DFT, but computed at the M06/6-31+g(d,p):PCM(CH_3CN)^{38–42} level of theory. In contrast to EnT-DB, which focuses solely on adiabatic S_0 - T_1 gaps as a target property

and prioritizes a broad chemical scope, Verde-DB is more focused on providing broader coverage of properties related to photoredox catalysis, such as excited-state reduction and oxidation potentials, or S_0 - S_1 energy gaps, at the cost of a narrower chemical scope. This chemical scope is more focused on molecules used in commercial applications, with additions of computationally generated derivatives. The molecular size distribution (**Figure 3.3**) has the same median as the EnT-DB dataset (30 atoms), while showing a pronounced tail towards larger molecules. Due to the specialized chemical scope targeted in the Verde-DB dataset, it exhibits a

strong bias towards molecules with a small S_0 - T_1 adiabatic gap, with a median value of 28.2 kcal/mol.

Finally, we develop our database targeting a broad chemical scope, following a protocol where the T_1 geometry is optimized using the optimized ground state structures as a starting point. Previously, we compiled a large database of DFT-computed structures for predicting bond dissociation enthalpies, which led to the creation of the ALFABET model. The ALFABET dataset was designed to cover a broad range of organic chemical space and comprises molecules with atoms of C, H, N, O, S, Cl, F, P, Br, and I. Building on this previous work, we compute the S_0 - T_1 adiabatic

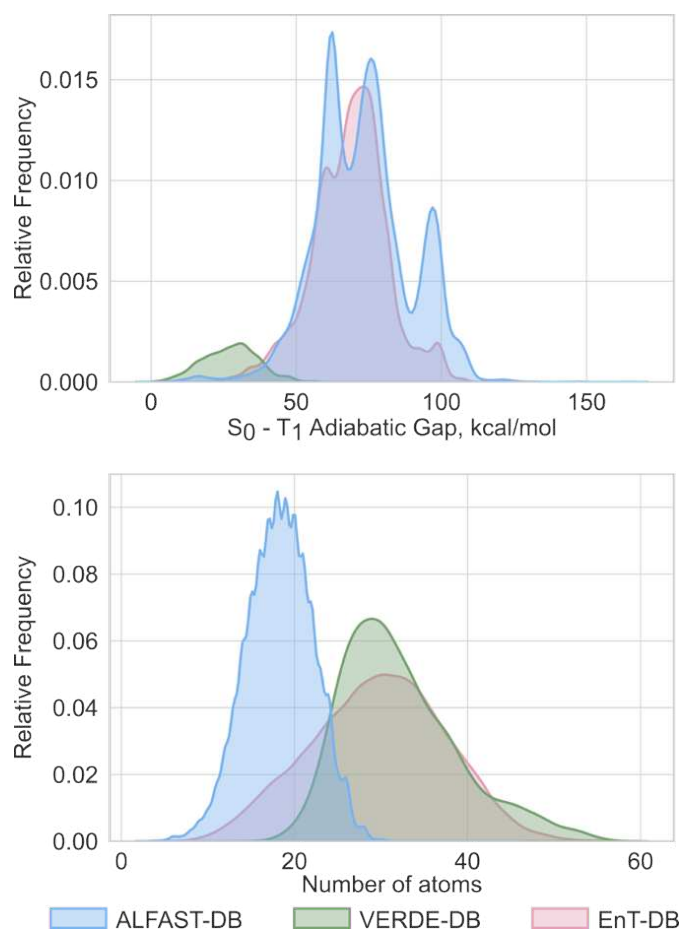


Figure 3.3: Total number of atoms and S_0 - T_1 adiabatic gap distributions of the 3 presented datasets.

gaps for molecules of interest in this database. This initial set comprises 57,736 molecules after removing radicals and retaining only those molecules with double, triple, or aromatic bonds. For the level of theory used in computation, we employ the M06-2X/def2-TZVP level of theory, in contrast to the one used to construct the EnT-DB dataset. Although the M06-2X functional had been identified as optimal for adiabatic triplet energy calculations by the Glorius group, challenges with convergence and computational costs led to the ω B97x-D3 functional being used instead. While recent work suggests that benchmarking of adiabatic triplet energies against experimental measurements may not be representative of physical reality and can lead to very large errors on systems that undergo high structural reorganizations, benchmarking against the adiabatic values computed at the “gold standard” DLPNO-CCSD(T) method with the cc-PVTZ basis set^{43–48} also showed that the M06-2X provides the best accuracy. Furthermore, the S_0 ground states in the original ALFABET dataset were computed at the M06-2X/def2-TZVP level, allowing us to avoid the significant computational cost mentioned by Glorius and coworkers; we thus computed the T1 minima at the same level of theory. Out of the 57,736 starting molecules, 53,137 achieved successful convergence, which were further curated down to 46,415 molecules after further quality assurance checks (see SI for further details). As shown in **Figure 3.3**, this database is more focused on smaller molecules than the EnT-DB or Verde-DB datasets, with a median number of atoms of 18. However, we maintain a similar distribution of S_0 -T₁ adiabatic gaps to EnT-DB, with a median of 72.6 kcal/mol. We refer to this new dataset as the ALFAST-DB.

3.3.3: Database Analysis and Comparison

Using the framework of the proposed fragmentation scheme, we analyze the fragment compositions of the three databases discussed above. First, fragments that did not contain a double, triple, or aromatic bond (using SMARTS pattern matching) were discarded. These fragments were then used to obtain the possible exciton localizations within each molecule across the databases (*vide infra*). This fragment-based breakdown allows for a more quantitative evaluation of the relevant chemical diversity and overlap between the three datasets.

DATABASE	ALFAST	EnT	Verde
ALFAST	15,410	2,146	0
EnT	2,146	18,909	0
Verde	0	0	3,286

Table 3.1: Number of shared fragments between databases. The diagonal values correspond to the total number of fragments of each database.

First, we evaluate the fragment overlap between databases (**Table 3.1**). We observe that while the ALFAST-DB has a larger number of molecules (46,415 vs 34,848), it has a slightly smaller diversity of unique fragments than the EnT database (15,410 vs 18,909). Between these two databases, there are 2,146 shared fragments, while neither database contains shared fragments with the Verde-DB. Next, we analyze the number of fragments formed for each molecule and quantify the percentage of shared fragments within this classification scheme (**Table 3.2**). Interestingly, a large number of molecules in each dataset contain a fragment from the small shared subset of 2,146 fragments, at 62.6% and 51.6% for ALFAST-DB and EnT-DB, respectively, indicating a significant overlap in chemical space. This overlap is thus more significant than one would expect if only the identity of the complete shared molecules (of which there are 1530) is considered. Almost all unique fragments

correspond to single-fragment molecules, and the EnT-DB has a higher composition of multi-fragment molecules (32.4% vs 15.5%). Notably, the Verde-DB is constituted solely of mono-fragment molecules (thus not shown for brevity). We posit that this may be a consequence of the more specialized scope considered when constructing the Verde-DB.

It is important to note that these statistics do not provide a basis for claiming the superiority of one dataset over the other. On the one hand, a larger number of unique fragments can be taken naively to represent increased chemical diversity of the constituent chromophores. On the other hand, a larger ratio of molecules to unique fragments provides a greater diversity of substituent effects on the chromophores, thus potentially improving the generalization characteristics of derived models.

	ALFAST-DB		EnT-DB	
	Single	Multiple	Single	Multiple
Unique	37.2% (17,289)	< 0.1% (35)	47.8% (16,649)	0.6% (212)
Shared (2,146)	47.2% (21,941)	15.4% (7,150)	19.8% (6,911)	31.8% (11,076)
Total	84.5% (39,230)	15.5% (7,185)	67.6% (23,560)	32.4% (11,288)

Table 3.2: Molecule breakdown of the ALFAST-DB and EnT-DB. Molecules are classified as either single-fragment or multiple-fragment, and based on whether at least one of the fragments is within the subset of fragments shared across the two databases (Shared) or all fragments of the molecule are unique. Numbers in parentheses are the absolute number of molecules.

Finally, to assess the effect of exciton localization mentioned previously, we utilize the Mulliken spin densities at the optimized T_1 states included in the ALFAST-DB to compute the total spin of each fragment obtained. The fragment with the highest spin was then selected as the “main fragment” of the molecule, effectively labeling each molecule with a main fragment, and a total spin of the main fragment. The distribution of these main fragment spins across the ALFAST-DB is shown in

Figure 3.4. While the majority of the database has a localized triplet, where the spin of the main fragment is over 1.7, three notable outlier regions require further discussion: molecules that have a main fragment with a spin over 2, molecules where the main fragment spin is between 0.7 and 1.3, and molecules whose main spin is below 0.65. We note that the choice of 1.7 as a threshold for triplet spin localization is an arbitrary value chosen for analysis and discussion of the generated database and should not be construed as a metric to define triplet spin localization. We leave rigorous benchmarking of this metric for future study; all computed values for fragment spins are available in the database files supplied in the Zenodo repository linked in Appendix B (DOI: 10.5281/zenodo.16563830).

We posit that the first region emerges as an artifact of the truncation during the total spin density calculation of the fragment. The second and third regions, although having more consequential root causes, comprise less than 1% of the molecules in ALFAST-DB (approximately 500 molecules). Thus their impact on the overall database is considered minor. The second region (between 0.7 and 1.3) corresponds to molecules where a charge transfer state was likely obtained. The third region, however, emerges as molecules whose main fragment was incorrectly assigned. This appears in cases where the obtained minima in the T_1 surface were not localized in a fragment containing a double, triple, or aromatic bond.

To illustrate this electron delocalization problem further, we analyzed representative examples from each of the outlier regions described above (**Figure 3.4B, C**). Examples were chosen such that the localization of the exciton can be

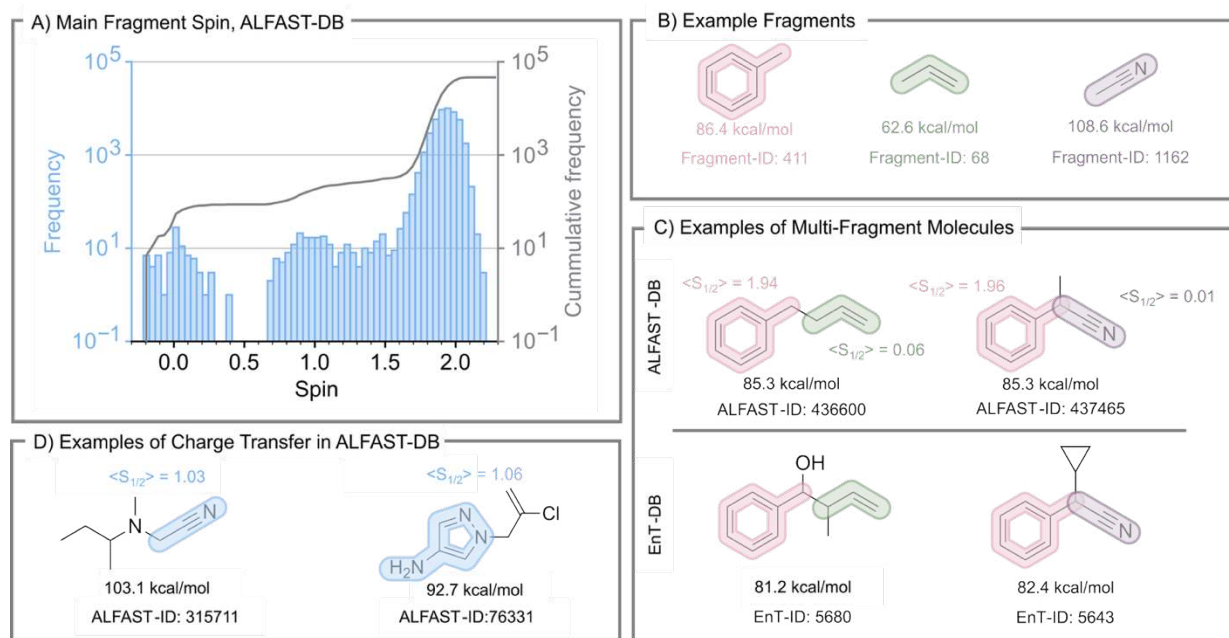


Figure 3.4: (A) Absolute (blue histogram) and cumulative frequency (grey line) of the distribution of the main fragment's spin within the ALFAST-DB. (B-D) Example DFT-computed S_0 - T_1 adiabatic gaps for fragments and for multi-fragment molecules within the ALFAST-DB and the EnT-DB. The spin of each fragment was computed from the atomic Mulliken Spin Densities. The example fragments (B) were computed at the same theory level as ALFAST-DB. We note that the EnT-Database does not explicitly include an ID number for each molecule; ID numbers were assigned in this work for organization.

confirmed based on the spin of the fragment. In addition, some example molecules where a charge transfer state was obtained are highlighted (**Figure 3.4D**). As shown in **Figure 3.4B**, the adiabatic S_0 - T_1 gap of toluene is lower than that of acetonitrile. As such, a molecule containing both moieties separated by an alkyl carbon chain should be able to access the triplet state of both moieties. In the selected examples (ALFAST-DB and EnT-DB, right column, **Figure 3.4C**), the computed value is more similar to the adiabatic gap of toluene. In the ALFAST-DB example, we can confirm that the exciton location is the phenyl ring. However, in this example, the overall high-throughput DFT computational protocol can be considered successful in approximating the adiabatic S_0 - T_1 gap, which is understood as the lowest free energy difference between the S_0 free energy surface and the T_1 free energy surface.

On the other hand, the examples highlighted with phenyl and alkene moieties (ALFAST-DB and Ent-DB, left column, **Figure 3.4C**) show the opposite result: out of the two exciton locations the minima with the phenyl-localized triplet was favored, which is higher in energy than that localized in the alkene (based on the difference between toluene and propene fragments), thus illustrating an unsuccessful application of the high-throughput DFT computational protocol. When the structure obtained from the S_0 optimization is used as a starting structure for the T_1 optimization, the resultant T_1 minima localizes the triplet in the phenyl moiety. Instead, the starting structure must be manually distorted to an alkene-like geometry (H-C-C-H dihedral angle of 90°) to localize the triplet on the alkene moiety and obtain a lower energy T_1 minima. Further work is necessary to elucidate which factors may bias the geometry optimization towards either of the T_1 minima, as well as how these factors affect the numerical energy value, in order to develop a high-throughput DFT protocol that finds the “true” lowest energy triplet in an automated manner without requiring manual geometry distortion.

The final charge transfer examples (**Figure 3.4D**) show two different cases. In the rightmost example, there are two clear, disconnected π -systems where the spin is localized, whereas the other example shows a case where not all the spin is localized in a region of the molecule containing a double, triple, or aromatic bond.

3.3.4: Architecture Selection and Model Extrapolation

Prior work by numerous groups has shown the effectiveness of Graph Neural Network models in predicting molecular properties. Here, we tested various message-passing neural network architectures based on the previously successful ALFABET model (further details are provided in Appendix B). Model evaluation was performed using a 64:16:20 train:validation: test split of the ALFAST-DB with randomized splitting. A 5-fold cross-validation strategy was used within the train-validation subsets. The validation loss function (mean absolute error, MAE) was tracked during training and used to select the best model for each validation fold. The final architecture and training strategy presented a cross-validation MAE (aggregated across folds) and test-set MAE (averaged across folds) of 1.94 and 2.00 kcal/mol, respectively. For the final predictions and the presented parity plots, the model was trained using the combined training and validation sets, resulting in a test-set MAE of 1.94. When trained on the EnT-DB using a similar splitting strategy, the cross-validation MAE and the averaged test-set MAE were 2.04 and 2.06 kcal/mol, respectively, with a final model test-set MAE of 2.37 kcal/mol. Compared to the random splitting results of the EnT-Decker GNN models, with MAEs of 1.97 and 1.93 kcal/mol for the Chemprop-D-MPNN and AttentiveFP-GNN models, respectively, it can be observed that our selected architecture achieves similar accuracies, regardless of the database on which it was trained.

After the architecture selection, the performance of the models on the other unseen datasets was evaluated to assess model extrapolation capabilities. As each

database is generated using a different DFT level of theory, a degradation in model performance is to be expected. This increase in model error stems from two separate causes: the uncertainty of the ML model in making predictions on unseen data, and the error incurred when comparing S_0 - T_1 adiabatic values computed using different DFT theory levels. To attempt to separate these two sources of error, each database's final model (trained with the respective combined training and validation sets) was then used to predict the values of each dataset (**Figure 3.5**). The prediction of each final model on its respective test set (**Figure 3.5 aa, bb, and cc**) provides an estimate of the model's uncertainty. Conversely, the predictions on the complete unseen datasets (**Figure 3.5ab, ac, ba, ca, bc, and cb**) provide insight into how the model extrapolates.

From this analysis, two clear trends are observed. First, when the architecture is trained on either ALFAST-DB or EnT-DB (i.e., the datasets with more general chemical space coverage) and used to predict the other (**Figure 3.5ab and ba**), a clear increase in the MAE is observed, while some degree of correlation (marked by R^2) is retained, on the other hand, when the training and prediction are performed using the more specialized Verde-DB and the other two remaining databases (**Figure 3.5ac, bc, ca and cb**) respectively, the models fail drastically at extrapolation. This behavior reproduces known trends regarding the importance of training set chemical space diversity on model performance.

To estimate the magnitude of the error incurred when comparing values computed with different DFT functionals, we analyze shared molecules between

ALFAST-DB (in gas phase) and EnT-DB (in acetonitrile). For this task, the canonical SMILES representations of molecules from both databases were obtained and used to search for common molecules, of which 1,530 were identified. An MAE and R^2 of 3.36 kcal/mol and 0.88, respectively, were obtained when comparing the S_0 - T_1 adiabatic gap of these 1,530 molecules (see **Figure S7**, Appendix B); a linear fit was also performed, yielding an MAE and R^2 of 2.63 kcal/mol and 0.92, respectively (see **Figure S7**, Appendix B). It should be noted that this MAE represents errors derived from directly comparing the results of different DFT methodologies, but also molecules where each method computed a different exciton location, molecules whose exciton is located on the same fragment but geometry optimizations led to different local minima, or molecules that would be more drastically stabilized in the triplet state by the acetonitrile solvent.

If then, the uncertainties in both model predictions and the underlying DFT methodology are combined, the model trained on ALFAST-DB achieves a net uncertainty of 5.30 kcal/mol, which is comparable to the MAE of 5.51 kcal/mol obtained when using the same model to make predictions on the EnT-DB (**Figure 3.5ab**). The same can be observed when the model trained on the EnT-DB extrapolates to the ALFAST-DB. This observation stays the same even when the cross-validated MAEs of the models are considered (1.94 and 2.04 kcals/mol for ALFAST-

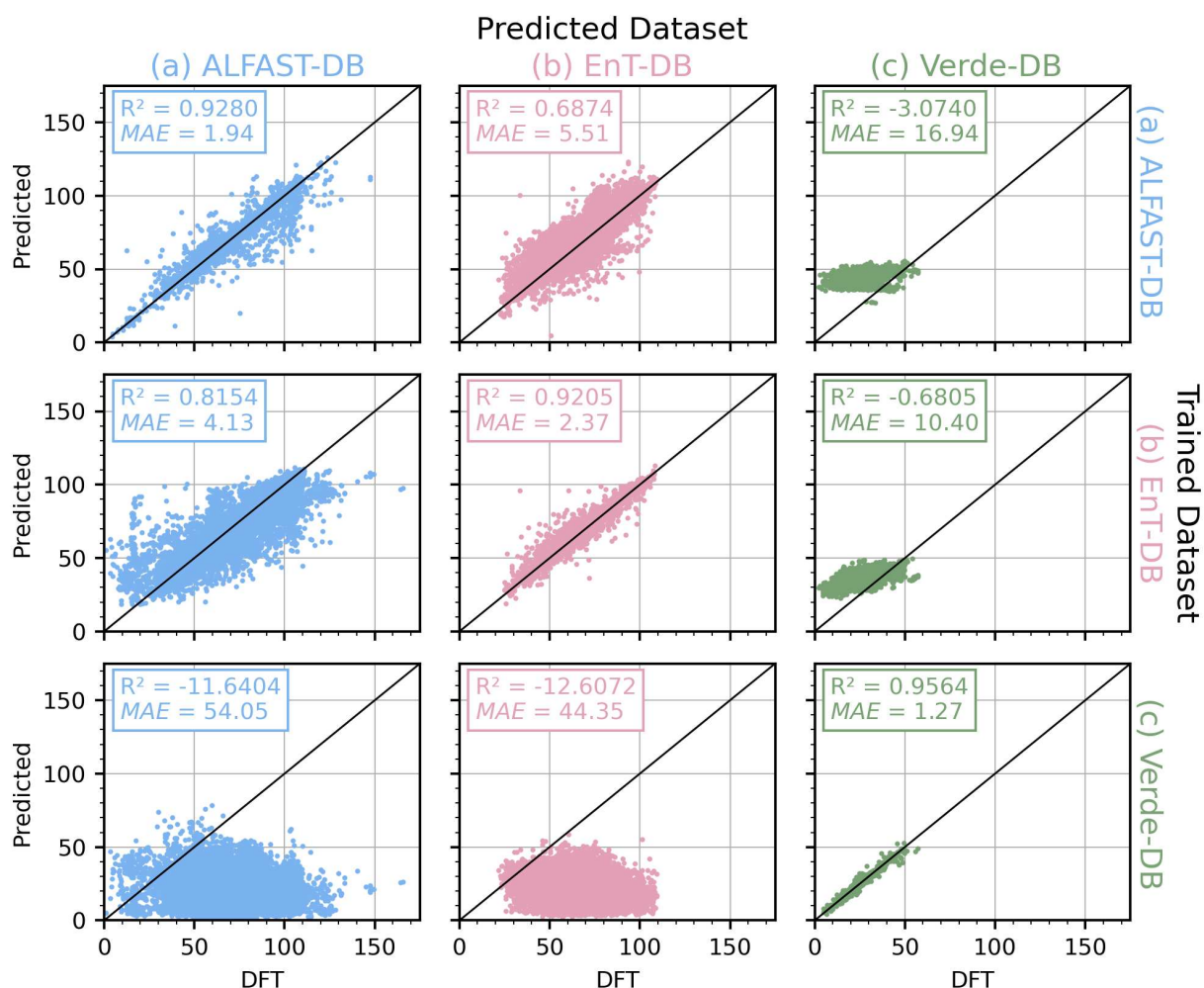


Figure 3.5: Model extrapolation across datasets. Parity plots of adiabatic S_0 - T_1 predictions in kcal/mol of each dataset (column) of the same model architecture trained each dataset (row). For example, parity plot “ab” corresponds to the architecture trained with the combined training and validation set of ALFAST-DB (80% of the database) and shows the prediction of EnT-DB. The predictions on the diagonal plots only shows the molecules in the test set (remaining 20% of the database) of each respective database. The metrics, R^2 and MAE values, correspond to the data presented in each plot. MAE values are in kcal/mol.

DB and EnT-DB, respectively). We note that the combination of both errors provides an approximation to the actual contribution to the error of the cross-dataset prediction. When we corrected the S_0-T_1 predicted using the MPGNN model with the linear fit derived from the shared molecules, a decrease in the MAE (**Figure S8** in Appendix B), although limited, can be observed.

In contrast, we observe that such “symmetric” behavior is not evident when considering the Verde-DB. As seen from the analysis above, the Verde-DB shares no fragments nor molecules with either of the other two databases, is constituted only of mono-fragment molecules, and is an order of magnitude smaller than the other two databases due to its more specialized nature.

Both the lack of shared fragments and the number of mono-fragment molecules may be responsible for the poor extrapolation to and from Verde-DB. To provide further insight into this, a detailed breakdown of errors in the ALFAST-DB-trained and EnT-DB-trained models (**Figure 3.6**) based on fragment presence was obtained.

First, it is reaffirmed that the absolute errors are significantly smaller and the distribution is narrower (**Figure 3.6**) when predicting the dataset on which the model was trained, as compared to the database to which it is being extrapolated. However, a clear difference is evident between the shared and non-shared fragment error distributions of the extrapolated dataset (**Figure 3.6**), where molecules without a shared fragment exhibit a larger error than those with a shared fragment. This involves

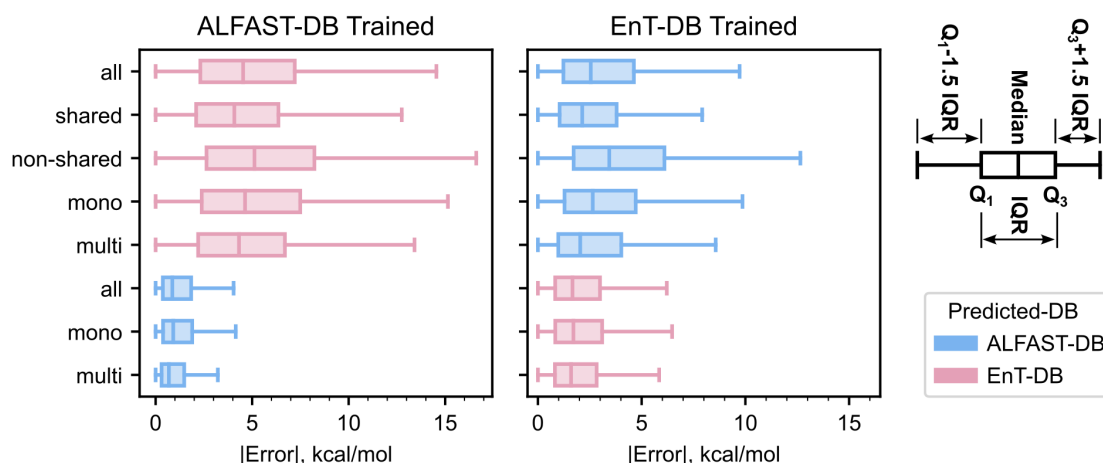


Figure 3.6: Boxplots with total (all) and marginal absolute error distributions of molecules based on whether they have a shared fragment between databases (shared vs non-shared) or whether they are mono- or multi-fragmented molecules (mono vs multi). Q1 and Q3 are the first and third quartile respectively and IQR stands for interquartile range. The error distributions of the dataset used for training are the distributions of the database's test set, while the distributions for the extrapolated database include the full extrapolated database.

a shift of the median of approximately 1-2 kcal/mol and a broadening of the error distribution. On the other hand, such consistent effects are not observed between mono- and multi-fragment distributions; interestingly, both models exhibit slightly better performance for multi-fragment molecules than for mono-fragment molecules, regardless of the predicted dataset.

One interpretation of this trend is that the model better fits multi-fragment molecules, as they are more information-dense. If indeed this is the case, the model should also exhibit lower errors for molecules that suffer from the exciton localization problem discussed earlier. It is to be noted that this interpretation assumes that the differences between the mono-fragmented and the multi-fragmented error distributions are not an artifact of the subset size (see molecule breakdown in Error! Reference source not found.), nor due to the lower chemical diversity of the main fragments in the multi-fragment molecules. Nonetheless, the consistent difference when extrapolating trained models to molecules with shared fragments vs molecules

with non-shared fragments, and the poor extrapolation to and from Verde-DB point towards the shared-fragment overlap as the predominant cause of the large prediction errors.

It is thus clear that one method to improve the generalizability of a model trained on a dataset of properties dependent on the exciton location is to increase the chemical scope of the fragments, instead of increasing the frequency of the fragments already present in the dataset.

3.3.5: Fragment-Based Models

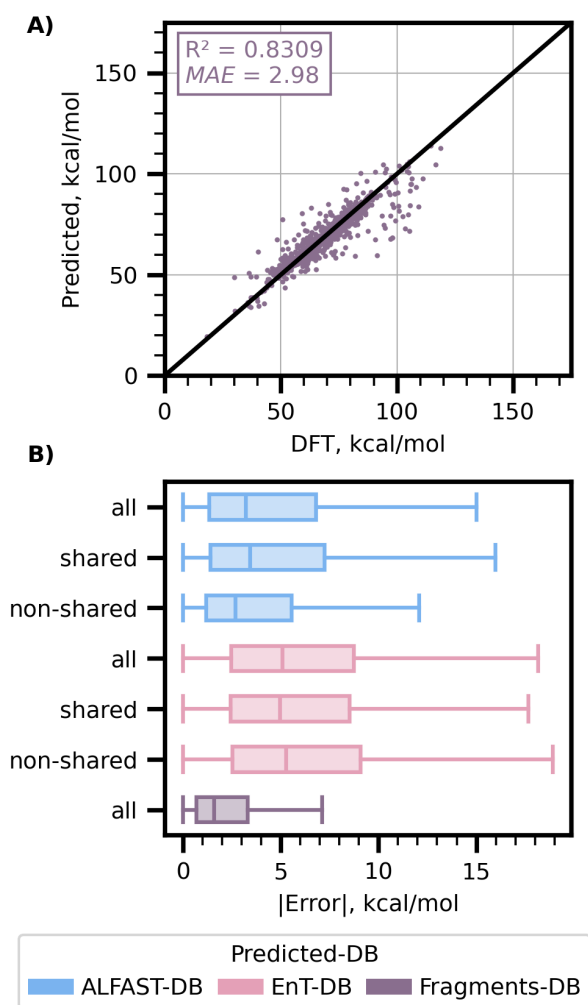


Figure 3.7: A) Predicted vs Computed S_0-T_1 adiabatic gaps of the model trained on the Fragments-DB. B) Boxplots with total (all) and marginal absolute error distributions of molecules based on whether they have a shared fragment in Fragments-DB (shared vs non-shared)

As fragment overlap between different datasets appears to directly correlate with the model's extrapolation ability, at a first-level approximation, it may be assumed that expanding the datasets with more fragments or even using the fragments as a surrogate prediction for the entire molecule may provide an interesting approach towards improving model performance.

To explore this angle, we created a dataset dubbed Fragments-DB which consists of DFT-computed adiabatic S_0 - T_1 gaps for each fragment present in ALFAST-DB or EnT-DB. We applied the fragmentation algorithm and included entries that appear at least twice in either of the two datasets, resulting in a total of 6,058 unique fragments. Following the same protocol for constructing ALFAST-DB, we obtained a total of 5135 S_0 - T_1 adiabatic gaps for structures in Fragments-DB.

Using this new Fragments-DB, we retrained the GNN model following the same approach as employed prior (**Figure 3.5**) to allow for direct comparison with the larger ALFAST dataset. Following this approach, the resulting GNN model performed slightly worse, yielding a test-set MAE of 2.98 kcal/mol and a coefficient of determination of 0.83 (**Figure 3.7A**).

The decrease in predictive accuracy compared to the ALFAST-DB trained GNN (**Figure 3.5aa**, 1.94 kcal/mol) is attributable to the composition of Fragments-DB. It is to note that due to the construction of Fragments-DB, only one instance of each fragment is present in the training set. Thus, a random splitting strategy does not have the same effect on model training with Fragments-DB as with ALFAST-DB. The

accuracy presented here should hence be evaluated against the 3.12 kcal/mol MAE of the out-of-sample splitting results presented by Glorius²⁶.

Intriguingly however, when the Fragments-DB-trained model was applied to predict the test set of ALFAST-DB, the R^2 and MAE were found to be 0.72 and 5.36 kcal/mol, worse than the Fragments' test-set MAE (2.98) or its cross-validation values (MAE of 3.07 and 3.26 for the aggregated validation and test set respectively, see SI). We then attempted to investigate the cause of this discrepancy by analyzing the errors in the context of fragment overlap (**Figure 3.7B**). Notably, the errors remain consistent between shared and non-shared fragments, for both ALFAST-DB and EnT-DB. There are several possible explanations for this behavior: a) the degradation in performance may be due to the excitation localization problem discussed previously, as the fragment based model is trained on individual chromophores only, and thus multi-chromophore molecules are out-of-distribution; b) the model may be failing to learn contributions from the whole molecule (*vide infra*, **Figure 3.8D**); c) the model fails at representing large molecules; d) it is known that certain functional groups more readily optimize to triplets than others (e.g. aryl groups); thus, it may be that the whole-molecule ALFAST model simply learns the identity of these groups more readily than the fragment model. We leave further investigation of these possibilities to future work.

To address the exciton localization problem while exploring the origin of deviations between these models, we employed a series of new training strategies that systematically consider the contributions of the fragments. We initially

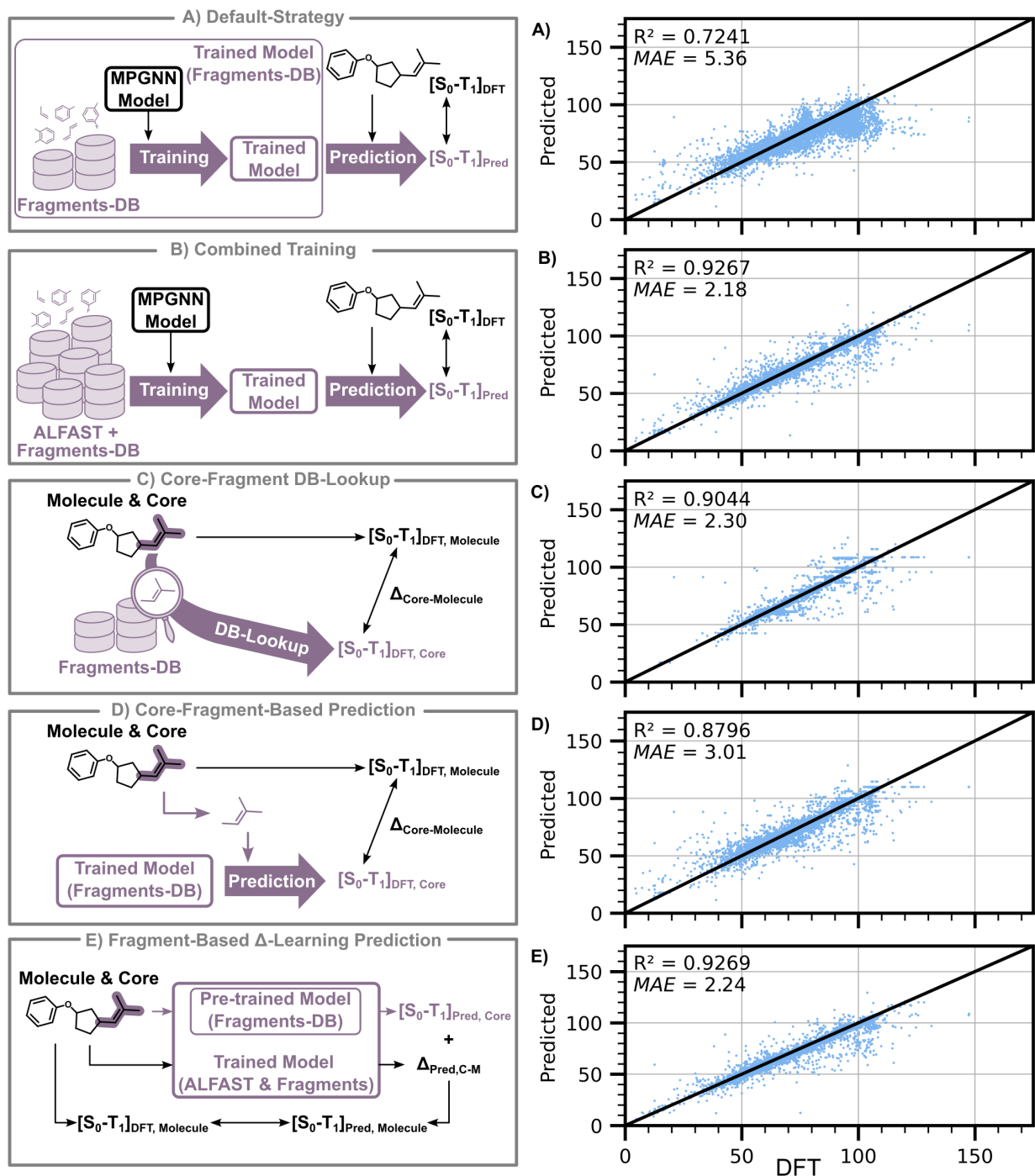


Figure 3.8: General illustration of the different "prediction" schemes discussed (A-E left) and their corresponding parity plots on the test set of ALFAST-DB (A-E right). All presented numbers are in kcal/mol except the R^2 coefficients which are dimensionless. In this figure "Core" is used to refer to the fragment where the exciton is localized. Reported MAE and R^2 are the metrics of the predictions of the ALFAST-DB test set.

investigated the simple strategy of dataset augmentation, where ALFAST-DB was naïvely augmented using Fragments-DB (**Figure 3.8B**), keeping only the entry from the

Fragments-DB when a molecule was present in both databases. When trained on the combined training and validation sets of each database, a test-set MAE (evaluated only on the test-set of ALFAST-DB) of 2.18 kcal/mol and an R^2 of 0.93 was obtained (cross validation results and performance on the combined test-set are provided in the SI). When compared with the model trained on whole molecules from the ALFAST-DB alone (MAE of 1.94, **Figure 3.5aa**), a slight decrease in performance is observed. The resulting 11% degradation is attributable to the differing chemical spaces represented by ALFAST-DB and Fragments-DB: the Fragments-DB is composed of smaller constituent chemical species and thus presents a different input space distribution than the ALFAST-DB. Nevertheless, this new model forms a more robust baseline for performance comparison.

We next evaluated the incurred error when the adiabatic S_0 - T_1 gap of the molecule is approximated as the adiabatic S_0 - T_1 gap of the main fragment. As such, we developed a “core-fragment DB lookup scheme”, where we approximate the value of the S_0 - T_1 gap for the whole molecule to be same as the DFT-computed value of the minimal size representation of the fragment where the excited triplet is localized (**Figure 3.8C**), providing an estimate of the error only due to the fragmentation approximation. Subsequently, the “core-fragment prediction scheme” employs a similar methodology; however, a Fragment-DB-trained GNN model is used for prediction instead of DFT (**Figure 3.8D**). Interestingly, for the GNN-based approach, the ALFAST-DB test set MAE is reduced from 5.36 to 3.01 kcal/mol by merely switching the target of the prediction from the whole molecule (**Figure**

3.8A) to the main fragment of the molecule (**Figure 3.8D**). The MAE further decreases to 2.30 kcal/mol when the DFT-calculated fragment adiabatic S_0 - T_1 gaps are used (**Figure 3.8C**). This observation is suggestive that, to a first order approximation, the calculation and prediction of just the fragment bearing the “correct” excitation can provide a reasonably accurate prediction.

Upon further analysis of the outliers in these approaches, it is found that the remaining part of the molecule can, in some instances, have a considerable effect. In particular, it was found that most such outliers consist of fragments that were originally part of ring systems, which is consistent with previous studies (for example, a (*Z*)-butene formed from the fragmentation of cyclopentene, for which errors can be as high as 9 kcal/mol)⁷. While this observation suggests that the final predictions could be improved by further refining the fragmentation scheme, an alternative and more robust approach would be the development of a model architecture that can learn the specific contributions of the neighboring environment not encapsulated by the fragment itself.

In light of these results, we attempted to utilize a Δ -learning scheme to reduce model error further. Here we keep the same core-fragment prediction scheme detailed above but adapt the architecture and the model training (detailed in sections S3.1.4 and S3.4 of Appendix B) to take as inputs both the molecule and the core and output two values: one associated to the adiabatic S_0 - T_1 gap for the provided core and a second value to account for the contribution of the rest of the molecule. From the SMILES or SMARTS of the core, a mask of the atoms and bonds in the provided core

of the molecule is created and used to iteratively update two distinct global states, one for the core (or main fragment) and another for the remaining atoms and bonds of the molecule. From each of these we obtain a prediction for the core and a correction, the addition of which leads to the final S_0-T_1 prediction. Historically, Δ -learning models have been used to learn error corrections, allowing low-level theory quantum mechanical methods to be matched with results from more accurate computational or experimental methodologies.⁴⁹ This methodology has proven successful in predicting diverse chemical properties, such as solvated ground-state redox potentials,⁵⁰ protein-ligand binding affinities,⁵¹ and dielectric constants.⁵² Here, we treat our core-fragment prediction model as a “base prediction” and predict a correction to it from the non-core atoms and bonds, thus considering atom contributions from the non-fragment constituent atoms. This adapted model follows a two-stage training process (fully detailed in Appendix B, Section S3), where the Fragments-DB is used during the first training stage to train the core-fragment prediction model weights. A second stage then follows, using the combined ALFAST and Fragments-DB to fine-tune the final predictions.

Pleasingly, upon implementation and training of this Δ -learning approach, a test set R^2 and MAE of 0.93 and 2.24 kcal/mol was obtained on the ALFAST-DB test set (**Figure 3.8E**), without a significant degradation in performance over the simplistic augmentation scheme ($R^2 = 0.93$, MAE = 2.18 kcal/mol; **Figure 3.8B**). More intriguingly, however, the results obtained for the last approach are comparable in accuracy to the performance of the GNN trained on and applied to predict exclusively

single-fragment molecules (**Figure 3.5cc** and **Figure 3.7**). To provide an additional baseline for validation of the predictive power of our developed model, we fine-tuned a large language model (LLM)-based predictor using a 102 million parameter RoBERTa⁵³-style model trained on 480 million SMILES strings the ZINC database.⁵⁴ As the LLM model is trained only to reconstruct SMILES strings, it is agnostic of molecular topology and atom connectivities and thus incorporates no fragment information directly. We find that the fine-tuned LLM model achieves an MAE of 6.17 kcal/mol, significantly worse than the task specialized Δ -learning approach, highlighting the improvements in accuracy gained by incorporating fragment and molecule graph information. Further details on LLM model training and evaluation can be found in Appendix B, Section S7 LLM Model Comparison.

As such, we can conclude that this approach not only removes the excitation localization problem but also can partially account for electronic effects contributions of non-fragment constituent atoms. While further investigations are needed to elucidate the precise quantitative role of the non-fragment constituents, this approach leads to a significant improvement in performance compared to the Ent-Decker and our original training approaches, providing a new baseline against which future GNN architectures for predicting excited state properties should be evaluated.

To demonstrate the applicability of the final Δ -learning approach, a test set of six experimentally known molecules that undergo intramolecular [2+2] cyclization

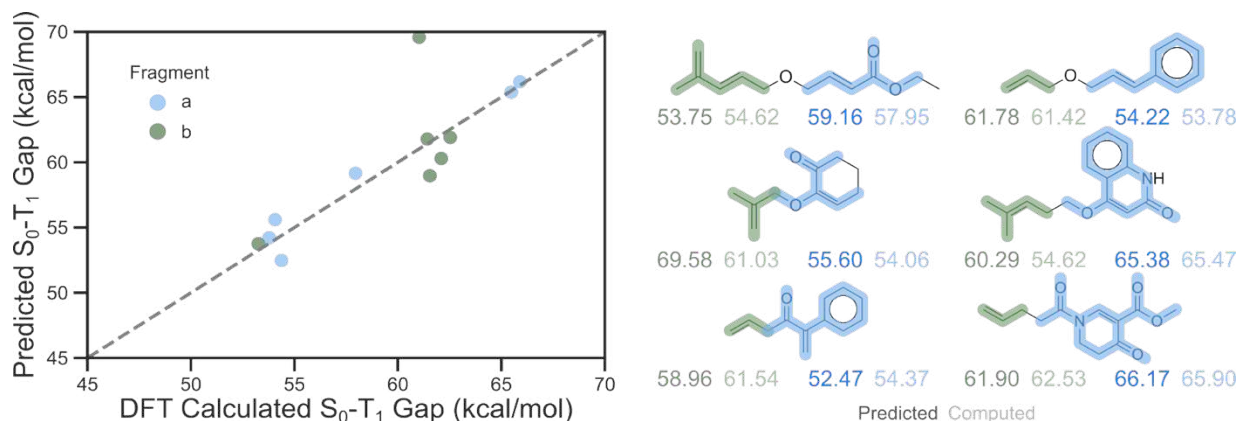


Figure 3.9: Model validation using a test set of six known molecules that can undergo [2+2] cycloaddition upon triplet energy sensitization, together with a parity plot of Δ -learning predictions against DFT computed values. Input fragments have been highlighted in green and blue for each molecule.

upon triplet energy sensitization was constructed (**Figure 3.9**).^{55–60} As these transformations inherently require two separated π -systems, two distinct excitations are anticipated, with spin densities localized on either one of the two fragments. Consistent with our expectations, we were able to computationally locate the triplet adiabatic structures for each chromophore at the M06-2X/def2-TZVP level of theory. Using the Δ -learning approach with the highlighted fragments as inputs, individual predictions can be provided for each chromophore. Comparison with the DFT-computed results yielded an MAE of 2.09 kcal/mol, which is in good agreement with the model accuracy of 2.24 kcal/mol reported for the external test set. Notably, only one large outlier was observed for the isobutene fragment, resulting in an error of 8.55 kcal/mol.

3.4: Conclusions

In this work, a fundamental problem in the context of photoredox and energy transfer catalysis is identified: the localization of the exciton. With this problem in mind, we constructed accurate high-throughput computational databases for the

development of machine learning models to predict the adiabatic S_0 - T_1 energy gap for small molecules. To do this, a chemically guided fragmentation algorithm was developed to address database analysis and curation. Next, its application on previously existing databases was showcased, highlighting the impact of exciton delocalization on both previously developed and newly built databases. By analyzing the impact of the fragmentation scheme on the EnT-DB, ALFAST-DB, and Verde-DB datasets, as well as the subsequent changes in performance of GNN models trained on these datasets, principles for designing and augmenting fragment-based datasets targeting the adiabatic S_0 - T_1 gap were elucidated. To improve database quality, the addition of diverse exciton localization moieties, “main fragments”, is preferable to the addition of duplicated versions of the same moieties, even when such moieties originate in diverse molecular species. Finally, based on these results, a promising strategy for developing ML models to predict the adiabatic S_0 - T_1 gap was developed, where the prediction of the adiabatic S_0 - T_1 gap for the key chemical moiety of the target molecule is decoupled from the rest of the molecule. For this strategy to be valid, the property of the molecule must be well approximated by the property of the minimal motif/fragment. We demonstrate here that such a model exhibits a similar error (MAE of 2.3 kcal/mol) to complex machine learning models trained on significantly larger databases (in the range of 1.93-2.26 kcal/mol).

We believe that the present work marks a clear direction for the development of newer fragment-based models and databases in the field, while also highlighting

the importance of incorporating domain knowledge into the prediction pipeline and data analysis. Although the final approach with a fragment-based Δ -learning model shows good predictive accuracy, we believe further improvements may be made through a better integration or design of the interaction between the core of the molecule and its substituents or the use of even more novel architectures. Future work may take advantage of newly developed GNN architectures that explicitly model this dependence, which may also improve model interpretability by allowing explicit fragment contributions to the adiabatic triplet energy to be directly quantified. Further, by incorporating newly released datasets of molecular properties and geometries spanning broad swathes of chemical space calculated using electronic structure methods (such as the OMol25⁶¹ or AIMNet2⁶² datasets), the accuracy and generalizability of subsequent models may be greatly improved. While dataset augmentation via application-specific quantum chemistry calculations or the use of sophisticated training strategies such as knowledge distillation may need to be employed, the potential for model improvement is substantial.

3.5: References

1. Strieth-Kalthoff, F., James, M. J., Teders, M., Pitzer, L. & Glorius, F. Energy transfer catalysis mediated by visible light: principles, applications, directions. *Chem. Soc. Rev.* **47**, 7190–7202 (2018).
2. Melchiorre, P. Introduction: Photochemical Catalytic Processes. *Chem. Rev.* **122**, 1483–1484 (2022).
3. Introduction: Photochemistry in Organic Synthesis. *Chem. Rev.* **116**, 9629–9630 (2016).

4. Cabanero, D. C. & Rovis, T. Low-energy photoredox catalysis. *Nat Rev Chem* **9**, 28–45 (2025).
5. Romero, N. A. & Nicewicz, D. A. Organic Photoredox Catalysis. *Chem. Rev.* **116**, 10075–10166 (2016).
6. Prier, C. K., Rankic, D. A. & MacMillan, D. W. C. Visible Light Photoredox Catalysis with Transition Metal Complexes: Applications in Organic Synthesis. *Chem. Rev.* **113**, 5322–5363 (2013).
7. Popescu, M. V. & Paton, R. S. Dynamic vertical triplet energies: Understanding and predicting triplet energy transfer. *Chem* **10**, 3428–3443 (2024).
8. Park, W., Komarov, K., Lee, S. & Choi, C. H. Mixed-Reference Spin-Flip Time-Dependent Density Functional Theory: Multireference Advantages with the Practicality of Linear Response Theory. *J. Phys. Chem. Lett.* **14**, 8896–8908 (2023).
9. Nelson, T. R. *et al.* Non-adiabatic Excited-State Molecular Dynamics: Theory and Applications for Modeling Photophysics in Extended Molecular Materials. *Chem. Rev.* **120**, 2215–2287 (2020).
10. Crespo-Otero, R. & Barbatti, M. Recent Advances and Perspectives on Nonadiabatic Mixed Quantum–Classical Dynamics. *Chem. Rev.* **118**, 7026–7068 (2018).
11. Kunze, L., Hansen, A., Grimme, S. & Mewes, J.-M. The Best of Both Worlds: Δ DFT Describes Multiresonance TADF Emitters with Wave-Function Accuracy at Density-Functional Cost. *J. Phys. Chem. Lett.* 1114–1125 (2025) doi:10.1021/acs.jpcllett.4c03192.
12. Guan, Y., Sowndarya, S. V. S., C. Gallegos, L., John, P. C. S. & S. Paton, R. Real-time prediction of ^1H and ^{13}C chemical shifts with DFT accuracy using a 3D graph neural network. *Chemical Science* **12**, 12012–12026 (2021).
13. Stubbs, C. D. *et al.* Predicting homopolymer and copolymer solubility through machine learning. *Digital Discovery* <https://doi.org/10.1039/D4DD00290C> (2024) doi:10.1039/D4DD00290C.
14. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
15. Kim, Y. *et al.* Physics-informed graph neural networks for predicting cetane number with systematic data quality analysis. *Proceedings of the Combustion Institute* **39**, 4969–4978 (2023).

16. Aldossary, A. *et al.* In Silico Chemical Experiments in the Age of AI: From Quantum Chemistry to Machine Learning and Back. *Advanced Materials* **36**, 2402369 (2024).
17. Sorkun, M. C., Khetan, A. & Er, S. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. *Sci Data* **6**, 143 (2019).
18. Kim, Y., Jung, H., Kumar, S., Paton, R. S. & Kim, S. Designing solvent systems using self-evolving solubility databases and graph neural networks. *Chem. Sci.* **15**, 923–939 (2024).
19. Gelžinytė, E., Öeren, M., Segall, M. D. & Csányi, G. Transferable machine learning interatomic potential for bond dissociation energy prediction of drug-like molecule. Preprint at <https://doi.org/10.26434/chemrxiv-2023-l85nf> (2023).
20. 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* **11**, 2328 (2020).
21. V, S. S. S., Kim, Y., Kim, S., John, P. C. S. & Paton, R. S. Expansion of bond dissociation prediction with machine learning to medicinally and environmentally relevant chemical space. *Digital Discovery* <https://doi.org/10.1039/D3DD00169E> (2023) doi:10.1039/D3DD00169E.
22. Casetti, N., Alfonso-Ramos, J. E., Coley, C. W. & Stuyver, T. Combining Molecular Quantum Mechanical Modeling and Machine Learning for Accelerated Reaction Screening and Discovery. *Chemistry – A European Journal* **29**, e202301957 (2023).
23. Lunger, J. R. *et al.* Towards atom-level understanding of metal oxide catalysts for the oxygen evolution reaction with machine learning. *npj Comput Mater* **10**, 1–11 (2024).
24. Ferri, P. *et al.* Approaching enzymatic catalysis with zeolites or how to select one reaction mechanism competing with others. *Nat Commun* **14**, 2878 (2023).
25. Abreha, B. G., Agarwal, S., Foster, I., Blaiszik, B. & Lopez, S. A. Virtual Excited State Reference for the Discovery of Electronic Materials Database: An Open-Access Resource for Ground and Excited State Properties of Organic Molecules. *J. Phys. Chem. Lett.* **10**, 6835–6841 (2019).
26. Schlosser, L., Rana, D., Pflüger, P., Katzenburg, F. & Glorius, F. EnTdecker – A Machine Learning-Based Platform for Guiding Substrate Discovery in Energy Transfer Catalysis. *J. Am. Chem. Soc.* <https://doi.org/10.1021/jacs.4c01352> (2024) doi:10.1021/jacs.4c01352.
27. Frisch, M. J. *et al.* Gaussian 16 Rev. C.01. (2016).

28. Ertl, P. An algorithm to identify functional groups in organic molecules. *Journal of Cheminformatics* **9**, 36 (2017).
29. Yadav, L. D. S. Ultraviolet (UV) and Visible Spectroscopy. in *Organic Spectroscopy* (ed. Yadav, L. D. S.) 7–51 (Springer Netherlands, Dordrecht, 2005). doi:10.1007/978-1-4020-2575-4_2.
30. Chai, J.-D. & Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Phys. Chem. Chem. Phys.* **10**, 6615–6620 (2008).
31. Scalmani, G. & Frisch, M. J. Continuous surface charge polarizable continuum models of solvation. I. General formalism. *The Journal of Chemical Physics* **132**, 114110 (2010).
32. Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics* **98**, 5648–5652 (1993).
33. Lee, C., Yang, W. & Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **37**, 785–789 (1988).
34. Vosko, S. H., Wilk, L. & Nusair, M. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Can. J. Phys.* **58**, 1200–1211 (1980).
35. Stephens, P. J., Devlin, F. J., Chabalowski, C. F. & Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* **98**, 11623–11627 (1994).
36. Neese, F. The ORCA program system. *WIREs Computational Molecular Science* **2**, 73–78 (2012).
37. Tingle, B. I. *et al.* ZINC-22—A Free Multi-Billion-Scale Database of Tangible Compounds for Ligand Discovery. *J. Chem. Inf. Model.* **63**, 1166–1176 (2023).
38. Miertuš, S., Scrocco, E. & Tomasi, J. Electrostatic interaction of a solute with a continuum. A direct utilization of AB initio molecular potentials for the prevision of solvent effects. *Chemical Physics* **55**, 117–129 (1981).
39. Clark, T., Chandrasekhar, J., Spitznagel, G. W. & Schleyer, P. V. R. Efficient diffuse function-augmented basis sets for anion calculations. III. The 3-21+G basis set for first-row elements, Li–F. *Journal of Computational Chemistry* **4**, 294–301 (1983).

40. Hariharan, P. C. & Pople, J. A. The influence of polarization functions on molecular orbital hydrogenation energies. *Theoret. Chim. Acta* **28**, 213–222 (1973).
41. 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 Account* **120**, 215–241 (2008).
42. Hehre, W. J., Ditchfield, R. & Pople, J. A. Self—Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian—Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules. *The Journal of Chemical Physics* **56**, 2257–2261 (1972).
43. Dunning, T. H., Jr. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *The Journal of Chemical Physics* **90**, 1007–1023 (1989).
44. Guo, Y. *et al.* Communication: An improved linear scaling perturbative triples correction for the domain based local pair-natural orbital based singles and doubles coupled cluster method [DLPNO-CCSD(T)]. *The Journal of Chemical Physics* **148**, 011101 (2018).
45. Saitow, M., Becker, U., Riplinger, C., Valeev, E. F. & Neese, F. A new near-linear scaling, efficient and accurate, open-shell domain-based local pair natural orbital coupled cluster singles and doubles theory. *J Chem Phys* **146**, 164105 (2017).
46. Riplinger, C., Pinski, P., Becker, U., Valeev, E. F. & Neese, F. Sparse maps—A systematic infrastructure for reduced-scaling electronic structure methods. II. Linear scaling domain based pair natural orbital coupled cluster theory. *The Journal of Chemical Physics* **144**, 024109 (2016).
47. Riplinger, C., Sandhoefer, B., Hansen, A. & Neese, F. Natural triple excitations in local coupled cluster calculations with pair natural orbitals. *J Chem Phys* **139**, 134101 (2013).
48. Riplinger, C. & Neese, F. An efficient and near linear scaling pair natural orbital based local coupled cluster method. *The Journal of Chemical Physics* **138**, 034106 (2013).
49. Ramakrishnan, R., Dral, P. O., Rupp, M. & Lilienfeld, O. A. von. Big Data meets Quantum Chemistry Approximations: The Δ -Machine Learning Approach. *J. Chem. Theory Comput.* **11**, 2087–2096 (2015).
50. Chen, X., Li, P., Hruska, E. & Liu, F. Δ -Machine learning for quantum chemistry prediction of solution-phase molecular properties at the ground and excited states. *Phys. Chem. Chem. Phys.* **25**, 13417–13428 (2023).

51. Yang, C. & Zhang, Y. Delta Machine Learning to Improve Scoring-Ranking-Screening Performances of Protein–Ligand Scoring Functions. *J. Chem. Inf. Model.* **62**, 2696–2712 (2022).
52. Grumet, M., von Scarpatetti, C., Bučko, T. & Egger, D. A. Delta Machine Learning for Predicting Dielectric Properties and Raman Spectra. *J. Phys. Chem. C* **128**, 6464–6470 (2024).
53. Liu, Y. et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. Preprint at <https://doi.org/10.48550/arXiv.1907.11692> (2019).
54. entropy/roberta_zinc_480m · Hugging Face. https://huggingface.co/entropy/roberta_zinc_480m.
55. Mayr, F., Brimioulle, R. & Bach, T. A Chiral Thiourea as a Template for Enantioselective Intramolecular [2 + 2] Photocycloaddition Reactions. *J. Org. Chem.* **81**, 6965–6971 (2016).
56. Hurtley, A. E., Lu, Z. & Yoon, T. P. [2+2] Cycloaddition of 1,3-Dienes by Visible Light Photocatalysis. *Angewandte Chemie International Edition* **53**, 8991–8994 (2014).
57. Skubi, K. L. et al. Enantioselective Excited-State Photoreactions Controlled by a Chiral Hydrogen-Bonding Iridium Sensitizer. *J. Am. Chem. Soc.* **139**, 17186–17192 (2017).
58. Herter, L., Koutsopetras, I., Turelli, L., Fessard, T. & Salomé, C. Preparation of new bicyclo[2.1.1]hexane compact modules: an opening towards novel sp³-rich chemical space. *Org. Biomol. Chem.* **20**, 9108–9111 (2022).
59. Graßl, R., Jandl, C. & Bach, T. Visible Light-Mediated Photochemical Reactions of 2-(2'-Alkenyloxy)cycloalk-2-enones. *J. Org. Chem.* **85**, 11426–11439 (2020).
60. Li, X., Jandl, C. & Bach, T. Visible-Light-Mediated Enantioselective Photoreactions of 3-Alkylquinolones with 4-O-Tethered Alkenes and Allenes. *Org. Lett.* **22**, 3618–3622 (2020).
61. Levine, D. S. et al. The Open Molecules 2025 (OMol25) Dataset, Evaluations, and Models. Preprint at <https://doi.org/10.48550/arXiv.2505.08762> (2025).
62. Training datasets for AIMNet2 machine-learned neural network potential. Carnegie Mellon University <https://doi.org/10.1184/R1/27629937.v2> (2024).

CHAPTER 4: Protein-Environment Interactions from Constituent Residues: Predicting Protein Solubility

4.1: Chapter Overview

In this chapter, we explore how the interaction of proteins with their environments arises as a function of their constituent amino acids. Proteins, large biomacromolecules, are consist of a backbone made up of amino acid residues joined in a linear chain through peptide bonds. Due to the dihedral degree of freedom in this peptide bond and the varied chemical compositions of amino acid side chains, this linear polypeptide adopts a folded conformation through both complex noncovalent (hydrogen bonding, pi-pi, cation-pi, anion-pi, hydrophilic/hydrophobic interactions) and covalent interactions (disulfide bridges), giving proteins complex, unique geometric folded structures that are deeply tied to both their function and interaction behavior. In this work, we make use of this relationship, using protein structure to directly inform model predictions of protein solubility.

Protein solubility is a somewhat nebulous concept; we utilize a binary definition here, with respect to a host expression system. I.e., a protein is “soluble” if it may be isolated in a folded state after expression in a given host organism; if expression fails, the protein is “insoluble”. This definition stands in contrast to conventionally used metrics for small molecule solubility, such as the octanol-water partition coefficient (logP) or the free energy change associated with solvation ($\Delta_{\text{solv}}G$); this discrepancy is due in part to the difficulty of measuring these traditional metrics for protein systems. The noncovalent interactions

responsible for a folded protein structure are fragile, and are easily disrupted by changes in solvent systems, making the measurement of $\log P$ infeasible, while the sensitivity of the calorimetric measurements needed to measure $\Delta_{\text{solv}}G$ also make this measurement infeasible at scale. Hence, we adopt the binary classification definition common in the literature.

Using an amino acid sequence-based binary solubility dataset from the literature, we construct a predicted structural dataset using the OpenFold protein structure prediction dataset. In parallel, we obtain experimental protein structures from the RCSB Protein Data Bank where available. We utilize a novel combined geometric-topological neural network prediction scheme to reduce the impact of structure prediction errors arising from the OpenFold prediction; a topological embedding network, TopoFormer, is trained using the experimental crystal structures and subsequently used to generate topological embeddings for the predicted protein structures. These embeddings are used in conjunction with protein structures as inputs to a rotationally equivariant model, TopoFormer, based on the SE(3) Transformer architecture. The TopoFormer model achieves state-of-the-art predictive accuracies on held out test sets. We apply the trained model to analyze the solubility of carbonic anhydrase mutants from the literature and show that predictions made by the TopoFormer model align with the results of conventional molecular dynamics simulations.

This collaborative research has been submitted for publication, with lead contributions from myself and additional contributions from Olivia Harman, Dr. Sumin Song, and Dr. Robert S. Paton, and Dr. Seonah Kim. I developed and created the databases, neural network architectures, performed model analysis, and wrote and edited the manuscript.

4.2: Introduction

The use of recombinant proteins in various industrial applications has grown rapidly in recent years, with strong newfound interest in using engineered proteins as peptide therapeutics,^{1,2} enzyme catalysts,^{3,4} and novel biomaterials.^{5,6} However, the industrial-scale production and utilization of proteins has long been hampered by poor protein solubility, defined here as the propensity of expressed proteins to form a solution-phase equilibria. This property is frequently conflated with aggregation propensity due to difficulties in experimentally characterizing protein-solvent equilibrium states: experimental labels of protein insolubility capture both an expressed protein's inability to form a solution phase equilibrium, as well as its tendency to form insoluble aggregates.⁷ Such insoluble aggregates consist of misfolded protein structures with amorphous secondary structure. These misfolded structures often contain solvent-exposed hydrophobic regions, which contribute to aggregation.^{8,9} This misfolding and loss of distinct structural identity, and thus catalytic function, preclude the use of expressed insoluble proteins in downstream applications. It is therefore necessary to rely on large-scale experimental campaigns to determine a protein's solubility, which wastes time, effort, and materials.

Hence, there is a need for accurate, explainable *in silico* models capable of predicting protein solubility, rather than relying on tedious experimental solubility determinations. Historically, *in-silico* methods were limited to long-timescale molecular dynamics (MD) simulations that leverage enhanced sampling methods.⁷ The recent development of machine learning (ML) techniques for predicting

molecular properties offers an attractive means to address this need for experimentation. However, many such models are data-hungry, requiring tens of thousands of initial training data points to generate reliable predictions;^{10–13} thus, careful dataset generation and curation become crucial to ensuring model generalizability and accuracy.^{14–16} While it is possible to use datasets of computationally predicted properties for model training,^{10,17,18} the results of *in-silico* property prediction calculations are subject to uncertainty, leading to compounded errors when used to train predictive models. Thus, experimentally measured properties have remained the gold standard as they mitigate the impacts of computational errors on model performance.

Due to these considerations, prior development of protein solubility datasets and prediction models has been limited, and few public datasets of protein solubility exist. One class of such datasets and models is oriented towards predicting solubility changes resulting from mutations in protein sequences. The Camsol,¹⁹ OptSolMut,²⁰ PonSol,^{21,22} and A3D²³ datasets, consisting of 56, 137, 433, and 129 datapoints, respectively, aim to quantify the positive and negative effects of mutations on protein solubility. An effort to compile such datasets, SoluProtMutDB, provides a convenient, curated central resource for dataset access and cross comparison²⁴.

On the other hand, datasets and models focusing on binary solubility predictions have also been developed. Due to the experimental difficulties in quantifying protein solubility, binary prediction datasets tend to be larger than those with continuous solubility labels. The PROSO II model employs a logistic function

trained on a binary-labelled dataset of over 82,000 protein sequences from various source organisms, achieving an accuracy of 64%.²⁵ Subsequent efforts using this dataset led to the development of the PaRSnIP²⁶ and DeepSol²⁷ models, which achieved accuracies of 74% and 77%, respectively. On the other hand, the eSOL dataset contains 3,173 binary solubility labels and corresponding sequences from expression experiments in *E. coli*, the most commonly used industrial protein expression system.²⁸ The ProteinSo²⁹ and SOLar³⁰ models were trained on the eSOL dataset (with the SOLart model also predicting solubilities in *S. cerevisiae* as a separate class of prediction targets), achieving 82% and 66% accuracy, respectively. However, to use protein structural information as model inputs, reference experimental structures must be obtained from the PDB or predicted using empirical codes. The PDBSOL³¹ binary solubility dataset consists of 64,598 protein structures, including a mix of expression experiments in *E. coli* and *S. cerevisiae*, and structural predictions from the ESMFold³² model for each protein. The ProtSolM multimodal model was then trained on both raw experimental protein structures and handcrafted features, achieving a predictive accuracy of 79% on the held-out test set.

With the widespread adoption of accurate, easy-to-run protein structure prediction models such as ESMFold³³ and Alphafold2,³⁴ predicted protein structures are readily available;³⁵ thus, accurate protein solubility models capable of using predicted structures as inputs must be developed. In contrast to the previously described works, only the SolART³⁰ and GATSol³⁶ models are designed to use predicted protein structures as inputs for binary protein solubility prediction tasks.

The SolART model shows a decrease in performance when evaluated against predicted structure inputs (66% to 51%). The GATSol model, on the other hand, uses precomputed distance maps derived from Alphafold-generated structures and node embeddings from large language models, achieving 79% accuracy. The DeepMutSol³⁷ model uses similar inputs (Alphafold-generated distance maps and protein sequence textual embeddings generated by a convolutional neural network) to predict solubility changes from amino acid point mutations, achieving a model accuracy of 65.4%. However, one major challenge in developing such models stems from the reliability of predicted protein structures. When used as model inputs, these poorly predicted regions introduce input noise in the form of geometric variability, reducing the accuracy of downstream model predictions. This geometric variability is due to inherent disorder in the protein structure or poor performance of the structure prediction model; it is often impossible to determine the source of this variability without a corresponding experimental crystal structure to compare with. Two recent advances in ML and computer science show promise in mitigating this problem: a) equivariant neural networks (ENNs) and b) topological data analysis. ENNs are a class of geometry-based neural networks (NNs) capable of maintaining the rotational equivariance of input geometric data throughout the NN's mathematical operations. This allows for raw coordinate data to be used as inputs to the NN model.³⁸ It has been shown that these NNs are more data efficient than their invariant counterparts,^{39,40} and achieve state-of-the-art predictive accuracies even when the prediction targets are rotationally invariant scalar quantities, such as molecular energies.^{41,42} The use of

ENN models thus enables the use of comparatively small training datasets while generating accurate predictions of scalar molecular properties.

However, using an ENN model architecture alone does not address the uncertainty associated with predicted protein structures, as geometric uncertainty from low-confidence structural predictions will be present in coordinate inputs to an equivariant model. Recent advances in topological data analysis provide a solution by utilizing global shape information to make inferences. Topological barcodes are a means of representing geometric data, such as a point cloud, to capture general structural properties across multiple dimensions and scales;⁴³ they have been used to describe physical phenomena as diverse as pore sizes in zeolites,⁴⁴ neuron maps in the brain,⁴⁵ and the flocking behavior of migratory birds.⁴⁶ The ability of these mathematical structures to capture key insights into the inherent structure of data has made them particularly useful for analyzing the impacts of shape on tasks such as NN information flow⁴⁷ and object segmentation.⁴⁸ Among the many implementations of such topological barcodes, the RipsNet NN architecture predicts the topological barcode from a given input geometry, improving the robustness of barcodes to input noise while maintaining an efficient computational cost.⁴⁹ This resistance to input perturbation allows topological barcodes to remain static for conformational variability in the 3D structure of a given protein; the topological barcode for regions of the protein with robust secondary structure remains invariant to structural change.

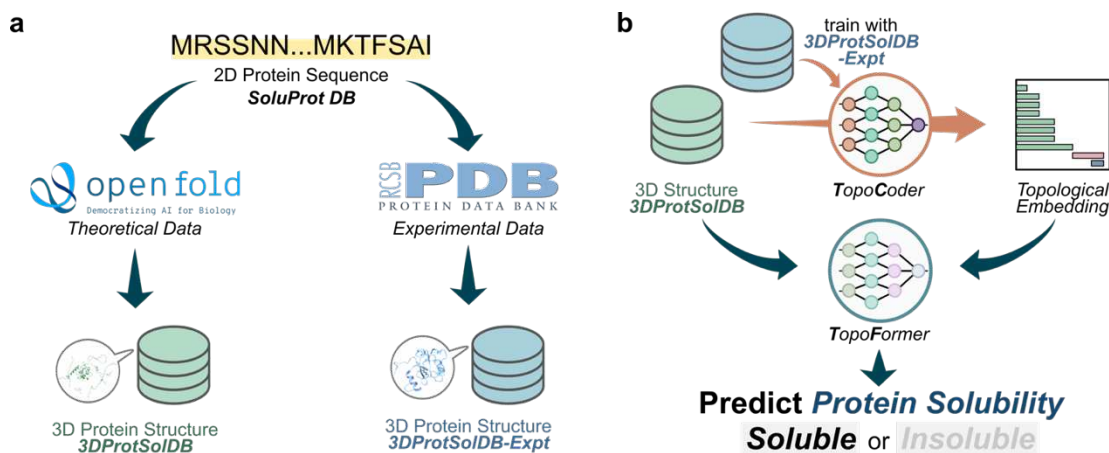


Figure 4.1: Overall scheme of the TopoFormer architecture and training procedure. a) Two newly developed 3D structural databases, **3DProtSolDB** and **3DProtSolDB-Expt**. The **3DProtSolDB** comprises OpenFold-predicted structures for 14,175 protein sequences from the SoluProt protein solubility database. A sequence search of the RSCB PDB for sequences in **3DProtSolDB** yielded 8,021 experimental structures, which were collected to form **3DProtSolDB-Expt**. b) Protein solubility prediction scheme via the **TopoCoder** and **TopoFormer** models. 3D Coordinates from **3DProtSolDB-Expt** were used to train a topological embedding model, **TopoCoder**. The trained **TopoCoder** model is used to generate embeddings for structures in the **3DProtSolDB** dataset. The 3D structures and predicted embeddings for structures in the **3DProtSolDB** are then used as inputs to an equivariant graph transformer architecture, **TopoFormer**.

These approaches motivated us to develop an accurate solubility prediction tool for protein systems using geometric and topological input data. We present a novel ML model to predict protein solubility from predicted 3D protein structures, using a topological barcode descriptor to address the inherent variability of these structures. Using 3D-predicted protein structures as model inputs introduces potential variability in their geometry. Thus, predicted topological representations of protein input structures are used as additional input embeddings to augment the geometric representations computed by an equivariant transformer model. In this work, 3D structural databases of predicted and experimental crystal structures, **3DProtSolDB** and **3DProtSolDB-Expt**, were created from amino sequences in the SoluProt binary solubility database (Figure 4.1a).⁵⁰ Then, an embedding model to generate topological barcodes for protein structures, which we term “**TopoCoder**”, was trained on experimental crystal structures from the **3DProtSolDB-Expt** database

and used to generate predicted topological barcode embeddings for structures from the **3DProtSolDB** dataset (**Figure 4.1b**). We show that the resultant barcode embeddings are more robust than those generated using traditional computational topology algorithms. A novel equivariant transformer architecture for proteins, which we term “**TopoFormer**”, is then trained on 3D structures from the **3DProtSolDB** dataset, and its state-of-the-art predictive accuracy is demonstrated on a held-out test set. The **TopoFormer** architecture uses a cross-attention mechanism to integrate topological barcode embeddings and geometric structure representations; the trained model learns which regions of the protein structure are most relevant to solubility prediction. Model predictions are evaluated using MD simulations, demonstrating that the trained **TopoFormer** model reproduces chemical insights regarding the impact of residue hydrophilicity and flexibility on overall protein solubility.

4.3: Results and Discussion

4.3.1: Generation and Validation of Protein Structural Databases

We used OpenFold⁵¹ to generate 3D structures for amino acid sequences in the full SoluProt dataset, which covers 11,436 training and 3,100 test proteins. The resulting **3DProtSolDB** contains 14,175 predicted 3D structures, labeled as soluble (7,099) and insoluble (7,076) (**Figure 4.2a**) after dropping 361 unreliable 3D structures (details in Methods, Protein Structural Database Generation). The OpenFold software package is a one-to-one, open-source reproduction of

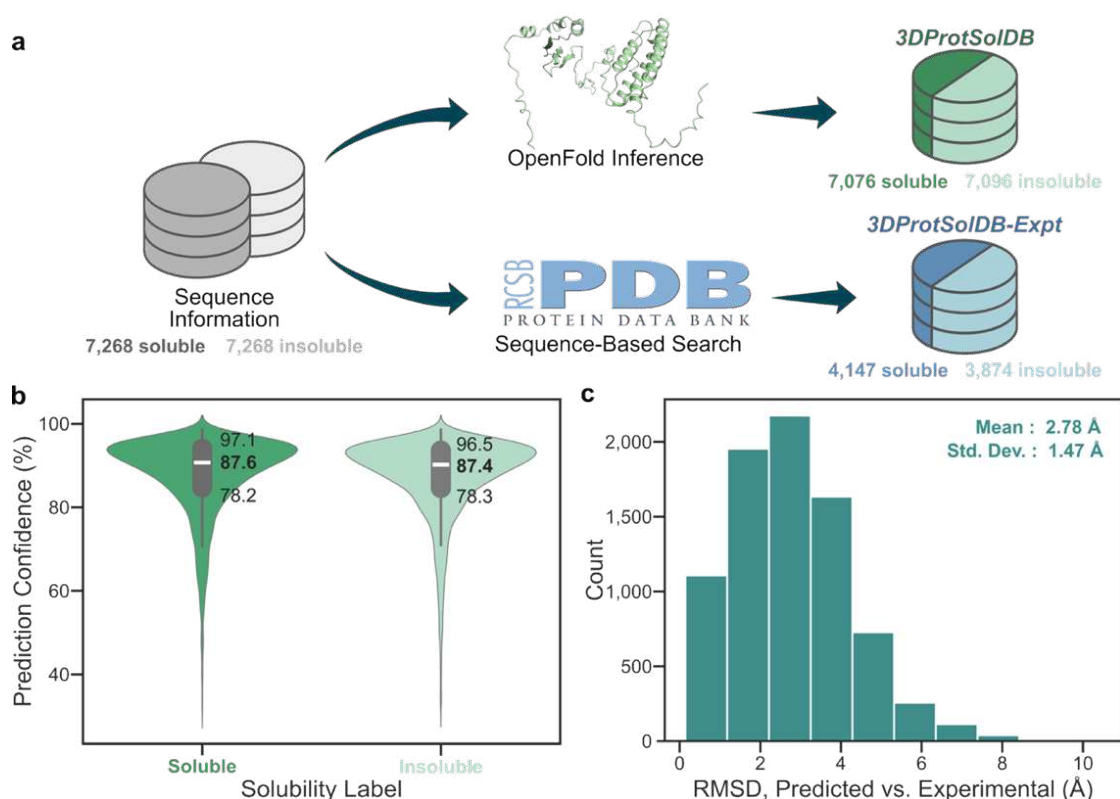


Figure 4.2: a) Generation of **3DProtSolDB** and **3DProtSolDB-Expt** structural databases from sequence information using OpenFold inference and RCSB PDB sequence similarity search, respectively. b) Distribution of OpenFold average prediction pLDDTs (predicted local distance difference test) for sequences labelled soluble and insoluble in **3DProtSolDB**. The interquartile ranges and mean (labelled bars) of average prediction pLDDTs are nearly identical for both soluble and insoluble predicted structures. c) Distribution of per-sequence average RMSD between predicted structures in the **3DProtSolDB** and corresponding experimental structures in the **3DProtSolDB-Expt**. RMSDs were calculated using the structure-based CEalign (Combinatorial Extension) algorithm.

AlphaFold2, offering increased accessibility and efficiency; workflows are provided to reproduce the predictive behavior of the original AlphaFold2 model. Since OpenFold is trained on experimental structures from the RCSB PDB,⁵² it is necessary to investigate whether OpenFold can generate reliable structural predictions for amino acid sequences labeled “insoluble” in the SoluProt dataset. Proteins that remain insoluble, uncrystallizable, and unsuitable for NMR cannot be reliably characterized experimentally. Thus, predicting their 3D structures is an “out-of-distribution” task for OpenFold. Since the OpenFold model is trained on experimental structures from the RCSB PDB, any uncharacterized insoluble structures would be

omitted from the OpenFold training set. Since these omitted structures may differ from those in the training set in sequence identity or structural motifs (as in *de novo-designed* proteins), structural predictions of these sequences are expected to contain regions with poorly predicted secondary structure, which often manifest as disordered regions.⁵³ Thus, we must first determine whether the amino acid motifs of “insoluble” labelled sequences are adequately represented in the OpenFold training set.

To do so, in parallel to the creation of the predicted **3DProtSolDB**, an experimental structural dataset, **3DProtSolDB-Expt**, was also constructed. By utilizing a BLAST sequence similarity search⁵⁴ of experimentally characterized protein structures deposited in the RCSB PDB, it is found that 8,021 of the 11,436 SoluProt training set sequences (71%) have experimental crystal structures deposited in the RCSB PDB, of which 4,147 are labeled soluble, while 3,874 are labeled insoluble (**Figure 4.2a**, details in Methods, Protein Structural Database Generation). Experimental structures for “insoluble” labelled proteins were found to be obtained from expression experiments in systems other than *E. coli*. Thus, ground-truth experimental crystal structures for “insoluble” labelled proteins were used during OpenFold training, and predicting structures for “insoluble” proteins is an “in-distribution” task (i.e., a task for which the trained ML model generates reliable predictions). The structures obtained via BLAST search were collected to form the **3DProtSolDB-Expt** database.

To further analyze the quality of the predicted structures in **3DProtSolDB** and **3DProtSolDB-Expt**, we first examine overall database statistics and then focus on specific example predictions. The OpenFold-reported pLDDT (predicted local distance difference test) is used as a metric of prediction confidence to identify disparities between soluble and insoluble training set sequences.⁵⁵ The average pLDDTs of the OpenFold predictions in **3DProtSolDB** were analyzed by binning predictions based on the interquartile range. It was found that their interquartile ranges span a narrow range of pLDDT values, with 75% interquartile, mean, and 25% interquartile of 97.1, 87.6, 78.2 units and 96.5, 87.9, and 78.3 units for soluble and insoluble proteins, respectively (**Figure 4.2b**). The 75% interquartile, mean, and 25% interquartile are thus nearly identical for both soluble and insoluble sequences. A Cohen's d test was then used to quantify differences between the average pLDDT for soluble and insoluble sequences: a value of 0.00 indicates identical distributions, values from 0.0 – 0.2 indicate small dissimilarity, values from 0.2 – 0.8 indicate medium to large dissimilarity, and values above 0.8 indicate high dissimilarity.⁵⁶ The Cohen's d test value between the average pLDDTs of the soluble and insoluble was found to be 0.04, indicating that the average per-sequence prediction confidence does not significantly differ between soluble and insoluble labelled sequences.

The RMSD between predicted and experimental protein structures is also compared for structures with corresponding entries in **3DProtSolDB-Expt** to evaluate the accuracy of the predicted structures. Corresponding structures were aligned using the CEAlign algorithm as implemented in BioPython with default

parameters (window size of 8, maximum gap size of 30).⁵⁷ The average per-structure RMSD between predicted and experimental structures was 2.78 Å, with a standard deviation of 1.47 Å, indicating good overall agreement (**Figure 4.2c**).

To further investigate any possible correlation between prediction confidence (pLDDT score) and geometric deviation between predicted and experimental structures (RMSD), we utilize quartile splits based on the average pLDDT value of **3DProtSolDB** predictions, as established above. 92% of **3DProtSolDB** structures are predicted with high overall confidence (above 80% pLDDT, which can be defined as “high confidence”), indicating that most of the **3DProtSolDB** secondary structures are reliable.⁵⁵ To further characterize the relationship between RMSD and pLDDT, an example structure from **3DProtSolDB** with a corresponding sequence match in **3DProtSolDB-Expt** was randomly selected from each split, and the per-residue RMSD was calculated between the two structures. When this per-residue RMSD is plotted against the per-residue pLDDT of the **3DProtSolDB** structure (Appendix C, **Figure S1**), the per-residue RMSD is lowest when the pLDDT is high, indicating a correlation between OpenFold prediction confidence and accurate structural prediction. Thus, protein structures with high overall pLDDT more closely resemble experimentally obtained counterparts.

4.3.2 Generation of Topological Barcode Embeddings (TopoCoder)

The relationship between prediction confidence (pLDDT) and the geometric deviation between predicted and experimental structures (RMSD) is further highlighted in structural superpositions of predicted structures for a single protein

(**Figure 4.3a**, left). Three different structural predictions with different starting seeds were obtained for a sequence chosen from the bottom (i.e., lowest confidence, 25%) average pLDDT quartile bin to obtain a representative sequence with low average prediction confidence. Within the structural prediction, two regions of high prediction confidence are observed (residues 1–401 and 461–527), separated by a region of low confidence (residues 402–460 in the grey block in **Figure 4.3a**). The high confidence (high pLDDT) sequence regions correspond to the regions of the predicted structure with well-defined secondary structure, while the low confidence regions lack well-defined structure (**Figure 4.3a**, right). Both high confidence regions show good structural alignment between the three predicted structures in isolation: residues 1–401, where all three structures are well superimposed in the initial generated structure show an average RMSD of 0.62 Å (**Table 4.1**); residues 461–527, where the three secondary structures are well superimposed when rotated in isolation (inset, Figure 5a), show an average RMSD of 0.24 Å. However, the presence of the low confidence region (residues 402–460) significantly degrades the overall alignment. No alignment between the three predicted structures is observed in the low-confidence region; however, adding this region increases the average RMSD across the whole protein to 22.6 Å (**Table 4.1**).

Residue Number	Structure 1-2 RMSD (Å)	Structure 1-3 RMSD (Å)	Average RMSD (Å)
0-401	0.74	0.49	0.62
402-460	No alignment found		
461-527	0.26	0.22	0.24
0-527	19.14	26.22	22.68

Table 4.1: Per-residue structural alignments for per-residue structural alignments for structural predictions of sample protein 100048.

It is not possible to determine whether such low-confidence regions are due to inherent disorder in the protein structure or to OpenFold's prediction error in an automated, high-throughput workflow. Even if manual comparison were used to analyze regions of disorder, not all structures in 3DProtSolDB have corresponding experimental structures, making direct validation of each predicted structure impossible. On the other hand, due to the limited data available for model training, discarding structures is undesirable. Therefore, we introduce an alternative to raw geometric coordinates as additional model inputs: the topological barcode. To mitigate the impact of inaccurate geometry predictions in the 3DProtSolDB, we employ tools from computational topology to characterize the overall *shape* (rather than the exact *geometry*) of protein structures. We use topological persistence barcodes to describe the overall structure of proteins and to complement the raw coordinates used as inputs to our equivariant predictive model. To associate the topology of experimental crystal structures with that of predicted protein structures, we develop a new topological embedding model, TopoCoder, to generate topological barcode embeddings for structures in the 3DProtSolDB-Expt. The trained TopoCoder model is then used to generate embeddings for structures in the 3DProtSolDB, which can serve as inputs to downstream models. Details of the TopoCoder architecture and model training can be found in the Materials and Methods, Topological Embedding Generation, and TopoCoder Model Development sections.

The neural network architecture of the TopoCoder model learns shape information from protein secondary structure and prevents perturbations in input

coordinates from affecting the predicted embedding vectors. A 3D protein structure is converted into a point cloud of alpha-carbon coordinates. This whole-protein point cloud serves as the model input; the TopoCoder model then predicts the topological barcode, which serves as an embedding vector. It is observed that even when a predicted whole protein structure with regions of indeterminate secondary structure is used as a model input, the output embeddings are similar to those generated for an experimental structure with well-defined secondary structure. The TopoCoder model thus learns to represent the regions of well-defined secondary structure while removing the influence of regions with poor secondary structure.

This “denoising” capability is illustrated by analyzing the differences in generated embeddings for different conformational coordinates of a single example system. A structure from the lowest prediction confidence quartile bin (using the split strategy employed in Results and Discussion, Generation and Validation of Protein Structural Databases) of the 3DProtSolDB with a corresponding experimental structure in the 3DProtSolDB-Expt was randomly chosen (**Figure 4.3b**). A structural alignment using the Matchmaker algorithm implemented in ChimeraX⁵⁸ finds 53 alignment pairs with the default settings, with a per-aligned-residue RMSD of 0.75 Å (**Figure 4.3b**). The average pLDDT of the predicted structure in 3DProtSolDB is 57.38, indicating low confidence in the prediction. Topological barcode embeddings were generated for each structure using the computed barcode vectors embedding and the trained **TopoCoder** model. For ease of comparison, we report deviations between

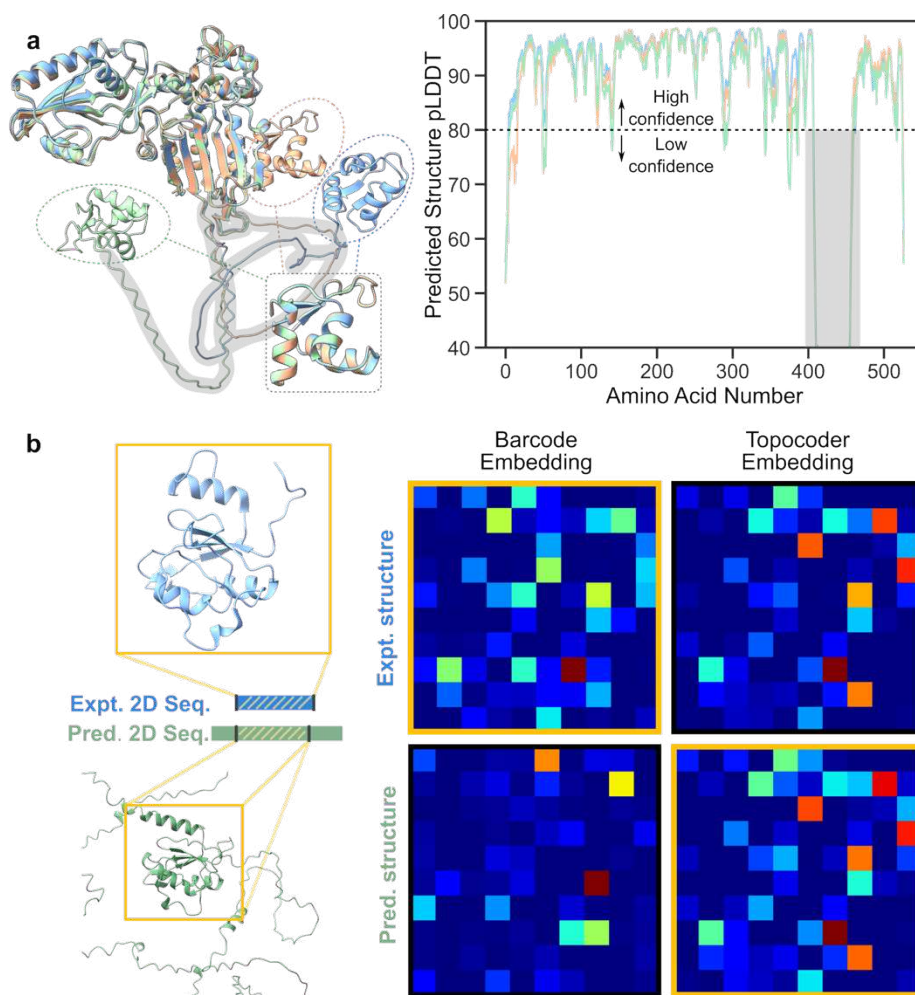


Figure 4.3: a) Superposition of 3 predicted structures for sample protein (ID: 100048) from **3DProtSolDB**. Per-residue pLDDT scores are plotted for 3 example predictions. b) Comparison of calculated topological barcodes vs. embeddings predicted by the trained **TopoCoder** model, for experimental and OpenFold predicted protein structures (i.e., structures from **3DProtSolDB** and **3DProtSolDB-Expt**, respectively, corresponding to the same amino acid sequence).

barcodes normalized to the maximum numerical value present in each barcode (**Figure 4.3b** heat map).

Comparing the calculated barcode vector embedding with **TopoCoder** generated embeddings for the experimental structure from **3DProtSolDB-Expt** reveals that the **TopoCoder** generated embeddings mirror prominent features of the barcode vector embedding. When subtracting the calculated barcode vector embedding from the **TopoCoder** generated embeddings for the experimental

structure, a normalized, summed mean absolute error of 1.7% is obtained. In contrast, when the calculated barcode vector method is used to generate embeddings for the predicted structure, several prominent features of the barcode are lost, as shown in Figure 5b. Yet, when the embeddings for the predicted structure are generated using the trained **TopoCoder** model, we observe that the overall pattern in the heat map is preserved. When compared individually to the calculated barcode vector embedding for the experimental structure, the normalized summed mean absolute error for the **TopoCoder** generated embeddings is 5.1%, approximately half that of the calculated barcode vector embedding, 10.3%. Thus, the **TopoCoder** embeddings are more robust to structural distortions in the poor-quality OpenFold predicted structure.

4.3.3 Equivariant Protein Solubility Prediction Model (TopoFormer) Training and Evaluation

Both the topological barcode embeddings generated by the TopoCoder model and the 3D structures in the 3DProtSolDB dataset were used as inputs to develop the equivariant protein solubility prediction model (TopoFormer, shown in Figure 3). Each protein structure was used to generate a graph representation in which individual amino acids are nodes. Features for each node were generated by concatenating relative solvent accessibility (RSA) computed by the DSSP software package^{59,60} using the accessible surface area values tabulated by Rost and Sander,⁶¹ , a one-hot encoding of the amino acid representing side chain identity, and a per-residue embedding generated using the pretrained ESM2 650M model³² to capture biological evolutionary context. Edges representing interactions between proximal amino acids

were then constructed by us using a 5 Å distance cutoff, as recommended in the literature,^{62–64} and implemented in the graphein software package.⁶⁵ In total, edges based on peptide bonds, hydrophobic interactions, hydrogen bond interactions, disulfide bonds, ionic interactions, aromatic interactions, and cation-pi interactions were constructed, and their types were one-hot encoded as bond features. The resultant featurized protein structure graph and global 3D topological barcode embeddings constitute the model inputs for each structure in the 3DProtSolDB.

We use two independent data splits to assess model performance: a random split and the original train-test split used in the SoluProt dataset. For the random split, the 3DProtSolDB was randomly partitioned into training, validation, and test sets at 80:10:10, yielding 11,340, 1,417, and 1,418 structures, respectively. In contrast, the train-test split from the SoluProt dataset is length balanced to reduce any correlation between amino acid sequence length and solubility. In addition, no train-test set pair has a global sequence identity above 25%, ensuring no data leakage is present during training. After removing failed structural predictions, 11,343 and 2,834 structures from the 3DProtSolDB correspond to sequences in the SoluProt training and test set, respectively. Initial model training and hyperparameter tuning were performed using only the random training and validation splits. After hyperparameter tuning, the training and validation splits were combined to train the final model architecture. The performance of the trained TopoFormer model was then evaluated on the held-out test set, which contains 726 insoluble and 692 soluble-labelled protein structures. The obtained hyperparameters were used to separately train the Topoformer model

on the 11,343 3DProtSolDB structures corresponding to the SoluProt training set; the model's performance was then assessed against a test set of 2834 structures from 3DProtSolDB corresponding to the SoluProt test set. The final model architecture consists of 5 TopoFormer layers (Appendix C, Section S1, Development of an Equivariant Protein Solubility Prediction Model), each with equivariant attention operations and one invariant cross-attention block. Equivariant layers use a maximum degree of 2, with 32 channels per degree, and 2 heads per attention layer. A layer norm was used after each TopoFormer layer. The AdamW optimizer with a learning rate of 0.0001 and weight decay of 0.01 using a mean absolute error (MAE) loss function was used for model training. The maximum training epochs were set to 130; early stopping was performed after 32 epochs based on validation set accuracy. Further training details can be found in Appendix C, Section S3 TopoFormer Network Details.

Confusion matrices of results of predictions on the test set structures for both splits are shown in **Figure 4.4** (left) using a classifier decision threshold of 0.5; the final model achieves a predictive accuracy of 81.4% (number of true positive predictions + number of true negative predictions/test set size) on the random split test set and 84.1% on the SoluProt test set. The trained TopoFormer model also achieves an F1 score (calculated as $2 \times (\text{true positive predictions}) / (2 \times (\text{true positive} + \text{false positive} + \text{false negative}))$) of 81.5 and 86.2 on the random split and SoluProt split test sets, respectively. The F1 score is the harmonic mean of precision and recall, where precision is the number of true positive predictions divided by the total number

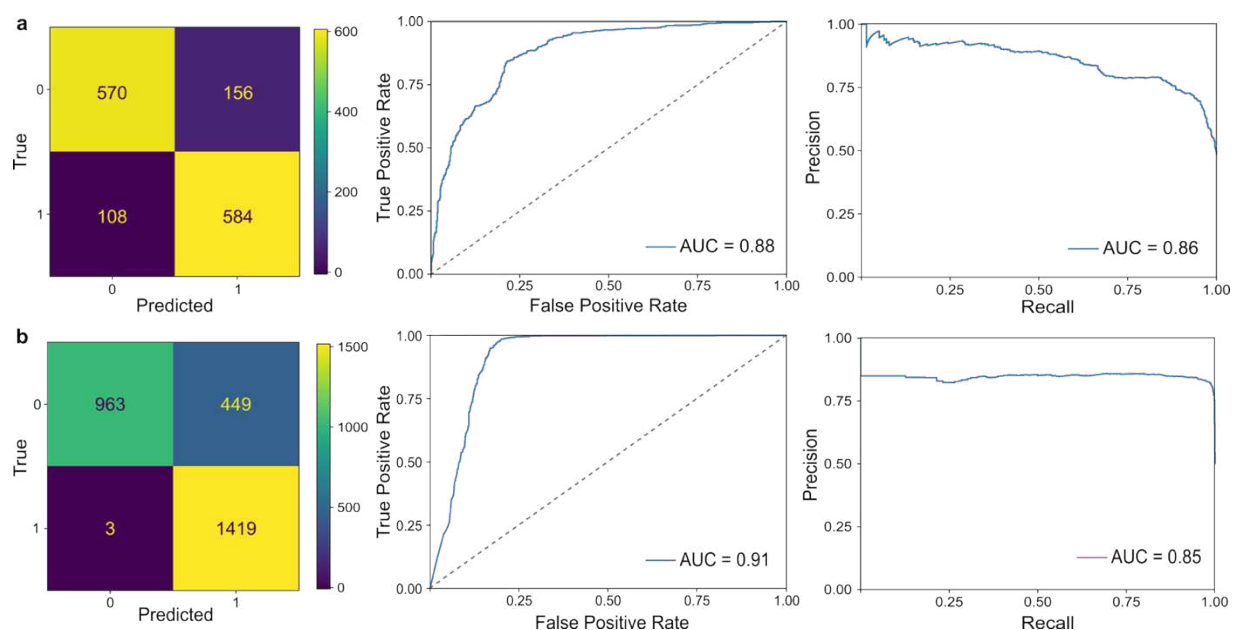


Figure 4.4: Confusion matrix (left), receiver operating characteristic curve (middle), and precision-recall curve (right) for Topoformer models trained using a (a) random split of the 3DProtSolDB and (b) original train-test splits from the SoluProt dataset. The random split partitions the 3DProtSolDB in an 80:10:10 ratio, resulting in split sizes of 11,340, 1,417, and 1,418 structures. The Topoformer model was initially trained and hyperparameter tuning performed using the train and validation sets from the random split; these splits were subsequently combined and used to train the final model with optimized hyperparameters. The same hyperparameters were then used to train an independent model on the original SoluProt training set; the performance of this model was evaluated against the SoluProt test set.

of positive predictions (including those not identified correctly), and recall is the number of true positive predictions divided by the total number of predictions labelled positive. The Matthews correlation coefficient (ie, Pearson correlation between two binary variables) was also calculated to be 0.63 and 0.71 for the trained TopoFormer model on the random split and the SoluProt split held-out test sets, respectively. Receiver operating characteristic (ROC) curves and precision-recall curves are shown in **Figure 4.4** (middle, right). The area under the ROC curves (AUC) is 0.88 and 0.91 for the random split and the SoluProt split held-out test sets, respectively, and the AUC for the PR curves is 0.86 and 0.85, respectively. These results show that the trained **TopoFormer** model accurately classifies soluble and insoluble proteins. It is also noted that for both splits, the number of true positive

labels is almost identical to the number of true negatives. However, the model predicts false positives and false negatives at approximately a 3:2 ratio in the random split, while overpredicting positive labels by almost 150:1 on the SoluProt test set, indicating a bias towards positive labels despite the balanced nature of the solubility class labels in the training data. Thus, the model overpredicts positive solubility, which is advantageous for its use as a predictive high-throughput screen preceding experimental verification: the overprediction ensures that potentially soluble proteins are not misclassified and removed from experimental screening. By tuning the classifier decision threshold, this bias toward positive predictions may be adjusted depending on user preference.

To determine the influence of each input component on **TopoFormer** performance, an ablation study was conducted. An alternative invariant model architecture was used to study the impact of the invariant components of the **TopoFormer** node inputs (topological barcode embeddings and ESM2 650M embeddings), and the equivariant architecture was benchmarked by removing these invariant inputs. The same train/test split as that used to train the final **TopoFormer** model was used for all ablation studies, and an invariant graph attention (GAT) network architecture was adopted as the surrogate model. The surrogate GAT model consists of 5 Graph Attention Network blocks with a 0.1 dropout rate and 4 attention heads each, as implemented in the DGL library, followed by a softmax output layer. ReLU activation functions were used in each layer, and the model was trained with the Adam optimizer at a learning rate of 0.0005 for 200 epochs, with a batch size of

32. The surrogate GAT model was first trained solely on topological embedding inputs. The trained model shows poor classifier performance, achieving an accuracy of only 57.7% on the held-out test set. To evaluate the importance of the ESM embeddings, we repeat this training procedure using only the ESM node embeddings as inputs. The trained model also performs similarly poorly, achieving an accuracy of 52.6%. We conclude that neither ESM embeddings nor topological barcodes alone are sufficient to adequately describe protein solubility; geometric inputs are also necessary to achieve good model accuracy. The equivariant SE(3) Transformer trained without barcode information achieves 58.2% accuracy; with barcode information, the accuracy increases to 74.3%. Using ESM-2 embeddings without barcode embeddings in the equivariant architecture results in a prediction accuracy of 69.5%, indicating that equivariant inputs, barcode embeddings, and ESM-2 embeddings are necessary to achieve the final predictive accuracy. The impact of using **TopoCoder**-predicted embeddings was also assessed; embeddings calculated from the data yield a model accuracy of 76.1%, demonstrating that predicted embeddings are necessary to maximize predictive accuracy. Lastly, model results were compared with random forest baseline models to validate the use of a deep learning architecture. Independent random forest models were trained using only ESM features, only non-ESM node features, and the full set of node features used in the Topoformer model. The model depth and number of leaves hyperparameters were optimized independently for each model using a grid search. We find that the full-node features random forest model performs best, achieving an accuracy of

55.4% on the SoluProt test set, indicating that a deep neural network architecture is necessary for predictive performance. Further details on random forest model benchmarking are provided in Appendix C, Section S3, Topoformer Network Details.

The results of model training were further analyzed by studying the solvent exposure of polar and nonpolar residues for correctly and incorrectly predicted test set samples. Numerous studies have shown that protein aggregation behavior in solution is correlated with interactions between solvent-exposed hydrophobic regions in protein structures.^{23,66} Thus, we analyze the combined fraction of solvent-exposed polar amino acids and buried non-polar amino acids as a proxy for this aggregation behavior. The computed RSAs for each residue in each test set structure are tabulated; a residue is considered solvent exposed if its RSA is above 0.20.⁶⁷ The percentage of combined solvent-exposed polar and buried non-polar amino acids is computed for both soluble and insoluble-labelled proteins from the test set. We refer to this metric as relative exposed hydrophilicity (REH). The correctly predicted soluble and insoluble proteins show an REH of 34.4 ± 5.7 and 31.3 ± 4.5 , respectively, while the incorrectly predicted soluble and insoluble proteins show an REH of 32.5 ± 3.7 and 32.7 ± 4.3 (Appendix C, **Figure S3**). The difference between these distributions was quantified using a Cohen's d test. It is found that, for correct predictions, the Cohen's d test value for soluble and insoluble proteins is 0.59, indicating a medium-to-large, significant difference in REHs.⁵⁶ The correctly predicted soluble structures thus have a greater number of solvent-exposed polar amino acids and buried non-polar amino acids, as shown by the larger mean REH

value (34.4 vs. 31.3). Performing a similar analysis yields Cohen's d values of 0.30 and 0.34 for comparisons between the correctly and incorrectly predicted proteins for both soluble and insoluble class labels, respectively, indicating a smaller but still significant difference. Since only the RSA for each residue is provided as a model input, the model lacks *a priori* knowledge of residue hydrophobicity/hydrophilicity, which is necessary for including residues when calculating the REH score, since only exposed hydrophilic and buried hydrophobic residues are included. Thus, we infer that this information is learned during model training, indicating that the model has captured this physical insight.

4.3.4 Structural Mapping of Solubility in Carbonic Anhydrase Mutants

The trained **TopoFormer** model was used to study the solubility of carbonic anhydrase mutants. Carbonic anhydrases are a class of enzymes that reversibly catalyze the hydration of carbon dioxide to a bicarbonate ion and a proton; as such, these enzymes have attracted recent attention for use in carbon sequestration and carbon dioxide upcycling.^{68,69} Recent work has focused on improving the industrial utility of carbonic anhydrase enzymes by selecting mutants tolerant of harsh catalytic conditions from the literature. However, candidate sequences with promising robustness must then be evaluated for high solubility to find industrially viable enzyme catalysts that can be produced at scale. This was performed through a top-down *in silico* screening strategy, where the Protein-Sol²⁹ and SOLpro⁷⁰ sequence-based solubility prediction models were used in tandem to make *a priori* solubility predictions for identified candidates. Candidates with high predicted solubilities

from both models are then selected for *in vitro* protein expression studies. Due to the predictive uncertainty of existing *in silico* methods (58% for Protein-Sol and 74% for SOLpro), two separate models must be employed to achieve high confidence during the initial screening step. Additionally, since neither model reports per-residue importance, information from the solubility predictions cannot be used in subsequent enzyme design steps, necessitating costly trial-and-error to optimize solubility during downstream expression.

In contrast to this methodology, we utilize our trained **TopoFormer** model to enable a bottom-up design strategy using a single predictive model for solubility prediction and the identification of target amino acid residues for subsequent mutation. Due to the attention mechanism in the **TopoFormer** model, we can identify amino acid residues that it considers most important for solubility prediction. We perform model inference and downstream analysis for two example carbonic anhydrase variants: CA1-TaCA (*Thermovibrio ammonificans* CA), experimentally found to be insoluble; and CA7-PabCA (*Pyrococcus abyssi* CA), experimentally found to be soluble. His-tags used for affinity chromatography were removed from the protein sequences, and OpenFold was used to generate predicted structures. The trained **TopoCoder** models were then used to generate 1D, 2D, and 3D topological embeddings for both predicted structures. Per-residue embeddings were simultaneously generated using the pre-trained ESM2 650M model.

We find that the **TopoFormer** model accurately predicts both CA1-TaCA and CA7-PabCA as soluble and insoluble, respectively. Attention weights were extracted

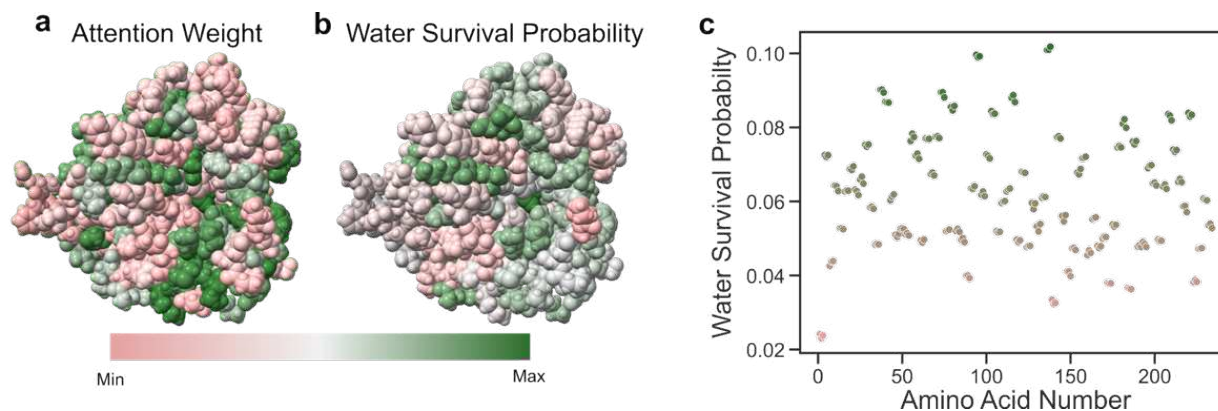


Figure 4.5: a) TopoFormer attention weights, b) MD water survival probability map for CA1-TaCA, and c) MD Normalized water survival probability per residue of CA1-TaCA.

from the penultimate cross-attention module in the **TopoFormer** architecture; the OpenFold predicted structure of CA1-TaCA, colored by **TopoFormer** attention weights, is shown in **Figure 4.5a**.

Simultaneously, MD simulations were performed to calculate the water survival probabilities of CA1-TaCA (experimentally insoluble in *E. coli*). Results are shown projected on the CA1-TaCA predicted structure in **Figure 4.5a** to compare with the solubility attention weights from **TopoFormer**; numerical results are shown in **Figure 4.5c**. When we compare the water survival probabilities with the OpenFold predicted structure of CA1-TaCA, we find they correlate well with the attention weights of the trained **TopoFormer** model. **Figure 4.5a** shows a clear qualitative relationship between the water survival probability and the **TopoFormer** attention weights (**Figure 4.5b**). While a linear fit of water survival probability and the extracted attention weights displays an R^2 of 0.15, we find that a random forest surrogate model with 10 trees, as implemented in scikit-learn,⁷¹ achieves an R^2 of 0.76 (further details in Appendix C, Section S6, Carbonic Anhydrase Prediction Results), indicating the presence of a correlation between **TopoFormer** attention weights and the water

survival probability. The high R^2 value suggests that the **TopoFormer** model accurately captures the physical interactions between water molecules and the protein target, despite having no prior information about the interactions.

Model predictions can thus be used to guide inverse design for downstream protein design by analyzing model attention weights alongside water-survival probabilities from MD studies. The attention weights indicate the patches most used by the model to predict solubility – thus, mutation and design efforts can target patches with high attention weights to most directly influence solubility. However, an integrated MD-attention weight analysis may provide the most comprehensive strategy for design efforts: by focusing on amino acid patches with high attention weights but low water survival probability times, specific protein regions may be prioritized for mutation campaigns. These amino acid patches indicate regions of the protein that the model predicts as key to protein solubility. It is believed that targeted mutations at these sites may allow large scale, precise tuning of protein solubility behavior. For example, the mutation of a hydrophilic glutamic acid residue to a hydrophobic valine in hemoglobin creates a hydrophobic patch, leading to protein aggregation and manifesting as sickle cell anemia.⁷² We hypothesize that such sites with outsize impacts on protein solubility may be identified by model attention weights, while being overlooked in traditional hydrophobicity patch analysis. Further rigorous analysis of this relationship is necessary, and exploration of its potential use in downstream predictive models will be conducted in future studies.

4.4: Conclusions

In this work, we present **TopoFormer**, a unified geometric-topological equivariant neural network model for predicting protein solubilities from OpenFold-predicted protein structures. The 3DProtSolDB dataset, comprising 14,175 OpenFold predictions and experimentally derived solubility labels, was compiled using amino acid sequences from the SoluProt dataset. In parallel, the **3DProtSolDB-Expt** dataset was compiled through an amino acid sequence similarity search of the RCSB PDB, using the same sequences as those used to create **3DProtSolDB**. A topological embedding model, **TopoCoder**, was trained on **3DProtSolDB-Expt** to generate predicted embeddings for structures in the **3DProtSolDB**. The generated embeddings capture secondary-structure shape information for protein sequences, even in regions of predicted disordered or low-confidence structure.

The predicted embeddings are used as inputs to the **TopoFormer** model, along with structural information. The **TopoFormer** model uses topological embeddings to perform cross-attention over geometric representations, achieving state-of-the-art predictive accuracy of 81.4% on a random split and 84.1% on the redundancy reduced SoluProt test split.. Upon analyzing model predictions using solvent accessibility scores, it is found that the model accurately identifies regions of protein structures that exhibit both positive and negative solvent interactions, highlighting the **TopoFormer** model's ability to learn protein-solvent interaction behavior. Finally, we validate the **TopoFormer** model by analyzing the solubility of carbonic anhydrase

mutants from the literature. Not only does the **TopoFormer** model accurately predict the solubility of the mutants, but an analysis of its attention weights also reveals that they are correlated with water survival probabilities from conventional MD simulations.

In summary, our results demonstrate the accuracy and interpretability of a novel **TopoFormer** model for predicting protein solubility using predicted OpenFold protein structures as model inputs. The trained **TopoFormer** and **TopoCoder** models are available at the following GitHub repository for use in solubility prediction tasks:

https://github.com/SabariKumar/combining_geom_topo

4.5: References

1. Albericio, F. & Kruger, H. G. Therapeutic Peptides. *Future Medicinal Chemistry* **4**, 1527–1531 (2012).
2. de Oliveira, E. C. L., Santana, K., Josino, L., Lima e Lima, A. H. & de Souza de Sales Júnior, C. Predicting cell-penetrating peptides using machine learning algorithms and navigating in their chemical space. *Sci Rep* **11**, 7628 (2021).
3. Raj, S. B., Ramaswamy, S. & Plapp, B. V. Yeast Alcohol Dehydrogenase Structure and Catalysis. *Biochemistry* **53**, 5791–5803 (2014).
4. Mallinson, S. J. B., Dessaux, D., Barbe, S. & Bomble, Y. J. Computer-Aided Engineering of a Non-Phosphorylating Glyceraldehyde-3-Phosphate Dehydrogenase to Enable Cell-Free Biocatalysis. *ACS Catal.* **13**, 11781–11797 (2023).
5. Arthur Jones, A. & D. Snow, C. Porous protein crystals: synthesis and applications. *Chemical Communications* **60**, 5790–5803 (2024).
6. Miserez, A., Yu, J. & Mohammadi, P. Protein-Based Biological Materials: Molecular Design and Artificial Production. *Chem. Rev.* **123**, 2049–2111 (2023).
7. Vihinen, M. Solubility of proteins. *ADMET and DMPK* **8**, 391–399 (2020).
8. Arakawa, T. & Timasheff, S. N. [3]Theory of protein solubility. in *Methods in Enzymology* vol. 114 49–77 (Academic Press, 1985).

9. Trimpin, S. & Brizzard, B. Analysis of Insoluble Proteins. *BioTechniques* **46**, 409–419 (2009).
10. Kim, Y., Jung, H., Kumar, S., Paton, R. S. & Kim, S. Designing solvent systems using self-evolving solubility databases and graph neural networks. *Chem. Sci.* **15**, 923–939 (2024).
11. V, S. S. S., Kim, Y., Kim, S., John, P. C. S. & Paton, R. S. Expansion of bond dissociation prediction with machine learning to medicinally and environmentally relevant chemical space. *Digital Discovery* <https://doi.org/10.1039/D3DD00169E> (2023) doi:10.1039/D3DD00169E.
12. Kim, Y. *et al.* Designing High-Performance Fuels through Graph Neural Networks for Predicting Cetane Number of Multicomponent Surrogate Mixtures. <https://www.sae.org/publications/technical-papers/content/2023-32-0052/> (2023).
13. Coley, C. Defining and Exploring Chemical Spaces. *Trends in Chemistry* **3**, (2020).
14. Axelrod, S. & Gómez-Bombarelli, R. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. *Sci Data* **9**, 185 (2022).
15. Corso, G. *et al.* Deep Confident Steps to New Pockets: Strategies for Docking Generalization. (2024).
16. Vita, J. A. & Schwalbe-Koda, D. Data efficiency and extrapolation trends in neural network interatomic potentials. (2023) doi:10.48550/arXiv.2302.05823.
17. John, P. C. S., 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. *Nature Communications* **11**, 2328 (2020).
18. Jung, H. *et al.* Enhancing predictive models for solubility in multicomponent solvent systems using semi-supervised graph neural networks. *Digital Discovery* **4**, 1492–1504 (2025).
19. Sormanni, P., Aprile, F. A. & Vendruscolo, M. The CamSol Method of Rational Design of Protein Mutants with Enhanced Solubility. *Journal of Molecular Biology* **427**, 478–490 (2015).
20. Tian, Y., Deutsch, C. & Krishnamoorthy, B. Scoring function to predict solubility mutagenesis. *Algorithms for Molecular Biology* **5**, 33 (2010).
21. Yang, Y., Zeng, L. & Vihinen, M. PON-Sol2: Prediction of Effects of Variants on Protein Solubility. *International Journal of Molecular Sciences* **22**, 8027 (2021).

22. Yang, Y., Niroula, A., Shen, B. & Vihinen, M. PON-Sol: prediction of effects of amino acid substitutions on protein solubility. *Bioinformatics* **32**, 2032–2034 (2016).
23. Zambrano, R. *et al.* AGGRESKAN3D (A3D): server for prediction of aggregation properties of protein structures. *Nucleic Acids Research* **43**, W306–W313 (2015).
24. Velecký, J. *et al.* SoluProtMutDB: A manually curated database of protein solubility changes upon mutations. *Computational and Structural Biotechnology Journal* **20**, 6339–6347 (2022).
25. Smialowski, P., Doose, G., Torkler, P., Kaufmann, S. & Frishman, D. PROSO II – a new method for protein solubility prediction. *The FEBS Journal* **279**, 2192–2200 (2012).
26. Rawi, R. *et al.* PaRSnIP: sequence-based protein solubility prediction using gradient boosting machine. *Bioinformatics* **34**, 1092–1098 (2018).
27. Khurana, S. *et al.* DeepSol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics* **34**, 2605–2613 (2018).
28. Niwa, T. *et al.* Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins. *Proceedings of the National Academy of Sciences* **106**, 4201–4206 (2009).
29. Hebditch, M., Carballo-Amador, M. A., Charonis, S., Curtis, R. & Warwicker, J. Protein-Sol: a web tool for predicting protein solubility from sequence. *Bioinformatics* **33**, 3098–3100 (2017).
30. Hou, Q., Kwasigroch, J. M., Rومان, M. & Pucci, F. SOLart: a structure-based method to predict protein solubility and aggregation. *Bioinformatics* **36**, 1445–1452 (2020).
31. Tan, Y., Zheng, J., Hong, L. & Zhou, B. ProtSolM: Protein Solubility Prediction with Multi-modal Features. Preprint at <https://doi.org/10.48550/arXiv.2406.19744> (2024).
32. Lin, Z. *et al.* Evolutionary-scale prediction of atomic level protein structure with a language model. (2022) doi:10.1101/2022.07.20.500902.
33. Lin, Z. *et al.* Language models of protein sequences at the scale of evolution enable accurate structure prediction. 2022.07.20.500902 (2022) doi:10.1101/2022.07.20.500902.
34. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
35. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* **50**, D439–D444 (2022).

36. Li, B. & Ming, D. GATSol, an enhanced predictor of protein solubility through the synergy of 3D structure graph and large language modeling. *BMC Bioinformatics* **25**, 204 (2024).
37. Wang, J. *et al.* Predicting the effects of mutations on protein solubility using graph convolution network and protein language model representation. *Journal of Computational Chemistry* **45**, 436–445 (2024).
38. Thomas, N. *et al.* Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds. (2018) doi:10.48550/arXiv.1802.08219.
39. Batzner, S. *et al.* E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat Commun* **13**, 2453 (2022).
40. Chen, Z. *et al.* Direct Prediction of Phonon Density of States With Euclidean Neural Networks. *Advanced Science* **8**, 2004214 (2021).
41. Köhler, J., Klein, L. & Noé, F. Equivariant Flows: Exact Likelihood Generative Learning for Symmetric Densities. (2020) doi:10.48550/arXiv.2006.02425.
42. Satorras, V. G., Hoogeboom, E. & Welling, M. E(n) Equivariant Graph Neural Networks. Preprint at <https://doi.org/10.48550/arXiv.2102.09844> (2022).
43. Ghrist, R. Barcodes: The persistent topology of data. *Bull. Amer. Math. Soc.* **45**, 61–75 (2008).
44. Lee, Y. *et al.* High-Throughput Screening Approach for Nanoporous Materials Genome Using Topological Data Analysis: Application to Zeolites. *J. Chem. Theory Comput.* **14**, 4427–4437 (2018).
45. Saggat, M., Shine, J. M., Liégeois, R., Dosenbach, N. U. F. & Fair, D. Precision dynamical mapping using topological data analysis reveals a hub-like transition state at rest. *Nat Commun* **13**, 4791 (2022).
46. Xian, L., Adams, H., Topaz, C. M. & Ziegelmeier, L. Capturing Dynamics of Time-Varying Data via Topology. Preprint at <https://doi.org/10.48550/arXiv.2010.05780> (2021).
47. Ballester, R., Casacuberta, C. & Escalera, S. Topological Data Analysis for Neural Network Analysis: A Comprehensive Survey. Preprint at <https://doi.org/10.48550/arXiv.2312.05840> (2024).
48. Glatt, R. & Liu, S. Topological Data Analysis Guided Segment Anything Model Prompt Optimization for Zero-Shot Segmentation in Biological Imaging. Preprint at <https://doi.org/10.48550/arXiv.2306.17400> (2023).

49. de Surrél, T. *et al.* RipsNet: a general architecture for fast and robust estimation of the persistent homology of point clouds. Preprint at <http://arxiv.org/abs/2202.01725> (2022).
50. Hon, J. *et al.* SoluProt: prediction of soluble protein expression in *Escherichia coli*. *Bioinformatics* **37**, 23–28 (2021).
51. Ahdritz, G. *et al.* OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. (2022) doi:10.1101/2022.11.20.517210.
52. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Research* **28**, 235–242 (2000).
53. Chang, L. & Perez, A. AlphaFold2 knows some protein folding principles. 2024.08.25.609581 Preprint at <https://doi.org/10.1101/2024.08.25.609581> (2024).
54. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
55. Guo, H.-B. *et al.* AlphaFold2 models indicate that protein sequence determines both structure and dynamics. *Sci Rep* **12**, 10696 (2022).
56. Lakens, D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Front. Psychol.* **4**, (2013).
57. Shindyalov, I. N. & Bourne, P. E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering* **11**, 739–747 (1998).
58. Meng, E. C. *et al.* UCSF ChimeraX: Tools for structure building and analysis. *Protein Science* **32**, e4792 (2023).
59. Joosten, R. P. *et al.* A series of PDB related databases for everyday needs. *Nucleic Acids Research* **39**, D411–D419 (2011).
60. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
61. Rost, B. & Sander, C. Conservation and prediction of solvent accessibility in protein families. *Proteins: Structure, Function, and Bioinformatics* **20**, 216–226 (1994).
62. Batatia, I., Kovács, D. P., Simm, G. N. C., Ortner, C. & Csányi, G. MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields. (2023) doi:10.48550/arXiv.2206.07697.

63. Mao, Z. *et al.* Molecule Graph Networks with Many-Body Equivariant Interactions. *J. Chem. Theory Comput.* **21**, 7954–7966 (2025).
64. Kumar, R., Romano, J. D. & Ritchie, M. D. CASTER-DTA: equivariant graph neural networks for predicting drug–target affinity. *Brief Bioinform* **26**, bbaf554 (2025).
65. Jamasb, A. *et al.* Graphein - a Python Library for Geometric Deep Learning and Network Analysis on Biomolecular Structures and Interaction Networks. *Advances in Neural Information Processing Systems* **35**, 27153–27167 (2022).
66. Fink, A. L. Protein aggregation: folding aggregates, inclusion bodies and amyloid. *Folding and Design* **3**, R9–R23 (1998).
67. Chen, H. & Zhou, H.-X. Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. *Nucleic Acids Research* **33**, 3193–3199 (2005).
68. Shen, J. & Salmon, S. Biocatalytic Membranes for Carbon Capture and Utilization. *Membranes* **13**, 367 (2023).
69. Di Fiore, A., Alterio, V., Monti, S. M., De Simone, G. & D’Ambrosio, K. Thermostable Carbonic Anhydrases in Biotechnological Applications. *International Journal of Molecular Sciences* **16**, 15456–15480 (2015).
70. Magnan, C. N., Randall, A. & Baldi, P. SOLpro: accurate sequence-based prediction of protein solubility. *Bioinformatics* **25**, 2200–2207 (2009).
71. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
72. Suhail, M. Biophysical chemistry behind sickle cell anemia and the mechanism of voxelotor action. *Sci Rep* **14**, 1861 (2024).

CHAPTER 5: Solvent-Solute Interactions from Atom and Bond Structure: Predicting Small Molecule Solubility

5.1: Chapter Overview

In this chapter, we explore how the interactions of solute molecules with solvents can be parameterized from individual molecular structures. Solubility is a key physical property, impacting chemical experimentation and discovery at almost every scale: solvent selection during reaction design can have large impacts on reaction rate, yield, and selectivity.

However, being a bulk phase property, solubility is challenging to measure using both experimental and computational means. Experimental datasets tend to be limited in solvent scope, and there exists a trade off between experimental accuracy and dataset size: larger datasets tend to have inconsistent or inaccurate measurements. On the other hand, the errors of *in-silico* techniques for computing solubility are poorly matched to experimental techniques. Thus, computational datasets and experimental datasets cannot be naively combined to build integrated models; rather, more sophisticated data augmentation techniques must be used to ensure dataset fidelity.

In this work, we introduce a dataset augmentation scheme based on semi-supervised distillation to reconcile errors between experimental and computational datasets, allowing data from both dataset types to be used in an integrated iterative training loop. A teacher model is trained on an initial restricted dataset consisting of experimental solubility measurements. This model is then used to predict solubility labels for

computationally generated data; molecules associated with low-error predictions are then added to the training dataset and used for subsequent rounds of model training. In this way, an integrated computational-experimental dataset is built, greatly expanding the model's chemical scope. The final student model achieves state-of-the-art accuracies on a broad range of chemical compounds. We show that predictions from our model can be used to correlate solubility with rates of chemical reaction and show another application in predicting the partition coefficients of lignin and drug-like molecules.

This collaborative research was published in the journal *Chemical Science*, with lead contributions from Dr. Yeonjoon Kim, and additional contributions from Hojin Jung, myself, Dr. Robert S. Paton, and Dr. Seonah Kim. I developed the self-distillation scheme and helped with neural network model development. I also helped with model analysis, and writing and editing of the manuscript.

5.2: Introduction

Solubility is a key molecular property that controls reactivity, catalytic activity, separation ability, and other molecular properties. In chemical synthesis, solvent selection influences the solubilities of reactants, intermediates, and products and impacts catalytic activity and product selectivity. It is crucial in designing catalytic reactions pertinent to pharmaceutical synthesis in solutions, such as functionalization through C-H activation.¹⁻⁶ In this regard, linear relationships have been elucidated between solvent properties (permittivity, polarity, etc.) and reaction rates for various organic reactions in different

solvents.^{2,7,8} Such linear solvation energy relationships (LSERs) inform solvent selection, leading to the maximal yield of target products.

In the pharmaceutical industry, solubilities in water and organic solvents are essential properties to consider during the screening and synthesis of drug candidates.^{9,10} Candidates having sufficient water solubility should be identified for high bioavailability in oral administration.¹¹ Water solubility is also relevant to the toxicity of drugs and pesticides on human health and the environment.^{12–14} Solubilities in organic solvents matter as well, especially for assessing the *in vivo* efficacy and safety of intravenous drugs dissolved in non-toxic organic solvents.^{10,15,16} Specifically, solubilities of drug-like molecules in chloroform and diethyl ether have been investigated for the simplified modeling of the polar environment around proteins, and membranes.^{17,18} In addition, solubility plays a critical role in emerging research areas such as sustainable chemistry and renewable energy. For instance, solvent selection is conducted in biomass upgrading to biofuels and renewable polymers to maximize catalytic activity.^{19–22} The optimal water-organic solvent systems enhance not only the conversion to target products but also their extraction from separation processes.^{19,20} Meanwhile, developing organic redox flow batteries is another promising research area for renewable energy storage, and it is important to design electrolytes highly soluble in water or organic solvents for high charge densities.^{23–25}

To date, solubilities of various solutes in water and organic solvents have been measured experimentally, and databases of experimental solubilities have been released. The available databases include AqSolDB,²⁶ Open Notebook Scientific Challenge,²⁷ Minnesota Solvation Database,^{28–30} FreeSolv,³¹ CompSol,³² and the

Journal of Chemical Information and Modeling's solubility challenge database.^{11,33,34}

Many computational methods have also been developed to predict solubilities *in silico*. Such methods include quantum mechanics (QM) or density functional theory (DFT) with implicit solvation models (e.g., Solvation Model based on Density – SMD),³⁵ molecular dynamics (MD) simulations, or QM-based thermodynamic equilibrium methods, e.g., the Conductor-like Screening Model (COSMO).^{36–38} A modified version of the COSMO model, COSMO for real solvents (COSMO-RS), incorporates statistical thermodynamics to treat bulk properties such as solubility more accurately.³⁹

For rapid and accurate solubility predictions, various predictive models have been developed by analyzing quantitative structure-property relationships (QSPRs)^{34,40–45} or via machine learning (ML) techniques.^{34,43,46–57} Current advanced ML models used graph neural networks (GNNs) combined with interaction layers^{48,54,58}, recurrent neural networks with attention layers,⁴⁶ and Natural Language Processing (NLP)-based transformers.⁵⁹ These models have achieved accuracies close to experimental uncertainties. Furthermore, the development of ML models has been expanded to predicting solubility limits at different temperatures,⁵³ solvation enthalpy, LSERs, and solute parameters,⁵² and generative models for designing molecules having optimal aqueous solubility.⁵³

The databases and models described above deal with either molar solubility ($\log S$) or Gibbs solvation free energy (ΔG_{solv}). $\log S$ is commonly used for solubility, but henceforth, this paper focuses on ΔG_{solv} , since ΔG_{solv} as well as $\log S$ quantifies solubility:

$$\Delta G_{solv} = -RT \ln K_{eq} = -RT \ln \frac{c_{solute}}{c'_{solute}}, \quad 5.1$$

where c_{solute} and c'_{solute} are the solute's concentration in the solvent and initial phase (before the solvation), respectively. ΔG_{solv} can also represent the solubility regardless of the solute's initial phase (solid, liquid, or gas):⁴⁰

$$\begin{aligned} \Delta G_{solv} &= \Delta G_{solid \rightarrow gas} + \Delta G_{gas \rightarrow solution} \\ \Delta G_{solv} &= \Delta G_{solid \rightarrow liquid} + \Delta G_{liquid \rightarrow solution} \end{aligned} \quad 5.2$$

Moreover, $\log S$ and ΔG_{solv} are related with each other:²⁹

$$\begin{aligned} \Delta G_{solv} &= -2.303RT \log(SV_m) \\ \Delta G_{solv} &= -2.303RT \log(S) + \log\left(\frac{P_{vap}}{P^0}\right), \end{aligned} \quad 5.3$$

where V_m and P_{vap} are molar volume and vapor pressure of a solute in the solid and liquid phase, respectively. While $\log S$ directly quantifies the amount of solutes, the significance of ΔG_{solv} lies in its ability to represent the thermodynamic stability of a solute in a solvent.

Further improvements are needed for accurate predictions of ΔG_{solv} for broad solvents and solutes. There are currently around 10,000 experimentally determined ΔG_{solv} in the most extensive database (e.g., **CombiSolv-Exp** in literature⁵⁴), but more data points (around >100,000) are desirable for reliable GNNs, due to their data-hungry nature.^{54,59} Moreover, there is a tradeoff between the size and reliability of

available solubility data. For example, a strictly curated dataset of reliable aqueous solubilities contains only hundreds of compounds.⁶⁰ Due to the scarcity of reliable experimental data, there have been attempts to pre-train against huge computational databases, followed by re-training against the experimental data.^{54,61} Such transfer learning can lose the information gained from the computational database after the model is re-trained against the small experimental database. In addition, QM solubilities systematically deviate from experimental values. The absolute errors of ΔG_{solv} for neutral solutes and solvents are 0.0-3.0 kcal/mol or higher for COSMO-RS, SMD, and other implicit solvation models.^{35,62-68} Therefore, developing a database seamlessly integrating theoretical and experimental values is a challenging but essential step.

Discrepancies between theoretical and experimental solubilities should be reconciled to build an integrated database. In other words, computational solubilities should have a fidelity as high as experimental ones despite the dependence of computational methods' accuracies on molecule size, constituent elements, functional groups, solvent properties, etc. Therefore, it is not feasible to merely combine experimental and computational databases. Each database has a different source and magnitude of errors and uncertainties,^{29,35,69,70} which would deteriorate the accuracy of predictive models. To overcome this limitation, a recent study added Gaussian-distributed random errors to the experimental datasets at different levels according to their uncertainties, leading to more accurate predictions than the models without Gaussian error distributions.⁷¹ Data augmentation and self-training

techniques have also been employed, such as semi-supervised distillation (SSD)^{72–78} for the reliable integration of databases from heterogeneous sources. These techniques have been successfully applied to various ML predictive models for image classification,^{73,74} natural language processing,⁷⁵ reaction classification,⁷⁶ and protein structures.⁷⁷

In this contribution, our own implementation of SSD was developed for our GNNs for solubilities. For reliable data integration, we refined the ΔG_{solv} values from COSMO-RS and SMD-DFT through the SSD, instead of directly adding values from these methods. Our approach dramatically amplified the database while correcting the discrepancies of different data sources. The resulting augmented database covers a much broader chemical space (11,213/1,445/932,509 solutes/solvents/data points) than that covered by experimental measurements (2,275/1,443/11,637 solutes/solvents/data points). Adopting the augmented database is powerful in enhancing GNNs' predictive accuracy, demonstrating the effectiveness of our approach. Moreover, we successfully applied our model to practical examples of solvent system design in reaction kinetics and separation. These would lead to more reliable evaluation of databases and predictive models than only one method.

5.3: Results and Discussion

5.3.1: Semi-supervised distillation scheme for developing solubility prediction models

We propose the following SSD scheme for solubility (**Figure 5.1**); first, the 'Teacher' ML model is trained against an experimental database of ΔG_{solv} (Step I).

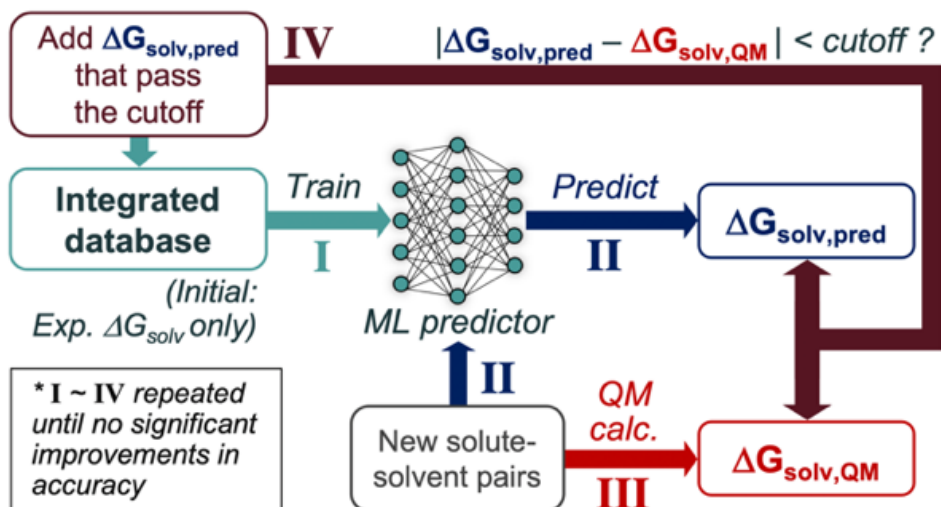


Figure 5.1: The semi-supervised distillation (SSD) scheme for the predictive models of solubility.

Second, ΔG_{solv} is predicted using the trained Teacher for the new solute-solvent pairs whose experimental ΔG_{solv} is unknown, resulting in $\Delta G_{\text{solv,pred}}$ (Step II). Meanwhile, QM-calculated ΔG_{solv} values are also obtained ($\Delta G_{\text{solv,QM}}$) for these new solute-solvent pairs (Step III). $\Delta G_{\text{solv,QM}}$ is referenced to determine which $\Delta G_{\text{solv,pred}}$ are used for the data augmentation. If the absolute difference between these two ($|\Delta G_{\text{solv,pred}} - \Delta G_{\text{solv,QM}}|$) is below a certain cutoff, the corresponding $\Delta G_{\text{solv,pred}}$ are added to the database (Step IV). This cutoff is to avoid the introduction of unreliable prediction data points from the Teacher. Notably, $\Delta G_{\text{solv,pred}}$ is added instead of $\Delta G_{\text{solv,QM}}$, enabling the data augmentation based on the inductive bias the ML model gained from the starting experimental database. Next, the ‘Student’ model is trained using the database combining the initial database and that from Teacher’s predictions. This procedure is repeated to add reliable data points to the integrated database gradually.

The SSD process described in **Figure 5.1** was applied to the GNNs for solubilities, constituting a one-of-a-kind approach in solubility predictions that

consolidate deep learning and heterogeneous data sources from experiments and QM calculations. The following sections describe the databases, QM methods, and architecture of the GNNs in detail.

Evaluation of quantum-mechanical methods and solubility databases

To accomplish data augmentation, first, we evaluated the QM methods used to provide reference $\Delta G_{\text{solv,QM}}$ during the SSD. 11,637 experimental ΔG_{solv} were collected and curated, resulting in **Exp-DB** (Figure 5.2A). Most data points in **Exp-DB** overlap with those in **CombiSolv-Exp**,⁵⁴ but our **Exp-DB** has additional 1,419 data points with the identical solute and solvent.

COSMO-RS and SMD-M06-2X/def2-TZVP were then benchmarked against **Exp-DB**. To assess COSMO-RS, we adopted **CombiSolv-QM**, the most extensive ΔG_{solv} database consisting of one million data points.⁵⁴ **QM-DB** (220,332 data points) was built to evaluate SMD-M06-2X/def2-TZVP which was elected among many SMD-DFT methods since it provided reliable results from calculating solubility-related

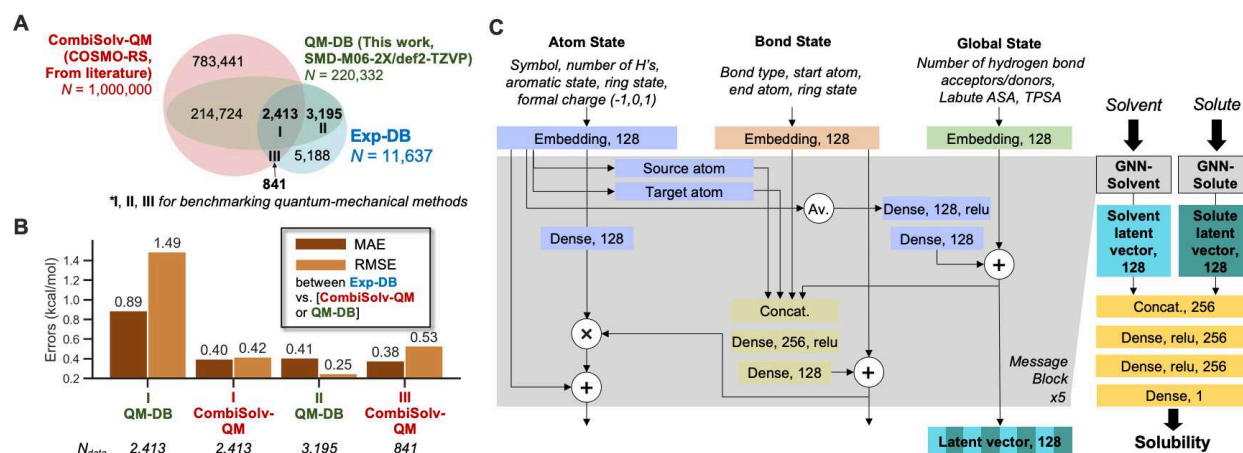


Figure 5.2: (A) The three databases for evaluating theoretical methods against experimental solubilities. (B) Accuracy comparisons of CombiSolv-QM and QM-DB for the data points overlapping with Exp-DB. (C) Architecture of the graph neural network for solubility. ASA: Accessible surface area, TPSA: Topological polar surface area.

properties (redox potentials of 174 organic molecules in water and acetonitrile).²⁵ 2,413 ΔG_{solv} values are available for all three databases (Region I), and 3,195 overlapped data points between **Exp-DB** and **QM-DB** (Region II). 841 solubilities in **Exp-DB** are available only in **CombiSolv-QM** (Region III) due to the unavailability of some solvents in SMD calculations.

Accuracies of the two theoretical methods were analyzed for I-III. In Region I, COSMO-RS is more accurate than SMD-M06-2X. Nonetheless, SMD-M06-2X shows notably high accuracy for the 3,195 data points in Region II, with an MAE and RMSE of 0.41 and 0.25 kcal/mol, respectively. Meanwhile, COSMO-RS showed a decent accuracy in Region III. These results show that each theoretical method has strengths and weaknesses regarding the scope of molecules and accuracies, and do not necessarily indicate the superiority of one method. Although **QM-DB** is less extensive than **CombiSolv-QM**, SMD-M06-2X can be used as a complementary method to COSMO-RS for explaining the errors of QM methods and ML models after model development.

With regards to computational cost, COSMO-RS is typically a more cost-efficient option for high-throughput calculations than SMD-DFT because COSMO-RS needs DFT calculations of a charge density profile only once per one solute/solvent. In contrast, SMD-DFT methods (e.g., SMD-M06-2X/def2-TZVP) need multiple geometry optimizations and thermochemistry calculations for the same solute when a solvent changes. SMD parameters are tabulated for 179 solvents, limiting the molecular scope for estimating ΔG_{solv} . However, SMD-M06-2X/def2-TZVP can show

higher accuracies than COSMO-RS for certain functional groups. These multiple theoretical methods would lead to more reliable evaluation of databases and predictive models than only one method.

5.3.2: Development of graph neural networks for solubility prediction

The GNNs for predicting solubility were constructed, as shown in **Figure 5.2C**. The model takes 2D molecular structures (SMILES strings) of solvent and solute as inputs, and each undergoes separate message passing. The overall architecture of two GNNs (GNN-Solvent and GNN-Solute) is inspired and modified from our previously implemented GNNs for predicting bond dissociation enthalpy and cetane number (reactivity of fuel compounds).^{59,79} It consists of three blocks representing a molecule's atom, bond, and global state. Initial atom, bond, and global features are embedded as 128-dimensional vectors and pass through five message-passing layers. This work did not consider stereochemical information as the atom features since ΔG_{soln} values of stereoisomers are scarce in the existing experimental databases (114 out of 11,637). In each layer, mathematical operations among feature vectors lead to mutual updates, such that the model captures the influence of local atom/bond environments and global molecular structures on solubility. Each GNN outputs a 128-dimensional latent vector for solvent and solute, respectively. These two vectors are concatenated and undergo additional dense layers to account for solute-solvent interactions, and finally, ΔG_{soln} is predicted.

Our model is unique compared to other GNNs recently developed for solubilities.⁵⁴ First, we minimized the number of atom features from 11 to five.

Second, the dimensions of hidden layers were also minimized. Our model has hidden layers with 128 and 256 nodes before and after concatenation, respectively, whereas 200 and 500-dimensional hidden layers were used in the literature⁵⁴ (details about the hyperparameter tuning can be found in Appendix D). Third, a separate global state block show comparable accuracies (*vide infra*), but the global updates within message-passing layers can incorporate the effects of long-range interactions on ΔG_{solv} into atom and bond features.⁸⁴ This leads to the reliable chemical explanation of atom-wise contributions to ΔG_{solv} using the modified version of Shapley additive explanation (SHAP) analysis⁸⁰ (*vide infra*). Meanwhile, other operations besides concatenation have also been reported in previous studies to consider molecular interactions, such as global convolution among molecules and graph-of-graphs neural networks.^{81,82} However, concatenating latent vectors was sufficient to achieve the accuracy close to experimental uncertainty (Mean absolute error of the test set around 0.2 kcal/mol, details in the next section).

We selected four global features after testing various molecular descriptors. Two are surface area descriptors: topological polar surface area (TPSA) and Labute accessible surface area (ASA). Each descriptor indicates different underlying chemistry of solutes/solvents. TPSA accounts for the molecule's viability to dipole-dipole interactions by quantifying the surface area of polar atoms, whereas Labute ASA is relevant to Van der Waals radii of atoms and encodes long-range dispersion interactions.⁸³ The Pearson correlation coefficients r between ΔG_{solv} in **Exp-DB** and the descriptors are -0.56 and -0.73 for solute's TPSA and Labute ASA, respectively (Close to ± 1 indicates stronger correlation). In contrast,

r between TPSA and Labute ASA is only 0.28, indicating that these two descriptors can independently explain ΔG_{solv} well and it is necessary to consider both descriptors for global features. Two additional descriptors were adopted: number of hydrogen bond donors and acceptors. They were also used in our predictive model for cetane number,⁷⁹ leading to the model with higher accuracy than that without these descriptors.

5.3.3: Semi-supervised distillation for self-evolving databases and graph neural networks

Building the GNN model and databases (**Figure 5.2**) was followed by training the model based on SSD (**Figure 5.3A**). The SSD was initiated by training the Teacher model using **Exp-DB** (Cycle 0). The trained model was then used for augmenting the database; new solute-solvent pairs were gathered from **CombiSolv-QM**, and their ΔG_{solv} was predicted using the Teacher model. The predicted values were compared with COSMO-RS solubilities stored in **CombiSolv-QM**. If the absolute difference between these two is below 0.2 kcal/mol, the corresponding data points were held in the augmented database (**Aug-DB-1**) with Teacher-predicted solubility values. It should be emphasized that the ML-predicted values are saved instead of ΔG_{solv} from COSMO-RS. This is for refining data points based on the solubility trends learned from **Exp-DB** while maintaining the reliability gained via labeled QM solubility values. The threshold value was set to 0.2 as the uncertainty of experimental measurements of ΔG_{solv} is typically up to 0.2 kcal/mol.^{29,35,69,70} If the deviation between ML and QM is below 0.2, it can be assumed that the difference is mainly from experimental uncertainty and the prediction from the Teacher is credible.

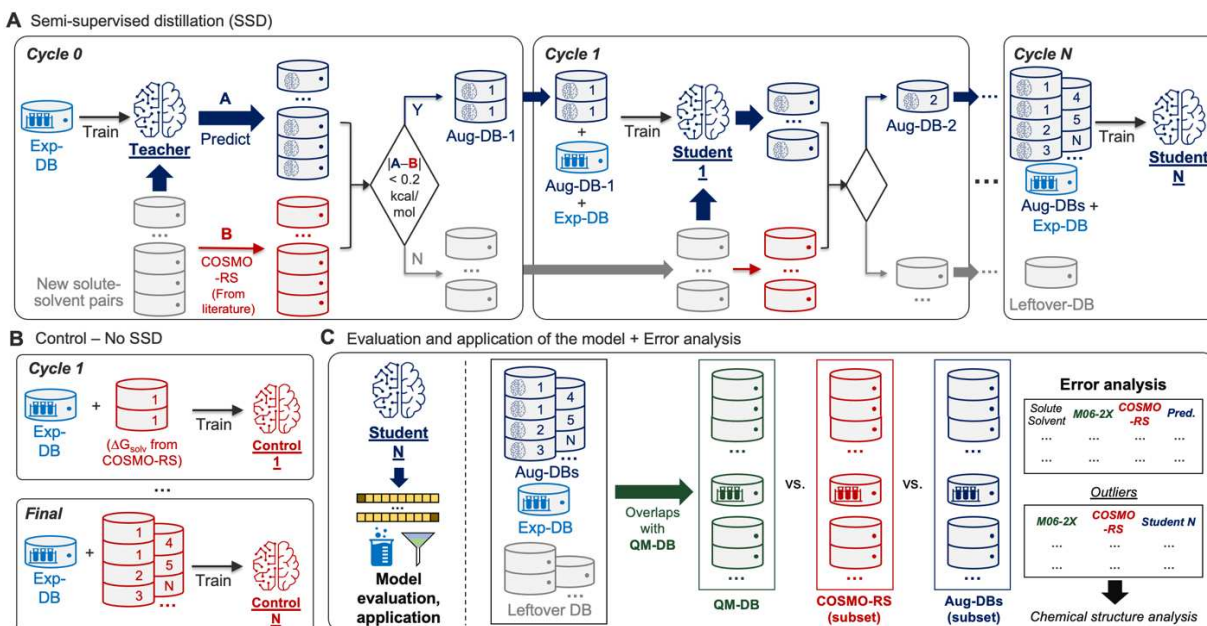


Figure 5.3: (A) Semi-supervised distillation (SSD) for self-evolving solubility databases and graph neural networks. (B) Control for comparing the accuracies of models with and without SSD. (C) A schematic description of evaluation, application, and error analysis of the model obtained from SSD.

Next, Student 1 was trained using the database combining **Aug-DB-1** and **Exp-DB** (Cycle 1), and the same procedure was carried out for the solute-solvent pairs that remained after extracting **Aug-DB-1**. Student 1 predicted ΔG_{solv} for the remaining pairs, and the predicted values were subject to the 0.2 kcal/mol cutoff, resulting in **Aug-DB-2**. These cycles were repeated multiple times, enabling the self-training of ML models. The database is grown gradually, and subsequent student models learn larger databases that contain ΔG_{solv} values refined based on the guidance from previous Students and COSMO-RS solubilities. Such gradual integration leads to better accuracy than combining the whole **CombiSolv-QM** with **Exp-DB** and re-training simultaneously (Details in the next section).

No trained weights of the GNN model are transferred from the previous cycle when training the Student model in the current cycle. Only the databases (**Aug-DBs**

and **Exp-DB**) are carried over, and each Student is trained from scratch at each cycle. In other words, the current Student is totally blind to the training results of previous Students. Therefore, at each cycle, the model learns new relationships between chemical structure and solubility that are not biased by previous cycles but are comprehensively applicable to all molecules from the previous and current cycles. This SSD scheme ensures that the new **Aug-DB-*i*** at the *i*-th cycle is integrated well with the databases accumulated from earlier cycles, and it shows no significant discrepancies and anomalies during the training. Recent studies expanded the databases using computational methods, but the model was trained sequentially or independently for the experimental and computational databases due to the discrepancies from heterogeneous data sources.^{53,54,59,76} Meanwhile, a few studies attempted the ‘ Δ -learning’ approach, where theory-experiment differences are trained and predicted.^{84,85} It should be emphasized that our approach is the first achievement of training the whole integrated database while carrying out the data augmentation and discrepancy corrections concurrently.

Ultimately, the N^{th} cycle yields the ‘Student *N*’ model and the integrated database containing **Exp-DB** and *N* **Aug-DBs**. The cycle was terminated when the errors of the **Exp-DB** test set did not show any more significant improvement. This stopping criterion finds the cycle when the remaining data points in **CombiSolv-QM** no longer synchronize well with the large **Aug-DBs** cumulated during previous cycles. The solute-solvent pairs not included in **Aug-DBs** were stored in the so-called **Leftover-DB**. As a result, the Student 35 model obtained after 35 SSD cycles led to

an optimal accuracy, with a total of 932,509 ΔG_{solv} in the integrated database (Details in **Figure 5.4**, *vide infra*).

5.3.4: Analysis of solubility model performance

Accuracies of the Student models from SSD were compared with those trained by the databases simply combining ΔG_{solv} values from experiments and COSMO-RS (Control, **Figure 5.3B**). The analysis on Control was performed at every SSD cycle to compare the increasing/decreasing trends of MAEs and RMSEs when the models are trained without/with SSD. All Control models are examined to demonstrate that the SSD approach in **Figure 5.3A** is feasible for maximizing the database size while minimizing the discrepancy between experimental and computational ΔG_{solv} and achieving the best accuracy.

The resulting model was then subject to subsequent evaluation, error analysis, and applications (**Figure 5.3C**). To evaluate the model's accuracy, mean absolute errors (MAEs), root-mean-square errors (RMSEs), and distributions of errors were investigated. For additional error analysis, we obtained the solute-solvent pairs in **QM-DB** that overlap with those in other databases (**Aug-DBs**, **Exp-DB**, **Leftover DB**). Next, we compared their ΔG_{solv} values acquired from four different sources: Experiments (if available), predictions from Student *N*, SMD-M06-2X/def2-TZVP, and COSMO-RS calculations. Outliers were identified from this comparison, and their chemical structures were analyzed to assess the strengths and weaknesses of each QM method or ML model.

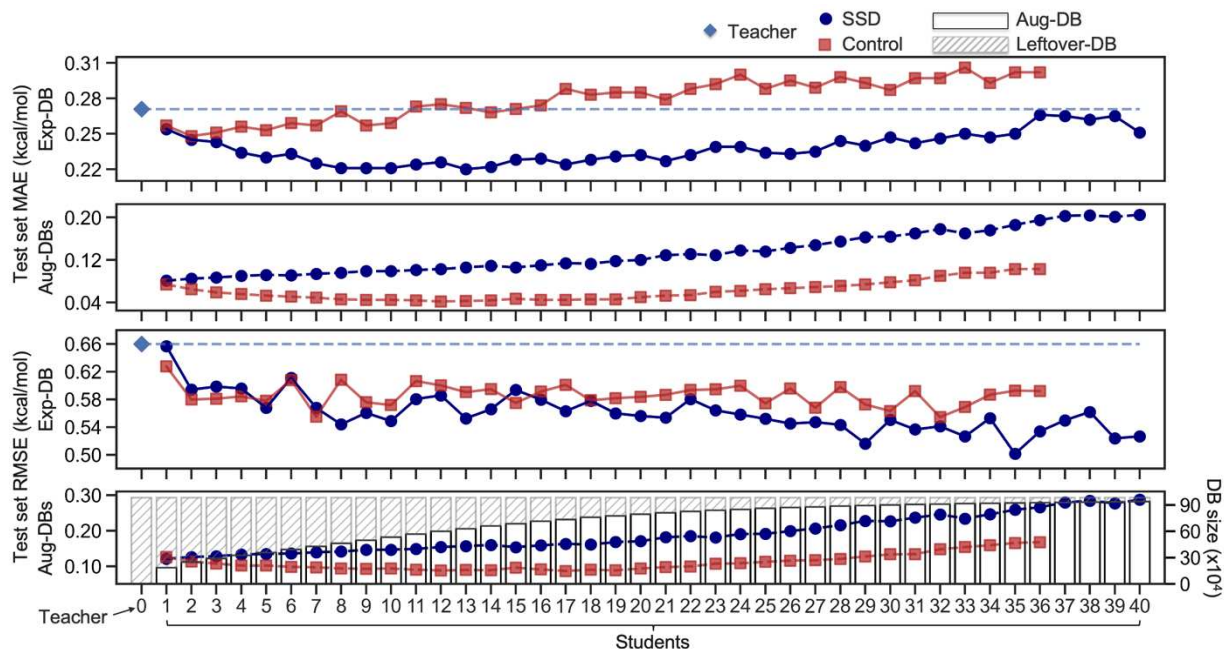


Figure 5.4: Mean absolute errors (MAEs) and root-mean-square errors (RMSEs) of test sets of Aug-DBs and Exp-DB during the SSD and the sizes of Aug-DBs.

Figure 5.4 illustrates the results from the SSD training (**Figure 5.3A**) of the GNNs shown in **Figure 5.2C**. The initial training to obtain the Teacher model resulted in the MAE of 0.27 kcal/mol for the test set of **Exp-DB**. As the SSD cycles proceeded, **Aug-DBs** gradually increased. Interestingly, the MAE for the **Exp-DB** test set reached a minimum at Student 13, while the database was grown from 11,637 to 639,925 data points. Student 13 achieved an MAE of 0.22 kcal/mol for the **Exp-DB** test set. This indicates that the SSD scheme works properly in the data augmentation while the model still captures experimental solubility trends. On the contrary, the MAEs did not decrease in Control models (0.27 kcal/mol for both Teacher and Control-13), demonstrating that simply merging solubilities from experiments and COSMO-RS is not advantageous for accurate predictions of experimental solubilities. Moreover, at the 13th SSD cycle, Control shows a discrepancy of 0.23 kcal/mol between test set MAEs of **Exp-DB** and **Aug-DBs** (0.27 vs. 0.04), whereas that from SSD is only 0.11

kcal/mol (0.22 vs. 0.11). In other words, more severe overfitting to **Aug-DBs** occurred in Control than in SSD.

It is hard to guarantee that 13 SSD cycles are sufficient to obtain the best model since the MAE is not the only metric for evaluating the accuracy. We analyzed RMSEs of Student models that show more irregular trends than MAEs. The initial Teacher training resulted in the RMSE of 0.66 kcal/mol for the test set of **Exp-DB**. As the SSD proceeds, the **Exp-DB** test set RMSE gradually decreases in general, while it fluctuates intermittently until Student 22. The accuracy of SSD models begins to surpass Control models in terms of RMSEs after Student 22. The best accuracy was achieved in Student 35 with an RMSE of 0.50 kcal/mol, whereas the RMSE of the 35th Control model is 0.59 kcal/mol. Although the MAE slightly increased from Student 13 to Student 35 (0.22-0.25 kcal/mol), the RMSE reaches a minimum with the larger database (932,509 data points) compared to Student 13 (639,925 data points). The SSD cycles after Student 35 did not effectively improve the accuracy. RMSE is a good metric for penalizing large errors of outliers,⁸⁶ indicating that Student 35 effectively alleviates prediction errors of **Exp-DB** outliers while maintaining reliable accuracy for other data points. It should be emphasized that MAE was used for the loss function (Details in Appendix D), but RMSE was also minimized during the later stages of SSD. This result implies the importance of including a large amount of data to minimize high prediction errors of outliers by iterating the SSD loop multiple times. Moreover, the best accuracy was obtained in Student 35 when the prediction accuracy was

assessed against the ‘external data set’ of 371 experimental partition coefficients (*vide infra* for details).

Meanwhile, the test set RMSEs of **Aug-DBs** increased (0.12-0.29 kcal/mol) from Teacher to Student 35. However, they are lower than the lowest test set RMSE of **Exp-DB** (0.50 kcal/mol), and the RMSE difference between **Exp-DB** and **Aug-DBs** decreased, indicating the mitigation of overfitting. **Aug-DBs** grow slower as SSD cycles proceed; the number of added data points is 183,432, 65,833, and 41,560 for the first three cycles, and gradually decreases from 34,013 to 3,168 until the 35th cycle (Details in Appendix D). The remaining solute-solvent pairs in the late SSD cycles are mostly problematic cases (Details in *Error analysis of solubility prediction*) whose true ΔG_{solv} are dubious; therefore, fewer solute-solvent pairs satisfy the cutoff. Different functional group distributions of **Aug-DBs** also influence the RMSE trends shown in **Figure 5.4**; there is the tradeoff between ‘over-generalization’ to existing functional groups and newly introduced ones (Details in Appendix D).^{87,88}

Moreover, the box plot in **Figure 5.5A** demonstrates that the SSD up to 35 cycles is beneficial to obtain an optimal model. We chose Students 13, 29, and 35, which resulted in the local minima of RMSEs during the SSD (**Figure 5.4**), in addition to Teacher. For the test set of **Exp-DB**, Student 13 shows more significant outliers (gray dots) with higher errors than the Teacher. The error of the first outlier becomes even higher in Student 29 than in Student 13. However, such errors of outliers become lowest in Student 35, which indicates the mitigation of overfitting through SSD. The outlying behavior is remedied in Student 35 while maintaining a lower MAE and similar

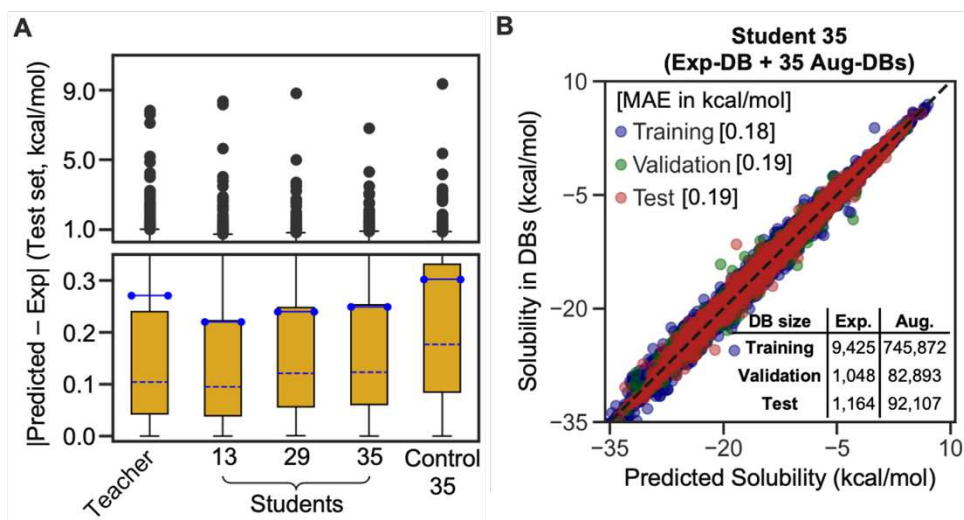


Figure 5.5: (A) Box plots of absolute error distributions for the test set of Exp-DB, for the four representative models from SSD and one Control model. (Yellow box: interquartile range, blue line: mean, blue dotted line: median, lower/upper bound of the error bar: 5th/95th percentile, gray dots: outliers beyond the 95th percentile.) (B) Parity plot of solubility values in the databases vs. those from the predictions of the best-case Student 35 model.

interquartile range (yellow box) compared to the Teacher. In contrast, the accuracy of Control 35 is even worse than Teacher, and their outliers also show higher errors. **Figure 5.5B** illustrates the parity plot of the solubility values from the databases vs. those from the predictions of Student 35. For the whole integrated database (**Exp-DB** + 35 **Aug-DBs**), Student 35 achieved balanced accuracies among the training, validation, and test sets, with MAEs of 0.18, 0.19, and 0.19 kcal/mol, respectively.

This model also showed comparable accuracies with the GNN models in the literature⁵⁴ for the **Exp-DB** test set, and higher accuracies than computational methods. **Table 5.1** summarizes the accuracies of GNNs from the literature⁵⁴, our GNNs (Student 35), COSMO-RS and SMD-M06-2X in predicting ΔG_{solv} values in **Exp-DB**. Student 35 showed comparable MAEs/RMSEs with the model in the literature (MAE: 0.20/0.22 kcal/mol, RMSE: 0.44/0.50 kcal/mol for ours/literature) with 150 more data points in the test set. Although a fair comparison was not possible due to

the different test sets, the MAE and RMSE differences of only 0.02 and 0.06 kcal/mol demonstrate the high reliability of our model and feasibility of SSD. In addition, Student 35 surpasses the accuracies of COSMO-RS and SMD-M06-2X. The MAEs and RMSEs of Student 35 are lower than those of COSMO-RS for the 3,254 and 5,608 data points in **CombiSolv-QM** and **QM-DB** overlapping with **Exp-DB**, respectively. Student 35 is still more accurate than COSMO-RS and SMD-M06-2X/def2-TZVP in all cases when these data points are categorized into training, validation, and test sets. It should be emphasized that our model predicts ΔG_{solv} more accurately while being computationally much less demanding (< 1 second for GNNs vs. several hours or days for QM methods).

Comparison with the GNNs in the literature (the best-case model) ^{a,b}	GNNs in the literature			Student 35			
	# of data points (Exp-DB)	Test set MAE (kcal/mol)	Test set RMSE (kcal/mol)	# of data points (Exp-DB)	Test set MAE (kcal/mol)	Test set RMSE (kcal/mol)	
	10,145 ^a [1,014 ^{a,c}]	0.20 ^a	0.44 ^a	11,637 [1,164 ^c]	0.22	0.50	
Exp-DB overlapping with CombiSolv- QM	# of data points	COSMO-RS		Student 35			
		MAE (kcal/mol)	RMSE (kcal/mol)	Training/validation/ Test	# of data points	MAE (kcal/mol)	RMSE (kcal/mol)
	3,254 ^a	0.40 ^a	0.67 ^a	Training	2,651	0.13	0.23
				Validation	286	0.26	0.59
				Test	317	0.22	0.39
				All	3,254	0.15	0.30
Exp-DB overlapping with QM-DB	# of data points	SMD-M06-2X/def2-TZVP		Student 35			
		MAE (kcal/mol)	RMSE (kcal/mol)	Training/validation/ Test	# of data points	MAE (kcal/mol)	RMSE (kcal/mol)
	5,608	0.62	0.88	Training	4,534	0.13	0.20
				Validation	498	0.21	0.42
				Test	576	0.22	0.39
				All	5,608	0.14	0.25

Table 5.1: Accuracies of the GNNs in the literature, our GNNs, COSMO-RS, and SMD-M06-2X in predicting ΔG_{solv} in Exp-DB, with the size of the databases used for evaluating accuracies. ^aFrom the literature.⁵⁴ ^bOf note, the experimental databases and test sets are not the same, and it is an indirect comparison. ^cTest set sizes.

To verify that the model is not prone to overfitting, we carried out 10-fold cross-validation as depicted in **Figure 5.6**. **Figure 5.6A** shows how the databases (**Exp-DB** and **Aug-DBs**) are split into 11 data subsets (Subsets **A – K**, each of which

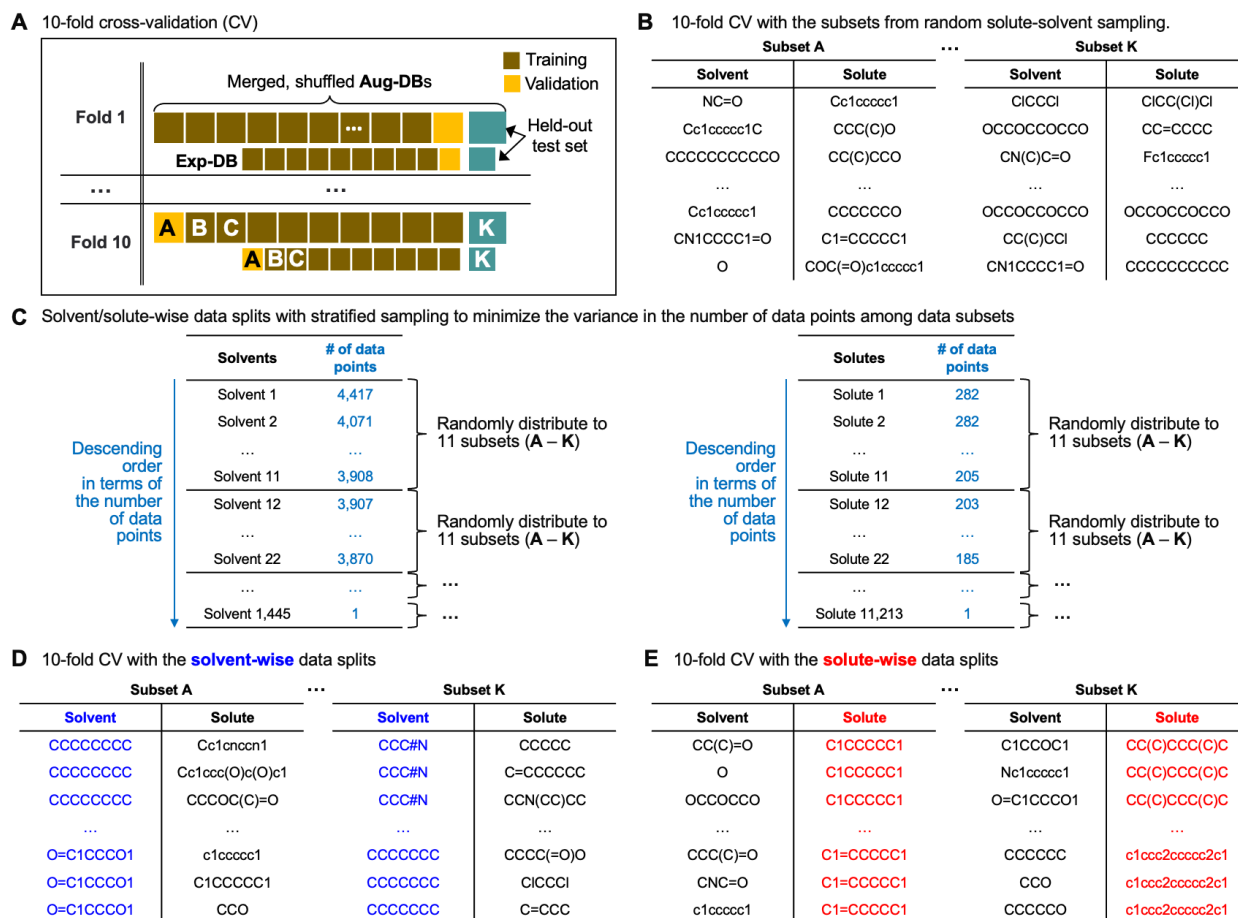


Figure 5.6: (A) Schematic illustration of the 10-fold cross-validation with data splits for **Exp-DB** and **Aug-DBs**. Data subsets obtained from (B) random solute-solvent sampling, (C) solvent-wise data splitting, and (D) solute-wise data splitting, with SMILES strings of some example molecules. (E) The stratified sampling scheme in solvent/solute-wise data splits to minimize the variance in the number of data points among different data subsets.

corresponds to either training/validation/held-out test set) to carry out the 10-fold cross-validation. This data set split was performed separately for **Aug-DBs** and **Exp-DB** to balance the ratio of the data points from data augmentation and experiments. Depending on the sampling methods, the constituent solutes/solvents in each subset can vary. The solute-solvent combinations can be sampled randomly, as shown in **Figure 5.6B**.

Meanwhile, solvent/solute-wise data splits are also possible with the stratified sampling **Figure 5.6C**; first, each bin of solvents/solutes was sorted in descending

order with respect to the number of data points. Then, every 11 bins with the similar number of data points were grouped, and the bins in each group were distributed randomly to 11 subsets **A – K**. Such a stratified sampling results in a certain solvent/solute included in only one subset (**Figures 5.6D** and **5.6E**). These are useful in assessing model accuracy when specific solvents or solutes in the validation/test set are unseen in the training set.

Figure 5.7 displays the **Exp-DB** test set RMSEs obtained from the 10-fold cross-validation of Teacher, three Students, and one Control model with the random solute-solvent sampling (**Figure 5.5B**). Means and standard deviations of 10 RMSEs were evaluated for each model. The mean of RMSEs decreases significantly from 0.679 to 0.577 kcal/mol when SSD proceeds from Teacher to Student 13 and is reduced further at Student 29 (0.549 kcal/mol). The accuracy of Student 35 (Mean of RMSEs: 0.546 kcal/mol) is slightly higher than Student 29, with a lower standard deviation among ten folds (0.023 and 0.021 kcal/mol for Student 29 and 35, respectively). The held-out test set from **Aug-DBs** was not considered for the

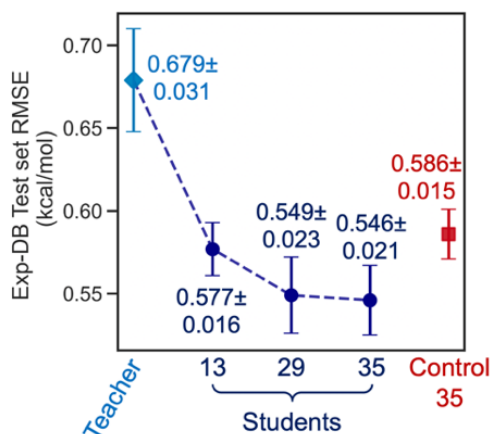


Figure 5.7: Model accuracies from the 10-fold cross-validation with the random solute-solvent sampling. RMSEs of the Exp-DB test set were evaluated for each of the 10 folds. The data points and error bars indicate the mean and standard deviation of 10 RMSEs, respectively.

evaluation since the sizes of **Aug-DBs** are different for all Student models. We also carried out the 10-fold cross-validation with the solvent-wise and solute-wise data splits (**Figures 5.5D** and **5.5E**). As a result, comparable accuracies were still obtained from these different methods of 10-fold cross-validation (Appendix D, Section S2).

All the above results demonstrate the effectiveness of SSD in improving the accuracy of the ground-truth experimental solubilities while augmenting the database. Our empirical results are consistent with mathematical proofs that several rounds of SSD enhance the accuracy of the held-out data and reduce overfitting.^{87,88} It has been verified that self-distillation amplifies the regularization of the space of trainable parameters if the model architectures for Teacher and Students are identical (**Figure 5.2C**).⁸⁸ When Students are trained using an extensive distilled database with a limited parameter space, their variance is reduced without significantly increasing its bias. In other words, the models' trainable parameters are neither overly sensitive to different training set batches nor biased to specific batches, and therefore overfitting is reduced. Meanwhile, too many rounds of SSD may over-regularize the model, leading to underfitting. These mathematical findings are in line with our SSD results shown in **Figure 5.4**. In addition, adding noises to the SSD model parameters showed minor output perturbations,⁸⁷ and the ablation analysis with removing the augmented data significantly degraded the model performance.⁷³ These previous studies further support the robustness of SSD with augmented datasets.

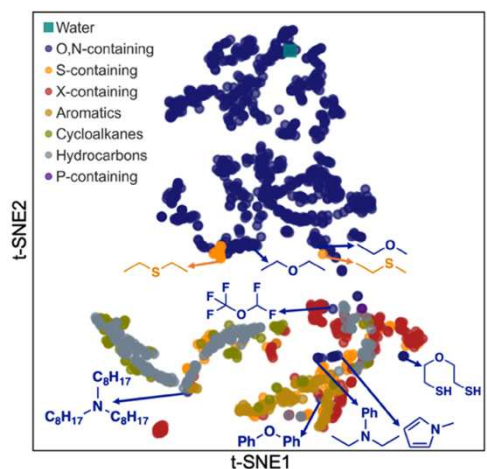


Figure 5.8: 2D plot of t-distributed stochastic neighbor embeddings (t-SNEs) for the latent vectors of 1,447 solvents obtained from Student 35 model.

Other variants of SSD were also attempted using **QM-DB**, or both **CombiSolv-QM** and **QM-DB** (Section S3 in Appendix D), but the SSD shown in **Figure 5.3A** showed the best accuracy. We also tested other variants of semi-supervised learning methods, such as noisy student self-distillation (NSSD); however, this led to higher prediction errors (See Section S4, Appendix D for detailed discussion).

Next, we carried out a clustering analysis of t-distributed stochastic neighbor embeddings (t-SNEs) of latent vectors for 1,447 solvents included in all the databases shown in **Figure 5.2A**. This analysis is to further verify the chemical feasibility of Student 35. 2D t-SNE coordinates were obtained for these solvents. Each solvent was categorized according to the priority of categories listed in the legend of **Figure 5.8**. For example, if a solvent contains both O and S, it is classified as ‘O,N-containing’ because O has higher priority than S. We identified certain clustering patterns among several categories: O,N-containing (upper side), halogen (X)-containing (mainly lower right), and hydrocarbon solvents (mainly lower left). O,N-containing solvents

exclusively occupy a specific region, possibly because they are solvents that can participate in hydrogen bonds and show characteristic solubility trends.

However, some O,N-containing solvents are located near other molecular groups, such as aromatics, hydrocarbons, and X-containing ones. These solvents contain oxygen or nitrogen, with the other atoms corresponding to the molecular groups they are close to. For instance, trioctylamine is in the cluster of Hydrocarbons since it has three alkyl chains having eight carbons per each. Pentafluorodimethyl ether was found adjacent to the X-containing cluster. Ethers, amines, and pyrroles with aromatic rings are placed around the group of aromatic solvents. Meanwhile, an ether with two thiol groups (2-mercaptoethyl ether) was found near S-containing solvents rather than O,N-containing ones, indicating that their behavior as a solvent is close to S-containing solvents rather than O,N-containing ones. Conversely, some sulfides (diethyl sulfide, ethyl methyl sulfide) are near their ether analogs, implying that their chemical behavior could be analogous to ethers.

5.3.5: Error analysis of solubility prediction

The error analysis was performed by comparing the **QM-DB** solubilities calculated in the SMD-M06-2X/def2-TZVP level of theory with those from COSMO-RS, experiments, and the Student 35 model. First, we analyzed **Exp-DB** solute-solvent pairs where SMD-M06-2X outperforms COSMO-RS and vice versa to identify the advantages and disadvantages of each theoretical method (**Figure 5.9A**). The left side of **Figure 5.9A** illustrates the five cases whose absolute error between ΔG_{solv} from experiment and SMD-M06-2X does not exceed 0.2 kcal/mol, whereas COSMO-RS

shows the worst performance. All these five cases correspond to polar solutes and solvents with halogen atoms, hydrogen bond donors and acceptors. SMD-M06-2X better reproduces the experimental solubilities of these molecules than COSMO-RS, which may be in part attributed to the halogenicity, hydrogen bond acidity, and basicity parameters used by SMD. For these five cases, the predictions of Student 35 should not be close to COSMO-RS; instead, it should reproduce experimental values, although COSMO-RS ΔG_{solv} were referred during the SSD. **2** and **4** show the jeopardy of the predicted value biased to COSMO-RS. However, for **1**, **3**, and **5**, Student 35's predictions are closer to the **Exp-DB** solubilities than those from COSMO-RS. Moreover, Student 35 showed the outliers with the lowest errors among other Students (**Fig. 5.5A**). This indicates that the distillation process (**Figure 5.3A**) effectively corrected the discrepancy between experiment and theory.

We also analyzed five different solute-solvent pairs for which COSMO-RS outperforms SMD-M06-2X (Right side of **Figure 5.9A**). They are solutes and solvents with low or no polarity or molecules with special moieties such as ozone. Investigating the two extreme cases shown in **Figure 5.9A** demonstrates the importance of accounting for multiple theoretical methods in assessing the results from SSD.

The analysis on SMD-M06-2X and COSMO-RS was then followed by the outlier analysis of Student 35 (**Figure 5.9B**). The outliers correspond to the gray dots in **Figure 5.5A**; Their extraordinary chemical structures indicate that they do not meaningfully deteriorate the model's accuracy. The top five outliers include solutes with multiple

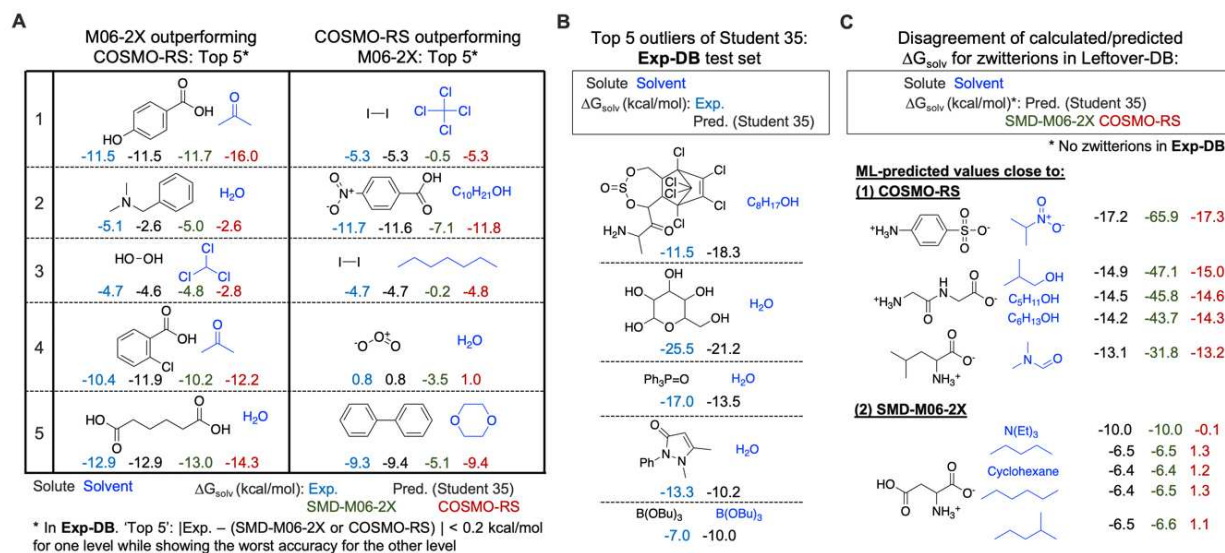


Figure 5.9: (A) Top 5 solute-solvent pairs in **Exp-DB** where the SMD-M06-2X/Def2-TZVP level outperforms COSMO-RS in calculating ΔG_{solv} , and vice versa. (B) Top 5 outliers of the Student 35 model when comparing the predicted ΔG_{solv} with those in the test set of **Exp-DB**. (C) In **Leftover-DB**, the disagreement of ΔG_{solv} among Student 35, SMD-M06-2X and COSMO-RS mainly occurs for zwitterion solutes which do not exist in **Exp-DB**.

complex rings, five hydroxy groups, and heteroatoms (P and B) that rarely appear in the whole database (932,509 data points). For example, the solutes with a P=O double bond and aromatic substituents appear only in 808 data points, and only 69 data points have solutes/solvents with B-O single bonds. These outliers occurred not because of the overfitting to **Aug-DBs**, but due to the chemical moieties rarely seen from the databases.

Further analysis was performed for **Leftover-DB** consisting of 57,721 solute-solvent pairs in **CombiSolv-QM** that were not included in the **Aug-DBs** but remained after 35 SSD cycles. Since **Leftover-DB** does not have experimental values, we compared their ΔG_{solv} values from Student 35, SMD-M06-2X and COSMO-RS for 14,053 out of 57,721 data points whose ΔG_{solv} from all the three models or methods are available. As a result, significant discrepancies were observed for 1,381 data points having zwitterionic solutes. Ten extreme cases are shown in **Figure 5.9C**. SMD-

M06-2X relatively overestimates ΔG_{solv} compared to the other two for the above five cases, whereas the ΔG_{solv} from COSMO-RS shows disagreement with Student 35 and SMD-M06-2X for the below five ones. It should be emphasized that no zwitterions are available in **Exp-DB**. However, 5,446 zwitterions were already included in **Aug-DBs** during the SSD; these species have no experimental ground-truth ΔG_{solv} values. Such a lack of data availability for zwitterions necessitates experimental measurements for their solubility values or additional reliable theoretical methods, possibly leading to a more extensive database from SSD, including zwitterions.

Although the above error analysis suggests room for improving our model, it is sufficiently reliable to be utilized in the practical design of solvent systems in various chemical processes such as catalysis and separation. The following sections demonstrate the application of our model to practical examples.

5.3.6: Application 1 – Linear Solvation Energy Relationships (LSERs) between solvation free energy and reaction rates

LSERs inform solvent selection to maximize the reaction rates in designing organic reactions.^{2,7,8} Here, we demonstrate the application of our ML model to discover LSERs in 11 organic reactions. Gibbs solvation free energies of the product(s) and reactant(s) ($\Delta G_{\text{solv}}(\mathbf{P})$ and $\Delta G_{\text{solv}}(\mathbf{R})$, respectively) were predicted using Student 35. If a reaction has two or more reactants or products, the sum of their ΔG_{solv} values was used as $\Delta G_{\text{solv}}(\mathbf{R})$ or $\Delta G_{\text{solv}}(\mathbf{P})$. These values and their differences [$\Delta G_{\text{solv}}(\mathbf{P}) - \Delta G_{\text{solv}}(\mathbf{R})$] were used as the descriptors to find highly positive or negative Pearson correlation coefficients r (i.e., close to 1 or -1) with experimental reaction rates in

different solvents. According to the Hammond postulate,⁸⁹ at least one of these three descriptors should show a high correlation (Details in Section S5, ESI[†]). The reaction rates were collected from the literature.^{2,90,91}

Figure 5.10 depicts the correlation for 11 organic reactions with varying solvents. For each reaction, we chose one descriptor that shows the strongest correlation with experimental reaction rates. $\Delta G_{\text{solv}}(\mathbf{P})$ is the best descriptor for the reactions **I**, **II**, and **III**, with Pearson ρ values of -0.95, -0.90, and -0.68, respectively, whereas $\Delta G_{\text{solv}}(\mathbf{R})$ was chosen for **IV~VII** ($\rho=0.80-0.99$). The rest four reactions (**VIII~XI**) can be explained by $\Delta G_{\text{solv}}(\mathbf{P}) - \Delta G_{\text{solv}}(\mathbf{R})$

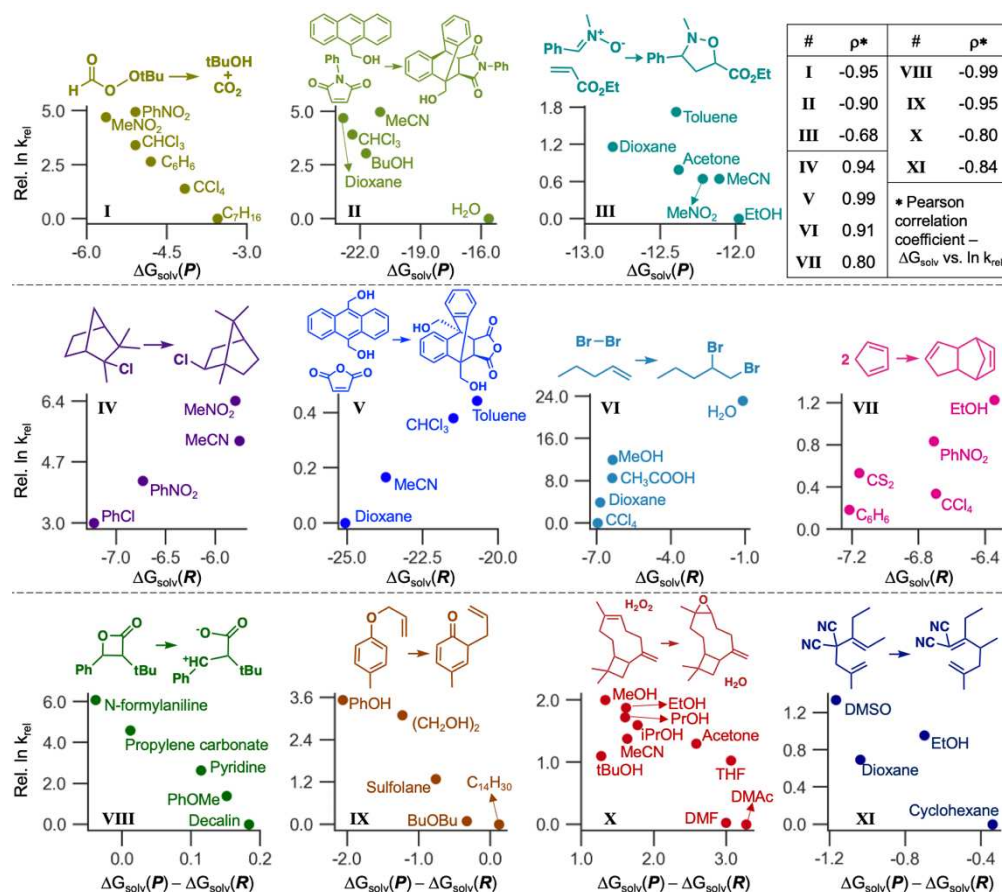


Figure 5.10: (A) Top 5 solute-solvent pairs in Exp-DB where the SMD-M06-2X/Def2-TZVP level outperforms COSMO-RS in calculating ΔG_{solv} , and vice versa. (B) Top 5 outliers of the Student 35 model when comparing the predicted ΔG_{solv} with those in the test set of Exp-DB. (C) In Leftover-DB, the disagreement of ΔG_{solv} among Student 35, SMD-M06-2X and COSMO-RS mainly occurs for zwitterion solutes which do not exist in Exp-DB.

as a descriptor (ρ from -0.99 to -0.80). Our ML model also showed reliable results ($\rho=-0.80$) in the complex reaction example, such as the epoxidation of β -caryophyllene investigated in 10 different solvents (**X**). These correlations are also chemically explainable (Section S5, Appendix D). While **Figure 5.10** shows only the best-case descriptor, ρ values for all three descriptors are available for each reaction (Section S5 Appendix D), with the reason for selecting a particular descriptor. Of note, some of the above 11 reactions were not performed at room temperature, whereas our GNN gives ΔG_{solv} at room temperature. Considering the temperature dependence of solubility would improve GNNs and LSERs, although the results in **Figure 5.10** already show decent correlations.

The ΔG_{solv} difference of only around 1 kcal/mol can significantly affect reactivity predictions (**Figure 5.10**), demanding a fast and accurate ML model. Designing solvent systems using our GNN is advantageous because it is fast compared to expensive QM calculations while being accurate. Solvents that maximize reaction rates can be designed by predicting $\Delta G_{\text{solv}}(\mathbf{R})$ and $\Delta G_{\text{solv}}(\mathbf{P})$, finding LSERs, and extrapolating the relationship for the new solvents in which experiments have not been performed. Although ΔG_{solv} of transition states is not considered here, our model enables rapid solvent screening before investigating the transition states.

5.3.7: Application 2 – Prediction of partition coefficients for lignin-derived monomers and drug-like molecules

Description of the Log P dataset		Prediction methods	Kendall τ	RMSE
Set A	300 data points (30 depolymerized lignin derivatives, 10 organic solvents)	GNN (Student 35)	0.87	0.51
		COSMO-RS	0.77	0.50
Set B	63 data points (17 drug-like compounds, 4 organic solvents)	GNN (Student 35)	0.70	1.15
		COSMO-RS	0.77	1.00

Table 5.2: Comparison of prediction accuracies of 363 partition coefficients (log P) for COSMO-RS and GNN for two datasets with their description.

As the second application example, we examined our GNN model by predicting 363 water-organic partition coefficients (log P) of depolymerized lignin derivatives and drug-like compounds (Sets A and B, **Table 5.2**). Their experimental log P values are available from the literature.⁹² Predicting log P would enable the design of solvents that effectively extract useful biomass-derived compounds and drug molecules with high yields. Here, we predicted the ΔG_{solv} values in water and organic solvents for Sets A and B and evaluated log P values using the formula: $\log P = (\Delta G_{\text{solv,water}} - \Delta G_{\text{solv,org}}) / 2.303RT$. Two metrics were used to assess accuracies: RMSE and Kendall rank correlation coefficient (τ). A τ value closer to 1 indicates a stronger rank correlation (i.e., higher accuracy). These metrics were chosen because the literature⁹² used the same metrics in assessing COSMO-RS.

To further verify the feasibility of SSD, we compared the accuracies of our Teacher and Students from SSD with the COSMO-RS method (**Figure 5.11A**). For both Sets, the latter Students show lower RMSEs than the earlier ones and Teacher, and their RMSEs become comparable to that of COSMO-RS, demonstrating the effectiveness of the SSD. Regarding Kendall τ of Set A, our GNN models even exceed the accuracy of COSMO-RS. For Set B, the

τ values of Students are lower than that of COSMO-RS. However, it displays an increasing trend for latter Students, indicating the strength of SSD.

Table 5.2 compares τ and RMSEs for COSMO-RS and our final GNN model, Student 35. For Set A, our model showed a higher τ than COSMO-RS (0.87 vs. 0.77), whereas τ from Student 35 and COSMO-RS is 0.70 and 0.77, respectively, for Set B. Our GNN resulted in a better correlation for Set A than COSMO-RS, while COSMO-RS performed slightly better in Set B. In terms of RMSE, our model achieved an RMSE almost identical to that from COSMO-RS for Set A. COSMO-RS showed better accuracy in Set B. Notably, Set B has fewer data points (63) than Set A (300), so Set A can assess model accuracies better than Set B; we achieve better rank correlation than COSMO-RS in Set A. **Figure 5.11B** depicts the parity plot of ML-predicted log P vs. experimental ones. Overall, the model shows predictions close to experimental ones. Log P of some cases in Set B is overestimated, but similar outliers were also found from the COSMO-RS results.⁹² Accuracies for these data points can be improved by considering ΔG_{solv} of ionic species for predicting distribution coefficients (log D) for acidic/basic solute molecules.

Another merit of our GNNs is chemical interpretability, which is possible by quantifying and analyzing atom-wise contributions based on the procedure described in our recent work.⁷⁹ This analysis was inspired by the Shapley additive explanation (SHAP) method.⁸⁰ The summation of quantified atom-wise contribution values equals the predicted ΔG_{solv} . We examined our SHAP method for GNNs for estradiol (**Figure 5.11C**) in two different organic solvents and water to explain a higher

experimental log P value in butanol-water (3.05) than in toluene-water (2.02). The aromatic moiety decreases the solubility in water because it is a polar solvent. The same moiety shows solvation stabilization in toluene, presumably due to non-covalent interactions between toluene and estradiol, whereas the methyl and hydroxyl groups in the aliphatic ring display unfavorable interactions with toluene. Estradiol in 1-butanol shows overall stabilization since either alkyl or hydroxy group in 1-butanol can stabilize aliphatic, aromatic, and hydroxy groups through dispersion

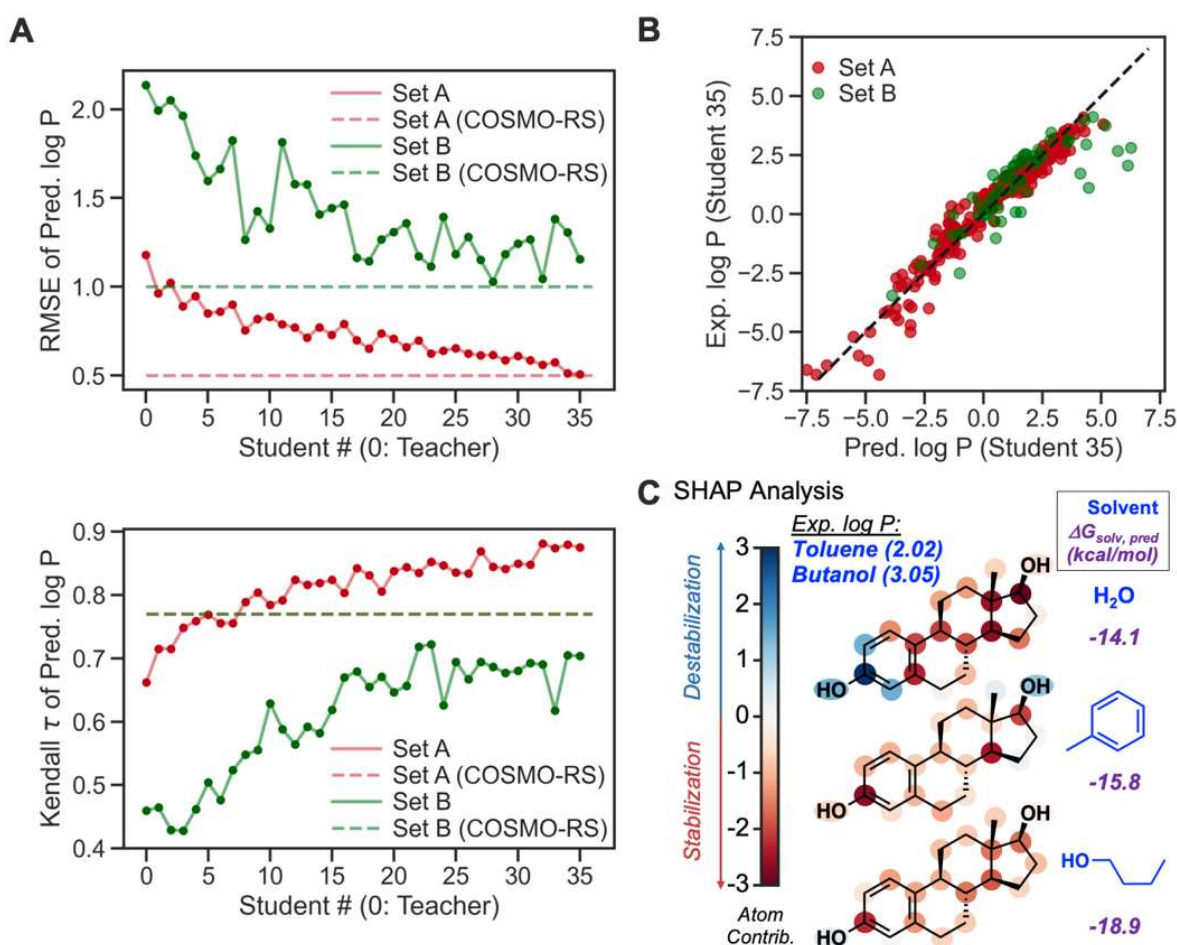


Figure 5.11: (A) Prediction accuracies of log P for Teacher and Student models when RMSE and Kendall rank correlation coefficient (τ) are used as metrics. (Red: Set A, Green: Set B) (B) A parity plot showing the experimental vs. predicted log P values for a total of 363 data points. (C) Atom-wise contribution values obtained using the Shapley additive explanation (SHAP) method, for ΔG_{solv} of estradiol in three different solvents. The SHAP analysis was conducted to chemically explain different log P values in two organic solvents.

interaction and hydrogen bonds. Such difference in atom-wise contributions and chemical interactions leads to higher solvation stabilization in 1-butanol than in toluene, and thus a higher log P. The quantified contribution values are consistent with chemical knowledge, enabling the chemistry-informed design of solvent systems.

All these results indicate that our GNNs reliably capture solubility trends and accurately predict log P values in different organic solvents. It should be emphasized that GNN predictions of log P take less than one second and are as accurate as COSMO-RS, whereas QM and COSMO-RS calculations are expensive (usually several hours or days per one molecule). Rapid and reliable log P predictions would lead to the computational solvent design for separation processes in organic, pharmaceutical synthesis, and renewable energy industries.

5.4: Conclusions

Solubility is a critical molecular property when designing chemical processes such as synthesis and separation in organic, pharmaceutical, and sustainable chemistry. Many ML models have been developed but lack reliable integrations of experimental and computational solubility databases to maximize the database size and, thus, prediction accuracy. To reduce the discrepancies among different data sources, semi-supervised self-training methodologies were adopted in solubility predictions, leading to self-evolving solubility databases and GNN predictive models. The resulting model showed reliable accuracy and was subsequently applied to practical examples of solvent selection in chemical reactions and separation

processes. All these results demonstrate the practical applicability of the developed model to the design of solvent systems in chemical processes.

Such approaches can be potentially improved by employing multiple QM methods during data augmentation. Particularly, additional QM calculations can be performed for transition states, radicals, ions, and other charged species in solution phase using reliable methods. Accuracies of predicting solubilities of zwitterions can also be improved. In terms of the molecular scope, the database adopted for model training can be expanded to organometallic complexes in addition to the organic molecules in the current database. Meanwhile, considering temperature effects on solubility in ML models should be pursued in the future to achieve the application of the model to a broader scope of chemistry. Lastly, predicting solubilities in multicomponent solvents is another challenge in developing future ML models, which would lead to the realistic modeling of mixtures utilized in various chemical reactions and separation processes.

5.5: References

1. Dalton, T., Faber, T. & Glorius, F. C–H Activation: Toward Sustainability and Applications. *ACS Cent. Sci.* **7**, 245–261 (2021).
2. Dyson, P. J. & Jessop, P. G. Solvent effects in catalysis: rational improvements of catalysts via manipulation of solvent interactions. *Catal. Sci. Technol.* **6**, 3302–3316 (2016).
3. Huxoll, F. *et al.* Solvent Selection in Homogeneous Catalysis—Optimization of Kinetics and Reaction Performance. *ACS Catal.* **11**, 590–594 (2021).
4. Hailes, H. C. Reaction Solvent Selection: The Potential of Water as a Solvent for Organic Transformations. *Org. Process Res. Dev.* **11**, 114–120 (2007).

5. Varghese, J. J. & Mushrif, S. H. Origins of complex solvent effects on chemical reactivity and computational tools to investigate them: a review. *React. Chem. Eng.* **4**, 165–206 (2019).
6. Moseley, J. D. & Murray, P. M. Ligand and solvent selection in challenging catalytic reactions. *Journal of Chemical Technology & Biotechnology* **89**, 623–632 (2014).
7. Slakman, B. L. & West, R. H. Kinetic solvent effects in organic reactions. *Journal of Physical Organic Chemistry* **32**, e3904 (2019).
8. Sherwood, J., Parker, H. L., Moonen, K., Farmer, T. J. & Hunt, A. J. N-Butylpyrrolidinone as a dipolar aprotic solvent for organic synthesis. *Green Chem.* **18**, 3990–3996 (2016).
9. Pinho, S. P. & Macedo, E. A. Solubility in Food, Pharmaceutical, and Cosmetic Industries. <https://doi.org/10.1039/9781847557681-00305> (2007) doi:10.1039/9781847557681-00305.
10. Jouyban, A. Review of the cosolvency models for predicting solubility of drugs in water-cosolvent mixtures. *J Pharm Pharm Sci* **11**, 32–58 (2008).
11. Llinàs, A., Glen, R. C. & Goodman, J. M. Solubility Challenge: Can You Predict Solubilities of 32 Molecules Using a Database of 100 Reliable Measurements? *J. Chem. Inf. Model.* **48**, 1289–1303 (2008).
12. Bergström, C. A. S., Charman, W. N. & Porter, C. J. H. Computational prediction of formulation strategies for beyond-rule-of-5 compounds. *Advanced Drug Delivery Reviews* **101**, 6–21 (2016).
13. Bergström, C. A. S. & Larsson, P. Computational prediction of drug solubility in water-based systems: Qualitative and quantitative approaches used in the current drug discovery and development setting. *International Journal of Pharmaceutics* **540**, 185–193 (2018).
14. Fioressi, S. E., Bacelo, D. E., Rojas, C., Aranda, J. F. & Duchowicz, P. R. Conformation-independent quantitative structure-property relationships study on water solubility of pesticides. *Ecotoxicology and Environmental Safety* **171**, 47–53 (2019).
15. Nayak, A. K. & Panigrahi, P. P. Solubility Enhancement of Etoricoxib by Cosolvency Approach. *International Scholarly Research Notices* **2012**, 820653 (2012).
16. Seedher, N. & Kanojia, M. Co-solvent solubilization of some poorly-soluble antidiabetic drugs. *Pharmaceutical Development and Technology* **14**, 185–192 (2009).
17. Newmister, S. A. *et al.* Structural basis of the Cope rearrangement and cyclization in hapalindole biogenesis. *Nat Chem Biol* **14**, 345–351 (2018).

18. Kraml, J. *et al.* Solvation Thermodynamics in Different Solvents: Water–Chloroform Partition Coefficients from Grid Inhomogeneous Solvation Theory. *J. Chem. Inf. Model.* **60**, 3843–3853 (2020).
19. Esteban, J., Vorholt, A. J. & Leitner, W. An overview of the biphasic dehydration of sugars to 5-hydroxymethylfurfural and furfural: a rational selection of solvents using COSMO-RS and selection guides. *Green Chem.* **22**, 2097–2128 (2020).
20. Huber, G. W., Chheda, J. N., Barrett, C. J. & Dumesic, J. A. Production of Liquid Alkanes by Aqueous-Phase Processing of Biomass-Derived Carbohydrates. *Science* **308**, 1446–1450 (2005).
21. Shen, Z. & Van Lehn, R. C. Solvent Selection for the Separation of Lignin-Derived Monomers Using the Conductor-like Screening Model for Real Solvents. *Ind. Eng. Chem. Res.* **59**, 7755–7764 (2020).
22. Kim, Y. *et al.* Prediction of Hydroxymethylfurfural Yield in Glucose Conversion through Investigation of Lewis Acid and Organic Solvent Effects. *ACS Catal.* **10**, 14707–14721 (2020).
23. Hollas, A. *et al.* A biomimetic high-capacity phenazine-based anolyte for aqueous organic redox flow batteries. *Nat Energy* **3**, 508–514 (2018).
24. Kucharyson, J. F., Cheng, L., Tung, S. O., Curtiss, L. A. & Thompson, L. T. Predicting the potentials, solubilities and stabilities of metal-acetylacetonates for non-aqueous redox flow batteries using density functional theory calculations. *J. Mater. Chem. A* **5**, 13700–13709 (2017).
25. S. V., S. S. *et al.* Multi-objective goal-directed optimization of de novo stable organic radicals for aqueous redox flow batteries. *Nat Mach Intell* **4**, 720–730 (2022).
26. Sorkun, M. C., Khetan, A. & Er, S. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. *Sci Data* **6**, 143 (2019).
27. Bradley, J.-C. *et al.* Open Notebook Science Challenge: Solubilities of Organic Compounds in Organic Solvents. *Nat Prec* 1–1 (2010) doi:10.1038/npre.2010.4243.3.
28. Marenich, A. V. *et al.* Minnesota Solvation Database (MNSOL) version 2012. <https://doi.org/10.13020/3eks-j059> (2020) doi:10.13020/3eks-j059.
29. Kelly, C. P., Cramer, C. J. & Truhlar, D. G. SM6: A Density Functional Theory Continuum Solvation Model for Calculating Aqueous Solvation Free Energies of Neutrals, Ions, and Solute–Water Clusters. *J. Chem. Theory Comput.* **1**, 1133–1152 (2005).

30. Thompson, J. D., Cramer, C. J. & Truhlar, D. G. New Universal Solvation Model and Comparison of the Accuracy of the SM5.42R, SM5.43R, C-PCM, D-PCM, and IEF-PCM Continuum Solvation Models for Aqueous and Organic Solvation Free Energies and for Vapor Pressures. *J. Phys. Chem. A* **108**, 6532–6542 (2004).
31. Mobley, D. L. & Guthrie, J. P. FreeSolv: a database of experimental and calculated hydration free energies, with input files. *J Comput Aided Mol Des* **28**, 711–720 (2014).
32. Moine, E., Privat, R., Sirjean, B. & Jaubert, J.-N. Estimation of Solvation Quantities from Experimental Thermodynamic Data: Development of the Comprehensive CompSol Databank for Pure and Mixed Solutes. *J. Phys. Chem. Ref. Data* **46**, 033102 (2017).
33. Llinas, A. & Avdeef, A. Solubility Challenge Revisited after Ten Years, with Multilab Shake-Flask Data, Using Tight (SD ~ 0.17 log) and Loose (SD ~ 0.62 log) Test Sets. *J. Chem. Inf. Model.* **59**, 3036–3040 (2019).
34. Llinas, A., Oprisiu, I. & Avdeef, A. Findings of the Second Challenge to Predict Aqueous Solubility. *J. Chem. Inf. Model.* **60**, 4791–4803 (2020).
35. 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* **113**, 6378–6396 (2009).
36. Boothroyd, S., Kerridge, A., Broo, A., Buttar, D. & Anwar, J. Solubility prediction from first principles: a density of states approach. *Phys. Chem. Chem. Phys.* **20**, 20981–20987 (2018).
37. Palmer, D. S., McDonagh, J. L., Mitchell, J. B. O., van Mourik, T. & Fedorov, M. V. First-Principles Calculation of the Intrinsic Aqueous Solubility of Crystalline Druglike Molecules. *J. Chem. Theory Comput.* **8**, 3322–3337 (2012).
38. Skyner, R. E., McDonagh, J. L., Groom, C. R., Mourik, T. van & Mitchell, J. B. O. A review of methods for the calculation of solution free energies and the modelling of systems in solution. *Phys. Chem. Chem. Phys.* **17**, 6174–6191 (2015).
39. Klamt, A. The COSMO and COSMO-RS solvation models. *WIREs Computational Molecular Science* **1**, 699–709 (2011).
40. Ran, Y., He, Y., Yang, G., Johnson, J. L. H. & Yalkowsky, S. H. Estimation of aqueous solubility of organic compounds by using the general solubility equation. *Chemosphere* **48**, 487–509 (2002).

41. Palmer, D. S. & Mitchell, J. B. O. Is Experimental Data Quality the Limiting Factor in Predicting the Aqueous Solubility of Druglike Molecules? *Mol. Pharmaceutics* **11**, 2962–2972 (2014).
42. Boobier, S., Hose, D. R. J., Blacker, A. J. & Nguyen, B. N. Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. *Nat Commun* **11**, 5753 (2020).
43. Yang, K. *et al.* Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **59**, 3370–3388 (2019).
44. Qiu, J., Albrecht, J. & Janey, J. Solubility Behaviors and Correlations of Common Organic Solvents. *Org. Process Res. Dev.* **24**, 2702–2708 (2020).
45. Lovrić, M. *et al.* Machine learning in prediction of intrinsic aqueous solubility of drug-like compounds: Generalization, complexity, or predictive ability? *Journal of Chemometrics* **35**, e3349 (2021).
46. Lim, H. & Jung, Y. Delfos: deep learning model for prediction of solvation free energies in generic organic solvents. *Chemical Science* **10**, 8306–8315 (2019).
47. Cui, Q. *et al.* Improved Prediction of Aqueous Solubility of Novel Compounds by Going Deeper With Deep Learning. *Front. Oncol.* **10**, (2020).
48. Pathak, Y., Laghuvarapu, S., Mehta, S. & Priyakumar, U. D. Chemically Interpretable Graph Interaction Network for Prediction of Pharmacokinetic Properties of Drug-Like Molecules. *Proceedings of the AAAI Conference on Artificial Intelligence* **34**, 873–880 (2020).
49. Sorkun, M. C., Koelman, J. M. V. A. & Er, S. Pushing the limits of solubility prediction via quality-oriented data selection. *iScience* **24**, (2021).
50. Francoeur, P. G. & Koes, D. R. SolTranNet–A Machine Learning Tool for Fast Aqueous Solubility Prediction. *J. Chem. Inf. Model.* **61**, 2530–2536 (2021).
51. Tang, B. *et al.* A self-attention based message passing neural network for predicting molecular lipophilicity and aqueous solubility. *J Cheminform* **12**, 15 (2020).
52. Chung, Y. *et al.* Group Contribution and Machine Learning Approaches to Predict Abraham Solute Parameters, Solvation Free Energy, and Solvation Enthalpy. *J. Chem. Inf. Model.* **62**, 433–446 (2022).
53. Vermeire, F. H., Chung, Y. & Green, W. H. Predicting Solubility Limits of Organic Solutes for a Wide Range of Solvents and Temperatures. *J. Am. Chem. Soc.* <https://doi.org/10.1021/jacs.2c01768> (2022) doi:10.1021/jacs.2c01768.

54. Vermeire, F. H. & Green, W. H. Transfer learning for solvation free energies: From quantum chemistry to experiments. *Chemical Engineering Journal* **418**, 129307 (2021).
55. Bilodeau, C. *et al.* Generating molecules with optimized aqueous solubility using iterative graph translation. *React. Chem. Eng.* **7**, 297–309 (2022).
56. Vassileiou, A. D. *et al.* A unified ML framework for solubility prediction across organic solvents. *Digital Discovery* **2**, 356–367 (2023).
57. Yu, J. *et al.* SolvBERT for solvation free energy and solubility prediction: a demonstration of an NLP model for predicting the properties of molecular complexes. *Digital Discovery* **2**, 409–421 (2023).
58. Lee, S. *et al.* Novel Solubility Prediction Models: Molecular Fingerprints and Physicochemical Features vs Graph Convolutional Neural Networks. *ACS Omega* **7**, 12268–12277 (2022).
59. John, P. C. S., 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. *Nature Communications* **11**, 2328 (2020).
60. Avdeef, A. Prediction of aqueous intrinsic solubility of druglike molecules using Random Forest regression trained with Wiki-pS0 database. *ADMET and DMPK* **8**, 29–77 (2020).
61. Panapitiya, G. *et al.* Predicting Aqueous Solubility of Organic Molecules Using Deep Learning Models with Varied Molecular Representations. *arXiv.org* <https://arxiv.org/abs/2105.12638v2> (2021) doi:10.1021/acsomega.2c00642.
62. Zhang, J., Tuguldur, B. & van der Spoel, D. Force Field Benchmark of Organic Liquids. 2. Gibbs Energy of Solvation. *J. Chem. Inf. Model.* **55**, 1192–1201 (2015).
63. Zhang, J., Zhang, H., Wu, T., Wang, Q. & van der Spoel, D. Comparison of Implicit and Explicit Solvent Models for the Calculation of Solvation Free Energy in Organic Solvents. *J. Chem. Theory Comput.* **13**, 1034–1043 (2017).
64. Takano, Y. & Houk, K. N. Benchmarking the Conductor-like Polarizable Continuum Model (CPCM) for Aqueous Solvation Free Energies of Neutral and Ionic Organic Molecules. *J. Chem. Theory Comput.* **1**, 70–77 (2005).
65. Tomasi, J., Mennucci, B. & Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chem. Rev.* **105**, 2999–3094 (2005).

66. Klamt, A., Jonas, V., Bürger, T. & Lohrenz, J. C. W. Refinement and Parametrization of COSMO-RS. *J. Phys. Chem. A* **102**, 5074–5085 (1998).
67. Eckert, F. & Klamt, A. Fast solvent screening via quantum chemistry: COSMO-RS approach. *AIChE Journal* **48**, 369–385 (2002).
68. Lovette, M. A. *et al.* Evaluation of Predictive Solubility Models in Pharmaceutical Process Development—an Enabling Technologies Consortium Collaboration. *Crystal Growth & Design* **22**, 5239–5263 (2022).
69. Nicholls, A. *et al.* Predicting Small-Molecule Solvation Free Energies: An Informal Blind Test for Computational Chemistry. *J. Med. Chem.* **51**, 769–779 (2008).
70. Geballe, M. T., Skillman, A. G., Nicholls, A., Guthrie, J. P. & Taylor, P. J. The SAMPL2 blind prediction challenge: introduction and overview. *J Comput Aided Mol Des* **24**, 259–279 (2010).
71. Kolmar, S. S. & Grulke, C. M. The effect of noise on the predictive limit of QSAR models. *J Cheminform* **13**, 92 (2021).
72. Orbes-Arteaga, M. *et al.* Knowledge distillation for semi-supervised domain adaptation. *arXiv.org* <https://arxiv.org/abs/1908.07355v1> (2019).
73. Xie, Q., Luong, M.-T., Hovy, E. & Le, Q. V. Self-training with Noisy Student improves ImageNet classification. *arXiv:1911.04252 [cs, stat]* <http://arxiv.org/abs/1911.04252> (2020).
74. Sohn, K. *et al.* FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. in *Advances in Neural Information Processing Systems* vol. 33 596–608 (Curran Associates, Inc., 2020).
75. He, J., Gu, J., Shen, J. & Ranzato, M. REVISITING SELF-TRAINING FOR NEURAL SEQUENCE GENERATION. (2020).
76. Wen, M., Blau, S. M., Xie, X., Dwaraknath, S. & Persson, K. A. Improving machine learning performance on small chemical reaction data with unsupervised contrastive pretraining. *Chem. Sci.* **13**, 1446–1458 (2022).
77. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
78. Magar, R. *et al.* AugLiChem: data augmentation library of chemical structures for machine learning. *Mach. Learn.: Sci. Technol.* **3**, 045015 (2022).

79. Kim, Y. *et al.* Physics-informed graph neural networks for predicting cetane number with systematic data quality analysis. *Proceedings of the Combustion Institute* **39**, 4969–4978 (2023).
80. Lundberg, S. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. Preprint at <https://doi.org/10.48550/arXiv.1705.07874> (2017).
81. Wang, H., Lian, D., Zhang, Y., Qin, L. & Lin, X. GoGNN: Graph of Graphs Neural Network for Predicting Structured Entity Interactions. in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence* 1317–1323 (2020). doi:10.24963/ijcai.2020/183.
82. Qin, S. *et al.* Capturing molecular interactions in graph neural networks: a case study in multi-component phase equilibrium. *Digital Discovery* **2**, 138–151 (2023).
83. Labute, P. A widely applicable set of descriptors. *Journal of Molecular Graphics and Modelling* **18**, 464–477 (2000).
84. Chen, X., Li, P., Hruska, E. & Liu, F. Δ -Machine learning for quantum chemistry prediction of solution-phase molecular properties at the ground and excited states. *Phys. Chem. Chem. Phys.* **25**, 13417–13428 (2023).
85. Hruska, E., Gale, A. & Liu, F. Bridging the Experiment-Calculation Divide: Machine Learning Corrections to Redox Potential Calculations in Implicit and Explicit Solvent Models. *J. Chem. Theory Comput.* **18**, 1096–1108 (2022).
86. Wang, H., Tang, Q. & Zheng, W. L1-Norm-Based Common Spatial Patterns. *IEEE Transactions on Biomedical Engineering* **59**, 653–662 (2012).
87. Zhang, L. *et al.* Be Your Own Teacher: Improve the Performance of Convolutional Neural Networks via Self Distillation. Preprint at <https://doi.org/10.48550/arXiv.1905.08094> (2019).
88. Mobahi, H., Farajtabar, M. & Bartlett, P. Self-Distillation Amplifies Regularization in Hilbert Space. in *Advances in Neural Information Processing Systems* vol. 33 3351–3361 (Curran Associates, Inc., 2020).
89. Hammond, G. S. A Correlation of Reaction Rates. *J. Am. Chem. Soc.* **77**, 334–338 (1955).
90. Solvent Effects on the Rates of Homogeneous Chemical Reactions. in *Solvents and Solvent Effects in Organic Chemistry* 165–357 (John Wiley & Sons, Ltd, 2010). doi:10.1002/9783527632220.ch5.

91. Steenackers, B., Neirinckx, A., De Cooman, L., Hermans, I. & De Vos, D. The Strained Sesquiterpene β -Caryophyllene as a Probe for the Solvent-Assisted Epoxidation Mechanism. *ChemPhysChem* **15**, 966–973 (2014).
92. Tshepelevitsh, S., Hernits, K. & Leito, I. Prediction of partition and distribution coefficients in various solvent pairs with COSMO-RS. *J Comput Aided Mol Des* **32**, 711–722 (2018).

CHAPTER 6: Reactive Behavior through Intermolecular Interactions: Predicting Blend Cetane Number

6.1: Chapter Overview

In this chapter, we detail a study of combustion efficiency dictated by intermolecular interactions between components of a biofuel surrogate blend consisting of two or more pure fuel surrogate molecules. This efficiency is numerically characterized by an experimentally derived quantity known as the blend cetane number, a measurement analogous to the octane number used to categorize gasoline fuels. The cetane number (CN) is hence a measurement of fuel reactivity as compared to that of the n-alkane cetane ($C_{16}H_{34}$). A higher CN fuel is more reactive than a lower CN fuel. While on the surface, this may appear to be beneficial, due to the design of a diesel engine, higher reactivity fuels can sometimes be less efficient and produce higher emissions than their low CN counterparts. Diesel engines rely on compression to ignite a premixed fuel-air mixture, in contrast to the spark plugs used for ignition in a gasoline engine. Studies in the literature have shown that lower CN fuels promote increased mixing and homogenization of the fuel-air mixture, mitigating the formation of particulate emissions.¹ Alternative designs using a mix of high and low-CN fuels have also been developed, such as the reactivity controlled compression engine (RCCI)². The ratio of high and low reactivity fuels is critical: too much high reactivity fuel affects the operable range of the RCCI engine, while too much low reactivity fuel affects thermal efficiency.²⁻⁴

Due to the interconnected relationship of CN and engine operating characteristics, knowledge of fuel CN is a critical parameter in engine design. While measurement of CN has recently been standardized through the use of standardized devices like the Cooperative Fuel Research (CFR) Engine, Ignition Quality Tester (IQT), and Advanced Fuel Ignition Delay Analyzer (AFIDA), a significant number of CN values reported in the literature use arbitrarily designed constant-volume combustion chambers similar to bomb calorimeters, injecting bias and variable aleatoric uncertainty into each measurement.

Thus, sophisticated curation strategies must be employed to build datasets suitable for model training. In a previous study, we employed a data weighting strategy to consolidate measurements from disparate experimental datasets published in the literature. A sample weighting approach was employed to preferentially weight experimental measurements conducted with the standardized methods mentioned above during model training; the resultant model achieved state-of-the-art predictive accuracies on CN prediction tasks for single component fuel surrogates.⁵

Due to the growing interest in multi-component fuel design strategies, an approach capable of robust predictions for surrogate fuel blends is necessary. In this work, we expand on our previous model to extend model predictions to multicomponent fuel surrogate blends: we utilize a mixing operator to combine predictions for individual blend components. The resulting model accurately captures deviations from nonlinearity in multicomponent biofuel surrogate blends of up to 12 different pure compound components. We introduce a novel analytic strategy to quantify deviations from nonlinearity, allowing the degree of deviation to be predicted from a single blend prediction.

This collaborative research effort was published as a Society of Automotive Engineers (SAE) Technical Paper. Equal lead contributions were made by myself and Dr. Yeonjoon Kim; additional contributions were made by Dr. Jaeyoung Cho, Dr. Nimal Naser, Wonjong Ko, Dr. Peter C. St. John, Dr. Robert L. McCormick, and Dr. Seonah Kim. I performed database curation and analysis. I also developed the analytical techniques for the neural network model, and helped to write and analyze the manuscript.

6.2: Introduction

As the regulations on the emission from medium-to-heavy duty vehicles get stringent, the combustion strategies in internal combustion (IC) engines get diversified, requiring a high-performance fuel (HPF) tailored to each application. Cetane number (CN) has been considered an important fuel property in designing HPFs as it affects the performance, thermal efficiency, and stability of IC engines. For example, partially premixed compression ignition (PPCI) engines need low-reactivity fuels (low CN) to increase the fuel-air mixing time, which effectively reduces both soot and nitric oxide emission.⁶ On the other hand, reactivity-controlled compression ignition (RCCI) engines require a high CN fuel to ignite the ambient fuel-air mixture with low-reactivity, typically natural gas, ethanol, or ammonia.⁴

In this regard, *a priori* prediction of CN has been of particular interest to both the automotive and refinery industries. One old study from DeFries et. al. correlated the CN with the number of different carbon types identified from the proton nuclear magnetic resonance (¹H NMR) spectroscopy.⁷ The derived correlation successfully predicted the CN of 67 diesel

fuels with a deterministic coefficient (r^2) of 0.992. This finding stimulated further efforts in the CN prediction based on molecular structural features such as the number of double bonds and carbon atoms.⁸ The group additivity model is one of the most comprehensive ways to correlate the molecular structure and the CN, while the contribution of each molecular fingerprint is still under investigation.⁹ A recent paper from our research group also utilized the group additivity for estimating the CN of ether-containing fuels based on their key molecular features (number of primary, secondary, tertiary, and quaternary carbons adjacent to or far from oxygen atoms), showing a remarkable accuracy with mean square error (MAE) of 6.14 on 55 saturated linear, branched, and cyclic ethers.¹⁰

Meanwhile, machine learning (ML) technologies are being recognized as a promising way of predicting CN. Santana et al.¹¹ provided one of the earliest examples of applying ML in CN prediction. Their model was based on the multilayer perceptron with back-static propagation, reproducing the experimental CN of 147 hydrocarbon fuels with a standard error of 8 CN units. The ML model was further exploited to cover the oxygenated fuels several years later by Saldana et al.¹², where the final model combining the feed-forward artificial neural network, support vector machines, and partial least squares regression showed the MAE of 4.7 on the 229 CN data points of hydrocarbons and oxygenates. Follow-up studies from other research groups^{13–16} have increased the size of the training dataset while keeping the prediction accuracy within a reasonable bound (~5 CN units). As an example, our previous study⁵ applied the graph neural networks (GNN) for CN prediction to leverage the successful results from the group additivity model. With a proper evaluation of uncertainty in the database, the GNN

predicted the CN of single compound fuels with an MAE of 2.4 CN units on the biggest dataset ever (630 single compound fuels), showing the competence of GNN in CN prediction.

Despite the recent progress in the CN prediction for single compound fuels, those for multicomponent fuels are still a challenging problem. Most prediction models for multicomponent fuels require the fuels' physical properties as input variables,¹⁷ which is not viable in the *de novo* design of HPFs. Otherwise, the CN of multicomponent fuels is estimated as the linear combination of each component's CN, so-called the linear mixing rule.^{18,19} However, the linear mixing rule of CN can sometimes fail due to the complicated synergistic effect between individual components in the fuel. The recent study from Fioroni et al.²⁰ suggested that the synergistic effect for research octane number (RON; inversely correlated with CN²¹) originated from the OH radical scavenging of low-reactivity components. Their simulation results indicate that the fuel that forms a resonance stabilized radical through OH addition reaction is the most effective in inducing the synergistic effect in RON of multicomponent fuels, which suggests the strong dependency of the blending nonlinearity on the structural features of individual components.

As an effort to resolve this complicated nature of multicomponent fuels' CN prediction, a previous study by Basu et al.²² utilized an artificial neural network for predicting the CN of 60 multicomponent diesel fuels based on their structural parameters observed from ¹H NMR. They divided the spectra from ¹H NMR into 17 regions to represent the structural features from different molecular groups (*n*-paraffins, *iso*-paraffins, olefins, cycloalkanes, and aromatics), based on which the ANN model shows reasonable accuracy

for both training and validation set ($r^2 = 0.91$ and 0.85 , respectively). Recently, Dalmiya et al.²³ applied convolutional neural networks (CNNs) in predicting the multicomponent fuel's CN. They measured the infrared spectra of multicomponent fuels on $2400 - 800 \text{ cm}^{-1}$ to characterize the molecular fingerprints of individual components. The measured spectra were directly used to train the CNN model and predict the CN of multicomponent fuel, which showed remarkable accuracy in reproducing the experimental CN of 88 single compounds and multicomponent fuel (mostly jet fuels) with R^2 over 0.93 . These two studies suggest the machine learning model is a promising way to connect the molecular fingerprints of individual components to the CN of multicomponent fuels.

In this study, we suggest a novel structure using GNN to predict the CN of multicomponent fuels. To do so, we assemble the largest database of experimental CN measurements to date. Then, using a linear approximation of nonlinear blending behavior with a learned weight matrix, our novel model can accurately predict the CN of fuel blends containing up to 12 distinct chemical species. Here, we present the structure of the model and analyze its performance on a subset of the compiled dataset. We also present a detailed analysis of the influence of the weight matrix in approximating nonlinear blending behavior. It is our hope that the model will serve as a valuable tool for fuel scientists to formulate surrogate blends with specific CNs and thus, ignition properties.

6.3: Results and Discussion

6.3.1: Nonlinear Blending Behavior

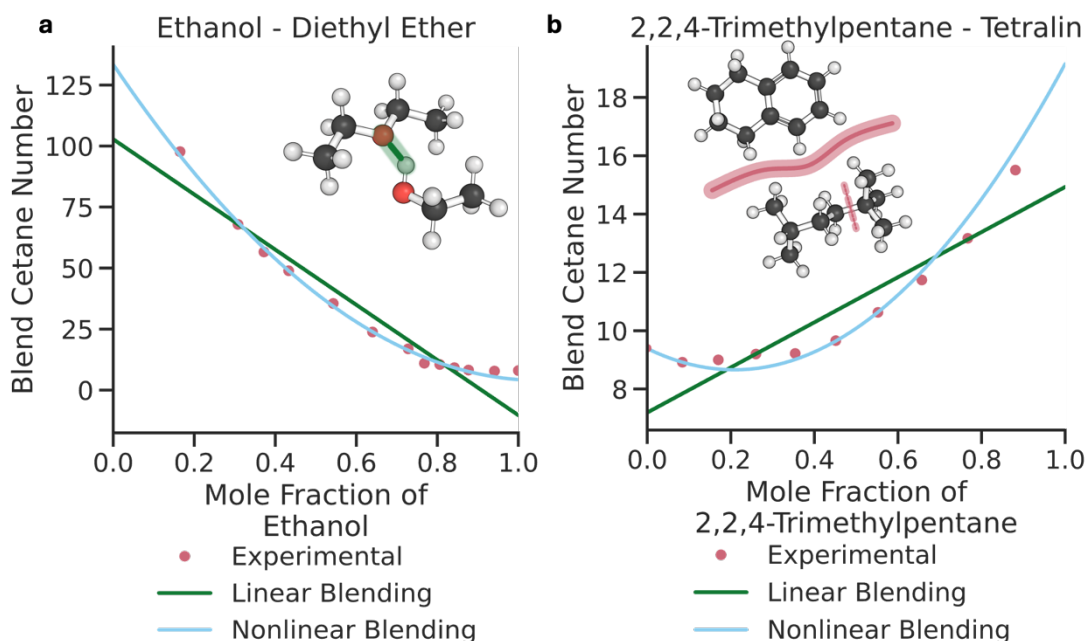


Figure 6.1: a) Nonlinear antagonistic blending behavior of ethanol-diethyl ether fuel blends. The experimentally observed cetane numbers can be reproduced by using a nonlinear approximation for the blending behavior. observed nonlinearities may be due in part to hydrogen bonding interactions between fuel components. b) Nonlinear antagonistic blending behavior of 2,2,4-trimethyl pentane (iso-octane) fuel blends. Dispersive Van der Waals interactions and kinetic factors may play a part in the observed experimental nonlinearity.

Fuel blends containing chemically diverse constituents often show nonlinear blending behavior, i.e., the CN of a fuel blend need not be a simple linear combination of the CN of its constituents. Since this behavior may be caused by a variety of factors, the interactions are difficult to model in a generalizable manner with a single, closed form expression. For example, as the concentration of ethanol is increased in a fuel blend consisting of ethanol and diethyl ether, the CN exhibits a non-linear decrease (**Figure 6.1a**), in contrast to a linear fit of the experimental data. It should be noted that the linear fit referenced does not account for prior knowledge of pure compound cetane numbers, and instead seeks to fit the full range of blending behaviors across constituent mole fractions. A

recent theoretical study has shown that the stability of the transition state involved in hydrogen abstraction and radical formation from fuel molecules is enhanced by hydrogen bonding interactions, thus enhancing the rate of reaction and increasing reactivity/CN.²⁴

On the contrary, a mixture of isooctane and tetralin exhibits first decreasing, then increasing cetane number as the isooctane concentration is increased (**Figure 6.1b**). This divergent behavior can be explained by the effects of dispersive Van der Waals interactions between the two compounds; in addition, the cleavage of a carbon-carbon bond can result in the formation of a highly stable tertiary butyl radical per inference using the ALFABET BDE prediction tool²⁵, accelerating the rate of downstream combustion reactions.

In addition, nonlinear blending behavior can be caused by the influence of downstream oxygenated intermediates. Recent work has shown that several competing pathways involving reactions of fuel species with the highly reactive peroxy radicals can lead to branching and chain propagation reactions, with the net effect of increasing reactivity at lower temperature regimes. These reactions involve the formation of a ketohydroperoxide intermediate and a hydroxyl radical.²⁶ On the contrary, at higher temperature regimes, oxygen addition is inhibited, and β -scission reactions become more prominent, leading to the observance of negative temperature coefficient behavior.²⁶ This complex variation of observed properties thus induces nonlinear blending behavior.

Thus, the problem we are faced with is twofold: any predictive model for blend CNs must incorporate both intramolecular information about the chemical kinetics, and information about intermolecular interactions. Here, we treat the intramolecular chemical

kinetics using the GNN embedding; previous work has shown that the molecular embedding utilized here is capable of learning the bond dissociation enthalpy of any bond in a molecule.²⁵ Since the bond dissociation enthalpy represents an energetic barrier to the slow unimolecular hydrogen abstraction step that constitutes the initial reaction of a fuel molecule during a combustion process, the neural network input embedding adequately captures the kinetics of the reaction. Details about our GNN architecture are explained in the following sections.

6.3.2: Formulation For Modeling Nonlinear Blending Behavior

As outlined in **Figure 6.1**, some mixtures exhibit non-linear relationships between CNs and mole fractions of constituent molecules, and the predictive model should capture

- **Vapor pressure of mixtures:**

$$P_{mixture} = \sum_i (\chi_i \gamma_i P_{pure,i})$$

- **Likewise, cetane number of mixtures:**

$$CN_{mixture} = \sum_i (\chi_i \gamma_i CN_{pure,i}),$$

where $\gamma_i CN_{pure,i} = CN_i = \sum_j (\chi_j w_{ij})$,

(χ_i : Mole fraction, γ_i : Activity coefficient at infinite dilution)

cf. Linear regression ($\gamma_i = 1$)

$$CN_{mixture} = \sum_i (\chi_i CN_{pure,i}),$$

- **Example: Ethanol – diethyl ether mixture**

$$W = \begin{bmatrix} 5.8 & 8.4 \\ 6.7 & 136.9 \end{bmatrix} \begin{array}{l} - CN_{pure,i}: \text{ethanol, diethyl ether} \\ - \text{Infinite dilution CN of:} \\ \text{Ethanol in diethyl ether} \\ \text{Diethyl ether in ethanol} \end{array}$$

* $i = 1$: Ethanol, $i = 2$: Diethyl ether,
 $\chi_1 = 0.64$, $\chi_2 = 0.36$, $CN_{mixture} = ?$

$$CN_1 = 5.8 \cdot 0.64 + 8.4 \cdot 0.36 = 6.71$$

$$CN_2 = 6.7 \cdot 0.64 + 136.9 \cdot 0.36 = 53.59$$

 $CN_{mixture} = 6.71 \cdot 0.64 + 53.59 \cdot 0.36 = 23.59$

(cf. Exp.: 23.8, Linear regression: 30.4)

Figure 6.2: Description of the assumptions and formulations for modeling non-linear blending behavior.

these trends. To this end, the governing equation regarding the description of mixtures was established and used in the ML model. Such formulation should be consistent with underlying physics and thermodynamics regarding mixtures. Of note, the deviations from the linearity are analogous to the non-ideal mixing behavior of vapor pressures of liquid mixtures. As shown in **Figure 6.2**, the extended Raoult's law formulates mixture's vapor pressure ($P_{mixture}$) as a function of mole fraction (c_i), activity coefficient at infinite dilution (g_i), and vapor pressure ($P_{pure,i}$) of each component.^{27–29} The activity coefficient g_i quantifies the deviation from the linearity; if it is 1, the equation corresponds to a simple linear model for ideal liquid solutions. Otherwise, it describes non-linearity.

Here, we hypothesize that the similar equation can also be applied to the CN of mixtures ($CN_{mixture}$), and accurate prediction of g_i leads to better description of the non-linear mixing behavior. $CN_{mixture}$ can be predicted by calculating the summation of $c_i g_i CN_{pure,i}$ values of each component (**Figure 6.2**). Inspired by the Wilson equation,³⁰ $g_i CN_{pure,i}$ term can be equated to the summed value of $c_j w_{ij}$, corresponding to a CN contribution of component i in a given mixture (Abbreviated as CN_i). Each row of the **W** matrix accounts for intermolecular interactions when one component i is infinitely diluted with respect to others (j). One example of **W** matrix is shown in **Figure 6.2**, for the ethanol – diethyl ether mixture. The diagonal elements ($w_{11}=5.8$, $w_{22}=136.9$) indicate CNs of pure ethanol and diethyl ether determined by the regression. The infinite dilution of ethanol in diethyl ether, and vice versa, is considered by the off-diagonal elements ($w_{12}=8.4$, $w_{21}=6.7$). Then, CN contributions of each compound (CN_1 and CN_2) are evaluated by the linear combination of each element in

a row multiplied by a mole fraction. Lastly, CN_{mixture} is calculated by a mole-fraction-weighted sum of CN_1 and CN_2 .

It should be noted that **W** can be asymmetric due to different infinite dilution effects from each component, which is useful for modeling the non-linear blending behavior. The formulation with the **W** matrix well reproduces the experimental CN of the mixture with 64% ethanol and 36% diethyl ether (23.59 and 23.8 from the regression and experiment, respectively). This estimation is more accurate than the linear regression which resulted in the CN of 30.4, ultimately leading to a better description of non-linear blending behavior than linear regression (**Figure 6.2**). Previous studies also adopted the similar approaches; they developed the reliable composition-based predictive models for octane and cetane numbers based on the analogy of CNs of blends with activity coefficients.^{18,31} A root-mean-square error of 1.25 was achieved for CNs of 203 diesel fuel mixtures.³² Here, we pursued the development of predictive models with broad applicability by using more extensive database consisting of 1,143 data points and modeling non-linearity as a function of molecular structures as well as mixture's composition.

6.3.3: Baseline Analysis for the Predictive Model Development

Before developing the actual ML model, baseline model analysis was performed for the model's feasibility. Specifically, we validated our hypothesis regarding the formulation shown in **Figure 6.2** for modeling non-linearity. The whole database consisting of 1,143 was used for the linear regression and the regression with **W** matrix, based on the equations shown in **Figure 6.2**. Adopting the **W** matrix showed a lower mean absolute error (MAE) of 0.74 compared to the linear regression (MAE: 1.86, **Figure 6.3**).

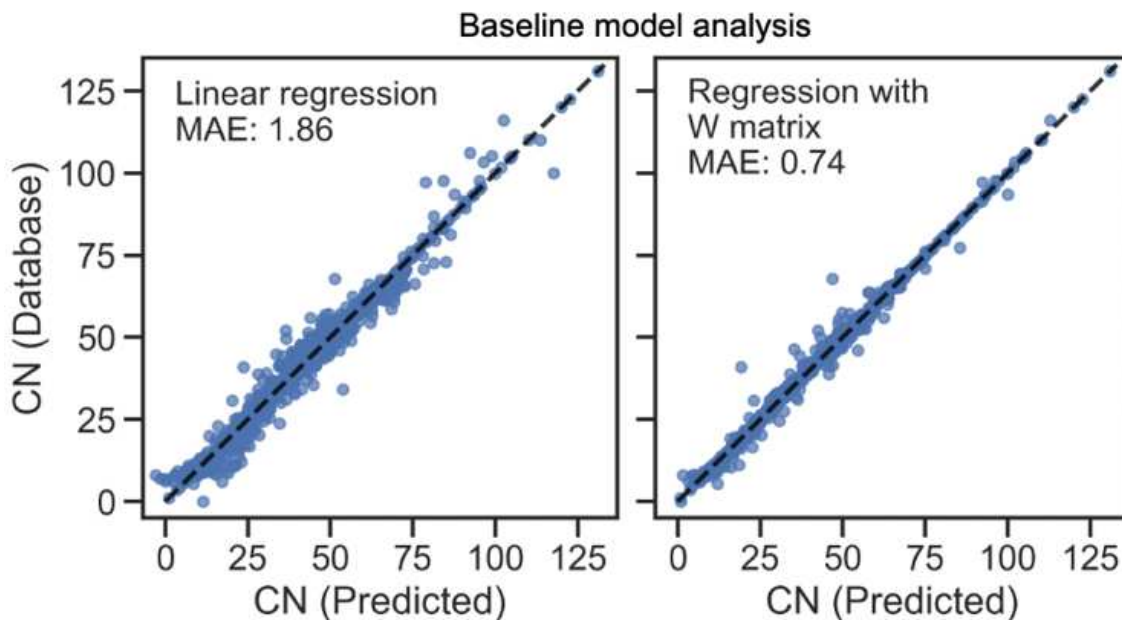


Figure 6.3: Results from the analysis of two baseline models: (1) Linear regression of mixtures' CNs in terms of mole fractions and CNs of each component. (2) Regression with using the $\mathbf{W}$ matrix.

The higher accuracy obtained from the $\mathbf{W}$ matrix demonstrates the feasibility of predicting non-linear blending behavior by considering infinite dilution effects, in analogy to the equations for activity coefficients. However, the size of $\mathbf{W}$ matrix grows proportionally to the number of components (N) in a mixture; it can be up to 12x12 within our database, and potentially even larger for mixtures with more than 12 components.

For efficient calculations of $\mathbf{W}$, we adopted a low-rank approximation scheme of $\mathbf{W}$ (Figure 6.4). The ($N \times N$) matrix $\mathbf{W}$ is obtained through a dot product of two matrices $\mathbf{U}^T$ and $\mathbf{V}$ whose sizes are ($N \times r$) and ($r \times N$) respectively. The rank, r , can be adjusted so that the number of calculated matrix elements is minimized while maintaining the accuracy of $\mathbf{W}$. The direct evaluation of $\mathbf{W}$ requires N^2 calculations, whereas low-rank approximations can reduce this number to $2rN$. To find an optimal number of r , we repeated the regression with $\mathbf{U}$ and $\mathbf{V}$ matrices with varying r . Figure 6.4 illustrates the MAEs from the regressions at different r . $r=1$ showed a MAE of 2.54, which is worse than that from linear regression, but a

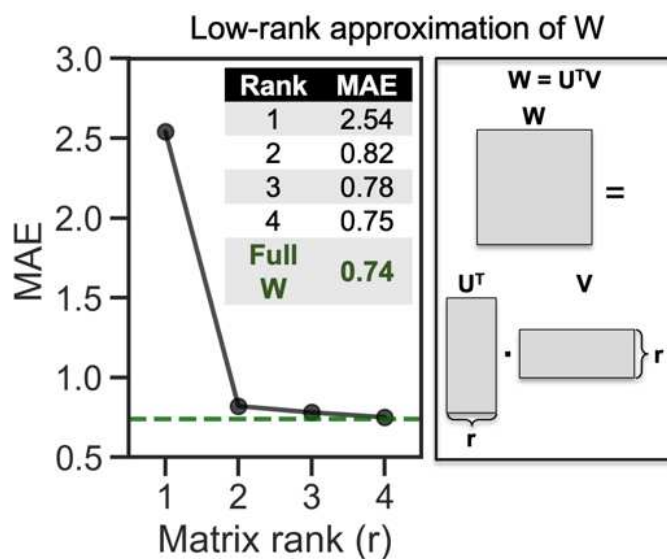


Figure 6.4: The low-rank approximation of W and accuracies from the regression with varying the rank r .

MAE of 0.82 was achieved at $r=2$. In other words, the rank of 2 is sufficient to obtain the accuracy comparable to that from the full W matrix (MAE: 0.74). Increasing r to 3 and 4 slightly improved the accuracy with the MAEs of 0.78 and 0.75, respectively, but at the cost of computational expense. Therefore, it can be concluded that $r=2$ is an optimal rank value that can achieve the best compromise between accuracy and computational costs.

The insights gained from the above baseline analysis were utilized in building graph neural networks (GNNs) for predicting CNs of mixtures with incorporating the W matrix. The GNNs are for embedding molecular structures of each component and considering ‘intramolecular’ structural effects on CN. The subsequent part of the model accounts for the ‘intermolecular’ influences on CN by constructing W with the $r=2$ dimensionality reduction. Details are discussed in the next section.

6.3.4: Graph Neural Networks and Molecular Representations

Figure 6.5 illustrates the architecture of our GNN. The first part consists of N parallel message-passing neural network (MPNN) blocks for encoding individual component molecules in a mixture and accounting for their chemical structural effects on CN. Of note, the model was trained to predict CNs of mixtures consisting of up to 12 components, as 12 is the maximum number of components in our curated database. However, in principle, the model architecture can be readily modified and expanded to any numbers of components. Two-dimensional molecular structures are given as inputs, with their atom and bond features. The model also inputs two global features (The number of hydrogen bond donors and acceptors: N_{HBD} , N_{HBA}). Such features influence fuels' physicochemical properties pertinent to CN, such as heat of vaporization, boiling point, reactivity, enthalpy of mixing, and non-linear mixing behavior of CNs (**Figure 6.1**).^{5,33,34}

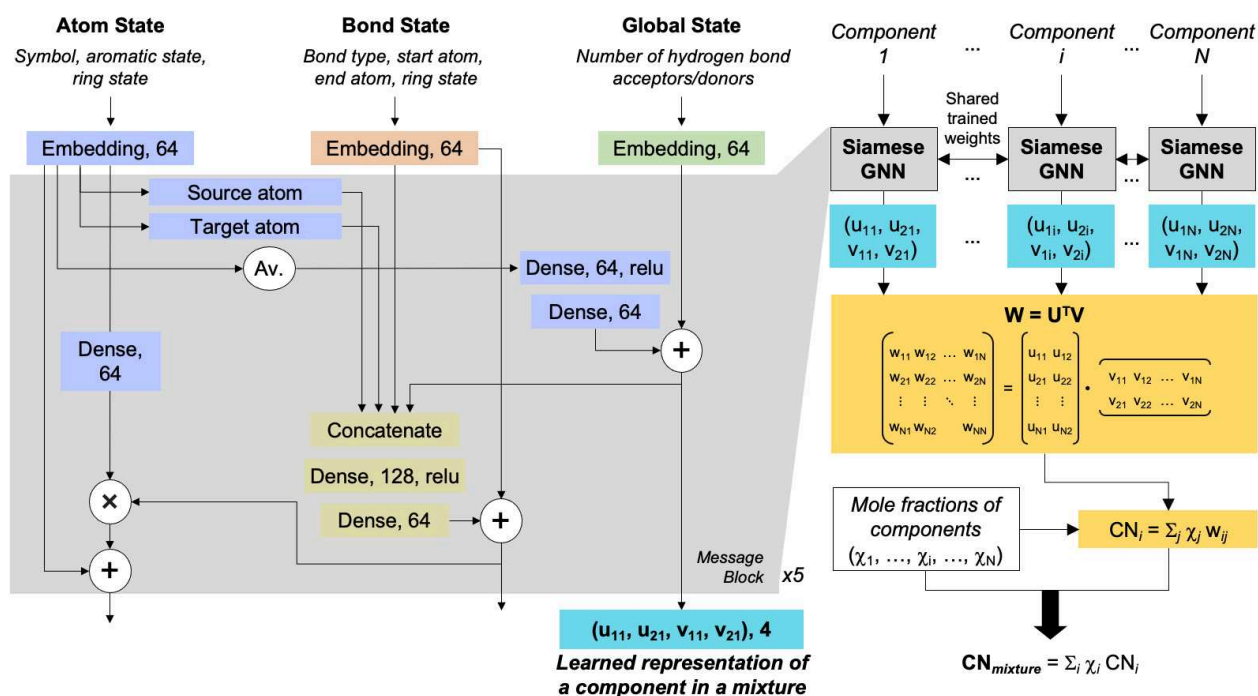


Figure 6.5: Architecture of our graph neural networks for predicting cetane numbers of mixtures

All these features are embedded as vectors and undergo message layers in which mathematical operations among the vectors are carried out. Passing through multiple layers enables the model to learn local and non-local environments of atoms and bonds, and thus, their impacts on CN. All the MPNN blocks share the trained weights, akin to Siamese neural networks, so that the whole GNN is trained consistently with respect to various compounds in the database. We successfully applied the similar MPNN architecture to predicting CNs of pure fuel compounds including 630 hydrocarbons and oxygenates, with an overall MAE of 5.34 from the 10-fold cross-validation.⁵ While the model for pure compounds has only one MPNN block and outputs the CN, here, each of MPNN blocks outputs a four-dimensional vector for component i ($u_{1i}, u_{2i}, v_{1i}, v_{2i}$, **Figure 6.5**) that is used for modeling multicomponent interactions. These vectors from the GNN blocks correspond to learned representations of a component in a mixture.

The dimension of four was chosen because the rank of 2 was found to be optimal in the baseline model analysis (**Figure 6.4**) for the low-rank approximation of $\mathbf{W}$ by multiplying two matrices ($\mathbf{U}$ and $\mathbf{V}$). The former and latter two elements comprise $\mathbf{U}$ and $\mathbf{V}$, respectively: (u_{1i}, u_{2i}) and (v_{1i}, v_{2i}) for component i . These four-dimensional vectors are collected from all MPNN blocks and then concatenated, resulting in two ($N \times 2$) matrices for rank-2 approximation: $\mathbf{U}=[(u_{11}, u_{21}), (u_{12}, u_{22}), \dots, (u_{1N}, u_{2N})]$ and $\mathbf{V}=[(v_{11}, v_{21}), (v_{12}, v_{22}), \dots, (v_{1N}, v_{2N})]$, where N is the number of components. Next, $\mathbf{U}$ and $\mathbf{V}$ are multiplied to obtain $\mathbf{W}=\mathbf{U}^T\mathbf{V}$. The elements of $\mathbf{W}$ are trained and adjusted so that it can account for the infinite dilution effects when one component is mixed with the other, and vice versa.

In summary, a multi-partite graph neural network was constructed, with parallel GNNs producing independent representations of each blend component molecule. The **W** matrix associated with rank-reduced approximations is then learned in conjunction with the molecular representations to predicted CNs of mixtures.

6.3.5: Model Training and Evaluation of Model Accuracy

The GNN depicted in **Figure 6.5** was then trained, and its accuracy was evaluated. The database was split into five parts for the 5-fold cross-validation (**Figure 6.6a**). Each fold consists of 216, 205, 223, 233, and 266 data points, respectively. The numbers of data points are not identical among the folds, since each mixture has different number of data points

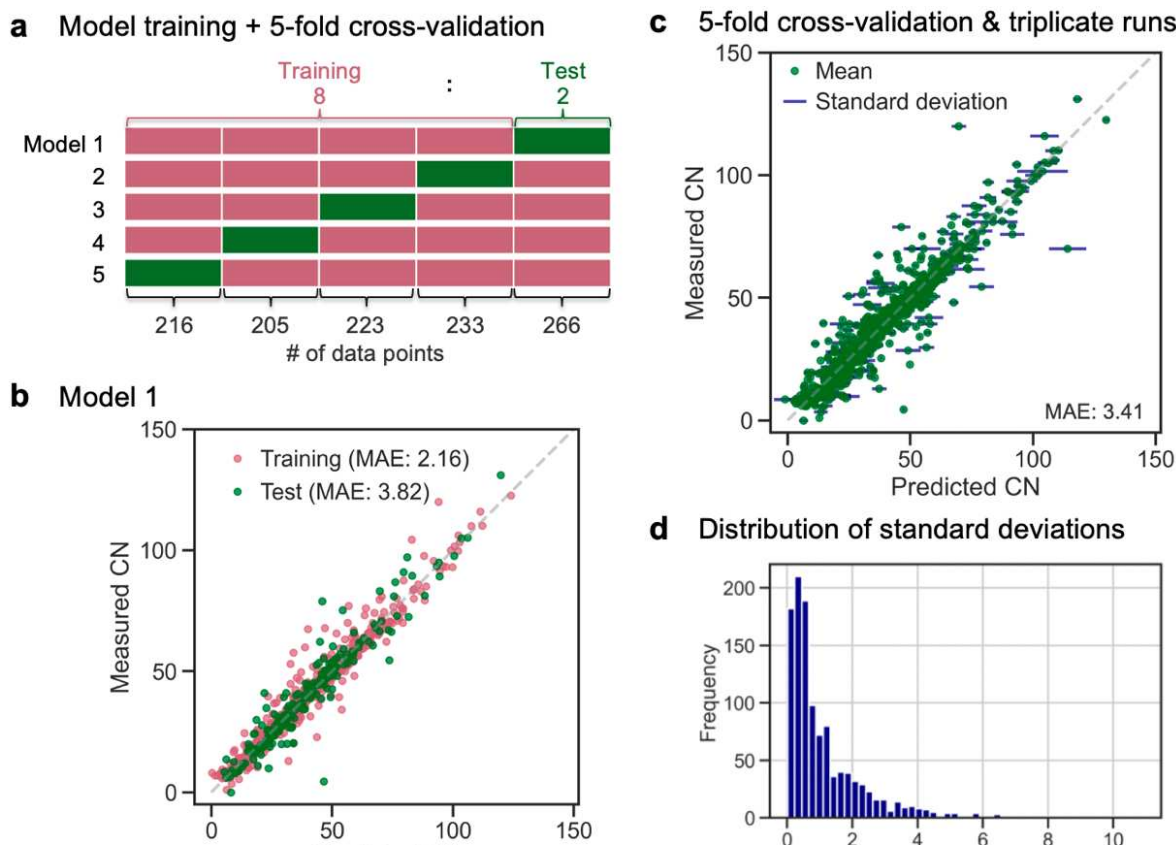


Figure 6.6: (a) Data splits for the 5-fold cross-validation, with the number of data points for each chunk of data. (b) A parity plot depicting the accuracy of the training and test sets of Model 1. (c) A parity plot of test sets from three repetitions of the 5-fold cross-validation, with the standard deviations of three predicted CN values per each data point. (d) A histogram showing the distribution of standard deviations of three predictions for the whole dataset (1,143 data points).

with varying mole fractions. For example, in **Figure 6.1a**, there are 13 data points for the ethanol – diethyl ether mixture, whereas for the iso-octane – tetralin mixture (**Figure 6.1b**), 10 different compositions exist in the database. The five-fold data split was carried out multiple times with changing random seeds, and the random split was chosen so that the deviation of data points among the folds is minimized. Then, five models were trained with changing training and test sets (Model 1 – 5, **Figure 6.6a**). The training:test set ratio was kept approximately 80:20.

Figure 6.6b shows the parity plot for Model 1. The MAEs for the training and test set are 2.16 and 3.82, respectively. It should be emphasized that this accuracy is comparable to or higher than previous studies,^{5,9,12–15,22,23,32,35} while the largest database was used (1,143 data points). Moreover, these MAEs are like the uncertainty of experimental CN measurements using standard devices such as CFR, IQT, and AFIDA. The typical reproducibility ranges of these devices are 2.8-4.8, 1.6-2.6, and 1.5-2.3 CN units, respectively, for CNs of diesel fuels ranging from 40 to 56.^{5,36–38} Model 1 was used for the subsequent analysis regarding how the model describes non-linear mixing behavior (*vide infra*). Furthermore, the 5-fold cross-validation was repeated for three times, and the accuracy of test sets from all folds were assessed, to confirm that the model is not susceptible to overfitting.

Figure 6.6c illustrates the results from the 5-fold cross-validation with triplicate runs, resulting in a total of 15 separate validation runs. Predicted CNs were obtained by averaging three CNs, and the standard deviation from three predictions were also plotted. Of note, in

Figure 6.6c, there are no data points from the training set, and the whole database is considered as the test set through the 5-fold cross-validation. The MAE is 3.41, showing no significant increase in the error when only test sets are used for the model evaluation. We also analyzed the standard deviations of predicted CN values from triplicate runs (**Figure 6.6d**); Most of the standard deviations (92.5 %) do not exceed 3.0, which is also comparable to the experimental uncertainties discussed above.

6.3.6: Interpretation of the W Matrix Trained for Modeling Nonlinear Blending Behavior

Next, we use a per-blend additive approximation to linear mixing behavior via the use of a learned weight matrix to model the intermolecular interactions. Approximation of nonlinear behavior with linear transformations has previously been well established in the fields of dynamical systems and process modeling. For instance, von Neumann and Koopmans' formulation of the Koopman operator for the time evolution of classical systems has been rigorously studied.^{39,40} While it is possible to construct an infinite dimensional Hilbert space operator that accurately describes the dynamics of the system, computer systems are incapable of treating such infinite-dimensional operations. Thus, the task becomes one of finding a high-fidelity, finite dimensional operator that approximates the linear, infinite dimensional operator, which in turn approximates a nonlinear operator.

Here, we are faced with a similar goal: the approximation of an accurate finite-dimensional, low rank linear transformation that describes the nonlinear blending behavior observed when different fuel compounds are mixed. Unlike the Koopman operator problem, however, we are not faced with the issue of resolving the behavior of our system along an

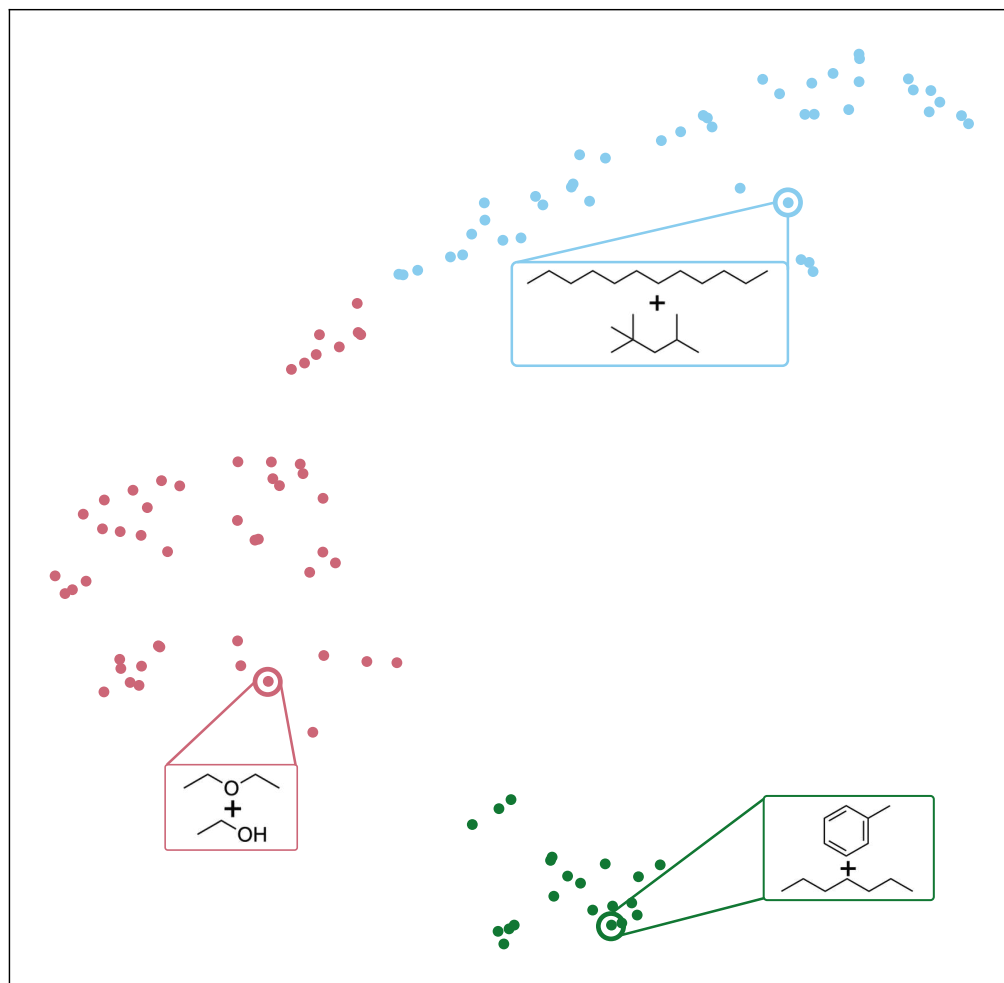


Figure 6.7: K-means clustering followed by 2D UMAP embedding of weight matrices for test set fuel blends. Clustering utilized 3 disparate clusters in the 12D latent space, corresponding to linear, synergistic, and antagonistic blending behavior respectively. The number of clusters used by the k-means algorithm was validated by silhouette score analysis (Appendix E, Figure S2).

infinite number of dynamical modes; instead, we only need to treat mixing between blend components, thus constraining our linear operator (written in the form of the square weight matrix **W**) to the number of blend components, 12. To further reduce the complexity of the interaction behavior, and to aid in interpretation, the weight matrix was further subjected to a low-rank approximation; it was found that using a four-rank matrix for CN prediction yielded a mean absolute error of only 0.01 CN more than the full-rank matrix (**Figure 6.4**).

Since the weight matrices represent a concentration-independent means of characterizing intermolecular interactions, they were then subjected to further analysis, to further investigate the non-linear nature of the blending interaction. Clustering and dimensionality reduction was performed using k-means clustering followed by Uniform Manifold Approximation and Projection (UMAP) embedding⁴¹ on the full rank learned weight matrices. It is observed that three distinct clusters are found (**Figure 6.7**); broadly, these correspond to blends that exhibit synergistic (cetane number higher than predicted by a linear model), antagonistic (cetane number lower than predicted by a linear model), and linear additive behavior. Since similar matrices are clustered together in the 12-dimensional

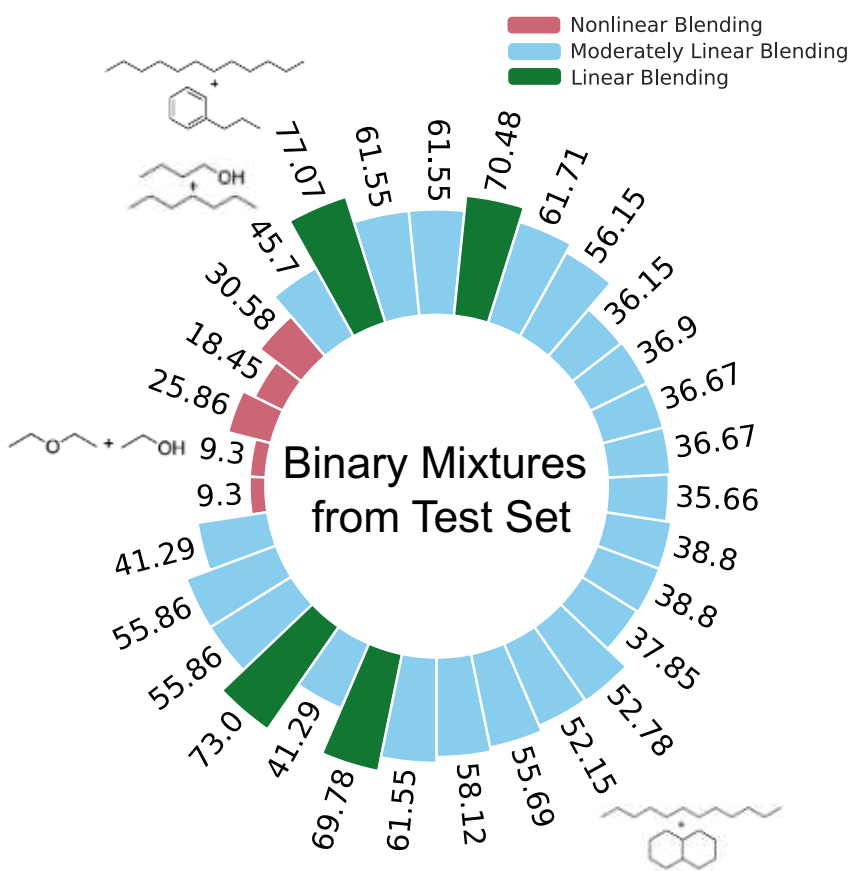


Figure 6.8: Scaled symmetry scores for binary mixtures in the test set fuel blends. Calculated symmetry scores were scaled to 100 for ease of interpretation.

latent space, the model can identify similar types of blending interactions and construct weight matrices accordingly.

This behavior was further analyzed by quantifying the degree of asymmetry in the learned weight matrices. As previously stated, the asymmetry in the weight matrices arises from infinite-dilution effects for each component – thus, the more asymmetric the weight matrix, the higher the deviation from linear blending behavior. Here, we study the degree of asymmetry for binary mixtures in the test set blends (binary mixtures were chosen to aid in interpretation of results). The degree of asymmetry is quantified by decomposing the matrix into purely symmetric and antisymmetric parts, and comparing the magnitude of each utilizing a Frobenius norm (written as $\|\bullet\|_{Frob}$) based symmetry score:

$$W_{sym} = \frac{1}{2}(W + W^T) \quad 6.1$$

$$W_{antisym} = \frac{1}{2}(W - W^T) \quad 6.2$$

$$Symmetry\ Score = \frac{\|W_{sym}\|_{Frob} - \|W_{antisym}\|_{Frob}}{\|W_{sym}\|_{Frob} + \|W_{antisym}\|_{Frob}} \quad 6.3$$

Blends with higher symmetry scores show linear mixing behavior, while those with lower symmetry scores show nonlinearity (**Figure 6.8**). For ease of interpretation, the symmetry scores were divided into three equally spaced broad classes: nonlinear blending

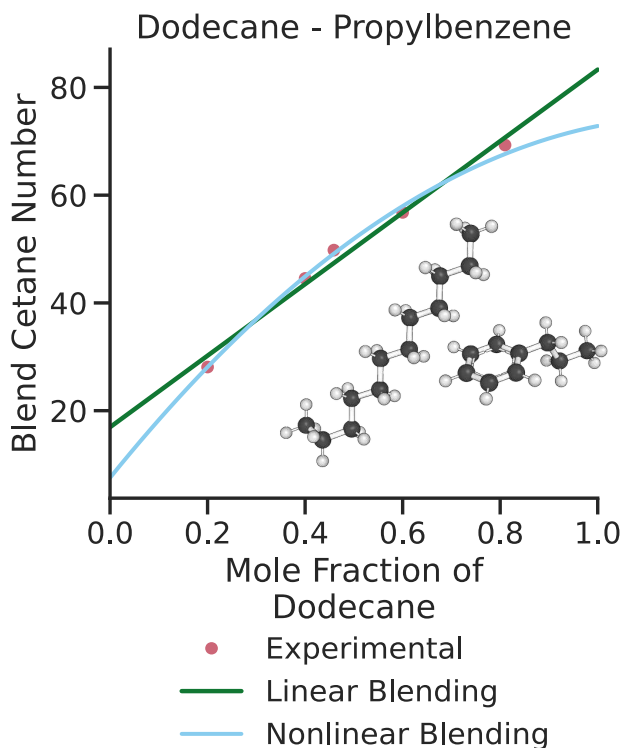


Figure 6.9: Linear blending behavior of the dodecane-propylbenzene fuel blend. By quantifying the asymmetry present in the learned weight matrix for this blend, this linear behavior can be predicted a priori.

(scaled symmetry score of 0-33), moderately linear blending (scaled symmetry score of 34-67), and linear blending behavior (scaled symmetry score of 67-100). As expected, hydrocarbon mixtures show greater symmetry, indicating tendencies toward linear blending behavior, while those mixtures containing oxygenated compounds show greater asymmetry and thus more nonlinear mixing behavior via intermolecular interactions. For example, for the linear blending behavior mixture of dodecane and propylbenzene shown in **Figure 6.9** obtained by varying the mole fraction of dodecane from 0.0 (pure propylbenzene) to 1.0 (pure dodecane), the calculated CN using the weight matrix based linear approximation is 46.6; the model predicted cetane number is 46.6, while the experimentally derived measurement is 44.5. The above analysis demonstrates that the governing equations for

describing mixtures (**Figure 6.2**) are valid and the model can distinguish the fuel blends that show linear or non-linear blending behavior.

6.4: Conclusions

We developed the GNNs that predict CN of multicomponent mixtures that can be fuel surrogates. Rapid predictions are possible as the model uses only the atom, bond, and molecular features generated from 2D structures. We achieved the high accuracy comparable to experimental uncertainties, with MAEs ranging from 3.4 to 3.8 from the 5-fold cross-validation. The standard deviation of predicted values did not exceed 3.0 for most cases when the model was trained three times. The model considers the effects of intermolecular interactions on CNs by adopting the formulation in analogy to the equations for activity coefficients in mixtures, leading to the prediction consistent with underlying physics and thermodynamics of mixtures. To further verify feasibility of the model, we found the relationship between the fuel's non-linear blending behavior and the values of trained weight matrix (**W**) in the hidden layer.

The GNNs for CN of mixtures manifest a successful expansion of our previous GNN for single compounds, by modeling both the influence of chemical structures and the infinite dilution of each molecule on CN. The model reliably reproduced the non-linear synergistic/antagonistic blending behavior of mixtures with respect to varying mole fractions of constituent molecules. Our approach will aid the design of new fuels with a high CN and can also be combined with predictive models for other fuel properties such as yield sooting index. Such integration of predictive models would enable the discovery of low-

emission and high-performance fuels, responding to the recently increasing demands on mitigating environmental impacts of vehicles while maintaining high engine performance.

6.5: References

1. Cho, K., Han, M., Sluder, C. S., Wagner, R. M. & Lilik, G. K. *Experimental Investigation of the Effects of Fuel Characteristics on High Efficiency Clean Combustion in a Light-Duty Diesel Engine*. <https://saemobilus.sae.org/papers/experimental-investigation-effects-fuel-characteristics-high-efficiency-clean-combustion-a-light-duty-diesel-engine-2009-01-2669> (2009) doi:10.4271/2009-01-2669.
2. Splitter, D., Hanson, R., Kokjohn, S. & Reitz, R. D. *Reactivity Controlled Compression Ignition (RCCI) Heavy-Duty Engine Operation at Mid-and High-Loads with Conventional and Alternative Fuels*. <https://saemobilus.sae.org/papers/reactivity-controlled-compression-ignition-rcci-heavy-duty-engine-operation-mid-high-loads-conventional-alternative-fuels-2011-01-0363> (2011) doi:10.4271/2011-01-0363.
3. Ickes, A., Wallner, T., Zhang, Y. & Ojeda, W. D. Impact of Cetane Number on Combustion of a Gasoline-Diesel Dual-Fuel Heavy-Duty Multi-Cylinder Engine. *SAE International Journal of Engines* **7**, 860–872 (2014).
4. Reitz, R. D. & Duraisamy, G. Review of high efficiency and clean reactivity controlled compression ignition (RCCI) combustion in internal combustion engines. *Progress in Energy and Combustion Science* **46**, 12–71 (2015).
5. Kim, Y. *et al.* Physics-informed graph neural networks for predicting cetane number with systematic data quality analysis. *Proceedings of the Combustion Institute* **39**, 4969–4978 (2023).
6. Naser, N. *et al.* On the effects of fuel properties and injection timing in partially premixed compression ignition of low octane fuels. *Fuel* **207**, 373–388 (2017).
7. DeFries, T. H., Kastrup, R. V. & Indritz, D. Prediction of cetane number by group additivity and carbon-13 Nuclear Magnetic Resonance. *Ind. Eng. Chem. Res.* **26**, 188–193 (1987).
8. Lapuerta, M., Rodríguez-Fernández, J. & de Mora, E. F. Correlation for the estimation of the cetane number of biodiesel fuels and implications on the iodine number. *Energy Policy* **37**, 4337–4344 (2009).

9. Dahmen, M. & Marquardt, W. A Novel Group Contribution Method for the Prediction of the Derived Cetane Number of Oxygenated Hydrocarbons. *Energy Fuels* **29**, 5781–5801 (2015).
10. Cho, J. *et al.* Bioderived ether design for low soot emission and high reactivity transport fuels. *Sustainable Energy Fuels* **6**, 3975–3988 (2022).
11. Santana, R. C. *et al.* Evaluation of different reaction strategies for the improvement of cetane number in diesel fuels. *Fuel* **85**, 643–656 (2006).
12. Saldana, D. A. *et al.* Flash Point and Cetane Number Predictions for Fuel Compounds Using Quantitative Structure Property Relationship (QSPR) Methods. *Energy Fuels* **25**, 3900–3908 (2011).
13. Guo, Z., Lim, K. H., Chen, M., Thio, B. J. R. & Loo, B. L. W. Predicting cetane numbers of hydrocarbons and oxygenates from highly accessible descriptors by using artificial neural networks. *Fuel* **207**, 344–351 (2017).
14. Kessler, T., Sacia, E. R., Bell, A. T. & Mack, J. H. Artificial neural network based predictions of cetane number for furanic biofuel additives. *Fuel* **206**, 171–179 (2017).
15. Li, R., Herreros, J. M., Tsolakis, A. & Yang, W. Machine learning-quantitative structure property relationship (ML-QSPR) method for fuel physicochemical properties prediction of multiple fuel types. *Fuel* **304**, 121437 (2021).
16. Schweidtmann, A. M. *et al.* Graph Neural Networks for Prediction of Fuel Ignition Quality. *Energy Fuels* **34**, 11395–11407 (2020).
17. Ladommatos, N. & Goacher, J. Equations for predicting the cetane number of diesel fuels from their physical properties. *Fuel* **74**, 1083–1093 (1995).
18. Ghosh, P. & Jaffe, S. B. Detailed Composition-Based Model for Predicting the Cetane Number of Diesel Fuels. *Ind. Eng. Chem. Res.* **45**, 346–351 (2006).
19. Tong, D., hu, C., Jiang, K. & Li, Y. Cetane Number Prediction of Biodiesel from the Composition of the Fatty Acid Methyl Esters. *Journal of Oil & Fat Industries* **88**, 415–423 (2011).
20. Fioroni, G. M. *et al.* Chemical kinetic basis of synergistic blending for research octane number. *Fuel* **307**, 121865 (2022).
21. Octane-Cetane Relationship. <https://apps.dtic.mil/sti/citations/AD0779512>.

22. Basu, B., Kapur, G. S., Sarpal, A. S. & Meusinger, R. A Neural Network Approach to the Prediction of Cetane Number of Diesel Fuels Using Nuclear Magnetic Resonance (NMR) Spectroscopy. *Energy Fuels* **17**, 1570–1575 (2003).
23. Dalmiya, A., Sheyyab, M., Mehta, J. M., Brezinsky, K. & Lynch, P. T. Derived cetane number prediction of jet fuels and their functional group surrogates using liquid phase infrared absorption. *Proceedings of the Combustion Institute* **39**, 1495–1504 (2023).
24. Zhang, X. *et al.* Theoretical Studies of the Hydrogen Abstraction from Poly(oxymethylene) Dimethyl Ethers by O₂ in Relation with Cetane Number Data. *ACS Omega* **4**, 19128–19135 (2019).
25. V, S. S. S., Kim, Y., Kim, S., John, P. C. S. & Paton, R. S. Expansion of bond dissociation prediction with machine learning to medicinally and environmentally relevant chemical space. *Digital Discovery* <https://doi.org/10.1039/D3DD00169E> (2023) doi:10.1039/D3DD00169E.
26. Wang, Z., Herbinet, O., Hansen, N. & Battin-Leclerc, F. Exploring hydroperoxides in combustion: History, recent advances and perspectives. *Progress in Energy and Combustion Science* **73**, 132–181 (2019).
27. Gautreaux Jr., M. F. & Coates, J. Activity coefficients at infinite dilution. *AIChE Journal* **1**, 496–500 (1955).
28. Maher, P. J. & Smith, B. D. Infinite Dilution Activity Coefficient Values from Total Pressure VLE Data. Effect of Equation of State Used. *Ind. Eng. Chem. Fund.* **18**, 354–357 (1979).
29. Sandler, S. I. Infinite dilution activity coefficients in chemical, environmental and biochemical engineering. *Fluid Phase Equilibria* **116**, 343–353 (1996).
30. Renon, H. & Prausnitz, J. M. Local compositions in thermodynamic excess functions for liquid mixtures. *AIChE Journal* **14**, 135–144 (1968).
31. Ghosh, P., Hickey, K. J. & Jaffe, S. B. Development of a Detailed Gasoline Composition-Based Octane Model. *Ind. Eng. Chem. Res.* **45**, 337–345 (2006).
32. Ghosh, P. & Jaffe, S. B. Detailed Composition-Based Model for Predicting the Cetane Number of Diesel Fuels. *Ind. Eng. Chem. Res.* **45**, 346–351 (2006).
33. Bondi, A. & Simkin, D. J. Heats of vaporization of hydrogen-bonded substances. *AIChE Journal* **3**, 473–479 (1957).

34. Lampe, A. I., Dittmar, A. K., Heyen, C. & Kiefer, J. Butanol as a potential biofuel: A spectroscopic study of its blends with n-decane and diesel. *Fuel* **222**, 312–318 (2018).
35. Gulder, O. L. & Glavincevski, B. Comments on ‘Prediction of cetane number by group additivity and carbon-13 nuclear magnetic resonance’. *Ind. Eng. Chem. Res.* **27**, 2192–2194 (1988).
36. ASTM D6890-18 - Standard Test Method for Determination of Ignition Delay and Derived Cetane Number (DCN) of Diesel Fuel Oils by Combustion in a Constant Volume Chamber.
<https://webstore.ansi.org/standards/astm/astmd689018?srsId=AfmBOooM4ufJdff-Gw459cDLlxXwZihxMpV7NF978MB1yqiGEcSnH0T>.
37. Inc, iTeh. ASTM D613-18ae1 - Standard Test Method for Cetane Number of Diesel Fuel Oil. *iTeh Standards* <https://standards.iteh.ai/catalog/standards/astm/5607433f-7088-4d1a-a46f-883b324aa4be/astm-d613-18ae1>.
38. Standard Test Method for Determination of Indicated Cetane Number (ICN) of Diesel Fuel Oils using a Constant Volume Combustion Chamber—Reference Fuels Calibration Method. <https://store.astm.org/d8183-18.html>.
39. Mardt, A., Pasquali, L., Wu, H. & Noé, F. VAMPnets for deep learning of molecular kinetics. *Nat Commun* **9**, 5 (2018).
40. Sidky, H., Chen, W. & Ferguson, A. L. High-Resolution Markov State Models for the Dynamics of Trp-Cage Miniprotein Constructed Over Slow Folding Modes Identified by State-Free Reversible VAMPnets. *J. Phys. Chem. B* **123**, 7999–8009 (2019).
41. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv.org* <https://arxiv.org/abs/1802.03426v3> (2018).

CHAPTER 7: Interactions of Molecular Ensembles with their Environments:

Conformationally-Aware Foundation Models for Property Prediction

7.1: Chapter Overview

In this chapter, we present a preliminary investigation examining how distributions of molecular conformations affect physical properties and highlight future studies to confirm our results. We focus our attention on characterizing the conformational distributions of macrocyclic peptides, an important class of molecules that have shown much recent promise as therapeutic agents. Notably, macrocyclic peptides show contradictory behavior to traditional structure-based heuristics for membrane permeability and bioavailability; recent studies have shown that this exceptional behavior may be due in large part to conformational flexibility. For instance, as a macrocyclic peptide passes through a phospholipid bilayer, such as the cell membrane, it moves from a hydrophilic environment to a hydrophobic one. Experimental studies have shown that the macrocyclic peptides change conformation as they move to a hydrophobic environment in such a way that shields exposed polar groups and limits their solvent accessibility, dynamically changing the peptide's logP.

We hypothesize that explicitly encoding this dynamic reparameterization will allow us to build a foundation model combining macrocyclic peptide electrostatics and geometric flexibility. Our preliminary model, PECAN (Peptide Electrostatics through Conformation-Agnostic Networks) is pretrained on conformational ensembles and energies from the literature. Downstream finetuning tasks then make use of learned weighted pooling

functions to learn property-specific reparameterizations for the conformational Boltzmann ensemble. Initial fine-tuned models show performance competitive with state-of-the-art models in the literature. However, detailed ablation studies, model architecture benchmarking, and model analysis are required to confirm our findings and further improve model performance.

7.2: Introduction

Conformational ensembles are an often overlooked but increasingly critical component of targeted molecular design. It is convenient, and often sufficient, to treat the minimum-energy geometry of a molecular species as representative for modeling tasks; for instance, often during mechanistic investigations, the minimum energy conformer is treated as the *de-facto* molecular structure. In reality, however, a multitude of structures exist described by the conformational Boltzmann ensemble, a probability distribution over all possible molecular structures. Experimental measurements are thus a weighted average over this ensemble.

The minimum energy structure corresponds to the peak of this probability distribution; as such, it is the most commonly present structure, and its properties dominate that of the conformational average measurement. The makeup of this distribution can be inferred computationally by conformational sampling, energetic calculations, and the application of an appropriate partition function. However, these calculations depend on the molecular environment: the minimum energy conformer in the gas phase need not be the minima in a solvent environment. For instance, hydrophobic interactions may lead to the

deweighting of conformations exposing polar atoms in a nonpolar solvent; the opposite may be true in a highly polar solvent.

This conformational reweighting has been heavily implicated in numerous chemical phenomena, especially in those shown by macrocyclic peptides, a class of molecules that have shown much recent promise as non-traditional therapeutic agents. Traditional drug design campaigns have been guided by heuristic rules such as Lipinski's Rule of 5, biasing efforts towards compounds with low molecular weights and specific lipophilicities. Macrocyclic peptide therapeutics represent departures from this paradigm, with recent blockbuster drugs such as semaglutide, liraglutide, and lorlatinib breaking most of these established heuristics.¹⁻⁵

This clinical and commercial success has prompted investigation into the unexpected properties of these compounds. Two interdependent examples that have been the subject of much recent study include membrane permeability and binding affinities, properties that have been shown to be dependent on chameleonic conformational changes in the peptide macrocycle (*vide infra*).^{2,6-10} To design molecules with therapeutic utility, both properties must be simultaneously optimized.

For permeability, macrocycle backbones must have enough flexibility to allow facile exposure/shielding of hydrogen bond donor-acceptor pairs, while also remaining small enough to allow for passive transport through membrane leaflets. Traversing the cell membrane phospholipid bilayer involves a transition from the polar, hydrophilic extracellular environment to the nonpolar, lipophilic membrane. This process thus favors a shift in the

conformational ensemble from one that maximizes the solvent exposure (i.e., stabilizing hydrogen bond interactions) of backbone amide groups, to one that favors conformers which shield polar amide groups from solvent exposure.¹¹⁻¹⁴ In addition, neither ensemble corresponds to that commonly used in computational studies, the global energetic minima at 298K in vacuum.¹⁵ Notably, several studies have shown poor correlations between cell permeability and calculated lipophilicities, suggesting that this dynamic ensemble shift is crucial to parameterizing membrane permeability.^{12,16-18}

On the other hand, peptide therapeutics must be able to bind strongly to their targets. Oftentimes, this binding is mediated by large-scale conformational changes in the peptide structure due to interactions with the target and the surrounding environment; in some cases, the bioactive form of the macrocyclic peptide can constitute as little as 4% of the population in solution.¹⁹ In a recent generalized study of macrocyclic peptides, Brueckner and coworkers analyzed computed strain energies for 156 protein-target cocrystal structures.²⁰ It was found that conventional position-based restraints used when modeling bound states yielded unphysical strain energies; instead, low energy conformational ensembles generated using the xGen software²¹ were found to fit the measured electron density while yielding a reasonable balance between peptide ligand strain energy and complex binding enthalpy. Notably, the generated conformational ensembles show marked structural deviations from those generated via conventional methods, indicating large scale induced conformational changes on binding. On the other hand, a targeted study on protein kinase CK2 inhibitors found that a designed 11-mer peptide shows strong inhibition of CK2 β .

On binding, this structure forms a β -hairpin shape which is highly disrupted in alanine scanning studies.^{22,23}

In contrast to the physical picture, existing neural network (NN) models for predicting the properties of macrocyclic peptides neglect this strong coupling to ensemble statistics. Two main classes of NN models exist: 2D QSAR and fingerprint-based models discard 3D structure, instead relying on 2D graph topology or carefully selected descriptors to infer ensemble physical properties, while single-conformer 3D models use a fixed geometry encoded by a message passing graph neural network or rotationally equivariant architecture. While the incorporation of shape information in 3D models adds an additional physically relevant inductive bias to the purely topological 2D models, 3D models remain agnostic to conformational redistribution; neither model class directly encodes the prediction task as function of an environment-dependent conformational ensemble.

The variation in conformer geometry gives rise to an additional challenge in input representation. A complex interplay exists between 3D shape, which determining steric properties and solvent exclusion behavior,^{24,25} topological flexibility, which encodes conformational rigidity,^{26–28} and electrostatics, which encodes how electronic charge is distributed across the ensembles and thus how it redistributes with changes in ensemble weights.^{29,30} No existing strategies incorporate all three representation types simultaneously; instead, existing models focus on individual representations, discarding information encoded by the other types, and ignoring the strong coupling of these three properties: shape determines whether conformers are sterically accessible, topology encodes which

parts of the macrocycle confer conformational flexibility, and electrostatics govern solvent interactions driving conformational reweighting.

Recently developed large-scale conformational datasets, however, may enable alternative approaches. The CREMP dataset developed by Grambow and coworkers³¹ presents a collection of over 31 million geometries corresponding to 36,198 unique macrocyclic peptide sequences. Sequences were selected via combinatorial enumeration of possible 4-, 5-, and 6-mer homodectic macrocycles consisting of canonical amino acids. In addition to side chain variation, L- and D- amino acids were considered, as well as N-methylation. Conformers are generated using the metadynamics-based CREST sampling tool;³² inputs to the CREST tool were obtained by optimizing 1000 conformers generated by the rdkit³³/MMFF94³⁴ with GFN2-xtb³⁵ using the ALPB solvent model in chloroform and selecting the lowest energy conformer. GFN2-xTB energies calculated at 298K are also provided for each conformer in the CREMP dataset. In total, 3.9 million CPU core-hours of

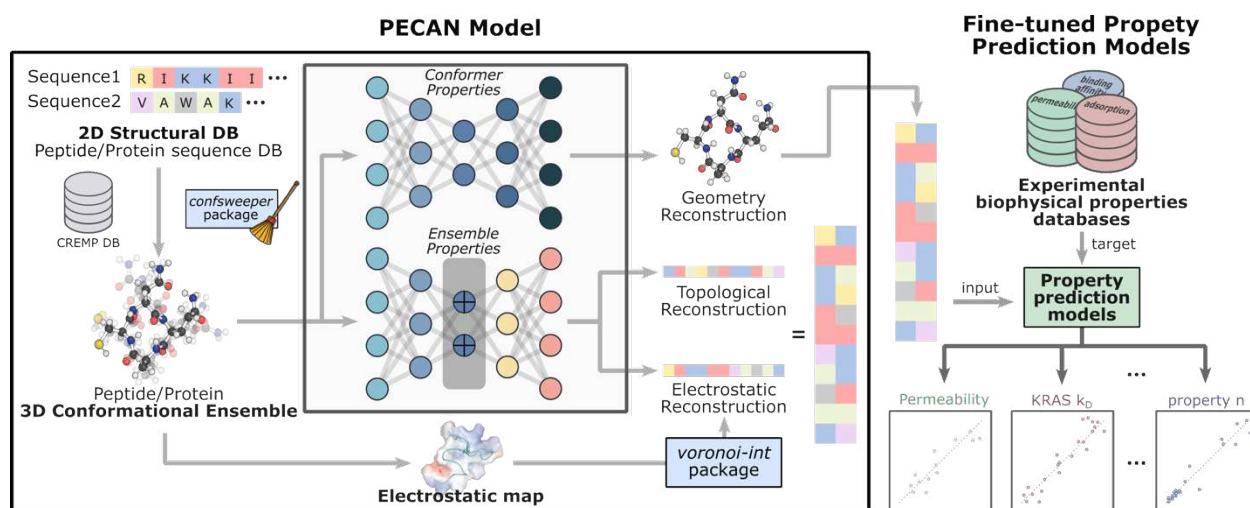


Figure 7.1: Overview of the PECAN model architecture. 3D conformational ensembles of macrocyclic peptides are used to pretrain the PECAN foundation model; ensembles from the literature (CREMP DB) are currently used for pretraining, but tools like the confsweeper package (currently under development) can be used to rapidly generate suitable conformational ensembles. Geometry reconstruction, topological reconstruction, and electrostatic reconstruction tasks are trained to generate a unified representation of macrocyclic peptide physical characteristics. The resultant model embeddings can be used for downstream fine-tuning prediction tasks, such as membrane permeability, binding affinity, etc.

compute time were expended to generate this dataset, making it a valuable resource for model training. Despite its availability, to the best of our knowledge, no large-scale discriminative model has been released the CREMP dataset for model training.

In this work, we present a unified foundation model bridging macrocyclic peptide structure, shape and electrostatics, PECAN (Peptide Electrostatics through Conformation-Agnostic Networks). The PECAN model uses a multiobjective pretraining strategy (**Figure 7.1**) to learn a) encodings of molecular geometry through pairwise distances; b) encodings of conformational ensemble electrostatics via electrostatic potentials projected onto spherical harmonics; and c) encodings of conformational ensemble topology via variance of topological persistence barcodes across the ensemble. Parallel trunks process per-conformation geometry-based embeddings and per-ensemble electrostatic and topology-based embeddings separately, allowing the same model to encode both per-conformer and ensemble aggregate properties. Conformational weights are explicitly encoded in the ensemble aggregation step; this weight is adjusted during downstream fine-tuning steps to reweight the ensemble for specific properties of interest. We will evaluate the performance of the PECAN foundation model using two examples of model fine-tuning for specific property prediction tasks: a) membrane permeability as measured by the PAMPA permeability assay;^{36–38} and b) binding affinity to the KRAS protein implicated in cell proliferation and differentiation.^{39,40} Preliminary investigations show that PECAN embeddings can be used in downstream models to predict macrocyclic membrane permeability; detailed optimization of the fine-tuned permeability model and training of the binding affinity fine-tuned model is ongoing, along with detailed analysis of the foundation

model and ablation studies. After training is concluded, we will perform analysis of learned model conformational weights to reveal the role of individual conformers on membrane permeability and binding affinity, with the end goal of using this information to guide downstream design of membrane-permeable and high-affinity macrocyclic peptide binders.

7.3: Results, Future Work, and Discussion

7.3.1: Dataset Generation and Conformer Ensemble Processing

The CREMP dataset curated by Grambow and coworkers was used as a starting point for database creation.³¹ The CREMP dataset consists of 31.3 million conformers corresponding to 36,198 macrocyclic peptide sequences sampled by CREST metadynamics simulations. Accompanying energetics are calculated at the GFN2-xTB level of theory at 298 K in ALPB chloroform. Peptide sequences consist of 4-, 5-, and 6-mers, with constituent amino acids sampled by combinatorial enumeration of canonical amino acids, L/D isomers, and N-methylation.

Conformers from the CREMP dataset were obtained to construct a pretraining dataset. H-atoms were then stripped, and features for both atoms (one-hot element encoding, one-hot hybridization encoding, formal charge, binary aromaticity, binary ring membership, one-hot chirality flag, xyz coordinates) and bonds (one hot bond type, binary conjugation, binary ring membership, xyz bond vector) were computed. To allow the testing of different equivariant architectures for the foundation model (*vide infra*), a molecular graph was constructed from the raw coordinate xyz files in the CREMP dataset using an 8 Å distance cutoff. Molecule indices were assigned to each conformer to keep track of peptide

sequence identity during pooling operations (*vide infra*); zero padding conformers (all zero positions) were used to account for differences in numbers of conformers. Such padding conformers were assigned Boltzmann weights of zero.

In this preliminary study, no filtering or deduplication of conformers was performed; rather, the raw outputs from CREST present in the CREMP dataset were used as is. This presents two potential related problems for production fine-tuning stemming from the cutoffs used by the CREST software to determine whether two conformers are duplicates or not. CREST uses three simultaneous cutoffs to do so: an energetic cutoff of 0.1 kcal/mol, an RMSD cutoff of 0.125 Å, and a rotational constant cutoff of 15.0 MHz. If two structures have an energetic difference over the energetic cutoff, an RMSD over the RMSD cutoff, and a rotational constant difference over the rotational constant cutoff, the structures are considered distinct conformers; if they instead have an RMSD over the RMSD cutoff, but energetic difference and rotational constant difference below the respective cutoff values, the structures are considered distinct rotamers. All other cases are considered to be duplicate conformers. Both conformers and rotamers are returned by default.

These cutoff values result in high coverage of the conformer space, but frequently result in nearly identical structures being counted as distinct conformers. For instance, the accuracy of most density function theory (DFT) functionals is limited to 2 kcal/mol; the literature standard for “chemical accuracy” is considered to be 1 kcal/mol. As such, the usage of default cutoff values for CREST in creating the CREMP dataset may result in over-representation of the conformational landscape, and filtering strategies, such as RMSD or

torsional fingerprint-based Butina clustering must be benchmarked to ensure overfitting does not occur. This analysis will be performed in future work.

The conformer structures from the CREMP dataset were used to generate targets for training the PECAN model. Three separate target classes are used for pretraining: a) pairwise distance reconstruction for encoding geometry; b) discretized electrostatic potential reconstruction for encoding electrostatics; c) persistence image reconstruction for encoding topology. Reconstruction proceeds through a dynamic masking strategy at training time, as commonly used in training transformer-based models.⁴¹ The distance matrix is generated by computing pairwise distances between heavy atoms within each conformer; since the resultant matrix is symmetric, only an upper triangular version is used as a label.

Discretized representations of the electrostatic potential for each conformer are also computed. Two distinct representations are computed and concatenated to obtain the final training target: a projection into spherical harmonics, as well as a per-atom representation. To compute the spherical harmonic projection, the per-conformer electrostatic potential is evaluated on a 200-point Fibonacci sphere centered at the molecule's center of mass, with radius $1.5\times$ the maximum heavy-atom distance from the center. The electrostatic potential is computed from Gasteiger charges⁴² obtained from rdkit via the Coulomb sum $V(r) = \sum_j \frac{q_j}{|r - r_j|}$. The potential on the sphere is projected onto real spherical harmonics up to $L = 4$, producing a 25-coefficient vector per conformer. The Boltzmann-weighted average over conformers is then taken as the pretraining target. The coordinate frame is the principal-axes

frame derived from the inertia tensor, making the target SE(3)-invariant without requiring a canonical alignment step. We note that the obtained coefficients can be negative, informing architecture of the prediction trunk. In addition, the per-atom electrostatic potential $V_i = \sum_{j \neq i} \frac{q_j}{|r_i - r_j|}$ is computed per conformer from Gasteiger charges and normalized to produce a unit-free, scale-invariant vector over atoms. The ESP-atom representation provides per-site electrostatic detail that the SH coefficients average over; together, the targets cover both the molecular-field-level and atom-site-level electrostatic structure.

A persistence image reconstruction is used to encode topological flexibility in the macrocycle loop. A thorough discussion on topological persistence images and barcodes is provided in Chapter 4.3, *Generation of Topological Barcode Embeddings (TopoCoder)*; for brevity, we will omit this discussion here. Here, only the 2D persistence is considered; due to the relatively consistent amide bond length, the 1D persistence (bond length-based) is trivial for macrocyclic peptides, and provides comparatively little information. Rather, the 2D persistence captures loops in the atom-distance graph that correspond to ring structures and transient hydrogen-bond cycles. We utilize the variance in the persistence diagram across the conformational ensemble for a given sequence as the prediction target, as the ensemble mean is dominated by the invariant backbone macrocycle ring, which contributes large, consistent 2D features regardless of conformation. The variance thus isolates barcode features that differ across conformers, corresponding to flexible sidechain loops and transient hydrogen bonds, which are the conformationally sensitive features. The persistence image used here is a 20×20 raster of the 2D persistence diagram, with atom

coordinates normalized by molecular diameter for scale invariance across peptide lengths and Gaussian kernel bandwidth $\sigma = 0.1$.

We note that the usage of CREMP conformers for model training is motivated by dataset availability and the computational expense required to generate high-quality conformational ensembles of macrocyclic peptides. To address this problem and facilitate efficient generation of conformer ensembles, we are currently developing the *confsweeper* package. *confsweeper* takes advantage of recent developments in GPU-based implementations⁴³ of the rdkit ETKDGV3 conformer generation methodology⁴⁴ and the development of fast, accurate machine learning interatomic potentials to rapidly generate conformers and energetic predictions for a wide variety of molecular species through massive parallelized random sampling.^{45–47} However, the cyclic nature of macrocyclic peptides introduces sampling challenges that cannot be overcome through brute-force sampling: in many cases, concerted backbone amide flips are necessary to generate energetically distinct conformers. This process involves a large energetic barrier that cannot be overcome through naïve Monte Carlo sampling. Instead, current work incorporates a directed Multiple Minima Monte Carlo sampling procedure combining the Monte Carlo rotational move set introduced by Dodd et. al.⁴⁸ for backbone dihedral exploration and side chain Cartesian jumps inspired by Goedecker’s minima hopping algorithm, using simulated annealing. Detailed benchmarking studies are necessary to ensure proper coverage of the conformational space of macrocyclic peptides, however, and we leave the use of *confsweeper* for input generation for future study.

Further, the use of Gasteiger charges to generate electrostatic targets presents another avenue for model improvement. While easy to use due to the convenient implementation in the rdkit, Gasteiger charges are conformationally invariant; the spherical harmonic coefficient representations in the current implementation vary across conformers due to the change in position of atoms, thus changing the denominator of the Coulomb sum. This problem is particularly relevant for macrocycles where N-methylation or ring strain affects electron density across conformers, where the current approximation underestimates conformational charge redistribution. The use of GFN2-xTB charges would alleviate this problem by explicitly calculating the charge redistribution on change in conformer geometry, and will be implemented in future work. Alternatively, we are currently developing the *voronoi-int* package to allow discretized partial charges to be calculated from the results of an Adaptive Poisson Boltzmann Solver (APBS) calculation.⁴⁹ The APBS software package calculates the electrostatic potential at discretized grid points, given an input structure; our *voronoi-int* package performs a nearest-neighbors spatial integration of this grid data given a set of “anchor points”, allowing spatially resolved electrostatic potential to be mapped back to heavy atoms. A detailed comparison of *voronoi-int* calculated electrostatics and GFN2-xTB calculated partial charges must be performed before the results of the *voronoi-int* software can be used for model training; we leave this work for future efforts.

7.3.2: PECAN Model Design and Training

The PECAN model is constructed to allow the reconstruction of both individual conformer level features as well as ensemble level features. We utilize a two-trunk architecture to do so, in which one trunk makes predictions for conformers, referred to here on as the conformer trunk, while the other trunk makes predictions for the ensemble, referred to here as the ensemble trunk. Both trunks share an initial common graph encoder structure; representations are then split into each trunk for specific conformer-level/ensemble-level prediction tasks. The conformer trunk consists of stacked linear layers and SiLU activations (multilayer perceptron) acting on node feature embeddings from the encoder graph layers; this simplistic architecture will be evaluated against more sophisticated alternatives, such as message passing GNNs and transformer-based architectures, in future studies. The conformer trunk is used for geometric reconstruction; the output MLP predicts masked elements of the upper triangular pairwise distance matrix, a rotationally invariant task.

For the prediction of ensemble-level properties, representations are subject to a learned pooling operation. Node and edge features from the encoder are subject to a Janossy pooling⁵⁰ operation using a Boltzmann-weighted aggregation scheme. For a given conformer, a trainable multilayer perceptron initialized with weights of unity maps a per-conformer embedding h_i to a scalar correction m_i . These weights are frozen during pretraining, and unfrozen for downstream fine-tuning tasks. The pooling weight for a conformer i is then given by:

$$\log w_i = m_i \cdot \log b_i \quad 7.1$$

Then, by applying a softmax function over all conformers, individual conformer weights are obtained. The molecule level representation is then given by :

$$z = \sum_i w_i \cdot h_i \quad 7.2$$

The Boltzmann weight acts as a physical prior weighting conformer importance, parameterized by the conformer energy. The MLP correction allows deviation from this energetic ranking when downstream property signals warrant it, allowing for dynamic reweighting of the conformer ensemble for property-specific prediction tasks. The molecule level representation is passed through parallel MLP output layers for each ensemble level pretraining prediction task. For the topological reconstruction task, MLP outputs are passed through a softplus activation function, matching the non-negative variance prediction target. Raw logit outputs are used for the electrostatic prediction task, as both spherical harmonic coefficients and the per-atom potentials can be negative.

A composite loss function and complementary training schedule is used to balance model training across the three target domains. Each loss type is added during one phase of training, a scheme motivated by loss scale at initialization: the untrained electrostatic and topology prediction heads produce loss contributions approximately 100× larger than the geometric head, which destabilizes encoder training in early epochs before useful gradient signal is available. Masking ratios are also increased sequentially between training phases to account for additional model parametric flexibility as additional heads are introduced. In

phase 1, only the geometric prediction head is active and a mask ratio of 0.15 is used, targeting convergence to a meaningful distance-prediction regime. Phase 2 introduces electrostatics after the encoder representations are stable; Phase 3 adds topology, whose sparse target (2D persistence image variance) provides the weakest training signal and benefits from an already-informed encoder. A per-term exponential moving average is used to further normalize the composite loss function, and per-term weighting factors are used to prioritize different representations. The per-term weighting factors are model hyperparameters, and must be rigorously benchmarked in future work to determine optimal values; we use values of $\lambda_{geo} = 1.0$, $\lambda_{esp_{sh}} = 0.5$, $\lambda_{esp_{atom}} = 0.5$, $\lambda_{topo} = 0.1$ for our preliminary study, reflecting a physically-motivated importance ranking: geometry is used as the primary pretraining signal, electrostatics inform target application modalities, and topology serves the auxiliary representation that enriches the latent representations without dominating training. The composite training loss function is thus:

$$L_{total} = \lambda_{geo} \cdot \frac{L_{geo}}{EMA_{geo}} + \lambda_{esp_{sh}} \cdot \frac{L_{esp_{sh}}}{EMA_{esp_{sh}}} + \lambda_{esp_{atom}} \cdot \frac{L_{esp_{atom}}}{EMA_{esp_{atom}}} + \lambda_{topo} \cdot \frac{L_{topo}}{EMA_{topo}} \quad 7.3$$

During model validation, the mask ratio is set to 1.0, ensuring full evaluation across target dimensions.

Multiple encoder structures were tested and evaluated for pretraining accuracy; we specifically sought to investigate the impact of incorporating explicit rotational equivariance in the encoder structure. As such, we benchmarked the use of both invariant and equivariant encoder architectures; we use EquiformerV2⁵¹ encoder layers for the equivariant architecture, and Graph Transformer⁵² layers with radial basis function edge embeddings

(similar to those in the SchNet architecture⁵³) for the invariant architecture. Hidden layer dimensions (channel dimension for EquiformerV2) of 128 were used throughout to maintain 1:1 comparisons between architectures. Preliminary pretraining studies show that the equivariant architecture outperforms the invariant architecture, producing validation losses that differ by one order of magnitude (0.006 vs 0.01 after 50 epochs). However, due to the additional overhead incurred by the equivariant encoder architecture, model training is significantly more expensive: on two nVIDIA GV100 GPUs, training the equivariant architecture completes in 22 hours, compared to 1.5 hours for the invariant architecture. Detailed evaluation of downstream model prediction accuracies is necessary to determine which model architecture is most suitable for use; studies are currently ongoing.

7.3.3: Preliminary Finetuning Studies

To verify whether latent embeddings from the pretrained PECAN model can be used for downstream property prediction tasks, we extract latent vectors from the ensemble prediction trunks and use them for downstream property prediction tasks. We emphasize here that the following results were obtained without any model finetuning. Hence, conformer ensem⁵⁴ble pooling weights reflect those obtained at the GFN2-xTB level of theory in ALPB chloroform at 298 K. As such, we expect marked improvement to these results when full model finetuning and evaluation is performed.

We focus our property prediction studies on two distinct prediction tasks where shifts in the conformational distribution drive property values: membrane permeability and binding affinity. In the case of membrane permeability, conformations that bury exposed

polar functional groups are favored during membrane traversal, while conformations that expose polar functional groups are favored in aqueous solution. On the other hand, peptide binding is driven by induced conformational changes which allow the peptide macrocycle to most closely complement the shape and electrostatics of a binding pocket. As such, conventional models that use single-conformer representations struggle to predict these property targets.

The recently developed CycPeptMP model uses a multiscale neural network framework to unify atom level, monomer level, and peptide level features.⁵⁴ The network is then trained on experimental data ($\log P_{\text{exp}}$ measured by PAMPA assay³⁸) from the CycPeptMPDB dataset³⁶ using multiscale data augmentation techniques, and achieves a mean absolute error (MAE) of 0.352 log units and an R^2 of 0.613 on a held out test set. The authors compare the performance of their multiscale model to that of the heterogeneous graph transformer introduced by Jiang et. al.⁵⁵ and the Fingerprint-Enhanced Graph Attention Network introduced by Choo et. al.⁵⁶ when trained on the same dataset. It was found that both models perform worse than the CycPeptMP model, achieving MAEs and R^2 values of 0.399 and 0.408 log units and 0.560 and 0.545 respectively. The CPMP model developed by Jiang et. al. uses a Molecule Attention Transformer⁵⁷ framework to encode interatomic distances in the molecular graph framework.⁵⁸ The CPMP model is trained on a curated consolidated dataset of experimental permeability measurements and achieves a MAE of 0.308 log units and an R^2 value of 0.671 on PAMPA prediction tasks.

To provide a baseline benchmark for the performance of our model, we obtained PAMPA permeability measurements from the Macrocycle-DB³⁷ corresponding to 8,466 experimental measurements from the CycPeptMPDB³⁶; the Macrocycle-DB is a comprehensive database aggregating experimental measurements of macrocycle physical properties from a variety of literature sources. We subsampled 500 random sequences, and generated conformers and associated single point energies using the CREST software package, mirroring the procedure used for conformer generation in the CREMP dataset.³¹ The pretrained equivariant encoder PECAN model was used to generate ensemble level embeddings for each sequence, and the resultant embeddings were used to train and evaluate a random forest model implemented in scikit-learn using default hyperparameters. A 90:10 train:test random split was used to generate training and test sets. The resulting random forest model achieves an accuracy of 0.182 log units and an R^2 value of 0.704 on the held out test set (**Figure 7.2a**).

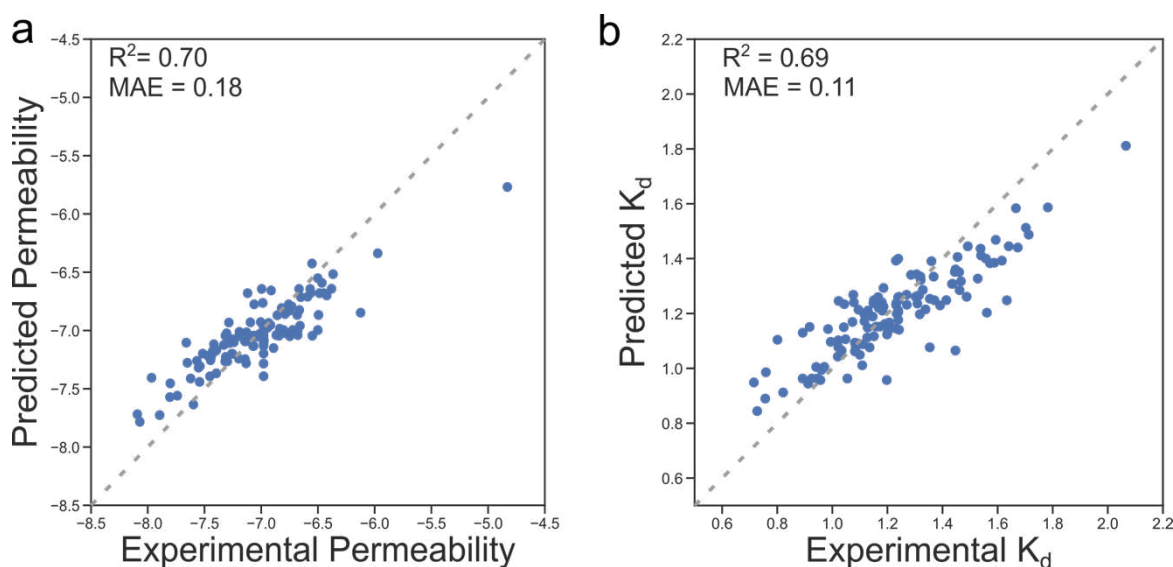


Figure 7.2: Preliminary property prediction model results using ensemble-level sequence embeddings from the pretrained equivariant PECAN model. a) Predictions for 50 test set peptide sequences subsampled from the CycPeptMPDB dataset; b) Predictions for 50 test set peptide sequences subsampled from the

We perform a similar study to model binding affinities for the KRAS protein. KRAS (Kirsten rat sarcoma virus oncogene) is a gene encoding the K-ras protein. In a slight abuse of notation, “KRAS” is frequently used in the literature to refer to the K-ras protein target instead of the encoding gene; we will adopt this convention here. The KRAS protein is a GPTase (ie., an enzyme that hydrolyzes guanosine triphosphate to guanosine diphosphate) and is part of the RAS/MAPK extracellular signaling pathway; binding of an extracellular signaling compound to receptor tyrosine kinase proteins in the cell membrane triggers a kinase cascade that results in the binding of an inactive KRAS-GDP to an activated receptor complex. This binding event triggers the unbinding of GDP, allowing the KRAS protein to bind GTP and become active. This activated KRAS complex can then perpetuate a subsequent kinase cascade as well as recruiting glucose transporters. As a result of this signaling pathway, the cell is instructed to proliferate or differentiate. Mutations in the KRAS gene can cause the KRAS protein to become permanently activated, triggering uncontrolled cell growth and proliferation; these mutations have been implicated in several cancers.^{59–61} As such, the development of peptide therapeutics that bind to and inactivate KRAS proteins has been the subject of much study.

Little public data exists on macrocyclic peptide KRAS binding affinities. However, Xu et. al. curated a dataset of 2757 experimental K_d values from a patent filed by Kawada et. al.^{39,40} Xu et. al. then benchmark a number of prediction models on this dataset; while errors are not reported, the best performing model (XGBoost) achieves a Spearman correlation coefficient of 0.82. We utilize a similar strategy to our membrane permeability prediction tasks to train a random forest property prediction model on this dataset. Once again, we

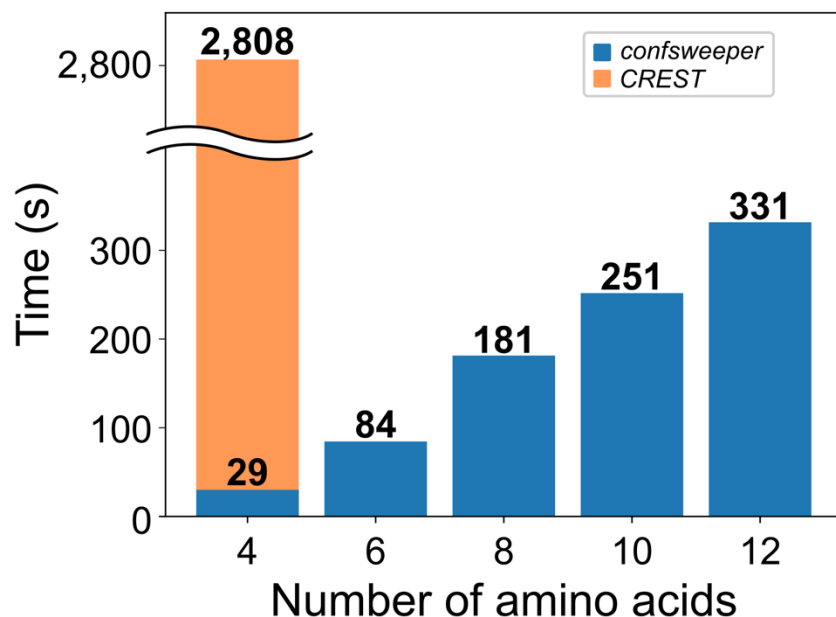


Figure 7.3: Timing performance of the *confsweeper* package compared to CREST

subsample 500 random peptide sequences, and use a 90:10 train:test split to train a random forest model using the default hyperparameters in scikit-learn. Our random forest model achieves an MAE of 0.11 units and an R^2 of 0.69 on this sampled dataset (**Figure 7.2b**).

Due to the randomized split strategy and limited dataset size, we acknowledge this result is likely an overestimate of model performance. However, a naïve approach is computationally expensive: generating CREST conformational ensembles for all 8,466 CycPeptMPDB peptide sequences would require the expenditure of approximately 1 million core-hours of CPU time. Ongoing work focuses on using our in-house *confsweeper* package to generate conformational ensembles for training; preliminary tests show a speedup of over 2 orders of magnitude compared to the CREST pipeline used to generate the CREMP dataset (**Figure 7.3**). Use of the accelerated *confsweeper* pipeline will allow for the full-scale CycPeptMPDB and KRAS inhibitor datasets to be used for model finetuning and analysis.

7.4: Conclusion

Macrocyclic peptides are a clinically promising therapeutic modality whose physicochemical properties are closely coupled to their conformational ensemble distribution. However, distinct physical and chemical events must occur for therapeutic effect: the macrocyclic peptide must move through the cell membrane, and then bind to a protein target. Different peptide conformers are favored for each phase of this process, with the conformational ensemble of the peptide in extracellular space (a hydrophilic environment) being distinct from that of the peptide as it moves through the cell membrane (a hydrophobic environment) or as it binds to the protein of interest (environment dictated by the protein binding pocket). Property prediction models that use only a single peptide conformer are agnostic to this behavior.

In this work, we present a preliminary effort to address this discrepancy using a model which explicitly parameterizes changes in the conformational distribution. The PECAN foundation model uses a multi-objective training strategy to encode conformational geometry, the electrostatics of the conformational ensemble, and the differences in topology across the conformational ensemble. A curriculum training strategy is used to selectively weight these factors during model training. Downstream property prediction tasks finetune per-conformer weights, learning the conformational reparameterization that is tied to a specific macrocyclic peptide property. Preliminary property prediction results on membrane permeability and KRAS protein binding affinities using raw latent vectors from a preliminary version of the PECAN model without finetuning are comparable to state-of-the-

art models in the literature, indicating that a final version of the PECAN model may accurately capture ensemble electrostatics and conformational flexibility in a manner that can be used for accurate downstream property prediction tasks.

7.4: References

1. Du, Y. *et al.* FDA-approved drugs featuring macrocycles or medium-sized rings. *Archiv der Pharmazie* **358**, e2400890 (2025).
2. Ohta, A. *et al.* Validation of a New Methodology to Create Oral Drugs beyond the Rule of 5 for Intracellular Tough Targets. *J. Am. Chem. Soc.* **145**, 24035–24051 (2023).
3. Tanada, M. *et al.* Development of Orally Bioavailable Peptides Targeting an Intracellular Protein: From a Hit to a Clinical KRAS Inhibitor. *J. Am. Chem. Soc.* **145**, 16610–16620 (2023).
4. Driggers, E. M., Hale, S. P., Lee, J. & Terrett, N. K. The exploration of macrocycles for drug discovery — an underexploited structural class. *Nat Rev Drug Discov* **7**, 608–624 (2008).
5. Marsault, E. & Peterson, M. L. Macrocycles Are Great Cycles: Applications, Opportunities, and Challenges of Synthetic Macrocycles in Drug Discovery. *J. Med. Chem.* **54**, 1961–2004 (2011).
6. Adaligil, E. & Fairbrother, W. J. NMR Spectroscopy for Studying Peptide Conformations and Cell Permeability. in *Approaching the Next Inflection in Peptide Therapeutics: Attaining Cell Permeability and Oral Bioavailability* vol. 1417 155–177 (American Chemical Society, 2022).
7. Faris, J. H. *et al.* Membrane Permeability in a Large Macrocyclic Peptide Driven by a Saddle-Shaped Conformation. *J. Am. Chem. Soc.* **146**, 4582–4591 (2024).
8. Wang, C. K. *et al.* Rational design and synthesis of an orally bioavailable peptide guided by NMR amide temperature coefficients. *Proceedings of the National Academy of Sciences* **111**, 17504–17509 (2014).
9. Ramelot, T. A., Palmer, J., Montelione, G. T. & Bhardwaj, G. Cell-permeable chameleonic peptides: Exploiting conformational dynamics in *de novo* cyclic peptide design. *Current Opinion in Structural Biology* **80**, 102603 (2023).

10. Bhardwaj, G. *et al.* Accurate de novo design of membrane-traversing macrocycles. *Cell* **185**, 3520–3532.e26 (2022).
11. Danelius, E. *et al.* Solution Conformations Explain the Chameleonic Behaviour of Macrocyclic Drugs. *Chemistry – A European Journal* **26**, 5231–5244 (2020).
12. Rossi Sebastiano, M. *et al.* Impact of Dynamically Exposed Polarity on Permeability and Solubility of Chameleonic Drugs Beyond the Rule of 5. *J. Med. Chem.* **61**, 4189–4202 (2018).
13. Atilaw, Y. *et al.* Solution Conformations Shed Light on PROTAC Cell Permeability. *ACS Med. Chem. Lett.* **12**, 107–114 (2021).
14. Nielsen, D. S. *et al.* Improving on Nature: Making a Cyclic Heptapeptide Orally Bioavailable. *Angewandte Chemie International Edition* **53**, 12059–12063 (2014).
15. Brethomé, A. V., Fletcher, S. P. & Paton, R. S. Conformational Effects on Physical-Organic Descriptors: The Case of Sterimol Steric Parameters. *ACS Catal.* **9**, 2313–2323 (2019).
16. Scott, D. E. *et al.* Systematic Investigation of the Permeability of Androgen Receptor PROTACs. *ACS Med. Chem. Lett.* **11**, 1539–1547 (2020).
17. Cantrill, C. *et al.* Fundamental aspects of DMPK optimization of targeted protein degraders. *Drug Discovery Today* **25**, 969–982 (2020).
18. J. McTiernan, T., B. Diaz, D., J. Saunders, G., Sprang, F. & K. Yudin, A. Navigating complex peptide structures using macrocycle conformational maps. *RSC Chemical Biology* **3**, 739–747 (2022).
19. Thepchatiri, P., Eliseo, T., Cicero, D. O., Myles, D. & Snyder, J. P. Relationship Among Ligand Conformations in Solution, in the Solid State, and at the Hsp90 Binding Site: Geldanamycin and Radicicol. *J. Am. Chem. Soc.* **129**, 3127–3134 (2007).
20. Brueckner, A. C. *et al.* Conformational Strain of Macrocyclic Peptides in Ligand–Receptor Complexes Based on Advanced Refinement of Bound-State Conformers. *J. Med. Chem.* **64**, 3282–3298 (2021).
21. Jain, A. N. *et al.* XGen: Real-Space Fitting of Complex Ligand Conformational Ensembles to X-ray Electron Density Maps. *J. Med. Chem.* **63**, 10509–10528 (2020).
22. Raaf, J. *et al.* First Structure of Protein Kinase CK2 Catalytic Subunit with an Effective CK2 β -Competitive Ligand. *ACS Chem. Biol.* **8**, 901–907 (2013).

23. Laudet, B. *et al.* Structure-based design of small peptide inhibitors of protein kinase CK2 subunit interaction. *Biochem J* **408**, 363–373 (2007).
24. Poongavanam, V. *et al.* Predicting the Permeability of Macrocycles from Conformational Sampling – Limitations of Molecular Flexibility. *Journal of Pharmaceutical Sciences* **110**, 301–313 (2021).
25. Begnini, F. *et al.* Cell Permeability of Isomeric Macrocycles: Predictions and NMR Studies. *ACS Med. Chem. Lett.* **12**, 983–990 (2021).
26. Nielsen, D. S. *et al.* Flexibility versus Rigidity for Orally Bioavailable Cyclic Hexapeptides. *ChemBioChem* **16**, 2289–2293 (2015).
27. Kamenik, A. S., Linker, S. M. & Riniker, S. Matching Simulations and Experiments of Conformationally Flexible Cyclic Peptides: Steps toward a Holistic View of Passive Membrane Permeability. in *Approaching the Next Inflection in Peptide Therapeutics: Attaining Cell Permeability and Oral Bioavailability* vol. 1417 137–154 (American Chemical Society, 2022).
28. Appavoo, S. D., Heller, N. W., van Campenhout, C., Saunders, G. J. & Yudin, A. K. Identification of “Structural Pin” Interactions and their Significance for the Conformational Control of Macrocyclic Scaffolds. *Angewandte Chemie International Edition* **63**, e202402372 (2024).
29. Kamenik, A. S. *et al.* Macrocycle Cell Permeability Measured by Solvation Free Energies in Polar and Apolar Environments. *J. Chem. Inf. Model.* **60**, 3508–3517 (2020).
30. Hung, T. I., Venkatesan, R. & Chang, C. Mechanistic Dissection of Conformational Transition of Bicyclic Peptide via Molecular Modeling and Deep Learning. 2026.03.12.711417 Preprint at <https://doi.org/10.64898/2026.03.12.711417> (2026).
31. Grambow, C. A., Weir, H., Cunningham, C. N., Biancalani, T. & Chuang, K. V. CREMP: Conformer-rotamer ensembles of macrocyclic peptides for machine learning. *Sci Data* **11**, 859 (2024).
32. Pracht, P., Bohle, F. & Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Phys. Chem. Chem. Phys.* **22**, 7169–7192 (2020).
33. RDKit: Open source cheminformatics. <https://doi.org/10.5281/zenodo.591637> doi:10.5281/zenodo.591637.

34. Halgren, T. A. Merck molecular force field. V. Extension of MMFF94 using experimental data, additional computational data, and empirical rules. *Journal of Computational Chemistry* **17**, 616–641 (1996).
35. Bannwarth, C., Ehlert, S. & Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **15**, 1652–1671 (2019).
36. Li, J. *et al.* CycPeptMPDB: A Comprehensive Database of Membrane Permeability of Cyclic Peptides. *J. Chem. Inf. Model.* **63**, 2240–2250 (2023).
37. Jiang, M. *et al.* Macrocycle-DB: a comprehensive database for macrocycle-based drug discovery. *Nucleic Acids Res* **54**, D1469–D1476 (2026).
38. Masungi, C. *et al.* Parallel Artificial Membrane Permeability Assay (PAMPA) combined with a 10-day multiscreen Caco-2 cell culture as a tool for assessing new drug candidates. *Pharmazie* **63**, 194–199 (2008).
39. Kawada, H. *et al.* Cyclic compound having selective inhibitory action on kras over hras and nras. (2024).
40. Xu, X. *et al.* HELM-GPT: de novo macrocyclic peptide design using generative pre-trained transformer. *Bioinformatics* **40**, btae364 (2024).
41. Zhuang, B. *et al.* A Survey on Efficient Training of Transformers. Preprint at <https://doi.org/10.48550/arXiv.2302.01107> (2023).
42. Gasteiger, J. & Marsili, M. A new model for calculating atomic charges in molecules. *Tetrahedron Letters* **19**, 3181–3184 (1978).
43. NVIDIA nvMolKit Documentation — nvMolKit. <https://nvidia-digital-bio.github.io/nvMolKit/index.html>.
44. Riniker, S. & Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Inf. Model.* **55**, 2562–2574 (2015).
45. Kovács, D. P. *et al.* MACE-OFF: Short-Range Transferable Machine Learning Force Fields for Organic Molecules. *J. Am. Chem. Soc.* **147**, 17598–17611 (2025).
46. Wood, B. M. *et al.* UMA: A Family of Universal Models for Atoms. Preprint at <https://doi.org/10.48550/arXiv.2506.23971> (2025).
47. Batatia, I. *et al.* A foundation model for atomistic materials chemistry. *J. Chem. Phys.* **163**, 184110 (2025).

48. Dodd, L. R., Boone, T. D. & Theodorou, D. N. A concerted rotation algorithm for atomistic Monte Carlo simulation of polymer melts and glasses. *Molecular Physics* **78**, 961–996 (1993).
49. Jurrus, E. *et al.* Improvements to the APBS biomolecular solvation software suite. *Protein Science* **27**, 112–128 (2018).
50. Murphy, R. L., Srinivasan, B., Rao, V. & Ribeiro, B. Janossy Pooling: Learning Deep Permutation-Invariant Functions for Variable-Size Inputs. Preprint at <https://doi.org/10.48550/arXiv.1811.01900> (2019).
51. Liao, Y.-L., Wood, B., Das, A. & Smidt, T. EquiformerV2: Improved Equivariant Transformer for Scaling to Higher-Degree Representations. Preprint at <https://doi.org/10.48550/arXiv.2306.12059> (2024).
52. Yun, S., Jeong, M., Kim, R., Kang, J. & Kim, H. J. Graph Transformer Networks.
53. Schütt, K. T. *et al.* SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. *arXiv.org* <https://arxiv.org/abs/1706.08566v5> (2017).
54. Li, J., Yanagisawa, K. & Akiyama, Y. CycPeptMP: enhancing membrane permeability prediction of cyclic peptides with multi-level molecular features and data augmentation. *Brief Bioinform* **25**, bbae417 (2024).
55. Jiang, Y. *et al.* Pharmacophoric-constrained heterogeneous graph transformer model for molecular property prediction. *Commun Chem* **6**, 60 (2023).
56. Choo, H. Y., Wee, J., Shen, C. & Xia, K. Fingerprint-Enhanced Graph Attention Network (FinGAT) Model for Antibiotic Discovery. *J. Chem. Inf. Model.* **63**, 2928–2935 (2023).
57. Maziarka, Ł. *et al.* Molecule Attention Transformer. *arXiv.org* <https://arxiv.org/abs/2002.08264v1> (2020).
58. Jiang, D., Chen, Z. & Du, H. Cyclic peptide membrane permeability prediction using deep learning model based on molecular attention transformer. *Front. Bioinform.* **5**, (2025).
59. Tam, I. Y. S. *et al.* Distinct Epidermal Growth Factor Receptor and KRAS Mutation Patterns in Non–Small Cell Lung Cancer Patients with Different Tobacco Exposure and Clinicopathologic Features. *Clin Cancer Res* **12**, 1647–1653 (2006).
60. Almoguera, C. *et al.* Most human carcinomas of the exocrine pancreas contain mutant c-K-ras genes. *Cell* **53**, 549–554 (1988).

61. Burmer, G. C. & Loeb, L. A. Mutations in the KRAS2 oncogene during progressive stages of human colon carcinoma. *Proceedings of the National Academy of Sciences* **86**, 2403–2407 (1989).

CHAPTER 8: Concluding Remarks and Future Directions

The field of computational chemistry has seen exponential growth in recent years due to the synthesis of conventional techniques for chemical simulation and accurate, generalizable machine learning models. Large scale neural network models trained to reproduce the results of chemical simulation balance accuracy, broad applicability, and ease of use; bleeding edge models such as AlphaFold3,¹ MACE,² and UMA³ have emerged as competitive surrogates for conventional simulations.

Yet despite these advances, the extrapolation of emergent scale behaviors from these models remains elusive. Bulk phase physical properties, such as viscosity, shear strength, and tensile properties remain difficult to predict from atom-scale representations, indicating a scale-level disconnect in model extrapolation behavior. Despite the dominant data-centric paradigm of Sutton's Bitter Lesson,⁴ the incorporation of physicochemical inductive biases remains a means to rectify this gap. In this work, I have presented examples that show building physically inspired neural network architectures allows models to recapitulate fundamental chemical understanding and provides a means for end users to interpret model predictions, "opening the black box". Through an ascending order of scale, I show how higher-order properties emerge as interactions between constituent building blocks: molecular properties emerge as a combination of atom and bond properties, and intermolecular interactions emerge as a combination of molecular interactions.

Future work will explore expanding this paradigm to increasing scales of interaction complexity. In Chapter 7, ongoing work modeling ensemble behavior of molecular species

is presented. The culmination of this work will map physical properties that are influenced by complex interaction behavior of macrocyclic peptide molecules with their surrounding environments to shifts in their ensemble statistics. Other ongoing efforts seek to model the interaction behavior of compounds across chemical space, representing chemical and structural similarity as a scaffold to model inverse property-activity relationships for both property-first molecular generation and informed training of large chemical foundation models. Together, these projects will expand the scale of modeled emergent interacting behavior, allowing ever more complex relationships between compounds in chemical space to be predicted from atom-level interaction behavior.

8.1: References

1. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
2. Batatia, I. *et al.* A foundation model for atomistic materials chemistry. *J. Chem. Phys.* **163**, 184110 (2025).
3. Wood, B. M. *et al.* UMA: A Family of Universal Models for Atoms. Preprint at <https://doi.org/10.48550/arXiv.2506.23971> (2025).
4. Sutton, R. The Bitter Lesson. *Incomplete Ideas* https://www.cs.utexas.edu/~eunsol/courses/data/bitter_lesson.pdf (2019).

APPENDIX A: Supplementary Materials for Chapter 2

S1. Benchmarking Computational Methods for Dataset Generation

In order to determine the optimal quantum mechanics (QM) calculation methodology for dataset generation, detailed benchmarking studies were performed to assess both accuracy and computational cost for a variety of QM methods. Based on results from prior studies, three density functional theory (DFT) functionals (B3LYP¹ with Grimme's D3 dispersion² [B3LYPD3], ω B97X-D³, and M06-2X⁴) and two basis sets (6-31G(d)⁵ and def2TZVP⁶) were selected. In addition to these DFT methods, the CBS-QB3 composite method⁷ and DLPNO-ccSD(T)⁸ coupled cluster calculations with the ccpVDZ⁹ basis set were also benchmarked. Finally, the DM21 neural functional¹⁰ with the ccpVDZ basis set was also benchmarked due to its strength in localizing spin density on radical species. All DFT and CBS-QB3 calculations were performed using Gaussian G16 B.01¹¹ with default grid and convergence settings. DLPNO-ccSD(T) calculations were performed with Orca 5.0.2¹² with default grid and convergence settings. DM21 calculations were performed using PySCF 2.0.1 software¹³ with default grid and convergence settings.

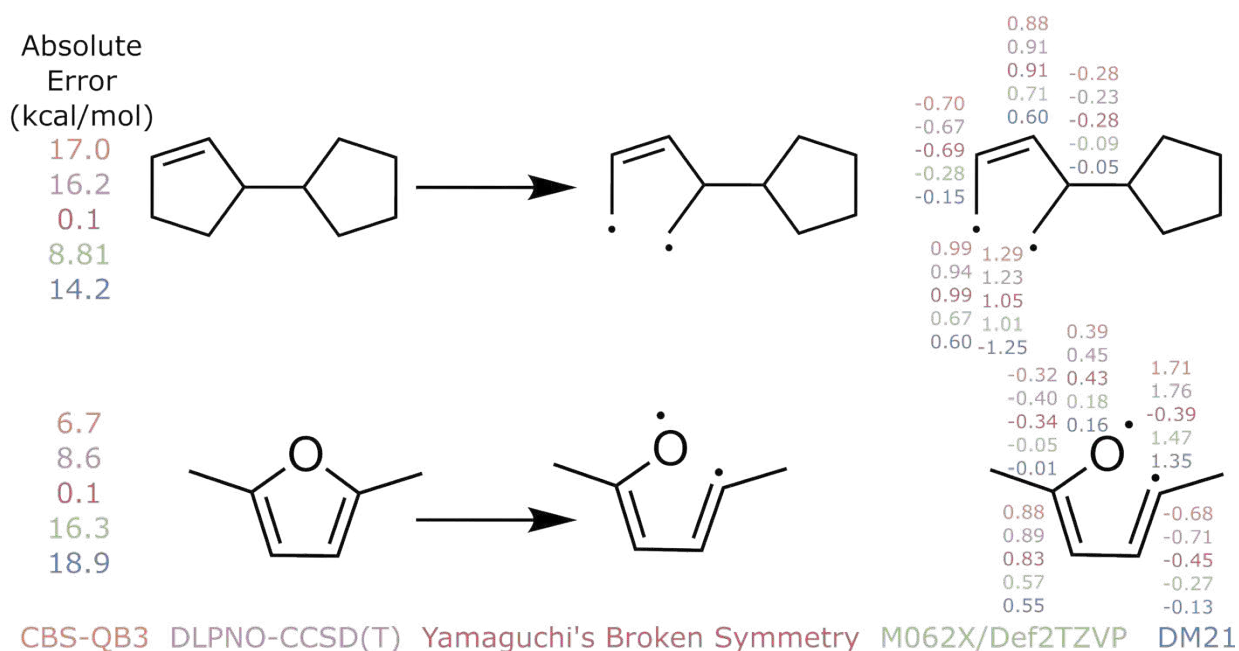


Figure S1: Mulliken spin densities for selected benchmark methods. Absolute error compared to experimental values is shown on the left.

To perform benchmarking calculations, 40 compounds with experimentally determined diradical formation enthalpies were obtained from the literature¹⁴⁻¹⁷. The ring-closed form of each compound was optimized in the singlet state and single-point energies calculated for all benchmarked levels of theory. Then, the target ring bond was deleted, forming a diradical, and the resultant geometry optimized using the unrestricted Kohn-Sham formalism as a triplet. The diradical formation enthalpy was then obtained using a single point calculation, per the following formula:

$$\Delta_f H^\circ(\text{Diradical}) = \Delta_f H^\circ(\text{Parent}) + \Delta H(\text{Ring opening})$$

For DLPNO-ccSD(T) calculations, the optimized geometries from M06-2X/def2TZVP and UM06-2X/def2TZVP calculations were used to obtain single point energies due to computational expense. The errors with respect to experimental measurements were

obtained along with calculation wall times. Since DM21 calculations were run in a jupyter notebook, accurate wall times reflective of solely the computational cost of running calculations were not obtained and are omitted here. To further study localization of the diradical spins, Mulikken spin densities were obtained for the outlier molecules shown in Figure 2a of the main text using M06-2X/def2TZVP, DM21/ccpVDZ, DLPNO-ccSD(T)/ccpVDZ and CBS-QB3 methods. A Yamaguchi broken symmetry calculation was also performed using the results from the M06-2X/def2TZVP calculation. The results of this spin density analysis are shown in Figure S1. As expected, the Yamaguchi correction values show the lowest error when compared to experimental results; however, the error shown by the selected M06-2X/def2TZVP method outperforms that of more costly post-Hartree Fock wavefunction methods or composite complete basis set extrapolation methods for one test case. Due to the computational expense of performing either the Yamaguchi broken symmetry calculation or post-Hartree Fock wavefunction methods for the entire ALFABET3 dataset, the M06-2X/def2TZVP method was adopted for full database construction.

S2. ALFABET3 Model Training

The ALFABET3 model was built on the same model architecture of the ALFABET2 model, which uses a modified version of the message passing framework introduced by Gilmer et al.¹⁸ In particular, node representations are initially updated by passing a concatenated vector of start and end node features through two dense layers separated by a ReLU activation function. This state vector is then used to update edge state using a residual connection with start and end node state. Edge states are updated by concatenating start and end atom state with the current edge state, and passing the

resultant vector through two dense layers separated by a ReLU activation function. All hidden state vectors are of dimension 128, resulting in dimensions of 256 after concatenation operations. The final BDE and BDFE predictions are obtained by passing the bond states through a dense readout layer and summing the result with a residual bond state embedding representing the per-molecule mean BDE/BDFE. Since robust hyperparameter tuning was performed during the development of the ALFABET2 model, only changes in model accuracy on changing number of layers was evaluated for hyperparameter tuning; it was found that the best performing model consisted of 6 message passing layers, mirroring that of the ALFABET2 model.

Outlier bonds were defined as test set bonds for which model predictions differed from calculated BDE values by over 10%. Using this definition, 676 model outliers were found, constituting 0.092% of the total ALFABET3 dataset; a parity plot of these results is shown in Figure S2a. In addition, it was found that several molecules contain multiple outlier bonds; a histogram of these results is shown in Figure S2b. Examples of the 16 largest outlier bonds can be seen in Figure S2d; all such bonds are present in molecules with a high degree of ring strain or uncommon substitution patterns. Finally, a histogram of outlier prediction errors is shown in Figure S2c; the standard deviation of prediction errors is found to be 16.5 kcal/mol.

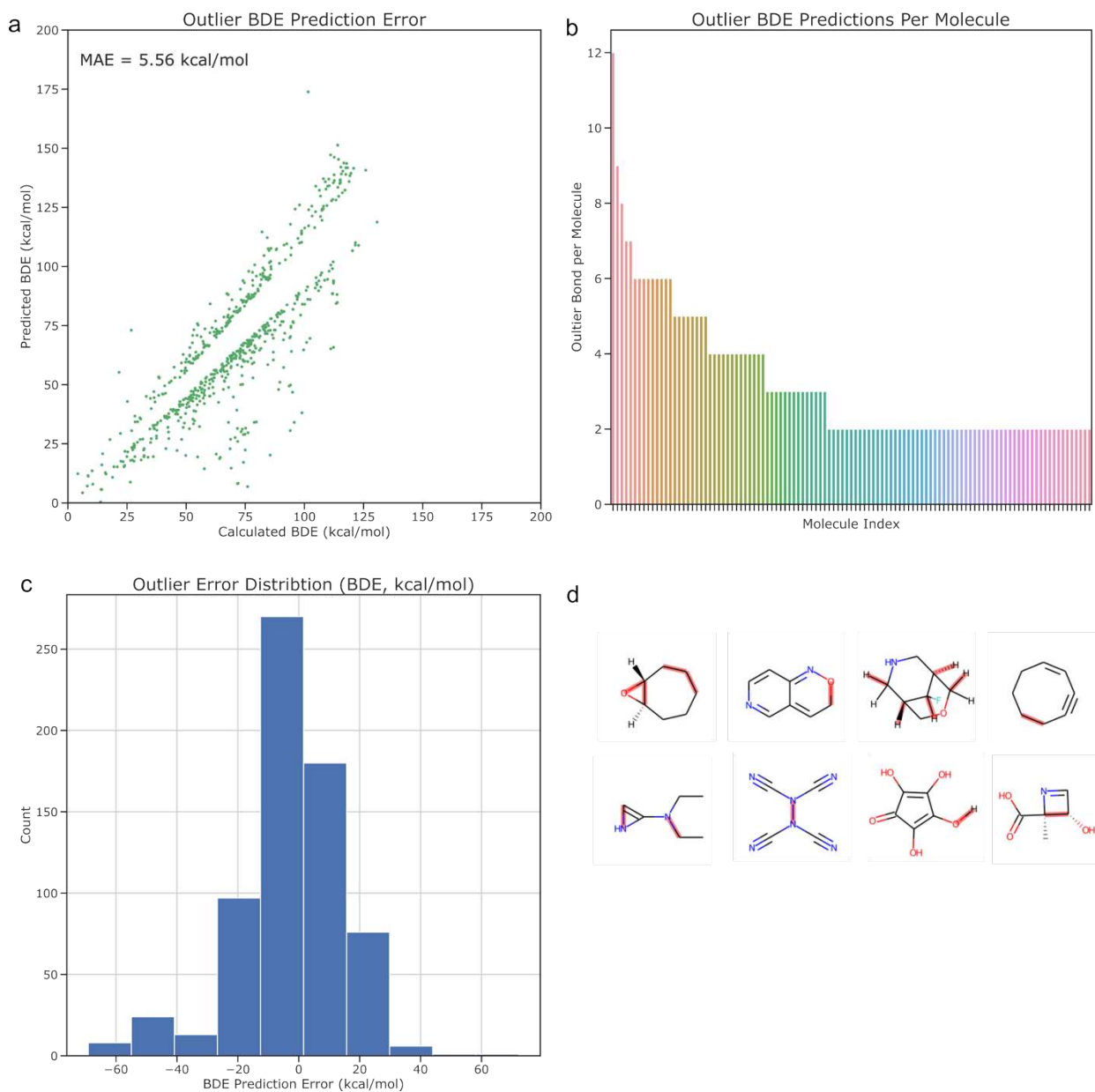


Figure S2: Statistics for model prediction outliers. a) Parity plot of predicted BDE vs. calculation results for test set model outliers (error of > 10%). b) Outlier bond counts per molecule c) Error distribution of predicted outliers d) Top 16 outlier bonds highlighted in red.

S3. Applications to Retrosynthesis Prediction

Ring systems serve as essential structural elements in medicinal chemistry and constitute fundamental patterns in numerous drugs currently available in the market. Methodologies involving ring forming reactions such as coupling reactions, substitutions

and cycloadditions allows chemists to construct cyclic systems which are prevalent in natural products and pharmaceuticals. Using a database of documented ring formations from the USPTO database, Bjerrum and coworkers built neural network based predictive model for determining retrosynthetic pathway for ring formations¹⁹. While this model would allow chemists to determine efficient synthetic routes for ring formations based on existing transformations, we would like to add a quantitative perspective by gaining insights into the bond dissociation energy of the ring bonds.

With the idea to analyze the bond dissociation energy for retrosynthetic ring opening reactions, we envision that the lowest bond dissociation would correspond to the ring forming bond. To the following hypothesis, we utilize a reduced dataset as reported by Bjerrum and coworkers. The dataset was trimmed down to 528 reactions which contain only intramolecular transformations, relating to O, S – substitutions reactions. The 528 products are encoded as graphs and the bond dissociations of all bonds in the molecule were calculated using the ALFABET3 model. A plot of the relative density of predicted BDEs for all

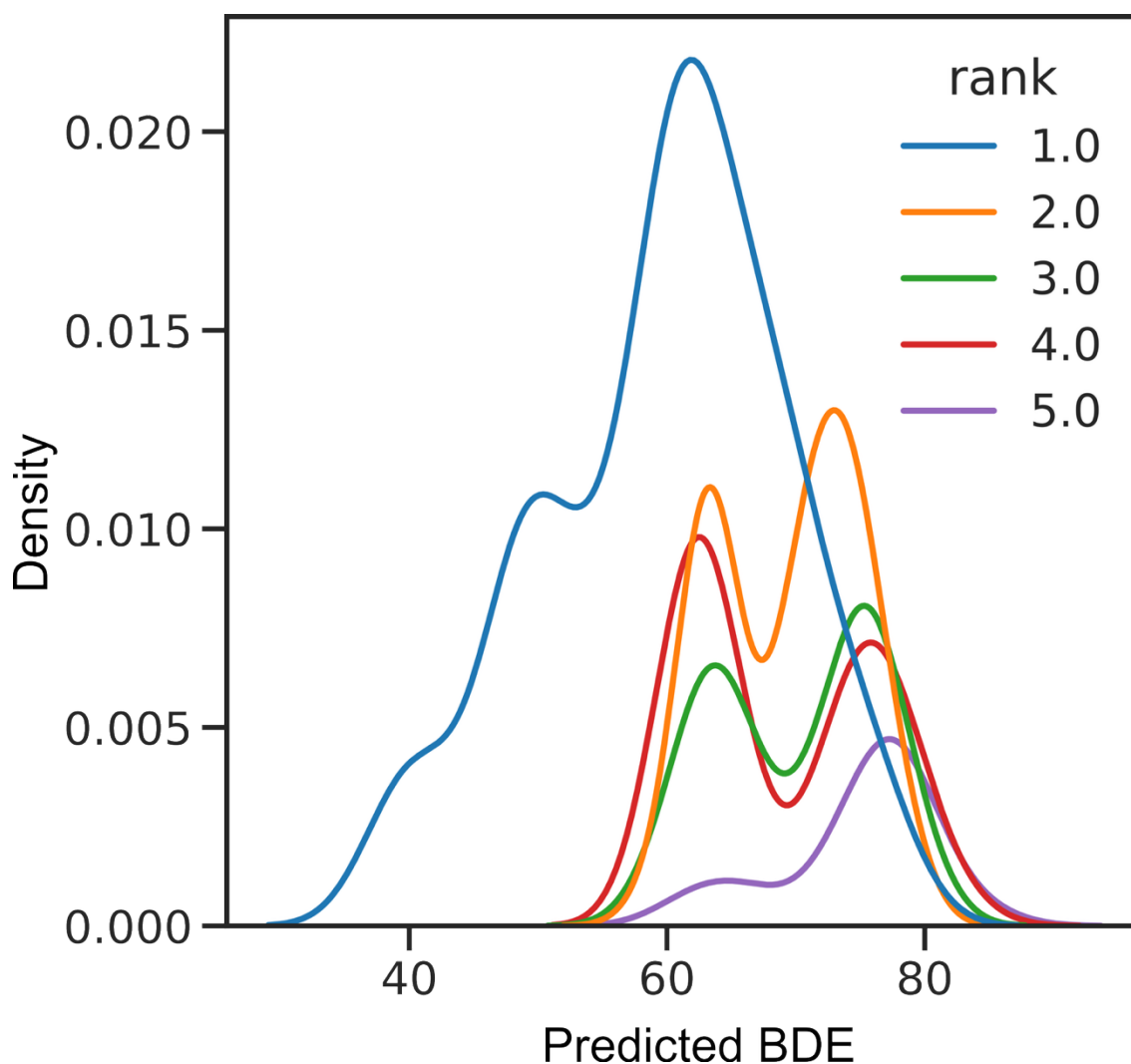


Figure S3: Density plot of predicted BDE for ring opening retrosynthesis pathways

reaction pathways is shown in Figure S3; each line corresponds to a reaction for which the n^{th} lowest BDE is the key retrosynthetic step.

As we aim to determine the intra-molecular ring breaking retrosynthetic pathway based on bond dissociation, we notice for 172 reactions the lowest BDE correspond to the proposed retrosynthetic pathway from the USPTSO database. Additionally, for 365 reactions the proposed retrosynthetic pathway is in the lowest five bond dissociations. We find that for a majority of reactions in which the proposed retrosynthetic pathway is in the lowest five bond dissociations, the lowest BDEs correspond to bonds in four/five membered rings, but do not correspond to the retrosynthetic pathway described in the USPTSO database. Given the qualitative accuracy of the lowest predicted BDE in determining retrosynthesis pathways, the bond dissociation energy provides a useful perspective to determining retrosynthetic pathway for intramolecular transformations.

APPENDIX B: Supplementary Materials for Chapter 3

S1. Database Generation: Computational Details

S1.1. ALFAST-DB and Fragments-DB

We screened the database ALFABET only for neutral molecules containing a double, triple or aromatic bond, through the usage of SMARTS and the pybel library of openbabel,[ref1] leading to a total of 57 736 molecules. As initial geometries for the triplet optimization we started from the already optimized singlet geometries (from ALFABET) and re-optimized the molecules in the triplet energy surface using Gaussian 16 at the M06-2X/def2-TZVP level with the nosymm keyword. Initially this led to a total of 39 225 calculations that converged to a minima while most of the remaining molecules failed either due to convergence during the optimization or reached a stationary point with an imaginary frequency. To automate the check of the calculation termination and to extract the geometries, atomic Mulliken spins we used the pyssian library.[ref2]

Then, we performed a second round of calculations. For this round we distorted the molecules with imaginary frequencies along the imaginary mode (the full distortion in the positive direction as determined by Gaussian). For molecules that failed at the frequency calculation stage the last geometry was used and for the other failed calculations the geometry with the lowest potential energy throughout the optimization was used. After the second round of calculations a total of 53 137 calculations had converged.

We then proceeded to screen the converged structures to ensure that the connectivity of the molecules remained the same (further detailed in the SI) which lowered

the total number of molecules to the final ALFAST-DB size, 46 415 molecules. This final geometry screen was carried out using, the openbabel and the networkx libraries. First, through openbabel's conversion from xyz format to SMILES the geometry was checked to ensure that the molecules did not dissociate into 2 different molecules during the optimization (presence of a "." in the SMILES) and the canonicalized SMILES in the database and the canonicalized SMILES of the xyz-to-SMILES conversion were tested for equality. As in some cases small variations in the bond-distance could lead to changes in the SMILES, two graph isomorphism checks were carried out using networkx. The first one, where only categorical node matches (atoms) were considered in the graph isomorphism between the graph of the initial SMILES and the graph of the SMILES obtained from the xyz-to-SMILES conversion. Then a more restrictive graph isomorphism with categorical matches for edges and nodes was considered, which provided similar results to a canonical SMILES comparison. Thus, we only removed molecules whose graph isomorphism with categorical node matching found no match between the initial SMILES and the SMILES perceived from the optimized geometry. The categories of the nodes were the atomic number of the atom it represented.

For the Fragments-DB due to the characteristics of this database ("monocore" small molecules with a restricted conformational variability) the initial geometries were generated from SMILES using openbabel's distance geometry implementation and initially relaxed using Conjugated Gradients with a stopping criteria of 350 steps or 1E-4 energy difference using MMFF94 when possible and for the molecules that failed the UFF forcefield was used instead. From these starting geometries a GFN2-XTB optimization with verytight

optimization criteria was carried out using XTB 6.7.1 [ref3] for both, the singlet and the triplet. The obtained triplet geometries were pre-screened using the same categorical node matching graph isomorphism check used for ALFAST-DB. Molecules that were not discarded were then used for the optimization at the M06-2X/def2-TZVP level using Gaussian 16 (respectively using the xtb-optimized singlet geometry as initial geometry for the singlet DFT optimization and the xtb-optimized triplet geometry for the triplet DFT optimization). At this stage, the same procedure as with ALFAST-DB was carried out.

Unfortunately, at a later stage in the study we found that a small subset of molecules that had undergone an H-transfer during the optimization were included in the curated dataset. We found that it is due to the initial SMILES not including explicit H atoms, therefore its graph representation would not have H-atoms and would be treated as a subgraph of the graph obtained from the optimized geometry. As these amount to a very small quantity of compounds (all corresponding a subset of the molecules with a main fragment with a total Mulliken of ~ 1 , see Figure 4a in the manuscript) relative to the database size we did not repeat the full study as we predicted them to have a low impact on the model performance.

S1.2. Verde-DB Extension

Our objective was to broaden the scope of our database by including light-responsive π -conjugated organic molecules with potential applications in organic redox flow batteries. To achieve this, we integrated new core structures including quinoline-5,8-dione and quinoxaline-5,8-dione, which feature systematic functionalization at each open site. We

utilized an in-house algorithm that systematically links 20 spacer and 11 terminal groups to generate the virtual library in line with the VERDE materials DB.

S2. Metrics and cross-validation schemes

To evaluate the performance of the models on unseen data we used a 80:20 train-validation:test split and we used a five-fold cross validation to select the optimal hyperparameters of each MPGNN architecture explored (ultimately being a training:validation ratio of 64:16, relative to the full dataset, per each fold). To ensure that the comparison of the results of each fold across the different models we decided to always use the same fold partitions. The splits of the combined ALFAST-DB and Fragments-DB were the combination of the respective splits of each database.

The metrics used to evaluate the performance of the models were the Mean Absolute Error (MAE), and the R^2 correlation coefficient between the predictions and the true values. These metrics were used as implemented in the tensorflow library and the general equations are shown below, where n is the number of samples y_{true} and y_{pred} are the true and predicted values of each sample.

$$1) MAE = \frac{\sum_i^n |y_{i,true} - y_{i,pred}|}{n}$$

$$2) R^2 = 1 - \frac{\sum_i^n (y_{i,true} - y_{i,pred})^2}{\sum_i^n (y_{i,true} - \overline{y_{true}})^2}$$

When cross-validation results are provided, they are accumulated across folds. For example, in the case of the MAE, per each fold's validation set the predictions and the absolute errors are

determined using only the model trained with the fold's training set, then the total absolute error per fold is added across folds and finally divided by the size of the dataset, mathematically:

$$3) MAE_{cv} = \frac{\sum_{k=0}^{folds} \sum_i^{n_k} |y_{i,true} - y_{i,pred}|}{n}$$

Finally, when test set metrics are provided they are the mean of the metric across different folds:

$$4) \overline{MAE_{Test}} = \frac{\sum_k^{folds} MAE_{Test,k}}{folds}$$

S3. Architecture Selection

S3.1. Candidate architectures

S3.1.1 Non-standard Layers and Operations Description

Dense Concatenation Layer: this layer is a composite layer (nfp.ConcatDense) that adds more weights to the model. First a concatenation in the axis of the number of features is carried out (which involves no new weights). As the nodes, bonds and global state are always considered the same size (embedding dimension), at least it duplicates the number of features. The outcome of the concatenation is fed to a Dense layer of size “2 x embedding dimension” with a ReLU activation and bias, whose outcome is then fed to a Dense layer of the same size of the embedding dimension with a lineal activation (in other words, no activation function) and bias. These last two Dense layers are the ones actually including new weights into the model.

Reduce Layer: This layer is a composite layer (nfp.Reduce) involving a reduction operation (in our case always a summation) that does not add any new weights to the model. It acts as a wrapper to allow the reduction of the messages for each node, without a self-update of each node, which has to be done afterwards.

Node Update: a node update operation (`nfp.NodeUpdate`) is a sequence of operations that overall updates the node states. First, a gather (`nfp.Gather`) layer is used to gather the end nodes of each edge in the graph, which are the sources of the messages. Edges in the graph in this context are always directional, and each bond is represented by two edges, one per direction. Next, a Dense Concatenation Layer between the edge states and the gathered end nodes is included, obtaining a message per edge. These are fed to a Reduce Layer coupled with the node states and the connectivity to aggregate the messages that each node receives. The aggregated messages are then fed sequentially to a Dense layer of size “2 x embedding dimension” with a ReLU activation and bias and then to a Dense layer with a lineal activation and bias obtaining a final form of the messages. These messages which are of the same dimensions as the node states are added to the node states to account for the self-update of the node. The outcome of all these steps is the updated state of the nodes. When a Global state is also passed as input, it is tiled to match the dimensions and also fed to the Dense Concatenation Layer.

Bond Update: a bond update operation (`nfp.BondUpdate`) is a sequence of operations that overall updates the bond states. First a gather (`nfp.Gather`) operation is used to gather the start nodes of each edge as well as the end nodes of each edge, which are then subsequently fed to a Dense Concatenation Layer with the bond state, whose outcome is then added to the bond state to obtain the updated bond state. When a Global state is also passed as input, it is tiled to match the dimensions and also fed to the Dense Concatenation Layer.

Global Update: When no previous Global state is defined, this operation pools the information from the nodes and edges of the graph generating an initial graph embedding of the same dimension as the bonds or nodes. When used to update the Global state, as with the Node Update or Bond Update, a final step involving a self-update through an addition operation is carried out. In this operation (`nfp.GlobalUpdate`) an attention mechanism is built into the pooling from the graph elements, thus the number of attention heads and the number of units needs to be specified and its product should match the embedding dimension used for the bonds and the atoms. We specifically used 1 attention head, and therefore the number of units and the embedding dimension are equal. First, in this operation nodes and bond states are concatenated (note that this is not a Dense Concatenation, but a concatenation across the dimension corresponding to the number of atoms and number of bonds instead of concatenating along the axis of the features). The outcome of this operation is the Graph Element state tensor, which is first fed to a query layer (a Dense layer of size 1 with bias and no activation function) and subsequently subjected to a softmax operation generating the attentions. Next, the Graph Element state tensor is also fed to a value layer (a Dense layer of the size of the embedding dimension with bias and no activation function) generating the values. The matrix multiplication of the values and the transposed attentions generates the final weighted graph element values are added together generating a global state update vector of the same dimension as the embedding dimension. Finally, if this operation is used to update a previously existing global state, an addition operation is carried out.

Masked Global Update: Exactly like Global Update but it takes as input also a list of node indices and edge indices with which to mask the attentions to ensure that only those nodes and edges contribute to the new global state. This is not an existing layer in the nfp package and the implementation is available with all the code associated to the present work available in our github.

Message Passing operation: A message passing operation, here, involves a Bond Update, Node Update and Global Update in this specific order to update the previous bond states, node states and global state.

S3.1.2 General Architecture

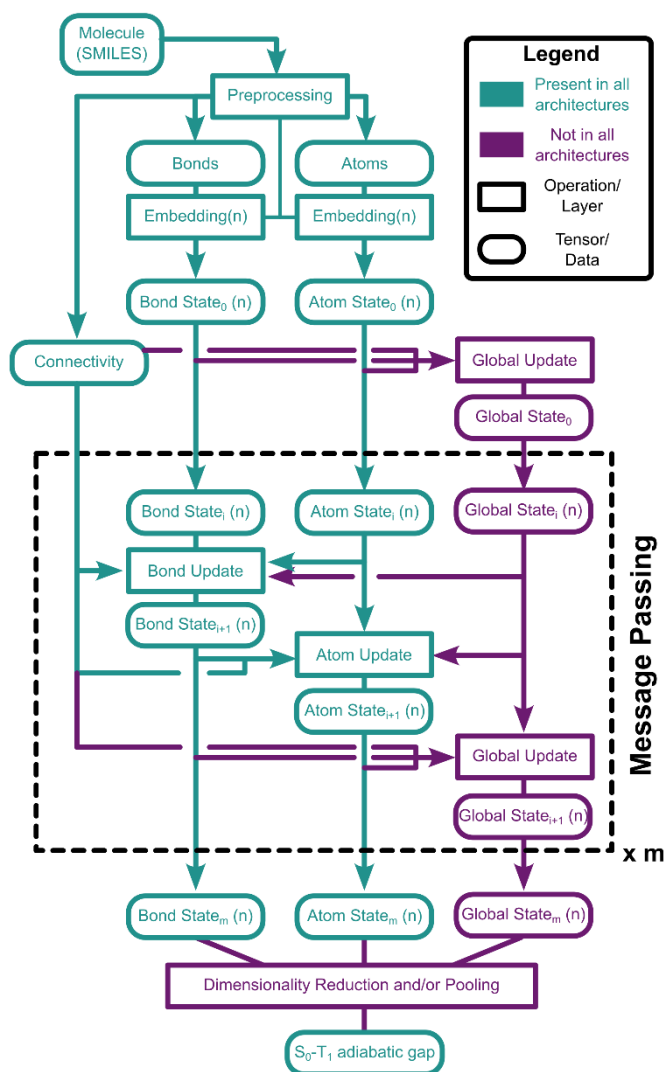


Figure S1: Diagram illustrating the common part of all the Message Passing Graph Neural Network architectures explored in the present work. “ n ” stands for the embedding dimension and “ m ” for the number of message-passing iterations.

A depiction of the general message passing architecture is used as the base for the explored model variations is shown above (Figure S1). The preprocessing used was the same across the architectures. It was implemented using the nfp SmilesBondIndexPreprocessor and the atom and bond features used for the preprocessor were:

- Atom features
 - Atom Symbol.
 - Number of Radical Electrons.
 - Formal Charge.
 - Chiral Tag.
 - Whether it is an aromatic atom or not.
 - The size of the smallest ring it belongs to.
 - Degree (Number of directly bonded neighbors)
 - The total number of H atoms bonded to.
- Bond features (each bond is represented by two directed edges one starting on each atom involved in the bond)
 - Atomic Symbol of the starting atom
 - Atomic Symbol of the end atom
 - Type of bond (single, aromatic, double or triple)
 - Whether it belongs to a ring or not.
 - Smallest ring the bond belongs to.

S3.1.3 Architectures

Architecture diagrams of the pooling and dimensionality reduction sections for models detailed in section S3.2 are provided below. “n” stands for the embedding dimension and “m” for the number of message passing iterations

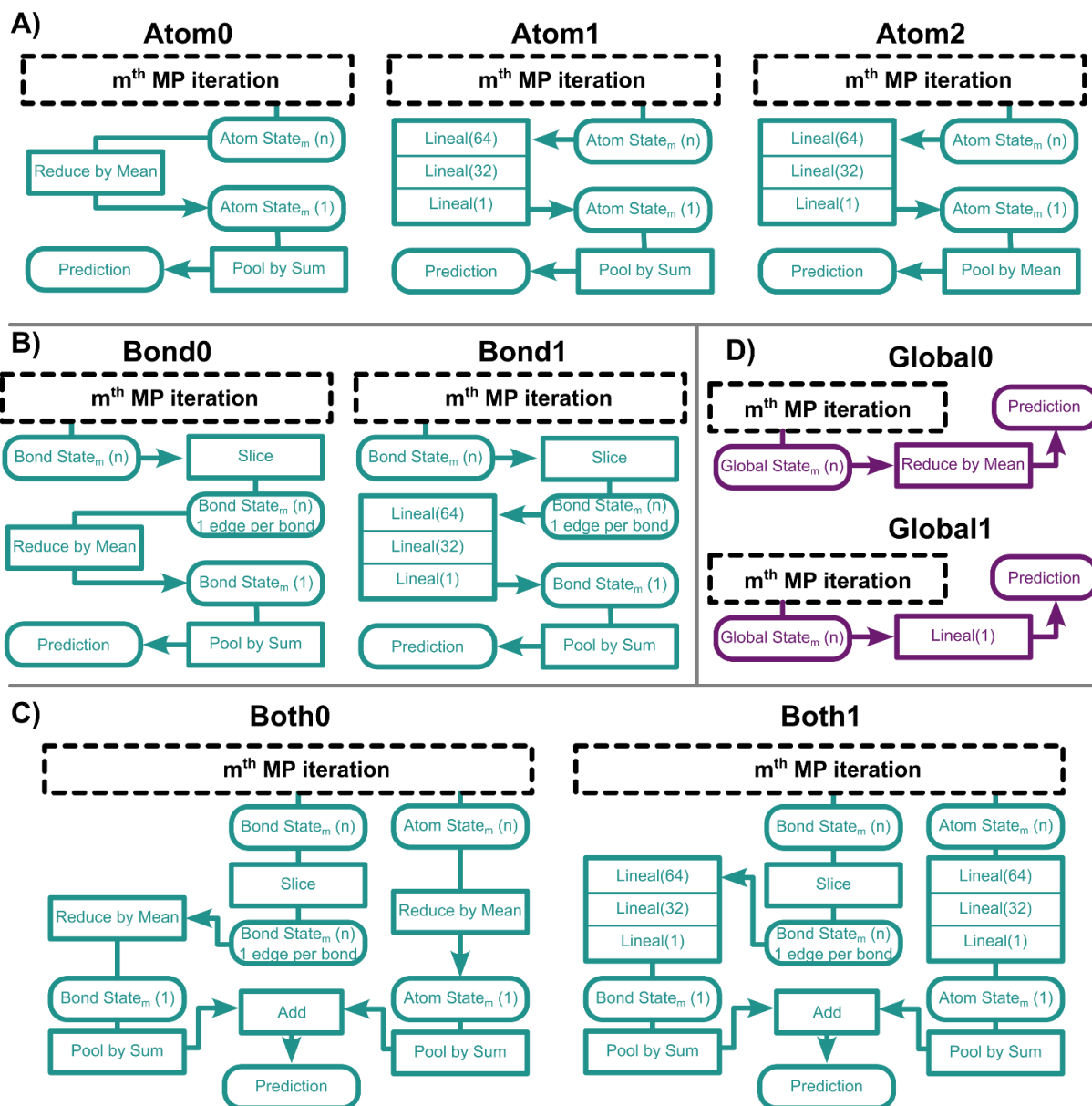


Figure S2: A) Atom pooling architectures. Predictions are generated by pooling node feature representations. B) Bond pooling architectures. Predictions are generated by pooling edge feature representations. C) "Both" pooling architectures. Predictions are generated using a sum of pooled node and edge representations. D) Global pooling architectures. Predictions are generated by doing a dimensionality reduction of the global states which pool the atoms and bonds data during the Message Passing iterations. MP stands for message passing. The number between brackets correspond to the output dimension. The same legend of Figure S1 applies.

S3.1.4 Delta Learning Model Architecture

The proposed delta learning model architecture and training was adapted from the Global1 architecture. Its architecture is almost identical, except for the presence of two extra global states, which are updated in parallel to the global state (See Figure S6). In this manner we have the original global state, to represent the full molecule and to interact in the updates of edges and nodes. This global state is always updated with a Global Update layer (nfp.GlobalUpdate). The two extra global states, one to represent the “core” (exciton location) and one to represent the “noncore” (remaining set of atoms and bonds) are instead updated using a Masked Global Update. After the message passing layers, the “core” and “noncore” states are fed each to a Lineal layer of size 1 to obtain respectively the “core energy gap” and the “noncore energy correction”. These two values are finally added together to obtain the final prediction. Further details on the custom training of this model are presented in the Delta Learning Model Details.

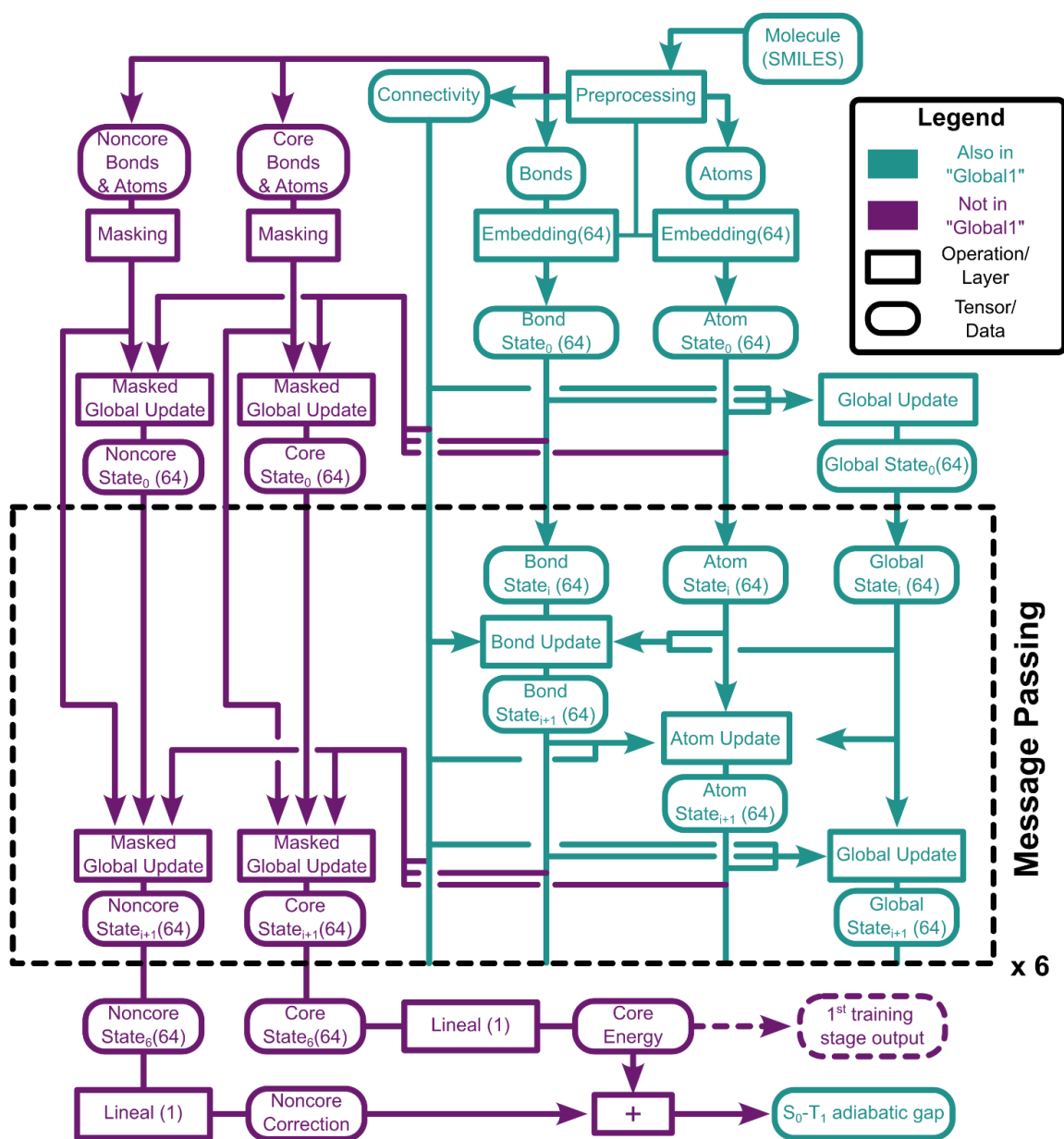


Figure S3: Diagram illustrating the final delta-fragment model architecture used in the present work. Numbers between brackets stand the embedding dimension used or output dimension of an operation/layer.

The preprocessor used for the training of this architecture is derived from nfp SmilesPreprocessor and was adapted to take as input the SMILES of the target molecule and the SMILES or SMARTS of the core as inputs. As for the target molecule, this preprocessor also builds a graph for the core and after obtaining the atom indices of the target molecule that match the core it finds a atom and bond mapping across both graphs and extracts a list

of node indices and edge indices in the core as well as another list of node indices and edge indices for the atoms not present in the core. A special argument “full” may be passed as core, to ensure that all atoms and bonds will be considered as part of the core. This special argument was included to simplify the first stage of the training.

S3.2. Hyperparameter tuning

The results of hyperparameter optimization are shown below. Selected hyperparameters for the final model are highlighted. All optimizations were carried out using an AdamW optimizer with an Inverse Time Decay learning rate scheduler and a constant weight decay rate. The batchsize was kept constant at 128 (molecules per batch). In all cases the loss function as the MAE. During model training the validation loss was tracked and the model weights with the best validation loss were also saved, and the values of such model are shown as “Best Loss” in Tables S1 and S2.

Table S1: Architecture selection. The initial learning rate was kept constant at 1E-3, the learning rate decay was kept at 1E-5 and the weight decay was kept at 1E-5.

Architecture	MP Layers	Embedding Dimension	5-fold mean values, kcal/mol			
			Final Loss (epoch 450)		Best Loss	
			Training	Validation	Training	Validation
atom0	2	16	2,85	3,53	2,89	3,42
		32	2,23	3,27	2,23	3,15
		64	1,76	3,10	1,79	3,02
		128	1,65	3,16	1,68	3,07
	3	16	2,25	3,05	2,25	3,00
		32	1,46	2,62	1,49	2,58
		64	1,12	2,62	1,11	2,50
		128	1,01	2,61	1,06	2,54
	4	16	1,90	2,80	1,94	2,74
		32	1,27	2,58	1,31	2,48
		64	1,00	2,48	1,03	2,42
		128	0,94	2,56	0,98	2,43
	5	16	1,62	2,67	1,66	2,55

		32	1,16	2,44	1,20	2,35
		64	0,88	2,35	0,93	2,32
		128	0,96	2,49	0,99	2,43
	6	16	1,56	2,67	1,60	2,52
		32	1,05	2,37	1,06	2,25
		64	0,83	2,30	0,88	2,25
		128	0,93	2,44	0,99	2,28
atom1	2	16	2,48	3,26	2,49	3,17
		32	1,91	3,10	1,94	2,94
		64	1,60	2,98	1,65	2,92
		128	1,56	3,02	1,60	2,97
	3	16	1,92	2,80	1,95	2,70
		32	1,37	2,53	1,42	2,46
		64	1,02	2,46	1,07	2,38
		128	1,04	2,59	1,09	2,52
	4	16	1,67	2,71	1,74	2,56
		32	1,17	2,45	1,21	2,33
		64	0,89	2,36	0,95	2,27
		128	0,90	2,44	0,93	2,37
	5	16	1,57	2,64	1,63	2,49
		32	1,05	2,37	1,10	2,23
		64	0,80	2,26	0,90	2,19
		128	0,84	2,29	0,87	2,25
	6	16	1,46	2,48	1,47	2,40
		32	0,97	2,25	1,00	2,18
		64	0,79	2,27	0,81	2,14
		128	0,79	2,27	0,80	2,17
atom2	2	16	2,50	3,23	2,51	3,03
		32	2,02	2,92	2,06	2,82
		64	1,71	2,92	1,84	2,76
		128	1,62	2,88	1,63	2,81
	3	16	1,93	2,61	1,95	2,53
		32	1,43	2,42	1,53	2,35
		64	1,16	2,31	1,18	2,26
		128	1,06	2,38	1,23	2,27
	4	16	1,75	2,52	1,76	2,38
		32	1,31	2,34	1,34	2,24
		64	1,08	2,30	1,15	2,18
		128	1,03	2,34	1,08	2,18
	5	16	1,64	2,46	1,66	2,34
		32	1,26	2,26	1,32	2,20
		64	1,04	2,23	1,10	2,12
		128	1,02	2,29	1,08	2,13
	6	16	1,58	2,43	1,61	2,31
		32	1,19	2,28	1,25	2,17

		64	1,01	2,17	1,09	2,09
		128	1,01	2,25	1,02	2,10
bond0	2	16	4,95	5,82	4,92	5,69
		32	4,29	5,34	4,29	5,21
		64	3,72	4,90	3,69	4,84
		128	3,44	5,07	3,49	4,82
	3	16	3,85	4,78	3,82	4,62
		32	2,84	4,13	2,89	3,92
		64	2,11	3,70	2,10	3,49
		128	2,05	3,60	2,09	3,50
	4	16	3,41	4,43	3,44	4,29
		32	2,43	3,69	2,45	3,59
		64	1,81	3,37	1,79	3,19
		128	1,74	3,26	1,70	3,21
	5	16	3,19	4,26	3,14	4,07
		32	2,19	3,59	2,17	3,42
		64	1,67	3,16	1,67	3,10
		128	1,62	3,36	1,65	3,15
	6	16	2,83	3,91	2,88	3,86
		32	1,99	3,41	2,00	3,27
		64	1,57	3,20	1,59	3,07
		128	1,62	3,20	1,64	3,12
bond1	2	16	4,64	5,48	4,64	5,42
		32	4,15	5,23	4,10	5,07
		64	3,59	5,09	3,63	4,91
		128	3,58	5,01	3,61	4,97
	3	16	3,49	4,79	3,56	4,35
		32	2,60	3,94	2,55	3,71
		64	2,11	3,73	2,10	3,52
		128	2,41	4,02	2,43	3,84
	4	16	3,17	4,10	3,18	4,04
		32	2,10	3,38	2,18	3,35
		64	1,77	3,25	1,78	3,18
		128	2,04	3,52	2,07	3,46
	5	16	2,89	3,89	2,91	3,80
		32	1,94	3,41	1,96	3,20
		64	1,70	3,10	1,67	3,01
		128	1,95	3,44	2,02	3,32
	6	16	2,71	3,83	2,70	3,65
		32	1,78	3,14	1,73	2,94
		64	1,49	3,14	1,51	2,86
		128	1,94	3,40	1,98	3,28
both0	2	16	3,16	3,93	3,17	3,77
		32	2,31	3,52	2,34	3,31
		64	1,86	3,24	1,89	3,21

	3	128	1,95	3,47	1,95	3,41
		16	2,28	3,15	2,29	3,07
		32	1,46	2,74	1,51	2,66
		64	1,18	2,73	1,20	2,66
	4	128	1,19	2,87	1,25	2,80
		16	1,96	2,99	1,92	2,76
		32	1,30	2,74	1,32	2,57
		64	0,97	2,63	0,99	2,53
		128	0,99	2,70	1,00	2,60
	5	16	1,74	2,67	1,77	2,65
		32	1,15	2,67	1,19	2,40
		64	0,93	2,58	0,94	2,48
		128	0,98	2,59	1,01	2,51
	6	16	1,59	2,69	1,61	2,54
		32	1,06	2,44	1,11	2,36
		64	0,88	2,43	0,91	2,35
		128	0,99	2,59	1,05	2,48
both1	2	16	2,65	3,41	2,65	3,32
		32	1,97	3,06	2,00	2,98
		64	1,74	3,11	1,72	2,99
		128	1,74	3,38	1,76	3,15
	3	16	1,94	2,94	1,99	2,75
		32	1,39	2,59	1,42	2,47
		64	1,06	2,48	1,07	2,41
		128	1,08	2,59	1,11	2,49
	4	16	1,69	2,61	1,68	2,53
		32	1,20	2,45	1,22	2,36
		64	0,91	2,38	0,94	2,26
		128	0,93	2,43	0,93	2,32
	5	16	1,58	2,63	1,59	2,50
		32	1,08	2,43	1,09	2,27
		64	0,85	2,23	0,86	2,20
		128	0,86	2,39	0,88	2,24
	6	16	1,48	2,55	1,51	2,41
		32	1,00	2,25	1,07	2,19
		64	0,79	2,27	0,83	2,14
		128	0,84	2,25	0,87	2,18
global0	2	16	1,94	2,59	1,98	2,49
		32	1,38	2,36	1,41	2,26
		64	1,02	2,43	1,05	2,22
		128	0,84	2,29	0,92	2,18
	3	16	1,63	2,40	1,66	2,28
		32	1,11	2,18	1,20	2,08
		64	0,77	2,08	0,84	2,04
		128	0,63	2,17	0,67	2,00

	4	16	1,49	2,34	1,50	2,14
		32	0,98	2,10	1,06	2,03
		64	0,66	2,03	0,74	1,93
		128	0,61	1,99	0,65	1,89
	5	16	1,39	2,16	1,46	2,09
		32	0,86	2,08	0,94	1,94
		64	0,58	1,95	0,63	1,88
		128	0,54	1,98	0,58	1,84
	6	16	1,29	2,14	1,33	2,06
		32	0,79	2,04	0,82	1,92
		64	0,55	1,93	0,66	1,83
		128	0,53	1,89	0,56	1,82
global1	2	16	1,64	2,37	1,66	2,26
		32	1,11	2,18	1,19	2,10
		64	0,74	2,05	0,92	2,00
		128	0,64	2,04	0,65	1,97
	3	16	1,48	2,24	1,53	2,13
		32	0,96	2,14	1,11	2,00
		64	0,65	2,04	0,72	1,90
		128	0,53	1,90	0,58	1,86
	4	16	1,39	2,34	1,43	2,09
		32	0,80	2,08	0,86	1,92
		64	0,54	1,88	0,58	1,83
		128	0,50	1,83	0,53	1,79
	5	16	1,25	2,08	1,29	2,03
		32	0,71	1,96	0,81	1,89
		64	0,52	1,83	0,59	1,79
		128	0,46	1,81	0,48	1,75
	6	16	1,15	2,13	1,22	2,01
		32	0,70	1,88	0,80	1,83
		64	0,48	1,79	0,58	1,76
		128	0,46	1,82	0,48	1,73

Table S2: Learning Rate optimization table. Embedding dimension was kept constant at 128 and the Global1 architecture was used.

MP Layers	Epochs	Learning Rate	Learning Rate Decay	Weight Decay	5-fold mean values, kcal/mol			
					Final Loss (Last epoch)		Best Loss	
					Training	Validation	Training	Validation
2	1000	0,00001	0,00001	0,000001	2,09	2,61	2,09	2,55
				0,00001	3,54	3,73	3,33	3,43
				0,0001	6,86	6,95	4,80	4,86
	450	0,00001	0,00001	0,000001	2,31	2,76	2,31	2,72

				0,00001	3,41	3,58	3,39	3,48
				0,0001	5,44	5,57	5,06	5,11
		0,0001	0,00001	0,000001	0,83	2,15	0,98	2,12
				0,00001	1,30	2,24	1,38	2,17
				0,0001	3,12	3,43	3,11	3,20
			0,1	0,000001	10,62	10,65	10,32	10,35
				0,00001	72,89	72,89	11,48	11,48
				0,0001	72,97	72,97	13,68	13,66
			0,3	0,000001	28,12	28,17	14,95	14,94
				0,00001	72,94	72,94	37,32	37,28
				0,0001	72,97	72,97	67,79	67,76
		0,001	0,00001	0,000001	0,53	2,08	0,61	1,99
				0,00001	0,63	2,08	0,67	1,97
				0,0001	1,50	2,43	1,54	2,09
			0,1	0,000001	4,66	4,74	4,66	4,74
				0,00001	8,75	8,79	5,07	5,13
				0,0001	72,97	72,97	6,63	6,68
			0,3	0,000001	6,81	6,86	6,77	6,82
				0,00001	71,85	71,86	7,73	7,76
				0,0001	72,97	72,97	8,75	8,76
6	450	0,00001	0,00001	0,000001	1,25	2,31	1,30	2,23
				0,00001	2,81	3,02	2,51	2,72
				0,0001	5,82	5,94	5,16	5,20
		0,00001	0,00001	0,000001	1,36	2,31	1,39	2,22
				0,00001	2,61	3,10	2,48	2,71
				0,0001	5,26	5,35	5,15	5,19
		0,0001	0,00001	0,000001	0,36	1,84	0,46	1,79
				0,00001	0,74	1,97	0,84	1,86
				0,0001	2,41	2,84	2,40	2,38
			0,1	0,000001	7,92	7,97	6,74	6,82
				0,00001	72,85	72,85	8,42	8,47
				0,0001	72,97	72,97	9,87	9,81
			0,3	0,000001	18,24	18,31	9,94	9,97
				0,00001	72,93	72,93	10,55	10,56
				0,0001	72,97	72,97	12,46	12,38
		0,001	0,00001	0,000001	0,42	1,80	0,45	1,75
				0,00001	0,45	1,77	0,45	1,72
				0,0001	1,26	2,19	1,27	1,89
			0,1	0,000001	3,25	3,41	3,16	3,34
				0,00001	8,03	8,07	3,98	4,07
				0,0001	72,97	72,97	5,34	5,37
			0,3	0,000001	4,71	4,79	4,55	4,64
				0,00001	69,10	69,14	5,81	5,86
				0,0001	72,97	72,97	7,72	7,71

S3.3. Final Model Metrics

The computed metrics of the selected architecture across the different datasets shown in the manuscript are presented below. The trainings followed the same protocol as the one indicated in S3.2. The presented results correspond to the “best” model weights during training, determined as the ones with the lowest validation loss value. In the case of the “final” models, the weights that produced the lowest loss value during training were selected. Here by “final” models we refer to the selected architecture trained with the combined training and validation data splits as training set of each database respectively while still keeping the test set data unseen during training (and only used for evaluation after the training).

Table S3: Computed metrics of the selected architecture with the chosen hyperparameters. Folds 0-4 are the individual results of each one of the individual training:validation folds. CV entries contain the averaged value across folds except for the validation set () which are the aggregated metrics instead of the averaged ones. The “final” entries are the models trained using the training and validation as training set, therefore no validation metrics were computed.*

MAE in kcal/mol		Training Set		Validation Set		Test Set	
Dataset	CV Fold	MAE	R ²	MAE	R ²	MAE	R ²
alfast	0	1,24	0,9726	1,97	0,9350	2,02	0,9275
	1	1,17	0,9744	1,95	0,9324	1,98	0,9274
	2	1,20	0,9718	1,92	0,9340	1,98	0,9277
	3	1,33	0,9685	1,94	0,9346	2,03	0,9265
	4	1,17	0,9747	1,90	0,9340	2,00	0,9268
	CV	1,22	0,9724	1,94*	0,9340*	2,00	0,9272
alfast	final	1,25	0,9682	--	--	1,94	0,9280
ent	0	0,96	0,9875	2,02	0,9303	2,03	0,9330
	1	0,89	0,9890	2,06	0,9261	2,08	0,9297
	2	1,14	0,9817	2,04	0,9323	2,05	0,9339
	3	1,15	0,9818	2,02	0,9358	2,07	0,9292
	4	1,04	0,9847	2,06	0,9314	2,07	0,9310
	CV	1,04	0,9849	2,04*	0,9312*	2,06	0,9314
ent	final	1,50	0,9774	--	--	2,37	0,9205
verde	0	0,84	0,9765	1,17	0,9548	1,20	0,9590
	1	0,63	0,9886	1,12	0,9684	1,12	0,9644
	2	0,85	0,9757	1,16	0,9578	1,13	0,9611
	3	0,70	0,9860	1,07	0,9586	1,11	0,9642
	4	0,44	0,9946	1,18	0,9608	1,16	0,9599
	CV	0,69	0,9843	1,14*	0,9604*	1,14	0,9617

verde	final	0,45	0,9950	--	--	1,27	0,9564
fragments	0	0,95	0,9783	2,68	0,8502	3,05	0,8287
	1	1,27	0,9579	2,98	0,8283	3,30	0,8048
	2	1,29	0,9600	3,34	0,7811	3,30	0,8008
	3	1,39	0,9453	3,29	0,7977	3,38	0,7759
	4	1,31	0,9569	3,07	0,8361	3,27	0,7987
	CV	1,24	0,9597	3,07*	0,8180*	3,26	0,8018
fragments	final	0,90	0,9813	--	--	2,98	0,8309
alfast- fragments	0	1,29	0,9664	1,92	0,9321	1,99	0,9246
	1	1,16	0,9730	1,92	0,9301	1,99	0,9272
	2	1,09	0,9775	1,91	0,9347	1,95	0,9293
	3	1,22	0,9692	1,94	0,9318	2,00	0,9241
	4	1,25	0,9721	1,93	0,9384	2,04	0,9265
	CV	1,20	0,9716	1,92*	0,9334*	1,99	0,9263
alfast- fragments	final	1,53	0,9674	--	--	2,21	0,9254

S3.4. Delta Learning Model Details

The training of this architecture, is also adapted from the training of the Global1 architecture. A two stage training is carried out. In the first stage of the training all weights associated to the noncore global state are frozen, and the “core energy” is taken as the model output and directly compared with the training S_0 - T_1 adiabatic gap values of the Fragments-DB (see Figure S3). In the second stage of the training all weights are allowed to optimized and the outcome of “core energy” + “noncore correction” is taken as the model output and compared again, against the S_0 - T_1 adiabatic gap values.

- **Stage 1: Core training**
 - **Preprocessor:** trained on training subset of the combined ALFAST-Fragments DB
 - **Training Dataset:** Fragments-DB training and validation sets used during training. The preprocessing is tweaked to ensure that all atoms, including H, are labeled as being part of the core.
 - **Optimizer:** AdamW
 - Inverse Time Decay Scheduler
 - Initial LR: 1E-3
 - LR Decay: 1E-5
 - Weight Decay: 1E-4
 - **Loss:** Mean Absolute Error

- **Epochs:** 200
- **Batch size:** 128
- **Stage 2:** Delta learning
 - **Preprocessor:** Same preprocessor
 - **Training Dataset:** The combined ALFAST-DB and Fragments-DB training and validation sets. The cores provided to the model in the case of ALFAST are the SMARTS of the main fragments while for Fragments are the full SMILES of the molecule (which is a valid SMARTS)
 - **Optimizer:** AdamW
 - Inverse Time Decay Scheduler
 - Initial LR: 1E-3
 - LR Decay: 1E-5
 - Weight Decay: 1E-4
 - **Loss:** Mean Absolute Error
 - **Epochs:** 450
 - **Batch size:** 128

Table S4: Computed metrics of the delta-learning based model. Folds 0-4 are the individual results of each one of the individual training:validation folds. CV entries contain the averaged value across folds except for the validation set (*) which are the aggregated metrics instead of the averaged ones. The “final” entries are the models trained using the training and validation as training set, therefore no validation metrics were computed.

MAE in kcal/mol		Training Set		Validation Set		Test Set	
Dataset	CV Fold	MAE	R ²	MAE	R ²	MAE	R ²
alfast-fragments	0	1,15	0,9743	1,90	0,9358	1,95	0,9275
	1	1,16	0,9764	1,96	0,9334	2,01	0,9274
	2	1,14	0,9748	1,90	0,9339	1,95	0,9277
	3	1,16	0,9766	1,97	0,9296	1,93	0,9265
	4	1,49	0,9720	2,19	0,9313	2,26	0,9268
	CV	1,22	0,9748	1,98*	0,9328*	2,02	0,9280
alfast-fragments	final	1,67	0,9639	--	--	2,26	0,9256

S4. Fragmentation analysis extended

S4.1. Database Comparison Details

Comparison of the chemical space overlap between EnT-DB and ALFAST-DB was performed via a comparison of the SMILES string representations of the constituent molecules. Constituent SMILES were canonicalized using rdkit’s MolToSmiles and MolFromSmiles functions. Then, an inner join of both database sets on the canonicalized SMILES strings was performed using Python. Figure S7 was obtained from the subset of common molecules determined through this procedure.

When fragment overlap analysis was carried out, each molecule was fragmented using the algorithm described in the main text, translated into SMILES and any duplicate of these fragments was discarded. Then the overlap between the unique pool of fragments of each dataset was assessed following the same procedure as for molecules.

S4.2. Main Spin Labeling: ALFAST-DB and Fragments-DB

For the molecules where we have the atomic Mulliken spin density at the triplet state (specifically, molecules in ALFAST-DB and Fragments-DB), the steps 1 and 2 of the fragmentation algorithm described in Section “Molecular Fragmentation Algorithm” (Figure 2 in the main text), were carried out and followed by adding together the atomic Mulliken spin density of each atom in the fragment, and the one with the highest value of total Mulliken spin was assigned as the moiety where the triplet is mostly localized.

Finally, in the provided ALFAST-DB the remaining unassigned molecules were labeled and assigned based on the smallest difference between the predicted S_0 - T_1 of each fragment and DFT value of the molecule. This “empirical” labeling was not used in the present work, and an additional column is provided with the dataset to allow researchers using the provided data to be able to differentiate between correct assignments and proposed (“empirical”) assignments.

S5. Additional Figures of database contents

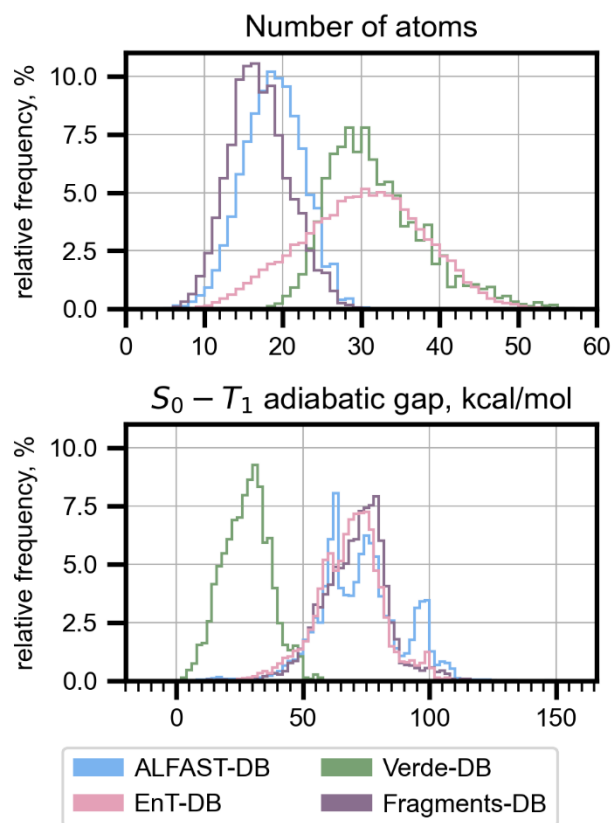


Figure S4: Total number of atoms and computed S_0-T_1 adiabatic gap distributions of the 4 discussed datasets. The bin size selected is of 1 kcal/mol for the S_0-T_1 adiabatic gap and of 1 atom for the number of atoms.

Table S5: Number of shared fragments between databases. The diagonal terms correspond to the total number of fragments of each database.

Databases	ALFAST	EnT	Verde	Fragments
ALFAST	15 410	2 146	0	3 764
EnT	2 146	18 909	0	3 659
Verde	0	0	3 286	0
Fragments	3 764	3 659	0	5 315

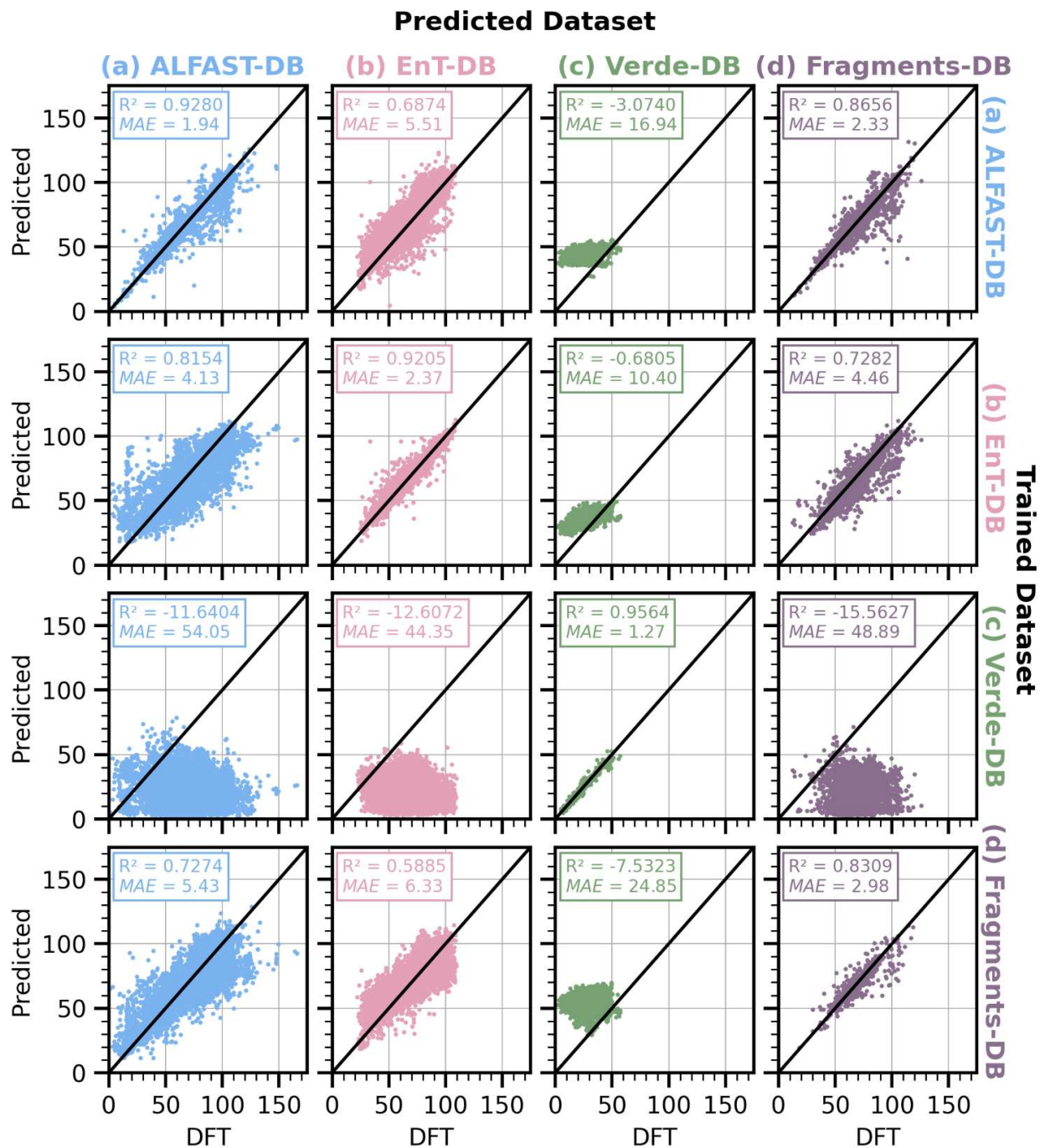


Figure S5: Model extrapolation across datasets (Extended version of Figure 5). Parity plots of adiabatic S_0-T_1 predictions in kcal/mol of each dataset (column) of the same model architecture trained each dataset (row). For example, parity plot “ab” corresponds to the predictions of the “final” ALFAST-DB (trained with ALFAST-DB’s training and validation splits as training dataset) and shows the prediction of the full EnT-DB. Parity plots in the diagonal (aa, bb, cc and dd) only show the predictions of each database’s test set. The metrics, R^2 and MAE values, correspond to the data presented in each plot, test set metrics for the diagonal elements and extrapolation metrics for the off-diagonal metrics.

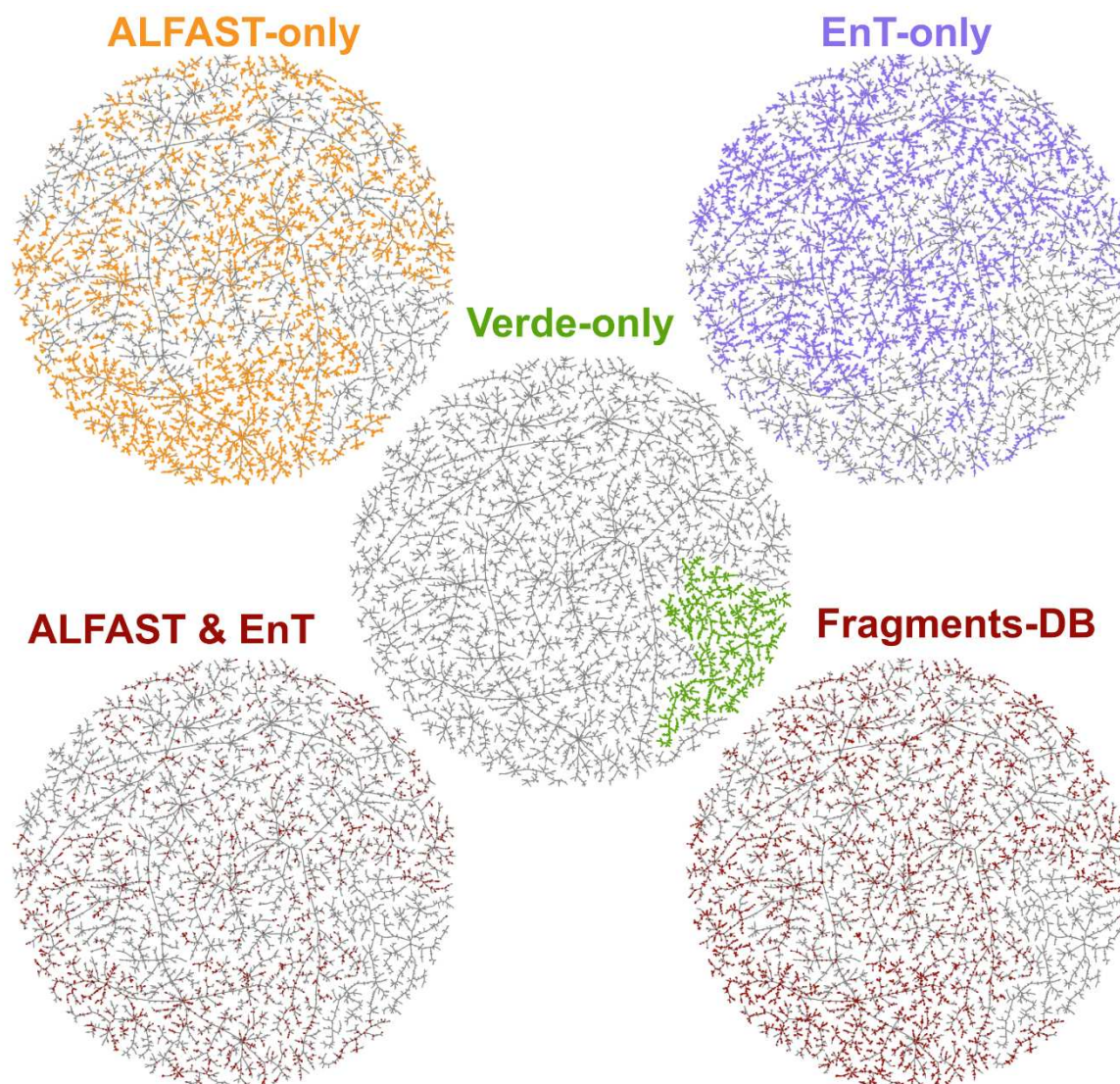


Figure S6: Tree maps of the fragments present in ALFAST-DB, EnT-DB and Verde-DB. Grey lines represent the connections to the most chemically similar fragment based on MHFP encoding and the layout was generated using the LSH forest indexing of the encoded fragments and kept constant across the five tree maps presented.

The tree maps presented in Figure S3 were computed using the python libraries: faerun 0.4.7, mhfp 1.9.6 and the tmap 1.0.6. In these tree maps each fragment is presented as a colored dot and the most similar fragments (based on the MinHash fingerprints of up to 6 bonds representations with a “n_permutations” parameter of 1024) [ref4] are connected through grey edges. The graph layout was generated from the LSHForest and kept constant for the 4

images shown in Figure S3. The nodes representing each fragment were then colored or removed from the graph to generate the four figures, where only the fragments present in each database subgroup are colored. For example, the fragments present in ALFAST-DB but not in EnT-DB nor in Fragments-DB nor in Verde-DB are shown in orange in the top left tree map. In Figure S6 “ALFAST & EnT” stands for fragments which are present in both, ALFAST-DB and EnT-DB but not in Verde-DB. The difference between the “ALFAST & EnT” and the “Fragments-DB” tree map correspond to the increase in chemical coverage of the Fragments-DB.

S6. Analysis of the effect of the theory-shift

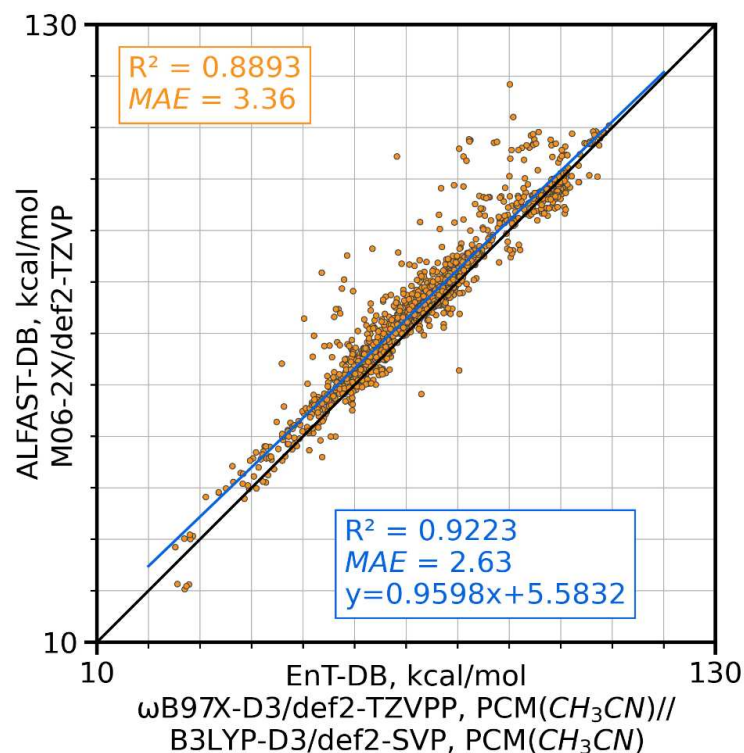


Figure S7: Parity plot and linear fitting of the molecules present in both, the EnT-DB and the ALFAST-DB.

From the Linear fitting obtained from the molecules present in both, EnT-DB and ALFAST-DB we attempted to apply a lineal correction to the parity plots ab and ba presented in figures 5 and S5 (in the manuscript and the present SI respectively. Figure S8A correspond shows the results of applying the lineal correction to the predicted values shown in Figure 5ab (and Figure S5ab), showing a small decrease in the MAE (from 4.13 to 3.95 kcal/mol). Figure S8B shows the results of applying the inverse of the lineal correction shown in Figure S7 to the predicted values in Figure 5ba (and Figure S5ba), showing a larger improvement of the MAE and the R² than Figure S8A. However, in both cases it is still far from the in-sample performance shown in Figures 5aa and 5bb (and also S5aa and S5bb).

The difference in the improvement (0.18 vs 1.2 kcal/mol for Figure S8A and B respectively) can be understood when coupled with the analysis of Figure 6 in the main text. The error distribution of molecules with shared fragments (Figure 6 “shared”, left), which could be expected to be more “in-sample” for the lineal correction, of the ALFAST-DB-trained model has a larger difference with the error distribution of the dataset it was trained on (Figure 6 “all”, left), and therefore it has more room for improvement. On the other hand the EnT-DB-trained model has a smaller difference between the EnT-DB error distribution (Figure 6 “all”, right) and the molecules sharing a fragment across databases (Figure 6 “shared”, right).

Finally, it is to note that a simple lineal correction to shift the theory level is not conceptually correct here. While ALFAST-DB contains the adiabatic S_0 - T_1 gap in vacuum, the EnT-DB was designed for the same property in a polar solvent (Acetonitrile). Molecules with a larger change in dipole moment between the two states are expected to display significant differences due to the stabilization of the solvent. This could lead to not only different energies but also guide the optimization to a different exciton localization.

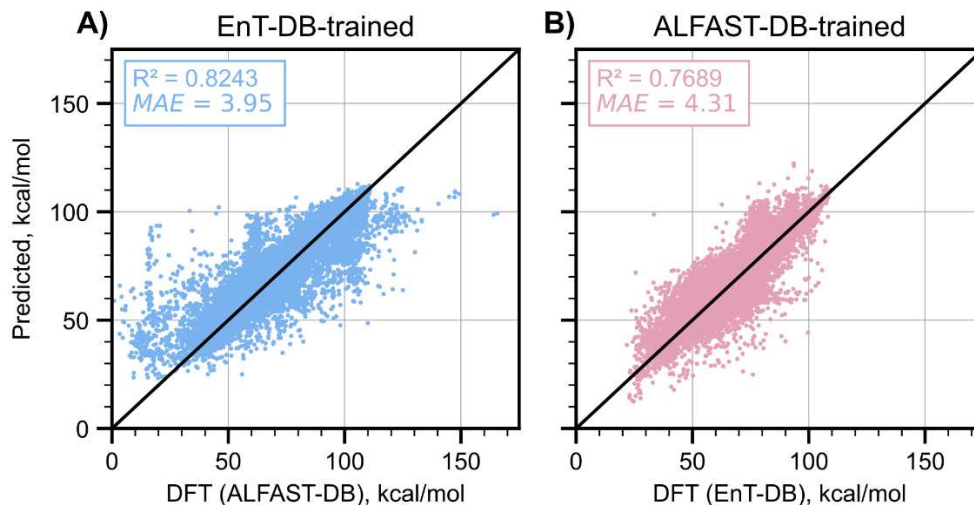


Figure S8: Parity plots of the linearly corrected (using the linear fitting of Figure S7) S0-T1 adiabatic gaps. A) Prediction of ALFAST-DB molecules from a model trained on EnT-DB and applying the lineal correction. B) Prediction of the EnT-DB molecules with a model trained on ALFAST-DB with a consequent lineal correction.

S7. LLM Model Comparison

To further assess the efficacy of our model, we compare its performance to a fine-tuned LLM model. We use a publicly available 108 million parameter RoBERTa model pre-trained on the 480 million SMILES strings available in the ZINC database; the trained model is available at https://huggingface.co/entropy/roberta_zinc_480m. To fine tune the model, we use an ANN readout regression head consisting of 4 linear layers each separated by a ReLU activation function (referred to henceforth as the “readout model”); the final output layer consists of a single linear layer. An embedding dimension of 128 was used for all linear layers. All model layers were implemented in pytorch 2.5.1. The huggingface transformers library was used for loading the pretrained model and generating embeddings for smiles strings corresponding to molecules in the ALFAST database; the generated embeddings were used as inputs for the readout regression model. We follow the train-validation-test split strategy used to train the final Δ -learning predictor model to train and evaluate the

readout model. The readout model was trained for 200 epochs using an AdamW optimizer with a learning rate of 10^{-3} and an MAE loss function as implemented in pytorch. The trained readout model achieves an MAE of 6.17 kcal/mol on the held out test set. Prediction results are shown along with prediction results from the Δ -learning model in Figure S9:

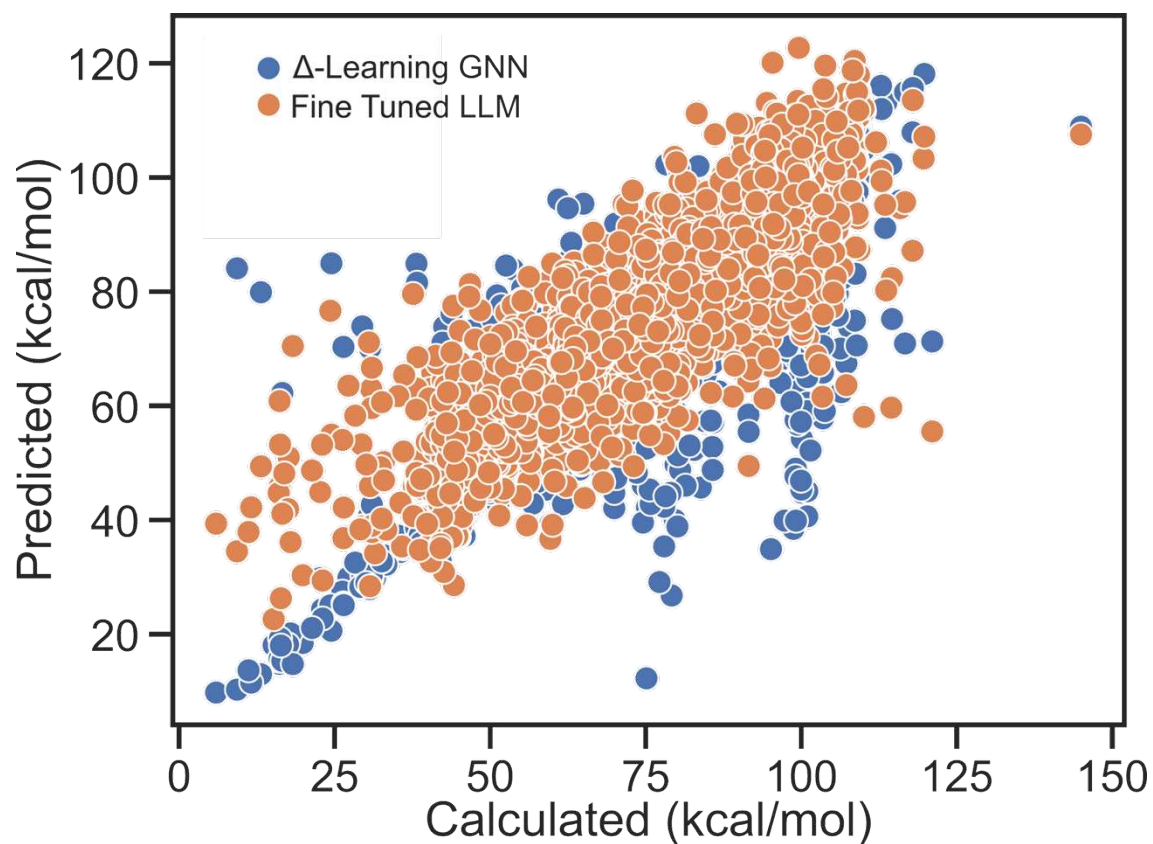
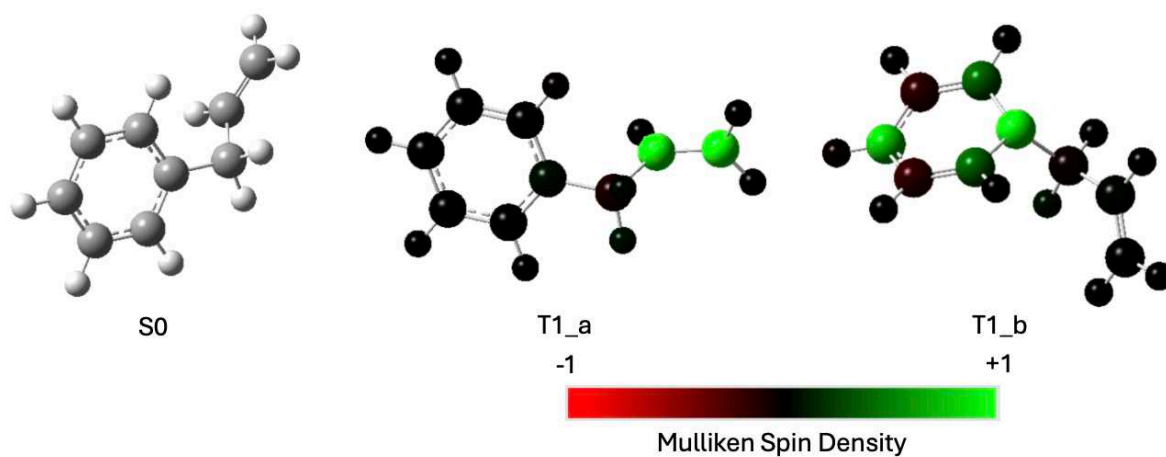


Figure S9: Prediction results for the fine-tuned LLM model vs. Δ -learning GNN model predictions

S8. XYZ coordinates



Mol1-S0

C	2.63141700	0.51890500	0.14912300
C	2.34470700	-0.83737600	0.18097800
C	1.03914500	-1.27734400	0.00351000
C	0.00486600	-0.37247300	-0.20829300
C	0.30320400	0.98822000	-0.23337100
C	1.60498800	1.43176800	-0.05901800
H	3.64752400	0.86443200	0.28806500
H	3.13725300	-1.55615300	0.34549500
H	0.82037500	-2.33861700	0.02923700
H	-0.49856200	1.70164700	-0.38588800
H	1.82025500	2.49242600	-0.08323200
C	-1.41755500	-0.84476000	-0.41827900
H	-1.45483000	-1.92561200	-0.25965900
H	-1.71753800	-0.66289200	-1.45386400
C	-2.39181000	-0.16883200	0.50129500

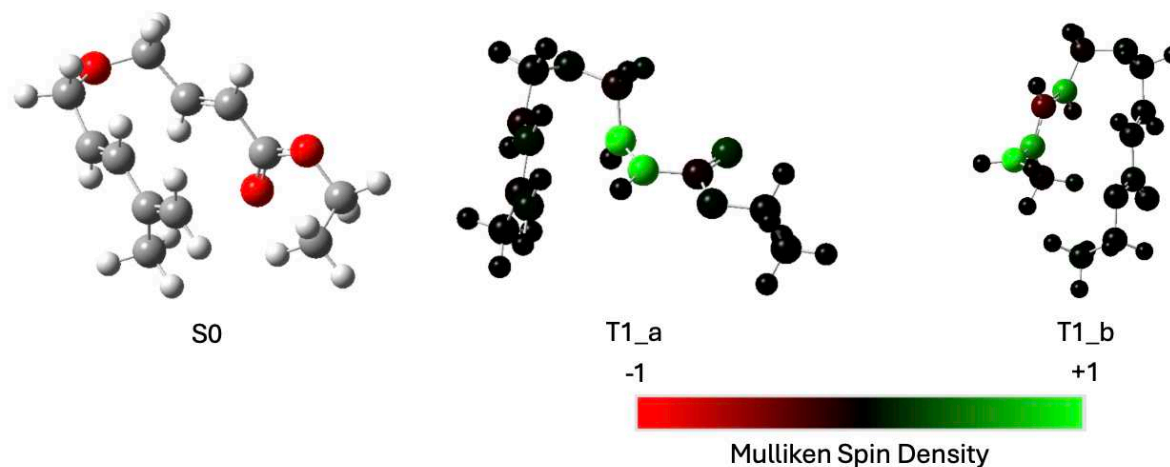
H	-2.20175600	-0.29088100	1.56420000
C	-3.42322200	0.55886000	0.10089800
H	-3.63255800	0.70508400	-0.95303000
H	-4.09459700	1.02875800	0.80761800

Mol1-T1_a

C	-2.72915200	0.32473300	-0.20135400
C	-2.28720500	-0.98924600	-0.20637100
C	-0.94647000	-1.27451300	0.02129800
C	-0.03165800	-0.25503300	0.25729700
C	-0.48655700	1.06247900	0.25729100
C	-1.82290700	1.35214900	0.03163900
H	-3.77277000	0.54935800	-0.37897900
H	-2.98525000	-1.79610200	-0.38955800
H	-0.60602400	-2.30344800	0.01447800
H	0.22100500	1.86493700	0.43495400
H	-2.15995000	2.38087100	0.03748700
C	1.42667200	-0.54974000	0.50442500
H	1.58883900	-1.63563600	0.46613100
H	1.69787600	-0.25236200	1.52427000
C	2.34314400	0.14172600	-0.45981500
H	1.94530900	0.38739900	-1.44022800
C	3.76973400	0.23240600	-0.21127600
H	4.44337000	-0.58182000	-0.45735700
H	4.21398900	1.11704200	0.23000000

Mol1-T1_b

C	-2.35517600	0.65457700	0.47855900
C	-2.30360100	-0.76248900	0.25677000
C	-1.21841400	-1.32129400	-0.31523500
C	-0.00542200	-0.48723700	-0.49377800
C	-0.25611200	0.96557000	-0.67348700
C	-1.35112100	1.50050200	-0.09306400
H	-3.22880500	1.09328900	0.93827700
H	-3.17700500	-1.36149900	0.48585300
H	-1.19472100	-2.37232000	-0.57907900
H	0.46753600	1.58010700	-1.19429900
H	-1.52065400	2.57044100	-0.11982000
C	1.18296900	-0.84872300	0.36537300
H	1.23396400	-1.93832300	0.45105300
H	1.03406000	-0.45335100	1.38064700
C	2.47389900	-0.32162900	-0.19168600
H	2.75886500	-0.70666200	-1.16712800
C	3.24609700	0.56932600	0.41063800
H	2.98196100	0.97786700	1.37962900
H	4.16607800	0.91883500	-0.03966800



Mol2-S0

O	-3.46629700	-1.06066400	0.25628900
C	-3.50518000	0.29535400	-0.16264900
H	-3.59550400	0.34686600	-1.25496200
H	-4.41866700	0.69798900	0.27714400
C	-2.45795700	-1.83529300	-0.35222400
H	-2.73814000	-2.87485600	-0.16347700
H	-2.45041000	-1.68134800	-1.43881300
C	-2.30012400	1.06076800	0.28858900
H	-2.13935300	1.10213300	1.36158400
C	-1.40408700	1.57562100	-0.54919000
H	-1.58214600	1.49431400	-1.61938500
C	-1.09138700	-1.59086900	0.21686800
H	-1.03977100	-1.41337100	1.28762300
C	0.03642000	-1.59046800	-0.48088800
H	0.06330000	-1.74259500	-1.55276200
C	-0.12662300	2.17547900	-0.16018800
C	0.72417000	2.57776900	-1.10656500

H	0.47537300	2.49375500	-2.15753600
H	1.68778800	3.00329700	-0.85666500
C	0.20657700	2.25982900	1.30038600
H	0.26761800	1.25788000	1.73428200
H	1.16344600	2.75702300	1.45047100
H	-0.56029700	2.81444900	1.84519400
C	1.31873900	-1.29137100	0.19169800
O	1.44570300	-1.02603300	1.36047300
O	2.34145900	-1.33302600	-0.67427000
C	3.62633200	-0.97018200	-0.15293000
H	4.33996900	-1.40455500	-0.85035200
H	3.75307400	-1.42885000	0.82703500
C	3.77011300	0.53503100	-0.07648000
H	4.76637900	0.79566500	0.28245200
H	3.62595500	0.98046100	-1.06085100
H	3.03251300	0.94952200	0.61052000

Mol2-T1_a

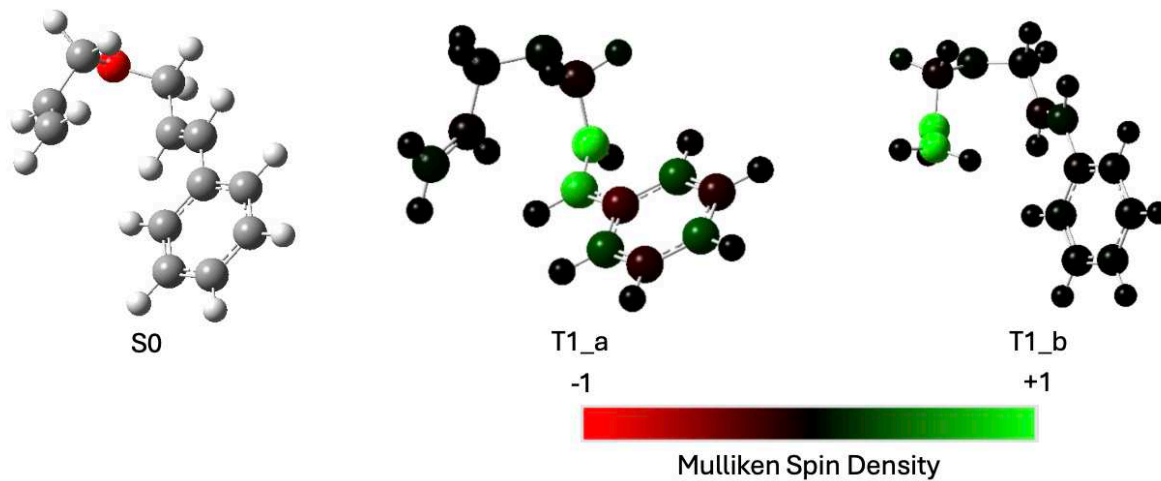
O	-2.04072400	-2.67881100	0.06453800
C	-2.93031200	-1.73684100	-0.50377800
H	-2.78256800	-1.68850800	-1.59036900
H	-3.92817900	-2.13559300	-0.31185500
C	-0.67726900	-2.37029500	-0.15543800
H	-0.13389500	-3.30622900	0.00134900
H	-0.51818400	-2.06161000	-1.19642500
C	-2.79156500	-0.36706200	0.08650100

H	-2.91425100	-0.30382900	1.16346600
C	-2.52221100	0.72356600	-0.63058300
H	-2.39872600	0.62676100	-1.70686900
C	-0.15325400	-1.31944900	0.77470700
H	-0.24653700	-1.50571100	1.83930100
C	0.64421400	-0.20339600	0.30501400
H	0.20956700	0.77642500	0.12968100
C	-2.33378000	2.06878500	-0.08176100
C	-2.04017200	3.08441900	-0.89548500
H	-1.95364500	2.93806900	-1.96516700
H	-1.88028500	4.08570300	-0.51681400
C	-2.44856000	2.24811500	1.40423300
H	-1.72767400	1.61244200	1.92575500
H	-2.26582400	3.28285000	1.68679500
H	-3.44217400	1.96150000	1.75513500
C	2.05759200	-0.38363100	0.00314500
O	2.64219000	-1.44041900	0.08829900
O	2.63867900	0.76220800	-0.39849300
C	4.03443700	0.68665900	-0.71144200
H	4.21947700	1.54811500	-1.35050400
H	4.22145400	-0.22741000	-1.27460200
C	4.88197900	0.73667100	0.54225300
H	5.93892600	0.74193500	0.27372000
H	4.66570000	1.64038200	1.11196800
H	4.68905800	-0.13436400	1.16649300

Mol2-T1_b

O	-3.50939500	-0.92463900	0.30168500
C	-3.50361600	0.48430600	0.09029400
H	-3.72096900	0.69806300	-0.96374000
H	-4.33389900	0.85398000	0.69430600
C	-2.57534800	-1.63294500	-0.48280800
H	-2.93443700	-2.66515000	-0.50452700
H	-2.56722500	-1.25650300	-1.51365600
C	-2.21176700	1.11661700	0.48467200
H	-1.93157500	1.02695900	1.52836000
C	-1.33229000	1.65194500	-0.42436600
H	-1.64768200	1.70203100	-1.46426300
C	-1.18641800	-1.61469000	0.08482100
H	-1.09095300	-1.81773100	1.14791400
C	-0.09116400	-1.37854200	-0.62606900
H	-0.12875900	-1.13318100	-1.68061000
C	-0.02524200	2.03463100	-0.14264000
C	0.87333200	2.45178900	-1.22141600
H	1.03612200	3.49947500	-1.44322900
H	1.44574800	1.71880200	-1.77694400
C	0.51120900	1.96234200	1.25920200
H	0.61475700	0.91931400	1.58307000
H	1.48919200	2.43514200	1.33007100
H	-0.16328500	2.45289800	1.96524900
C	1.23672500	-1.34346300	0.02096600
O	1.46450600	-1.61687500	1.17148300

O	2.18126600	-0.93885900	-0.84515900
C	3.51850300	-0.85008200	-0.33466500
H	4.15000500	-0.86925300	-1.22085300
H	3.72171700	-1.73197900	0.27247700
C	3.72405800	0.41931900	0.46413400
H	4.77215500	0.50721800	0.75274900
H	3.45389800	1.29349400	-0.13031900
H	3.11627900	0.40204100	1.36712200



Mol3-S0

O	3.40146600	-0.61371800	-0.39546100
C	2.36803100	-1.57937700	-0.46241300
H	2.63574400	-2.22119300	-1.30396200
H	2.36494800	-2.19188900	0.44824700
C	3.23817600	0.34517700	0.63531200
H	2.66295500	-0.07199500	1.46975600
H	4.24419100	0.57096400	0.99810700
C	1.01472500	-0.97373700	-0.66632600

H	0.91467000	-0.32213600	-1.52886500
C	-0.01160800	-1.17387800	0.15427000
H	0.12182000	-1.84996100	0.99610400
C	2.59110000	1.60620100	0.14090900
H	3.08269000	2.07655000	-0.70579800
C	1.51015100	2.15250900	0.67543000
H	0.99266000	1.68068300	1.50392200
H	1.09781900	3.08205500	0.30411200
C	-1.33967200	-0.55524500	0.04035600
C	-1.54783400	0.61762300	-0.68986200
C	-2.42768100	-1.13082500	0.69716400
C	-2.81084800	1.17906400	-0.77862300
H	-0.70799800	1.10231000	-1.17211600
C	-3.69311000	-0.57119200	0.60667200
H	-2.27611500	-2.03276900	1.27864300
C	-3.88967000	0.58489400	-0.13493200
H	-2.95452900	2.08995900	-1.34556800
H	-4.52601600	-1.03671500	1.11763600
H	-4.87512600	1.02659800	-0.20428200

Mol3-T1_a

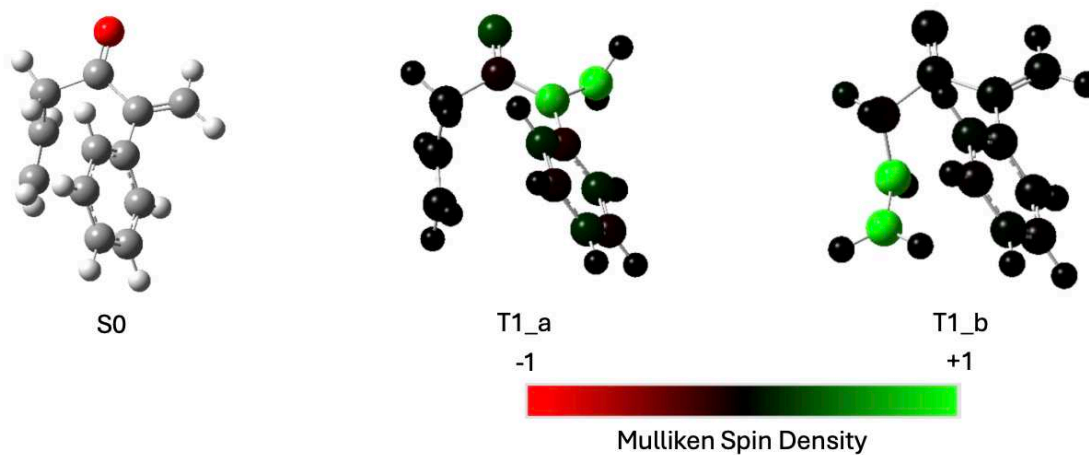
O	-2.95462200	-1.47912100	-0.13924200
C	-1.55666100	-1.37756700	-0.34093100

H	-1.19986900	-2.41298200	-0.35943000
H	-1.34325500	-0.93468400	-1.32213200
C	-3.67032500	-0.32059700	-0.51144200
H	-3.46003000	-0.06439900	-1.55815700
H	-4.72340300	-0.59990600	-0.43596800
C	-0.85379400	-0.60298500	0.72996700
H	-1.07694300	-0.87801600	1.75639700
C	0.09846000	0.46014500	0.44297000
H	-0.27623000	1.48046900	0.39443100
C	-3.40300700	0.87059400	0.36131100
H	-3.51632200	0.70885600	1.42941500
C	-3.08663200	2.06930500	-0.10604200
H	-2.96502500	2.24181000	-1.17003800
H	-2.94302600	2.91789100	0.55040100
C	1.48162400	0.25803500	0.20945200
C	2.06752300	-1.02783200	0.24791700
C	2.32795100	1.35460400	-0.07486600
C	3.41699000	-1.19906600	0.01371900
H	1.44149400	-1.88382800	0.46947900
C	3.67451500	1.17287400	-0.30756900
H	1.89830700	2.34880100	-0.10759500
C	4.23097100	-0.10445400	-0.26603700
H	3.84495400	-2.19288000	0.04880300
H	4.30269600	2.02773400	-0.52331800
H	5.28795000	-0.24422800	-0.44904900

Mol3-T1_b

O	3.44287700	-0.58504300	-0.49409700
C	2.41762500	-1.54529000	-0.33212800
H	2.65236400	-2.33492600	-1.04856400
H	2.45872800	-1.97393600	0.67773500
C	3.34932900	0.49995300	0.41292800
H	3.13124500	0.13329900	1.42444400
H	4.35118200	0.94207000	0.42667600
C	1.04785000	-0.99707500	-0.59113700
H	0.90214200	-0.53326700	-1.56206300
C	0.04662200	-1.06717200	0.28263200
H	0.22033700	-1.56791200	1.23248700
C	2.32836000	1.52059700	0.01329000
H	2.38646100	1.88365200	-1.00917600
C	1.37320900	2.08721900	0.94505000
H	0.37460200	1.67455200	1.04556600
H	1.61886500	2.94381200	1.56279300
C	-1.30352900	-0.52028600	0.08819800
C	-1.56359500	0.49819400	-0.83273100
C	-2.36220000	-1.00569100	0.85672400
C	-2.84651300	0.99338700	-0.99575800
H	-0.74794900	0.92085600	-1.40718000
C	-3.64793000	-0.51304100	0.69208700
H	-2.17118000	-1.78521400	1.58491700
C	-3.89542900	0.48636200	-0.23789400
H	-3.02893400	1.78600200	-1.70997500

H	-4.45693500	-0.90787000	1.29302800
H	-4.89673800	0.87627300	-0.36547500



Mol4-S0

C	1.16324400	-2.04256600	-1.14670300
H	2.19941800	-2.32207600	-1.28619000
C	0.84156100	-1.07544400	-0.29207300
C	1.97407400	-0.38322800	0.43115900
O	3.02092200	-0.94602700	0.62794700
H	0.40399500	-2.57366100	-1.70714500
C	-0.55937600	-0.64636100	-0.07008700
C	-1.40153900	-0.37981400	-1.14830300
C	-1.06629100	-0.51964200	1.22257100
C	-2.71698000	0.00600300	-0.94000500
H	-1.00644900	-0.45170800	-2.15407700
C	-2.38170900	-0.13354000	1.43328100
H	-0.42787400	-0.74134400	2.07038100

C	-3.21015400	0.13351500	0.35159500
H	-3.35611100	0.21662900	-1.78775100
H	-2.76159700	-0.04633600	2.44293400
H	-4.23530200	0.43911500	0.51484500
C	1.78120600	1.07809200	0.81260300
H	0.90703600	1.18833100	1.45456100
H	2.67204900	1.37667100	1.36462500
C	1.60072400	1.91721400	-0.42207400
H	2.46290600	1.98669000	-1.07876500
C	0.47371000	2.53171500	-0.74654000
H	-0.40686300	2.46343300	-0.11630800
H	0.39059400	3.11681100	-1.65322400

Mol4-T1_a

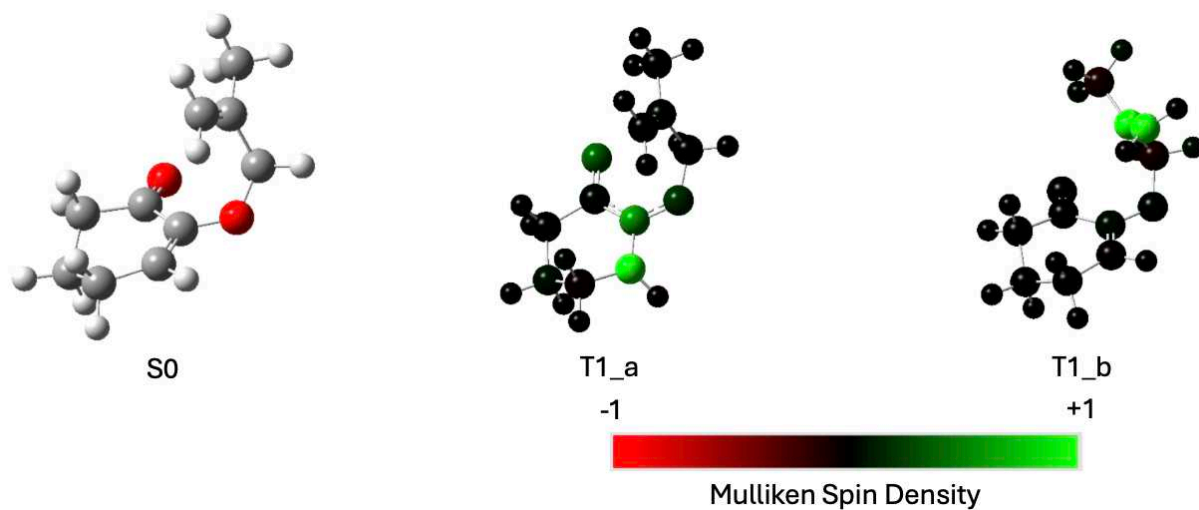
C	1.03232600	-2.38851800	-0.64035900
H	1.37780100	-2.43507500	-1.66500900
C	0.77484800	-1.10020000	0.00115200
C	1.95219100	-0.35376800	0.46233600
O	3.02630200	-0.92172400	0.56574000
H	1.13806200	-3.27956300	-0.03601300
C	-0.59655000	-0.61403200	0.09991000
C	-1.49761700	-0.86306800	-0.94531500
C	-1.06986800	0.06349500	1.23171400
C	-2.79969800	-0.39931700	-0.88654200
H	-1.15519300	-1.40807800	-1.81585600
C	-2.37898100	0.51011700	1.29608900

H	-0.41763000	0.20093400	2.08394400
C	-3.24530400	0.29459000	0.23194500
H	-3.47262900	-0.58318900	-1.71393700
H	-2.72792500	1.01943700	2.18500700
H	-4.26590300	0.65065000	0.28095200
C	1.87459400	1.16125500	0.63635100
H	1.04476900	1.46706100	1.26778100
H	2.81018400	1.45999900	1.10873100
C	1.73189200	1.79982800	-0.71782700
H	2.58624600	1.69686000	-1.38056100
C	0.64201600	2.42442700	-1.13769500
H	-0.23098800	2.52413600	-0.50130700
H	0.58369700	2.85177600	-2.13021100

Mol4-T1_b

C	1.44204600	-2.24134600	-0.57999600
H	2.49897500	-2.45178600	-0.68048800
C	1.04008700	-1.07617500	-0.08021100
C	2.10694900	-0.07683400	0.30883000
O	3.24488100	-0.42598200	0.49259500
H	0.73003000	-2.99853300	-0.88356100
C	-0.39471100	-0.72939100	0.05912100
C	-1.26004600	-0.85369700	-1.02684200
C	-0.90849600	-0.28539900	1.27685100
C	-2.60487100	-0.53784400	-0.90057700
H	-0.86310400	-1.17965200	-1.98047400

C	-2.25272000	0.03159100	1.40484900
H	-0.25318000	-0.20518100	2.13618900
C	-3.10461500	-0.09051200	0.31539800
H	-3.26196000	-0.63301700	-1.75540700
H	-2.63638700	0.37062600	2.35831900
H	-4.15247900	0.16095900	0.41360600
C	1.71366600	1.39557500	0.39463500
H	1.10341900	1.54830400	1.28744400
H	2.65013900	1.94301600	0.53144200
C	0.96007300	1.86523700	-0.81121400
H	1.44983900	1.75546800	-1.77458700
C	-0.33094400	2.52183500	-0.72513100
H	-0.40593200	3.59660600	-0.60095700
H	-1.25692700	1.96280700	-0.80657000



Mol5-S0

C	-1.53832700	-1.37067000	-0.07771100
C	-0.51745300	-0.60527400	-0.47825500

C	-1.75151500	1.40080700	0.46685900
C	-3.03169700	0.62648100	0.17545900
C	-2.82813300	-0.85223200	0.47835400
H	-1.41134100	-2.44486800	-0.16350500
H	-1.47331600	1.27554600	1.51987100
H	-1.84815100	2.46662600	0.26714100
H	-3.86087200	1.02808900	0.75876800
H	-3.28919500	0.74637000	-0.88027300
H	-2.83208500	-1.01178800	1.56397700
H	-3.65886000	-1.44299300	0.08667200
C	-0.60032000	0.88162200	-0.36513500
O	0.19375500	1.62545800	-0.89134900
O	0.56301500	-1.21250700	-1.03838000
C	1.84967100	-0.60940900	-0.95815400
H	1.90226800	0.25757100	-1.61471100
H	2.52692000	-1.37363900	-1.34672400
C	3.33408900	0.84911800	0.46040600
H	2.92621000	1.77600600	0.04957000
H	3.68790300	1.03303300	1.47324600
H	4.19015600	0.57091200	-0.16049400
C	2.28594600	-0.22113900	0.43010900
C	1.78605500	-0.78785300	1.52076900
H	2.12719900	-0.49433500	2.50572700
H	1.02910500	-1.55883700	1.46235700

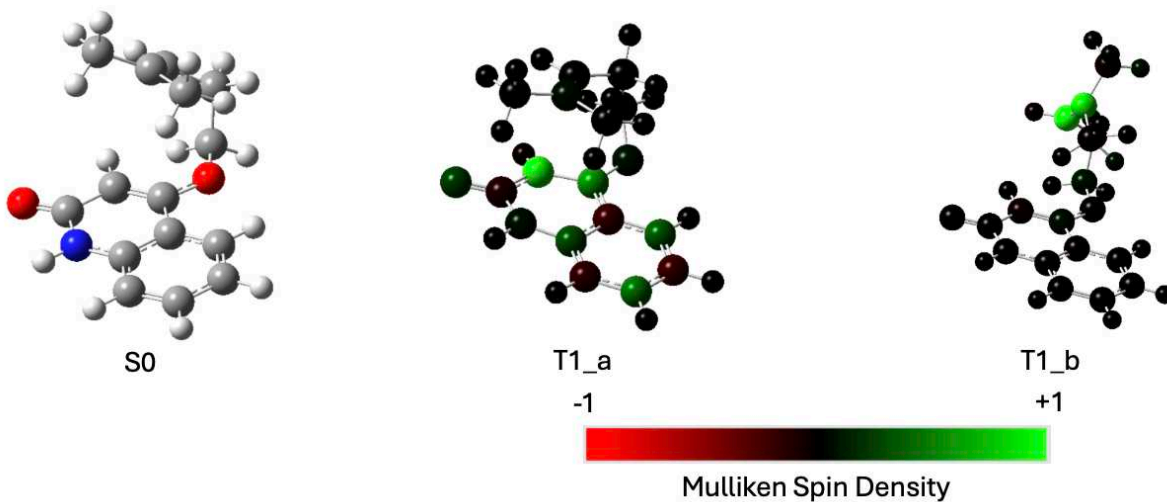
Mol5-T1_a

C	-1.70678200	-1.40464100	-0.24751200
C	-0.54916000	-0.52510900	-0.52408900
C	-1.91971300	1.46631500	0.19424600
C	-3.12066400	0.52593500	0.20123000
C	-2.71022100	-0.87158600	0.70651100
H	-2.01453700	-2.07876700	-1.03934200
H	-1.69333000	1.78816900	1.21560800
H	-2.12934600	2.37498300	-0.37150200
H	-3.91351400	0.93698900	0.82848400
H	-3.52866500	0.42734000	-0.80805100
H	-2.26794900	-0.76200900	1.70391700
H	-3.57177000	-1.53257900	0.78842400
C	-0.61677200	0.89301800	-0.35277700
O	0.33994000	1.65661700	-0.54990200
O	0.51829400	-1.21412000	-0.90093500
C	1.81699700	-0.61608300	-0.96715600
H	1.79222600	0.25605100	-1.61600900
H	2.44080100	-1.38915100	-1.41960400
C	3.46105600	0.76806100	0.33271900
H	3.06322200	1.71005000	-0.05207600
H	3.89183100	0.93781400	1.31761700
H	4.25788800	0.44513300	-0.34295900
C	2.35892400	-0.24336500	0.38704300
C	1.89031800	-0.77150900	1.50802000
H	2.30220000	-0.48833700	2.46803300
H	1.08117200	-1.49187400	1.50474800

Mol5-T1_b

C	-1.28438600	-1.39267600	-0.27762000
C	-0.43589300	-0.43389200	-0.66112000
C	-1.81470200	1.23442800	0.67067400
C	-2.99593900	0.31337200	0.38745700
C	-2.55978400	-1.14458700	0.46773700
H	-1.02406000	-2.41427500	-0.53443900
H	-1.42020200	1.03367500	1.67402200
H	-2.07986400	2.28927000	0.62050900
H	-3.80895100	0.50985800	1.08709700
H	-3.37542400	0.51902900	-0.61698400
H	-2.41699500	-1.43263800	1.51683600
H	-3.34233100	-1.80134400	0.08204200
C	-0.68695500	0.99444400	-0.30917100
O	-0.02431300	1.89786800	-0.76252900
O	0.63696600	-0.75724100	-1.43046300
C	1.90222900	-0.17647800	-1.07584700
H	1.96657400	0.84021700	-1.46070000
H	2.62537700	-0.80505400	-1.60630200
C	2.81368700	0.97902800	1.03571900
H	2.44280600	1.90939300	0.60077600
H	2.63716100	0.99400700	2.11221200
H	3.90263800	0.95789400	0.89264700
C	2.15178400	-0.19918400	0.39854300
C	2.06466400	-1.47215600	1.10300900

H	1.16528400	-1.77148700	1.62740600
H	2.89852100	-2.16735800	1.11251400



Mol6-S0

C	0.37265400	1.10048600	-1.03969500
C	-0.02377900	-0.18383500	-0.88993500
C	-1.33096100	-0.49836000	-0.33596900
C	-2.16253000	0.56870700	0.03054200
C	-0.48277800	2.21177600	-0.66475300
H	-1.12513000	-2.62301800	-0.43211600
H	1.34751000	1.38121000	-1.40267100
C	-1.77767800	-1.80976100	-0.14750200
C	-3.42498300	0.31206000	0.57487200
H	-2.30954800	2.63060400	0.11398300
C	-3.84503000	-0.98855500	0.75128400
C	-3.02216000	-2.05787600	0.39151500
H	-4.06164000	1.14256000	0.85396700
H	-4.82311600	-1.17814200	1.17396400

H	-3.36021600	-3.07486600	0.53435400
O	-0.17673300	3.38499900	-0.76860600
O	0.68677800	-1.27696500	-1.20694600
N	-1.71855900	1.85548200	-0.14905100
C	1.97451100	-1.19060500	-1.81245600
H	2.00436600	-2.01692600	-2.52299900
H	2.07351000	-0.25985900	-2.37194700
C	3.07897300	-1.35345800	-0.77362300
H	2.90328500	-2.28960100	-0.24174900
H	4.02028100	-1.47148000	-1.31610600
C	3.17728400	-0.19019300	0.16989600
H	3.80596100	0.63149100	-0.16372400
C	2.52306600	-0.03740900	1.31896400
C	1.57189800	-1.04267100	1.89877300
H	1.52788000	-1.97239600	1.33678200
H	1.83926600	-1.26719600	2.93426700
H	0.56170000	-0.62024200	1.91857500
C	2.63676800	1.23436400	2.10788200
H	1.66462500	1.73515000	2.14841900
H	2.93370500	1.02926600	3.13945900
H	3.35557700	1.92278500	1.66653900

Mol6-T1_a

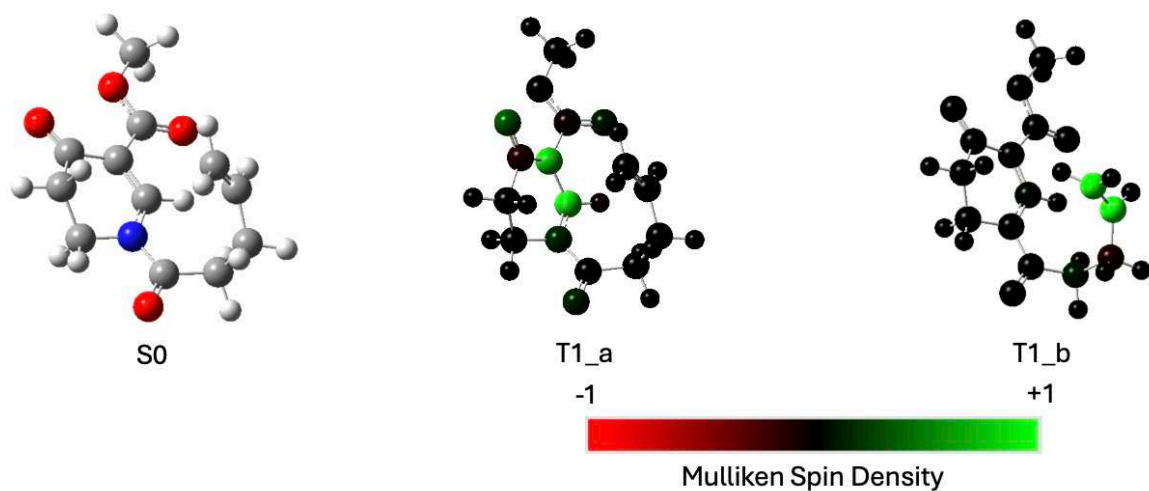
C	0.83952200	0.45272800	-1.54872800
C	0.08882000	-0.72096300	-1.04553600
C	-1.17069300	-0.56085000	-0.50405000
C	-1.65902200	0.76981200	-0.27160100
C	0.37538000	1.78330300	-1.24364800
H	-1.63990100	-2.65256900	-0.26303200
H	1.64496300	0.36200500	-2.25978900
C	-2.00279400	-1.64731000	-0.10168200
C	-2.89663500	0.96925500	0.29460800
H	-1.15837100	2.78736900	-0.38213000
C	-3.69698500	-0.11776800	0.66526600
C	-3.23327000	-1.41986700	0.46875400
H	-3.24076400	1.98194900	0.46921100
H	-4.66677100	0.05807700	1.10891000
H	-3.84925600	-2.25921300	0.76385000
O	0.98243200	2.79823500	-1.54779000
O	0.59063600	-1.96086100	-1.22879800
N	-0.82328700	1.85213200	-0.55834600
C	1.97951100	-2.16212100	-0.98505400
H	2.14263000	-3.21524200	-1.20479000
H	2.58079700	-1.56927800	-1.68087300
C	2.36510300	-1.84356200	0.45975100
H	1.65201400	-2.33612800	1.12022400
H	3.34006200	-2.30830100	0.63657900
C	2.47383900	-0.37427000	0.73107000
H	3.28387300	0.12536100	0.20368400

C	1.71136700	0.37930400	1.53126900
C	0.54610700	-0.12528000	2.32722400
H	0.31904600	-1.17415800	2.15395500
H	0.73087700	0.02204200	3.39488900
H	-0.35002500	0.45131300	2.07868400
C	1.98788400	1.84686100	1.68542700
H	1.10087700	2.43160000	1.42849000
H	2.23202500	2.07948900	2.72564300
H	2.80757600	2.17614600	1.04919800

Mol6-T1_b

C	0.03649600	1.64688800	0.54924100
C	0.19033400	0.35474500	0.91850900
C	1.34960900	-0.40828900	0.48704400
C	2.29295400	0.23433500	-0.32569200
C	1.01534700	2.32468500	-0.28259900
H	0.81089600	-2.23102400	1.46381100
H	-0.81693200	2.24674300	0.81777800
C	1.54765500	-1.74694800	0.83835700
C	3.41822000	-0.46553400	-0.77264800
H	2.76945500	2.01971200	-1.25633200
C	3.59402600	-1.78407800	-0.41299500
C	2.65809000	-2.43311900	0.39496800
H	4.14262000	0.03884600	-1.40013500
H	4.46756900	-2.31826600	-0.76342300
H	2.80430900	-3.46809000	0.67133500

O	0.93160900	3.48412200	-0.64087800
O	-0.65210100	-0.35741300	1.68282400
N	2.09340300	1.54759100	-0.67374800
C	-1.86434200	0.20650900	2.17077400
H	-2.00488500	-0.23609000	3.15698000
H	-1.76619300	1.28464800	2.29665000
C	-3.04322400	-0.15497700	1.26850200
H	-3.10607200	-1.24532900	1.22257000
H	-3.94690000	0.19809600	1.78067200
C	-2.96455300	0.40880400	-0.11761000
H	-3.09636200	1.48070400	-0.23644100
C	-2.83101500	-0.44583000	-1.29609800
C	-4.04419900	-1.16933500	-1.79146600
H	-4.93152200	-0.53555600	-1.74331800
H	-3.91209400	-1.50727600	-2.82105600
H	-4.25780200	-2.06319200	-1.18672000
C	-1.47484400	-0.94618400	-1.67984900
H	-1.06807800	-1.63004000	-0.91971500
H	-1.50213200	-1.48621700	-2.62720000
H	-0.75909600	-0.12451500	-1.76542600



Mol7-S0

C	1.17787900	0.32477600	-0.22036800
C	0.17968300	2.36928600	0.81590200
C	-0.90589400	2.24454200	-0.23608000
C	-0.08813800	-0.00448700	-0.57474800
H	-0.16828100	1.93346800	1.75801900
H	-1.84218700	2.68679800	0.08849600
H	-0.60028500	2.73736500	-1.16382300
H	-0.25541500	-1.00678600	-0.93764200
H	0.42846100	3.41341700	0.99349600
C	1.44736400	1.63111300	0.41648700
C	2.17188300	-0.76013600	-0.41632600
C	-2.47212300	0.43306200	-0.81105400
N	-1.15760400	0.83117300	-0.53697900
O	2.53842100	2.08044400	0.65935100
O	1.88693400	-1.85062100	-0.85465600
O	-3.36088800	1.24827400	-0.77914900
O	3.40303700	-0.42404400	-0.05170000

C	4.38729400	-1.44161100	-0.21549500
H	4.45515500	-1.73834800	-1.26112900
H	5.32076900	-1.00188300	0.12109500
H	4.13387200	-2.31522100	0.38345400
C	-2.77116100	-1.02621500	-1.07822700
H	-1.98464500	-1.53749300	-1.63006600
H	-3.66437400	-1.01994200	-1.69984100
C	-3.08290100	-1.76568100	0.23910300
H	-3.86987500	-1.22131200	0.76457200
H	-3.48292600	-2.74765100	-0.02355100
C	-1.89043300	-1.93212800	1.13135900
H	-1.08987900	-2.56757200	0.75803600
C	-1.75551300	-1.34914300	2.31309300
H	-2.53411500	-0.71046400	2.71619300
H	-0.87071300	-1.49527800	2.91890900

Mol7-T1_a

C	0.98215800	0.42680500	-0.36145400
C	-0.34015000	2.30746400	0.67801500
C	-1.21396600	2.14227000	-0.55740200
C	-0.23511400	0.02996000	-1.13655300
H	-0.88867900	1.97310600	1.56238900
H	-2.19922500	2.57353900	-0.41502800
H	-0.74765400	2.59946300	-1.43119100
H	-0.30428600	-1.00685500	-1.41180900
H	-0.07152200	3.35191500	0.83639300

C	0.94839200	1.50740700	0.60178400
C	2.12773200	-0.49684800	-0.49152300
C	-2.64310600	0.14244800	-0.66367700
N	-1.38191100	0.71283300	-0.83078100
O	1.89293000	1.77566500	1.33142500
O	1.96506500	-1.66907200	-0.74103100
O	-3.60359800	0.83961100	-0.42053700
O	3.31233600	0.07518700	-0.34088800
C	4.43138600	-0.80513000	-0.43894700
H	4.45297400	-1.28216100	-1.41783200
H	5.30728300	-0.18166200	-0.29285700
H	4.37267000	-1.57362500	0.33026700
C	-2.75987200	-1.36347200	-0.71117200
H	-2.13475300	-1.81642500	-1.47910100
H	-3.79965900	-1.56340300	-0.96151700
C	-2.45265900	-1.97354000	0.67478100
H	-3.05543900	-1.45376600	1.42251500
H	-2.78836500	-3.01308300	0.65271300
C	-1.00517200	-1.93415600	1.06241700
H	-0.32410700	-2.53740800	0.46693000
C	-0.51604500	-1.21669800	2.06512400
H	-1.16054700	-0.60211700	2.68516300
H	0.53933000	-1.22754100	2.30831600

Mol7-T1_b

C	1.13907500	0.32822500	-0.23475900

C	0.14904600	2.40529500	0.75527200
C	-0.97355100	2.22370300	-0.24842300
C	-0.12933900	-0.02260800	-0.56157600
H	-0.16080100	2.02745200	1.73473700
H	-1.90530100	2.65776900	0.09968700
H	-0.71479700	2.69555700	-1.20094400
H	-0.29167700	-1.03196100	-0.90795800
H	0.39680600	3.45885700	0.86746300
C	1.41053100	1.65605300	0.35457200
C	2.14316600	-0.74819500	-0.43132400
C	-2.52457800	0.38112700	-0.76145200
N	-1.20993800	0.79813800	-0.50233900
O	2.50338800	2.12133300	0.55562200
O	1.87083300	-1.83392500	-0.88778200
O	-3.41945900	1.18916200	-0.73819100
O	3.36576200	-0.41077900	-0.03986700
C	4.35801900	-1.42101000	-0.20065500
H	4.44301500	-1.70513100	-1.24852700
H	5.28396100	-0.98026100	0.15486500
H	4.10036800	-2.30285400	0.38431100
C	-2.80043300	-1.08409300	-1.00894800
H	-2.03671200	-1.56648600	-1.61636600
H	-3.73551700	-1.10235300	-1.56569400
C	-2.99677500	-1.85512000	0.32113300
H	-3.76938600	-1.33760300	0.89467000
H	-3.38607800	-2.84185000	0.05947900

C	-1.75931700	-2.00416800	1.14881600
H	-1.07145300	-2.80692900	0.89691000
C	-1.45903300	-1.13111900	2.26887400
H	-1.76009200	-1.39005300	3.27784900
H	-0.82782400	-0.25598100	2.15845000

S8. References

- [ref1] (a) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. *J. Cheminform* 2011, 3, 33. (b) Open Babel, version 3.1.1, compiled from source code obtained at <https://github.com/openbabel/openbabel>
- [ref2] Pérez-Soto, R.; Besora, M.; Maseras, F.; pyssian v1.1.0, 2025. maserasgroup-repo/pyssian: 1.1.0 (v1.1.0). Zenodo. <https://doi.org/10.5281/zenodo.14801551>
- [ref3] Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, P.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S. *WIREs Comput. Mol. Sci.* 2020, 11, e01493
- [ref4] Probst, D.; Reymond, J.-L.; *J. Cheminformatics* 2018, 10, 66.

APPENDIX C: Supplementary Materials for Chapter 4

S1. Computational Methods

S1.1 Protein Structural Database Generation

14,536 amino acid sequences (7,268 soluble, 7,268 insoluble) from the SoluProt¹ dataset were chosen as the starting point for database creation. The SoluProt dataset consists of curated amino acid sequence-solubility label pairs curated from the experimental TargetTrack database using a combination of text based searches and sequence identity clustering with the MMseqs2 software. The training and test sets are curated independently, and are both length-balanced and constructed to ensure low overlap between the sequence identity in the train and test sets. The training set consists of 11,436 amino acid sequences, while the test set contains 3,100 sequences. All sequences were used as inputs to generate structural predictions using the OpenFold protein structure software. The Fireworks² workflow orchestration software package was used to generate multiple sequence alignments (MSAs) with the MMseqs2³ software package in a parallelized manner. The generated MSAs were then used as inputs into the OpenFold software to generate structural predictions using the pretrained parameters of the original Alphafold2 model. MSAs were generated against the Uniref90,⁴ mgnify,⁵ and PDB70 databases.⁶ The recommended defaults for monomer prediction were used, with the following exceptions: 1) inference time, low-memory attention was enabled to improve GPU memory usage; 2) chunking was disabled; 3) model tracing was enabled to enhance

runtime efficiency. This resulted in the creation of the **3DProtSolDB**, comprising 14,175 predicted 3D structures (7099 soluble, 7076 insoluble). Structure generation for 361 sequences failed during the relaxation steps of the OpenFold pipeline; these structures were omitted from further study. Concurrently, a sequence similarity-based search of experimental crystal structures deposited in the RCSB PDB was performed for each sequence in the SoluProt dataset using BLAST queries implemented in the pyPDB⁷ Python library. Matched protein structures (sequence identity > 0.9) from any expression system were collected, with the highest sequence identity structure saved in the case of multiple matches. The resulting experimental structural dataset, comprising 8,021 structures (4,147 soluble and 3,874 insoluble), is referred to as **3DProtSolDB-Expt**.

S1.2 Topological Embedding Generation Model Development (TopoCoder)

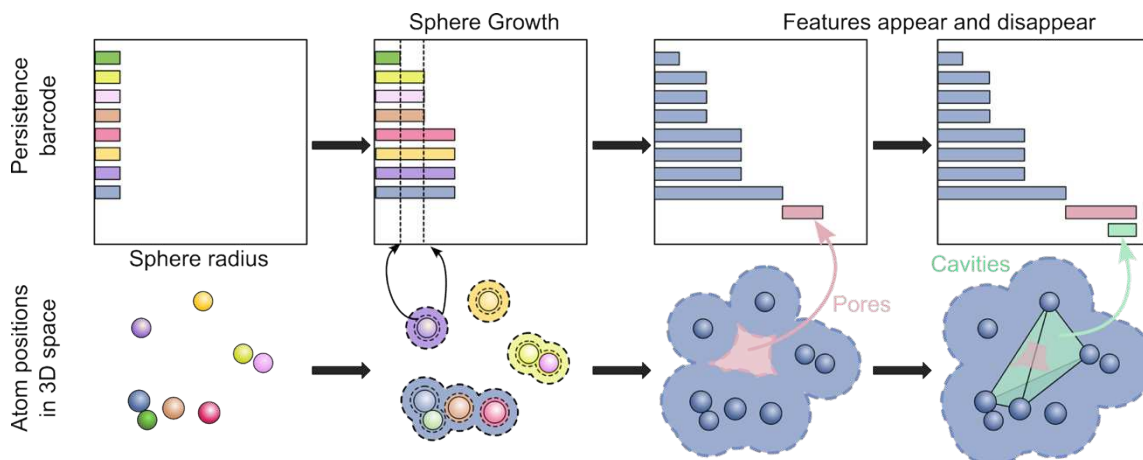


Figure S1: Overall process of generating topological barcodes comparing atom positions in 3D space. Spheres of radius zero are placed at each alpha carbon position and grown. At the intersection, the growth of one of the spheres is stopped, and the associated radius is thus recorded. The appearance and disappearance of 2D pores and 3D cavities are also noted. The resultant structure is called a birth-death diagram; the start and end radii of each bar are plotted against each other to yield the persistence diagram, which in turn is then vectorized to yield the final topological barcode.

Vectorized topological barcodes were generated to encode shape information of the generated protein structures for all 3D protein structures in the **3DProtSolDB-Expt**. The process of generating topological barcodes for protein structures is illustrated in **Figure S1**: a sphere of radius 0 Å is placed at the alpha carbon of each amino acid residue, representing the “birth”. These spheres are then grown simultaneously. When a sphere intersects with another, they merge, and the growing process continues. The radii at which intersections occur are noted, representing the “death” for each sphere; thus, at infinite radius, only one sphere consisting of the union of all initial spheres exists. During this growth process, radii corresponding to the birth and death of 2D pores and 3D cavities are also recorded. This collection of birth and death intervals in each dimension represents the topological persistence intervals for each protein. A bar plot of these intervals for each observed sphere, pore, and cavity, in each dimension, is known as the birth-death diagram. Due in part to the

infinite-dimensional nature and low information density of this representation, the birth-death diagram is transformed by plotting the birth radius vs. the death radius for each bar, yielding a persistence diagram. However, the resulting persistence diagram can be sensitive to the discretization scheme used to vectorize the radius-growth process. A variety of methods have been developed to transform the persistence diagram into a more compact, noise-resistant format, commonly referred to as the topological barcode. Here, we benchmarked two distinct discretization methods for generating topological barcodes: persistence image vectorization⁸ and Atol vectorization.⁹ We find that the persistence image vectorization scheme yields barcodes with low variance across protein structures, and thus, Atol vectorization was used to generate topological barcodes.

The process of generating topological barcodes in this manner is deterministic for a given input structure. To reduce the dependence of generated barcodes on variations in input coordinates, we develop a protein embedding model based on the general RipsNet architecture. The RipsNet model implements a permutation invariant NN architecture, DeepSets,¹⁰ to predict a topological barcode from input geometric coordinates, which can be used for subsequent predictive tasks. The predicted barcodes offer significantly greater resistance to input noise than those generated via the radius growth process detailed above and achieve state-of-the-art performance when used as inputs for downstream image segmentation and object detection tasks.

We built and tested two separate barcode embedding architectures that take alpha carbon positions as inputs, and output a topological barcode: one, termed **TopoCoder-S**, used the same architecture hyperparameters as the RipsNet model, while the other, termed **TopoCoder**, used a larger embedding space to enhance model expressivity. Both models consist of stacks of DeepSets convolutional layers, following the DeepSets architecture to enforce permutation invariance when computing representations of the alpha-carbon inputs. These node representations are summed and pooled, and the resultant vectors are passed through three dense layers. Further architecture details can be found in Section S2.

To train the **TopoCoder** models, the Gudhi¹¹ software package was used to generate 1D, 2D, and 3D Atol vectors for each protein in the **3DProtSolDB-Expt**. The Atol vectors were used as labels for NN training; the model inputs were alpha-carbon positions for each protein structure. The **TopoCoder** architecture yielded more accurate topological barcode predictions than the **TopoCoder-S** architecture, as evidenced by a lower L1 (MAE) loss. Additionally, using separate models with distinct weights for each barcode dimension also yielded more accurate topological barcode predictions. Each model was trained for 300 epochs with early stopping, using an Adam optimizer with a learning rate of 0.0005 and a batch size of 1000, with a training-validation-test split of 0.8, 0.1, and 0.1, respectively, and an L1 (MAE) loss function. Full details of model training are provided in Section S4. The trained **TopoCoder** models were then used to generate 1D, 2D, and 3D topological embeddings for protein structures in the **3DProtSolDB** via forward inference. We note that Since

TopoCoder predicts topological barcodes from alpha-carbon geometry without using solubility labels, and is trained on 3DProtSolDB-Expt (not the SoluProt training set directly), TopoCoder training does not constitute direct leakage of solubility labels into the TopoFormer training pipeline.

S1.3 Development of an Equivariant Protein Solubility Prediction Model (TopoFormer)

The generated topological embeddings (from **TopoCoder**) and 3D structures from the **3DProtSolDB** were used as inputs to train an equivariant transformer NN model to predict protein solubilities. Conventional graph neural networks (GNNs) use vector-valued representations of nodes and edges; when the input is rotated in 3D space, the node and edge representations remain invariant. We applied an equivariant NN model that uses higher-order tensors to represent nodes and edges of protein graphs, allowing direct geometric representations that maintain rotational equivariance. As model inputs are rotated, the intermediate representations and outputs of the equivariant neural model rotate accordingly.

The SE(3) Transformer¹² is an equivariant NN architecture that can directly use geometric information as input. Equivariance is enforced by representing geometric features as linear combinations of spherical harmonics; a learned convolutional filter is then constructed via a radial weighting function. Scalar, vector, and higher-order features are combined throughout the network using tensor products to represent node and edge features. This architecture has been shown to outperform other equivariant architectures on a variety of prediction tasks;¹³ however, due to the large

number of tensor product operations and the radial basis encoding, it incurs a high computational cost. To address this, researchers at NVIDIA implemented optimizations in memory usage and caching, which greatly reduced the computational overhead during model training.¹⁴

We incorporate an invariant attention mechanism¹⁵ into SE(3) equivariant Transformer blocks (**Figure S2**) to enable topological embeddings to influence 3D geometric representations. We exploit the rotational invariance of topological embeddings when inputting coordinates: the topological barcodes used to train the **TopoCoder** model remain unchanged under rotation. In each layer, 1D, 2D, and 3D topological embeddings are concatenated and passed through a multi-head self-attention layer. The result is used as a query vector in a cross-attention layer, using scalar (0-order) node features from the SE(3) Transformer block as key and value vectors. Equivariant convolutional operations in subsequent layers ensure mixing of the resultant 0-order scalar features with higher-order features. Finally, scalar node states are pooled via a summing operation, then passed through a linear readout layer, followed by a softmax filter to obtain the final solubility prediction. Attention

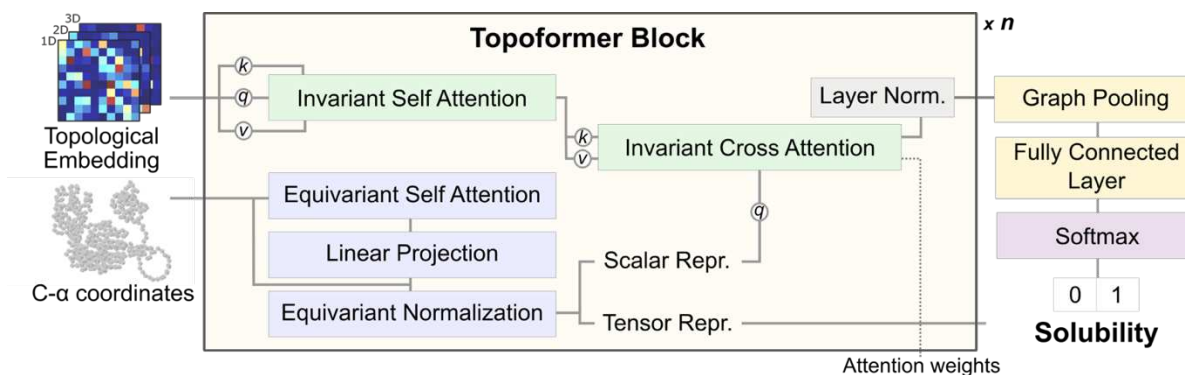


Figure S2: **TopoFormer** architecture diagram. The **TopoFormer** architecture combines scalar geometric representations generated by the equivariant operations with invariant topological embeddings generated by the **TopoCoder** model, utilizing an invariant cross-attention mechanism. Since only scalar representations are treated by the cross-attention mechanism, overall rotational equivariance is preserved. The results of the cross-attention mechanism are propagated to higher-order

weights for the scalar node features from the penultimate cross-attention layer are also obtained for model interpretability. Preliminary hyperparameter tuning was performed to identify optimal model parameters; further details, including detailed hyperparameters and model architecture, can be found in Section S4.

S1.4 Molecular Dynamics (MD) Simulation of Water Survival Probability

MD simulations using the OpenMM¹⁶ software package were performed to validate model predictions for selected example protein sequences. OpenFold-predicted crystal structures were obtained for the chosen sequences using the procedure detailed in **Protein Structural Database Generation**. The OpenFold predicted structure for the protein of interest was placed in a cubic solvent box that is 5 Å larger in each direction than the biggest dimension (i.e., 1 nm padding) using the pdbfixer software package.¹⁷ Equilibration was performed using the Amber14SB¹⁸ force field for 10,000 steps of 0.002 ps from 0 K to 292 K in the NVT ensemble. Following equilibration, a production run of 50,000,000 steps with a time step of 0.002 ps was performed in the NPT ensemble, yielding a total of 100 nanoseconds. The OpenMM LangevinMiddleIntegrator and MonteCarlo barostat were used for heating and barostating. Barostating was applied every 200 steps during the NPT simulations. Per-residue water survival probabilities were then calculated as the proportion of production simulation time in which a water molecule remains within 7.5 Å of a given residue. All simulations were run in triplicate; the reported calculated values were averages across the simulations. Thus, the water survival probability serves as a proxy for studying interactions between solvent water molecules and amino acids.^{19,20} The 7.5 Å radial cutoff was

determined by calculating the distance at which the derivative of the radial distribution function of water.

S2: 3DProtSolDB Database Details

To further investigate the impact of OpenFold model prediction error on the generated structures in the **3DProtSolDB**, we analyze the correlation between the prediction error (measured via per-alpha-carbon RMSD with corresponding experimental structures in **3DProtSolDB-Expt**). The binning strategy adopted in Results and Discussion, Generation and Validation of Protein Structural Databases in the manuscript was used to separate protein structures based on the average per-sequence prediction pLDDT. The frequency distribution of proteins in each bin with respect to total dataset size is shown in Fig. S3; it is found that the proteins in the quartile range corresponding to the lowest average confidence bin makes up only 0.48% of the database.

In addition, we note that the following database and model versions were used to generate the 3DProtSolDB database.

RCSB PDB – Accessed March 28, 2023

UniRef90 – Accessed March 20, 2023

Mgnify – Accessed March 22, 2023 – Version 2023_02

PDB70 – Accessed April 7, 2023

Openfold – Version 1.0.1

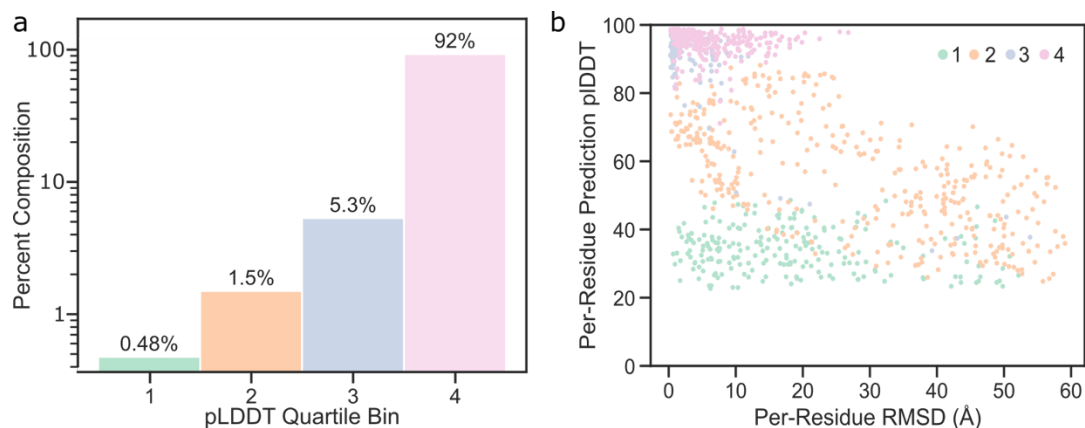


Figure S3: Binned pLDDT scores for OpenFold predicted structures in 3DProtSol. a) Percent composition of each per-sequence pLDDT confidence bin. The highest confidence bin makes up 92% of the total dataset. b) Per-residue pLDDT vs. RMSD with respect to experimental structure for sequences with corresponding matches in 3DProtSol-Expt. A qualitative correlation between low pLDDT score and high RMSD is seen for points above 30 Å RMSD – no point with RMSD over 30 Å has pLDDT greater than 80.

S3: Topocoder Network Details

The **Topocoder** model is a permutation invariant, rotationally invariant model based on the RipsNet framework developed by de Surrel et. al²¹, which in turn is based on the DeepSets²² framework introduced by Zaheer et. al. in order to mirror the physical symmetries inherent in the topological barcode generation process. The topological barcode is invariant to permutation and rotation of the input point cloud, as the filtration used to generate the barcode utilizes pairwise distances between input points.

The **Topocoder** consists of stacks of permutation invariant Janossy pooling blocks, followed by a permutation invariant pooling layer and dense layers. Janossy pooling enforces permutation invariance by permuting an h dimensional input embedding all possible ways, thus requiring weights to be computed for $h!$ possible inputs; this is computationally intractable²³. Thus, initial Janossy pooling blocks are constructed using the sum pooling method detailed by Zaheer et. al. to ensure tractability. Alpha carbon positions are used to

construct the initial point cloud input, which is then passed through a sequence of 5 Janossy pooling layers to generate pointwise representations for each alpha carbon in the input. A sum pooling layer is then implemented to combine representations and ensure permutation invariance, and the result passed through 3 dense layers to generate the final topological embedding.

To train the model, we benchmark the embeddings generated two different topological barcode discretization schemes: the Atol vector⁹ and the persistence image²⁴. Previous work by de Surrel et. al. was performed by training the RipsNet model against persistence image barcodes; thus, persistence image barcodes were generated for the alpha carbon point clouds for structures in the **3DProtSol-Expt** dataset. However, it was found that the generated barcodes show low variance across all structures in the dataset. Thus, the Atol vector was adopted in lieu of persistence image barcodes. We find that the calculated representations show much greater per-structure variance, and we utilize the Atol vectors to train the **Topocoder** model.

The final model architecture is detailed in Table S1:

Table S1: Topocoder architecture hyperparameters

Parameter	Value
Number of Janossy Pooling Layers	5
Janossy Pooling Layer Dimensions	[1200, 600, 300, 150, 50]
Janossy Pooling Layer Bias	True
Number of SumPooling Layers	1
Number of Dense Layers	3
Dense Layer Dimensions	[50, 75, 100]
Dense Layer Bias	True
Sigmoid Readout Layer	False
Optimizer	AdamW

Learning Rate	0.005
Number of Epochs	100
Output Embedding Dimension	100
Loss Function	L1 Norm

S4: Topoformer Network Details

The **Topoformer** model is an equivariant transformer model that allows scalar invariant input features to interact with scalar geometric representations obtained from equivariant attention. The model is composed of stacks of **Topoformer** blocks, followed by a global graph pooling layer, readout layer, and softmax layer outputting the final solubility prediction. Each **Topoformer** block is composed of two trunks: an equivariant transformer trunk modelled around the SE(3) Transformer¹² architecture which transforms geometric data and features from the 3D protein structure, and an invariant self-attention trunk which transforms topological representations. Scalar representations from each trunk are then combined using a cross-attention mechanism, allowing the topological representations to combine with geometric representations. The combined representations from each trunk are passed through a layer norm as suggested by Brody et. al²⁵, and the combined full-tensor rank representation passed to the next block. Residuals are thus combined outside of this layer, allowing for faster training time convergence as suggested by Xiong et. al²⁶

The equivariant trunk in each **Topoformer** block consists of a single SE(3) Transformer block. In this block, scalar-weighted edge-wise equivariant messages are propagated to update a target node, between nodes within a local neighborhood of 5 Å of the target. A linear self-interaction layer allows each node to attend to itself. Scalar (ie. invariant)

attention weights are computed using a scaled inner product of trainable query and key weight matrices, which are in turn computed using the tensor direct sum of the cartesian product of the linear map between each order representation, for each input feature. Thus, the interaction of scalar ($l=0$) node features with vector ($l=1$) edge features would be represented by the direct sum of 3 key and query matrices respectively. Scalar features from the final equivariant convolution are used to generate key and value matrices to combine trunk representations. Dimensions of 64 are used to encode each matrix element (channel dimension). Due to the use of the inner product, the resultant attention weight is rotationally invariant. The inputs to the first Topoformer block equivariant trunk consist of scalar node features (amino acid one-hot encoding, computed RSA, and per-residue ESM embedding) and both scalar ($l=0$) (edge type) and vector ($l=1$) features.

The invariant trunk consists of a single cross attention block. An input vector is passed through three separate linear projection layers to generate key, value and query matrices. An inner product between key and query matrices is used to generate scalar attention weights, and the value projection rescaled using the calculated weights. The scaled vector is passed through a linear intermediate projection layer and an activation layer. The resultant vector is then used to generate a cross-attention query matrix to combine representations from both trunks.

After representation updates in the Topoformer layers, the resultant vectors are passed through an equivariant graph pooling layer to aggregate representations from each node, resulting in a single graph-level representation for scalar and higher order representations. Scalar representations are then passed through a single layer MLP

(multilayer perceptron), which functions as a readout layer, and a softmax layer to generate the final solubility prediction. Node-level attention weights are obtained from the penultimate cross attention layer. Full model hyperparameters can be found in Table S2:

Table S2: Topoformer architecture hyperparameters

Parameter	Value
General Parameters	
Automatic Mixed Precision	True
Batch Size	25
Max Epochs	100
Training Epochs	35
Optimizer	AdamW
Learning Rate	0.0001
Weight Decay	0.01
Topoformer Layers	5
Equivariant Trunk (per Topoformer block)	
Low-memory mode	True
Radial Basis Caching	True
Total Degrees	2
Channels per Degree	64
Channel division	2
Invariant Trunk	
Input dimension	300
Number of linear projection layers	1
Projection layer hidden dimension	300
Attention heads	2
Dropout	0.1

The loss plot for final model training is shown in Figure S4.

The Relative Exposed Hydrophilicity metric was used to analyze model results. A plot of computed REH values for both insoluble and soluble labelled test set samples is shown in Figure S5.

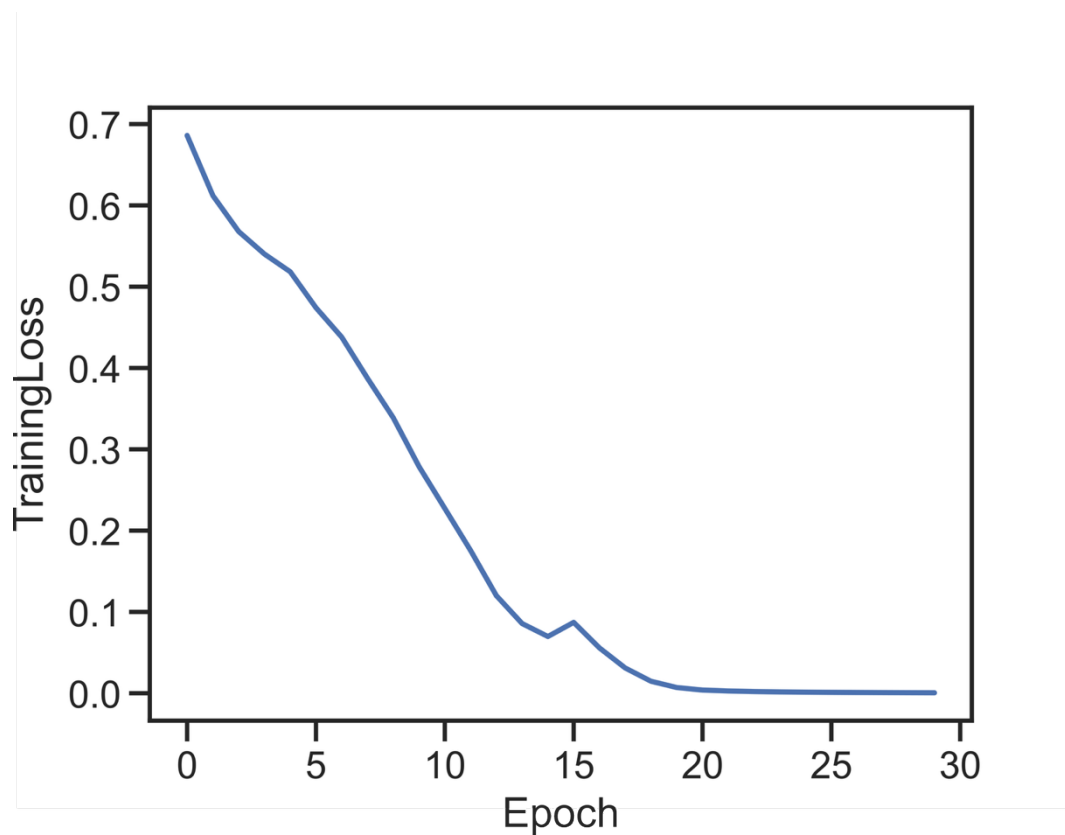


Figure S4: Final TopoFormer model training loss. Final TopoFormer model was trained on both the initial training and validation sets used during hyperparameter tuning

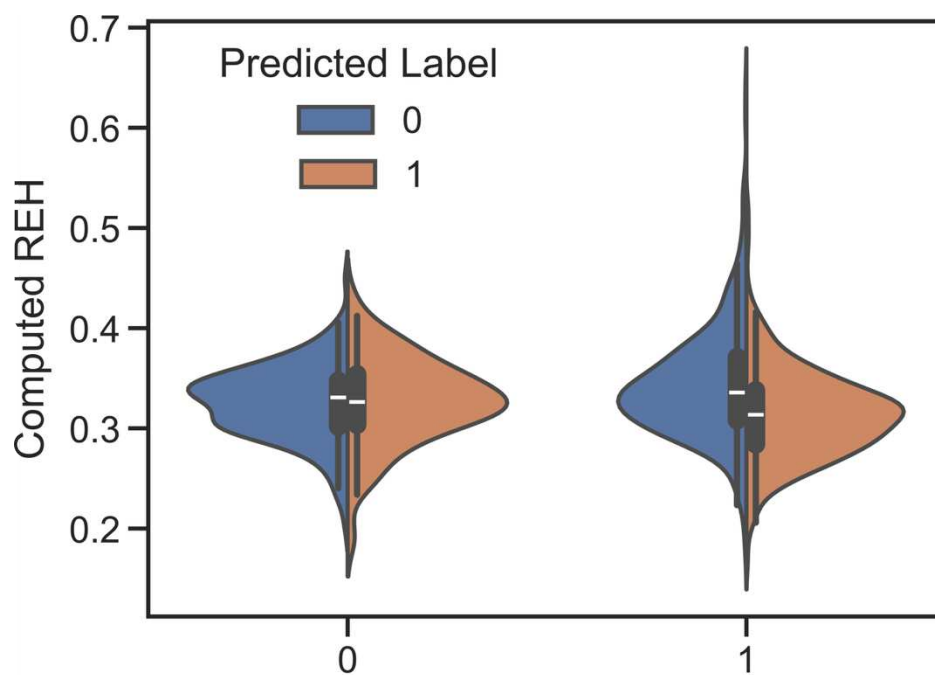


Figure S5: Distribution of per-structure REH values for soluble and insoluble labelled proteins in the 3DProtSol test set.

Timing for each stage of the inference pipeline was obtained for a representative protein selected from the SoluProt test dataset consisting of 223 amino acid residues. Timings were obtained on a machine with 64 cores and two Nvidia A6000 graphics cards. The timings for each stage are detailed in Table S3:

Table S3: Timing details for inference of a single test protein.

Step	Description	Time (s)
1a	ESM model loading	8.54
1b	ESM inference	0.31
2	Persistence diagram inference	0.69
3	Graph construction	1.54
4	Solubility inference	0.46
Per sample total		3.0

The performance of the Topoformer model was benchmarked against a random forest baseline model. Due to the rotational invariance of the random forest model, geometric features were omitted from the study to ensure fair comparison. Calculated node features for each protein from the SoluProt training set were used as inputs; the random forest was trained using only ESM features, only structural node features, and both ESM and structural node features. Barcodes were used for all training runs. Model hyperparameters were tuned independently for each model using a grid search; final hyperparameter choices are given in Table S4. The results of random forest model training are shown in Figure S6:

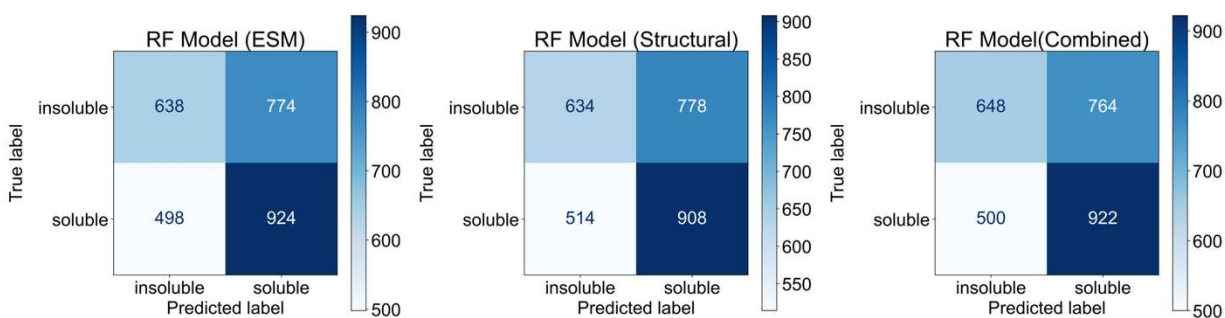


Figure S6: Random forest baseline results. The model trained on both ESM and structural features performs best, achieving an accuracy of 55.4%.

Table S4: Random forest hyperparameters.

Model	Depth	Leaves
ESM Only	20	5
Structural Only	20	1
Full	-1	3

S5: Molecular Dynamics Calculation Details

Benchmarking calculations were performed to assess the impact of solvent box padding distances in order to verify that periodic images remain independent over the course of the simulation with the 1nm padding distance used for production water survival probability calculations. Two additional MD simulations were performed using padding distances of 2nm and 3nm, with all other parameters kept the same as production simulation. These simulations indicate that water survival probabilities increase slightly with larger padding distances, but the effect is not substantial. In the 2nm padding simulation, water survival probability increases modestly as the padding expands from 10 Å to 20 Å. The maximum water survival probability observed with 2nm padding was approximately 0.25, compared to about 0.1 in the 1nm padding simulation. For the 3nm padding simulation, water survival

probability values changed minimally as the padding increased from 10 Å to 30 Å, with the highest values rising only slightly from 0.1 to approximately 0.22. This modest increase is attributed to the larger margin around the protein, which permits water molecules to remain in proximity to the protein for longer durations and provides additional space.

Preliminary benchmarking calculations were performed to determine the distance of first water solvation shell surrounding the CA1-TaCA protein for water survival probability calculations using the SurvivalProbability algorithm implemented by the MDAnalysis software package, with initial radial cutoffs of 4 Å and 5 Å. The following equation is used to perform the calculation:

$$P(\tau) = \frac{1}{T} \sum_{t=1}^T \frac{N(t, t + \tau)}{N(t)}$$

where T is the simulation time, τ is the timestep, N(t) is the number of particles at time t, and N(t, t+ τ) is the number of particles between times t and t+ τ . For a survival time indicating a solvation shell around the protein, the plot should be flat, which indicates a high survival probability. No nonzero horizontal asymptote was obtained at 5Å, indicating the need for benchmarking the cutoff distance choice (Figure S7).

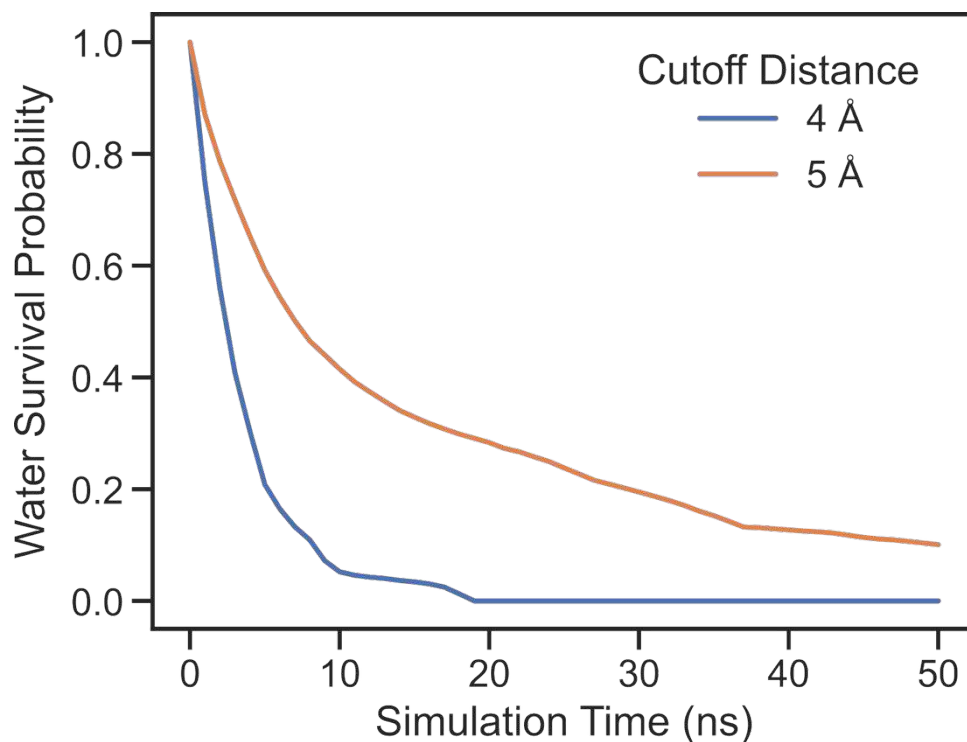


Figure S7: Calculated water survival probability for CA1-TaCA. No asymptote is obtained for the 5 Å cutoff radius.

The radial distribution function (RDF) was then calculated to determine the appropriate cutoff distance, indicated by the minimum radius at which a horizontal asymptote is obtained. The RDF was calculated using the RDF class as implemented in MDAnalysis, which in turn implements the following formula to calculate the RDF between particles a and b :

$$g_{ab}(r) = (N_a N_b)^{-1} \sum_{i=1}^{N_a} \sum_{j=1}^{N_b} \langle \delta(|r_i - r_j| - r) \rangle$$

The first derivative of the RDF function was then computed using the np.gradient function in the numpy software package. A clear horizontal asymptote in the first derivative of the RDF

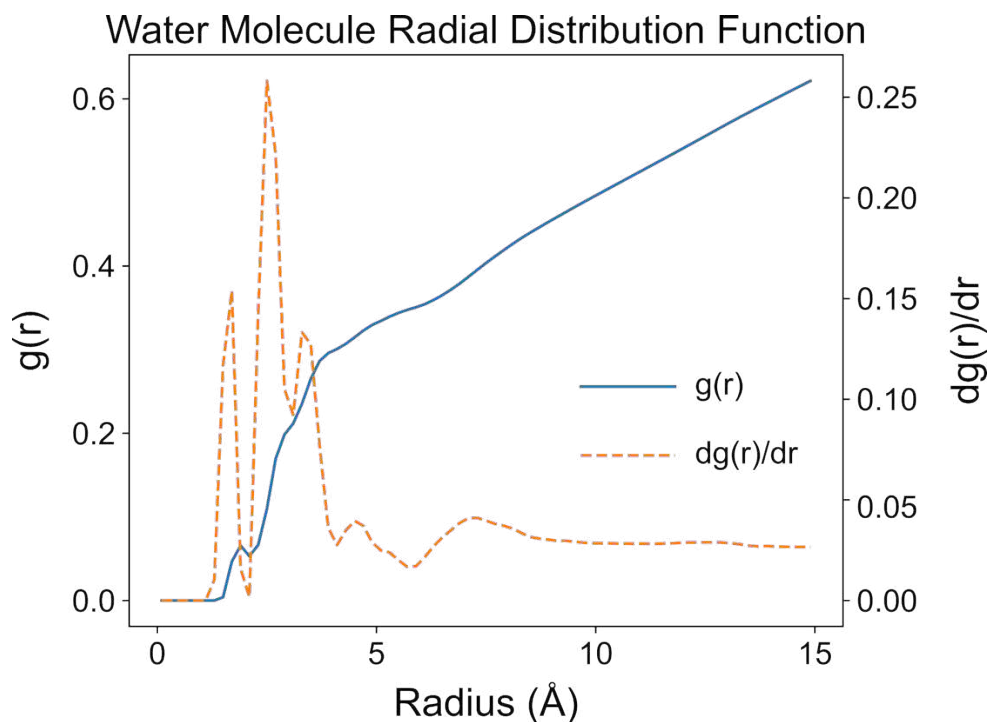


Figure S8: Calculated RDF for CA1-TaCA. The experimental procedure detailed in the main text, Materials and Methods: Molecular Dynamics Simulation of Water Survival Probability was used to perform the MD simulation.

is obtained at 7.5 Å, indicating the distance of the solvation shell from the protein. This distance was then used as the radial cutoff for further water survival probability calculations.

S8: Carbonic Anhydrase Prediction Results

To investigate the correlation between model predictions and water survival probabilities calculated using molecular dynamics (MD) simulations, we utilize two different surrogate models: a linear regression and a random forest model, both implemented using the scikit learn package. Inputs to each model consist of the scalar per-residue attention weights from the TopoFormer model; the surrogate model is then trained to predict the water survival probability. The linear regression model was fit using the sklearn model defaults, while the random forest model used 10 trees and sklearn default parameters. The fit model was then used to generate predictions for all residues. Prediction results for both surrogate models

are shown in Figure S7; the linear regression model achieves an R^2 of 0.15, indicating no correlation, while the random forest achieves an R^2 of 0.76. Thus, a nonlinear relationship exists between the trained TopoFormer attention weights and the MD-based water survival probability.

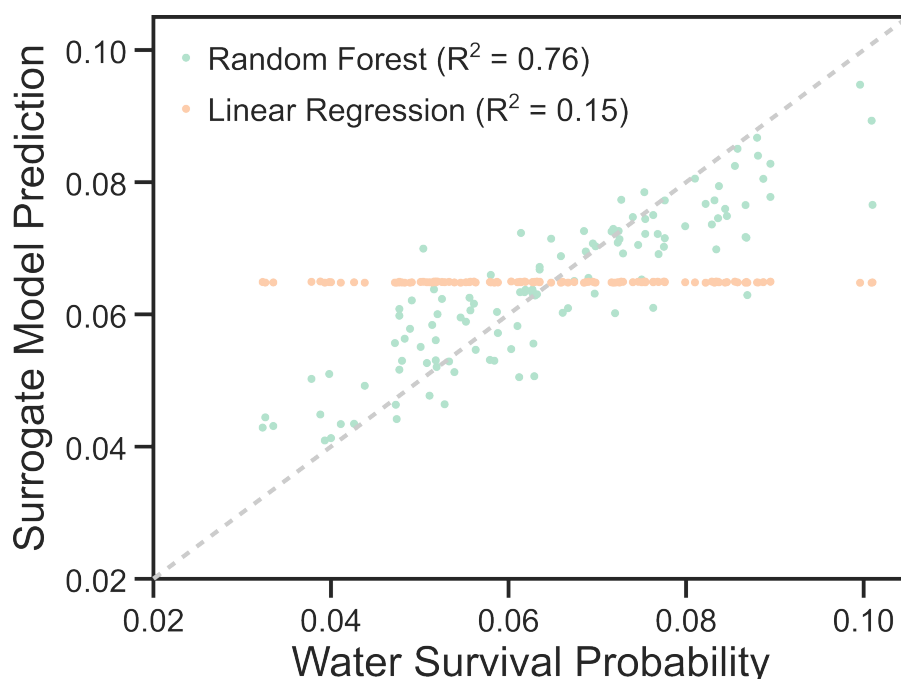


Figure S9: Surrogate model predictions of water survival probability, using TopoFormer attention weights as inputs.

S9: References

1. Hon, J. *et al.* SoluProt: prediction of soluble protein expression in Escherichia coli. *Bioinformatics* **37**, 23–28 (2021).
2. Jain, A. *et al.* FireWorks: a dynamic workflow system designed for high-throughput applications. *Concurrency and Computation: Practice and Experience* **27**, 5037–5059 (2015).
3. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**, 1026–1028 (2017).
4. Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B. & Wu, C. H. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).

5. Richardson, L. *et al.* MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Research* **51**, D753–D759 (2023).
6. Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473 (2019).
7. Gilpin, W. PyPDB: a Python API for the Protein Data Bank. *Bioinformatics* **32**, 159–160 (2016).
8. Adams, H. *et al.* Persistence Images: A Stable Vector Representation of Persistent Homology. *Journal of Machine Learning Research* **18**, 1–35 (2017).
9. Royer, M., Chazal, F., Levrard, C., Umeda, Y. & Ike, Y. ATOL: Measure Vectorization for Automatic Topologically-Oriented Learning. in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics* 1000–1008 (PMLR, 2021).
10. Zaheer, M. *et al.* Deep Sets. Preprint at <https://doi.org/10.48550/arXiv.1703.06114> (2018).
11. The GUDHI Project. *GUDHI User and Reference Manual*. (GUDHI Editorial Board, 2020).
12. Fuchs, F. B., Worrall, D. E., Fischer, V. & Welling, M. SE(3)-Transformers: 3D Roto-Translation Equivariant Attention Networks. (2020) doi:10.48550/arXiv.2006.10503.
13. Thölke, P. & Fabritiis, G. D. TorchMD-NET: Equivariant Transformers for Neural Network based Molecular Potentials. Preprint at <https://doi.org/10.48550/arXiv.2202.02541> (2022).
14. DeepLearningExamples/DGLPyTorch/DrugDiscovery/SE3Transformer/images/se3-transformer.png at master · NVIDIA/DeepLearningExamples. *GitHub* <https://github.com/NVIDIA/DeepLearningExamples/blob/master/DGLPyTorch/DrugDiscovery/SE3Transformer/images/se3-transformer.png>.
15. Vaswani, A. *et al.* Attention is All you Need. in *Advances in Neural Information Processing Systems* vol. 30 (Curran Associates, Inc., 2017).
16. Eastman, P. *et al.* OpenMM 8: Molecular Dynamics Simulation with Machine Learning Potentials. *J. Phys. Chem. B* **128**, 109–116 (2024).
17. Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput Biol* **13**, e1005659 (2017).
18. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).

19. Luise, A., Falconi, M. & Desideri, A. Molecular dynamics simulation of solvated azurin: Correlation between surface solvent accessibility and water residence times. *Proteins: Structure, Function, and Bioinformatics* **39**, 56–67 (2000).
20. Levine, B. G., Stone, J. E. & Kohlmeyer, A. Fast analysis of molecular dynamics trajectories with graphics processing units—Radial distribution function histogramming. *Journal of Computational Physics* **230**, 3556–3569 (2011).
21. de Surrel, T. *et al.* RipsNet: a general architecture for fast and robust estimation of the persistent homology of point clouds. Preprint at <http://arxiv.org/abs/2202.01725> (2022).
22. Zaheer, M. *et al.* Deep Sets. Preprint at <https://doi.org/10.48550/arXiv.1703.06114> (2018).
23. Murphy, R. L., Srinivasan, B., Rao, V. & Ribeiro, B. Janossy Pooling: Learning Deep Permutation-Invariant Functions for Variable-Size Inputs. Preprint at <https://doi.org/10.48550/arXiv.1811.01900> (2019).
24. Adams, H. *et al.* Persistence Images: A Stable Vector Representation of Persistent Homology. *Journal of Machine Learning Research* **18**, 1–35 (2017).
25. [2305.02582] On the Expressivity Role of LayerNorm in Transformers’ Attention. <https://arxiv.org/abs/2305.02582>.
26. Xiong, R. *et al.* On Layer Normalization in the Transformer Architecture. Preprint at <https://doi.org/10.48550/arXiv.2002.04745> (2020).

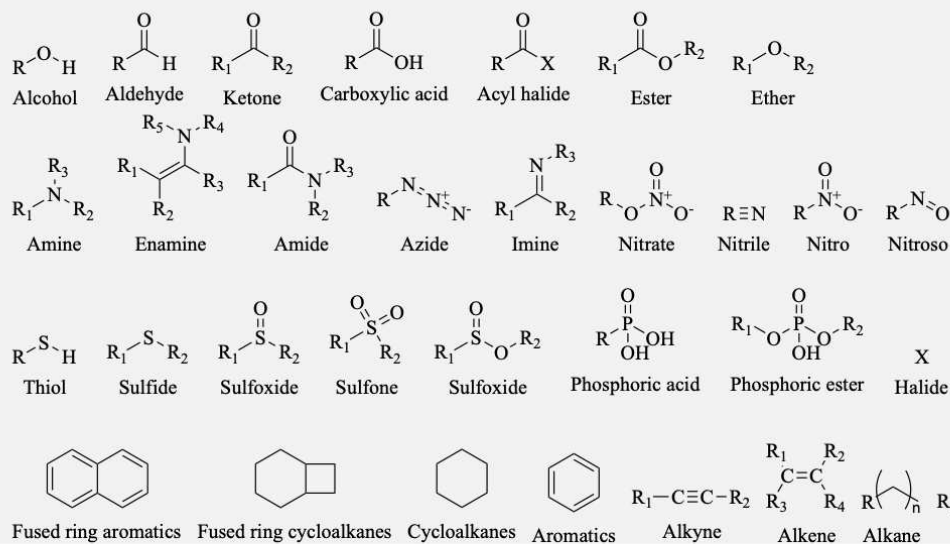
APPENDIX D: Supplementary Materials for Chapter 5

S1. Effects of functional group distributions of solutes/solvents in Aug-DBs on the Root-mean-square Errors (RMSEs) of SSD models

Table S1. The number of data points for each Aug-DB.

Aug-DB-#	1	2	3	4	5	6	7	8	9	10	11	12	13	14
# of data points	183,432	65,833	41,560	34,013	32,474	34,611	34,421	35,701	34,792	35,617	34,134	31,074	30,626	29,906
Aug-DB-#	15	16	17	18	19	20	21	22	23	24	25	26	27	28
# of data points	26,324	27,968	21,448	23,263	16,959	19,311	14,931	15,951	12,215	10,901	12,160	8,927	7,142	8,521
Aug-DB-#	29	30	31	32	33	34	35	36	37	38	39	40		
# of data points	7,705	6,346	5,603	5,605	4,522	3,708	3,168	3,664	3,280	3,233	2,961	2,249		

Solutes



Solvents

Water / O₂N-containing / S-containing / X-containing / Aromatics / Cycloalkanes / Hydrocarbons / P-containing

Fig. S1 Functional group categories of solutes and solvents in **Exp-DB** and **Aug-DBs** used for analyzing the effects of functional group distributions of databases on the RMSEs of SSD models.

Table S2. Solute-solvent functional groups that are overlapped between **Aug-DB** and **Exp-DB**, with the number of data points and mean absolute errors for each category, for three **Aug-DBs**.

FGs overlapped between Aug-DB-18 and Exp-DB (solute.solvent)	# of data points in Exp-DB	MAE
Thiol.Water	1	0.13
PhosphoricAcid.X-containing	1	1.65
Nitrile.Water	2	0.44
PhosphoricAcid.Aromatics	2	1.58
Cycloalkanes.Cycloalkanes	3	0.26
Total	9	0.73
FGs overlapped between Aug-DB-22 and Exp-DB (solute.solvent)	# of data points in Exp-DB	MAE
PhosphoricAcid.Hydrocarbons	3	1.13
Cycloalkanes.Aromatics	6	0.19
etc.X-containing	2	1.22
Peroxide.Water	1	0.13
PhosphoricAcid.X-containing	1	1.65
Thiol.Cycloalkanes	1	0.61
Cycloalkanes.S-containing	1	0.12
PhosphoricAcid.Aromatics	2	1.58
Total	17	0.74
FGs overlapped between Aug-DB-29 and Exp-DB (solute.solvent)	# of data points in Exp-DB	MAE
Sulfide.Cycloalkanes	1	1.21
Cycloalkanes.Aromatics	6	0.19
etc.X-containing	2	1.22
Ester.Water	26	1.16
Aldehyde.Water	7	0.43
Cycloalkanes.Hydrocarbons	13	0.31
Sulfide.Water	11	0.28
Arene.Cycloalkanes	4	0.23
Nitro.Cycloalkanes	1	0.65
Nitro.Water	12	1.13
Oxygen.O,N-containing	10	0.81
Total	93	0.74

The increasing/decreasing RMSEs for the **Exp-DB** test set (**Figure 5.3**) are attributed to several factors. During the initial SSD cycles (~Cycle 15), more than 25,000 data points were added to the database (Table S1). In other words, the sizes of **Aug-DBs** until Cycle 16

are relatively more extensive than those after Cycle 16. Such larger **Aug-DBs** indicate that not sufficient data points have been added yet to accommodate new solutes/solvents having new functional groups (FGs) unseen by the model, presumably leading to the fluctuating RMSEs until Cycle 16 (Fig. 3). During the Cycles from 17 to 29, fewer data points were added (around 7,700 – 23,000, Table S1), and the RMSEs showed less fluctuation. However, an increase in RMSE compared to the previous cycle was still observed in Students 18 and 22 (Fig. 3). To explain these intermittent rises in RMSE, we hypothesized that RMSE increases if the added **Aug-DB** in the current cycle is less overlapped with **Exp-DB** in terms of FGs of solutes and solvents.

This hypothesis was verified by, first, assigning FGs of solutes and solvents in **Exp-DB** and **Aug-DBs**, based on the categories defined in **Figure S1**. The assignment was performed automatically by matching substructure patterns between the simplified molecular-input line-entry system (SMILES) of a molecule and the SMILES arbitrary target specification (SMARTS) string of FGs. If a molecule has multiple FGs, the one with higher priority was assigned according to the priority order shown in **Figure S1**. 33 categories were defined for solutes. For solvents, eight categories are sufficient to distinguish different FGs, according to the results of t-distributed stochastic neighbor embedding (t-SNE) clustering analysis shown in **Figure 6.7**. After assigning FGs, we counted the number of data points in **Aug-DBs** whose FGs of solutes and solvents overlap with those in **Exp-DB**. This analysis was performed for **Aug-DB-18** and **Aug-DB-22** which led to an increase in RMSE for Students 18 and 22, and for **Aug-DB-29** which resulted in a decreasing RMSE for Student 29. Table S2 tabulates the number of data points whose solute-solvent FG pairs in **Aug-DB** overlap with

those in **Exp-DB**, with the MAE between experimental and COSMO-RS solubilities. The ΔG_{solv} values from COSMO-RS were referred to here since it was used as a reference method during the SSD.

For **Aug-DB-18** and **Aug-DB-22**, FGs of only nine and 17 data points overlap with **Exp-DB**, whereas 93 data points in **Aug-DB-29** have common FGs with **Exp-DB**. As can be seen in Students 18 and 22, fewer data points having common FGs with **Exp-DB** leads to an increased RMSE, whereas Student 29 showed a decreasing trend because of a higher overlap of **Aug-DB-29** with **Exp-DB**. Of note, the accuracies of the reference COSMO-RS method are similar in all these three cases (MAEs of 0.73-0.74 kcal/mol). Therefore, it can be deduced that RMSEs of Student models are affected by the number of data points intersecting with the FGs of **Exp-DB** rather than the accuracy of the reference method.

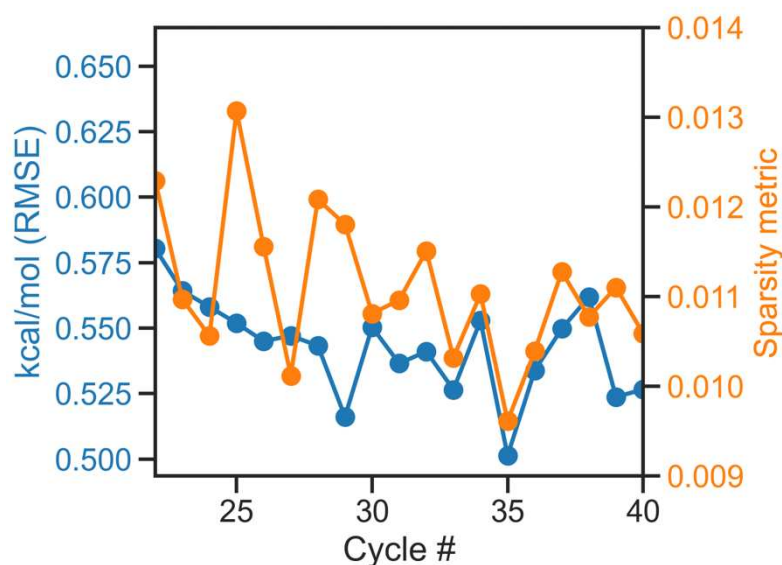


Fig. S2 The root-mean-square errors (RMSEs) vs. sparsity metric values in later SSD cycles (Cycle 22 – 40).

Meanwhile, after Cycle 30, the number of data points for the accumulated **Aug-DBs** exceeds 900,000, and the model is presumably prone to the ‘over-regularization’ by the solvents and solutes that frequently appeared in the accumulated **Aug-DBs**. In this case, the appearance of new solute-solvent pairs can improve the RMSEs of the **Exp-DB** test set for Students. For the quantitative validation of this hypothesis, we devised a metric for the sparsity in the distribution of augmented data points of each cycle as $\frac{\# \text{ of data points in the current Aug-DB}}{\# \text{ unique solutes} \times \# \text{ unique solvents}}$. The metric value is 1.0 if all possible solute-solvent pairs exist in the accumulated database for the given solutes and solvents, and a value closer to 0.0 indicates more introduction of unseen solutes/solvents to the database. As shown in Fig. S2, from Cycle 30 to 37, the trend of the sparsity metric is mostly in line with the trend of RMSEs. This result demonstrates that adding unseen solute/solvent molecules (i.e., a decreased metric value) leads to SSD models that predict solubilities of problematic solute-solvent pairs in **Exp-DB** more accurately, and thus lower RMSEs.

S2. Detailed results of the 10-fold cross-validation using data subsets obtained from different data splitting methods

Table S3. Detailed information about the data set splits for the 10-fold cross-validation (**Fig. 6**) of the Student 35 models.

Data set	# of data points (Exp-DB)	# of data points (35 Aug-DBs)
Fold 1	1,048	82,893
Fold 2	1,048	82,888
Fold 3	1,048	82,886
Fold 4	1,047	82,882
Fold 5	1,047	82,877
Fold 6	1,047	82,874
Fold 7	1,047	82,871
Fold 8	1,047	82,868
Fold 9	1,047	82,865

Fold 10	1,047	82,861
Held-out test set	1,164	92,107

Table S4. Model accuracies of the 10-fold cross-validation with the random solute-solvent data set split for Student 35.

Exp-DB MAE (Validation sets from 10 folds, kcal/mol) ^a	Exp-DB MAE (Held-out test set, kcal/mol) ^b	Exp-DB RMSE (Validation sets from 10 folds, kcal/mol) ^a	Exp-DB RMSE (Held-out test set, kcal/mol) ^b
0.22	0.25±0.01	0.44	0.55±0.02

^aErrors evaluated using all 10 validation sets collected from 10 folds. ^bMean and standard deviation of MAEs/RMSEs from predicting solubilities of a held-out test set using 10 different models.

Table S3 lists the number of data points for each fold and held-out test set when the random solute-solvent sampling was performed. Each fold has almost the same number of data points so that the training:validation:test set ratio of 72:8:9 is maintained during each run of the 10-fold cross-validation. Table S4 summarizes the accuracies of the 10-fold cross-validation when the data splits in Table S3 were used. All MAEs and RMSEs of the 10 validation sets and held-out test set do not deviate significantly from **Exp-DB** test set MAEs (0.25 kcal/mol) and RMSEs (0.50 kcal/mol) evaluated when the best-case model was used. These results indicate the Student 35 models are scarcely susceptible to overfitting.

Table S5. Detailed information about the solvent-wise and solute-wise data set splits for the 10-fold cross-validation of the Student 35 models.^a

Data set	Solvent-wise split				Solute-wise split			
	Exp-DB		35 Aug-DBs		Exp-DB		35 Aug-DBs	
	N _{solvent}	N _{data}	N _{solvent}	N _{data}	N _{solute}	N _{data}	N _{solute}	N _{data}
Fold 1	130	706	26	84,468	205	1,218	1,022	86,689
Fold 2	132	1,127	25	82,040	228	961	1,012	86,083
Fold 3	130	1,101	26	84,466	216	895	998	85,755
Fold 4	131	1,131	25	81,570	197	960	1,002	85,121
Fold 5	131	979	26	83,769	214	1,219	991	84,473
Fold 6	132	1,065	26	85,212	174	1,063	991	84,446

Fold 7	132	648	26	83,482	208	801	981	83,765
Fold 8	131	1,043	26	82,410	214	849	963	82,970
Fold 9	131	770	26	84,672	223	1,391	954	81,273
Fold 10	132 ^b	2,079 ^b	26 ^b	84,771 ^b	198	1,196	939	80,360
Held-out test set	131	988	26	84,012	198	1,084	938	79,937

^aThe number of data points varies among different split data sets because of different number of available data points for each unique solute/solvent. However, the stratified data set split was carried out so that the variance of the number of data points is minimized while being randomly sampled. ^bThis split data set contains the water solvent.

Table S6. Model accuracies of the 10-fold cross-validation with solvent/solute-wise, stratified data set split for Student 35.

	Exp-DB MAE (Validation sets from 10 folds, kcal/mol) ^a	Exp-DB MAE (Held-out test set, kcal/mol) ^b	Exp-DB RMSE (Validation sets from 10 folds, kcal/mol) ^a	Exp-DB RMSE (Held-out test set, kcal/mol) ^b
Solvent-wise split	0.55 [0.27 ^c]	0.28±0.02	1.23 [0.54 ^c]	0.50±0.02
Solute-wise split	0.34	0.30±0.01	0.67	0.55±0.02

^aErrors evaluated using all 10 validation sets collected from 10 folds. ^bMean and standard deviation of MAEs/RMSEs from predicting solubilities of a held-out test set using 10 different models. ^cMAEs/RMSEs evaluated without the data points having the water solvent.

Similar analysis was carried out under the solvent/solute-wise data splits. This is to verify that the model accuracies are maintained when solubilities involving certain solutes and solvents are unseen in the training set and evaluated in the validation or test sets. **Table S5** shows the number of data points in each subset of data when the solvent/solute-wise data splits were conducted. Although the differences in the number of data points are higher than those in **Table S3**, such deviations were minimized through the stratified sampling described in **Figure 6.5C**. Fold 10 in the solvent-wise split has particularly more data points, since it includes the water solvent for which more data points are available than for other solvents.

Table S6 summarizes the accuracies from the 10-fold cross-validation with the solvent/solute-wise data splits. In the solvent-wise splits, water solvent was included in one of the validation sets, leading to high errors than those without the water solvent. This result indicates that it is challenging to obtain desirable accuracies when the water solvent is unseen because the chemical behavior of water as a solvent is significantly different from other organic solvents. Similar trends were also observed in the literature for the water solvent.¹ The validation set MAE and RMSE without water (0.27 and 0.54 kcal/mol, respectively) is comparable to those from the best case model (0.25 and 0.50 kcal/mol, respectively), demonstrating the robustness of the model even in the solvent-wise data splits. Test set MAE and RMSE are 0.28 and 0.50 kcal/mol with the standard deviations of 0.02 kcal/mol when the prediction was carried out for 10 times, which also indicates the low overfitting tendency of the model. In the solute-wise splits, MAEs and RMSEs are slightly higher than the other above cases, but they showed an increment of errors only around 0.1 kcal/mol.

Table S7. Prediction accuracies from the 10-fold cross-validation of Student 35 with the solvent-wise data split, for the 10 representative solvents in the **Exp-DB** held-out test set.

Exp-DB test set solvent	RMSE (kcal/mol)	MAE (kcal/mol)	# of data points
Heptane	0.48	0.20	211
Diethyl ether	0.55	0.37	140
Butyl acetate	0.38	0.26	97
Hexadecene	0.62	0.28	90
Methyl acetate	0.46	0.34	87
N-Formylmorpholine	0.44	0.37	61
Cyclopropane	0.48	0.33	53
Methyl isobutyl ketone	0.50	0.31	23
Propanenitrile	0.30	0.25	20
Dipropyl ether	0.26	0.21	16

Table S7 shows the prediction errors for the 10 representative solvents in the **Exp-DB** test set when the solvent-wise data splits were performed. The MAEs and RMSEs range from 0.20 – 0.37 kcal/mol and 0.26 – 0.62 kcal/mol, respectively. These accuracies are comparable to those in the literature around 0.1 kcal/mol differences (MAEs: 0.10 – 0.27 kcal/mol, RMSEs: 0.13 – 0.51 kcal/mol), although the test set solvents are different.¹

S3. Other control models to demonstrate the feasibility of the SSD shown in Fig. 2A

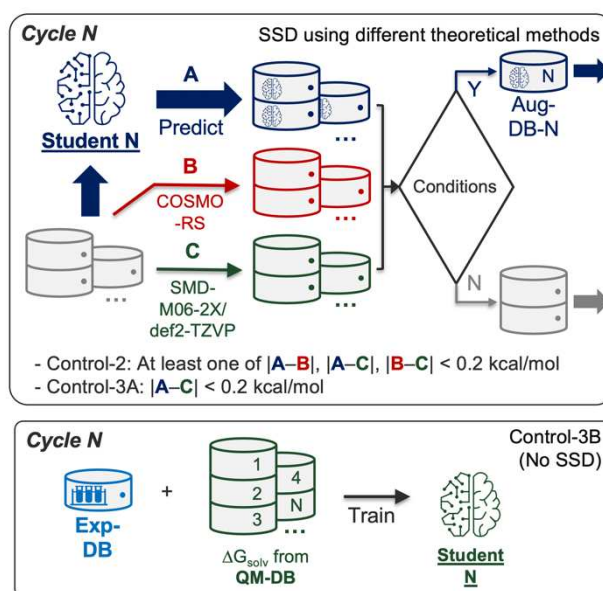


Fig. S3 Description of other Control models examined to demonstrate the feasibility of the SSD shown in **Fig. 2A**.

Table S8. Comparison of accuracies of Control models with the models from the SSD shown in **Fig. 2A**.

	SSD shown in Fig. 2A		Control-2		Control-3A		Control-3B	
Cycle #	Exp-DB test set RMSE (kcal/mol)	# of data points in the accumulated database	Exp-DB test set RMSE (kcal/mol)	# of data points in the accumulated database	Exp-DB test set RMSE (kcal/mol)	# of data points in the accumulated database	Exp-DB test set RMSE (kcal/mol)	# of data points in the accumulated database
0 ^a	0.66	11,637						
1	0.59	195,069	0.64	228,488	0.68	33,290	0.68	33,290
2	0.60	260,902	0.65	286,984	0.65	41,417	0.71	41,417
3	0.60	302,462	0.65	319,845	0.70	45,757	0.68	45,757
35^b	0.50^b	932,509^b						
36	0.53	936,173	0.54 ^c	937,531 ^c				

37	0.55	939,453	0.56 ^c	941,048 ^c	
38	0.56	942,686	0.53 ^c	944,004 ^c	

^aThe Teacher model trained using only **Exp-DB**. ^bThe best-case model. ^cControl-2 scheme was applied, starting from the Student 35 model and 35 accumulated **Aug-DBs** obtained from the original SSD.

Figure S3 shows the other Control models devised for comparing the accuracies of GNN models trained using both **CombiSolv-QM** and **QM-DB** (Control-2) or **QM-DB** only (Control-3A and Control-3B). **CombiSolv-QM** and **QM-DB** are described in **Figure 6.1A**. In Control-2, three sets of solubility values were compared during the SSD cycles (**A**: Student-predicted, **B**: Calculated using COSMO-RS, **C**: Calculated using SMD-M06-2X/def2-TZVP). If at least one of $|A-B|$, $|A-C|$, or $|B-C|$ is below the 0.2 kcal/mol threshold, the corresponding Student-predicted value is added to the Nth **Aug-DB** in Cycle N. Control-2 was to examine whether incorporating multiple theoretical methods can lead to more extensive databases with improved accuracies. On the other hand, Control-3A carried out the SSD using only the SMD-M06-2X/def2-TZVP solubility values (**C**) stored in **QM-DB** to investigate the feasibility of employing DFT with implicit solvation models during the SSD. The training using **QM-DB** DG_{solv} values without the SSD was also performed (Control-3B).

Table S8 summarizes the model accuracies and database sizes of the above three Control models and compares them with those obtained from the SSD depicted in Fig. 2A. For Control-2, three cycles were conducted, starting from the Teacher model. The number of data points in Cycle 3 is higher than that from the original SSD (319,845 vs. 302,462). All the resulting three models showed slightly lower **Exp-DB** test set RMSEs (0.64 – 0.65 kcal/mol) than that from Teacher (0.66 kcal/mol) but higher RMSEs than those from the original SSD (0.59 – 0.60 kcal/mol). We also applied the Control-2 scheme to the best-case

Student 35 model and proceeded with three Control-2 cycles to seek possibilities of improving the accuracy from Student 35. More data points were added compared to the original SSD referencing only the COSMO-RS solubilities; however, no further decreases in RMSEs were observed from Student 35. The results from Control-2 indicate that referring to the single largest database, **CombiSolv-QM**, most effectively augments the database and improves the accuracy of models from the SSD.

Meanwhile, Control-3A and Control-3B were examined. In Control-3A, the sizes of **Aug-DBs** are less extensive than the SSD using **CombiSolv-QM**, and the lower accuracies were observed when **QM-DB** was employed for the SSD. Simply combining **Exp-DB** and **QM-DB** did not also improve the accuracy; the same trends were observed in the Control models (Fig. 3) where **Exp-DB** and **CombiSolv-QM** solubilities were combined without the SSD. These results demonstrate the importance of utilizing a large and comprehensive computational solubility database when conducting SSD. Although **QM-DB** did not show higher accuracies, possibly due to the less extensive database (220,332 vs. 1,000,000 for **QM-DB** vs. **CombiSolv-QM**), further SMD-DFT calculations can potentially lead to better accuracies. Moreover, the current **QM-DB** was useful in analyzing the strengths and weaknesses of COSMO-RS, SMD-DFT, and GNN models (Fig. 8 and **Error analysis of solubility prediction** Section in the Main Text). Such analysis informs the possible future work where heterogeneous data sources are adopted for developing predictive models.

S4. Results of training graph neural networks with noisy student self-distillation (NSSD)

Table S9. Comparison of accuracies for SSD and NSSD models.

Model description	Mean absolute error (kcal/mol, Exp-DB)			Mean absolute error (kcal/mol, Aug-DB-1)		
	Training	Validation	Test	Training	Validation	Test
Student 1, SSD	0.08	0.23	0.25	0.04	0.16	0.17
Student 1, 10% dropout, stochastic depth with survival probability of 90%	0.36	0.43	0.43	0.39	0.39	0.40
Student 1, 5% dropout, stochastic depth with survival probability of 20%	0.22	0.27	0.28	0.12	0.13	0.13

^a Survival probability (p) indicates the probability of the ‘final’ layer. The probability decreases gradually throughout the five global update layers with the same interval, based on the given formula: $\mathbf{v}_{i+1, \text{updated}} = \mathbf{v}_i + p_i \mathbf{v}_{i+1}$ where $p_i = 1 - i[(1-p) / 4]$ ($i = 0, 1, 2, 3, 4$), and $\mathbf{v}_i$ is the global feature vector at the i -th layer. For example, if $p=0.8$ (80%), The survival probabilities per each layer are: 1.0, 0.95, 0.9, 0.85, 0.8.

During the training using NSSD, noise is introduced to the model by applying dropout and stochastic depth methods to the hidden layers of the model. NSSD was effective in ML models for image classification because partially dropping the information from hidden layers would be helpful for handling the variance among different images with the same label. In this regard, we also tested multiple NSSD models in solubility predictions with different dropout rates (for all GNN layers and readout layers) and survival probabilities of stochastic depth (for the residual connection part of global feature vector). In all cases, NSSD showed higher prediction errors for **Exp-DB** than SSD (i.e., no noise was introduced to the model, **Table S9**). That is because dropout and stochastic depth can presumably cause errors in recognizing a molecule. The model can miss the information about key structural features related to solubility due to introducing noise to the model. In contrast, for images, if some part is lost, the model can still recognize and classify them. As a result, the SSD

method was chosen throughout this study instead of NSSD for the development of self-evolving solubility databases and GNNs.

S5. Supplementary information for Application 1 – Linear Free Energy Relationships between solvation free energy and reaction rates of organic reactions

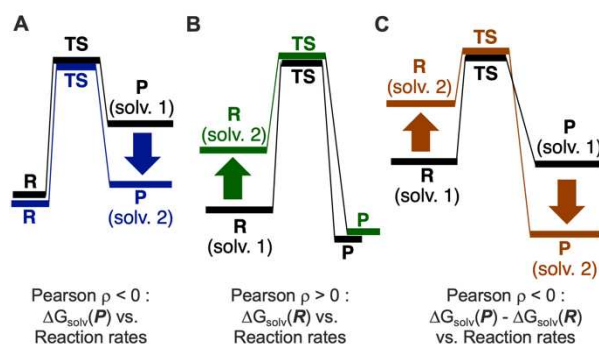


Fig. S4 Schematic energy diagrams of the reactions where (A) higher solvation stabilization of the product leads to more product formation, and (B) higher solvation destabilization of the reactant leads to faster reaction. (C) Designing new solvents can also affect both relative free energies of both reactant and product and thus the reaction rates.

Finding linear free energy relationships between solvation free energy and reaction rates based on the Hammond Postulate

In principle, a negative Pearson correlation should be found for $\Delta G_{\text{solv}}(\mathbf{P})$ vs. reaction rates for reactions in which the preferential stabilization of the product(s) plays a key role in lowering the activation barrier by increasing the thermodynamic driving force (**Figure S4A**). Meanwhile, for some reactions, higher reaction rates can be achieved by selective destabilization of reactants by a solvent (**Figure S4B**). Such a correlation leads to a positive Pearson r between $\Delta G_{\text{solv}}(\mathbf{R})$ and reaction rates. The cases in Figs. S4A and S4B mainly occur when the structure of a transition state is analogous to that of reactant(s) and product(s) of an elementary reaction, respectively, according to the Hammond Postulate. In addition, product stabilization and reactant destabilization can be considered together by using $\Delta G_{\text{solv}}(\mathbf{P}) - \Delta G_{\text{solv}}(\mathbf{R})$ as a descriptor (**Figure S4C**). Changing the sign of $\Delta G_{\text{solv}}(\mathbf{R})$ and adding to ΔG_{solv}

(**P**) enables the quantification of the influences on the reaction rates by both reactants and products, resulting in a negative Pearson r with reaction rates.

Chemical explanation of the linear relationships shown in Fig. 9

The reaction **I** is the dissociation of tert-butylperoxyaldehyde into tert-butyl alcohol and CO₂. Our results suggest that solvents with higher polarity (nitrobenzene, nitromethane, and chloroform) better stabilize the tert-butyl alcohol product than non-polar solvents (benzene, tetrachloromethane and heptane) and thus show higher rates. Lower Pearson r values were obtained from the other two reactions compared to **I**, but they display good negative correlations except for one solvent (MeCN and toluene for **II** and **III**, respectively). Of note, reaction **V** is analogous to **II** except for having more polar reactants than **II**. In this case, using non-polar solvents such as toluene show the most reactant destabilization and the highest reaction rate. The effect of different functional groups for the same reaction was captured by our GNN model, leading to the identification of a strong positive correlation ($r=0.99$). In contrast, high reaction rates were achieved when polar solvents such as water or ethanol were used with non-polar reactants (Br₂, pentene, and cyclopentadiene) for reactions **VI** and **VII**, respectively.

Reaction **VIII** is a ring opening to decarboxylate the reactant and form an alkene whose reaction rates were measured in five solvents. A non-polar solvent, decalin, shows the lowest reaction rate, whereas the fastest reaction was observed in a polar N-phenylformamide solvent. This is consistent with the fact that the zwitterionic product (**P**) is more polar than

the reactant (**R**), so a polar solvent would be favorable to stabilize the product more than the reactant. The Cope rearrangement (**IX**), in five different solvents was also investigated. Two solvents with hydroxyl groups (ethylene glycol and phenol) showed higher reaction rates than other solvents. This is because the ketone group in the product can form hydrogen bonds with alcoholic solvents, leading to product stabilization and faster reactions.

Table S10. Pearson correlation coefficients between experimental reaction rates vs. all three ΔG_{solv} descriptors for 11 organic reactions, with the reason of choosing one descriptor.

Reaction #	Descriptors			Explanation
	$\Delta G_{\text{solv}}(\mathbf{P})^a$	$\Delta G_{\text{solv}}(\mathbf{R})^b$	$\Delta G_{\text{solv}}(\mathbf{P}) - \Delta G_{\text{solv}}(\mathbf{R})^c$	
I	-0.95	-0.66	-0.63	Reactions can be driven by the thermodynamic stability of products and can be either single or multiple steps depending on the solvent.
II	-0.90	-0.95	0.99	
III	-0.68	0.50	-0.76	$\Delta G_{\text{solv}}(\mathbf{P}) - \Delta G_{\text{solv}}(\mathbf{R})$ is the best descriptor, but $\Delta G_{\text{solv}}(\mathbf{P})$ was chosen because the reactant contains nitroso group (R-N=O) which rarely appears in the database used for model training.
IV	0.94	0.94	0.82	Not a single-step reaction, so $\Delta G_{\text{solv}}(\mathbf{P}) - \Delta G_{\text{solv}}(\mathbf{R})$ is less reliable.
V	0.97	0.99	-0.95	$\Delta G_{\text{solv}}(\mathbf{P}) - \Delta G_{\text{solv}}(\mathbf{R})$ also shows a comparably strong correlation, but the reactant destabilization by solvents may be a key factor.
VI	0.91	0.91	-0.59	Not a single-step reaction, so $\Delta G_{\text{solv}}(\mathbf{P}) - \Delta G_{\text{solv}}(\mathbf{R})$ is less reliable.
VII	0.93	0.80	0.71	The reactant rapidly dimerizes even in mild conditions (room temperature), and the product is stable in most conditions except at high temperatures (> 150°C). Therefore, the reactant destabilization by solvents may be a key factor.
VIII	-0.49	-0.39	-0.99	Reactions where the kinetics are highly consistent with the Hammond postulate, and thus the generic $\Delta G_{\text{solv}}(\mathbf{P}) - \Delta G_{\text{solv}}(\mathbf{R})$ descriptor shows the best correlation. Also, all of them are elementary reactions.
IX	-0.61	0.09	-0.95	
X	-0.07	0.64	-0.80	
XI	-0.68	-0.59	-0.84	

^a Close to -1 if product stabilization plays a key role for higher reaction rates.

^b Close to 1 if reactant destabilization plays a key role for higher reaction rates.

^c Close to -1 if both product stabilization and reactant destabilization play a key role for higher reaction rates.

References

1. F. H. Vermeire and W. H. Green, *Chem. Eng. J.*, 2021, **418**, 129307.