

# Linear Graphlet Models for Accurate and Interpretable Cheminformatics

Michael Tynes,<sup>\*,†,‡,¶,§</sup> Michael G. Taylor,<sup>†</sup> Jan Janssen,<sup>†</sup> Daniel J. Burrill,<sup>†,‡</sup>  
Danny Perez,<sup>†</sup> Ping Yang,<sup>\*,†</sup> and Nicholas Lubbers<sup>\*,||</sup>

<sup>†</sup>*Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA*

<sup>‡</sup>*Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM 87545, USA*

<sup>¶</sup>*Analytics, Intelligence, and Technology Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA (Current address)*

<sup>§</sup>*Department of Computer Science, The University of Chicago, Chicago, IL 60637, USA (Current address)*

<sup>||</sup>*Computer, Computational, and Statistical Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA*

E-mail: [mtynes@uchicago.edu](mailto:mtynes@uchicago.edu); [pyang@lanl.gov](mailto:pyang@lanl.gov); [nlubbers@lanl.gov](mailto:nlubbers@lanl.gov)

## Abstract

Advances in machine learning have given rise to a plurality of data-driven methods for estimating chemical properties from molecular structure. For many decades, the cheminformatics field has relied heavily on structural fingerprinting, while in recent years much focus has shifted leveraging highly parameterized deep neural networks which usually maximize accuracy. Beyond accuracy, machine learning techniques need intuitive and useful explanations for the predictions of models and uncertainty quantification techniques so that a practitioner might know when a model is appropriate to apply to new data. Here we show that linear models built on unfolded molecular-graphlet-based fingerprints attain accuracy that is competitive with the state of the art while retaining an explainability advantage over black-box approaches. We show how to produce precise explanations of predictions by exploiting the relationships between molecular graphlets and show that these explanations are consistent with chemical intuition, experimental measurements, and theoretical calculations. Finally we show how to use the presence of unseen fragments in new molecules to adjust predictions and quantify uncertainty.

## Graphical TOC entry

1. Induce subgraphs



2. Regress

$$F(\text{graph}) = \beta_{\text{blue}} + \beta_{\text{orange}} + \dots + \beta_{\text{yellow}}$$


✓ Accuracy

3. Combine



✓ Interpretability

4. Adjust


$$F(\text{graph}) = \beta_{\text{blue}} + \beta_{\text{orange}} + \dots + \beta_{\text{yellow}}$$

✓ Uncertainty Quantification

# 1 Introduction

Property prediction from molecular graphs is a long-established method that has evolved into a complex discipline, especially so following the recent revolutions in machine and deep learning (DL).<sup>1-3</sup> DL approaches such as message passing and graph neural networks have come to overshadow other machine learning (ML) methods due to their state-of-the-art performance across a variety of tasks.<sup>4-6</sup> However, these performance improvements have come at the expense of increased training cost and decreased model interpretability; addressing these drawbacks is an active area of research.<sup>7-13</sup> Considering the limitations of DL, we show here that prediction 2D fragments still has a place in the modern chemistry-applied ML arsenal by revisiting fragment generation, and constructing new approaches to model building, chemical explanation, and uncertainty quantification.

Model interpretability is often crucial in scientific ML contexts to increase trust in models, to understand when and why they fail, and to enhance scientific understanding. Methods for interpreting ML models for chemistry and materials problems are reviewed thoroughly in Refs 12,14, which noted that some models are intrinsically interpretable (*e.g.*, by examining the coefficients on each feature in a linear regression model), whereas some models are black boxes (*e.g.*, neural networks) to which some *post-hoc* interpretability method need to be applied. These *post-hoc* methods include local surrogates that examine the contribution of input features near a given input point<sup>15</sup> and similar local explanations called SHapley Additive exPlanations (SHAP) based on game-theory.<sup>16-18</sup> Methods like SHAP adapted to DNNs and new methods that explain network predictions through mathematical functions of network gradients have been applied to molecular prediction to attribute predictions to atoms,<sup>19</sup> bonds,<sup>20</sup> and higher-order molecular subgraphs.<sup>21</sup>

Some argue that because DNNs often provide the most accuracy across ML models and that numerous *post-hoc* DNN explanation methods exist, these methods should be favored over in-

trinsically explainable models, which are commonly thought to be comparatively weak predictors.<sup>12</sup> However, there is evidence that this perceived accuracy-interpretability tradeoff is often over-exaggerated and is sometimes orthogonal to observed trends.<sup>13</sup> Furthermore, many of these *post-hoc* black box interpretability methods have theoretical weaknesses that are empirically borne out by counter-intuitive and untrustworthy explanations. For example Many explanation methods are not locally Lipschitz continuous, meaning that extremely small perturbations in model input can yield large changes in explanations, a phenomenon which makes explanations appear inconsistent to a human observer.<sup>22</sup> Other *post-hoc* methods can be manipulated to produce arbitrary explanations.<sup>23</sup> These kinds of findings have lead to calls to use simpler, explainable models when possible.<sup>13</sup>

Intrinsically intrinsically explainable models often have comparable predictive power to black box models when constructed carefully.<sup>13</sup> This trend has been demonstrated recently in in a materials context in a set of experiments where interpretable (multi-)linear models applied material property prediction problems achieved accuracy close to state-of-the-art nonlinear approaches.<sup>24</sup> One of these models was constructed by counting atomic n-grams present in a crystal unit cell lattice and assigning coefficients to their presence inspired by the cluster expansion. An analogous representation for organic molecules is what we call the atom-induced molecular subgraph, or graphlet, representation, wherein a molecule is represented by constituent *n*-atom connected subgraphs. A similar representation was recently developed by Ref 25 and used to sample and characterize large chemical spaces of more than billions of molecules.<sup>26-28</sup> Here we show that a like representation can be combined with linear models for competitive and interpretable prediction.

Inspired by the many-body expansion (MBE),<sup>29-31</sup> we approximate molecular properties as functions of molecular graphlets organized by their body-order, *i.e.*, the number of atoms in each graphlet. We show that so constructed linear models perform competitively

with nonlinear blackbox models across a variety of structure-property prediction tasks. We then show that this approach can naturally be used to produce coherent additive explanations by projecting higher-order model coefficients onto atoms or bonds within their molecular context, and empirically find correlation between these projections and both chemical intuition and chemical theory. We then examine how graphlet train and test statistics can be used to estimate distribution shift<sup>32</sup> and thereby quantify prediction uncertainty.

## 2 Methods

Using the principle of the many-body expansion (MBE), we aim to write a property of a molecule as a linear combination of coefficients associated with all of the graphlets of the molecule weighted by their number of occurrences in the molecular graph. This is illustrated in Figure 1. Here we outline the mathematical framework for constructing and counting these fragments (Section 2.2). We then discuss hierarchical regression (Section 2.3), show how graphlet coefficients can be combined to give model explanations (Section 2.4), and finally, how the presence of unseen fragments in molecules not seen during training can be used to both adjust model predictions and quantify uncertainty (Section 2.5). Finally, we give a few remarks on our implementation (Section 2.7), which is open-source and freely available.

### 2.1 Graphlet Fingerprint Approach

Graphlets are defined as isomorphism classes of connected subgraphs induced by choosing a set of nodes and all of the edges connecting those nodes in a graph. We define a graphlet fingerprint as a vectors of counts of occurrences of graphlets in a molecular graph. This is similar to other fingerprinting approaches, which enumerate molecular subgraphs of a given family up to some size. For example, so-called Daylight-like fingerprints enumerate linear paths on the molec-

ular graph up to a maximum path length with optional path branching;<sup>33</sup> Extended Connectivity FingerPrints (ECPF or Morgan, the latter after Morgan’s original formulation<sup>34</sup>) enumerate atom-centered radial subgraphs up out to a maximum radius.<sup>35</sup> In contrast, rather than restricting the process that generates subgraphs, we base our machine learning counts of all molecular graphlets up to some size  $N$ .

Graphlet fingerprints, like other fingerprint approaches, are built upon pre-defined type labels assigned to the atoms and bonds in a molecular graph. We label nodes by atomic species, formal charge, and aromaticity. This implies that every node type has an precise degree which is constant across all instances in all molecular graphs. Edges are labeled according to bond type as either single, double, triple, or aromatic. As a result, the graphlet statistics form a weighted version of  $D^k$  degree statistics.<sup>36</sup> This rich typing scheme helps to model information more efficiently at lower maximum graphlet size; as an example, with only species based typing an N+ atom with 4 bonds is not distinguished from an N atom with 3 bonds until using a graphlet size of at least 5. Including more information in atom and bond labels allows identification of such chemically distinct systems at far smaller graphlet sizes (in this case, a size of one).

Graphlets are enumerated by a recursive algorithm similar to the explicit subgraph enumeration routine described in Refs. 37 and 25, during which we identify and count membership in isomorphism classes through our hashing technique described in Equation 5.

### 2.2 Graphlet Fingerprint Mathematical Description

More mathematically, we construct graphlet fingerprints as histograms of members of induced-subgraph isomorphism classes present in a molecular graph as follows. We consider a molecular graph to be set of atoms and a set of bonds between those atoms, that is  $\mathcal{M} = (\mathcal{A}, \mathcal{B})$ . We label a subgraph of  $\mathcal{M}$  induced by choosing a a subset of atoms  $\mathcal{S} \subseteq \mathcal{A}$  and all of the bonds between them as  $\mathcal{M}[\mathcal{S}]$ .

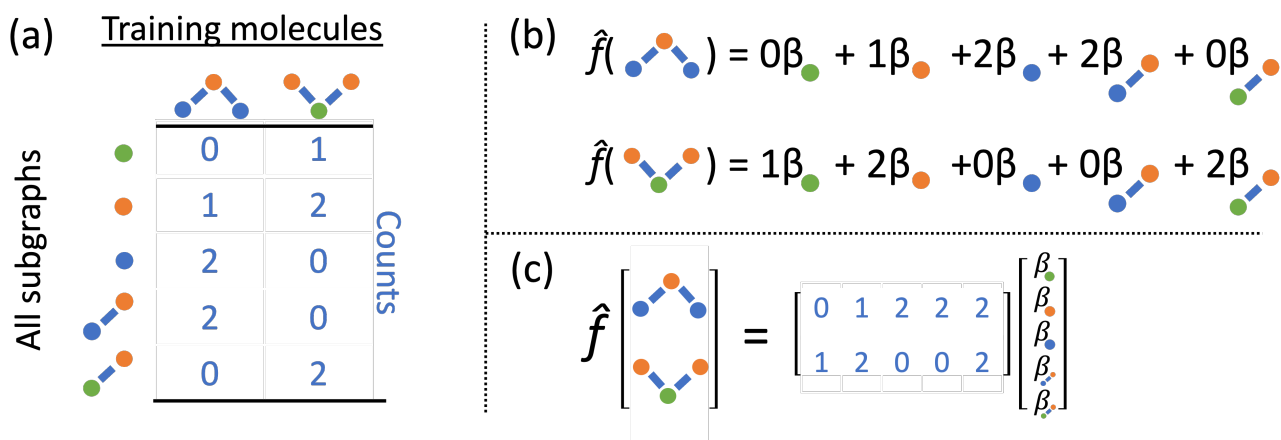


Figure 1: Illustration of graphlet featurization and linear model construction (a) All induced subgraphs up to size 2 are counted in a set of 2 training molecules (b) form of a linear model fit to predict some molecular property from counts shown in (a). (c) The matrix formulation of (b)

We denote the family of sets of atoms for whom the induced subgraph is connected as

$$\mathcal{P}(\mathcal{M}) = \{\mathcal{S} \subseteq \mathcal{A} : \mathcal{M}[\mathcal{S}] \text{ is connected}\}. \quad (1)$$

Loosely speaking, one can think of  $\mathcal{P}(\mathcal{M})$  as playing the role of a cumulant expansion over the induced graphs formed by the power set of  $\mathcal{A}$ . It will be useful to consider graphlets restricted to a given number of atoms,  $N$ , as

$$\mathcal{P}^N(\mathcal{M}) = \{\mathcal{S} : \mathcal{S} \subseteq \mathcal{A}, |\mathcal{S}| \leq N\}. \quad (2)$$

The graphlet fingerprint is the histogram  $\mathbf{C}^N$  of the graphlets with respect to isomorphism classes labeled by  $\mathcal{H}$  up to subgraph size  $N$ . With the Iverson brackets  $\llbracket \cdot \rrbracket$  representing the indicator function, the components of  $\mathbf{C}^N$  are

$$\mathbf{C}_{\mathcal{H}}^N(\mathcal{M}) = \sum_{\mathcal{S} \in \mathcal{P}^N(\mathcal{M})} \llbracket \mathcal{M}[\mathcal{S}] \in \mathcal{H} \rrbracket. \quad (3)$$

To efficiently track graphlet isomorphism classes  $\mathcal{H}$  we build an integer-valued labeling function  $H$  and produce counts of these labels. That is, we construct a concrete histogram  $\mathbf{c}$  with components labeled  $h$  as

$$\mathbf{c}_h^N(\mathcal{M}) = \sum_{\mathcal{S} \in \mathcal{P}^N(\mathcal{M})} \llbracket H(\mathcal{M}[\mathcal{S}]) = h \rrbracket \quad (4)$$

To do so we require pre-defined atom labels,  $h_{\text{atom}}(\mathcal{M}[\{i\}])$ , and pair labels,  $h_{\text{bond}}(\mathcal{M}[\{i, j\}])$ , as well as a labeling function  $h_{\text{rec}}$ , which can identify a histogram  $\mathbf{c}_h$  by sorting the labels in the histogram, pairing them with the accompanying count. Finally, we construct a recursive labeling function  $H$  with  $h_{\text{rec}}$  using  $h_{\text{atom}}$  and  $h_{\text{bond}}$  as base cases:

$$H(\mathcal{M}[\mathcal{S}]) = \begin{cases} h_{\text{atom}}(\mathcal{M}[\mathcal{S}]), & |\mathcal{S}| = 1 \\ h_{\text{bond}}(\mathcal{M}[\mathcal{S}]), & |\mathcal{S}| = 2 \\ h_{\text{rec}}(\mathbf{c}^{|\mathcal{S}|-1}(\mathcal{M}[\mathcal{S}])), & |\mathcal{S}| \geq 3. \end{cases} \quad (5)$$

In plain words, we label graphlet isomorphism classes by their own histograms of induced graphlets: triplets are labeled in terms of bonds and atoms, and four-point graphlets are labeled in terms of triplets, bonds, and atoms, *etc.* Whether or not the concrete labeling function  $H$  is a faithful realization of the abstract isomorphism classes  $\mathcal{H}$  is a complex question related to the long-standing Graph Reconstruction Conjecture,<sup>38–41</sup> which is notably true for some particular classes of graphs, false for others, and not settled for a great many cases. For the molecules and subgraph sizes studied here we have found no counterexamples.

## 2.3 Hierarchical Regression

We explore fitting linear regression models hierarchically, first to subgraphs with  $|\mathcal{S}| = 1$ , and then  $|\mathcal{S}| = 2$ , and so on. Let us think of the graphlet histograms  $\mathbf{c}^N$  of graphlets up to  $|\mathcal{S}| = N$  for a molecule as members of a space  $C^N$ . This space can be decomposed as a direct sum of vector spaces  $V^n$

$$C^N = \bigoplus_{n=1}^N V^n. \quad (6)$$

where the components of vectors  $\mathbf{v}^n \in V^n$  are counts of graphlets of size precisely equal to  $n$ . Using this notation, we construct an order- $N$  hierarchical model to predict  $y$  as

$$y \approx F_N(\mathbf{c}^N) = \sum_{n=1}^N f_n(\mathbf{v}^n) \quad (7)$$

Using  $\hat{Y}_n = F_n(\mathbf{c}^n)$ , each constituent model  $f_n$  is trained to minimize the same loss function evaluated against the residual  $y - \hat{Y}_{n-1}$ , that is to minimize  $\mathcal{L}(y_n, f_n(\mathbf{v}^n))$  with

$$y_n = \begin{cases} y, & n = 1 \\ y - \hat{Y}_{n-1}, & n > 1. \end{cases} \quad (8)$$

Put in less mathematical words, using the graphlet approach, we can build a function up by first applying regression to the graphlet counts generated by atoms, and then to the graphlet counts generated by bonds, and then to the graphlet counts generated by connected triples, and so on, up to some graphlet size  $N$ , where the model at size  $n$  learns a correction to the model at size  $n - 1$ . This same hierarchical approach can be analogously applied to the other 2D graph fingerprints we examine. For path-based fingerprints, the hierarchical levels indexed by  $n$  correspond to the number of steps in the graph walks or, equivalently, the number of bonds. For circular fingerprints, the hierarchical levels  $n$  indicate the set of fragment features with radius equal to  $n$ .

## 2.4 Interpretation Projections

We produce local (per molecule) interpretations of our graphlet-based linear models by exploiting the inclusion of smaller graphlets within larger graphlets. Using the graphlet inclusion relationships in a particular molecular graph, we project the linear model coefficients associated with each graphlet onto the molecule’s atoms or bonds. The projected values describe the contribution of each atom or bond to the model prediction, given its context within the molecular graph. These atom- or bond-projected values sum to the prediction value on this molecule. A visualization of the inclusion relationship structure is presented in Figure. 2. The remainder of this section describes how we produce the pictured graph and use it to perform projections in formal notation.

We consider the directed acyclic graph (DAG) of inclusion relationships between graphlets of varying size, defined as

$$G_{SS'}(\mathcal{M}) = \{(\mathcal{M}[\mathcal{S}], \mathcal{M}[\mathcal{S}']) : \mathcal{S} \subset \mathcal{S}'\} \quad (9)$$

and equivalently described by the adjacency matrix  $\mathbf{G}$  with elements given by

$$G_{SS'}(\mathcal{M}) = \llbracket \mathcal{S} \subset \mathcal{S}' \rrbracket. \quad (10)$$

For brevity, we will omit the  $\mathcal{M}$  and write this matrix as  $G_{SS'}$ , but the matrix remains associated with a particular molecule.

We principally we deal with the inclusions of size  $n$  graphlets within size  $n + 1$  graphlets, which form an  $N$ -partite DAG with partitions for each graphlet size from  $n = 1, \dots, N$ . Moving forward, we will call the partitions levels. Much like in a feed-forward neural network, each adjacent pair of levels is connected by a set of edges. These edges are a subset of those in  $\mathbf{G}$ : the adjacency matrix  $\mathbf{G}^n$  connecting nodes from level  $n + 1$  to level  $n$  corresponds to

$$G_{SS'}^n = G_{\llbracket |\mathcal{S}|=n \rrbracket, \llbracket |\mathcal{S}'|=n+1 \rrbracket}, \quad (11)$$

where the Iverson brackets  $\llbracket \cdot \rrbracket$  subscripts indicate taking only rows and columns of  $\mathbf{G}$  that respectively correspond to size  $n$  and size  $n + 1$  fragments. A column of  $\mathbf{G}^n$  describes which

size  $n$  graphlets are included in each size  $n + 1$  graphlet. We use the matrices  $\mathbf{G}^n$  to perform our projections from higher to lower levels of the DAG. The DAG described by these matrices, and their upward analogs to be introduced shortly, is visualized for a fictitious molecule in Figure

We want our projections onto atoms or bonds to sum to the prediction of the linear model, so we want each  $\mathbf{G}^n$  to be sum conserving. To accomplish this, when projecting contributions from a size  $n + 1$  graphlet to its size  $n$  graphlets, we evenly distribute this contribution across all of the size  $n$  graphlets. Mathematically, we can ensure this by normalizing the columns of  $\mathbf{G}^n$  to sum to 1. We write the column-normalized adjacency matrix as  $\hat{\mathbf{G}}^n$  with columns defined as

$$\hat{\mathbf{g}}_{\mathcal{S}'}^n = \frac{1}{\mathbf{1} \cdot \mathbf{g}_{\mathcal{S}'}^n} \mathbf{g}_{\mathcal{S}'}^n \quad (12)$$

where  $\mathbf{1}$  is the vector of ones.

A linear model with weights  $\beta$  acting on graphlet histogram  $\mathbf{c}^N$  to estimate  $y$  is written as,

$$\hat{y} = \beta \cdot \mathbf{c}^N. \quad (13)$$

Here, every model coefficient  $\beta$  is associated a graphlet isomorphism class and is multiplied by the number of occurrences of that graphlet class in a molecule being summed. We can think of this model as a sum over the coefficients associated with every individual occurrence of a graphlet induced by  $\mathcal{S}$  in a molecule, written as

$$\hat{y} = \sum_{\mathcal{S} \in \mathcal{P}^N(\mathcal{M})} \beta[\mathcal{S}]. \quad (14)$$

When projecting “downwards” from larger to smaller fragments, denote the projection value on a set of atoms as  $\alpha[\mathcal{S}]$ . We write the vector of these  $\alpha$  and  $\beta$  values associated with all  $\mathcal{S}$  of size  $n$  in a molecule as  $\vec{\alpha}^n$  and  $\vec{\beta}^n$ . With this notation, we define the projection from level  $n + 1$  to  $n$  as

$$\vec{\alpha}^n = \vec{\beta}^n + \hat{\mathbf{G}}^n \cdot \vec{\alpha}^{n+1} \quad (15)$$

with the recursive base case

$$\vec{\alpha}^N = \vec{\beta}^N. \quad (16)$$

Equation 15 is sufficient to produce atom-level explanations by computing  $\vec{\alpha}^1$

For bond-level explanations, we introduce the reverse “upwards” projection from level  $n - 1$  to  $n$ . To do so, we reverse the direction of the edges in the DAGs described above. The edges are weighted by the total “valence” (in graph terms, the total edge weight) of one graphlet within another. Loosely speaking these valence weights are the counts of bonds subsumed in the larger graphlet. More formally, the matrix  $\mathbf{K}$  with elements given by

$$K_{\mathcal{S}'\mathcal{S}} = \llbracket \mathcal{S}' \subseteq \mathcal{S} \rrbracket \sum_{b \in \mathcal{B}} \llbracket b \in \mathcal{S} \rrbracket \llbracket b \notin \mathcal{S}' \rrbracket w_b \quad (17)$$

where  $w_b$  gives the weight of an ordinary edge (bond) in the molecule,  $b$ . Note that the sparsity structure of  $\mathbf{K}$  is the same as  $\mathbf{G}^\top$  (another way of observing that the DAG is reversed) and only the edge weights differ. We then define  $\mathbf{K}^n$  is analogously to  $\mathbf{G}^n$  to only have support between levels  $n$  and  $n - 1$ . We then column-normalize  $\mathbf{K}^n$  in the same sense as Eq. 12, producing  $\hat{\mathbf{K}}^n$  with columns summing to 1. In the case where  $n = 2$ , edges toward atom pairs connected by integral bonds are weighted by the number of electron pairs, and aromatic bonds have weight of  $\frac{3}{2}$ . This allows us to define natural “upward” projections  $\omega[\mathcal{S}]$  as

$$\vec{\omega}^n = \hat{\mathbf{K}}^n \cdot (\vec{\omega}^{n-1} + \vec{\beta}^{n-1}) \quad (18)$$

$$\vec{\omega}^1 = 0. \quad (19)$$

A diagram of this scheme for interpretability projections is shown in Figure 2. As a concrete realization of this definition, a 2-graphlet (bond) containing a carbon and nitrogen that are double-bonded will receive  $\frac{1}{2}$  of the carbon-associated  $\beta$  and (2 out of 4 total bonds) and  $\frac{2}{3}$  of the nitrogen-associated  $\beta$  (2 out of 3 bonds) in an upward projection.

We combine the upward and downward projections at level  $n$ , denoted by  $\vec{\chi}^n$ , as

$$\vec{\chi}^n = \vec{\alpha}^n + \vec{\omega}^n. \quad (20)$$

For any level  $n$ ,

$$\hat{y} = \vec{\mathbf{1}} \cdot \vec{\chi}^n. \quad (21)$$



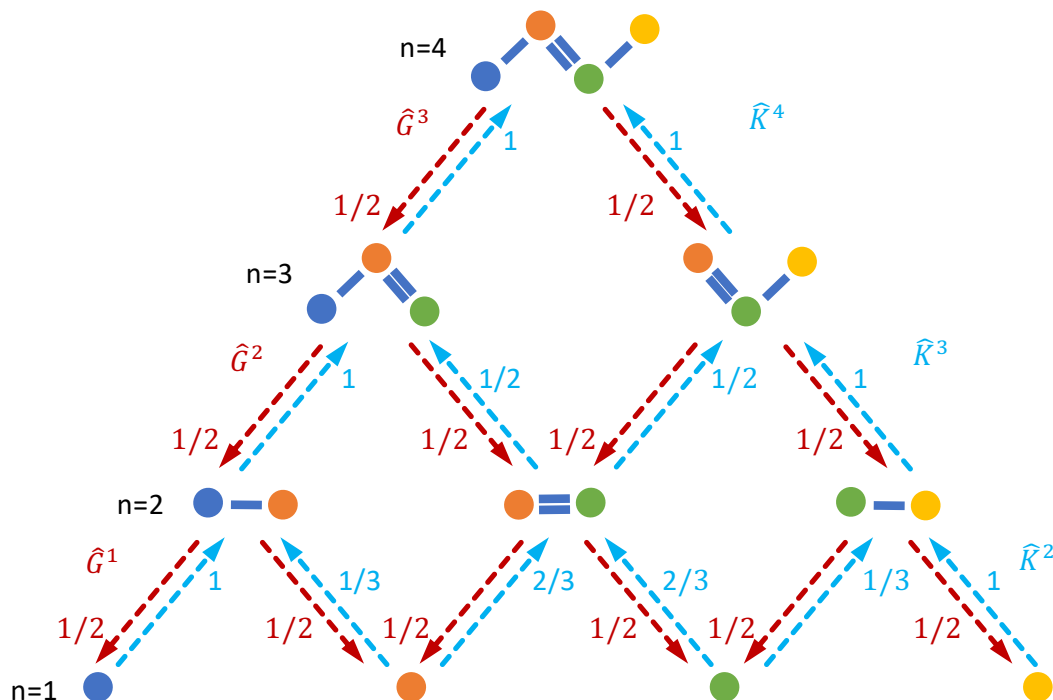


Figure 2: Illustration of interpretability scheme based on substructure graphs. A linear model associates some contribution to each fragment in the molecule. By tracking the inclusion relationships between subgraphs (red/blue arrows), we can create normalized matrices  $\hat{\mathbf{G}}^n$  and  $\hat{\mathbf{K}}^n$  which can be used to move predictions between many-body levels. Of particular interest are interpretability projections to atoms,  $n = 1$ , and bonded pairs,  $n = 2$ .

We primarily consider  $\bar{\chi}^1$  which corresponds to breaking up the prediction  $\hat{y}$  into atom-centered terms and  $\bar{\chi}^2$  corresponding  $\hat{y}$  decomposed into bond-centered terms, although one can compute  $\bar{\chi}^n$  for any  $n = 1..N$ .

## 2.5 Uncertainty quantification and prediction adjustment based on unseen graphlets

When evaluating graphlet-based regression models on new molecules, it is likely that these molecules will contain graphlets not present during training. We can use the presence of these unseen fragments both to construct uncertainty quantification (UQ) measures and, when appropriate, adjust our predictions to account for systematic biases introduced by the absence of fitted model coefficients associated with the unseen graphlets. These uncertainty metrics can be useful in active learning<sup>42</sup> and Bayesian optimization<sup>43</sup> workflows to discover new molecules and materials.

We examine various methods of constructing uncertainty metrics based on unseen graphlets. While yet more approaches are possible, we explored using the total number of unseen graphlets, the fraction of unseen graphlets, and an auxiliary uncertainty regression models to weight the relative importance of unseen fragments of size  $s$ .

We can exploit statistical information in the distribution of graphlet coefficients to adjust predictions when a test molecule has unseen graphlets. We examine this in the context of predicting energies, where each graphlet is associated with a coefficient that can be thought of as the energy contribution of each graphlet. Ignoring unseen graphlets present in a new molecule causes the magnitude of molecule's energy to be mispredicted. We adjust for this bias by finding the mean coefficients  $\tilde{\beta}_s$  for each size  $s = 1, \dots, N$ , constructing a histogram of counts  $d_s^N$  of unseen all fragments of size  $s$ , regardless of the fragment structure. The adjusted prediction is then written as  $\hat{y}^{\text{adj}} =$

$$\beta \cdot \mathbf{c}^N + \tilde{\beta} \cdot \mathbf{d}^N.$$

## 2.6 Overview of data sources

We evaluate models built using graphlet fingerprints on fifteen molecular datasets from several sources to assess its general applicability.

To examine regression performance, we first examine prediction on atomization energies from the QM9<sup>44</sup> dataset and compare performance against a range of methods applied to this dataset by Ref. 45. We then evaluate our method's performance on solubility prediction using four datasets from Ref 46 and compare our results to those therein. Finally, we evaluate regression performance on nine drug discovery related quantities in datasets from Ref 47 and compare performance to leaderboards hosted online.<sup>48</sup>

To evaluate our interpretability projections, we qualitatively examine solubility predictions on the datasets from Ref 46 and, more quantitatively, correlate bond-projected energies to bond dissociation energies calculated on a set of molecules from Ref 49.

## 2.7 Implementation

We implemented our fingerprints using a custom python code, `minervachem`, which we have open-sourced and made freely available [link]. `Minervachem` uses `RDKit`<sup>50</sup> and `networkx`<sup>51</sup> to represent molecules and graphlets. Graphlet counts are represented as `scipy`<sup>52</sup> sparse matrices for model fitting. Linear and hierarchical modelling procedures in are implemented with `scikit-learn`.<sup>53</sup> Nonlinear models are implemented with Light Gradient Boosting Machine (LightGBM) library.<sup>54</sup> LightGBM model hyperparameters were optimized using the Fast Library for Auto Machine Learning (FLAML)<sup>55</sup> and include the number of boosted tree estimators, the maximum number of leaves per estimator, the maximum number of samples per leaf, the fraction of features considered by each tree, the learning rate, and L1 and L2 regression parameters. Visualizing projected coefficients uses `RDKit` plotting methods. Visualizing projection DAGs is done in `networkx`.

# 3 Results

Our results show competitive model predictive performance, strong interpretability, and uncertainty quantification that is well-correlated to absolute error. In Sec. 3.1, we see that graphlet-based linear models fit to DFT atomization energies exceed the performance of both linear and nonlinear models built on other fingerprints and exceed the accuracy of these DFT calculations with respect to experiment. In Sec. 3.2, we show that projecting coefficients from these models to bonds gives bond-level attributions that are correlated with DFT-derived bond dissociation energies. In Sec. 3.3, we see predictive performance on solubilities in various solvents that is competitive with nonlinear models from the literature, and in Sec. 3.4 that interpreting the coefficients from these models projected to atoms agrees with chemical intuition. In Sec. 3.6, we use information about unseen fragments to improve prediction quality on unseen molecules by up 38% in a fragment holdout experiment. Finally, in Sec. 3.7, we use unseen fragment information to construct uncertainty quantification metrics that show strong correlation with absolute prediction error.

## 3.1 High Accuracy on Diverse Chemical Systems

We present computational experiments on the QM9 dataset that are designed to evaluate the performance of graphlet fingerprint-based linear models compared to other fingerprinting methods and to nonlinear modeling approaches. We restricted the target QM9 property to atomization energy as a case study to examine our graphlet-based linear modeling approach, as our approach was inspired by many-body expansion energy models. For linear regression models, we fit  $L_2$ -regularized models on graphlet fingerprint representations with maximum graphlet size  $s$  ranging from 1 to 9. For comparison with a nonlinear model, we included gradient boosting as implemented by the LightGBM library.<sup>54</sup> To compare with other fingerprinting methods, we included `RDKit` and Morgan fingerprints as implemented

in the RDKit library,<sup>50</sup> using count-based unfolded representations with branching allowed for RDKit fingerprints. We also varied  $s$  from 1 to 9 for these fingerprints, where  $s$  is the maximum number of bonds present in an RDKit fingerprint and the maximum radius in a Morgan fingerprint. At each fragment size, we performed a hyperparameter optimization and measured test performance on a .64/.16/.2 train/val/test split. For the ridge regression model we searched for the  $L_2$  strength parameter. For the LightGBM we used the FLAML procedure,<sup>55</sup> which aims to intelligently find hyperparameters (which are listed in Section 2.7) given a wall-time budget, which we set to 30min for each level in the hierarchical models and  $s \times 30\text{min}$  in the non-hierarchical case. We also compare our model to the best fingerprint-based model for the atomization energy task presented in Ref 45.

Overall, graphlet-based linear models show stronger performance than both nonlinear models and models constructed with other fingerprints, as seen in Figure 3 which summarizes our QM9 experiments. (Additional learning curves including training performance are given in Supplementary Fig.S1 and Fig.S2). First considering hierarchical models, performance improves consistently with fragment size to a mean absolute error (MAE) of less than  $5 \frac{\text{kcal}}{\text{mol}}$  for all models. The best of these models uses graphlet fingerprints and attains a test MAE of  $1.74 \frac{\text{kcal}}{\text{mol}}$ . Although this is not quite the  $1 \frac{\text{kcal}}{\text{mol}}$  widely considered to be chemical accuracy,<sup>56</sup> it is less than the error of the DFT used to calculate  $\Delta H_{at}$ .<sup>45</sup> The RDKit fingerprint-based model closely follows this performance, with an MAE of  $2.15 \frac{\text{kcal}}{\text{mol}}$ . The performance of these fingerprints is likely similar because they capture similar chemical information, *i.e.*, atom-induced subgraphs vs. branched-path fingerprints which can be thought of as edge-induced subgraphs. We hypothesize that this is because they are a richer representation than the Morgan fingerprints, having many more features at a given size, whereas the RDKit fingerprints, being walk-based, include multiple fingerprint elements that map to identical sets of atoms. graphlets out-perform the other fingerprints be-

cause they include a more complete representation of the possible substructures; Morgan fingerprints in many cases do not directly capture. The next best performing hierarchical model is the LightGBM with graphlet fingerprints, which is superior to the linear model at smaller fragment sizes but saturates in performance at around 1000 fragments. This is possibly because LightGBM does not capture the additivity of energies reflected in the many body expansion and thus cannot effectively leverage large numbers of graphlet fingerprint bits, a hypothesis that is consistent with the divergence of non-hierarchical LightGBM models with fragment count. This may also be attributed to the reduced number of fragments used at each fragment size within the hierarchical model compared to fitting on all fragments at once in the non-hierarchical case. Hierarchicality also eliminates feature correlation induced by the inclusion of smaller fragments within large ones.

All models constructed here outperform the best fingerprint-based model performance reported in Ref 45, most by 1-2 orders of magnitude, regardless of fingerprint type, highlighting the importance of carefully selecting fingerprint parameters. Ref 45 uses binary ECP4 (Morgan) fingerprints out to radius 4 folded to a fingerprint length of 1024. At radius 4, we find roughly 500,000 Morgan fragments in the training set when using unfolded fingerprints (see Supplementary table S1 for the exact number of fragments observed at each size for each fingerprint type). This close to 500x ratio between unique fragments and fingerprint entries likely explains the limitations of the model from Ref 45.

## 3.2 Interpreting energy models on bonds

Energy models built in Sec. 3.1 provides an opportunity to investigate whether the bond-level projections  $\bar{\chi}^2$  (defined in Eq. 20) of an energy prediction correlate with bond dissociation energies (BDEs). We examine the relationship between  $\bar{\chi}^2$  and both experimentally and theoretically-derived BDEs, in both cases

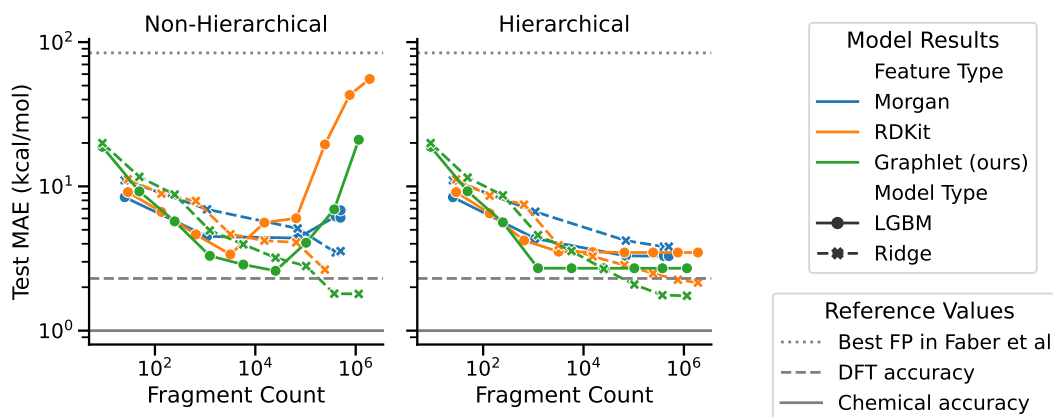


Figure 3: Test set performance vs. fragment count by model type, feature type, and hierarchicity on the QM9  $\Delta H_{at}$  task. The horizontal axis is shown in number of fragments rather than fingerprint size because the size parameter has different meanings across fingerprint types. The dashed horizontal line at  $2.3 \frac{\text{kcal}}{\text{mol}}$  indicates the accuracy of the DFT calculations that produced these  $\Delta H_{at}$  values compared to experiment.<sup>45</sup> The solid horizontal line at  $1 \frac{\text{kcal}}{\text{mol}}$  gives a common benchmark of “quantum chemical accuracy” at  $1 \frac{\text{kcal}}{\text{mol}}$ . The dotted horizontal line at  $84.2 \frac{\text{kcal}}{\text{mol}}$  shows the best MAE attained by fingerprint-based models in Ref 45.

using a linear model fit with graphlet fingerprints up to size 7 on approximately 128,000 molecules from QM9 (the details of the split construction are discussed in Section 3.2.2). To examine these

### 3.2.1 Experimental BDEs

We first consider a few experimental BDEs from simple molecules obtained from Ref 57. Both the bond-projected energy predictions  $\bar{\chi}^2$  and experimental BDEs for three example molecules—ethane, ethene, and ethyne—are shown in Figure 4. Both quantities, reported in  $\frac{\text{kcal}}{\text{mol}}$ , appear on the same scale. The expected trend of increasing C-C bond energy with bond order is captured. This can be explained largely in terms of the explicit single, double, and triple bond fragment coefficients. More interestingly, the subtle trend in C-H bond energy with C-C bond order is also partially captured. In this case, the proximity to the higher energy higher order bonds through the inclusion with these bonds in higher order graphlets, raised the values of  $\bar{\chi}_{\text{C-H}}^2$ .

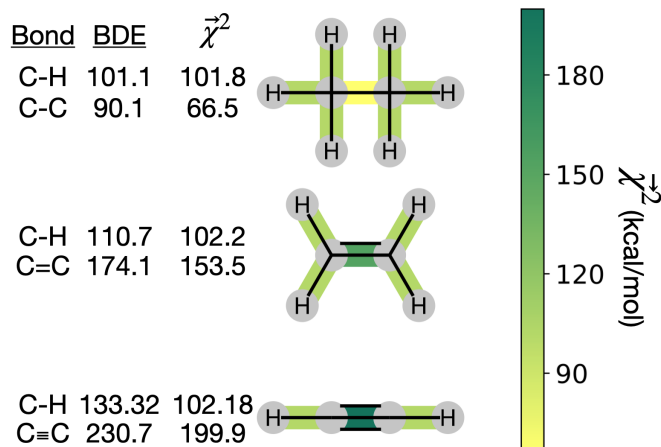


Figure 4: Bond-level model interpretability: The projection of a  $\Delta H_{at}$  model onto bonds ( $\bar{\chi}^2$ ) along experimental bond dissociation energies (BDE) for Ethane, Ethene, and Ethyne. Energies are in units of  $\frac{\text{kcal}}{\text{mol}}$ .

### 3.2.2 Theoretical BDEs

Though illustrative, recapitulating the relative strength bonds in only a few simple molecules is not a sensitive probe of the interpretability scheme. To ask whether bond-level projections are well-aligned with BDEs in a statistically significant sense, we turn to the large theoretical BDE dataset presented in Ref 49. This dataset includes single-bond BDEs for

roughly 40,000 molecules (and 200,000 associated BDEs) calculated at the M06-2X/def2-TZVP level of theory. Of these, approximately 5,000 molecules (with roughly 50,000 associated BDEs) are present in QM9. For consistency with QM9, we recalculated the dissociation energies of the bonds in these molecules at the B3LYP/6-31G(2df,p) level of theory. The calculations converged for over 99% of these roughly 50,000 bonds, and serve as a reference for our bond-projected predictions  $\bar{\chi}^2$ .

We construct a holdout set of BDEs to evaluate whether the correlation between  $\bar{\chi}^2$  is subject to a generalization gap. Half of the 5,000 molecules present in both the theoretical BDE dataset and QM9 were held out. We then trained a graphlet-based hierarchical linear model to all QM9 molecules except those present in this holdout set, using a maximum graphlet size of 7.

The bond-projected predictions  $\bar{\chi}^2$  from these models show reasonable agreement with the theoretically calculated BDEs, especially considering that only single bond dissociations are present. On the held out bonds, we attain a Pearson  $r$  of 0.46 over all of the bonds, shown in Figure 5. (Notably, there is very little generalization gap in these correlations, as shown in Supplementary Fig. S3 and Fig. S4) To test whether this correlation is driven by the relative strengths of single bonds between each element pair (an instance of Simpson’s paradox<sup>58</sup>), we separate the data by element pairs and compute the correlations, shown in Table 1. Within the element pairs, the strength of the relationship between  $\bar{\chi}^2$  and BDE varies widely with relatively strong performance for C-C bonds, weak performance for H-O bonds, and moderate performance for the remaining element pairs. Thus our interpretability scheme recapitulates trends even within (some) individual bond types. In particular, heavy-atom to heavy-atom bond energies are better correlated with the BDE in comparison to hydrogen-heavy-atom bonds; the variance the bond explanation for hydrogen atoms is noticeably smaller than the variance of bond explanations for heavy atoms. When interpreting these correlations, it is important to remember that this is a test of empirical cor-

relation between qualitatively similar phenomena; the model was not trained in any way to predict BDEs - rather, it predicts total energies, and the bond-wise interpretation of these predictions is significantly correlated to the BDE. The model is unaware of open-shell molecules, radicals, or ions that are produced in breaking those bonds.

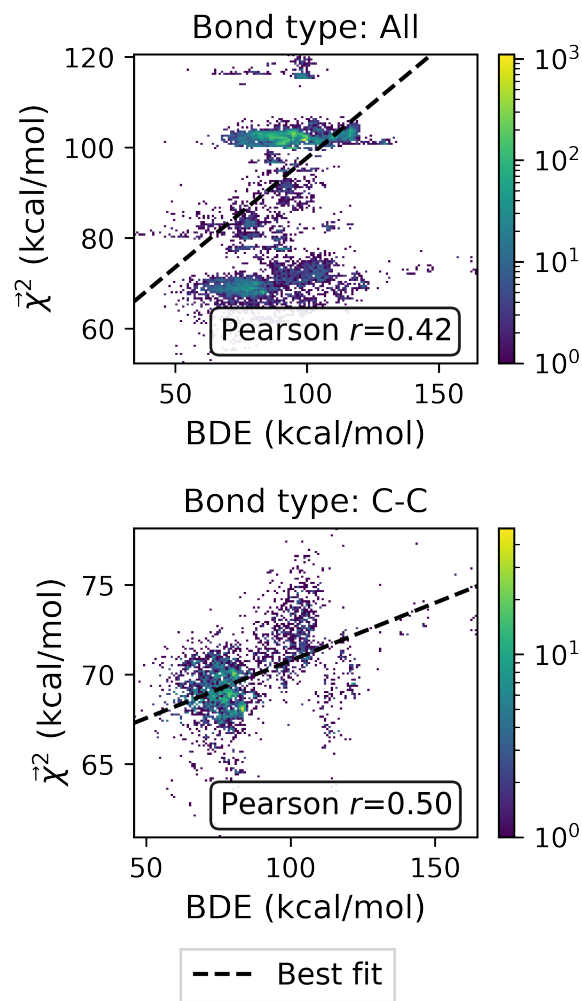


Figure 5: Relationship between Bond Dissociation Energies (BDEs) computed with DFT and the bond-level interpretations  $\bar{\chi}^2$  for a model fit to atomization energy on the QM9 dataset.

### 3.3 Competitive performance and model interpretation for solubility prediction

To evaluate the applicability of graphlet-based linear models beyond energy prediction, we

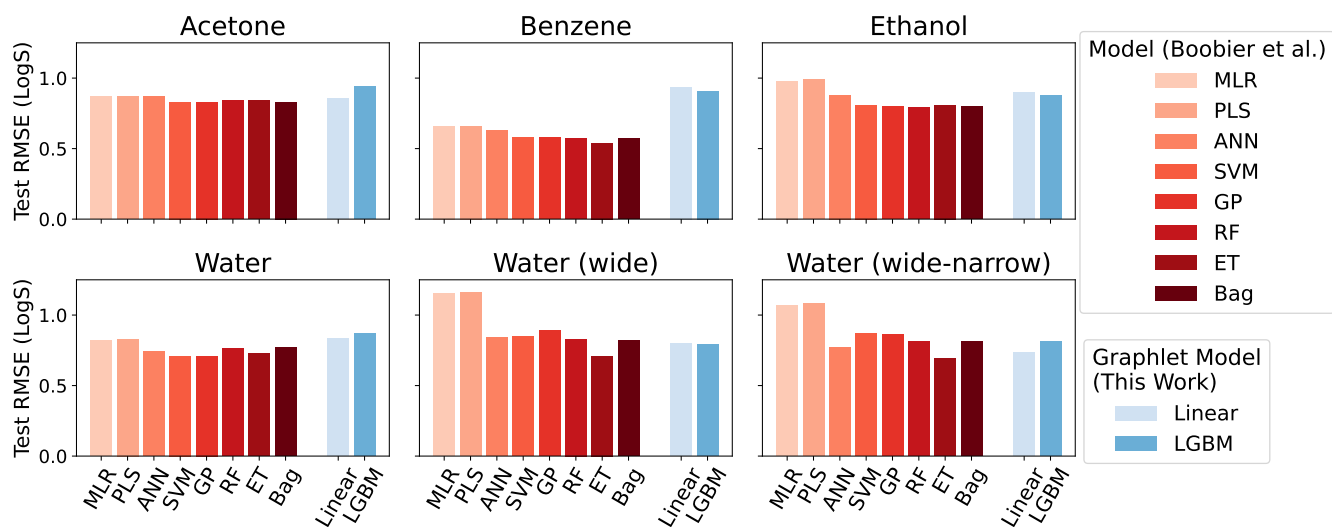


Figure 6: Model performance on the solubility datasets from Ref 46. Root mean squared errors are in units of log molarity. Models from Ref 46 are shown in shades of red in the left hand side of each panel, models from this work are shown in shades of blue, offset on the right hand side of each panel.

Table 1: Correlation coefficients between bond-level interpretations  $\bar{\chi}^2$  of the linear model and theoretical bond dissociation energies by element pair. Coefficients are calculated on molecules from the holdout set, and the number of bonds is denoted by  $n$ . Each correlation is statistically significant, with  $p < 10^{-12}$ , except for the smallest category of H-O bonds, for which the correlation is not significant ( $p > .05$ ).

| Bond | $n$    | Pearson $r$ |
|------|--------|-------------|
| All  | 24,695 | 0.4198      |
| C-H  | 15,980 | 0.1414      |
| C-C  | 4,699  | 0.4968      |
| C-N  | 1,064  | 0.3043      |
| C-O  | 1,440  | 0.2266      |
| H-N  | 990    | 0.2298      |
| H-O  | 522    | -0.0584     |

evaluate them on a dataset of hundreds of experimental log solubilities in four solvents: acetone, benzene, ethanol, and water, as presented in reference 46. We compare our model performance directly with ML model performance presented in reference 46 using the same datasets and prediction tasks: Molecular solubilities in each solvent are considered separate tasks and each have their own ML models, with

the water solubility prediction task broken into three tasks based on varying  $\log S$  cutoffs—(1) “Water”: only molecules with  $-4 < \log S < -1$  are included, (2) “Water (wide)”: all molecules are included, and (3) “Water (wide-narrow)”: only the test set is filtered to  $-4 < \log S < -1$ . We use the same train/test splitting procedure as reported in Ref 46 and further split the training set into an 80/20 train/val split to search optimize the maximum graphlet size and model hyperparameters. For linear models we searched for the  $L_2$  parameter in a range of  $10^{-5}$  to  $10^2$ , and LightGBM was optimized with FLAML with a time budget of 2 min.

Overall, we found that the graphlet fingerprints coupled with linear models predict small molecule solubility in four solvents competitively with nonlinear models from reference 46, which were trained on expensive DFT- and experiment-based features. Figure 6 shows test RMSE (log molarity) for each model presented in Ref 46 and for graphlet-based linear and LightGBM models.

Graphlet-fingerprint-based models are competitive on all datasets save for Benzene, and are among the best for the Water (wide) and Water (wide-narrow) sets, while being both inexpensive and easy to interpret based on structural motifs (see Figure 7 in Section 3.4).

Compared to nonlinear LightGBM models, linear models are unexpectedly strong on these tasks. This demonstrates that, surprisingly, our molecular representation coupled with linear models is useful outside of the context of predicting extensive properties like energy.

The interesting sub-result of improved performance of the graphlet-based models when moving from the Water to Water (wide-narrow) suggests a robustness to overfitting. In the former task, only molecules of log-solubility ranging from -4 to -1 are included. In the latter task, the test set is the same, but the training set additionally includes molecules in the wider log  $S$  range of -12 to 2. In principle, the test task is statistically identical, but in the wide-narrow version, more information is given for training. Most of the models from Ref 46 nonetheless perform worse on the test task in the wide-narrow version; the new information somehow confounds the models. In contrast, our graphlet approach behaves intuitively - when given more data, it makes strictly better predictions on the same test set.

We note again the relative expense of our approach to the models in Ref 46; because these models rely on features that involve Density Functional Theory and experimental measurements, applying the model to an arbitrary new chemical can be limited by the speed to calculate, find, or measure these quantities. In contrast, a fingerprinting approach such as graphlet fingerprints can be applied to completely new molecules in timescales far less than a second. For these tasks, there are on the order of hundreds to thousands of graphlet features; precise counts are given in Supplementary Table S2.

### 3.4 Atom-wise Interpretation of Solubility Models

Here, we examine the interpretability of graphlet fingerprints using linear models by computing the atom-projections of the predictions  $\bar{\chi}^1$  and examining the qualitative agreement between structural trends in the projections and chemical intuition about solubility. In particular, we choose propyl and benzyl back-

bones, by themselves and in combination with alcohol, amine, and chloro functional groups. Figure 7 shows the interpretation  $\bar{\chi}^1$ , that is, the atom-level-projected contributions from the solubility model. Note how each functional group contributes to the overall molecular solubility. As expected, alcohol and amine groups are shown to be responsible for increasing solubility, and chloro groups are responsible for lowering solubility. Supplementary Fig. S5 shows interpretations for additional molecules selected from the intersection of the Acetone and Water solubility datasets.

### 3.5 ADMET Leaderboards

To further assess general applicability of graphlet-fingerprint-based models, we evaluate their performance on nine drug-discovery-relevant regression tasks from the Therapeutic Data Commons (TDC).<sup>47,48</sup> These tasks include the prediction of variety of biochemical attributes relevant to drug drug design, including chemical properties such as lipophilicity and aqueous solubility along with human-biological observables such as toxicity and half-life in blood which are not chemically absolute. These properties are contained in datasets of roughly 1,000 to 10,000 molecules. We followed the same train/val/test split recommended by the TDC. This is a challenging generalization evaluation which holds out all molecules built upon a particular scaffold (molecular backbone) and conducting training and hyperparameter optimization on the remaining molecules. The performance is averaged over 5 random training/validation splits. We fit both the LightGBM and Ridge models with graphlet fingerprints.

Table 2 shows our performance compared to the existing leaderboard entries. A visualization of all of the models performance for all leaderboard tasks is present in Supplemental Fig. S6 and Fig. S7. Models using graphlet fingerprints score in the upper half of the leaderboard for seven out of nine of the tasks. Notably, all these high-scoring models used the LightGBM regressor; ridge regression did not perform impressively in these tests and was in

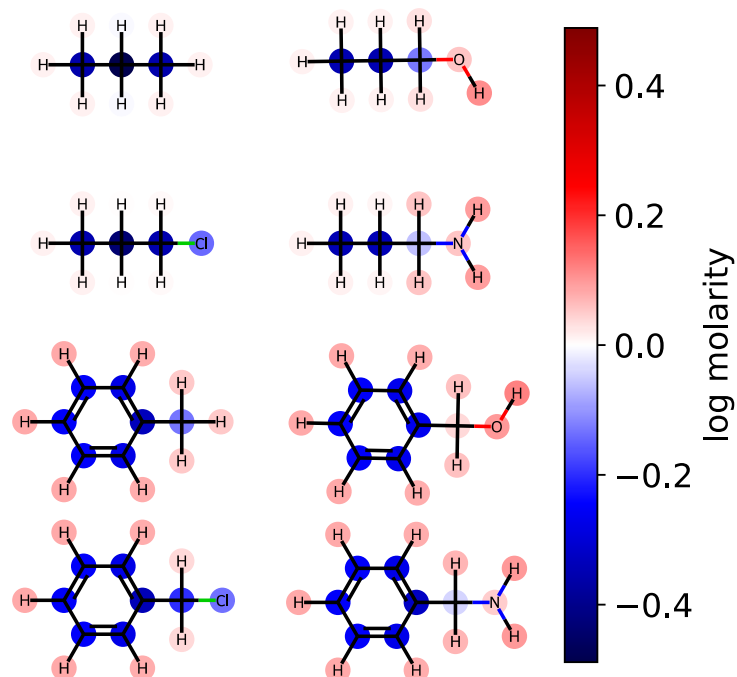


Figure 7: Linear model contributions projected to the atom level for selected backbones and functional groups. Colors show the contribution to the predicted log solubility (measured in molarity). Contributions of the functional groups to the overall solubility agree qualitatively with chemical intuition.

the lower half of the leaderboard for five of the nine tasks. This is surprising in the context of ridge regression’s strong performance on solubility prediction in Section 3.3, but less surprising in that we expect non-linear, non-extensive models to perform better on such properties.

### 3.6 Exploiting interpretability to account for new information

Here we evaluate the effectiveness of the adjustment based on unseen fragments discussed in Section 2.5. We conducted a series of experiments on the QM9 dataset holding out molecules with graphlets of size  $\leq 2$  that appeared in at least 1000 and at most 100,000 molecules—22 fragments in total. We expect small graphlets to have large influence on model performance, as their coefficients tend to be larger in our models. Visualizations of these fragments and their counts in QM9 can be found in Supplemental Fig. S8. For each held-out fragment, we fit a linear model with graphlet fingerprints up to size 5 on molecules

that did not contain this held out fragment. We measured the raw and adjusted performance on molecules containing the held-out fragment. Fig 8 shows the aggregated hold-out molecule predictions from these 22 experiments. Panel (a) shows that models make drastic errors when predicting on molecules with unseen small fragments, yielding an MAE of  $90.20 \frac{\text{kcal}}{\text{mol}}$ , and panel (b) shows that the adjustment reduced error by 52% to  $42.96 \frac{\text{kcal}}{\text{mol}}$  and improved  $R^2$  by over 38% from 0.67 to 0.93 by exploiting the simple assumption that unknown fragments are similar in nature to known ones on average. This proof-of-concept result demonstrates how an interpretable model can be readily manipulated to incorporate further knowledge and intuition.

### 3.7 Uncertainty Quantification

As discussed in Section 2.5, the presence of unseen fragments in new molecules can also be used to quantify model uncertainty about this molecule. There are numerous ways one could



Table 2: Performance of models using graphlet fingerprints compared to those present in the TDC leaderboards. Ranks are computed after our models are included. For tasks scored with MAE, a lower score is better and this is reflected in the ranking. The ranking order is reversed for tasks ranked by Spearman’s  $\rho$ , where higher scores are better.

| Task      | LightGBM Perf. | LightGBM Rank | Ridge Perf. | Ridge Rank | Perf. Metric      |
|-----------|----------------|---------------|-------------|------------|-------------------|
| LD50      | 0.603          | 3/19          | 0.632       | 8/19       | MAE               |
| AqSol     | 0.803          | 5/16          | 1.105       | 15/16      | MAE               |
| Lipo      | 0.519          | 5/17          | 0.516       | 4/17       | MAE               |
| PPBR      | 8.729          | 6/17          | 10.699      | 15/17      | MAE               |
| Caco2     | 0.316          | 7/19          | 0.306       | 6/19       | MAE               |
| VDss      | 0.500          | 8/17          | 0.448       | 13/17      | Spearman’s $\rho$ |
| Half Life | 0.217          | 12/18         | 0.229       | 11/18      | Spearman’s $\rho$ |
| CL-Hepa   | 0.341          | 13/16         | 0.349       | 12/16      | Spearman’s $\rho$ |
| CL-Micro  | 0.525          | 14/18         | 0.600       | 4/18       | Spearman’s $\rho$ |

utilize this information for UQ, including 1) using the count of unseen fragments or 2) using their frequency. One could also 3) build an explicit model of uncertainty based on unseen fragments. We evaluate all three of these approaches on a graphlet-based linear model trained on a small sample of 1,000 random molecules and their atomization energies from the QM9 dataset. The remaining molecules are used as a test set for all three UQ methods.

The explicit uncertainty model is a linear regression mapping the number of fragments of each size to the absolute residuals. This model is fit with non-negative least squares, guaranteeing non-negative residual prediction and giving coefficients with a natural interpretation as the contribution a single unseen fragment of a given size to the model uncertainty in units of the regression target. The resulting model coefficients are given in Supplementary Table S3.

We measure the performance of these uncertainty quantification methods with both correlation coefficients and confidence curves. Confidence curves (CCs) show how the model error changes as data points (here, molecules) with the highest uncertainty are excluded from the test set.<sup>59,60</sup> Confidence curves for different UQ metrics are compared quantitatively by comparing their integrals in the so-named AUC metric: the less area under the CC the better the performance of the UQ metric. To give a more intuitive meaning to the AUC, the Area Under the Confidence Oracle (AUCO)

metric presented in reference 59 considers the area between the confidence curve and an oracle curve, that is the true ordering of the points by decreasing absolute error and is the best-case CC for a UQ metric. As AUCO approaches zero, the confidence curve approaches the oracle curve, so smaller AUCO values are better. To provide an even more intuitive functional of the CC, we consider both an oracle and an anti-oracle which randomly discards points. This serves as a baseline that any well-performing UQ metric should outperform. Because the anti-oracle throws away points randomly, the anti-oracle has an expected CC equal to the test set MAE. The area between the anti-oracle and oracle (ABAO) thus represents a baseline AUCO. The  $CC_{\text{eff}}$  metric is then defined as

$$CC_{\text{eff}} = 1 - \frac{\text{AUCO}}{\text{ABAO}}. \quad (22)$$

Values of  $CC_{\text{eff}}$  close to one occur in the best case when the CC approaches the oracle CC, values near zero occur when the UQ metric is no better than random guessing, and negative values occur when the UQ metric is worse than random assignment. In this way, we can think of  $CC_{\text{eff}}$  as being related to the AUC in the same way that  $R^2$  relates to  $MSE$ ; A perfect  $CC_{\text{eff}}$  is 1, an uninformative  $CC_{\text{eff}}$  is 0.

All of the proposed measures of uncertainty based on unseen fragments have moderate to strong correlation with absolute residuals, shown in Figure 9. Confidence curves and  $CC_{\text{eff}}$

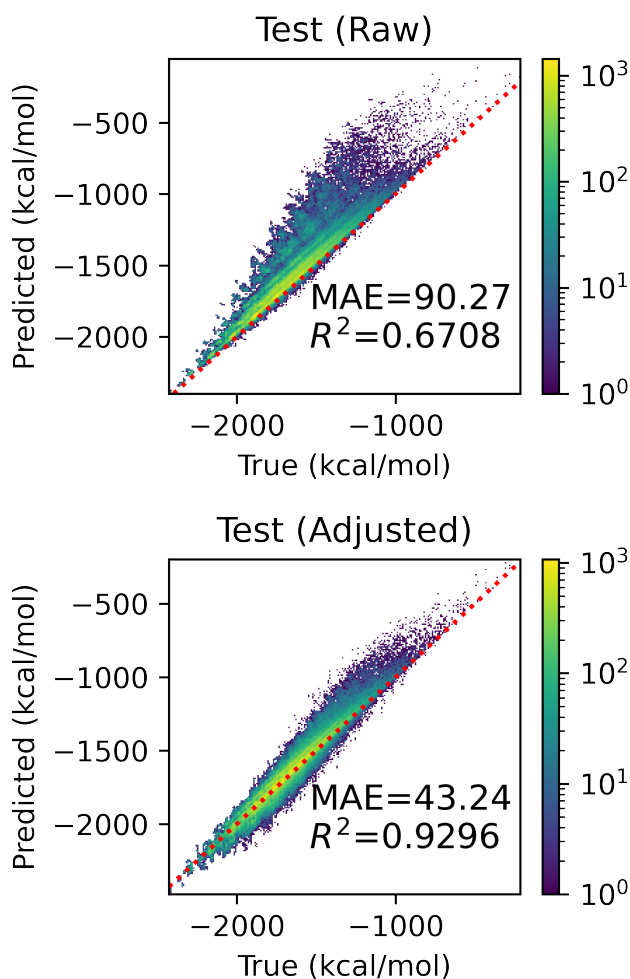


Figure 8: Performance improvement from adjustment based on unseen fragments. Both panels show the predicted and true atomization energies for every held-out fragment, coalesced into one figure. The top panel shows the raw predictions and the bottom shows the predictions after applying the adjustment.

values are shown in Figure 10. The fraction of unseen fragments performs the strongest under both correlation coefficients and our confidence curve efficiency metric.

## 4 Discussion

Regression on molecular graphlets is a simple yet powerful technique yielding consistent strong performance at little computational cost; even our  $s = 9$  models of QM9 with over 1M features (graphlet isomorphism classes) were

trained on a single computing node in less than 48 hours, including hyperparameter search. Many of the datasets examined in this work can be trained in the order of minutes. When carefully constructed and utilized, models on molecular graphlets are also highly locally interpretable via projected-coefficient visualizations, and the presence of unseen fragments can be used to adjust for model biases and quantify model uncertainty. In many cases, linear models built on graphlets are comparable to non-linear ones—we prefer the former due to their stronger interpretability.

Previous work has explored the interpretability of ML models built on fingerprints. Some research<sup>17,18</sup> examines SHAP values on the sub-graphs corresponding to fingerprint fragments. We caution that this has notable disadvantages. In addition to the inconsistency of SHAP discussed in Ref 22, SHAP explanations can include contributions from the absence of a particular fragment, effectively saying “the fact that this fragment was not present shifted this prediction from the mean by this much.” Explaining properties of molecules based on what they are not is reasonable from a statistical perspective but does not provide an explanation tied only to the given molecular graph which we found unintuitive in our own explorations of this approach. Interpreting linear model coefficients instead focuses only on the fragments that are present, as the zero-counts will remove the coefficients of the absent training fragments from the prediction. We also note that care must be taken to always use unfolded fingerprints when attempting to explain model predictions, or else the a one-to-many correspondence between model inputs and molecular fragments<sup>61–63</sup> significantly complicates interpretation, if not rendering it fully impossible. We also observed in our initial work that direct interpretation of the contribution of every molecular graph fragment present in a molecule is confounded by the inclusion relationships between graph fragments. This lead us to develop our method of projecting contributions using these inclusion relationships to atom- or bond-level contributions. We note that a similar interpretation method to our projection scheme is

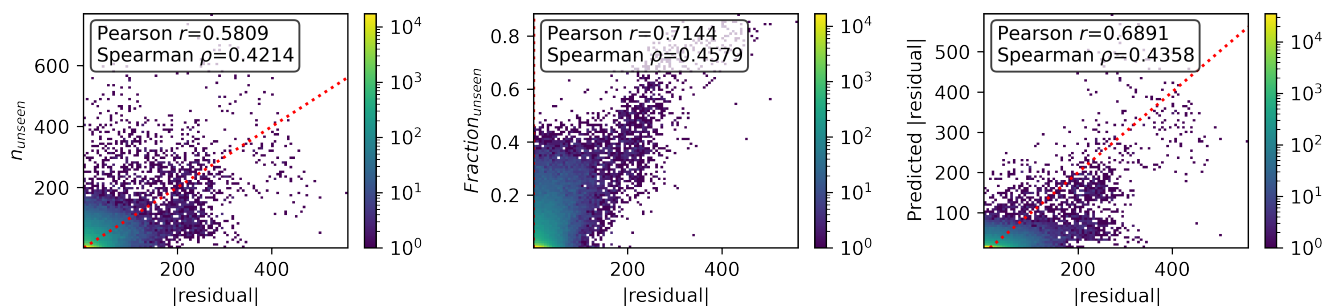


Figure 9: Correspondence of various metrics for uncertainty with error. a) Number of unseen fragments in a test molecule b) Fraction of unseen fragments in a test molecule c) Calibrated UQ model which takes into account both the number of unseen graph fragments and their sizes.

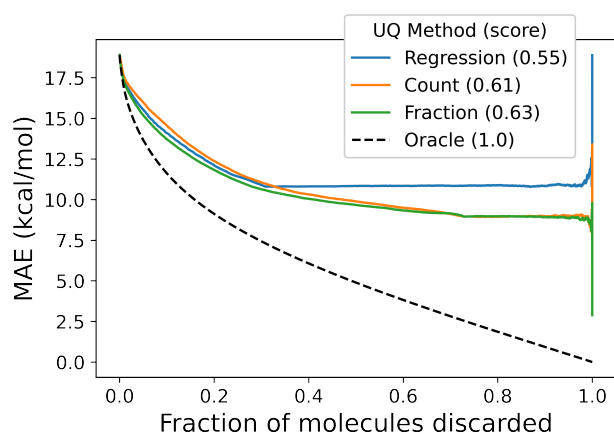


Figure 10: Confidence curves comparing uncertainty metrics to an oracle. Each curve shows the expected error in the test dataset as a curve by systematically dropping tests points with the highest uncertainty values; lower curves are better. The “oracle” curve shows the error distribution when points are dropped in order of their error; this is the best possible confidence curve for this error distribution.

presented in Ref 64, wherein the SHAP attributions of all of the fingerprint fragments containing a given atom are summed, giving an atom-level SHAP contribution.

Some similarity may be noted between our work and that of atom-in-molecule-ones (AMONs),<sup>65</sup> because each involves analysis of substructures. AMONs constitute a framework for for the comparison of 3D configurations; in that lens, they are a composition of a selective (as opposed to exhaustive) fragmentation into subgraphs, and molecular similarity kernels.<sup>66</sup> 3D information about target molecules

is typically used for contexts where the target property varies with respect to the input coordinates- for example, conformational energy variations; the cheminformatics applications presented in this work are distinct because they do not depend on conformation.

We note some advantages of graphlet fingerprints over other fingerprints, some of which are noted in Ref 25. Graphlet fingerprints may be considered at once more complete than Morgan-like fingerprints and more compact or less redundant than RDKit fingerprints. This is visible in the feature counts in Fig. 3, also shown in the Supplementary Information, Table 1. Due to the radial nature, many substructures have no direct representation in Morgan fingerprints. Notably Morgan fingerprints do not explicitly represent bonds, which important chemical and physical meanings. Bond-level interpretations like those in 3.2 are impossible with Morgan fingerprints. Likewise, RDKit fingerprints cannot directly represent atoms: paths of length one–bonds–are the smallest fragments in RDKit fingerprints. RDKit fingerprints are also redundant in their representation of fragments in individual molecules when multiple bond paths connect the same set of atoms. For example, in a molecule with a ring containing  $n$  atoms, there is precisely one graphlet-based induced subgraph containing exactly those atoms, yet RDKit fingerprints will produce  $n-1$  fingerprint elements containing that same set of atoms, each one missing one bond from the ring. This leads to many-to-one correspondence between model coefficients and atom subsets which presents a

challenge to directly interpreting these coefficients. This redundancy may also challenge machine learning methods as the fingerprint vectors will be highly correlated, even when models are fit hierarchically by fragment size. Hierarchical fitting helps to alleviate this redundancy by assigning model contributions to the lowest fragment size possible.

Regarding computational efficiency, we note that Morgan fingerprints are less costly to compute on large molecules with large fragment sizes due to their highly restrictive radial fragment definition. Graphlet fingerprints scale similarly in cost with molecule and fragment size to RDKit fingerprints. This is unsurprising, as the latter can be thought of as edge-induced subgraphs rather than node-induced subgraphs.

We note that identifying and counting graphlets has long shown promise in other areas of machine learning. A kernel based on the number of shared graphlets between graphs has been used to predict functional residues of proteins.<sup>67</sup> Due to the potentially combinatoric cost of counting all graphlets on arbitrarily connected graphs, these methods often incorporate random sampling of small graphlets.<sup>68</sup> Examining the symmetry relationships between nodes within graphlets has been exploited to understand protein-protein interactions<sup>69</sup> and interactions between MicroRNA and small molecules.<sup>70</sup> Various spectral methods based on graphlets have been developed<sup>71</sup> and applied to problems such as biological network comparison.<sup>72,73</sup> Recently, graphlet substructure relationships have been used to improve performance of graph neural networks.<sup>40</sup>

## 5 Conclusion

In this manuscript, we have compared the performance of molecular graphlet fingerprints coupled with linear and nonlinear regressors to a variety of molecular featurization techniques from the literature. These include similar topological fingerprints such as the RDKit and Morgan fingerprints on QM9, hand-crafted DFT and experimental features on solubility datasets, and a variety of methods, including

deep learning methods such as attention-based and graph neural networks on the ADMET regression tasks from the Therapeutic Data Commons. We find that the graphlet approach fairs better than other topological fingerprint techniques, and is generally comparable in accuracy to the other techniques in the literature. This result gives counterpoint to recent efforts advocating for the use of black-box algorithms followed by post-hoc interpretability algorithms.<sup>12</sup>

At the same time, we have shown that the transparent nature of fingerprint techniques comes with many additional advantages. For one, we show that hierarchical linear modeling in the graphlet approach, using a many-body expansion hypothesis, produces a somewhat more accurate model which is far more stable to the maximum graphlet size hyperparameter. We also show how graphlet inclusion relationships can be used to assign precise interpretations which decompose the model prediction across the input molecular structure. This was shown to produce reasonable correlation with chemical theory and chemical intuition in the case of both 2-body (bond) and 1-body (atom) projections. Finally, we showed how the interpretability of graphlet-fingerprint-based linear models provides natural methods for uncertainty quantification as well as model adjustments which address distribution shift, in particular, adjustments that estimate the effect of new topological structures arising in test molecules.

Future work could take on a variety of directions. For one, having shown good performance on a very wide array of tasks, the graphlet featurization approach is suitable for application to new chemical datasets. Methodologically, the uncertainty quantification and interpretability methods discussed in this work are but the tip of the iceberg; a variety of more complex schemes could be explored, and some of them might prove to produce better interpretations or more well-calibrated uncertainty estimates. The problem of modeling uncertainty and constructing interpretations when using these features in combination with external features (such as ab-initio properties) re-

mains unsolved; in some sense, any feature that is itself a function of the molecular structure cannot present information that is orthogonal to the graphlet histogram, and so a method to project the information gained by models through these features back onto the molecular graph would need to be devised in order to extend the notions of interpretations presented here. In this work, we have concentrated on modeling data whose domain is scalar, that is the prediction target for each molecule is a single number. However, the graphlet distribution can be localized to their location in the graph, and so the graphlet technique could be modified to predict information such as the partial atomic charges within a molecule.

Finally, we remind the reader that we have released the code for our approach as an open source package, `minervachem`, at [github.com/lanl/minervachem](https://github.com/lanl/minervachem), along with tutorials outlining how to build models using the methods described here. We hope that future work by ourselves and others will be made available through the library.

## 6 Acknowledgements

We thank the following individuals, in no particular order, for helpful discussions concerning this work, including Alon Perkus, Eric Mojlness, Pieter Swart, Nathan Lemons, Alice E.A. Allen, Sakib Matin, Kipton Barros, and Joshua Schrier.

Los Alamos National Laboratory (LANL) is operated by Triad National Security, LLC, for the National Nuclear Security Administration of U.S. Department of Energy (contract no. 89233218CNA000001). We acknowledge the support by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Separation Science Program under contract number 2022LANLE3M1 and Heavy Element Chemistry Program under contract number 2022LANLE3M2. M.T. acknowledges funding from student fellowships sponsored by the Seaborg Institute and the Center for Nonlinear Studies (CNLS). M.T. is also supported by U.S. Department of Energy, Office of Sci-

ence, Office of Advanced Scientific Computing Research, Computational Science Graduate Fellowship under Award Number(s) DE-SC0023112. D.J.B. acknowledges a postdoctoral fellowship with the Center of Nonlinear Studies at LANL. We acknowledge the LANL's Director's Postdoc Fellowship (M.G.T.: LANL-LDRD-20210966PRD4 ; J.J.: LANL-LDRD-20220815PRD4) This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 using NERSC award BES-ERCAP0022265. Computational experiments were conducted in part using LANL's CCS-7 Darwin cluster.

## References

- (1) Hann, M.; Green, R. Chemoinformatics — a new name for an old problem? *Current Opinion in Chemical Biology* **1999**, *3*, 379–383.
- (2) Willett, P. Chemoinformatics: a history. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *1*, 46–56.
- (3) Engel, T. Basic Overview of Chemoinformatics. *Journal of Chemical Information and Modeling* **2006**, *46*, 2267–2277, PMID: 17125169.
- (4) Walters, W. P.; Barzilay, R. Applications of deep learning in molecule generation and molecular property prediction. *Accounts of chemical research* **2020**, *54*, 263–270.
- (5) Wieder, O.; Kohlbacher, S.; Kueneemann, M.; Garon, A.; Ducrot, P.; Seidel, T.; Langer, T. A compact review of molecular property prediction with graph neural networks. *Drug Discovery Today: Technologies* **2020**, *37*, 1–12.
- (6) Li, Z.; Jiang, M.; Wang, S.; Zhang, S. Deep learning methods for molecular representation and property prediction. *Drug Discovery Today* **2022**, 103373.
- (7) Gupta, V.; Choudhary, K.; Tavazza, F.; Campbell, C.; Liao, W.-k.; Choudhary, A.; Agrawal, A. Cross-property deep transfer learning framework for enhanced predictive analytics on small materials data. *Nature communications* **2021**, *12*, 6595.
- (8) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nature communications* **2019**, *10*, 2903.
- (9) Vermeire, F. H.; Green, W. H. Transfer learning for solvation free energies: From quantum chemistry to experiments. *Chemical Engineering Journal* **2021**, *418*, 129307.
- (10) Wang, Z.; Dai, Z.; Póczos, B.; Carbonell, J. Characterizing and avoiding negative transfer. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019; pp 11293–11302.
- (11) Hoffmann, N.; Schmidt, J.; Botti, S.; Marques, M. A. Transfer learning on large datasets for the accurate prediction of material properties. *arXiv preprint arXiv:2303.03000* **2023**,
- (12) Wellawatte, G. P.; Gandhi, H. A.; Seshadri, A.; White, A. D. A Perspective on Explanations of Molecular Prediction Models. *Journal of Chemical Theory and Computation* **2023**, *19*, 2149–2160.
- (13) Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* **2019**, *1*, 206–215.
- (14) Oviedo, F.; Ferres, J. L.; Buonassisi, T.; Butler, K. T. Interpretable and explainable machine learning for materials science and chemistry. *Accounts of Materials Research* **2022**, *3*, 597–607.
- (15) Ribeiro, M. T.; Singh, S.; Guestrin, C. "Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016; pp 1135–1144.
- (16) Lundberg, S. M.; Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **2017**, *30*.
- (17) Sanches-Neto, F. O.; Dias-Silva, J. R.; Keng Queiroz Junior, L. H.; Carvalho-Silva, V. H. "py SiRC": Machine Learning Combined with Molecular Fingerprints to Predict the Reaction Rate Constant of

the Radical-Based Oxidation Processes of Aqueous Organic Contaminants. *Environmental Science & Technology* **2021**, *55*, 12437–12448.

- (18) Ding, Y.; Chen, M.; Guo, C.; Zhang, P.; Wang, J. Molecular fingerprint-based machine learning assisted QSAR model development for prediction of ionic liquid properties. *Journal of Molecular Liquids* **2021**, *326*, 115212.
- (19) Ishida, S.; Terayama, K.; Kojima, R.; Takasu, K.; Okuno, Y. Prediction and interpretable visualization of retrosynthetic reactions using graph convolutional networks. *Journal of chemical information and modeling* **2019**, *59*, 5026–5033.
- (20) Mastropietro, A.; Pasculli, G.; Feldmann, C.; Rodríguez-Pérez, R.; Bajorath, J. EdgeSHAPer: Bond-centric Shapley value-based explanation method for graph neural networks. *Iscience* **2022**, *25*.
- (21) Xiong, P.; Schnake, T.; Gastegger, M.; Montavon, G.; Muller, K. R.; Nakajima, S. Relevant Walk Search for Explaining Graph Neural Networks. **2023**,
- (22) Alvarez-Melis, D.; Jaakkola, T. S. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049* **2018**,
- (23) Sixt, L.; Landgraf, T. A rigorous study of the deep Taylor decomposition. *arXiv preprint arXiv:2211.08425* **2022**,
- (24) Allen, A. E.; Tkatchenko, A. Machine learning of material properties: Predictive and interpretable multilinear models. *Science advances* **2022**, *8*, eabm7185.
- (25) Bellmann, L.; Penner, P.; Rarey, M. Connected subgraph fingerprints: representing molecules using exhaustive subgraph enumeration. *Journal of chemical information and modeling* **2019**, *59*, 4625–4635.
- (26) Bellmann, L.; Penner, P.; Rarey, M. Topological similarity search in large combinatorial fragment spaces. *Journal of Chemical Information and Modeling* **2020**, *61*, 238–251.
- (27) Bellmann, L.; Penner, P.; Gastreich, M.; Rarey, M. Comparison of combinatorial fragment spaces and its application to ultralarge make-on-demand compound catalogs. *Journal of Chemical Information and Modeling* **2022**, *62*, 553–566.
- (28) Bellmann, L.; Klein, R.; Rarey, M. Calculating and optimizing physicochemical property distributions of large combinatorial fragment spaces. *Journal of Chemical Information and Modeling* **2022**, *62*, 2800–2810.
- (29) Yao, K.; Herr, J. E.; Parkhill, J. The many-body expansion combined with neural networks. *The Journal of chemical physics* **2017**, *146*.
- (30) Lubbers, N.; Smith, J. S.; Barros, K. Hierarchical modeling of molecular energies using a deep neural network. *The Journal of chemical physics* **2018**, *148*.
- (31) Batatia, I.; Batzner, S.; Kovács, D. P.; Musaelian, A.; Simm, G. N.; Drautz, R.; Ortner, C.; Kozinsky, B.; Csányi, G. The design space of E (3)-equivariant atom-centered interatomic potentials. *arXiv preprint arXiv:2205.06643* **2022**,
- (32) Quinonero-Candela, J.; Sugiyama, M.; Schwaighofer, A.; Lawrence, N. D. *Dataset shift in machine learning*; Mit Press, 2008.
- (33) Landrum, G. Fingerprints in the RDKit. [https://www.rdkit.org/UGM/2012/Landrum\\_RDKit\\_UGM.Fingerprints.Final.pptx.pdf](https://www.rdkit.org/UGM/2012/Landrum_RDKit_UGM.Fingerprints.Final.pptx.pdf), [Online; accessed 4-Oct-2023].
- (34) Morgan, H. L. The generation of a unique machine description for chemical structures—a technique developed at chemical abstracts service. *Journal of chemical documentation* **1965**, *5*, 107–113.

- (35) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling* **2010**, *50*, 742–754.
- (36) Mahadevan, P.; Krioukov, D.; Fall, K.; Vahdat, A. Systematic topology analysis and generation using degree correlations. *ACM SIGCOMM Computer Communication Review* **2006**, *36*, 135–146.
- (37) Wernicke, S. Efficient detection of network motifs. *IEEE/ACM transactions on computational biology and bioinformatics* **2006**, *3*, 347–359.
- (38) Ulam, S. *A collection of mathematical problems*; Interscience Publishers: New York, 1960.
- (39) Bondy, J. A.; Hemminger, R. L. Graph reconstruction—a survey. *Journal of Graph Theory* **1977**, *1*, 227–268.
- (40) Bouritsas, G.; Frasca, F.; Zafeiriou, S.; Bronstein, M. M. Improving graph neural network expressivity via subgraph isomorphism counting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2022**, *45*, 657–668.
- (41) Bollobás, B. Almost every graph has reconstruction number three. *Journal of Graph Theory* **1990**, *14*, 1–4.
- (42) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less is more: Sampling chemical space with active learning. *The Journal of chemical physics* **2018**, *148*.
- (43) Wang, K.; Dowling, A. W. Bayesian optimization for chemical products and functional materials. *Current Opinion in Chemical Engineering* **2022**, *36*, 100728.
- (44) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data* **2014**, *1*, 1–7.
- (45) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; Von Lilienfeld, O. A. Prediction errors of molecular machine learning models lower than hybrid DFT error. *Journal of chemical theory and computation* **2017**, *13*, 5255–5264.
- (46) Boobier, S.; Hose, D. R.; Blacker, A. J.; Nguyen, B. N. Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. *Nature communications* **2020**, *11*, 1–10.
- (47) Huang, K.; Fu, T.; Gao, W.; Zhao, Y.; Roohani, Y.; Leskovec, J.; Coley, C. W.; Xiao, C.; Sun, J.; Zitnik, M. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *NeurIPS* **2021**,
- (48) Therapeutics Data Commons: ADMET Leaderboards. [https://tdcommons.ai/benchmark/admet\\_group/overview/](https://tdcommons.ai/benchmark/admet_group/overview/), Accessed: 2023-07-24.
- (49) St. John, P. C.; Guan, Y.; Kim, Y.; Etz, B. D.; Kim, S.; Paton, R. S. Quantum chemical calculations for over 200,000 organic radical species and 40,000 associated closed-shell molecules. *Scientific Data* **2020**, *7*, 244.
- (50) Landrum, G. RDKit: Open-Source Cheminformatics. <http://www.rdkit.org>, [Online; accessed 2-Feb-2020].
- (51) Hagberg, A.; Swart, P.; Schult, D. *Exploring network structure, dynamics, and function using NetworkX*; 2008.
- (52) Virtanen, P. et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **2020**, *17*, 261–272.
- (53) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.



- (54) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* **2017**, *30*.
- (55) Wang, C.; Wu, Q.; Weimer, M.; Zhu, E. FLAML: A Fast and Lightweight AutoML Library. MLSys. 2021.
- (56) Bogojeski, M.; Vogt-Maranto, L.; Tuckerman, M. E.; Müller, K.-R.; Burke, K. Quantum chemical accuracy from density functional approximations via machine learning. *Nature communications* **2020**, *11*, 5223.
- (57) Blanksby, S. J.; Ellison, G. B. Bond dissociation energies of organic molecules. *Accounts of chemical research* **2003**, *36*, 255–263.
- (58) Simpson, E. H. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **1951**, *13*, 238–241.
- (59) Scalia, G.; Grambow, C. A.; Pernici, B.; Li, Y.-P.; Green, W. H. Evaluating Scalable Uncertainty Estimation Methods for Deep Learning-Based Molecular Property Prediction. *J. Chem. Inf. Model.* **2020**, *60*, 2697–2717.
- (60) Huang, X.; Yang, J.; Li, L.; Deng, H.; Ni, B.; Xu, Y. Evaluating and Boosting Uncertainty Quantification in Classification. *arXiv preprint arXiv:1909.06030* **2019**,
- (61) Cortés-Ciriano, I.; Ain, Q. U.; Subramanian, V.; Lenselink, E. B.; Méndez-Lucio, O.; IJzerman, A. P.; Wohlfahrt, G.; Prusis, P.; Malliavin, T. E.; van Westen, G. J., et al. Polypharmacology modelling using proteochemometrics (PCM): recent methodological developments, applications to target families, and future prospects. *MedChemComm* **2015**, *6*, 24–50.
- (62) Cortes-Ciriano, I. Bioalerts: a python library for the derivation of structural alerts from bioactivity and toxicity data sets. *Journal of cheminformatics* **2016**, *8*, 1–6.
- (63) Murrell, D. S.; Cortes-Ciriano, I.; Van Westen, G. J.; Stott, I. P.; Bender, A.; Malliavin, T. E.; Glen, R. C. Chemically Aware Model Builder (camb): an R package for property and bioactivity modelling of small molecules. *Journal of cheminformatics* **2015**, *7*, 1–10.
- (64) Humer, C.; Heberle, H.; Montanari, F.; Wolf, T.; Huber, F.; Henderson, R.; Heinrich, J.; Streit, M. ChemInformatics Model Explorer (CIME): exploratory analysis of chemical model explanations. *Journal of Cheminformatics* **2022**, *14*, 21.
- (65) Huang, B.; von Lilienfeld, O. A. Quantum machine learning using atom-in-molecule-based fragments selected on the fly. *Nature chemistry* **2020**, *12*, 945–951.
- (66) Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R. Machine learning of accurate energy-conserving molecular force fields. *Science Advances* **2017**, *3*, e1603015.
- (67) Vacic, V.; Iakoucheva, L. M.; Lonardi, S.; Radivojac, P. Graphlet kernels for prediction of functional residues in protein structures. *Journal of Computational Biology* **2010**, *17*, 55–72.
- (68) Shervashidze, N.; Vishwanathan, S.; Petri, T.; Mehlhorn, K.; Borgwardt, K. Efficient graphlet kernels for large graph comparison. *Artificial intelligence and statistics*. 2009; pp 488–495.
- (69) Wang, X.-D.; Huang, J.-L.; Yang, L.; Wei, D.-Q.; Qi, Y.-X.; Jiang, Z.-L. Identification of human disease genes from interactome network using graphlet interaction. *PloS one* **2014**, *9*, e86142.

- (70) Guan, N.-N.; Sun, Y.-Z.; Ming, Z.; Li, J.-Q.; Chen, X. Prediction of potential small molecule-associated microRNAs using graphlet interaction. *Frontiers in pharmacology* **2018**, *9*, 1152.
- (71) Kondor, R.; Shervashidze, N.; Borgwardt, K. M. The graphlet spectrum. Proceedings of the 26th Annual International Conference on Machine Learning. 2009; pp 529–536.
- (72) Pržulj, N. Biological network comparison using graphlet degree distribution. *Bioinformatics* **2007**, *23*, e177–e183.
- (73) Windels, S. F.; Malod-Dognin, N.; Pržulj, N. Graphlet Laplacians for topology-function and topology-disease relationships. *Bioinformatics* **2019**, *35*, 5226–5234.

# Supporting Information for: Linear Graphlet Models for Accurate and Interpretable Cheminformatics

Michael Tynes,<sup>\*,†,‡,¶,§</sup> Michael G. Taylor,<sup>†</sup> Jan Janssen,<sup>†</sup> Daniel J. Burrill,<sup>†,‡</sup>  
Danny Perez,<sup>†</sup> Ping Yang,<sup>\*,†</sup> and Nicholas Lubbers<sup>\*,||</sup>

<sup>†</sup>*Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA*

<sup>‡</sup>*Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM 87545,  
USA*

<sup>¶</sup>*Analytics, Intelligence, and Technology Division, Los Alamos National Laboratory, Los  
Alamos, NM 87545, USA (Current address)*

<sup>§</sup>*Department of Computer Science, The University of Chicago, Chicago, IL 60637, USA  
(Current address)*

<sup>||</sup>*Computer, Computational, and Statistical Sciences Division, Los Alamos National  
Laboratory, Los Alamos, NM 87545, USA*

E-mail: [mtynes@uchicago.edu](mailto:mtynes@uchicago.edu); [pyang@lanl.gov](mailto:pyang@lanl.gov); [nlubbers@lanl.gov](mailto:nlubbers@lanl.gov)

# S1 QM9

## S1.1 L2 hyperparameter optimization

Following preliminary experiments, we found the below ranges for the linear regression regularization  $\alpha$ , and noted that underregularized models are much slower to converge so we did not use them. For the nonhierarchical linear models, we searched for  $\alpha$  in a range of  $10^{-2}$  to  $10^{-1}$  for morgan substructures and a range from  $10^{-1}$  to  $10^2$  for graphlet and RDKit fingerprints. The ranges were motivated by slower runtimes for small values of  $\alpha$  for the latter two types of fingerprint. For similar reasons, for hierarchical linear models we searched the range  $10^{-2}$  to  $10^2$  for fragment sizes  $s \leq 4$ ,  $10^0$  to  $10^3$  for fragment sizes  $s \leq 8$  and from  $10^1$  to  $10^4$  for fragment size  $s = 9$ .

## S1.2 Fragment Counts by fingerprint type

Table S1: Number of fragments identified during training for each fingerprint type, for each maximum fragment size. For Graphlet fingerprints, size is the maximum number of atoms included in a fragment corresponding to a fingerprint element. For RDKit fingerprints, it is the maximum number of bonds. For Morgan, it is the Morgan radius.

| Fragment Size $s$ | Morgan      |              | RDKit       |              | Graphlet    |              |
|-------------------|-------------|--------------|-------------|--------------|-------------|--------------|
|                   | # Fragments | (Cumulative) | # Fragments | (Cumulative) | # Fragments | (Cumulative) |
| 0                 | 25          | (25)         | -           | -            | -           | -            |
| 1                 | 1,068       | (1,093)      | 29          | (29)         | 9           | (9)          |
| 2                 | 68,669      | (69,762)     | 106         | (135)        | 40          | (49)         |
| 3                 | 329,419     | (399,181)    | 518         | (653)        | 198         | (247)        |
| 4                 | 99,133      | (498,314)    | 2,530       | (3,183)      | 992         | (1,239)      |
| 5                 | 2,162       | (500,476)    | 11,998      | (15,181)     | 4,536       | (5,775)      |
| 6                 | 0           | -            | 50,041      | (65,222)     | 19,524      | (25,299)     |
| 7                 | -           | -            | 178,608     | (243,830)    | 77,599      | (102,898)    |
| 8                 | -           | -            | 514,229     | (758,059)    | 270,792     | (373,690)    |
| 9                 | -           | -            | 1,145,734   | (1,903,793)  | 776,315     | (1,150,005)  |

## S1.3 Learning curves

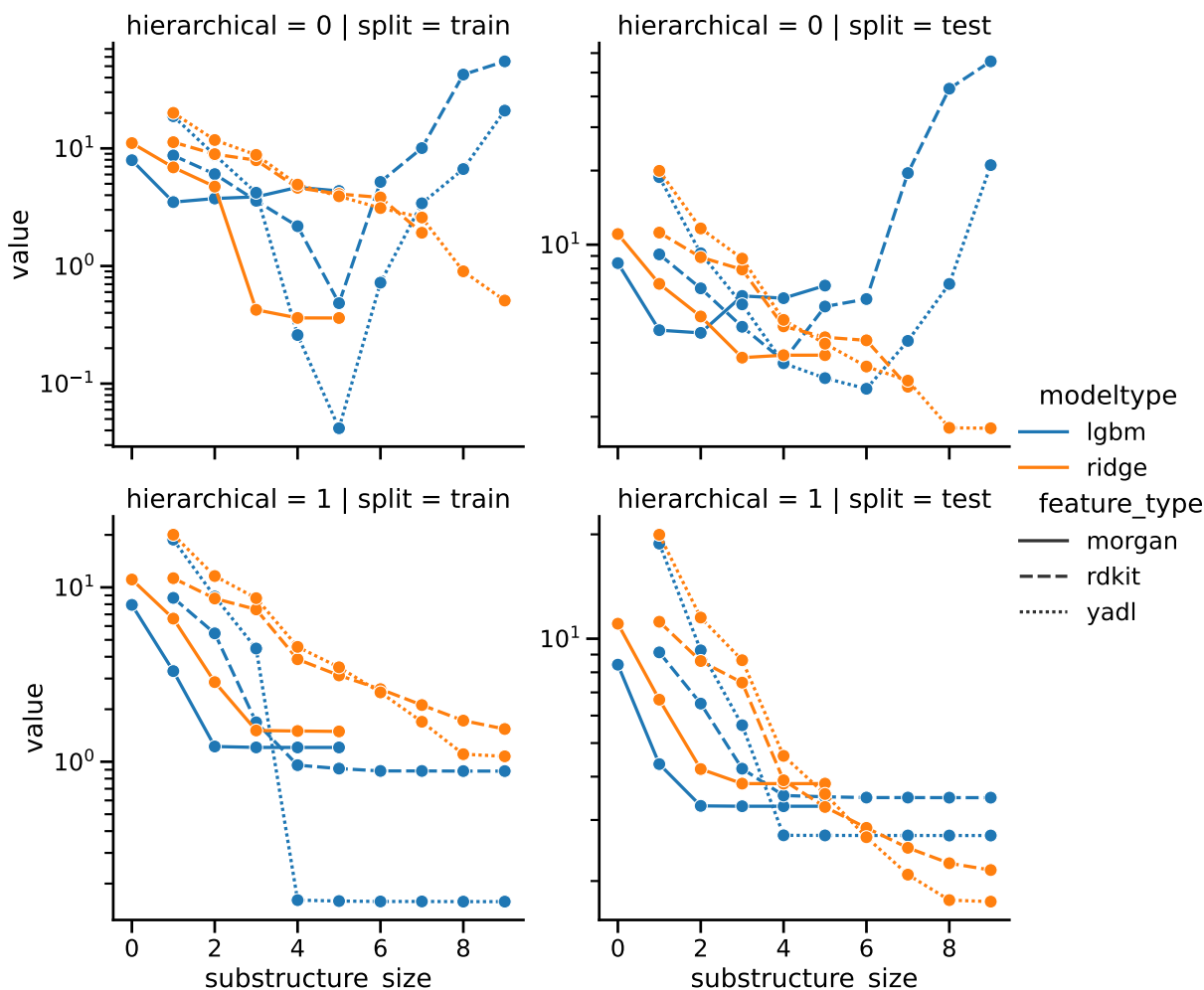


Figure S1: Learning curves for all model and fingerprint types by maximum size (`substructure_size`) for each fingerprint type. For Graphlet fingerprints (ours), size is the maximum number of atoms included in a fragment corresponding to a fingerprint element. For RDKit fingerprints, it is the maximum number of bonds. For morgan, it is the morgan radius.

## S2 Bond Dissociation Energies

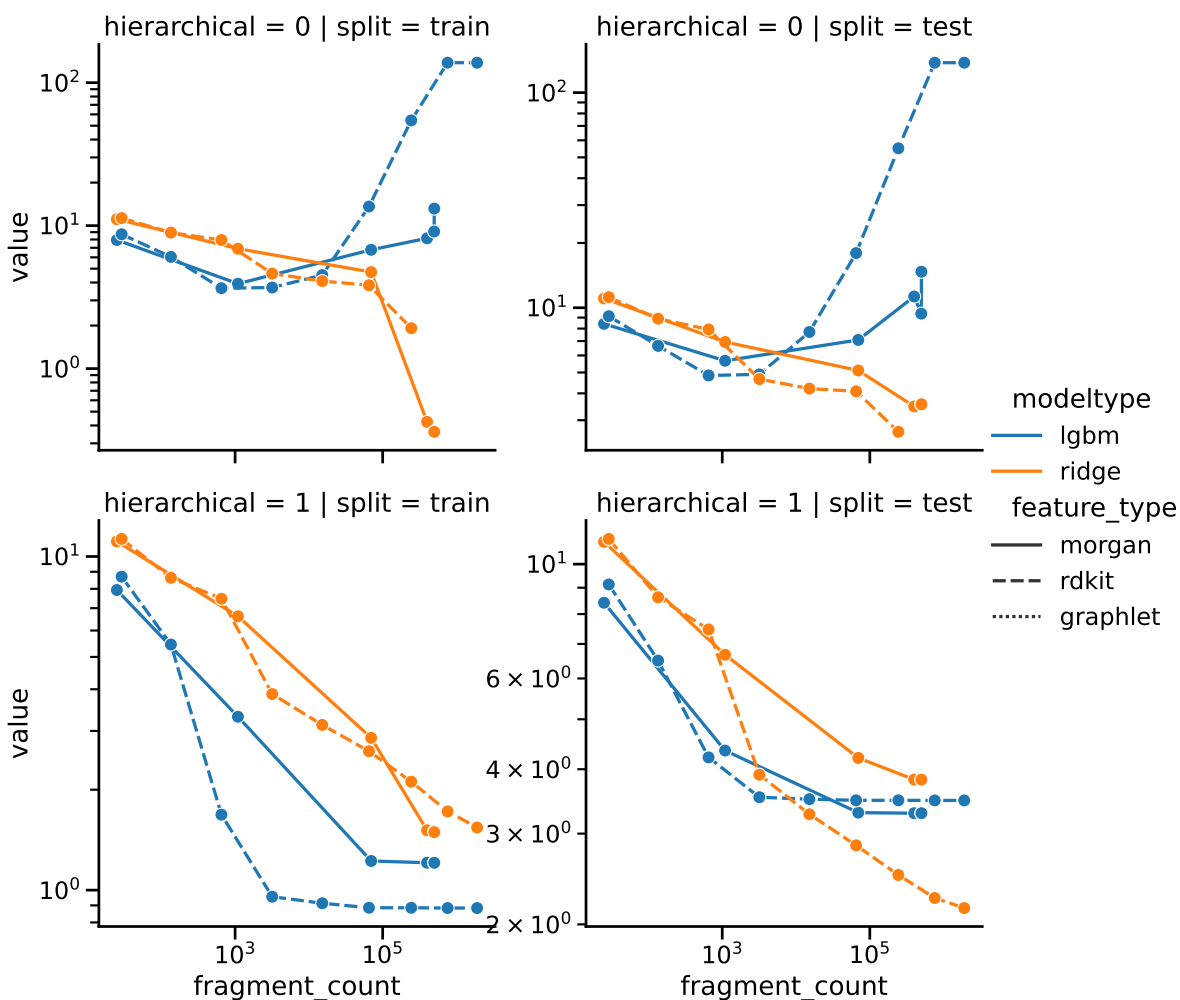


Figure S2: Learning curves for all model and fingerprint types by number of fragments identified during training. Each point corresponds to a choice of maximum size in figure S1, but here the horizontal axis has the same meaning for each curve.

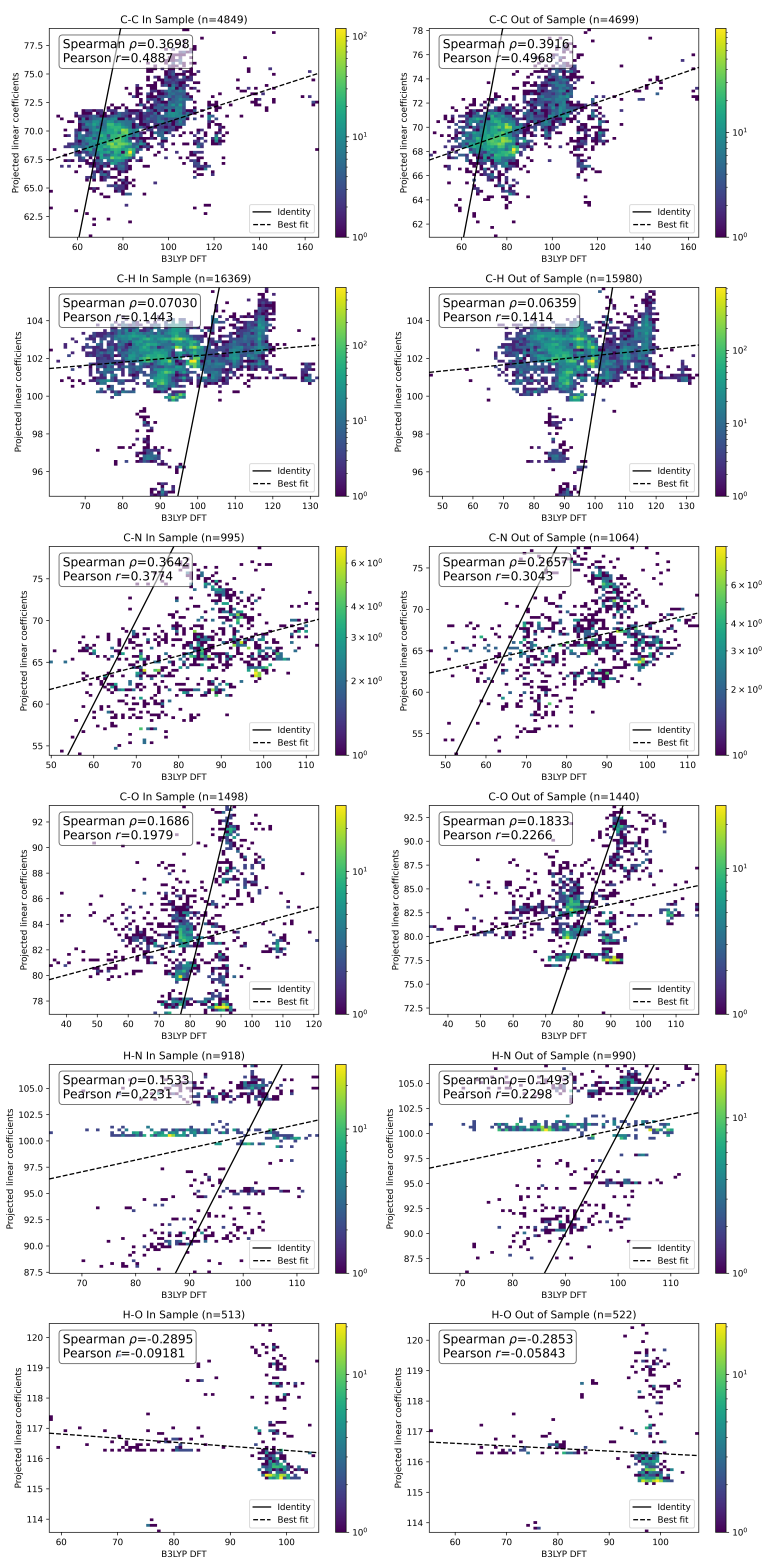


Figure S3: Relationships between bond-projected coefficients  $\bar{\chi}^2$  and bond dissociation energies by bond type and dataset split.

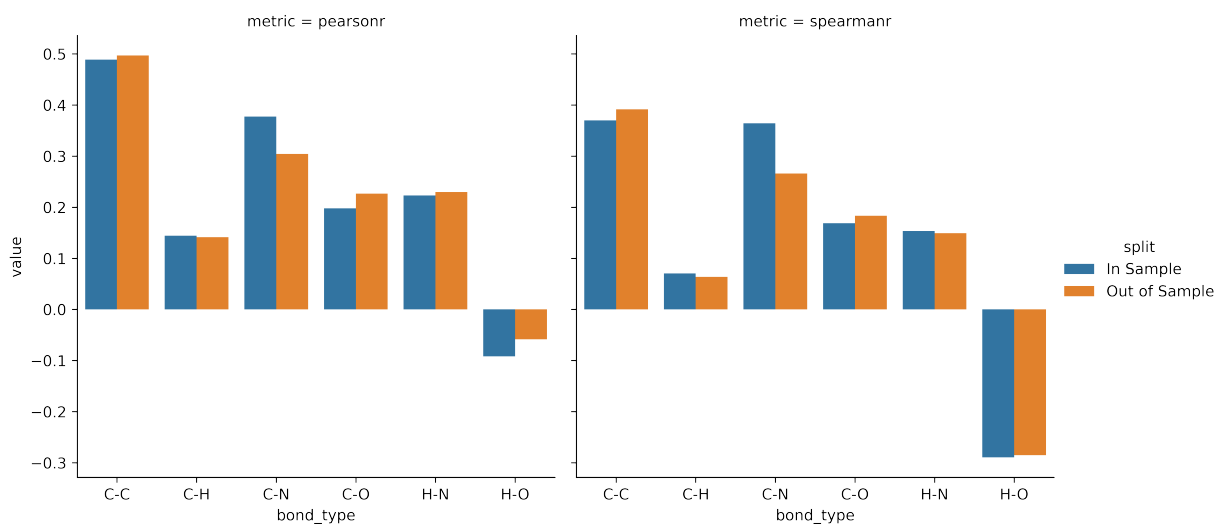


Figure S4: Correlation coefficients between bond-projected coefficients  $\chi^2$  and bond dissociation energies by bond type and dataset split.



## S3 Graphlet Fingerprint Fragment Counts for Solubility Tasks

Table S2: Graphlet fingerprint feature counts for solubility tasks

| Fragment Size | Dataset |         |         |       |            |          |         |
|---------------|---------|---------|---------|-------|------------|----------|---------|
|               | Acetone | Benzene | Ethanol | Water | Water-wide | Water-WN | Average |
| 1             | 25      | 22      | 26      | 24    | 24         | 24       | 24      |
| 2             | 116     | 101     | 127     | 118   | 125        | 125      | 118     |
| 3             | 451     | 354     | 514     | 515   | 570        | 570      | 495     |
| 4             | 0       | 1222    | 1932    | 0     | 0          | 0        | 525     |
| Total         | 592     | 1699    | 2599    | 657   | 719        | 719      | 1164    |

## S4 Solubility interpretation

Figure S5 shows atom-level coefficient projections constructed following Section 3.3.1 for five example molecules selected for structural diversity in acetone and water. These projected coefficients sum to the model prediction on each molecule, and can be thought of as the contributions of individual atoms to the model prediction after taking account its context in the molecular graph. By comparing the same molecule under different solvents, one can examine how structural motifs contribute to solubility in the different solvents. In this regard, the projections presented in Figure S5 largely agree with chemical intuition, e.g., carbon rings (molecules a and b) and chains (molecule d) explicitly lead to lower predictions of solubility in water than they do in acetone.

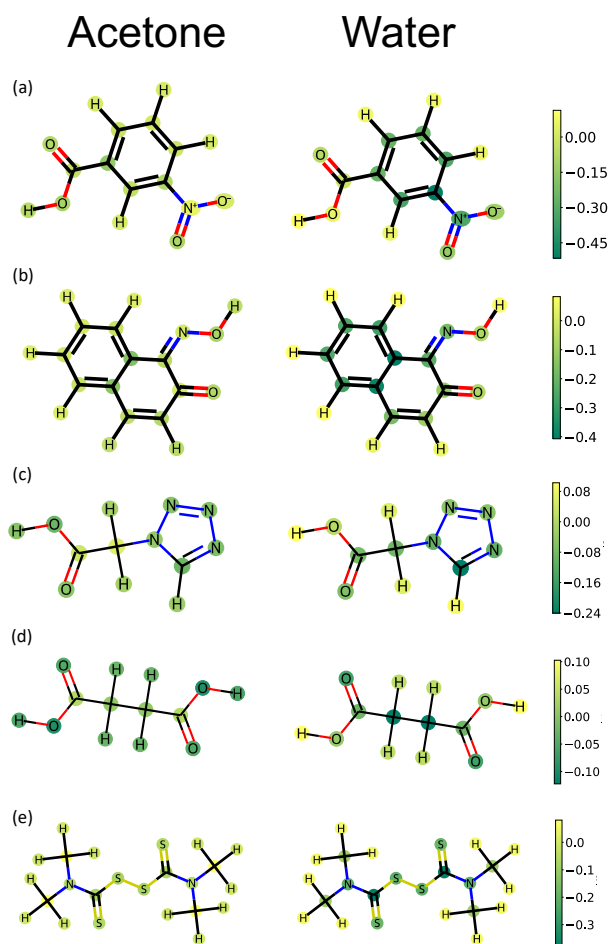


Figure S5: Linear model contributions projected to the atom level for molecules from the acetone and water sets presented in Ref 1. Colors show the contribution to the predicted log solubility (measured in molarity). Each column is based on predictions from the model fit on the corresponding solvent. Atoms in each row are colored using the same color map.

## S5 Leaderboard Results

Here we show performance results for our models compared to those on the existing TDC leaderboards<sup>2</sup> as bar plots.

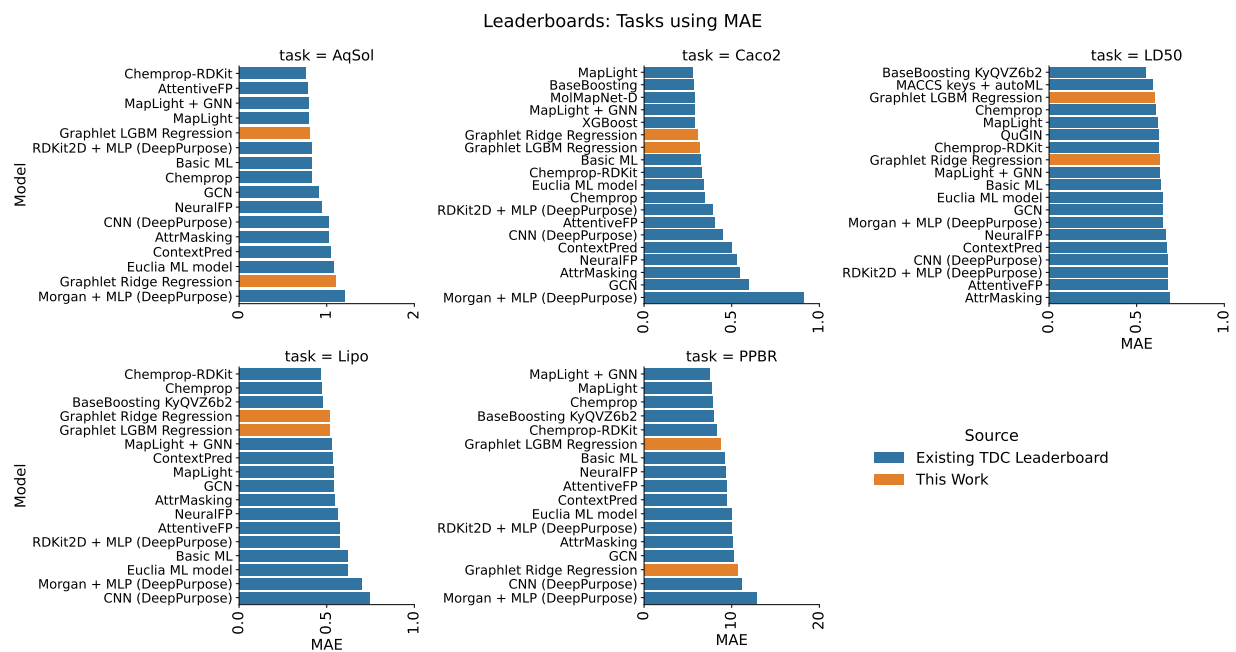


Figure S6: ADMET Performance barplots for leaderboards where task performance is measured in mean absolute error (MAE) (lower is better)

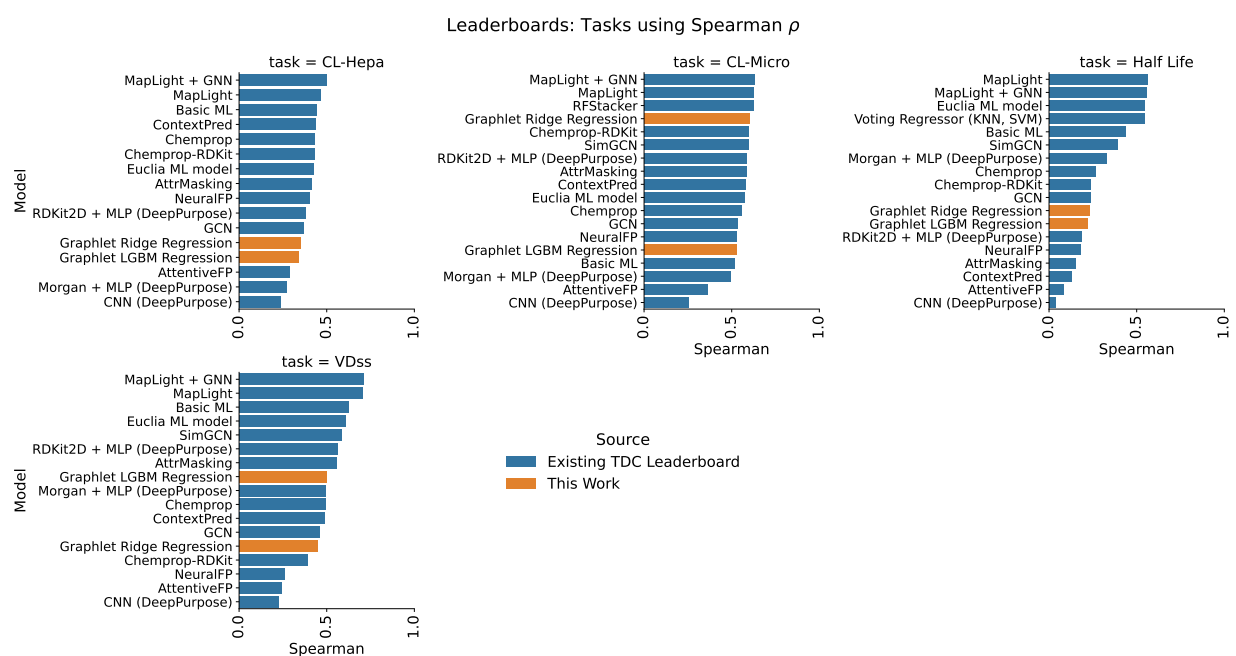


Figure S7: ADMET Performance barplots for leaderboards where task performance is measured using Spearman's rank-rank correlation coefficient  $\rho$  (higher is better)

## S6 Holdout fragments for adjustment experiments



Figure S8: Molecular graphlets and their corresponding counts in QM9 that were used to construct fragment holdout sets for experiments in section 3.5

## S7 Uncertainty Quantification

### S7.1 UQ regression coefficients

Table S3: Error model coefficients for the calibrated uncertainty model.

| Fragment Size | Error Model Coefficient (eV) |
|---------------|------------------------------|
| 1             | 0                            |
| 2             | 0                            |
| 3             | 18.7                         |
| 4             | 0                            |
| 5             | 0.441                        |
| 6             | 1.52                         |
| 7             | 0                            |

## References

- (1) Boobier, S.; Hose, D. R.; Blacker, A. J.; Nguyen, B. N. Machine learning with physico-chemical relationships: solubility prediction in organic solvents and water. *Nature communications* **2020**, *11*, 1–10.
- (2) Therapeutics Data Commons: ADMET Leaderboards. [https://tdcommons.ai/benchmark/admet\\_group/overview/](https://tdcommons.ai/benchmark/admet_group/overview/), Accessed: 2023-07-24.