THE JOURNAL OF PHYSICAL CHEMISTRY A
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCA Article

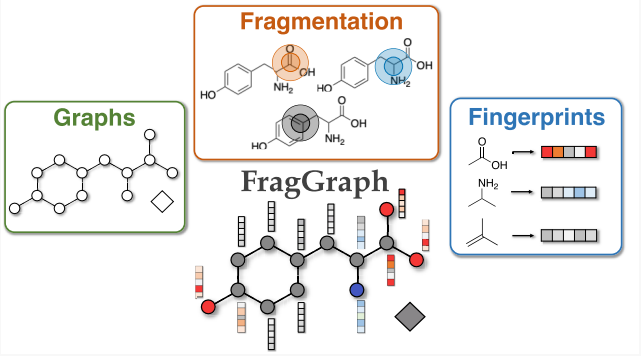
A Fragmentation-Based Graph Embedding Framework for QM/ML

Eric M. Collins* and Krishnan Raghavachari*

Cite This: J. Phys. Chem. A 2021, 125, 6872–6880 Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: We introduce a new fragmentation-based molecular representation framework "FragGraph" for QM/ML methods involving embedding fragment-wise fingerprints onto molecular graphs. Our model is specifically designed for delta machine learning (Δ-ML) with the central goal of correcting the deficiencies of approximate methods such as DFT to achieve high accuracy. Our framework is based on a judicious combination of ideas from fragmentation, error cancellation, and a state-of-the-art deep learning architecture. Broadly, we develop a general graph-network framework for molecular machine learning by incorporating the inherent advantages prebuilt into error cancellation methods such as the generalized Connectivity-Based Hierarchy. More specifically, we develop a QM/ML representation through a fragmentation-based attributed graph representation encoded with fragment-wise molecular fingerprints. The utility of our representation is demonstrated through a graph network fingerprint encoder in which a global fingerprint is generated through message passing of local neighborhoods of fragment-wise fingerprints, effectively augmenting standard fingerprints to also include the inbuilt molecular graph structure. On the 130k-GDB9 dataset, our method predicts an out-of-sample mean absolute error significantly lower than 1 kJ/mol compared to target G4(MP2) calculated energies, rivaling current deep learning methods with reduced computational scaling.



1. INTRODUCTION

The field of quantum chemistry has undergone a vast number of computational advances, allowing for the study of an ever-expanding range of chemical systems at varying levels of accuracy, rivaling experiments for many small molecules.[1−5] Although the most sophisticated *ab initio* methods are approaching the exact solution of the Schrödinger equation, such state-of-the-art methods cannot be used for systems larger than ∼10 atoms given the current computational resources. Thus, the quantum chemist's toolbox has been filled with a range of more tractable computational methods, e.g., DFT at different rungs of complexity, introducing significant approximations that cause a loss in accuracy.[6] Despite this deterioration in performance, DFT remains the primary workhorse for electronic structure studies due to its broad availability, high computational speed, and wide range of applicability. Improving the accuracy of DFT remains one of the primary challenges in quantum chemistry.

Of particular interest to this work is computational thermochemistry, which focuses on quantifying changes in energy associated with various chemical processes, giving insights into chemical reactivities, stability, and spontaneity. The accuracy of the computed quantities is, however, highly dependent on the level of theory chosen. Despite the success of modern density functionals, these methods are still inadequate for many demanding applications that require a highly accurate treatment of electron correlation effects. For an acceptable accuracy, at least one of two conditions must be met: (1) chemically accurate energies of all species in a reaction or (2) a high degree of error cancellation between reactants and products.

The direct calculation of accurate energies of all reactant and product species requires a highly sophisticated level of theory, such as composite wave function theory approaches (cWFTs).[1−5,7−12] Mid-level cWFTs, such as the Gaussian-n (Gn) and complete basis set (CBS) approaches, typically utilize DFT for geometry optimization and then calculate the electronic energy from a series of more sophisticated electron correlation methods such as MP2 and CCSD(T), providing a means to study moderately sized molecules within 1−2 kcal/mol of experiments.[8,10,11,13] Although cWFTs provide a good approximation to CCSD(T) with a large or complete basis set, many are still limited in applicability due to the steep scaling cost

Received: July 9, 2021
Revised: July 20, 2021
Published: August 3, 2021

of calculating accurate electronic energies (e.g., $O(N^7)$ for CCSD(T)), leading computational chemists to rely on the use of density functionals.

Alternatively, the errors of more approximate methods can be exploited under the assumption that structurally similar molecules will lead to similar deviations from the "exact" energy or property. This can be seen anecdotally through reactions in which only a small chemical transformation (where products and reactants are very structurally similar) takes place, e.g., isomerization reactions or conformer energy differences. For such reactions, inexpensive levels of theory can approach accuracy similar to those from more sophisticated methods. This idea was recognized and popularized in the 1970s by Pople et al. with the introduction of the (isodesmic) bond separation reaction, in which a molecule is separated into its constituent heavy atom bonds.[14] The number of bonds of a given formal type is retained throughout a given chemical transformation, and the associated energy change (heat of bond separation) could be calculated accurately with inexpensive levels of theory employing modest basis sets.[14]

Since the conception of the isodesmic scheme, these ideas have been explored further and applied to a wide range of applications in computational thermochemistry.[15] For example, we have developed the Connectivity-Based Hierarchy (CBH) of error cancellation schemes that provides an automated protocol to generate isodesmic-type reactions, which increasingly preserve the chemical environment on both sides of a reaction.[16,17] CBH reactions can be used to eliminate certain systematic errors present in approximate levels of theory via a corrective term derived from the reactants and products of the CBH reaction calculated at low and high levels of theory. The CBH approach has been utilized to study a broad range of thermochemical problems with an accuracy comparable to cWFT methods at the cost of low-fidelity DFT calculations, including heat of formation of charged and neutral organic and biomolecules,[16−22] redox potentials,[23] pK_as,[24] and bond dissociation energies.[25] Thus, where the direct computation of accurate energies is not possible, the exploitation of systematic error cancellation provides a viable (and often the only) alternative to achieve high accuracies in thermochemistry.[15] In this context, we note that related ideas using multiple levels of theory have also been developed via the fragmentation-based hybrid QM/QM approach for the study of large molecules to perform electronic structure calculations that would otherwise be computationally prohibitive.[26−30] The CBH can be considered as a systematic fragmentation-based approach that is particularly tuned for error cancellation.

A range of parallel developments has emerged more recently in a completely different context from the growth of artificial intelligence technology and application of machine learning (ML) techniques to the prediction of chemical properties at a greatly reduced cost. One of the more significant frameworks in this field is a hybrid QM/ML method, or Δ-ML[31] (and its multilevel generalized version CQML[32]), combining low-cost quantum chemical methods with ML models to mitigate the inaccuracies introduced from such approximations. The success of these QM/ML methods, however, is dependent on the architecture of the model, such as artificial feed-forward neural network (ANN) models or kernel ridge regression (KRR), along with the chosen input representation—commonly referred to as molecular descriptors or fingerprints (FP)—which typically encode the presence or frequency of a set of substructures in the form of a bit vector. Thus, substantial research efforts have focused on designing a suitable numerical description of molecules for common chemistry-related ML tasks.

Nevertheless, many of the most successful representations thus far have been designed to replace QM through standard ML techniques by describing the composition and the 3D structure of a chemical system as either an encoded vector through popular fingerprinting algorithms, such as Morgan FP or ECFP,[33] or as a higher dimensional tensor through machine learning interatomic potentials.[34,35] Such models are typically benchmarked against large datasets of DFT-calculated properties. Although these ML models can achieve mean absolute errors (MAEs) below the threshold of "chemical accuracy" (∼1 kcal/mol), the reference values being reproduced (typically DFT) are still significantly inaccurate compared to experiments or more sophisticated CCSD(T)-based cWFTs such as G4 or G4(MP2).[36−38]

While the aforementioned ML models are still useful in some large-scale screening processes in which DFT calculations are too computationally expensive, the problems related to insufficient accuracy of DFT still remain in computational thermochemistry. To mitigate the accuracy loss, different variants of ML models have recently been developed as hybrid QM/ML methods, in which the ML model is tasked to learn the difference between a baseline (typically DFT) and a target (experimental value or more accurate level of theory).[35,37,39,40] In this context, we note that many learning models developed earlier were designed to produce energy (or other properties) solely from a set of atomic numbers $\{Z_1,\cdots, Z_n\}$ and Cartesian coordinates $\{r_1,\cdots, r_n\}$.[34,35,41] While such a simplistic brute-force approach may be appropriate for learning the patterns in the total energy of a molecule, the differences between two levels of theory do not necessarily contain the same patterns. Instead, the present work focuses on designing a physically insightful molecular representation specifically for QM/ML by expanding on the well-established ideas from computational thermochemistry, i.e., error cancellation and fragmentation.

Our goal in this work is to develop a general graph-network framework for combining the strategies from machine learning to further enhance the inherent advantages prebuilt into error cancellation methods. In particular, we will combine the systematic behavior of CBH fragments arising from local connectivities from the molecular structure to build the molecular descriptors for ML. Our framework is termed "FragGraph", or FG(CBH) for short, to denote that it uses fragments from CBH in conjunction with a molecular graph network. The localized fragment-based molecular descriptors (*vide infra*) automatically encode the knowledge about chemical bonds to yield a physically motivated method that incorporates chemical insights. The FragGraph framework is specifically developed for Δ-ML with the central goal of correcting DFT deficiencies to achieve high accuracy. The specific demonstration in this work is based on the CBH-2 rung of the hierarchy (*vide infra*) that can potentially be extended to higher rungs in future work. Nevertheless, the performance of our CBH-2 based ML models for theoretical thermochemistry is quite comparable to the best in the literature (*vide infra*). We also note that while this initial calibration is for theoretical thermochemistry, the FragGraph framework can potentially be used to investigate a broad range of other electronic and spectroscopic properties as well. Finally, the FragGraph framework is designed to incorporate any atom-wise or bond-wise electronic descriptors that can be derived from the baseline calculation for ML. While such electronic descriptors were not needed for theoretical

https://doi.org/10.1021/acs.jpca.1c06152
J. Phys. Chem. A 2021, 125, 6872−6880

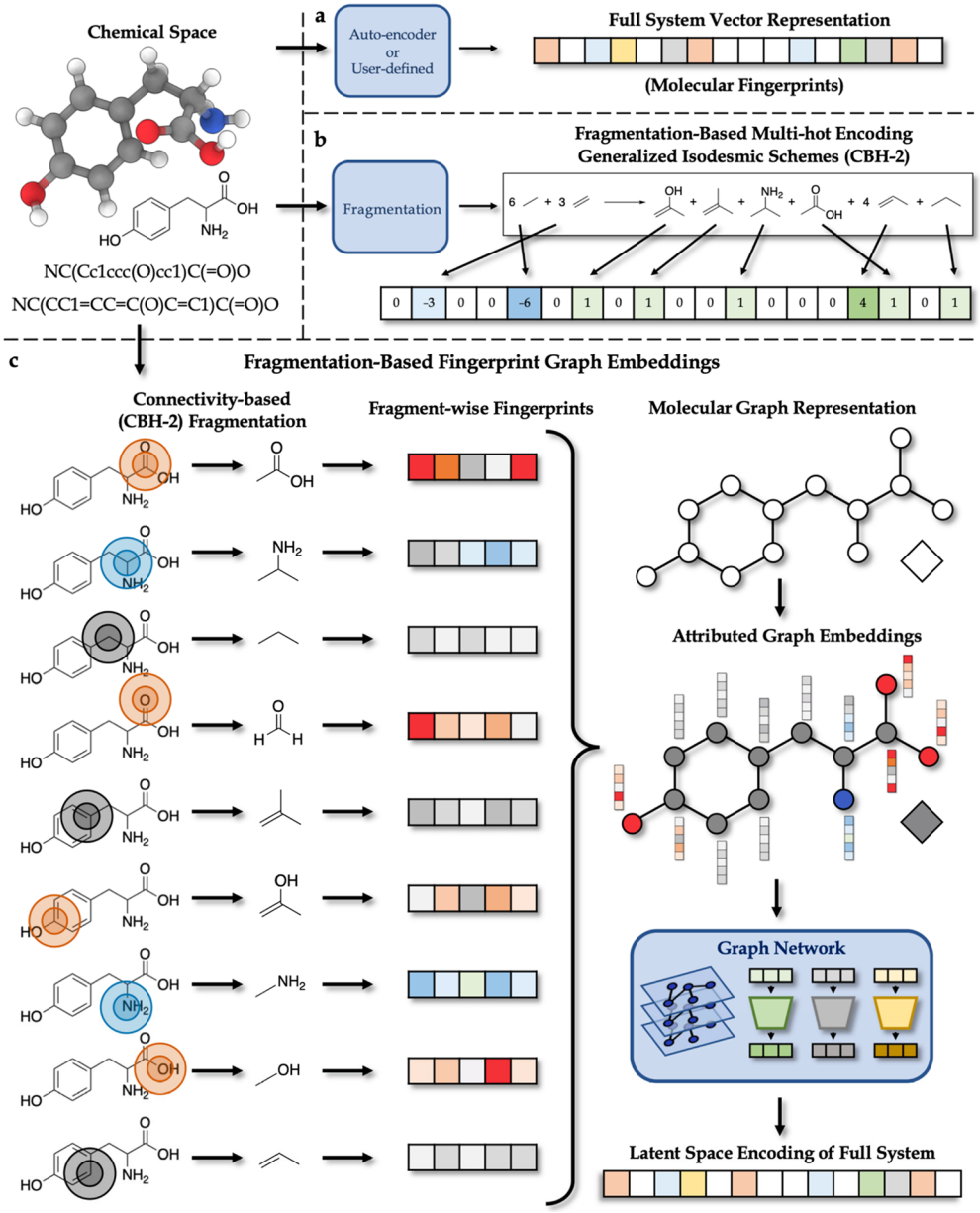


Figure 1. Overview of fingerprint encodings referenced in this study. (a) Traditional fingerprint representations generated through an autoencoder or user-defined, (b) MLCBH representation constructed from the generalized isodesmic scheme,[16] and (c) the FragGraph localized embedding procedure; first the full molecule is fragmented via the CBH-2 scheme, then fragment-wise representations are generated, and finally these fingerprints are embedded into a graph representation of the full molecule and passed through a graph network.

thermochemistry, they may be important for other properties such as redox potentials or electronic excited states.

2. METHODS

2.1. Representation. Standard machine learning techniques often fall short for more challenging tasks, typically resulting in a shift to "deep learning" architectures featuring a large number of stacked units, e.g., hidden layers or convolution operations, capable of learning complex patterns in the data.[41−43] Deep learning models have been revolutionary for a wide range of applications, including deep neural networks (DNN) for standard vector representations,[33,44−48] recurrent neural networks (RNN) for temporal sequences such as speech recognition and natural language processing,[49] and convolutional neural networks (CNN) for image classification.[50] They have also been adapted and showcased in scientific applications, such as protein-folding,[51,52] drug design,[53,54] and synthesis planning.[55,56] A recent review from Google's DeepMind[57] unified these deep learning building blocks and presented a general framework for graph networks (GN, neural networks that operate on graphs).

Defined broadly, a graph is a mathematical data structure consisting of nodes connected by edges that describe information about entities (nodes) and the relationships between them (edges). GN perform edge- and node-wise operations on these graphs to facilitate pattern recognition and connect graph structures to observable trends. These models have already shown great success in general AI applications as much of the structured data in the world can be represented as graphs. Molecular systems are no exception, as chemists typically visualize molecules and molecular reactions through skeletal formulae drawn as a set of atoms with lines between

https://doi.org/10.1021/acs.jpca.1c06152
J. Phys. Chem. A 2021, 125, 6872−6880

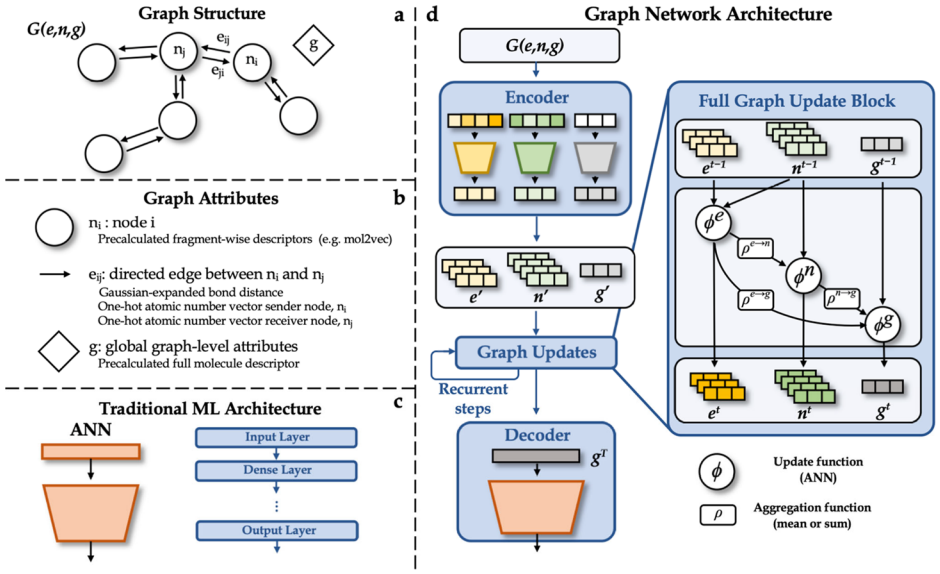


Figure 2. Diagrams of all machine learning architectures and representations used in this work: (a) graph structure $G(e,n,g)$ consisting of edge attributes, node attributes, and the global graph-level vector, (b) initial graph attributes for the graph-based representations, (c) a standard artificial feed-forward neural network (ANN) or a multilayer perceptron (MLP) model with one or more dense layers, which consist of a fully connected layer followed by an activation layer, and (d) graph network architecture used in this work containing three graph blocks: an encoder, updater, and decoder.

them to represent bonds. Indeed, a chemical graph theory has been utilized for decades in cheminformatics to study structure−property relationships and has benefitted immensely from recent advances in deep learning, leading to the development of molecular graph networks, such as MPNN,[42] MEGNet,[43] and SchNet.[41]

Molecular graph networks have many unique advantages over other deep learning architectures. First of all, most ML models require input representations of constant size across all of the chemical space. Graph networks bypass this restriction since graph operations are performed pairwise between nodes along their edges, allowing computations on graphs of arbitrary size. In addition, deep learning approaches, other than GN, receive either fixed-size images (CNN) or sentences (RNN) as input. Pixels of an image can be represented as nodes in a fixed 2D grid-structured pattern with edges connecting neighboring pixels (up, down, left, and right), while sentences are sequences denoted as a linearly directed graph with nodes representing each word. Graphs, on the other hand, are unordered sets with arbitrary structures, which utilize internal coordinates and can feature any number of edges around a single node. Consequently, molecular graphs are invariant to the permutation of atomic indices as well as isometric transformations such as translations and rotations, eliminating the need for any data augmentation or larger training sets to ensure that the models learn such invariances.

Typical methods for constructing molecular representations (Figure 1a) include either a handcrafted fingerprint algorithm or through unsupervised learning with an autoencoder or a clustering method. The former relies heavily on chemical intuition by requiring the user to choose which atomic or molecular attributes are important for the problem at hand. Recent developments from our group proposed a class of fragmentation-based representations termed ML(CBH) or simply MLCBH,[38] in which a system is broken apart into smaller fragments based on the generalized isodesmic schemes of the Connectivity-Based Hierarchy.[16,17] CBH reactions are characterized by deconstructing the molecule into smaller n diameter fragments, corresponding to the nth rung on CBH, as well as their overlaps, to satisfy the inclusion−exclusion principle. Once the full reaction scheme is constructed, the coefficients of the fragments along with their overlaps are multi-hot encoded into a vector of all possible fragments (Figure 1b). The ML(CBH-2) representation achieves a generalization error (out-of-sample MAE) within 0.5 kcal/mol of the CCSD(T)-based cWFT method G4 on the 1k-G4-C9 dataset of (HCNOClS)-containing molecules, outperforming other molecular descriptors, such as the Coulomb matrix[34] (2.77) and bag of bonds[58] (0.81), while featuring a shorter input vector length.

This work extends the ideas of fragmentation and the ML(CBH-2) representation into a graph theoretic framework by encoding molecular fingerprints of fragments onto nodes of a graph (Figure 1c). First, a molecule is decomposed into a set of nodes centered on nonhydrogen atoms. Each node represents an atom-centered fragment defined by CBH-2, including only the immediately connected heavy atoms saturated with hydrogens to preserve the original hybridization. Coincidentally, these CBH-2 fragments align with the maximum diameter considered in ECFP2,[33] both defining each heavy atom's neighborhood within one bond. The FragGraph representation includes a node for each fragment along with edge connections between two nodes if the two atom centers are adjacent in the parent molecule. Next, each fragment is passed individually through a pretrained encoder or algorithm to generate fragment-wise fingerprints. Last, fragment-wise fingerprints are embedded onto their respective nodes. The nodes of the attributed graphs in this work (Figure 2a,b) contain fragment-wise representations calculated from a pretrained mol2vec model (N = 300),[48] an unsupervised natural language processing (NLP)-inspired model that treats Morgan substructures as "words" and molecules as "sentences." The mol2vec model was trained on a compiled database of 20 million biorelevant molecules to return a high-dimensional dense representation of substructures, which can be summed together and used as a molecular descriptor. The directed edges of FragGraphs are encoded with

https://doi.org/10.1021/acs.jpca.1c06152
J. Phys. Chem. A 2021, 125, 6872−6880

one-hot atomic number vectors for the sending and receiving nodes along with the Gaussian-expanded bond distance. Finally, graph-wise global attribute vectors are initialized as the mol2vec representation for the full system. Note that this approach could work for any molecular descriptor in place of mol2vec to generate fragment-wise node embeddings and global feature vectors.

2.2. Architecture. Graph networks are a general class of architectures that map an input graph to an output graph. A full graph block, as defined by ref 57, is a series of attribute updates to the edges, nodes, and global vectors (in that order).[57] These updates take the form of a standard ANN (Figure 2c), which maps a vector to another vector. The full architecture of our graph model (Figure 2d) is composed of three GN blocks resembling the Encode-Process-Decode model of ref 57 with the process step implemented in a manner analogous to the update function of message-passing neural networks (MPNN). The first block GN_{ENC} consists of three independent ANNs (Φ), in which each vector (edges, nodes, and global vectors) of a graph $G(e, n, g)$ is encoded into a latent space graph representation $G'(e', n', g')$

$$G'(e',n',g') = GN_{ENC}(G(e, n, g))$$

$$GN_{ENC} = (\Phi^{e}_{ENC}, \Phi^{n}_{ENC}, \Phi^{g}_{ENC})$$

$$e'_{ij} = \Phi^{e}_{ENC}(e_{ij})$$

$$n'_{i} = \Phi^{n}_{ENC}(n_{i})$$

$$g' = \Phi^{g}_{ENC}(g)$$

Next, latent space graphs are passed to the graph update block, GN_{PROC}, where edges, nodes, and global vectors are updated through a series of message passing steps. Each attribute update consists of two components: (1) the aggregation function ρ and (2) the update function Φ. Graphs contain directed edges to define which vectors to broadcast for neighboring vector updates. In the case of multiple vectors being broadcast to the same location, an aggregation function is applied, typically chosen to be sum, average, or maximum (as appropriate). This ensures that atomic indices of the molecular graphs satisfy permutation invariance. For the update step t, sender nodes n_i are broadcast to update each of the connected edges e_{ij}, then updated edges are aggregated ($\bar{e}^{t}_{j}$) to update common receiving nodes n_j, and finally the updated node vectors are aggregated ($\bar{n}^{t}$) to update the global vector g^t. Note that before passing through their respective update function, the aggregated vector is concatenated with the current vector undergoing the update

$$G^{t}(e^{t}, n^{t}, g^{t}) = GN_{PROC}(G^{t-1}(e^{t-1}, n^{t-1}, g^{t-1}))$$

$$GN_{PROC} = (\Phi^{e}_{PROC}, \Phi^{n}_{PROC}, \Phi^{g}_{PROC})$$

$$e^{t}_{ij} = \Phi^{e}_{PROC}(e^{t-1}_{ij}, n^{t-1}_{i})$$

$$\bar{e}^{t}_{j} = \rho^{e\to n}(e^{t}_{ij})$$

$$n^{t}_{j} = \Phi^{n}_{PROC}(\bar{e}^{t}_{j}, n^{t-1}_{j})$$

$$\bar{n}^{t} = \rho^{n\to g}(n^{t}_{j})$$

$$g^{t} = \Phi^{g}_{PROC}(\bar{n}^{t}, g^{t-1})$$

After the T graph updates, the global vector g^T is taken as the final latent space representation of the full molecule to be used for ML, in this case, paired with $\Phi^{g}_{DEC,}$ e.g., a standard ANN or other decoder function. Thus, the first two GN blocks, GN_{ENC} and GN_{PROC}, can be viewed as a molecular representation encoder, transforming an input graph to a single latent space vector

$$Y = GN_{DEC}(G^{T}(e^{T}, n^{T}, g^{T}))$$

$$GN_{DEC} = (\Phi^{g}_{DEC})$$

$$y = \Phi^{g}_{DEC}(g^{t})$$

In total, the Encode-Process-Decode model has nine neural networks, three for each GN block, returning a graph as the final output. The GN architecture used in this work consists of seven independent neural networks disregarding the final node and edge vectors by learning the full molecule energy (or energy differences) as a global attribute. The final graph outputs could be utilized further to learn other properties that depend on each node or edge.

3. RESULTS AND DISCUSSION

3.1. Results on the G4(MP2)-GDB9 Dataset. Both ANN and GN models were trained on up to 117k training molecules with the remaining 13k of the GDB9 dataset acting as the out-of-sample generalization set.[36,59] Learning curves for five models (Figure 3) were tested with the same 13k test set for every training point on each curve. The models include two standard molecular representations, mol2vec and ML(CBH-2), each paired with an ANN. The other three are graph network models: a simple graph model starting with only atomic number and bond length information and two fragment-embedded graph models: a skeletal graph that is locally connected through bonding interactions of fragments alone and a complete graph in which every node is connected to every other node. All graph

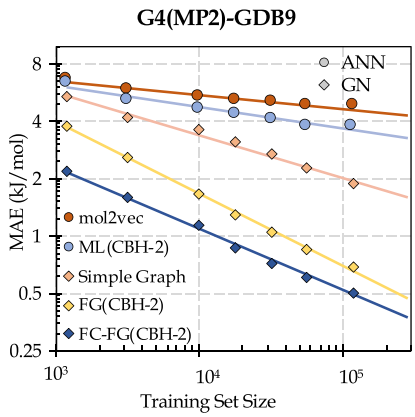


Figure 3. Generalization performance (out-of-sample mean absolute error in kJ/mol) of five representations on the G4(MP2)-GDB9 dataset. Simple feed-forward ANNs are used with traditional molecular representations mol2vec and MLCBH2, while graph networks (GN) are paired with the simple graph representation (atomic number) and fragment-wise embedded graphs FragGraph. FG(CBH-2) models use either locally connected (FG) or fully connected (FC-FG) update steps.

https://doi.org/10.1021/acs.jpca.1c06152
J. Phys. Chem. A 2021, 125, 6872−6880

networks employ the same architecture and hyperparameters. Fragment-embedded graphs were the top-performing models in this study, reaching a generalization error between 0.50−0.68 kJ/mol on the unseen test set. By comparison, mol2vec fared much worse with an MAE of 4.8 kJ/mol and the previous fragmentation-based fingerprint ML(CBH-2) had an MAE of 3.8 kJ/mol. The simple graph model performed better than the standard representation methods at an MAE of 1.7 kJ/mol, indicating that the graph network structure itself is responsible for a significant part of the improvement in overall performance.

FG(CBH-2) is a combination of standard mol2vec, ML(CBH-2), and simple graph models but performs significantly better than all three methods, indicating that the graph network model is learning more complex patterns about the relationships between embedded fragments. Graph updates on the FragGraph representation not only capture the graph structure of the neighborhood but also the local representations of nearby nodes allowing for learning of perturbations from the initial mol2vec vectors.

Learning curves can give useful insights into the generalizability of a model as well as its training efficiency. Typically, these curves are approximately linear in the log−log space due to the inverse power law relationship between the generalizability of a model and the number of data points in the training set.[58,60] At the asymptotic limit, the generalization error should scale as approximately $1/t$, where t is the number of training samples. In practice, a faster $1/t^2$ scaling is observed during training at medium training set sizes and a large deviation from this pattern is seen with small training set sizes due to a large amount of overfitting. Learning curves for all methods in this work are approximately linear in the log−log space, indicating that the training set sizes are sufficiently large to learn generalizable patterns in our data. Moreover, GN models feature steeper slopes than ANN models, showcasing their ability to generalize with smaller datasets. Indeed, embedded graph models outperform the best simple graph counterpart (117k training data points) with a factor of 10−20 less training data (starting around 5k to 10k) and cross the 1 kJ/mol threshold with merely 17k and 50k training data points for the two models. Additionally, even the simple graph-based model provided with information about atomic numbers and interatomic distances reaches a generalization error within the typical range of chemical accuracy (1 kcal/mol) after being trained on 10k training data points, outperforming the standard ANN-based models given the full 117k training set.

Although this work primarily focuses on eliminating the systematic errors of DFT, models utilizing less expensive baseline calculation, i.e., semi-empirical method PM7, were also trained and compared in Table 1. FG(CBH-2) models reduced the out-of-sample errors to just under 0.5 kcal/mol, which is around 5−6% of the baseline PM7 MAE of 7.99 kcal/mol. As a comparison, the models trained on the B3LYP baseline achieved around 3−4% of the uncorrected DFT MAE of 4.63 kcal/mol. The increase in performance from the standard mol2vec representation to the FC-FG(CBH-2) models was similar, decreasing the MAE by a factor of 10, for both PM7 and B3LYP baselines. These results indicate that our fragmentation-based framework could potentially be useful in other applications, such as virtual screening, when paired with less expensive quantum chemical methods.

Table 1. Generalization Performance (Mean Absolute Errors in kcal/mol) of FG(CBH-2) and ANN(mol2vec) Models Compared to Baseline Methods

model	number of molecules on which MAE is based	PM7	B3LYP
uncorrected	13,026	7.99	4.63
ANN(mol2vec)	13,026	3.96	1.15
FG(CBH-2)	13,026	0.50	0.16
FC-FG(CBH-2)	13,026	0.38	0.12

Graph networks have an inherent affinity for generalization due to their large amount of parameter sharing in the encoder and update blocks. Each block is composed of three ANN update functions: Φ^e, Φ^n, and Φ^g, which maps each of the edges, nodes, and global vectors to a corresponding updated vector based on the surrounding neighborhood. As an example, the full 130k GDB9 dataset contains 870k nodes connected by approximately 2M edges in the locally connected FragGraph representation and 9M edges for the fully connected variant. Accordingly, the number of vectors passed through each Φ in the encoder block is equal to the total number of edges, nodes, and global vectors in the training set, while this number is multiplied by the number of recurrent update passes T for the graph update block. As a result, each of the ANN update functions in the encoder and update blocks is able to learn from a larger pool of data, leading to a more widespread generalization.

3.2. Comparison to Other Methods. Machine learning-based interatomic potentials (MLIP) and deep tensor neural networks (DTNN) are among the current state-of-the-art deep learning methods. In the context of graph networks, these methods employ complete graphs with nodes representing every atom of a chemical system. For example, one of the best methods, SchNet,[41] starts with a set of atomic numbers individually mapped to latent space vectors. Each atomic vector is updated through multiple convolution layers based on the distances to every other atom in the molecule. After convolutions, the final vector representation for each atom is converted to an atom-wise contribution to the total energy and summed to give the energy of the molecule. Similar ideas are present in many machine learning-based atomic potentials, which learn energy contributions from individual units of a molecular system.[34,35] In contrast, the FG(CBH-2) model more closely resembles a molecular representation encoder, which learns a global vector to represent the full molecule starting from local representations of its fragments.

Although SchNet was originally designed for standard ML, with the task of reproducing the total energy of a molecule from atomic numbers and coordinates, it has more recently been applied as a hybrid QM/ML protocol, performing with an MAE of 0.1 kcal/mol on the G4(MP2)-GDB9 dataset.[37,61] While the best FG(CBH-2) models in this work are competitive in performance with field-leading models, our aim is more focused on improving existing descriptors through the ideas of fragmentation and the relational inductive biases from graph networks. Nonetheless, the FragGraph representation features a few advantages, including reduced scaling of the number of edges, decreased computational complexity, and the ability to generalize to more difficult problems through the choice of fragment-wise vectors. Unlike complete graphs in MLIP, FragGraph was designed to simplify the system into individual coarse-grained units. In this work, these units are nonhydrogen atoms. Additionally, the fragment-wise vectors on these units encode information about the surrounding environment, reducing the number of required edges for graph updates in order to learn a sufficient representation. Since a graph's edges

https://doi.org/10.1021/acs.jpca.1c06152
J. Phys. Chem. A 2021, 125, 6872−6880

dictate the total number of update operations, the computational scaling of graph-based methods can be approximated by the number of edges. For the GDB9 dataset, the average number of edges was calculated (Figure 4) for three separate cases:

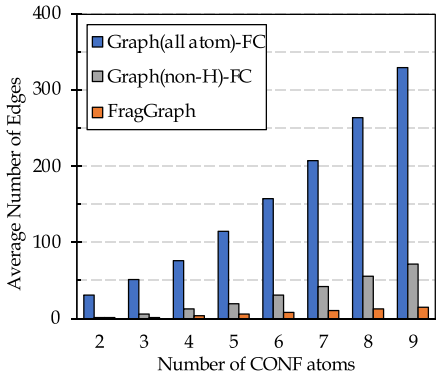


Figure 4. Comparison of the scaling of the average number of edges for different graph-based deep learning representations.

complete graphs of fully connected atoms Graph(all atom)-FC, complete graphs of fully connected non-H atoms Graph(non-H)-FC, and our locally connected FragGraph. In the fully connected non-H atom case, the number of edges grows as $N(N-1)$, where N is the number of nodes (corresponding to CONF atoms). The number of neighbors in the FragGraph representation grows approximately linearly with each node added, introducing 2−4 new edges for linear or branched molecules. At the asymptotic limit, the edges of complete graphs scale as $O(N^2)$, while locally connected graphs scale as $O(N)$. The inclusion of hydrogen nodes in Graph(all-atom)-FC, as done in the SchNet model, significantly increases the total number of edges for a fully connected graph. Thus, the FragGraph representation provides a more efficient molecular structure encoding while performing similar to other deep learning methods, at a reduced cost while requiring much fewer edges.

3.3. Fingerprint Similarity. Pivotal to the success of structure−activity relationships is describing structural characteristics of molecules as a vector through molecular fingerprinting. Molecular fingerprints are essential tools for cheminformatics studies describing chemical diversity in chemical space as well as virtual screening and similarity searching in drug design. Chemical fingerprint space is a multidimensional conceptual space in which dimensions represent properties of the molecular structure or the feature vector of a fingerprint. In this context, molecules are placed on the coordinates corresponding to their feature vector, and the distance between two points can be used as a measure of similarity. The distribution of these distances is therefore important for understanding how well a given molecular representation can distinguish chemicals with similar structures. Pairwise distances between all vectors were calculated for both mol2vec and FG(CBH-2) to compare their distributions (Figure 5). Since mol2vec is essentially the starting point to the FG(CBH-2) model, a shift to a wider distribution indicates a higher differentiation between different molecules. Furthermore, FragGraph models can eliminate deficiencies of the base-level fragment-wise representation chosen. For mol2vec, there were over 2 million pairwise distances of 0.0, mainly due to the inability to distinguish some constitutional isomers. The graph network-based models could learn directly from the connectivity

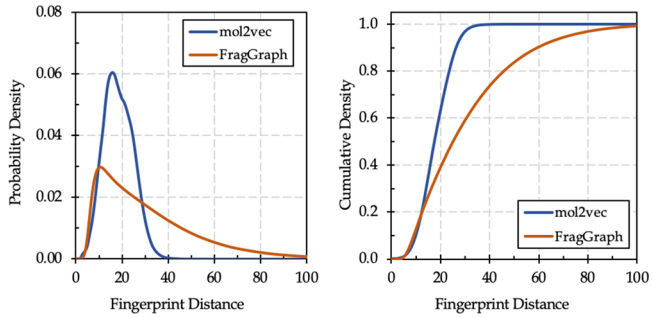


Figure 5. Comparison of pairwise distance distributions for standard mol2vec- and FG(CBH-2)-generated molecular representations starting from fragment-wise mol2vec graph embeddings based on CBH-2.

of the fragments to the point that no two molecules in our dataset were represented by the same vector. The drastic increase in performance from mol2vec (4.8 kJ/mol) to FG(CBH-2) utilizing the mol2vec feature vector (0.5−0.68 kJ/mol) can be attributed, at least in part, to this broader distinction of similarly structured molecules. Additionally, the FragGraph framework is capable of interfacing with virtually any molecular representation or fingerprinting method for the fragment-wise vector embeddings, potentially improving many other popular descriptors. Once a molecule is coarse-grained into a graph, nodes can be embedded with any vector or properties of the molecular fragment the unit represents, allowing the FragGraph representation to be applied to more complex problems where structural information may be insufficient.

4. CONCLUSIONS

Molecular graph networks provide a deep learning framework allowing for the accurate prediction of thermochemistry. Many strategies in computational thermochemistry, such as fragmentation and error cancellation, have not yet been explored fully within the realm of machine learning, despite the potential for leading to new strategies for designing molecular representations. In this context, our work provides a new foundation for QM/ML-based methods with the following conclusions:

- Deep learning and the FragGraph representation can be used to obtain accuracy in the sub-kJ/mol range compared to G4(MP2). Since the deviation of G4(MP2) to experimental values is an order of magnitude higher than this range, errors from the model are predominately from the reference calculation.
- FragGraph provides a general framework that combines fragmentation, error cancellation, and graph networks. These fragmentation-based embeddings could be used in conjunction with a wide range of fragmentation schemes and currently available fingerprints. As shown above, the standard mol2vec representation was drastically improved through the use of FG(CBH-2) in terms of overall performance and fingerprint similarity.
- FG(CBH-2) is competitive in performance to current state-of-the-art QM/ML methods, which have a steeper computational scaling. FragGraph provides a computational cost reduction from the simplification of molecular systems through chemical environment embeddings, implicit hydrogens, and locally connected graphs.

https://doi.org/10.1021/acs.jpca.1c06152
J. Phys. Chem. A 2021, 125, 6872−6880

- The FragGraph framework is not limited to computational thermochemistry and can potentially be used to investigate a broad range of other electronic and spectroscopic properties as well. The incorporation of additional atom-wise or bond-wise electronic descriptors from the baseline calculation may be useful in such applications that we plan to pursue in the future.

AUTHOR INFORMATION

Corresponding Authors

Eric M. Collins − *Department of Chemistry, Indiana University, Bloomington, Indiana 47405, United States;* orcid.org/0000-0002-9113-1705; Email: colliner@indiana.edu

Krishnan Raghavachari − *Department of Chemistry, Indiana University, Bloomington, Indiana 47405, United States;* orcid.org/0000-0003-3275-1426; Email: kraghava@indiana.edu

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpca.1c06152

Notes

The authors declare no competing financial interest.

Additional information and sample code for input representation generation for this article may be accessed at https://github.com/colliner/FragGraph.

ACKNOWLEDGMENTS

We acknowledge support from the National Science Foundation grants CHE-1665427 and CHE-2102583 at Indiana University. The computations carried out in this work were enabled in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute.

REFERENCES

(1) Martin, J. M. L.; de Oliveira, G. Towards standard methods for benchmark quality *ab initio* thermochemistry—W_1 and W_2 theory. *J. Chem. Phys.* **1999**, *111*, 1843−1856.

(2) Boese, A. D.; Oren, M.; Atasoylu, O.; Martin, J. M. L.; Kallay, M.; Gauss, J. W_3 theory: Robust computational thermochemistry in the kJ/mol accuracy range. *J. Chem. Phys.* **2004**, *120*, 4129−4141.

(3) Karton, A.; Sylvetsky, N.; Martin, J. M. L. W4-17: A Diverse and High-Confidence Dataset of Atomization Energies for Benchmarking High-Level Electronic Structure Methods. *J. Comput. Chem.* **2017**, *38*, 2063−2075.

(4) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. A fifth-order perturbation comparison of electron correlation theories. *Chem. Phys. Lett.* **1989**, *157*, 479−483.

(5) Karton, A. A computational chemist's guide to accurate thermochemistry for organic molecules. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2016**, *6*, 292−310.

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

(7) Das, S. K.; Chakraborty, S.; Ramakrishnan, R. Critical benchmarking of popular composite thermochemistry models and density functional approximations on a probabilistically pruned benchmark dataset of formation enthalpies. *J. Chem. Phys.* **2021**, *154*, No. 044113.

(8) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gaussian-4 theory. *J. Chem. Phys.* **2007**, *126*, No. 084108.

(9) Tajti, A.; Szalay, P. G.; Császár, A. G.; Kállay, M.; Gauss, J.; Valeev, E. F.; Flowers, B. A.; Vázquez, J.; Stanton, J. F. HEAT: High accuracy extrapolated *ab initio* thermochemistry. *J. Chem. Phys.* **2004**, *121*, 11599−11613.

(10) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gaussian-4 theory using reduced order perturbation theory. *J. Chem. Phys.* **2007**, *127*, 124105.

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

(12) Montgomery, J. A., Jr.; 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.* **2000**, *112*, 6532−6542.

(13) Simmie, J. M.; Somers, K. P. Benchmarking Compound Methods (CBS-QB3, CBS-APNO, G3, G4, W1BD) against the Active Thermochemical Tables: A Litmus Test for Cost-Effective Molecular Formation Enthalpies. *J. Phys. Chem. A* **2015**, *119*, 7235−7246.

(14) Hehre, W. J.; Ditchfield, R.; Radom, L.; Pople, J. A. Molecular orbital theory of the electronic structure of organic compounds. V. Molecular theory of bond separation. *J. Am. Chem. Soc.* **1970**, *92*, 4796−4801.

(15) Chan, B.; Collins, E.; Raghavachari, K. Applications of isodesmic-type reactions for computational thermochemistry. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2021**, *11*, No. e1501.

(16) Ramabhadran, R. O.; Raghavachari, K. Theoretical Thermochemistry for Organic Molecules: Development of the Generalized Connectivity-Based Hierarchy. *J. Chem. Theory Comput.* **2011**, *7*, 2094−2103.

(17) Ramabhadran, R. O.; Raghavachari, K. The Successful Merger of Theoretical Thermochemistry with Fragment-Based Methods in Quantum Chemistry. *Acc. Chem. Res.* **2014**, *47*, 3596−3604.

(18) Ramabhadran, R. O.; Raghavachari, K. Connectivity-Based Hierarchy for Theoretical Thermochemistry: Assessment Using Wave Function-Based Methods. *J. Phys. Chem. A* **2012**, *116*, 7531−7537.

(19) Ramabhadran, R. O.; Sengupta, A.; Raghavachari, K. Application of the Generalized Connectivity-Based Hierarchy to Biomonomers: Enthalpies of Formation of Cysteine and Methionine. *J. Phys. Chem. A* **2013**, *117*, 4973−4980.

(20) Sengupta, A.; Raghavachari, K. Prediction of Accurate Thermochemistry of Medium and Large Sized Radicals Using Connectivity-Based Hierarchy (CBH). *J. Chem. Theory Comput.* **2014**, *10*, 4342−4350.

(21) Sengupta, A.; Ramabhadran, R. O.; Raghavachari, K. Accurate and Computationally Efficient Prediction of Thermochemical Properties of Biomolecules Using the Generalized Connectivity-Based Hierarchy. *J. Phys. Chem. B* **2014**, *118*, 9631−9643.

(22) Collins, E. M.; Sengupta, A.; AbuSalim, D. I.; Raghavachari, K. Accurate Thermochemistry for Organic Cations via Error Cancellation using Connectivity-Based Hierarchy. *J. Phys. Chem. A* **2018**, *122*, 1807−1812.

(23) Maier, S.; Thapa, B.; Raghavachari, K. G4 accuracy at DFT cost: unlocking accurate redox potentials for organic molecules using systematic error cancellation. *Phys. Chem. Chem. Phys.* **2020**, *22*, 4439−4452.

(24) Thapa, B.; Raghavachari, K. Accurate pK_a Evaluations for Complex Bio-Organic Molecules in Aqueous Media. *J. Chem. Theory Comput.* **2019**, *15*, 6025−6035.

(25) Debnath, S.; Sengupta, A.; Raghavachari, K. Eliminating Systematic Errors in DFT via Connectivity-Based Hierarchy: Accurate Bond Dissociation Energies of Biodiesel Methyl Esters. *J. Phys. Chem. A* **2019**, *123*, 3543−3550.

(26) Raghavachari, K.; Saha, A. Accurate Composite and Fragment-Based Quantum Chemical Models for Large Molecules. *Chem. Rev.* **2015**, *115*, 5643−5677.

(27) Mayhall, N. J.; Raghavachari, K. Molecules-in-Molecules: An Extrapolated Fragment-Based Approach for Accurate Calculations on Large Molecules and Materials. *J. Chem. Theory Comput.* **2011**, *7*, 1336−1343.

(28) Zhang, D. W.; Zhang, J. Z. H. Molecular fractionation with conjugate caps for full quantum mechanical calculation of protein−molecule interaction energy. *J. Chem. Phys.* **2003**, *119*, 3599−3605.

(29) Gadre, S. R.; Shirsat, R. N.; Limaye, A. C. Molecular Tailoring Approach for Simulation of Electrostatic Properties. *J. Phys. Chem.* **1994**, *98*, 9165−9169.
(30) Collins, M. A.; Cvitkovic, M. W.; Bettens, R. P. A. The Combined Fragmentation and Systematic Molecular Fragmentation Methods. *Acc. Chem. Res.* **2014**, *47*, 2776−2785.
(31) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087−2096.
(32) Zaspel, P.; Huang, B.; Harbrecht, H.; von Lilienfeld, O. A. Boosting Quantum Machine Learning Models with a Multilevel Combination Technique: Pople Diagrams Revisited. *J. Chem. Theory Comput.* **2019**, *15*, 1546−1559.
(33) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(34) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326−2331.
(35) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat. Commun.* **2019**, *10*, 2903.
(36) Narayanan, B.; Redfern, P. C.; Assary, R. S.; Curtiss, L. A. Accurate quantum chemical energies for 133 000 organic molecules. *Chem. Sci.* **2019**, *10*, 7449−7455.
(37) Ward, L.; Blaiszik, B.; Foster, I.; Assary, R. S.; Narayanan, B.; Curtiss, L. Machine learning prediction of accurate atomization energies of organic molecules from low-fidelity quantum chemical calculations. *MRS Commun.* **2019**, *9*, 891−899.
(38) Collins, E. M.; Raghavachari, K. Effective Molecular Descriptors for Chemical Accuracy at DFT Cost: Fragmentation, Error-Cancellation, and Machine Learning. *J. Chem. Theory Comput.* **2020**, *16*, 4938−4950.
(39) von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Exploring chemical compound space with quantum-based machine learning. *Nat. Rev. Chem.* **2020**, *4*, 347−358.
(40) Rupp, M. Machine learning for quantum mechanics in a nutshell. *Int. J. Quantum Chem.* **2015**, *115*, 1058−1073.
(41) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet − A deep learning architecture for molecules and materials. *J. Chem. Phys.* **2018**, *148*, 241722.
(42) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry.In *Proceedings of the 34th International Conference on Machine Learning*; Cornell University *arXiv e-prints* 2017, arXiv:1704.01212.
(43) Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.* **2019**, *31*, 3564−3572.
(44) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular graph convolutions: moving beyond fingerprints. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 595−608.
(45) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **2017**, *3*, 434−443.
(46) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci.* **2016**, *2*, 725−732.
(47) O'Boyle, N. M.; Sayle, R. A. Comparing structural fingerprints using a literature-based similarity benchmark. *J. Cheminf.* **2016**, *8*, 36.
(48) Jaeger, S.; Fulle, S.; Turk, S. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *J. Chem. Inf. Model.* **2018**, *58*, 27−35.
(49) Hewamalage, H.; Bergmeir, C.; Bandara, K. Recurrent Neural Networks for Time Series Forecasting: Current status and future directions. *Int. J. Forecast.* **2021**, *37*, 388−427.
(50) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. In ImageNet Classification with Deep Convolutional Neural Networks, *Advances in Neural Information Processing Systems*, 2012; Pereira, F.; Burges, C. J. C.; Bottou, L.; Weinberger, K. Q. Eds. Curran Associates, Inc.
(51) Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. Protein−Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.* **2017**, *57*, 942−957.
(52) Xu, J. Distance-based protein folding powered by deep learning. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 16856.
(53) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The rise of deep learning in drug discovery. *Drug Discovery Today* **2018**, *23*, 1241−1250.
(54) Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; et al. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discovery* **2019**, *18*, 463−477.
(55) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018**, *51*, 1281−1289.
(56) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* **2019**, *10*, 370−377.
(57) Battaglia, P. W.; Hamrick, J. B.; Bapst, V.; Sanchez-Gonzalez, A.; Zambaldi, V.; Malinowski, M.; Tacchetti, A.; Raposo, D.; Santoro, A.; Faulkner, R. et al. Relational inductive biases, deep learning, and graph networks. *arXiv e-prints;* Cornell University 2018, arXiv:1806.01261.
(58) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, No. 058301.
(59) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci Data* **2014**, *1*, 140022.
(60) Müller, K.-R.; Finke, M.; Murata, N.; Schulten, K.; Amari, S. A Numerical Study on Learning Curves in Stochastic Multilayer Feedforward Networks. *Neural Comput.* **1996**, *8*, 1085−1106.
(61) Dandu, N.; Ward, L.; Assary, R. S.; Redfern, P. C.; Narayanan, B.; Foster, I. T.; Curtiss, L. A. Quantum-Chemically Informed Machine Learning: Prediction of Energies of Organic Molecules with 10 to 14 Non-hydrogen Atoms. *J. Phys. Chem. A* **2020**, *124*, 5804−5811.

https://doi.org/10.1021/acs.jpca.1c06152
J. Phys. Chem. A 2021, 125, 6872−6880