Graph to Activation Energy Models Easily Reach Irreducible Errors but Show Limited Transferability

Sai Mahit Vadaddi,[†] Qiyuan Zhao,[*,‡] and Brett M. Savoie[*,†]

[†]*Davidson School of Chemical Engineering, Purdue University, West Lafayette, IN, 47906*

[‡]*Department of Medicinal Chemistry, University of Michigan, Ann Arbor, MI, 48109*

E-mail: zhaoqiy@med.umich.edu; bsavoie@purdue.edu

Abstract

Activation energy characterization of competing reactions is a costly, but crucial step for understanding the kinetic relevance of distinct reaction pathways, product yields, and myriad other properties of reacting systems. The standard methodology for activation energy characterization has historically been a transition state search using the highest level of theory that can be afforded. However recently, several groups have popularized the idea of predicting activation energies directly, based on nothing more than the reactant and product graphs, a sufficiently complex neural network, and a broad enough dataset. Here, we have revisited this task using the recently developed Reaction Graph Depth 1 (RGD1) transition state dataset and several newly developed graph attention architectures. All of these new architectures achieve similar state-of-the-art results of ∼4 kcal/mol mean absolute error on withheld testing sets of reactions but poor performance on external testing sets composed

of reactions with differing mechanisms, reaction molecularity, or reactant size distribution. Limited transferability is also shown to be shared by other contemporary graph to activation energy architectures through a series of case-studies. We conclude that an array of standard graph architectures can already achieve results comparable to the irreducible error of available reaction datasets but that out-of-distribution performance remains poor.

1 Introduction

Reaction activation energies (E_a) and heats of reaction (ΔH_r) are essential to understanding of reactivity in applications ranging from biofuel utilization,[1–6] drug design,[7–10] and materials stability.[11–13] Collecting this information from experiments is costly,[14–18] making it highly desirable to develop predictive methods that can be used prior to synthesis to expedite hypothesis formation and optimization.[19–25] Over the past several decades, quantum chemistry has delivered many algorithms for localizing transition states (TSs) and characterizing activation energies;[26–31] however, finding transition states remains relatively expensive for on-the-fly and high-throughput applications. It would be a qualitative advance if reaction properties like activation energy could be directly calculated without first localizing a transition state.

In the past few years several groups have shown the feasibility of predicting activation energies from only the reactant and product graphs (Fig. 1A). Early examples focused on summarizing changes between reactants and products using expert-generated features[32–36] and molecular fingerprints.[37–40] With the advent of larger reaction datasets,[41–43] several standard practices have been identified, including the necessity to avoid including reverse reactions in testing datasets (i.e., having an example in the testing set that was seen during training with the reactant and product switched),[44] the potential usefulness of the heat of reaction (ΔH_r) as an input feature,[38,44] and the advantage of learnable reaction fingerprints over pre-defined expert fingerprints.[44–46] In paral-

lel, prediction strategies are also being developed based on three-dimensional featurizations of the reactant and product,[47,48] relatively inexpensive information from approximate levels of theory,[49] and additional features from quantum chemistry.[39,50–55] The cost of additional features can nullify the advantage of using a graph-based approach (albeit, with potentially higher transferability). In this sense, the ideal model would be able to achieve high accuracy based solely on the reactant and product graphs. The negotiation of these trade-offs remains a live issue.[44,56,57]

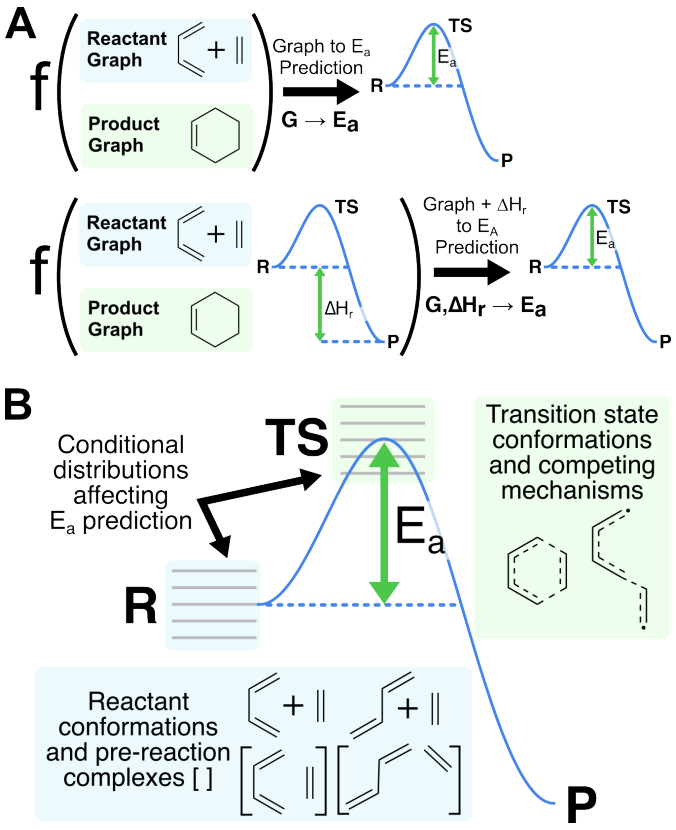

Figure 1: Overview of the graph to activation energy (G2Ea) prediction task. (a) A minimal featurization of this problem consists of only using the reactant and product graphs, while sometimes relatively inexpensive but informative features like heat of reaction (ΔH_r) are also used. (b) E_a depends on the reference initial state and the sampling distribution used to localize the transition state. Illustrative examples are shown for the sequential versus concerted Diels-Alder mechanisms. Comparisons across datasets and predictions on unseen reactions will be out-of-distribution if such factors are not consistently sampled.

Despite many practical demonstrations of the graph-to-activation-energy (G2Ea) concept, sev-

eral challenges persist that limit the usefulness of these models as drop-in replacements for quantum-chemistry based TS searches. One challenge is that the scarcity of large reaction datasets has limited convincing out-of-distribution tests of the transferability of G2Ea models. Under the assumption that reaction mechanisms are conserved, G2Ea models should be capable of extrapolating to unseen substrates involved in classes of reactions that have been directly trained on, but this remains a largely untested hypothesis.

A more fundamental challenge that is specific to the G2Ea learning task is that the prediction problem is underdetermined with respect to the manner in which computational E_{a} datasets are currently generated (Fig. 1B). While experiments measure an effective activation energy based on the Boltzmann average over all accessible conformations and transition states, computationally derived activation energies are typically extracted from a single pair of energies corresponding to a particular reactant and TS conformation that are not always uniquely defined. Available datasets have no guarantee of conformational completeness or of having found a globally minimum energy barrier for the observed reactions. This will lead to an irreducible error for any G2Ea model trained on available quantum-chemistry derived datasets, since predictions are conditioned on the conformational distribution of reactants and TSs used during curation. For example, the Conformer-Rotamer Ensemble Sampling Tool (CREST)[58] algorithm was used during the curation of the Reaction Graph Depth 1 (RGD1) dataset[41] to find the minimum energy conformer of the isolated reactants, and a protocol specific to Yet Another Reaction Program (YARP) was used to select up to three conformations for double-ended TS searches.[59] These choices will show up as inductive biases in models trained on RGD1 and any incomplete conformational sampling will show up as an irreducible error when predicting on unseen reactions.

Here, the transferability of G2Ea models has been revisited using the recently developed RGD1 dataset and an adaptation of the graph attention architecture.[60–62] The motivation for this study was that the size and mechanistic diversity of the RGD1 dataset potentially allows for the training

of more data-demanding architectures with better transferability. Using a graph attention architecture we are able to train models that reliably approach the estimated irreducible error of RGD1 and perform well on withheld reactions drawn from the same distribution. Nevertheless, these models show minimal transferability in external testing scenarios, oftentimes performing worse than naïve mean-predicting models. Through several comparative case-studies it is shown that this behavior is shared by models using closely related featurizations, and other contemporary graph to property architectures recently published for use in the G2Ea prediction task.

2 Methods

Several variations of the Edge-featured Graph Attention Network[62] (EGAT) architecture are developed here for the prediction of activation energies from reactant and product graphs. The main elements shared by all the architectures are described in the Model Overview, Input Features, and Description subsections (Sections 2.1-2.3) and the differences between models are described in Learning Tasks subsection (Section 2.4). These models are compared with Chemprop,[63] a directed message passing neural network (D-MPNN) model that has previously been trained for the G2Ea task. The training and implementation details of Chemprop in this work followed those supplied by the developers through their distributed code,[63] additional can be found in the SI (Section 2).

2.1 Model Overview

Edge-featured Graph Attention Networks (EGAT) are a subset of Graph Attention Networks (GAT).[60–62] The basic idea behind EGATs is to use the features of each edge to create an attention score that is used to weight the information mixing between nodes. Each pass through an EGAT layer results in the mixing of information between nodes as determined by learnable attention scores. Thus, the use of n EGAT layers results in the mixing of information from nodes up to n

edges away into the fingerprint of each node. A fingerprint of the whole reaction graph is obtained by pooling the node and edge fingerprints, which can be used for reaction property prediction.

Here, the EGAT architecture is adapted for E_a prediction by featurizing each atom as a node and each bond as an edge. The reactant and product are separately passed to the model, converted to individual fingerprints, and then the fingerprint difference is used to predict E_a subject to several small architectural variations (See Section 2.4). The relative atom-mapping between reactants and products affects the prediction of the model, but the architecture includes an intermediate pooling operation to ensure the prediction is invariant to the absolute numbering of the atom sequence. Differences in atom mapping reflect distinct bond-changes and so the atom-mapping awareness is an important feature of the model that is not captured by models based on simpler Morgan fingerprints.

The edge topology that is used for the reactant and product corresponds to the union of the bonds that are presented in either of the species. This results in molecular fingerprints for the reactant and product that have the same number of edges when taking the fingerprint difference. Thus, when processed by the EGAT model, bonds that are only physically present in the reactant or product (i.e., they are formed or broken in the reaction) are still present as edges in the product and reactant graphs, respectively, albeit with a special feature indicating that the bond was broken or formed.

2.2 Input Features

The input to the EGAT architecture is an ordered set of features for each atom and bond in the reactant and product. Table 1 and 2 list the input atom (i.e., node) features, $\mathbf{n}$, and bond (i.e., edge) features, $\mathbf{e}$, fed into the EGAT model, respectively, and their comparisons with input features of Chemprop. Many of the features could be incorporated as either distinct numeric values (i.e., integers) or categorical values (i.e., one-hot vectors). In such cases we elected to use integers over

one-hot vectors wherever possible to simplify the representation. For example, hybridization in EGAT is represented using a four bit one-hot vector (i.e., this is clearly a categorical feature), but the number of bonded hydrogens are featurized with integers (i.e., this is a case that could be treated either numerically or categorically). Wherever one-hots are used, they are transposed and concatenated with the other features during the input preprocessing. The resulting dimensions for the featurized node and edge vectors are 17x1 and 14x1, respectively.

The rationale for some of the new features is as follows. A node's proximity to a bond that is broken or formed is novel to the EGAT model, which denotes the effect it may have on the reaction as atoms closer to the reaction center exhibit more local changes than those farther away. Node features corresponding to the number of bonded CHON atoms were also directly featurized, even though this is an implicit feature that might be learned from the convolution. The novel edge features include a set of binary features corresponding to whether the bond was broken, formed, if the bond order changed, or was unchanged. These features are necessary to use the union of the reactant and product graphs as the inputted graph topology to the model. For example, the reactant may have some edges with bond order zero, because these might only exist in the product, but it is nonetheless potentially useful for the model to mix information along these edges given the fact that a bond forms or breaks between such atoms in the reaction.

2.3 Model Description

The overall model architecture consists of four EGAT layers that yield compressed fingerprints of the reactant(s) and product(s), which are then used to predict the E_a using a feed-forward stack (Fig. 2). Even though there are four total EGAT layers, only two are distinct, with the first being unique and the last three sharing weights. The first is responsible for embedding the raw node and edge features in a higher dimensional space so it has different internal dimensions from the other three. The three-fold application of the second EGAT layer results in mixing of information

Table 1: Input Node Features in the EGAT and Chemprop Models

Features	**Chemprop**	**EGAT**
Atom Type	100x1 one-hot vector	Integer
Number of Bonds	6x1 one-hot vector	-
Charge	5x1 one-hot vector	Integer
Neighboring H	5x1 one-hot vector	Integer
Neighboring C	-	Integer
Neighboring N	-	Integer
Neighboring O	-	Integer
Distance to Nearest Reacting Atom	-	Float
Hybridization	5x1 one-hot vector	4x1 one-hot vector [1]
Aromaticity	One-hot Value	One-hot Value
Atomic Mass	Mass/100	Float
Chirality	4x1 one-hot vector	3x1 one-hot vector
Ring Atom	-	One-hot value

1. The bits correspond to whether the atom is s,sp,sp2,or sp3 hybridized. These should be expanded if applying the model beyond second row chemistry.

Table 2: Input Edge Features in the EGAT and Chemprop Models

Features	**Chemprop**	**EGAT**
Bond Type	4x1 one-hot vector	5x1 one-hot vector [1]
Conjugated	One-hot Value	One-hot Value
In Ring	One-hot Value	One-hot Value
Stereochemistry	4x1 one-hot vector	3x1 one-hot vector
Bond Change	-	4x1 one-hot vector[2]
Change in Bond Order	-	One-hot value

1. Four of the vector values determine the order of the bond, while the additional value determines whether the bond is aromatic.
2. One of the bits denotes whether there is a bond break, a second denotes whether a bond forms, a third denotes whether the bond order changes, a fourth denotes whether the bond is unchanged.

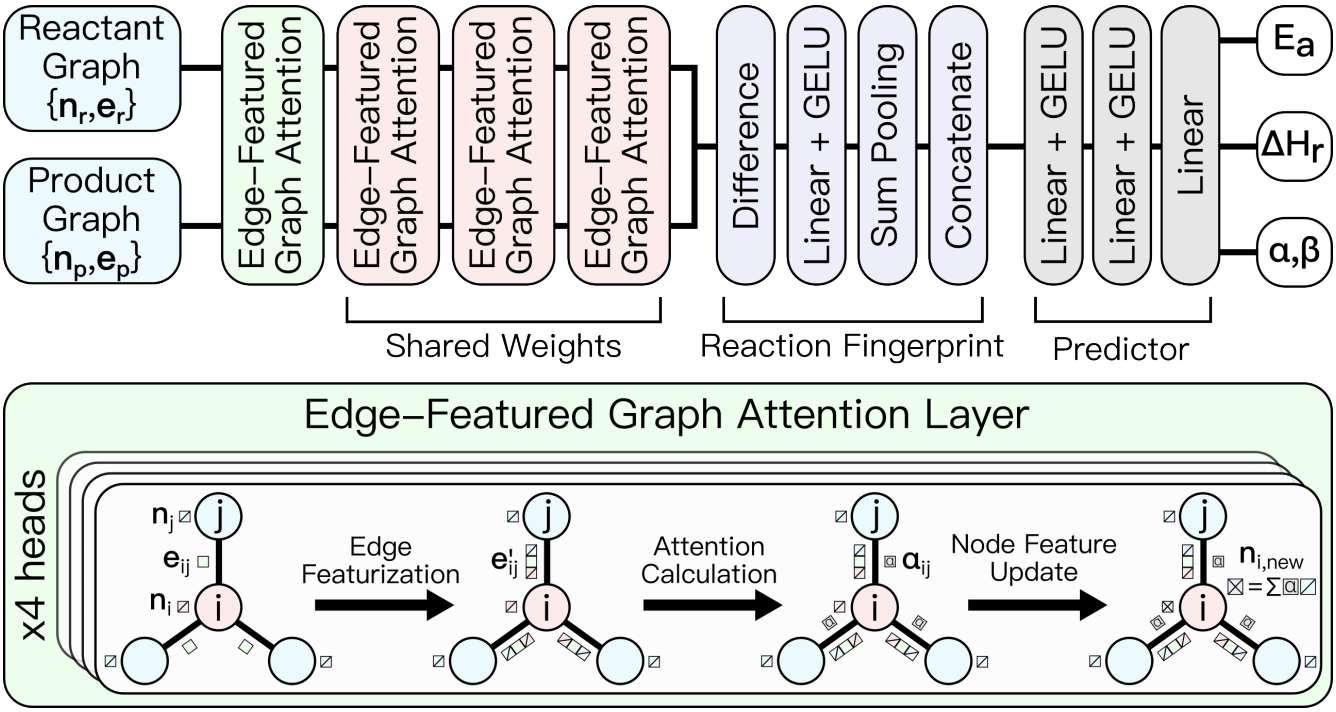

Figure 2: Overview of the EGAT model architecture developed in this work. The model consists of a stack of edge-featured graph attention layers (green and red) that generate the derived features for each edge and node in the reactant and product. The differences in these features are then transformed and pooled to generate a global fixed-size reaction fingerprint (purple) that is used as an input for the E_a prediction task (gray). Depending on the formulation of the learning task, the final predictor may be trained to solely predict E_a, E_a and ΔH_r in multi-task fashion, or α and β the linear free energy parameters associated with E_a. The lower inset illustrates the edge-featured graph attention mechanism applied to a single node and its edges.

from nodes separated by up to four edges away. Experiments on validation data showed no further benefit from using more EGAT layers or distinct weights in consecutive layers.

Within each EGAT layer, each edge is embedded based on both the edge and node features of the atoms connected by the edge. For an edge connecting nodes, i and j, the edge features, $\mathbf{e}_{ij}$, and connected node features, $\mathbf{n}_i$ and $\mathbf{n}_j$, are embedded as a concatenated vector

$$\mathbf{e}'_{ij} = \mathbf{A}\left(\mathbf{n}_i||\mathbf{e}_{ij}||\mathbf{n}_j\right) + \mathbf{a} \tag{1}$$

where $||$ refers to the catenation operation, $\mathbf{A}$ is a learnable linear transformation, and $\mathbf{a}$ is a bias term. In the first EGAT layer, $\mathbf{A}$ has dimensions of 512x48, where 48 is the size of the catenated inputted node and edge features (Tables 1 and 2) and 512 is the embedding dimension such that $\mathbf{e}'_{ij}$ has dimensions of 512x1. In subsequent EGAT layers, $\mathbf{A}$ has dimensions of 512x1536, where 1536 is the size of the catenated embedded node and edge features, and $\mathbf{e}'_{ij}$ has unchanged dimensions of 512x1. Recall that edges are defined based on the union of the bonds in the reactant and product graphs, such that an edge may exist in the reactant or product where no physical bond exists. The embedded edges are directional (i.e., $\mathbf{e}'_{ij} \neq \mathbf{e}'_{ji}$) and so there are twice as many embedded edges as connected nodes.

Each node is embedded into a size 512x1 space using a learnable linear projection of its node features according to

$$\mathbf{n}'_i = \mathbf{B}\mathbf{n}_i + \mathbf{b} \tag{2}$$

where $\mathbf{B}$ is a learnable linear transformation and $\mathbf{b}$ is a 512x1 bias vector. In the first EGAT layer, $\mathbf{B}$ has dimensions of 512x17, where 17 is the size of the inputted node features (Table 1) and 512 is the embedding dimension such that $\mathbf{n}'_i$ has dimensions of 512x1. In subsequent EGAT layers, $\mathbf{B}$ has dimensions of 512x512, where 512 is the size of the embedded node features after the first EGAT layer, and $\mathbf{n}'_i$ has unchanged dimensions of 512x1.

Each graph edge attention layer performs a non-linear transform and mixing operation between nodes connected by edges. The embedded edges, $\mathbf{e}'_{\mathrm{ij}}$, are transformed to create attention scores that determine the degree of mixing between nodes along each attention head. Here, each EGAT layer has four heads, meaning that the embedded 512x1 node feature vector, $\mathbf{n}'_{\mathrm{i}}$, is reshaped into four 128x1 vectors, $\mathbf{n}'^{,h}_{\mathrm{i}}$, each of which is mixed with the corresponding reshaped vectors of neighboring nodes, $\mathbf{n}'^{,h}_{\mathrm{j}}$, based on the attention values of each head, α^{h}_{ij}. The mathematical description of these steps is as follows.

The edge vectors are reshaped into a 128x1x4 tensor, $\mathbf{e}'^{,h}_{\mathrm{ij}}$, where each column vector corresponds to a query for each head in the attention mechanism. This matrix is non-linearly transformed using the LeakyReLU function and projected along learnable directions by each head according to

$$\epsilon_{ij} = \mathbf{C}\,\mathrm{LeakyReLU}\left(\mathbf{e}'^{,h}_{\mathrm{ij}}\right) \tag{3}$$

where $\mathbf{C}$ is a 1x128x4 learnable tensor, with each row vector corresponding to the key for each head in the attention mechanism, and ϵ_{ij} is a 1x1x4 tensor holding the logits for each head. The ϵ_{ij} values for each head are then softmax normalized over all edges originating at node i to obtain the attention scores

$$\alpha_{\mathrm{ij}} = \mathrm{softmax}\left(\epsilon_{\mathrm{ij}}\right) = \frac{e^{\epsilon_{\mathrm{ij}}}}{\sum_{k\in N(\mathrm{i})} e^{\epsilon_{\mathrm{ik}}}} \tag{4}$$

where $N(\mathrm{i})$ refers to all edges originating at node i, the operations are performed per-element, and α_{ij} is a 1x1x4 tensor. The softmax guarantees that $\sum_{k\in N(\mathrm{i})} \alpha_{\mathrm{ik}} = \mathbf{1}$ such that the attention scores associated with each head can be interpreted as mixing probabilities along each edge. Finally, the inputted 512x1 node feature vectors, $\mathbf{n}'_{\mathrm{i}}$, are reshaped into 128x1x4 tensors, $\mathbf{n}'^{,h}_{\mathrm{i}}$, where each row vector plays the role of a value in the attention mechanism. The updated node features are calculated as attention-weighted mixtures of the neighboring node features according to

$$\mathbf{n}_{\mathrm{i,out}}^{h} = \sum_{k \in N(\mathrm{i})} \alpha_{ik} \mathbf{n}_{k}^{\prime,h} \tag{5}$$

where $\mathbf{n}_{\mathrm{i,out}}^{h}$ is a 128x1x4 tensor that is reshaped to a 512x1 vector, $\mathbf{n}_{\mathrm{i,out}}$, before being returned by the layer. Each EGAT layer returns both the edge features, $\mathbf{e}_{\mathrm{ij}}^{\prime}$, and node features, $\mathbf{n}_{\mathrm{i,out}}$, for use by subsequent layers.

The nodes and edges of the reactant and product graphs are separately transformed by the EGAT layers to yield a set of 512x1 node and edge vectors for the reactant and product. The same stack of layers and weights are used for the reactant and product graphs. Reaction node and edge features are obtained based on the differences between these vectors,

$$\mathbf{n}_{\mathrm{i,rxn}} = \mathbf{n}_{\mathrm{i,product}} - \mathbf{n}_{\mathrm{i,reactant}} \tag{6}$$

$$\mathbf{e}_{\mathrm{ij,rxn}} = \mathbf{e}_{\mathrm{ij,product}} - \mathbf{e}_{\mathrm{ij,reactant}} \tag{7}$$

Where the product and reactant labels refer to the vector features outputted by the final EGAT layer. The reaction features are then subjected to a final projection and non-linear transform

$$\mathbf{n}_{\mathrm{i,rxn}}^{\prime} = \mathrm{GeLU}\left(\mathbf{D}\mathbf{n}_{\mathrm{i,rxn}} + \mathbf{d}\right) \tag{8}$$

$$\mathbf{e}_{\mathrm{ij,rxn}}^{\prime} = \mathrm{GeLU}\left(\mathbf{F}\mathbf{e}_{\mathrm{ij,rxn}} + \mathbf{f}\right) \tag{9}$$

where $\mathbf{D}$ and $\mathbf{F}$ are 512x512 learnable matrices and $\mathbf{d}$ and $\mathbf{f}$ are 512x1 bias vectors. At this stage there are a variable number of node and edge features depending on the size and topology of the reactant and product graphs. A fixed size 1024x1 reaction fingerprint, $\mathbf{FP}$, that is invariant to

the absolute atomic numbering is obtained sum pooling the reaction node and edge features and concatenating the result

$$\mathbf{FP} = \sum_{\mathrm{i}} \mathbf{n}'_{\mathrm{i,rxn}} || \sum_{\mathrm{ij}} \mathbf{e}'_{\mathrm{ij,rxn}}. \tag{10}$$

FP is used as an input to a stack of three feed-forward layers, with hidden dimensions of [256,128,1], a GeLU activation function after the first and second layers, and the last linear layer mapping to one output in the simplest model. Several small variations of the **FP** predictor stack were explored, depending on the formulation of the learning task as described next.

2.4 Learning Tasks

Four models were developed using the shared architecture for generating reaction fingerprints described in the previous section but with minor variations in the predictor stack. The basic model predicts E_{a} as a scalar output of the model using only the reactant and product graphs. The **FP** predictor stack for this model has dimensions of [256,128,1], with the final layer predicting E_{a}. We refer to this model in the results as the graph to E_{a} model (G$\rightarrow E_{\mathrm{a}}$).

Three other models were developed that use ΔH_{r} as an additional input feature or as an additional prediction target. The first uses ΔH_{r} as an additional input to the **FP** predictor stack, so that it has dimensions of [257,128,1], with the final layer predicting E_{a}. We refer to this model in the results section as the graph and ΔH_{r} to E_{a} model (G,$\Delta H_{\mathrm{r}} \rightarrow E_{\mathrm{a}}$). The second uses ΔH_{r} as an indirect feature by having the model predict E_{a} using the Bell-Evans-Polanyi[64] (BEP) relationship

$$E_{\mathrm{a}} = \alpha \Delta H_{\mathrm{r}} + \beta \tag{11}$$

where α and β are constants predicted by the model. The predictor stack for this model has dimensions of [256,128,2], where α and β are predicted in the final layer and ΔH_{r} is used

to calculate E_{a}. The rationale for this physics-inspired formulation is that it might show better transferability due to the well established linear free-energy relationships expected for reactions sharing a common mechanism. We refer to this model in the results as the graph to BEP model (G→BEP).

The last model uses a multi-task predictor stack with dimensions of [256,128,2] that predicts both ΔH_{r} and E_{a} in the last layer. Whereas both of the previous models require ΔH_{r} to make a prediction, the rationale for multi-task training is that it indirectly informs the model of the underlying mechanistic information associated with ΔH_{r} without requiring it at the time of prediction. This is a common form of transfer learning. We refer to this model in the results as the graph to ΔH_{r} and E_{a} model (G→ $\Delta H_{\mathrm{r}}, E_{\mathrm{a}}$).

2.5 Training Details

The loss function for training the single-task models (G→ E_{a}; G,ΔH_{r} → E_{a}; G→BEP) was the mean absolute error in E_{a} prediction

$$L = \frac{1}{N_{\mathrm{batch}}} \sum_{\mathrm{i}}^{N_{\mathrm{batch}}} |E_{\mathrm{a,i,0}} - E_{\mathrm{a,i,p}}| \tag{12}$$

where i runs over all samples in the batch, N_{batch} is the number of samples per batch, $E_{\mathrm{a,i,0}}$ refers to the reference activation energy, and $E_{\mathrm{a,i,p}}$ refers to the predicted activation energy. For the multi-task model (G→ $\Delta H_{\mathrm{r}}, E\mathrm{a}$) trained to predict both ΔH_{r} and E_{a}, the loss function was weighted to prioritize E_{a} accuracy according to

$$L = \frac{1}{N_{\mathrm{batch}}} \sum_{\mathrm{i}}^{N_{\mathrm{batch}}} 0.8|E_{\mathrm{a,i,0}} - E_{\mathrm{a,i,p}}| + 0.2|\Delta H_{\mathrm{r,i,0}} - \Delta H_{\mathrm{r,i,p}}| \tag{13}$$

where $\Delta H_{\mathrm{r,i,0}}$ refers to the reference heat of reaction, $\Delta H_{\mathrm{r,i,p}}$ refers to the predicted heat of reaction, and all other symbols have the same meaning as in Eq. 12.

All EGAT models were trained using the Adam optimizer and a batch size of 50. The learning rate, η, was initially set to 5e-4, linearly increased each update step for 10 epochs to 1e-3, followed by an exponential decay to a minimum of 1e-5. Early stopping was applied to terminate training if the validation loss did not decrease in 30 consecutive epochs. Further information on the hyperparameters for each model can be found in the Supplemental Information. All Chemprop models were trained using the optimized parameters found via the hyperopt package listed in the Supplemental Information (SI) section of the Heid et. al.[46] paper.

2.6 Data

The Chemprop and EGAT models used in this paper were trained on the RGD1 dataset,[41] which contains around 177,000 reactions with up to ten heavy atoms consisting of carbon, hydrogen, nitrogen and oxygen. In brief, the RGD1 dataset was generated by a graph-based enumeration of ∼700,000 reactions involving reactants sampled from PubChem.[20,65] A reaction conformational sampling strategy[20,58] was applied to generate up to three conformations for each reaction that were used to initialize double-ended TS searches,[29] followed by Berny optimization,[66] and intrinsic reaction coordinate[67] (IRC) validation at the GFN2-xTB[68] level of theory. The GFN2-xTB optimized TSs were further refined using the Gaussian16[69] quantum chemistry engine at the B3LYP-D3/TZVP[70,71] level of theory with D3 dispersion.[72] The DFT-level TS were classified as intended or unintended using an XGBoost[73] model that uses geometric features of the TS and the GFN2-xTB level information to classify the TS as intended or unintended (i.e., whether the TS corresponds to the reaction that was used to seed the TS search). This model exhibits a testing set accuracy of $\sim 95\%$ in a previous smaller testing set.[20] For a detailed description of the RGD1 database, we direct readers to our previous publication.[41]

Several data processing steps were applied to prepare the RGD1 data for training the activation energy models. Firstly, reactions were pruned by the uniqueness of the reactant and product

InChIKeys[74] to eliminate the presence of cases where reactions may have the same reactant and product but only differ by tautomerization. This was done to prevent possible data leakage by training on one tautomer while including another in the testing set. Secondly, only the minimum activation energy and its corresponding heat of reaction was listed for each reaction. This step is required because RGD1 contains multiple conformationally distinct TSs for many of the reactions. In total 135,455 distinct reactions passed these filters. Random splitting was used to generate 80/10/10 training (101,236 reactions), validation (12,139), and testing splits (12,139). The splits included only one direction of each reaction (i.e., the data was not augmented by including the reverse versions of reactions in training or testing). Unless otherwise stated, all accuracies are reported for prediction accuracy on the RGD1 testing set.

3 Results and Discussion

3.1 Overall RGD1 Performance

The overall performance of the various EGAT architectures was tested by predicting the activation energy of the 12.3k reactions in the RGD1 testing split (Fig. 3A). There is remarkably little variation across the different EGAT architectures; all models show uniformly low mean absolute prediction errors (MAE) of ∼4 kcal/mol. None of the models show a significant systematic bias based on the negligible mean signed errors (Fig. S1A). The G→ E_a model that doesn't use the ΔH_r information intuitively shows the largest MAE, but only performs ∼0.5 kcal/mol worse on average than G→ BEP, the best EGAT model. The models were tested as ensembles of five independently trained models of each class. The standard deviation in the mean testing split performance of the individual models comprising the ensemble provides an estimate of the testing performance uncertainty (error bars in Fig. 3A). The mean performance uncertainty is also within

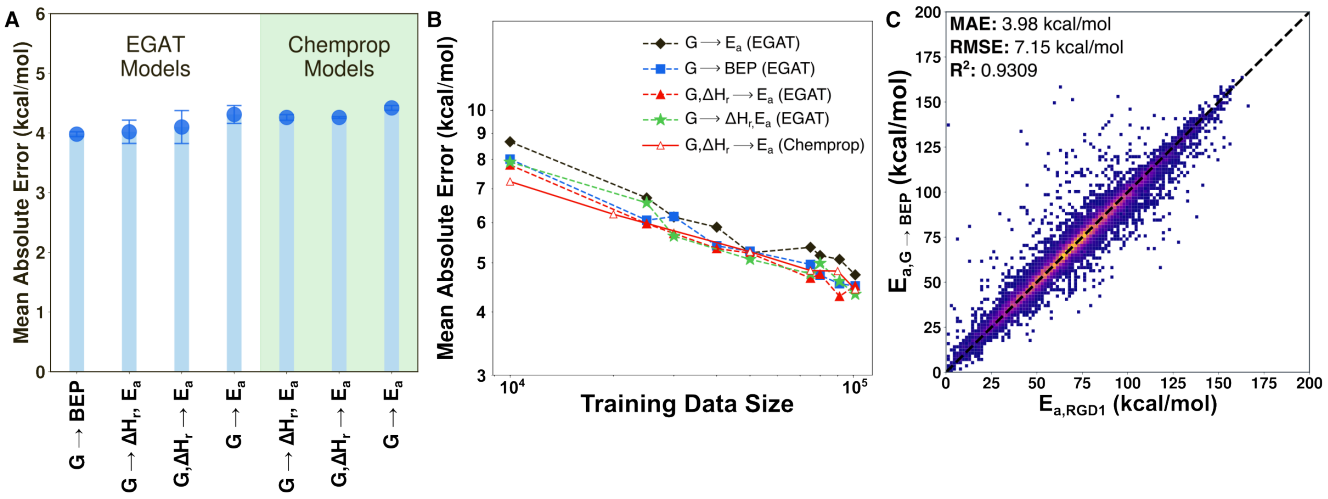

Figure 3: Performance of the G2Ea architectures on the RGD1 dataset. (A) Mean performance of each EGAT and Chemprop architecture on the RGD1 testing set ranked by accuracy. Error bars correspond to the standard deviation in mean performance across the ensemble. Each datapoint reflects the mean over five models trained with independent starting weights but the same training data. (B) Testing set error versus training dataset size for the EGAT architectures and best overall Chemprop model. Linear scaling is expected on a log-log plot. Each datapoint represents the best performance of a single model. (C) Parity plot showing the performance of best EGAT model (G→BEP) on the testing set reactions. Individual pixels are colored by the density of datapoints (purple to yellow is low to high).

∼0.5 kcal/mol in all cases.

The performance of analogous Chemprop models trained and tested on RGD1 provide a useful reference (Fig. 3A). Similar to the EGAT architectures, the Chemprop model that eschews ΔH_{r} information performs the worst on average, but only by ∼0.5 kcal/mol compared with the multitask Chemprop architecture (G→ $\Delta H_{\mathrm{r}}, E\mathrm{a}$). The best EGAT architecture marginally outperforms the best Chemprop architecture in this comparison, but we do not consider this difference significant for reasons that are further explored below.

How should we interpret all of the architectures approaching a MAE of ∼4 kcal/mol on the RGD1 testing split? To answer this the reader should consider that there are several known sources of irreducible error in RGD1—indeed to some extent all available computational TS datasets suffer from these issues. These sources include the inaccuracy of the underlying DFT method (B3LYP-D3/TZVP), incomplete conformational sampling of TSs and reactants, and the potential inclusion of TSs that correspond to unintended reactions (i.e., they connect a reactant and product that are different than the label). Because we are testing on DFT-level E_{a} values and not experimental values, the absolute DFT errors should not contribute to the irreducible error for this particular learning task; but incomplete conformational sampling and mislabeled TSs still represent sources of irreducible error. Below we will provide some lower bound estimates of the latter errors, but cumulatively they are expected to easily amount to an uncertainty of 4 kcal/mol across the entire dataset. As such, we interpret the prediction errors of all the models as effectively approaching the irreducible error of this prediction task.

If the models are approaching the irreducible error of this particular learning task, then that should be evident in the known reducible sources of error, such as dataset size, approaching zero. To test this, the training data size versus testing set accuracy learning curves were generated by training individual models for each architecture on subsets of the training data and evaluating their performance (Fig. 3B). Between training sample sizes of 10k and 40k, the models show

large error reductions (MAE reductions from 8 kcal/mol to 5.5 kcal/mol in the largest cases), but for sizes between 75k and 100k, the models show average improvements of less than 1 kcal/mol. Extrapolating the power law scaling evident in these curves, we estimate that increasing training data from order 100k to 1m samples would only further reduce errors by $\sim$ 0.1 kcal/mol for the best EGAT and Chemprop architectures. The minimal errors associated with training data size are consistent with the interpretation that the models are approaching the prediction accuracy limit associated with the irreducible error of this task.

Only looking at the mean performance obscures the number of outliers predicted by all the models. The presence of outliers are already evidenced by the significantly lower median absolute errors—between 2-3 kcal/mol for all EGAT and chemprop models—compared with the MAEs (Fig S1A). To illustrate some individual outliers, a parity plot is presented that shows the per sample accuracy of the G$\rightarrow$ BEP model on the testing set (Fig. 3B). The median absolute deviation (MAD = median $(|X_{\mathrm{i}} - \mathrm{median}\,(X_{\mathrm{i}})\,|)$) has been used as a robust estimator for the width of the error distribution that is minimally affected by outliers. Assuming normal statistics the corresponding estimate for the standard deviation in the testing errors for this model is 3.40 kcal/mol. The estimated standard deviation in prediction errors is similar when calculated for the underlying reaction classes (Fig. S4A), reactions with distinct molecularity (Fig. S4B), and across model architectures. The interested reader can also find violin plots of the MAEs of different populations of the testing set reactions differentiated by reaction type (Fig S5c) and molecularity (Fig S5D) in the SI. As an estimate of the number of outliers, there are a total of 796 testing samples (out of 12,139 total, or $\sim$ 7%) with absolute errors greater than 10.2 kcal/mol (3x the standard deviation as estimated from the median absolute deviation). Based on the testing set size and assuming normal statistics only $\sim$ 14 samples (i.e., $\sim$0.1% of the testing split) would be expected with errors this large. An analogous calculation using the absolute percent errors rather than the absolute errors produces a large number of percentage based outliers that is driven mainly

by low barrier reactions and not necessarily unphysical predictions (Fig. S6), and these are not further analyzed here.

3.2 Sources of Error

Although the EGAT architectures achieve mean accuracies on RGD1 comparable to the best G2EA models previously published, we consider it useful spend some time elaborating the obstacles to further accuracy improvements (Fig. 4). First we analyzed the conservation of poorly predicted outliers across the different architectures by histogramming testing samples by their error percentile for each EGAT model and calculating the overlapping samples in each bin (Fig. 4A). This analysis reveals that the architectures tended to perform poorly on the same samples, whereas the other bins show overlaps that are more consistent with chance (the expected overlap is $\sim 1\%$ for this number of bins and models).

It is possible that the outliers represent unusually hard samples for all models, but the more likely scenario is that the outliers are cases affected by incomplete conformational sampling or with unintended transition states (i.e., the known sources of irreducible error in RGD1). To test this hypothesis, we took the 100 lowest error testing samples and 100 highest error testing samples based on the G$\rightarrow$ BEP model performance and re-investigated their transition states. For these 200 reactions from the RGD1 testing set, transition state searches were re-performed using 10 rather than 3 reaction conformers (i.e., the RGD1 curation protocol) and DFT-level IRC calculations were performed on the RGD1 TSs.

The additional conformational sampling was done to estimate the error associated with incomplete sampling by calculating the average reduction in E_{a} for the top 100 and bottom 100 samples (Fig. 4B). For example, if the additional conformations merely rediscovered the same TS or higher energy TSs, then the reduction would be zero, but if they led to the discovery of lower barrier TSs, then the reduction would be positive. This experiment revealed that incomplete conformational

sampling disproportionately affects the poorly predicted samples, with the most and least accurate populations showing mean E_a reductions of ∼2.5 kcal/mol and ∼6.5 kcal/mol, respectively. The reaction with the largest E_a reduction of ∼30 kcal/mol occurs in the poorly predicted samples versus a maximum reduction of ∼13 kcal/mol in the top 100 predictions (See Fig. S6 for other relevant analyses). For the bottom 100 predictions that were intended after the IRC calculation (i.e., the 27 reactions where unintended TSs are not a confounding factor as revealed by the analysis in the next paragraph) the G→BEP model underestimates E_a by 21.8 kcal/mol on average compared with a negligible mean signed error on the whole dataset (Fig. S1A), meaning that the model recognizes these reactions as being conformationally undersampled. The E_a reduction of ∼2.5 kcal/mol observed in the best predictions can be considered a lower bound on the irreducible error of RGD1, assuming no other factors contribute to the error and that unintended reactions can be perfectly filtered. Thus while conformational sampling errors significantly contribute to the irreducible error for G2Ea prediction on RGD1, alone they cannot explain the majority of outliers that comprise the worst predictions.

IRC calculations on the top 100 and bottom 100 samples were performed to investigate the prevalence of unintended reactions in the two populations (Fig. 4C). Recall that IRC calculations are expensive at the DFT level and so RGD1 was curated using a machine learning model to filter intended and unintended TSs based on IRC calculations performed at the GFN2-xTB level.[41] Despite the high accuracy of this model, presumably some unintended reactions remain in the dataset. This experiment revealed that unintended reactions are disproportionately represented in the bottom 100 predictions (73%) compared with the best predictions (1%). This suggests that a majority of the outliers in Fig. 3C are in fact unintended reactions. This result motivated us to perform a larger IRC study on the 1000 worst predicted samples from the training split, which returned a similarly high proportion (63.8%) of unintended reactions (Figs. S8). Based on these tests, unintended reactions likely comprise only a few percent of RGD1, but they are the cause of

most poorly predicted outliers. Recalculating the MAE without the three sigma outliers results in an improvement of $\sim 1-2$ kcal/mol across the models, which can be taken as an estimate of the irreducible error associated with imperfect filtering of unintended reactions from RGD1.

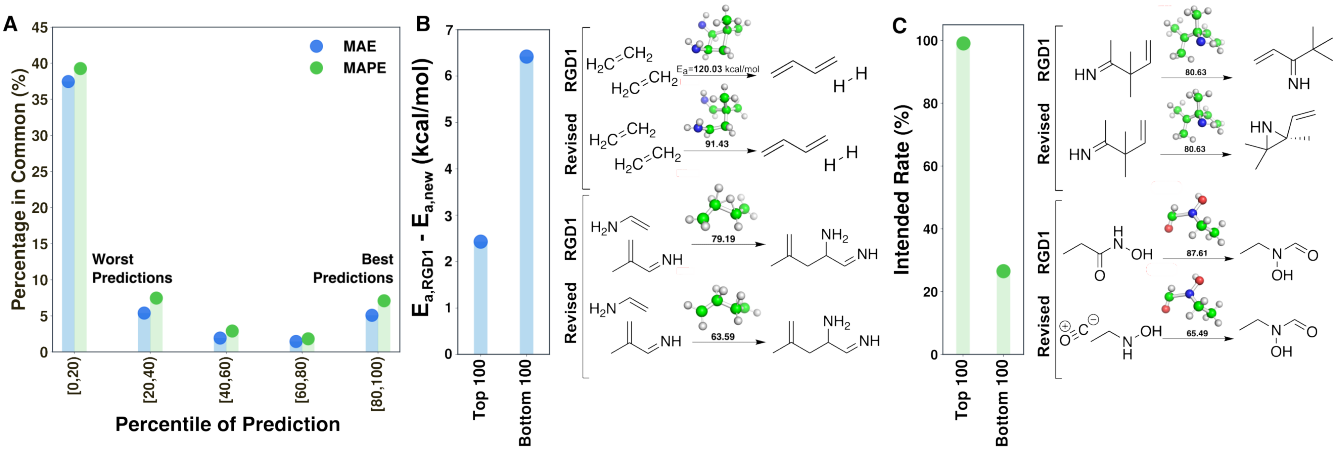

Figure 4: Investigating the origin of outlier behavior in G2Ea models. (A) Comparison of the fraction of testing set reactions that are consistently predicted accurately and inaccurately by each EGAT model. The rank of each reaction was determined by accuracy, binned by performance pentile, then the membership in each pentile was compared across models (percentage in common). (B) The mean E_a reduction after expanding the conformational sampling for the 100 best and 100 worst predicted testing set reactions by the G→BEP model. Two examples showing newly discovered TSs with lower barriers (revised) are shown. (C) The intended rates determined by IRC calculation for the RGD1 TSs of the 100 best and 100 worst predicted testing set reactions by the G→BEP model. Two illustrative examples are shown where the TS connected a different product (top) and different reactant (bottom) after the IRC calculation (revised). In the latter case, E_a is also incorrect because of the misidentified reactant.

3.2.1 Model Transferability

Graph-based neural-network models have become notorious in many contexts for overfitting and poor out-of-distribution performance.[44,56,57] Although the models trained on RGD1 show excellent testing performance on unseen reactions, this is a large dataset and reactions typically involve a small number of bond changes and conserved mechanisms. This means that even if the testing set involves unseen reactions in terms of reactants or products, it is not expected to necessarily

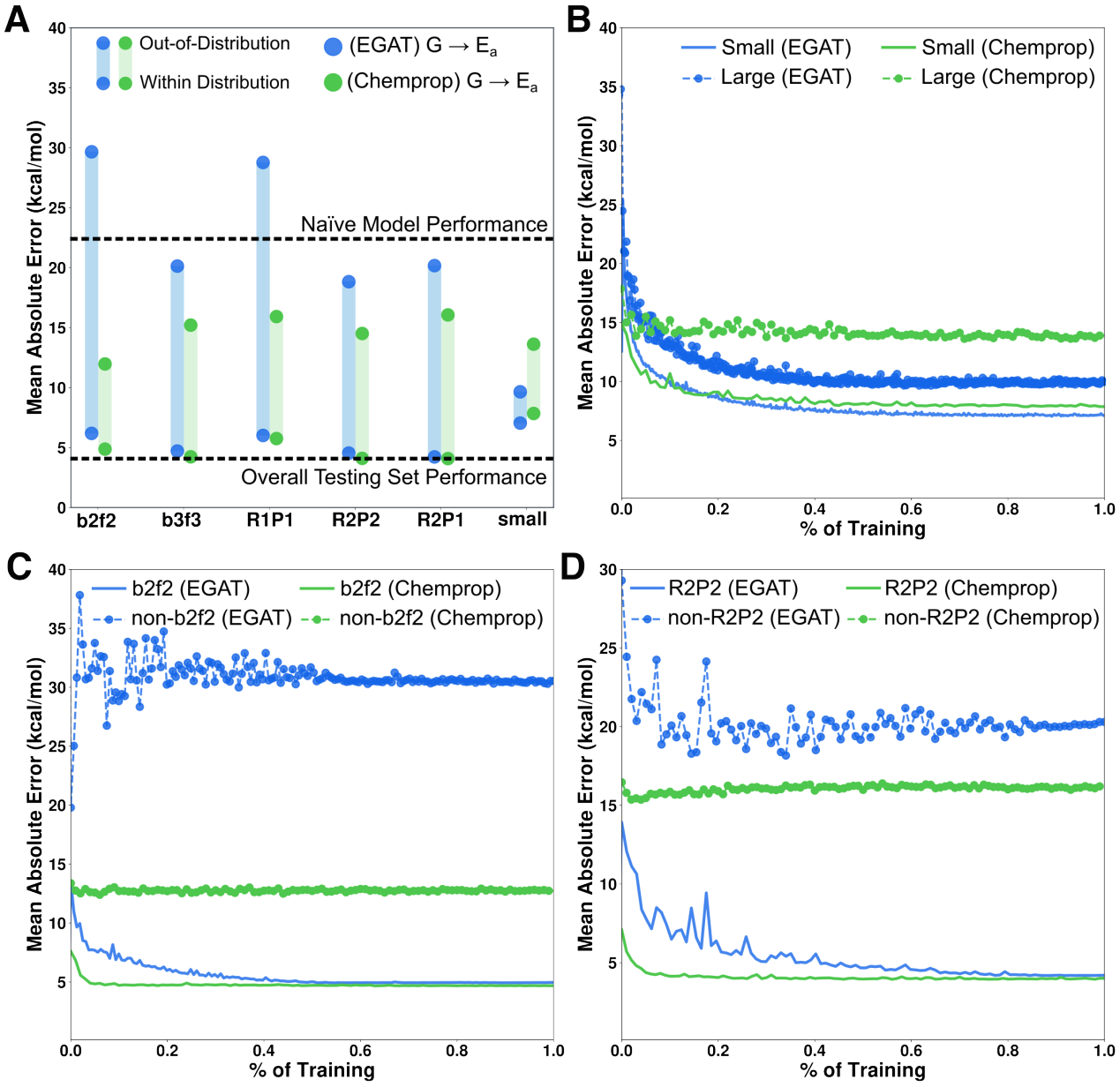

Figure 5: Transferability case-studies for G→ E_a models. (A) Mean performance of models trained on subsets of the RGD1 dataset indicated by the x-axis label. These models were tested on the RGD1 testing set reactions with the same reaction type as in training (within distribution) and reactions types excluded from training (out-of-distribution). (B) Learning curves for EGAT and Chemprop G→ E_a models trained on small reaction (3-7 heavy atoms) and evaluated on both small and large (8-10 heavy atoms) reactions from the validation split.(C) Learning curves for EGAT and Chemprop G→ E_a models trained on b2f2 data and evaluated on both b2f2 and non-b2f2 reactions from the validation split. (D) Learning curves for EGAT and Chemprop G→ E_a models trained on R2P2 data and evaluated on both R2P2 and non-R2P2 reactions from the validation split.

present novel reactivity (e.g., in terms of new types of bonds being broken and formed) that is not seen elsewhere in the training data.

To interrogate these architectures with a more rigorous test of out-of-distribution performance, several case studies were performed where the EGAT and Chemprop architectures were trained on subsets of the training data split that excluded one or more classes of reactions (Fig. 5). Three factors were used to classify different subpopulations within RGD1—the number of bonds broken and formed in the reaction (bnfn), the molecularity (number of species) of the reactant and product (RnPn), and the size of the reactant (more than seven heavy atoms was considered large). The manner in which RGD1 was generated resulted in reactions involving breaking two bonds and forming two bonds (b2f2) being the most abundant class of reactions and b3f3 reactions being the second most abundant with 22,538 and 68,615 reactions in the training dataset, respectively. In terms of molecularity, R1P1 (i.e., unimolecular) and R2P2 (i.e., bimolecular reactant and bimolecular product) reactions are the most abundant with 38,240 and 16,817 reactions in the training dataset, respectively. Similarly, the heavy atom cutoff was selected to make the small and large subpopulations approximately evenly matched with 40,018 and 57,746 training set reactions, respectively. In these case studies, the models were trained on one subpopulation of the training split but tested on the original testing split. The performance on the class of reactions included in training is referred to as "within distribution" and the class of reactions excluded from training is referred to as "out-of-distribution" in the discussed comparisons. All performance is reported for the testing set of reactions unless indicated otherwise.

The gap between the within distribution and out-of-distribution performance is large in all of the cases studies (Fig. 5A). The accuracy drop for within distribution performance is small in most cases and generally mirrors the reduced amount of training data available. In contrast, the accuracy drop for the out-of-distribution reaction classes is much larger than can be explained from reduced training data and reflects the qualitative failure of these architectures to predict E_a for

reactions that differ from the training distribution in either the number and kind of bonds changes, molecularity, or size. We had hypothesized that the models would extrapolate better across some of these classes than others. This was incorrect—the transferability to unseen reaction populations is consistently poor, with accuracy often worse than the naïve mean-predicting model. We had also hypothesized that the multitask or transfer learning architectures might show better transferability, but this also wasn't the case (see Fig. S8 for the combined results with all architectures). All the EGAT architectures perform poorly in the out-of-distribution prediction tasks with only a marginal benefit from the architectures with access to ΔH_{r}. With the exception of large reactants, Chemprop generally shows significantly lower errors than EGAT in out-of-distribution tasks. But neither architecture achieves out-of-distribution performance that could be relied on for applications.

The learning curves for these case studies also illustrate the limited transferability of the learned reaction representation to out-of-distribution prediction (Figs. 5B-D). These learning curves show the performance of the G$\rightarrow E_{\mathrm{a}}$ EGAT and Chemprop architectures on the validation split as they are trained. A normalized x-axis is used because the Chemprop and EGAT architectures train over differing numbers of epochs. The within distribution validation curves look typical, with an asymptotic approach to a saturation value that is near the testing set performance shown in Figure 5A. However, the out-of-distribution samples in the validation split show remarkably little reduction in error throughout the training. In some of the cases, a marginal initial reduction occurs that can be interpreted as the models learning generic reaction fingerprint features relevant to all reactions, but these curves still plateau at much larger values. Additional case studies were performed where models that were pretrained on one reaction class were then retrained on an excluded reaction class (Fig. S9). In all cases, these models show catastrophic forgetfullness after a single training epoch in predicting validation reactions from their original training class.

These case studies highlight the extremely limited transferability of these models to classes of reactions that are nevertheless expected to share many essential mechanistic features with the

training reactions. For example, many reaction mechanisms are expected to be conserved regardless of molecularity. Similarly, many small and large molecules should exhibit comparable activation energies for reactions with conserved mechanisms. The lack of transferability in these cases studies speaks to a gap in the current architectures. Based on the class-based reaction modeling of experts, this gap should be addressable, but it will require better regularization or additional architectural developments to promote mechanistic transferability.

4 Conclusions

This study has revisited the chemical graph to activation energy (G2Ea) prediction problem using an edge-featured graph attention (EGAT) architecture. This was motivated by the recent development of the relatively large ($\sim$ 176k) RGD1 organic chemistry reaction dataset, which enabled benchmarking across a broad swathe of reaction types and interrogation of out-of-distribution performance using case-studies with partial training on subsets of RGD1. Another motivation was to contribute an additional open-source architecture for other research groups to experiment with on their own. To the credit of the Chemprop developers, they have opened up their code for adaptation and comparison. Expanding the pool of publicly available models is critical to resolve where fundamental obstacles exist. The summary observations from these experiments are that it is relatively straightforward for G2Ea models to approach the irreducible error of RGD1—estimated to be $\sim$ 4 kcal/mol—but that out-of-distribution performance is often worse than naïve models.

Sources of irreducible error for the G2Ea prediction task warrant more detailed consideration moving forward. Incomplete conformational sampling results in the curation of TSs that are not the lowest barrier possible, and thus out-of-distribution for training and prediction. Likewise, unintended TSs (i.e., true saddle points that nonetheless correspond to reactions that are different from the labeled reactants and products) are commonly produced by TS search algorithms. Due to

the incorrect input features, unintended samples are intrinsically unpredictable by G2Ea models. Both sources of irreducible error were observed in the RGD1 dataset and are expected to be found in any other computational dataset of conformationally complex reactants without explicit protocols for mitigation. Although these errors are "irreducible" from the perspective of performance in the G2Ea task for a given dataset, they are reducible from the perspective of curating better datasets. For example, there are many possibilities for developing better filters for unintended transition states—from more complex models to using more informative featurizations for the unintended/intended classification task—but they have yet to be implemented. We estimated that the majority of prediction outliers in the current study were in fact unintended reactions and that perfect filtering would reduce the irreducible error for the G2Ea task by $\sim 1-2$ kcal/mol with the remaining $\sim 2-3$ kcal/mol being accounted for by incomplete conformational sampling.

The discussion of errors also bears on the relative advantage of formulating E_{a} prediction as a G2Ea task. For example, unintended transition states and biased conformational sampling are not intrinsic problems for some three-dimensional formulations of E_{a} prediction. For example, a model with a three dimensional reactant featurization could plausibly learn the contribution to E_{a} associated with a particular conformer and thus reduce that source of error. Or consider an E_{a} model that performs E_{a} estimation identically to quantum chemical approaches but instead uses ML atomic potentials.[75] Because such a model is formulated to predict reactive potential energy surfaces, it can learn from both intended and unintended TSs and all reactant and product conformations. This formulation may sound much more expensive than directly predicting E_{a} from a pair of chemical graphs, but the advent of GPU-compatible routines for performing TS searches on ML-potentials may render such cost differences moot. These differing formulations of the E_{a} prediction problem can also be expected to affect model tranferability. Better benchmarks will ultimately be required to resolve the accuracy vs cost Pareto front of different E_{a} prediction formulations being gestured toward by this discussion.

Within the context of G2Ea models, these results also suggest several pathways for improvement. First, the fact that the overall accuracy of all models uniformly approached the irreducible error of RGD1 shows that these models have spare complexity to learn broader classes of reactions. RGD1 only contains closed-shell neutral reactants containing CHON elements, but extensions to ionic, radical, and other elements are expected to be successful. Second, the current featurization does not directly consider the reaction conditions (e.g., solvent) or the availability of a catalyst. It is plausible that a condition fingerprint could be catenated to the reaction fingerprint to predict how E_a would be modulated by environment, but to our knowledge the requisite data for a convincing attempt at this does not yet exist. Lastly, our current experiments found marginal advantage from using complementary information sources like ΔH_r during training and prediction in the large data limit. More advantages might become apparent for datasets possessing a broader range of reaction mechanisms, datasets with lower irreducible error for the G2Ea task, or by using auxiliary information sources beyond ΔH_r. These and adjacent opportunities suggest that the field is far from determining the ultimate performance of G2Ea models.

References

(1) Demirbas, A.; Arin, G. An Overview of Biomass Pyrolysis. *Energy Sources* **2002**, *24*, 471–482.

(2) Akbar Ali, M.; Violi, A. Reaction Pathways for the Thermal Decomposition of Methyl Butanoate. *J. Org. Chem.* **2013**, *78*, 5898–5908.

(3) Kang, P.-L.; Shang, C.; Liu, Z.-P. Glucose to 5-Hydroxymethylfurfural: Origin of Site-Selectivity Resolved by Machine Learning Based Reaction Sampling. *J. Am. Chem. Soc.* **2019**, *141*, 20525–20536.

(4) Mohan, D.; Pittman Jr, C. U.; Steele, P. H. Pyrolysis of wood/biomass for bio-oil: A critical review. *Energy Fuels* **2006**, *20*, 848–889.

(5) Corma, A.; Iborra, S.; Velty, A. Chemical routes for the transformation of biomass into chemicals. *Chem. Rev.* **2007**, *107*, 2411–2502.

(6) Zhao, Q.; Savoie, B. M. Deep reaction network exploration of glucose pyrolysis. *Proc. Natl. Acad. Sci.* **2023**, *120*, e2305884120.

(7) Reymond, J.-L.; Awale, M. Exploring Chemical Space for Drug Discovery Using the Chemical Universe Database. *ACS Chem. Neurosci.* **2012**, *3*, 649–657.

(8) Virshup, A. M.; Contreras-García, J.; Wipf, P.; Yang, W.; Beratan, D. N. Stochastic Voyages into Uncharted Chemical Space Produce a Representative Library of All Possible Drug-Like Compounds. *J. Am. Chem. Soc.* **2013**, *135*, 7296–7303.

(9) Blakemore, D. C.; Castro, L.; Churcher, I.; Rees, D. C.; Thomas, A. W.; Wilson, D. M.; Wood, A. Organic synthesis provides opportunities to transform drug discovery. *Nat. Chem.* **2018**, *10*, 383–394.

(10) Montgomery, J. High-Throughput Discovery of New Chemical Reactions. *Science* **2011**, *333*, 1387–1388.

(11) Heiskanen, S. K.; Kim, J.; Lucht, B. L. Generation and Evolution of the Solid Electrolyte Interphase of Lithium-Ion Batteries. *Joule* **2019**, *3*, 2322–2333.

(12) Takenaka, N.; Suzuki, Y.; Sakai, H.; Nagaoka, M. On Electrolyte-Dependent Formation of Solid Electrolyte Interphase Film in Lithium-Ion Batteries: Strong Sensitivity to Small Structural Difference of Electrolyte Molecules. *J. Phys. Chem. C* **2014**, *118*, 10874–10882.

(13) Chamas, A.; Moon, H.; Zheng, J.; Qiu, Y.; Tabassum, T.; Jang, J. H.; Abu-Omar, M.; Scott, S. L.; Suh, S. Degradation Rates of Plastics in the Environment. *ACS Sustain. Chem. Eng.* **2020**, *8*, 3494–3511.

(14) Huang, Z.; Chen, M. S.; Woroch, C. P.; Markland, T. E.; Kanan, M. W. A framework for automated structure elucidation from routine NMR spectra. *Chem. Sci.* **2021**, *12*, 15329–15338.

(15) Firet, N. J.; Smith, W. A. Probing the Reaction Mechanism of CO_2 Electroreduction over Ag Films via Operando Infrared Spectroscopy. *ACS Catal.* **2017**, *7*, 606–612.

(16) Lansford, J. L.; Vlachos, D. G. Infrared spectroscopy data- and physics-driven machine learning for characterizing surface microstructure of complex materials. *Nat. Commun.* **2020**, *11*, 1513.

(17) Fine, J.; Kuan-Yu Liu, J.; Beck, A.; Alzarieni, K. Z.; Ma, X.; Boulos, V. M.; Kenttämaa, H. I.; Chopra, G. Graph-based machine learning interprets and predicts diagnostic isomer-selective ion–molecule reactions in tandem mass spectrometry. *Chem. Sci.* **2020**, *11*, 11849–11858.

(18) Fang, Y.; Li, J.; Chen, Y.; Lu, Q.; Yang, H.; Wang, X.; Chen, H. Experiment and Modeling Study of Glucose Pyrolysis: Formation of 3-Hydroxy-γ-butyrolactone and 3-(2H)-Furanone. *Energy Fuels* **2018**, *32*, 9519–9529.

(19) Zhao, Q.; Savoie, B. M. Simultaneously improving reaction coverage and computational cost in automated reaction prediction tasks. *Nat. Comput. Sci.* **2021**, *1*, 479–490.

(20) Zhao, Q.; Savoie, B. M. Algorithmic explorations of unimolecular and bimolecular reaction spaces. *Angew. Chem., Int. Ed.* **2022**, *61*, e202210693.

(21) Barter, D.; Clark Spotte-Smith, E. W.; Redkar, N. S.; Khanwale, A.; Dwaraknath, S.; Persson, K. A.; Blau, S. M. Predictive stochastic analysis of massive filter-based electrochemical reaction networks. *Digital Discovery* **2023**, *2*, 123–137.

(22) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine learning in computer-aided synthesis planning. *Acc. Chem. Res.* **2018**, *51*, 1281–1289.

(23) Suleimanov, Y. V.; Green, W. H. Automated Discovery of Elementary Chemical Reaction Steps using Freezing String and Berny Optimization Methods. *J. Chem. Theory Comput.* **2015**, *11*, 4248–4259.

(24) Zimmerman, P. M. Automated discovery of chemically reasonable elementary reaction steps. *J. Comput. Chem.* **2013**, *34*, 1385–1392.

(25) Vernuccio, S.; Broadbelt, L. J. Discerning complex reaction networks using automated generators. *AIChE J.* **2019**, *65*, e16663.

(26) Ásgeirsson, V.; Birgisson, B. O.; Bjornsson, R.; Becker, U.; Neese, F.; Riplinger, C.; Jónsson, H. Nudged Elastic Band Method for Molecular Reactions Using Energy-Weighted Springs Combined with Eigenvector Following. *J. Chem. Theory Comput.* **2021**, *17*, 4929–4945.

(27) Henkelman, G.; Uberuaga, B. P.; Jónsson, H. A Climbing Image Nudged Elastic Band Method for Finding Saddle Points and Minimum Energy Paths. *J. Chem. Phys.* **2000**, *113*, 9901–9904.

(28) Zimmerman, P. M. Growing string method with interpolation and optimization in internal coordinates: Method and examples. *J. Chem. Phys.* **2013**, *138*, 184102.

(29) Zimmerman, P. Reliable transition state searches integrated with the growing string method. *J. Chem. Theory Comput.* **2013**, *9*, 3043–3050.

(30) Maeda, S.; Harabuchi, Y.; Takagi, M.; Taketsugu, T.; Morokuma, K. Artificial force induced reaction (AFIR) method for exploring quantum chemical potential energy surfaces. *Chem. Rec.* **2016**, *16*, 2232–2248.

(31) Simm, G. N.; Reiher, M. Context-driven exploration of complex chemical reaction networks. *J. Chem. Theory Comput.* **2017**, *13*, 6108–6119.

(32) Abdelfatah, K.; Yang, W.; Vijay Solomon, R.; Rajbanshi, B.; Chowdhury, A.; Zare, M.; Kundu, S. K.; Yonge, A.; Heyden, A.; Terejanu, G. Prediction of Transition-State Energies of Hydrodeoxygenation Reactions on Transition-Metal Surfaces Based on Machine Learning. *J. Phys. Chem. C* **2019**, *123*, 29804–29810.

(33) Wang, S. et al. Universal transition state scaling relations for (de)hydrogenation over transition metals. *Phys. Chem. Chem. Phys.* **2011**, *13*, 20760–20765.

(34) Wang, S.; Vorotnikov, V.; Sutton, J. E.; Vlachos, D. G. Brønsted–Evans–Polanyi and Transition State Scaling Relations of Furan Derivatives on Pd(111) and Their Relation to Those of Small Molecules. *ACS Catal.* **2014**, *4*, 604–612.

(35) Van Santen, R. A.; Neurock, M.; Shetty, S. G. Reactivity Theory of Transition-Metal Surfaces: a Brønsted-Evans-Polanyi Linear Activation Energy-Free-Energy analysis. *Chem. Rev.* **2010**, *110*, 2005–2048.

(36) Bragato, M.; von Rudorff, G. F.; von Lilienfeld, O. A. Data enhanced Hammett-equation: reaction barriers in chemical space. *Chem. Sci.* **2020**, *11*, 11859–11868.

(37) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(38) Ismail, I.; Robertson, C.; Habershon, S. Successes and challenges in using machine-learned activation energies in kinetic simulations. *J. Chem. Phys.* **2022**, *157*, 014109.

(39) Li, X.; Zhang, S.-Q.; Xu, L.-C.; Hong, X. Predicting Regioselectivity in Radical C-H Functionalization of Heterocycles through Machine Learning. *Angew. Chem., Int. Ed.* **2020**, *59*, 13253–13259.

(40) Choi, S.; Kim, Y.; Kim, J. W.; Kim, Z.; Kim, W. Y. Feasibility of Activation Energy Prediction of Gas-Phase Reactions by Machine Learning. *Chem. Eur. J.* **2018**, *24*, 12354–12358.

(41) Zhao, Q.; Vaddadi, S. M.; Woulfe, M.; Ogunfowora, L. A.; Garimella, S. S.; Isayev, O.; Savoie, B. M. Comprehensive exploration of graphically defined reaction spaces. *Sci. Data* **2023**, *10*, 145.

(42) Rudorff, G. F.; Heinen, S. N.; Bragato, M.; Lilienfeld, O. A. Thousands of reactants and transition states for competing E2 and SN2 reactions. *Mach. Learn.: Sci. Technol* **2020**, *1*, 045026.

(43) Grambow, C. A.; Pattanaik, L.; Green, W. H. Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry. *Sci. Data* **2020**, *7*, 137.

(44) Spiekermann, K. A.; Pattanaik, L.; Green, W. H. Fast Predictions of Reaction Barrier Heights: Toward Coupled-Cluster Accuracy. *J. Phys. Chem. A* **2022**, *126*, 3976–3986.

(45) Grambow, C. A.; Pattanaik, L.; Green, W. H. Deep Learning of Activation Energies. *J. Phys. Chem. Lett.* **2020**, *11*, 2992–2997.

(46) Heid, E.; Green, W. H. Machine Learning of Reaction Properties via Learned Representations of the Condensed Graph of Reaction. *J. Chem. Inf. Model.* **2022**, *62*, 2101–2110.

(47) Heinen, S.; von Rudorff, G. F.; von Lilienfeld, O. A. Toward the Design of Chemical Reactions: Machine Learning Barriers of Competing Mechanisms in Reactant Space. *J. Chem. Phys.* **2021**, *155*, 064105.

(48) Heinen, S.; von Rudorff, G. F.; von Lilienfeld, O. A. Transition state search and geometry relaxation throughout chemical compound space with quantum machine learning. *J. Chem. Phys.* **2022**, *157*, 221102.

(49) Zhao, Q.; Anstine, D. M.; Isayev, O.; Savoie, B. M. Δ^2 machine learning for reaction property prediction. *Chem. Sci.* **2023**,

(50) Marques, E.; de Gendt, S.; Pourtois, G.; van Setten, M. J. Improving Accuracy and Transferability of Machine Learning Chemical Activation Energies by Adding Electronic Structure Information. *J. Chem. Inf. Model.* **2023**, *63*, 1454–1461.

(51) Friederich, P.; dos Passos Gomes, G.; De Bin, R.; Aspuru-Guzik, A.; Balcells, D. Machine learning dihydrogen activation in the chemical space surrounding Vaska's complex. *Chem. Sci.* **2020**, *11*, 4584–4601.

(52) Migliaro, I.; Cundari, T. R. Density Functional Study of Methane Activation by Frustrated Lewis Pairs with Group 13 Trihalides and Group 15 Pentahalides and a Machine Learning Analysis of Their Barrier Heights. *J. Chem. Inf. Model.* **2020**, *60*, 4958–4966.

(53) Mikami, K. Interactive-quantum-chemical-descriptors enabling accurate prediction of an activation energy through machine learning. *Polymer* **2020**, *203*, 122738.

(54) Döntgen, M.; Fenard, Y.; Heufer, K. A. Atomic Partial Charges as Descriptors for Barrier Heights. *J. Chem. Inf. Model.* **2020**, *60*, 5928–5931.

(55) Jorner, K.; Brinck, T.; Norrby, P.-O.; Buttar, D. Machine learning meets mechanistic modelling for accurate prediction of experimental activation energies. *Chem. Sci.* **2021**, *12*, 1163–1175.

(56) Gerwen, P. v.; Wodrich, M. D.; Laplaza, R.; Corminboeuf, C. Reply to Comment on 'Physics-based representations for machine learning properties of chemical reactions'. *Mach. learn.: Sci. Technol.* **2023**, *4*, 048002.

(57) Spiekermann, K. A.; Stuyver, T.; Pattanaik, L.; Green, W. H. Comment on 'Physics-based representations for machine learning properties of chemical reactions'. *Mach. learn.: Sci. Technol.* **2023**, *4*, 048001.

(58) Pracht, P.; Bohle, F.; Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169–7192.

(59) Zhao, Q.; Hsu, H.-H.; Savoie, B. Conformational Sampling for Transition State Searches on a Computational Budget. *J. Chem. Theory Comput.* **2022**, *18*, 3006–3016.

(60) Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. International Conference on Learning Representations. 2018.

(61) Kamiński, K.; Ludwiczak, J.; Jasiński, M.; Bukala, A.; Madaj, R.; Szczepaniak, K.; Dunin-Horkawicz, S. Rossmann-toolbox: a deep learning-based protocol for the prediction and design of cofactor specificity in Rossmann fold proteins. *Briefings Bioinf.* **2021**, *23*, bbab371.

(62) Wang, Z.; Chen, J.; Chen, H. EGAT: Edge-Featured Graph Attention Network. Artificial Neural Networks and Machine Learning – ICANN 2021. Cham, 2021; pp 253–264.

(63) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyz-

ing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388.

(64) Evans, M. G.; Polanyi, M. Inertia and driving force of chemical reactions. *Trans. Faraday Soc.* **1938**, *34*, 11–24.

(65) Kim, S. Pubchem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.* **2021**, *49*, D1388–D1395.

(66) Schlegel, H. B. Optimization of equilibrium geometries and transition structures. *J. Comput. Chem.* **1982**, *3*, 214–218.

(67) Maeda, S.; Harabuchi, Y.; Ono, Y.; Taketsugu, T.; Morokuma, K. Intrinsic reaction coordinate: Calculation, bifurcation, and automated search. *Int. J. Quantum Chem.* **2015**, *115*, 258–269.

(68) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.

(69) Frisch, M. J. et al. Gaussian 16 Revision C.01. **2016**, Gaussian Inc. Wallingford CT.

(70) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *J. Phys. Chem.* **1994**, *98*, 11623–11627.

(71) Schäfer, A.; Huber, C.; Ahlrichs, R. Fully optimized contracted Gaussian basis sets of triple zeta valence quality for atoms Li to Kr. *J. Chem. Phys* **1994**, *100*, 5829–5835.

(72) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys* **2010**, *132*, 154104.

(73) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016; pp 785–794.

(74) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. ChemInf.* **2015**, *7*, 1–34.

(75) Dylan Anstine, O. I., Roman Zabatyuk AIMNet2: A Neural Network Potential to Meet your Neutral, Charged, Organic, and Elemental-Organic Needs. *ChemRxiv* **2023**,

5 Data and Code Availability

The authors declare that the data supporting the findings of this study are available within the paper and its supplementary information files. Pretained EGAT models and corresponding training scripts are available on Github (XXX To be filled upon publication XXX). Training and testing set splits and all relevant data for the figures are detailed on Figshare (XXX To be filled upon publication XXX).

Acknowledgements

The work performed by S.M.V., Q.Z., and B.M.S was made possible by the Office of Naval Research (ONR) through support provided by the Energetic Materials Program (MURI grant number:

N00014-21-1-2476, Program Manager: Dr. Chad Stoltz). B.M.S also acknowledges partial support for this work from the Purdue Process Safety and Assurance Center.

Author contributions statement

Q.Z. and B.M.S. conceived and designed the study. Q.Z. developed the initial iterations of the model. Q.Z. and S.M.V. pre-processed the data into training, testing, and validation sets. S.M.V. performed the tests and analyzed the data. All authors prepared the manuscript.

Competing interests

The authors declare no competing interests.