# Supporting Information for *CoeffNet*: Predicting activation barriers through a chemically-interpretable, equivariant and physically constrained graph neural network

Sudarshan Vijay,[1,2] Maxwell C. Venetos,[1,2] Evan Walter Clark Spotte-Smith,[1,2] Aaron D. Kaplan,[2] Mingjian Wen,[3] and Kristin A. Persson[4,1,a)]

[1)] *Department of Materials Science and Engineering, University of California, Berkeley, 210 Hearst Memorial Mining Building, Berkeley, CA, 94720 USA*

[2)] *Materials Science Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA, 94720 USA*

[3)] *Department of Chemical and Biomolecular Engineering, University of Houston, Houston, Texas 77204, United States*

[4)] *The Molecular Foundry, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA, 94720 USA*

(Dated: 4 January 2024)

[a)] Electronic mail: kristinpersson@berkeley.edu

## S1. CONVERTING TO AND FROM ORTHONORMAL BASIS

In this section, we illustrate the steps needed to change the coefficient matrix from the non-orthonormal basis in which it was computed into an orthonormal basis (which has the properties listed Section II).

1. Diagonalise the overlap matrix, $\mathbf{S}^{\mathrm{DFT}}$ as $\mathbf{S}^{\mathrm{DFT}}\Lambda = \lambda\Lambda$
2. Determine $\mathbf{X} = \mathbf{S}^{-1/2} = \lambda\Lambda^{-1/2}\lambda$.
3. Compute the orthogonal Hamiltonian matrix, $\mathbf{F} = \mathbf{X}^{\intercal}\mathbf{F}^{\mathrm{DFT}}\mathbf{X}$
4. Determine the eigenvalues and eigenvectors of $\mathbf{F}$. The eigenvalues are identical to those obtained from DFT (i.e. in non-orthonormal basis). The eigenvectors are the coefficient matrix elements, $\mathbf{C}$, in orthonormal basis.

All operations are available within the class `coeffnet.predata.matrices.BaseMatrices`.

Molecular orbitals predicted by the model can be returned to the non-orthonormal basis in which the DFT calculation was performed by,

$$\mathbf{C}' = \mathbf{S}^{-1/2}\mathbf{C} \tag{1}$$

where $\mathbf{C}'$ is the coefficient matrix in non-orthonormal basis.

## S2. PRACTICAL ASPECTS OF D-MATRIX ROTATION

Different electronic structure codes order orbitals and coordinate axes in different ways. In this section, we describe the transformation needed to ensure that the orbitals computed with `Q-Chem` are consistent with those used as spherical harmonics in the neural network (and with `e3nn`, the Python package we use to compute tensor products).

### A. Modifications to the inputs of the DFT calculation

All calculations were run using the `pymatgen` / `atomate` / `fireworks` framework. To ensure that the $\mathbf{D}-$matrix of rotation is consistent with the axis of rotation of `e3nn`, we switch the $x$, $y$ and $z$ axes with $z$, $x$ and $y$ axes. For a given `pymatgen` molecule, the following modification needs to be made:

```
coordinates = np.array(molecule.cart_coords)
coordinates[:, [0, 1, 2]] = coordinates[:, [2, 0, 1]]

molecule = Molecule(
    species=molecule.species,
    coords=coordinates,
    charge=molecule.charge,
    spin_multiplicity=molecule.spin_multiplicity,
)
```

### B. Modifications to choice of basis set

In this work, we use basis sets where all basis functions are computed in spherical basis sets. All calculations were run using the `purecart=1111` option in `Q-Chem`.

## S3. MULTIPLE MOLECULAR ORBITALS

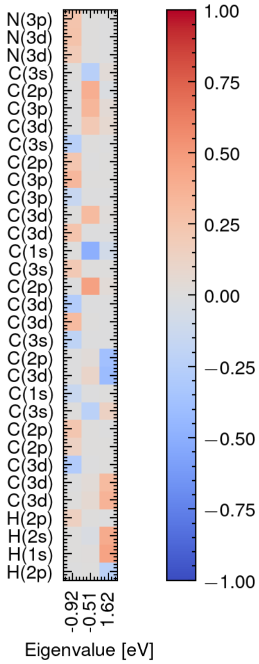
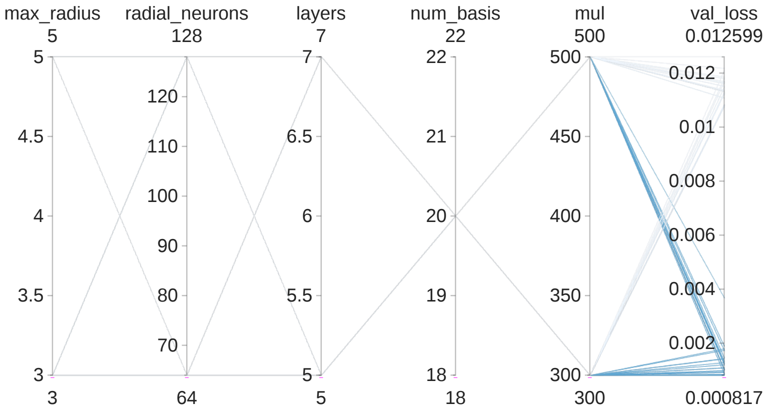


FIG. S1: Coefficients of selected molecular orbitals of pyridine

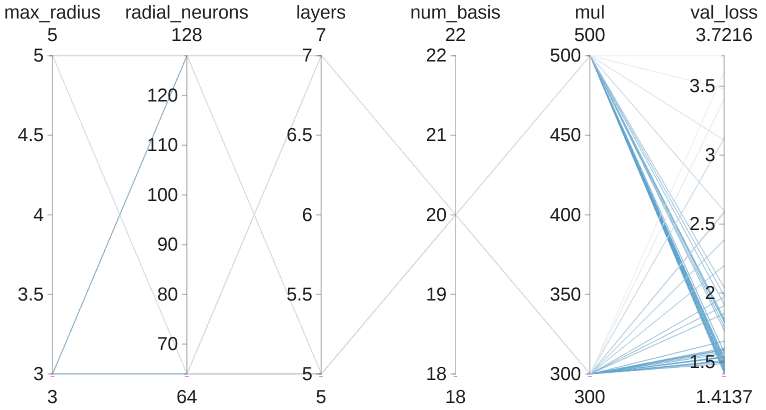
While the example of the molecular orbitals of water in Figure 1 illustrates the role that the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) has on reactivity, other frontier orbitals might be needed for more complex molecular systems. In this section, we illustrate the need for using more than one molecular orbital simultaneously for describing reactivity by studying the coefficients of molecular orbitals of pyridine, a ringed molecule with more than one reactive center.[1]

Figure S1 shows the coefficients of selected atomic orbitals for the HOMO, LUMO and one orbital lower than the HOMO (so called, HOMO-1) for pyridine. Using just the HOMO of pyridine in a model designed to predict reactive quantities might lead to the incorrect description that it is the carbon atoms that are predominantly important for reactivity. However, incorporating information from an eigenvalue narrowly lower in energy than the HOMO provides more information. HOMO-1 shows the lone-pairs of nitrogen as well as the atomic orbitals of (a different) carbon atom being relevant for reactivity. Since the presence of multiple molecular orbitals is relevant for a complete description, more complex systems may be treated with a model with allows for the inclusion of multiple molecular orbitals.

## S4. EFFECT OF PARAMETERS AND HYPERPARAMETERS



(a) Parallel coordinates plot for validation loss for predicting relative energies



(b) Parallel coordinates plot for validation loss for predicting coefficients of the HOMO.

FIG. S2: Tests for optimal hyperparameters

In this section, we discuss the effect that the hyperparameters have on the validation loss of the model for predicting activation barriers and coefficients of the molecular orbital of the HOMO. Figure S2 shows a parallel coordinates plot with the different tunable hyperparameters. *max_radius* is the maximum radius used to perform the convolution, *radial_nuerons* is the number of hidden-neurons for learning the the radial components, *layers* is the number of non-linearities used for the scalar components, *num_basis* is the number of basis functions used for the edge attributes and *mul* is the number of hidden basis functions for the same irreps as the node inputs.

From Figure S2(a,b) it is clear that the largest determination of a reduced validation loss comes from *mul*. This reduction in loss is expected with a greater number of spherical harmonics of each irrep in the hidden layer.

## S5. DATASET SPLITS

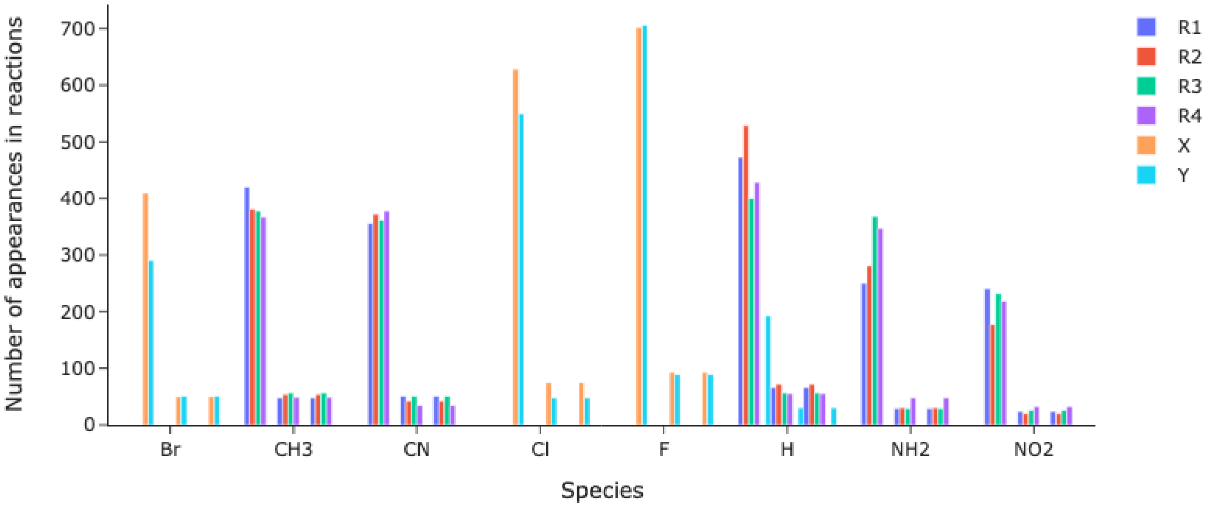


FIG. S3: Occurence of different species in the train (leftmost bars), validation (middle bars) and test (rightmost bars). Halogens (Br, Cl and F) are present in the reaction only as attacking and leaving species.

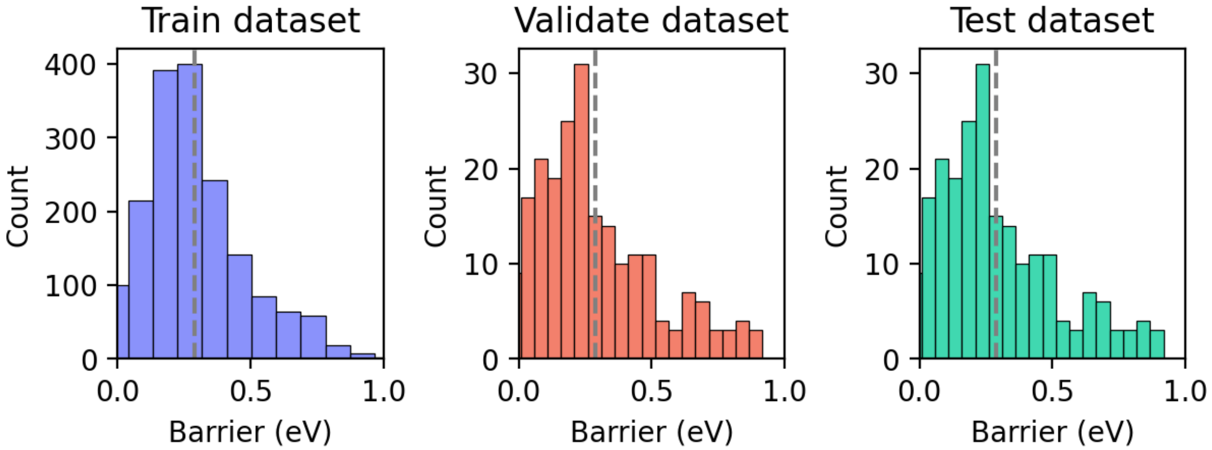


FIG. S4: Distribution of the barriers in the train, validation and test datasets. The dashed grey line indicates the mean of the distribution (train: 0.291 eV, validation: 0.287 eV and test: 0.287 eV.)

Figure S3 shows the prevalence of different species in the train, validation and test dataset. Through a random split, it appears that each species is represented well and equally in all datasets. Figure S4 shows the distribution of the activation barriers for the train, validate and test datasets. The mean of all distributions (dashed grey line) is very similar across all datasets.

## S6. BASELINE PERFORMANCE

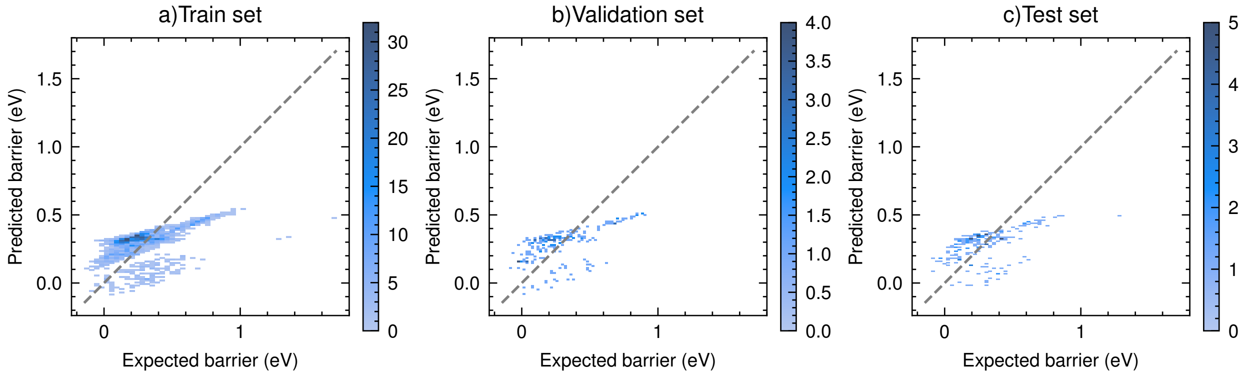


FIG. S5: Performce of the baseline model (see text) on the barrier prediction task for a) train b) validation c) test sets.

In this section, we discuss the performance of a baseline model. Through this baseline, we can gauge how accurate *CoeffNet* is in comparison to a simpler model.

We construct a baseline model by constructing linear relations between the activation barrier and the energy difference between the reactant and product. This type of linear relation is often referred to as a Brønsted-Evans-Polanyi (BEP) relation.[2,3] Figure S5 shows the performance of such a BEP model on the same $S_N2$ dataset used in the manuscript to test the performance of *CoeffNet*. Overall, the baseline model predictions are not very accurate. The relatively low mean absolute error (MAE) comes from the fact that the predictions are clustered around $0 - 0.5$ while the real barriers are in a similar range.

## S7. ANALYSIS OF REQUIRED TRAINING DATA

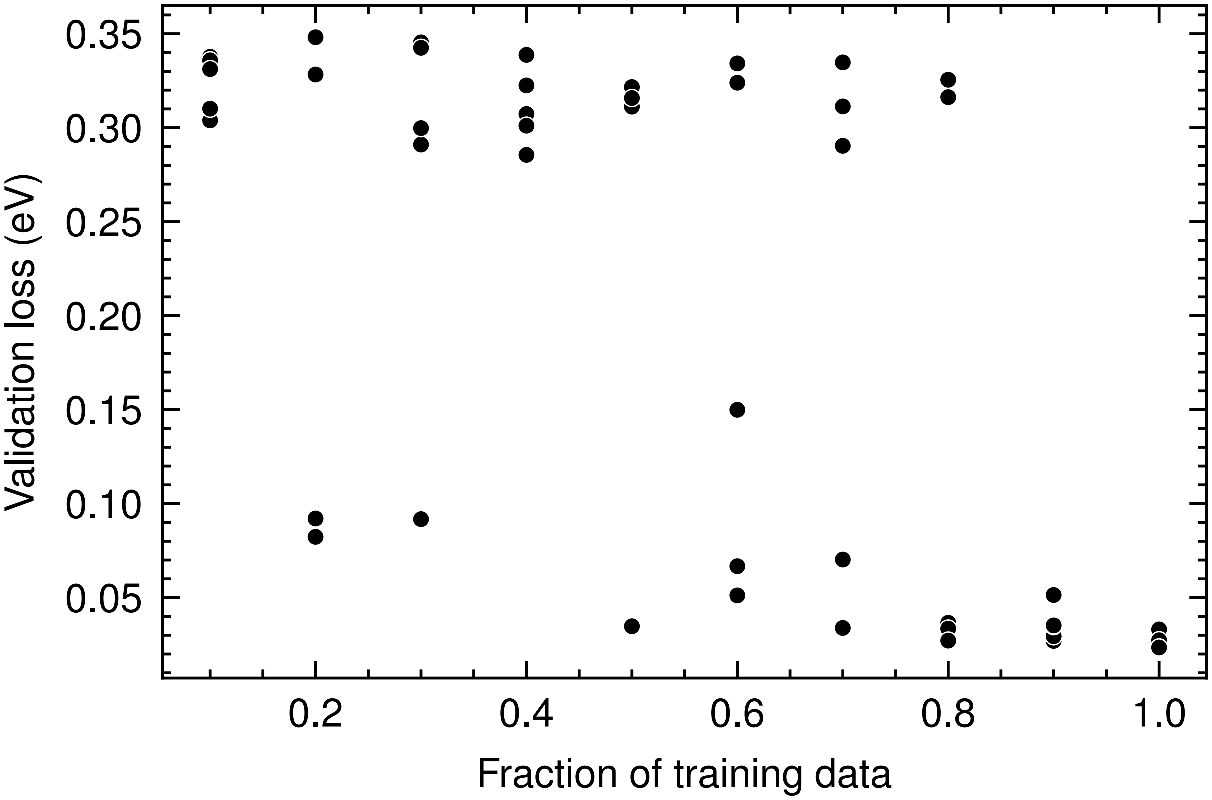


FIG. S6: Variation of the validation loss against the fraction of training data. Four different fractions of the training data were sampled (black points).

In this section, we discuss the effect of the number of training data points on the validation loss. Figure S6 shows the change in the validation loss ($y$-axis) with the fraction of the training dataset used in the main text if a large fraction is used (greater than 0.9, which is approximately 1600-1800 data points), we find that the performance is very similar to that reported in the main text. However, for smaller fractions of the dataset, we find that the performance varies markedly with the dataset that is used. This difference is likely due to the fact that smaller fractions do not "see" certain elements of the dataset (such as triple bonds of carbon and nitrogen in CN) and hence lead to larger errors.

## S8. ACTIVATION BARRIER PERFORMANCE ON DEF2-TZVP

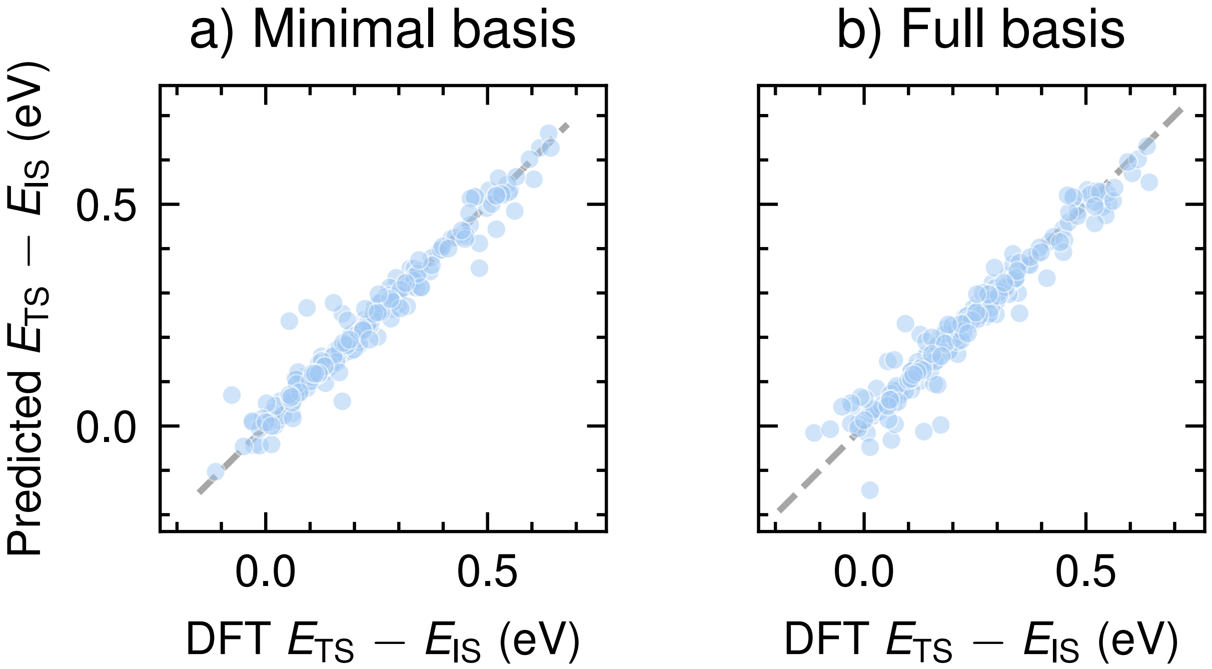


FIG. S7: Comparison of predicted activation barriers and DFT computed barriers on the test set for a) minimal-basis representation (i.e. only $s$ and $p$ basis functions, and b) full-basis (i.e. as computed with DFT) for the def2-TZVP basis set

## S9. TRAINING CURVES FOR COEFFICIENT MATRIX PREDICTION

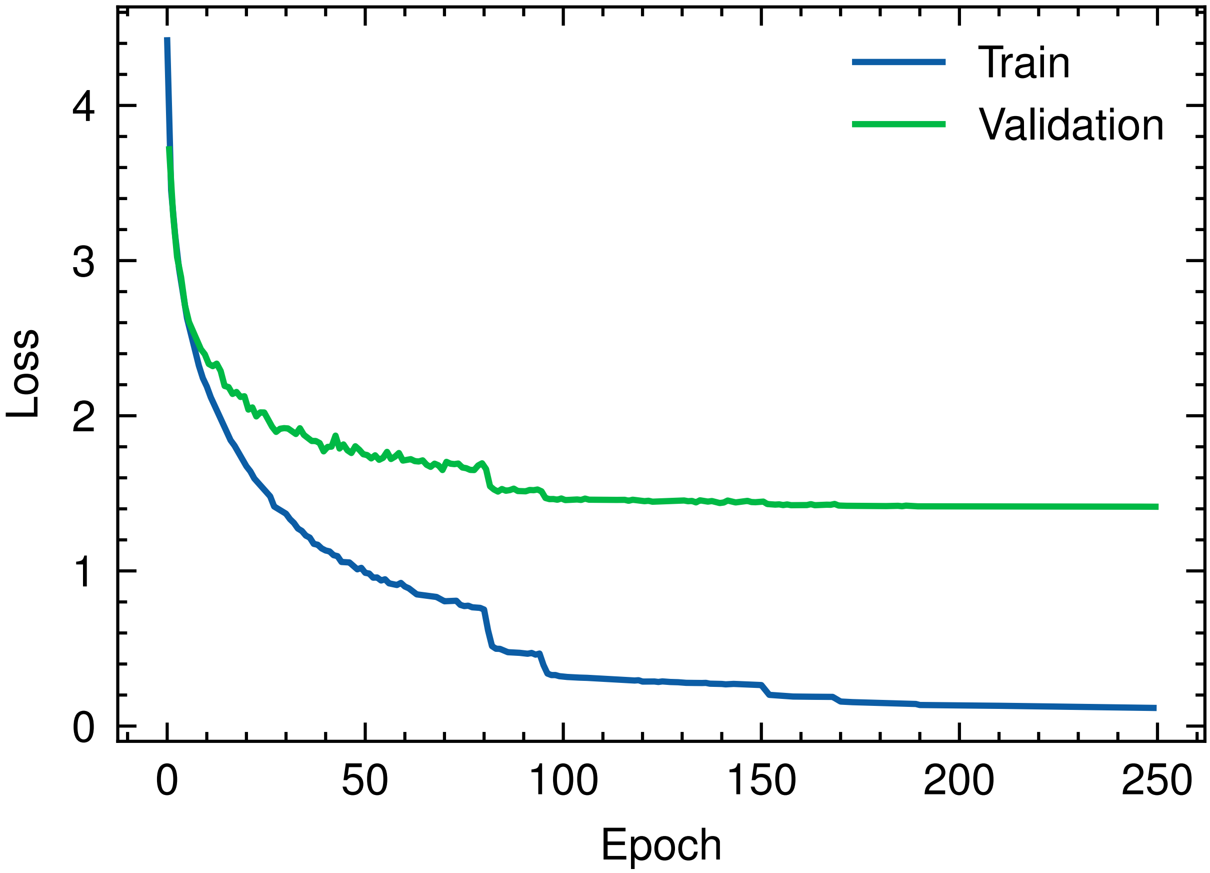


FIG. S8: Training curves showing the *unsigned*-MAE loss function

## S10. TRAINING ON A DATASET WITH DIVERSE CHEMICAL REACTIONS

Our main goal in this manuscript is to illustrate that the coefficients of molecular orbitals serve as a chemically expressive descriptor for tasks involving predicting activation barriers at sufficient accuracy to be used for applications studying chemical reactivity. Typically, errors in prediction of activation barriers need to be at least of the order of magnitude of the errors associated with DFT predictions, which are typically less than 0.05 eV.[4]

Given the scarcity of datasets containing activation energies needed to reach these high accuracies, we chose to focus on datasets like Ref 5 which have high quality data for a specific *type* of reaction. The added benefit of the approach is that we can analyze the *predicted* coefficients of the molecular orbitals at the transition state, allowing us to interpret our results.

In this section, we test the performance of *CoeffNet* on activation barriers generated from diverse chemical reactions. Instead of our original goal of high-accuracy activation barrier prediction, we investigate if *CoeffNet* can generalize to more diverse datasets out-of-the-box, simply by increasing the number of parameters of the model.

### A. Details of the dataset

We recompute the dataset generated by Ref 6 and Ref 7 to generate coefficients for molecular orbitals. We choose reactions where the difference between the number of bonds in the reactant and the product are between -1 and 1, i.e., at most one bond is formed or broken in the reaction. Overall, this choice generates $\approx 14,000$ activation barriers. We compute the coefficient matrices for the reactants and products using the computational setup as described in Section VI.

### B. Multi-eigenstate mode of *CoeffNet*

In order to train a model for a diverse dataset, it is preferred to use more parameters. The straightforward way of including more useful parameters into *CoeffNet* is to use coefficients of more than one molecular orbital as inputs to the model. In this section, we train *CoeffNet* using the coefficients from three eigenstates, the HOMO, HOMO-1 and the LUMO orbitals.

Note that this mode of *CoeffNet* is more expensive than that used in the main text, where the coefficients of only one orbital (typically the HOMO) is used as inputs to *CoeffNet*. Practically, this mode of operation can be used by setting the `idx_eigenvalue` flag in the input `yaml` file used by *CoeffNet*.

## C. Results

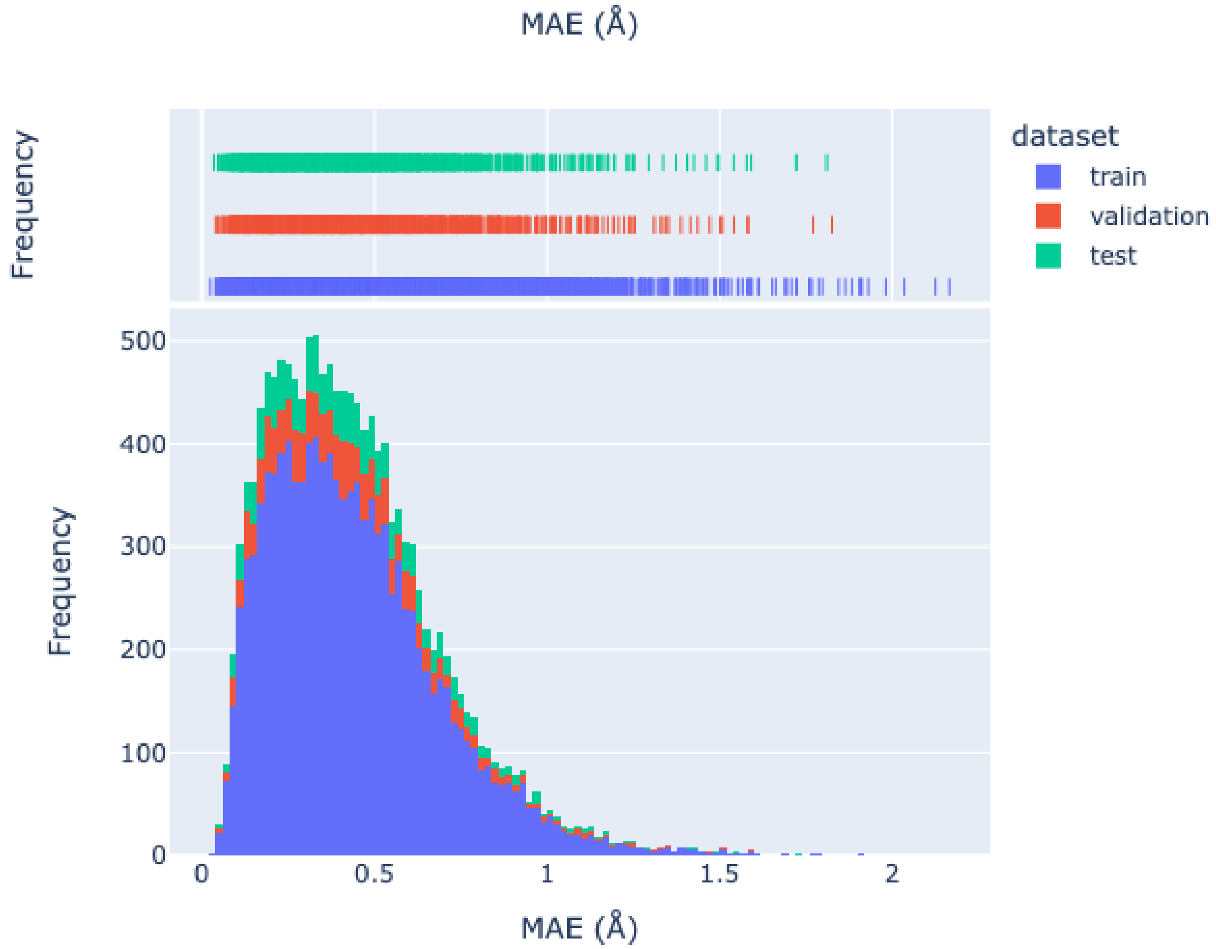


FIG. S9: Histogram of the mean absolute error for approximate prediction of the positions are atom centers for the transition state (equivalent to the analysis of Figure 4). Results are reported for train (purple), validation (red) and test (green) datasets.

Figure S9 shows the mean absolute error (MAE) for prediction of the transition state structure using the same procedure as illustrated in the main text. The train:validation:test

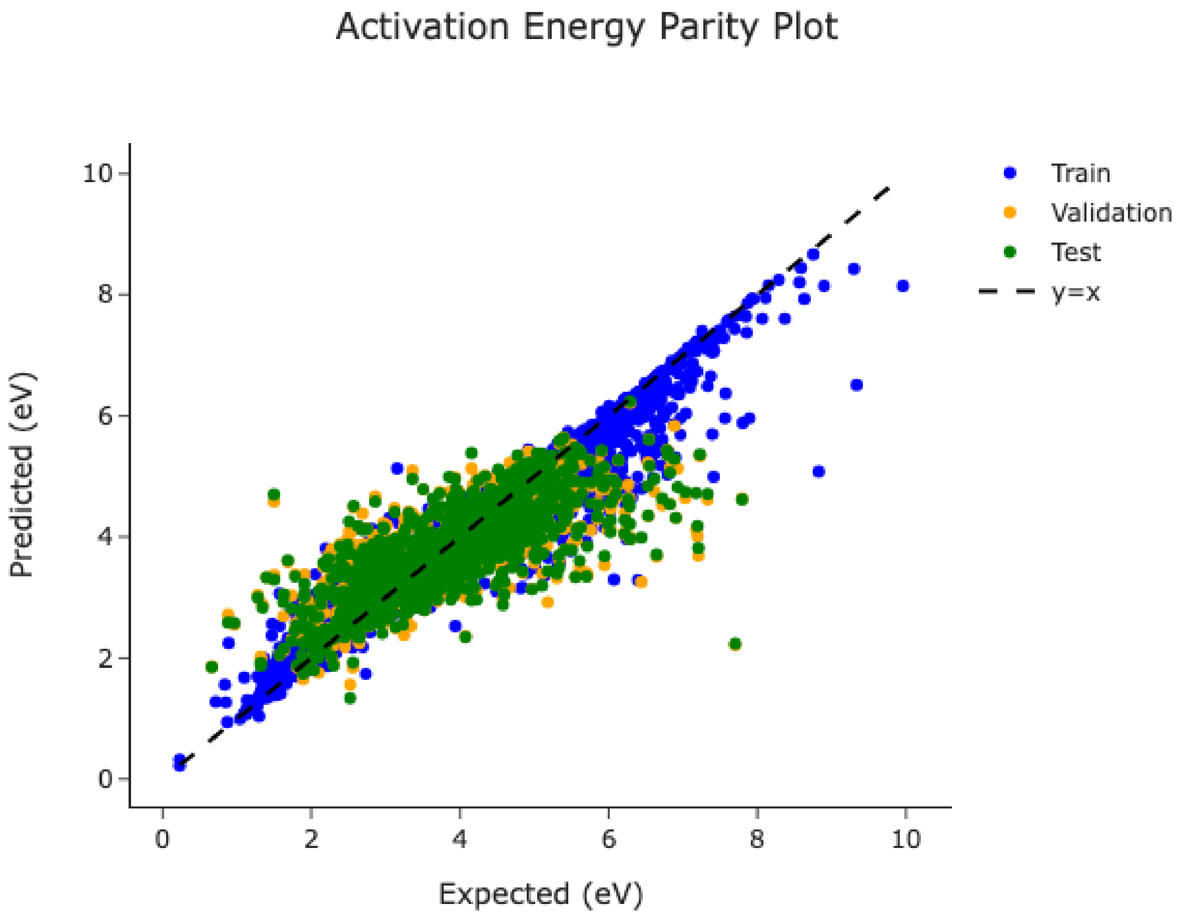


FIG. S10: Predicted ($y$-axis) and expected ($x$-axis) activation barriers for the train (in blue), validation (in yellow) and test (in green) datasets

split is chosen to be 80:10:10, identical to what is used in the main text.

As expected, the MAE for this diverse dataset is larger than that of the more ordered dataset studied in the main text. However, the errors for most structures are well below 0.5Å and only a small fraction of reactions (less than 20) have MAE greater than 1Å. Furthermore, all train, validation and test datasets have similar MAE distributions, indicating that the data splits are indicative representations of the dataset. Given that *CoeffNet* only requires somewhat *reasonable* transition state structures (and not accurate ones) to predict the activation energies, we believe that no further modification in methods is needed prior to training.

Figure S10 shows the results of the activation energy prediction task using *CoeffNet*. The blue points show the error for the training set, the yellow for the validation and the green for the test set. The mean absolute error is 0.42 eV on a dataset which spans ≈ 10 eV. We do not perform extensive hyperparameter tests as our goal is simply to show that *CoeffNet* in its current form can generalize to larger datasets by increasing the number of parameters in the model. Instead, we apply the procedure as used in the main text for the $S_N2$ dataset (see Section VI).

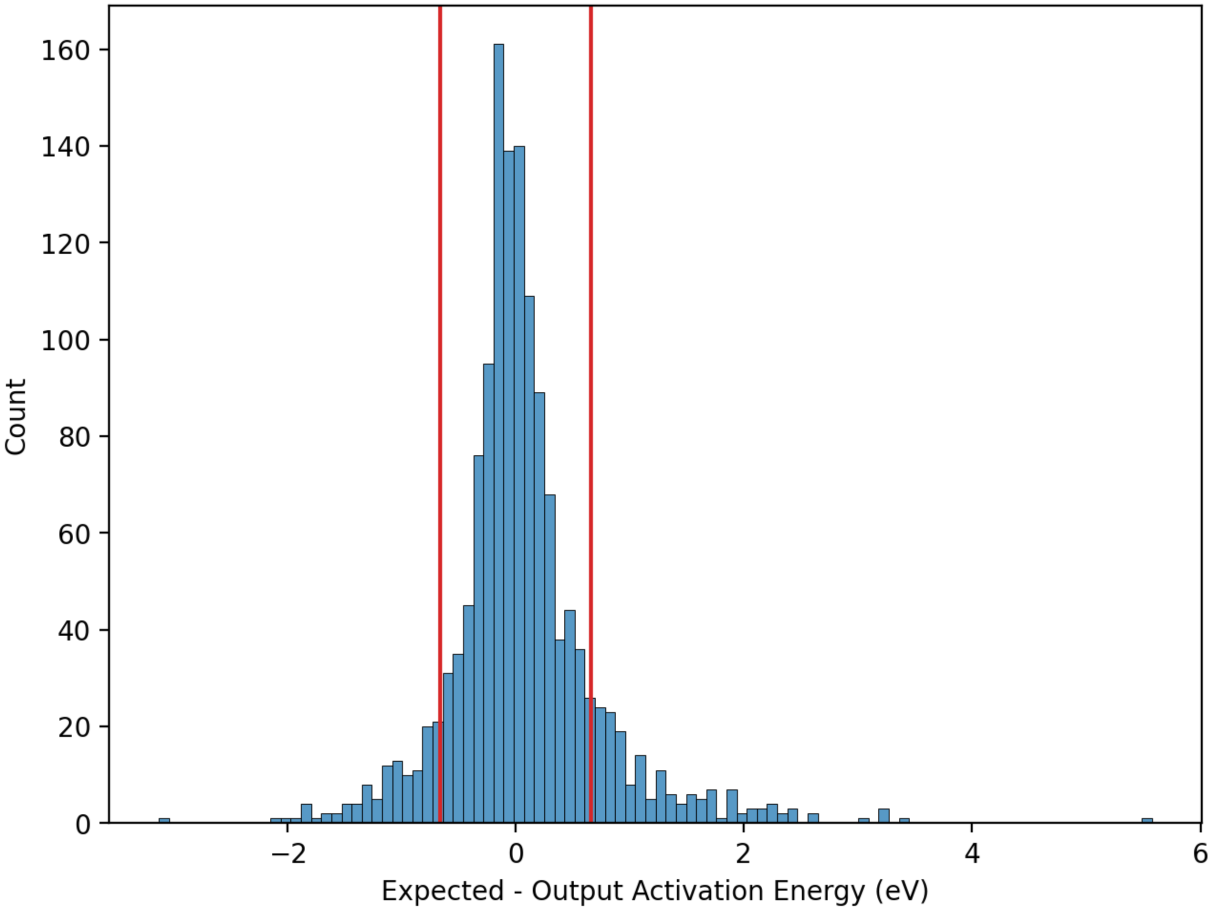


FIG. S11: Difference between the expected (computed) and the output (from *CoeffNet*) activation energy for the test set; red lines indicate the standard deviation of the error (= 0.68 eV)

A common use-case of *CoeffNet* is to act as a sieve for high-throughput reaction network studies. Instead of performing computationally intensive DFT calculations and transition state searches on all pathways of a reaction network, *CoeffNet* is used to predict *relevant* pathways where the activation barriers are lower than a chosen threshold. DFT calculations and transition state searches are performed only for those reactions which have *CoeffNet* predicted activation energies below this chosen threshold. This strategy of pre-sieving a large compositional space using *CoeffNet* and then refining activation barrier predictions with DFT methods is crucial to rapid understanding of complex reaction networks.

In order for this pre-seiving strategy to work accurately in practice, we require a low standard-deviation of the error as well as a reasonably low mean absolute error. Note that the absolute value of the activation barrier is less important than the spread of the error, as the true activation barrier will be computed through DFT calculations if the predicted value from *CoeffNet* is lower than a chosen threshold.

Figure S11 shows a histogram of the difference between expected (computed using DFT) and output (from *CoeffNet*) activation barriers for the test set of the dataset from Ref 6. The vertical red lines indicate the standard devation of the errors, which is 0.68 eV. Given the small deviation of the errors from the mean, we expect that *CoeffNet* would be able to accurately predict when barriers would be too large to be chemically interesting (and hence, not worth computing with DFT) and those that are sufficiently small (and hence, worth computing with DFT).

## REFERENCES

[1] T. Stuyver and S. Shaik, “Unifying conceptual density functional and valence bond theory: The hardness–softness conundrum associated with protonation reactions and uncovering complementary reactivity modes,” Journal of the American Chemical Society **142**, 20002–20013 (2020).

[2] T. Z. Gani and H. J. Kulik, “Understanding and Breaking Scaling Relations in Single-Site Catalysis: Methane to Methanol Conversion by FeIV=O,” ACS Catalysis **8**, 975–986 (2018).

[3] P. N. Plessow and F. Abild-Pedersen, “Examining the linearity of transition state scaling relations,” Journal of Physical Chemistry C **119**, 10448–10453 (2015), publisher: American Chemical Society.

[4] A. D. Kaplan, C. Shahi, P. Bhetwal, R. K. Sah, and J. P. Perdew, “Understanding density-driven errors for reaction barrier heights,” Journal of Chemical Theory and Computation **19**, 532–543 (2023).

[5] G. F. von Rudorff, S. N. Heinen, M. Bragato, and O. A. von Lilienfeld, “Thousands of reactants and transition states for competing e2 and s2 reactions,” Machine Learning: Science and Technology **1**, 045026 (2020).

[6] C. A. Grambow, L. Pattanaik, and W. H. Green, “Deep Learning of Activation Energies,” Journal of Physical Chemistry Letters **11**, 2992–2997 (2020), publisher: American Chemical Society.

[7] C. A. Grambow, L. Pattanaik, and W. H. Green, “Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry,” Scientific data **7**, 137 (2020).