Supplementary Information (SI) for Digital Discovery.

Distortion/Interaction Analysis via Machine Learning

Samuel G. Espley[a], Samuel S. Allsop[a], David Buttar[b], Simone Tomasi[c], and Matthew N. Grayson[a]*

a) Department of Chemistry, University of Bath, Claverton Down, Bath, BA2 7AY, UK. *M.N.Grayson@bath.ac.uk
b) Data Science and Modelling, Pharmaceutical Sciences, R&D, AstraZeneca, Macclesfield, UK.
c) Chemical Development, Pharmaceutical Technology & Development, Operations, AstraZeneca, Macclesfield, UK.

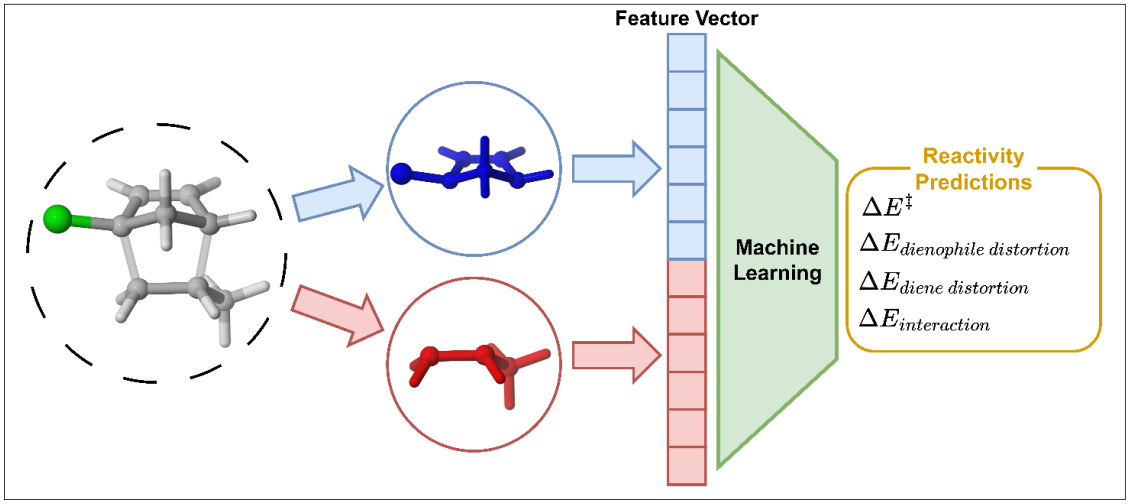

Table of Contents

1. Dataset Generation

Four different datasets were used throughout this work and are as follows: Nitro-Michael addition (ds1), Diels-Alder (ds2), [3+2] cycloaddition (ds3), and dimethyl malonate Michael addition (ds4). Descriptions of these datasets can be found in S1.1, S1.2, S1.3, and S1.4, respectively. Additionally, an extra dataset was created from literature examples. Its creation is outlined in section S1.5. Any SQM and DFT calculations run as part of this work were performed using Gaussian 16 (versions A.01 and C.03).[1,2] The Gaussian 16 computed output files are publicly available in Dataset for "Distortion/Interaction Analysis via Machine Learning" in the University of Bath Research Data Archive (accessible at: https://doi.org/10.15125/BATH-01398). All structures visualised within this work were created using CYLView.[3]

1.1. Nitro-Michael addition Dataset (ds1)

This nitro-Michael addition dataset is a fully enumerated, published dataset of ground state reactants (GSs) and transition structures (TSs).[4] These structures were obtained from the associated data archive (https://doi.org/10.15125/BATH-01092) and subsequently used for distortion/interaction analysis. For this, the TSs were split into distorted Michael acceptor and distorted nitromethane-derived nucleophile before single point energy (SPE) calculations were performed. The IEFPCM solvent model[5] was used for this with toluene as the selected solvent to match that of the original work.[4]

1.2. Diels-Alder Dataset (ds2)

This Diels-Alder dataset was a published dataset of GSs and TSs.[6] These structures were obtained from the associated data archive (https://doi.org/10.15125/BATH-01229) and subsequently used for distortion/interaction analysis. For this, the TSs were split into distorted diene and distorted dienophile before SPE calculations were performed. These calculations on the distorted structures were performed in the gas phase to match that of the original work.[6]

1.3. [3+2] Cycloaddition Dataset (ds3)

This [3+2] cycloaddition dataset was a published dataset of DFT optimised GSs and TSs.[7] These structures were extracted from the associated GitHub page (https://github.com/coleygroup/dipolar_cycloaddition_dataset). This dataset is comprehensive and provides a significantly varied view into the bioorthogonal [3+2] click reaction; however, it does present some challenges. While the dataset has the same [3+2] reactivity throughout, the atom mapping is arbitrary across the dataset. This means that equivalent atoms across reactions may not have the same atom number. This provides a challenge with utilising a 'common atoms' approach that has previously worked well for predicting reaction barriers via ML.[4,6] To determine the atom mapping for the TSs, we utilised the nature of cycloadditions having two reaction centres. Using these two reaction centres, we were able to determine connectivity between the four reacting atoms and thus determine the atom mapping of the dipole and dipolarophile in the TS. Once we obtained the atom mapping for the TS, we could calculate the GS and distorted GSs atom mapping for each individual system. The code used to perform this separation and determination of common atoms is available on the associated GitHub (https://github.com/the-grayson-group/distortion-interaction_ML). The process for separation is outlined in section S2. The TSs were split into distorted dipole and distorted dipolarophile before SPE calculations were performed. These calculations on the distorted structures were performed in the same solvent system used in the original research.

1.4. Dimethyl malonate Michael addition Dataset (ds4)

A dimethyl malonate Michael addition dataset of TSs and their equivalent GSs was created via enumeration over a common backbone in the same manner as Grayson et al.[4] The common backbone was built and enumerated over four different positions using the Custom R-Group Enumeration feature of Schrödinger's Maestro.[8] The functional groups used to create the enumerated dataset are in Fig. S1. This enumeration resulted in 1000 unique Michael addition reactions. For this enumeration, only the Michael acceptor was varied; the dimethyl malonate nucleophile remained constant across the entire dataset.

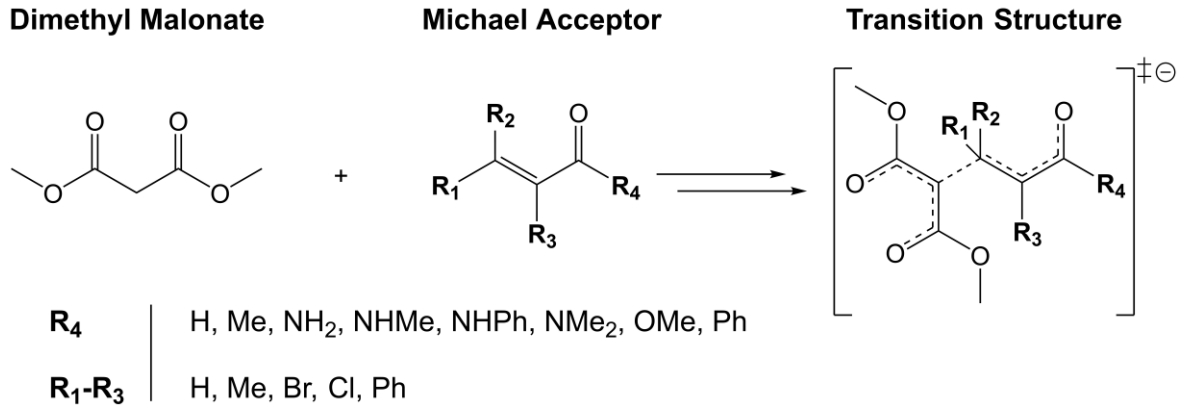

Fig. S1 - Representation of how the dimethyl malonate Michael addition dataset was created with functional groups used.

Upon completion of the enumeration, Schrödinger's MacroModel[9] was used to conformationally search all GSs and TSs using the forcefield OPLS3e[10] to obtain the lowest energy conformer for each respective structure. With all conformationally searched lowest energy conformers obtained, the structures were then optimised with AM1[11] and ωB97X-D/def2-TZVP[12,13] to either GSs or TSs. All 1000 TSs were successfully optimised with both AM1 and ωB97X-D/def2-TZVP. All optimisation calculations were performed in the gas phase with SPE correction calculations performed on the optimised structures in solvent. The solvent model was the integral equation formalism of the polarisable continuum model (IEFPCM) with water.[5] These TSs were subsequently used for distortion/interaction analysis. For this, the TSs were split into distorted dimethyl malonate nucleophile and distorted Michael acceptor before SPE calculations were performed. The IEFPCM solvent model was used for this with water as the selected solvent to match that of the GS and TS calculations.

1.5. Further Literature Datasets

To further validate our models trained on ds2, we took two high impact instances of distortion/interaction analysis being used for reactivity insight in Diels-Alder reactions from the literature. One study investigated the reactivity of cycloalkenones with cyclic dienes[14] while the other explored the reactivity and stereoselectivity of cyclopropene Diels-Alder reactions.[15] Fig. S2 shows the structures that make up the reactions in these literature test sets. In total, this resulted in 23 unseen datapoints that have previously been studied to elucidate information about Diels-Alder reactivity.

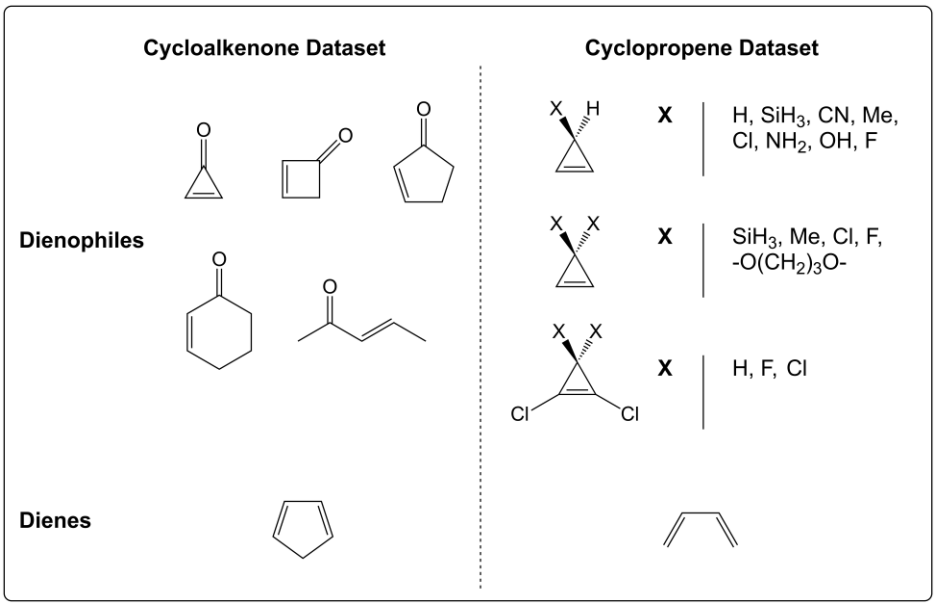

Fig. S2 - External test set created from data found in references 14 and 15.

2. Distortion Calculations

Information on the separation of distorted structures from the TS can be found in the methodology section of the main paper. Extra information/code can also be found in the associated GitHub repository (https://github.com/the-grayson-group/distortion-interaction_ML). With the distorted structures separated, the SPE calculations were performed using Gaussian 16 (versions A.01 and C.03) at both AM1 and ωB97X-D/def2-TZVP levels of theory. These calculations were performed exactly as the equivalent GS structure calculations were performed i.e., if the GS was optimised in the gas phase, then the SPE calculation of the distorted structure was also performed in the gas phase. All Gaussian 16 computed distortion output files are publicly available in Dataset for "Distortion/Interaction Analysis via Machine Learning" in the University of Bath Research Data Archive (accessible at: https://doi.org/10.15125/BATH-01398).

Once the calculations were completed, energies were extracted using the python package GoodVibes[16]. These energies were then used to calculate the distortion energies (Equation S1). When obtained, these distortion energies for all components in the reaction were then summated to yield $\Delta E_{distortion}$ which was subsequently used alongside $\Delta E^{\ddagger}$ to determine the interaction energy for the system (Equation S2).

$$\Delta E_{distortion\ energy} = E_{distorted\ structure} - E_{GS\ structure} \quad \text{(S1)}$$

$$\Delta E_{interaction} = \Delta E^{\ddagger} - \Delta E_{distortion} \quad \text{(S2)}$$

3. Data Spread

3.1. ds1

Figures S4-S7 are histograms to show the spread of $\Delta G^{\ddagger}$, $\Delta E^{\ddagger}$, nitromethane-derived nucleophile and Michael acceptor distortion energies, and interaction energies respectively for the nitro-Michael addition dataset (ds1).[4] Blue indicates the AM1 energies and red indicates the DFT (ωB97X-D/def2-TZVP) energies.

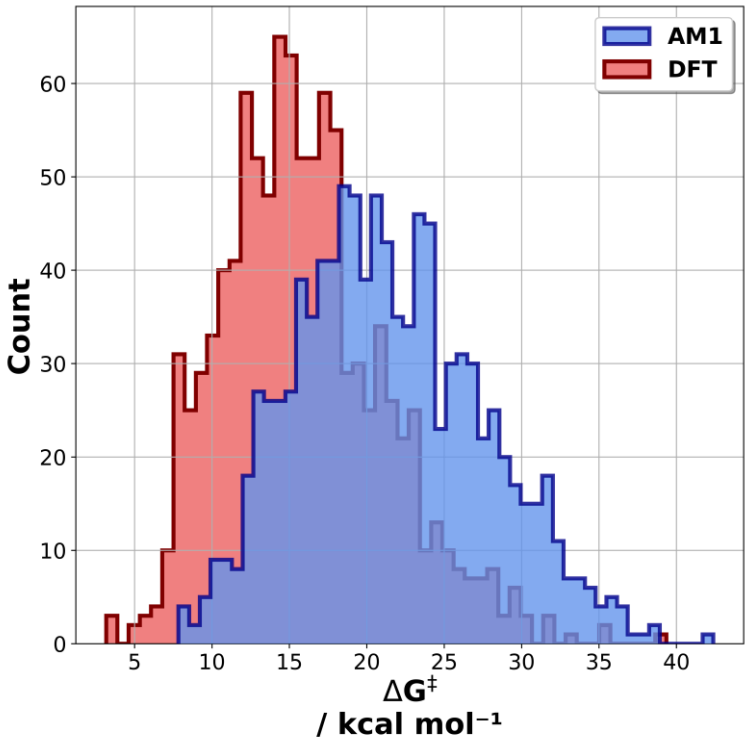

Fig. S3 - Histogram showing spread of AM1 (Blue) and DFT (Red) $\Delta G^{\ddagger}$ for the nitro-Michael addition dataset (ds1).

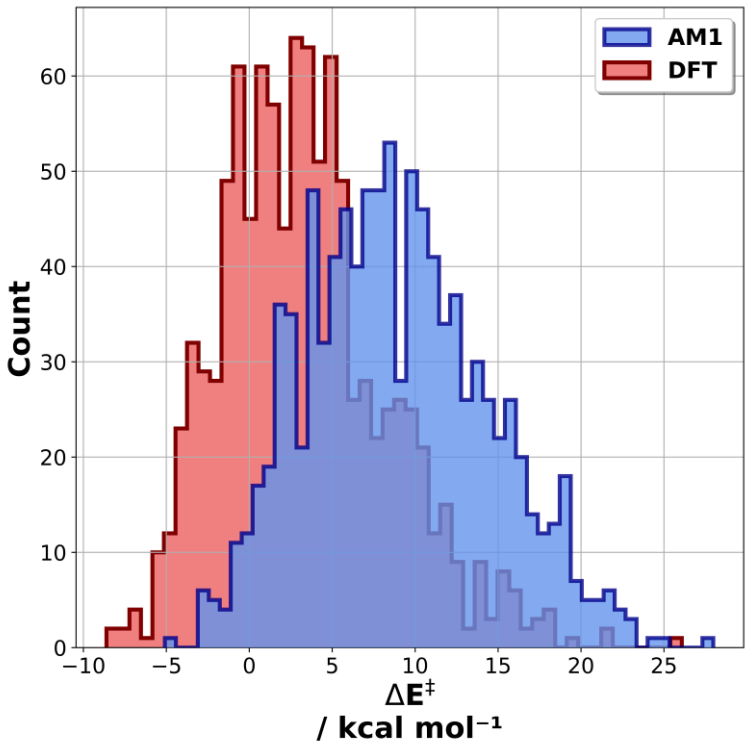

Fig. S4 - Histogram showing spread of AM1 (Blue) and DFT (Red) $\Delta E^{\ddagger}$ for the nitro-Michael addition dataset (ds1).

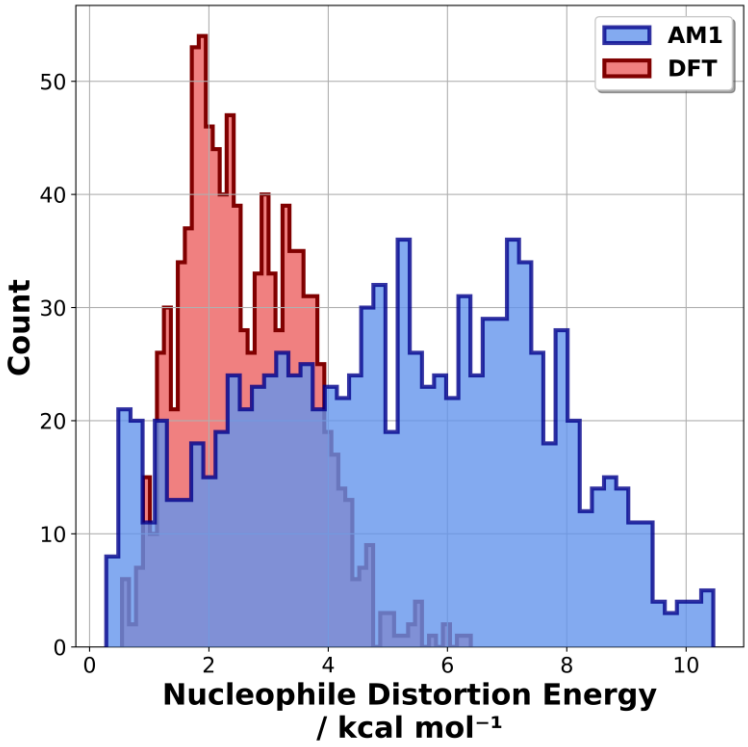

Fig. S5 - Histogram showing spread of AM1 (Blue) and DFT (Red) nucleophile distortion energies for the nitro-Michael addition dataset (ds1).

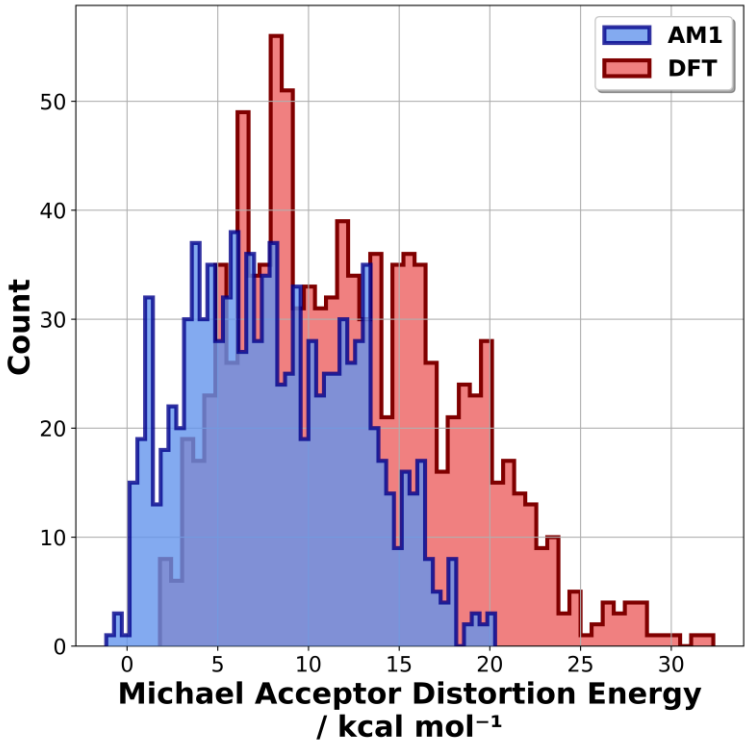

Fig. S6 - Histogram showing spread of AM1 (Blue) and DFT (Red) Michael acceptor distortion energies for the nitro-Michael addition dataset (ds1).

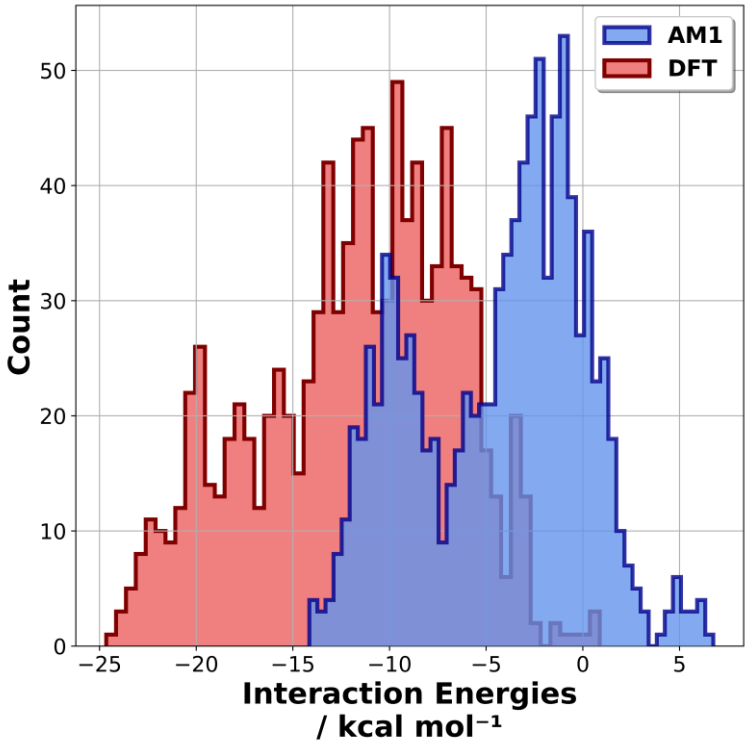

Fig. S7 - Histogram showing spread of AM1 (Blue) and DFT (Red) interaction energies for the nitro-Michael addition dataset (ds1).

3.2. ds2

Figures S8-S12 are histograms to show the spread of $\Delta G^{\ddagger}$, $\Delta E^{\ddagger}$, diene and dienophile distortion energies, and interaction energies respectively for the Diels-Alder dataset (ds2).[6] Blue indicates the AM1 energies and red indicates the DFT (ωB97X-D/def2-TZVP) energies.

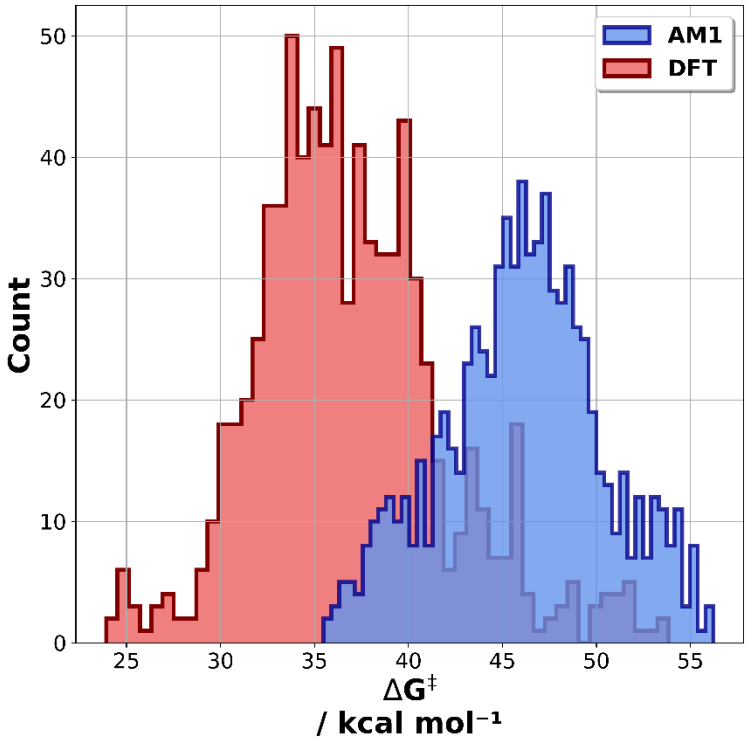

Fig. S8 - Histogram showing spread of AM1 (Blue) and DFT (Red) $\Delta G^{\ddagger}$ for the Diels-Alder dataset (ds2).

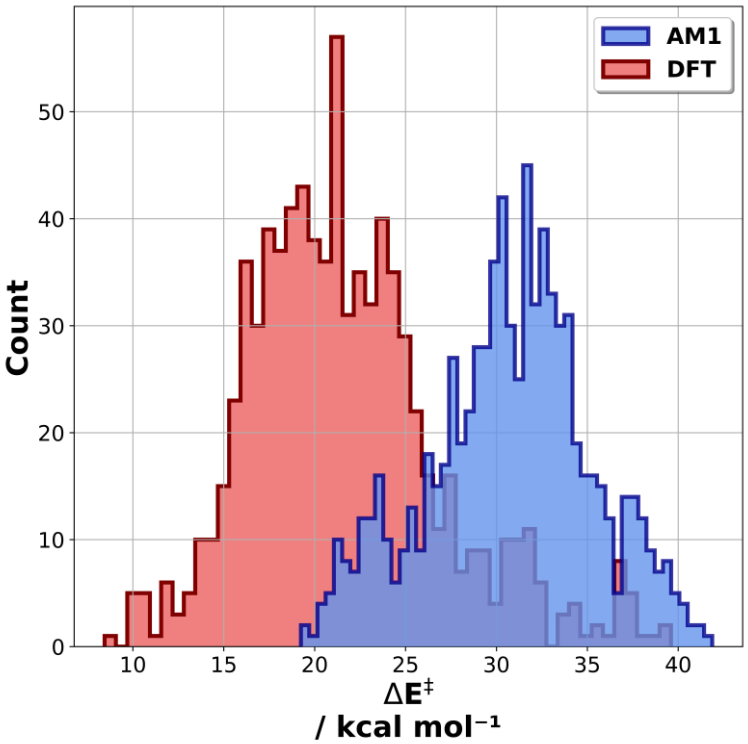

Fig. S9 - Histogram showing spread of AM1 (Blue) and DFT (Red) $\Delta E^{\ddagger}$ for the Diels-Alder dataset (ds2).

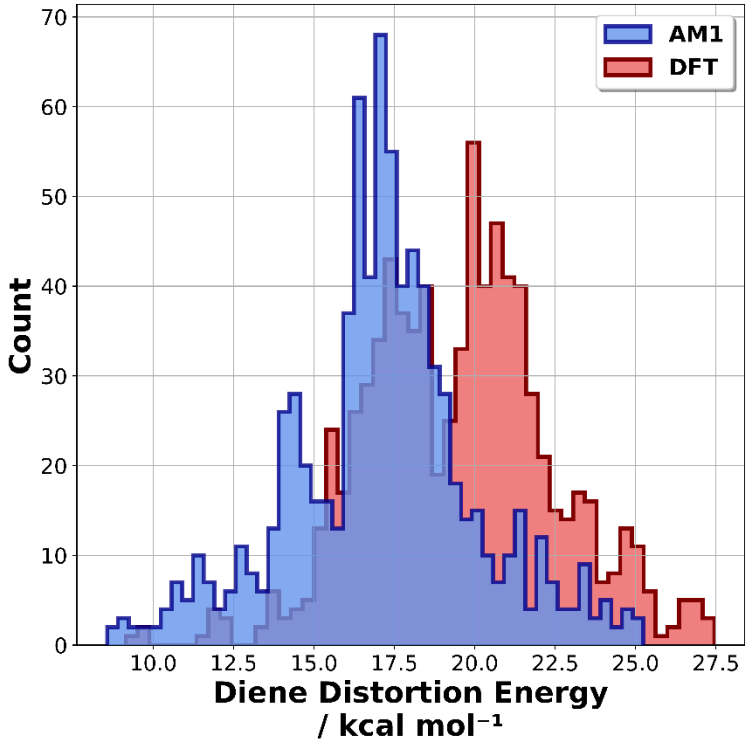

Fig. S10 - Histogram showing spread of AM1 (Blue) and DFT (Red) diene distortion energies for the Diels-Alder dataset (ds2).

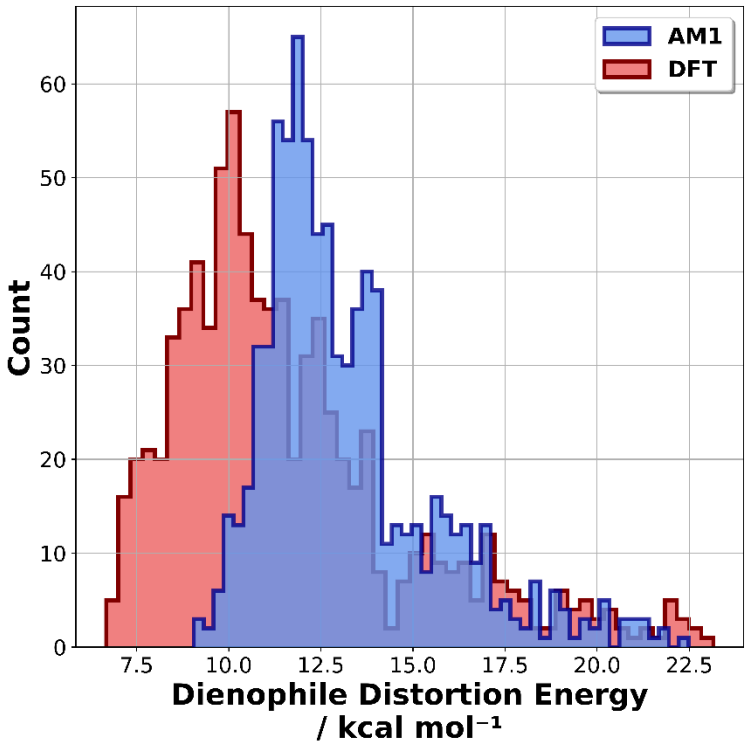

Fig. S11 - Histogram showing spread of AM1 (Blue) and DFT (Red) dienophile distortion energies for the Diels-Alder dataset (ds2).

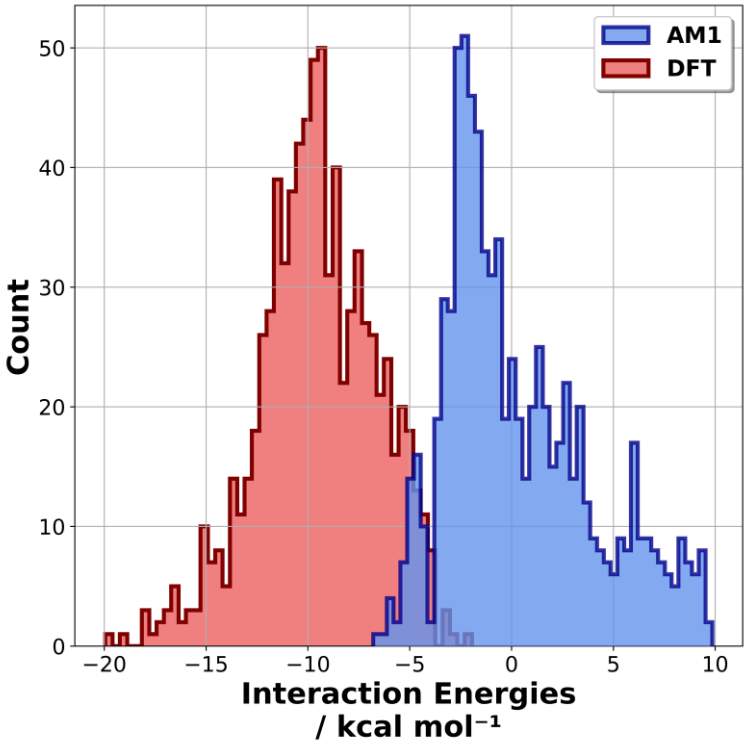

Fig. S12 - Histogram showing spread of AM1 (Blue) and DFT (Red) interaction energies for the Diels-Alder dataset (ds2).

3.3. ds3

Figures S13-S17 are histograms to show the spread of $\Delta G^{\ddagger}$, $\Delta E^{\ddagger}$, dipole and dipolarophile distortion energies, and interaction energies respectively for the [3+2] cycloaddition dataset (ds3). Blue indicates the AM1 energies and red indicates the DFT (ωB97X-D/def2-TZVP) energies.

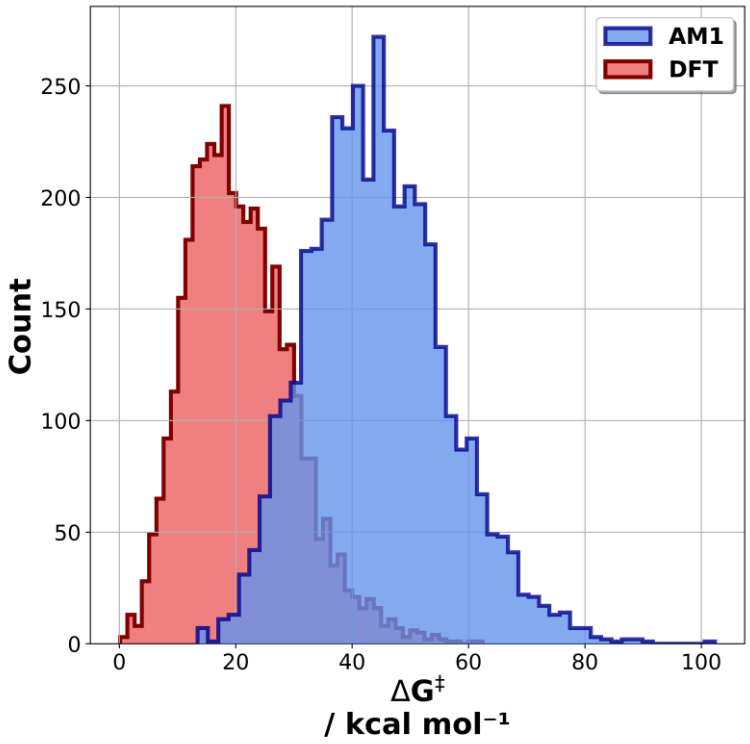

Fig. S13 - Histogram showing spread of AM1 (Blue) and DFT (Red) $\Delta G^{\ddagger}$ for the [3+2] cycloaddition dataset (ds3).

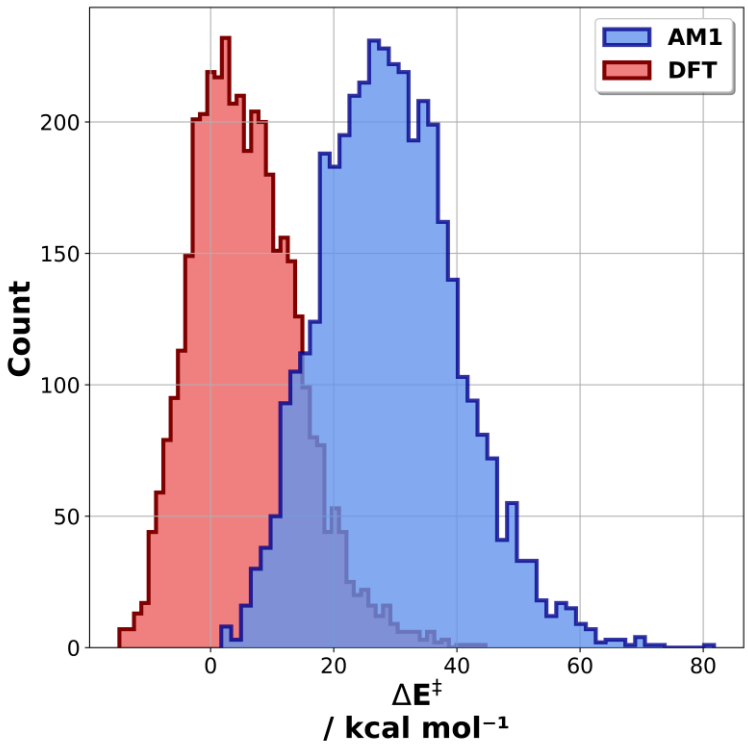

Fig. S14 - Histogram showing spread of AM1 (Blue) and DFT (Red) $\Delta E^{\ddagger}$ for the [3+2] cycloaddition dataset (ds3).

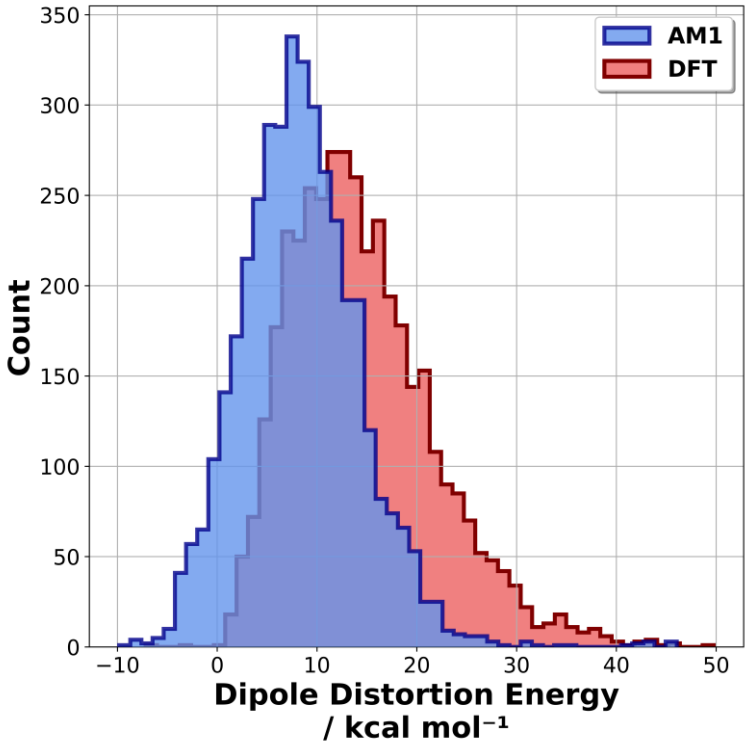

Fig. S15 - Histogram showing spread of AM1 (Blue) and DFT (Red) dipole distortion energies for the [3+2] cycloaddition dataset (ds3).

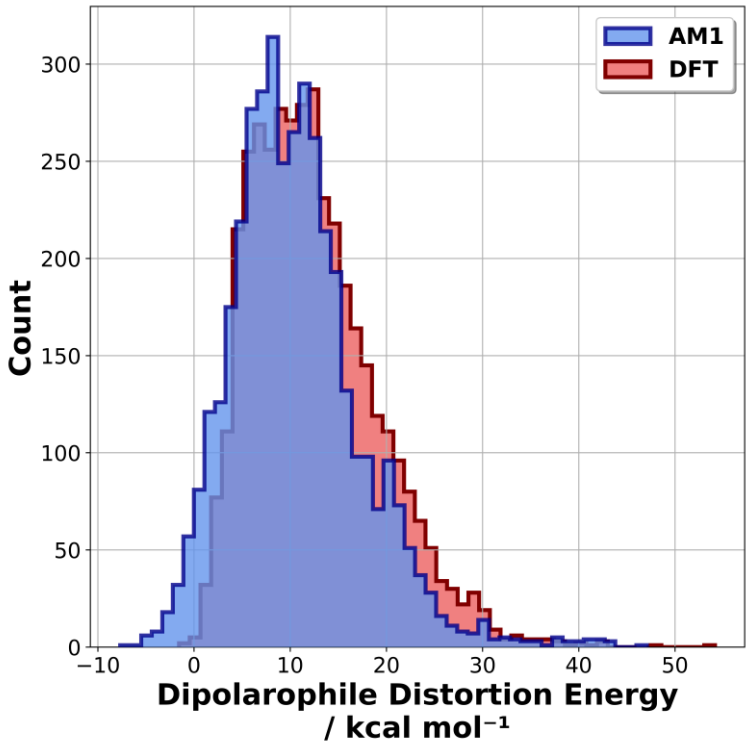

Fig. S16 - Histogram showing spread of AM1 (Blue) and DFT (Red) dipolarophile distortion energies for the [3+2] cycloaddition dataset (ds3).

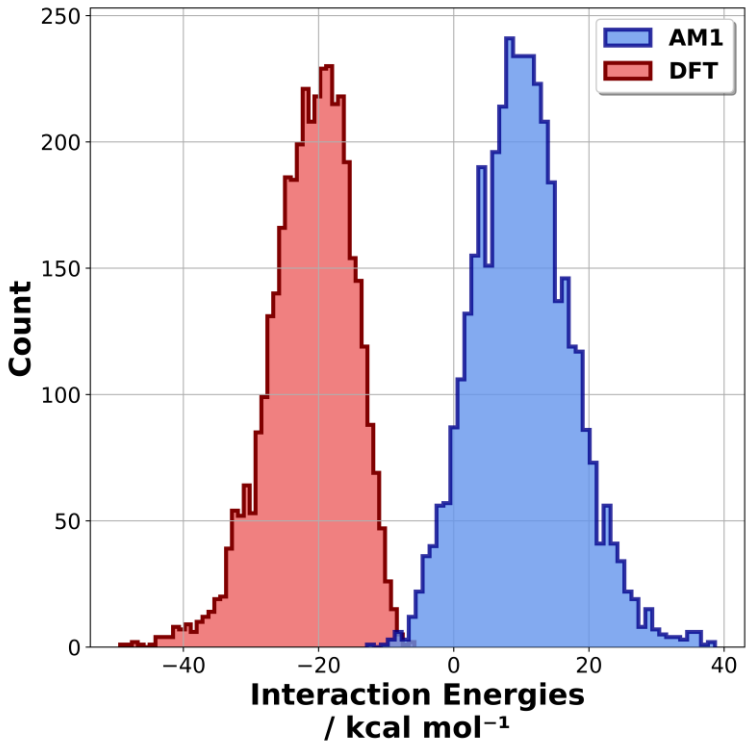

Fig. S17 - Histogram showing spread of AM1 (Blue) and DFT (Red) interaction energies for the [3+2] cycloaddition dataset (ds3).

3.4. ds4

Figures S18-S22 are histograms to show the spread of $\Delta G^{\ddagger}$, $\Delta E^{\ddagger}$, dimethyl malonate (nucleophile) and Michael acceptor distortion energies, and interaction energies respectively for the dimethyl malonate Michael addition dataset (ds4). Blue indicates the AM1 energies and red indicates the DFT (ωB97X-D/def2-TZVP) energies.

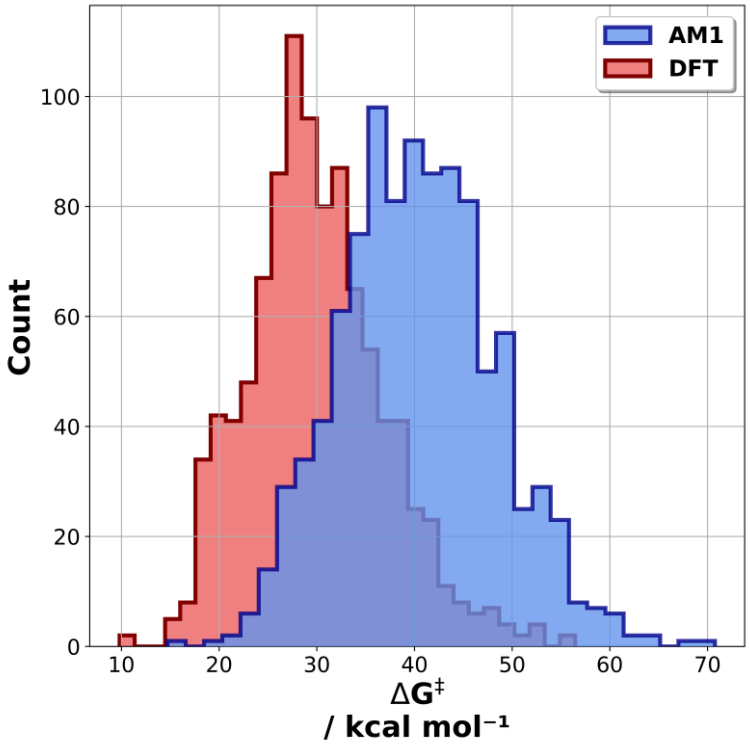

Fig. S18 - Histogram showing spread of AM1 (Blue) and DFT (Red) $\Delta G^{\ddagger}$ for the dimethyl malonate Michael addition dataset (ds4).

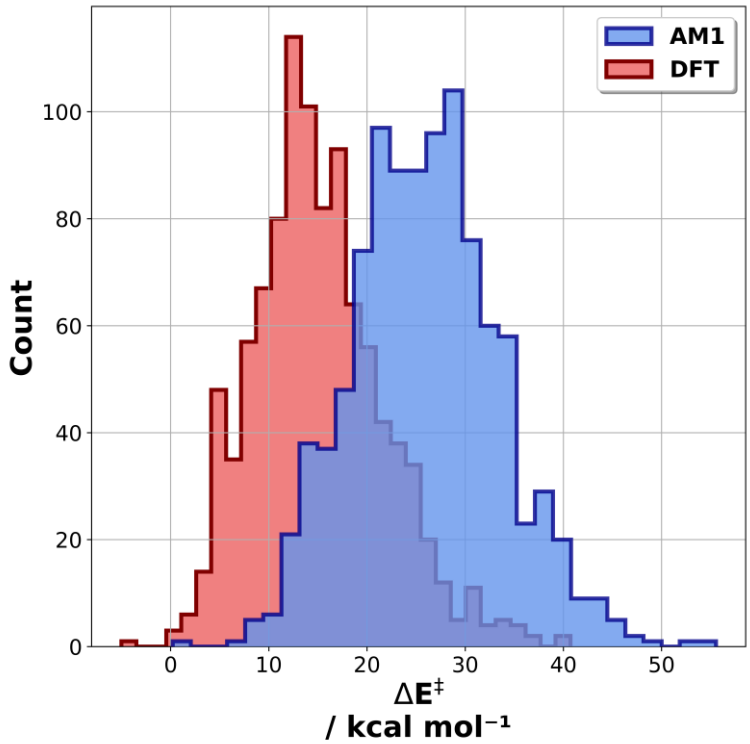

Fig. S19 - Histogram showing spread of AM1 (Blue) and DFT (Red) $\Delta E^{\ddagger}$ for the dimethyl malonate Michael addition dataset (ds4).

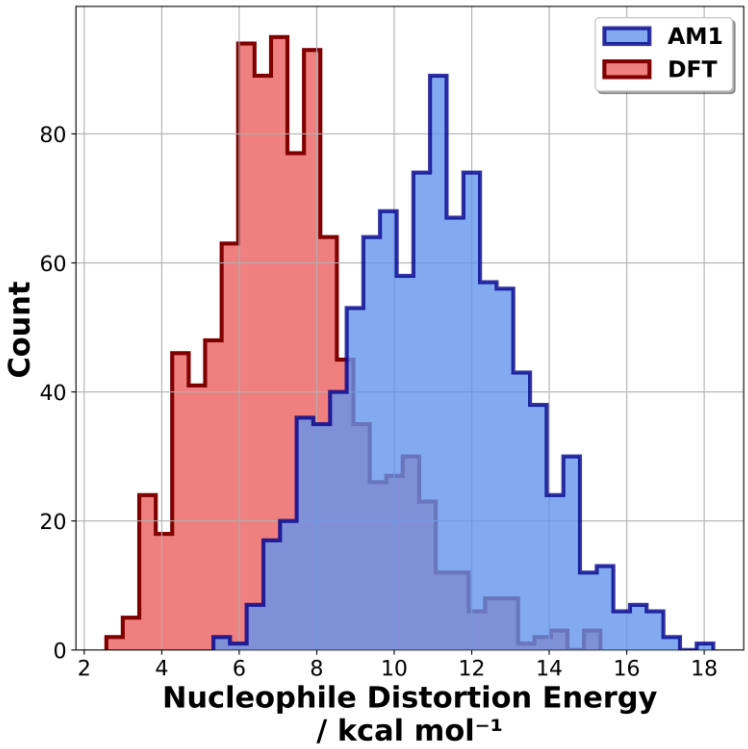

Fig. S20 - Histogram showing spread of AM1 (Blue) and DFT (Red) nucleophile distortion energies for the dimethyl malonate Michael addition dataset (ds4).

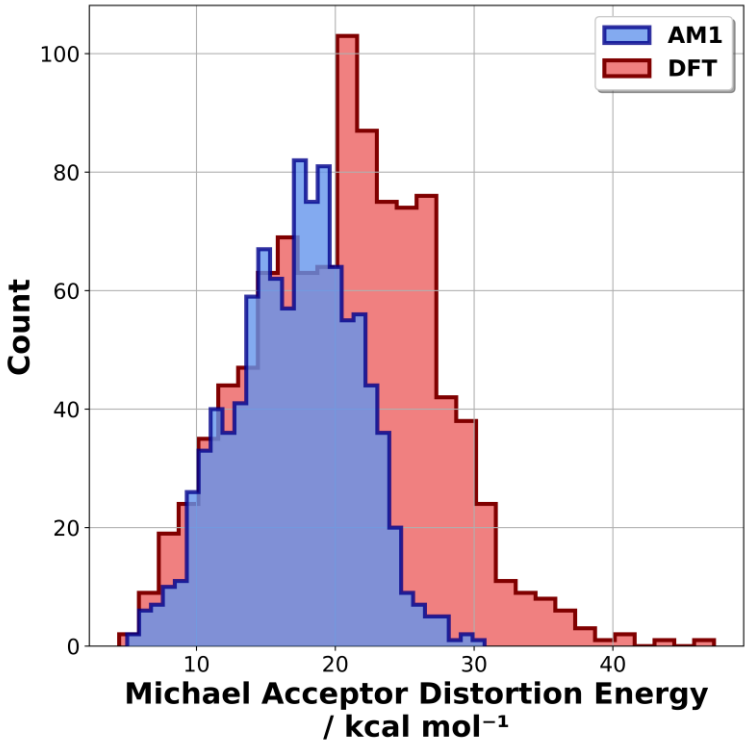

Fig. S21 - Histogram showing spread of AM1 (Blue) and DFT (Red) Michael acceptor distortion energies for the dimethyl malonate Michael addition dataset (ds4).

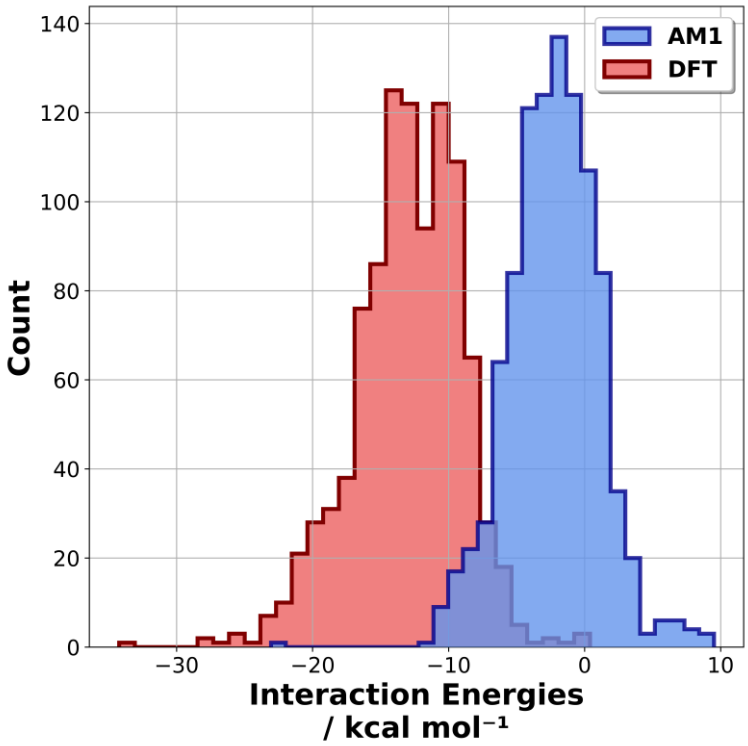

Fig. S22 - Histogram showing spread of AM1 (Blue) and DFT (Red) interaction energies for the dimethyl malonate Michael addition dataset (ds4).

4. Machine Learning (ML)

The methods used in this work were ridge regression, kernel ridge regression (KRR) with the polynomial kernel, support vector regression (SVR) with the radial basis function kernel (RBF), a 2-layer neural network (NN), and a 4-layer NN due to their performance in previous barrier prediction tasks.[4,6] In our ML approach, we utilise a common atom-based feature extraction. This requires a pre-existing knowledge of the common atoms in the system across a dataset.

4.1. Feature Extraction

A similar approach was used as in our previous work[4,6] for feature extraction, however the method of extraction was adapted for multiple datasets, along with different python packages utilised. All features are derived from the semi-empirical AM1 Gaussian 16 calculations[1,2] which avoids the high computational cost associated with running DFT calculations. The python code used to extract features for the four datasets used in this work along with accompanying information about usage can be found in the associated GitHub page (https://github.com/the-grayson-group/distortion-interaction_ML). Several different physical organic features were extracted for all of the different datasets based upon common atoms. Fig. S23 shows the positions of the atoms for which features were taken. The complete list of extracted features and their associated python packages can be found in Table S1, with the features selected as model inputs shown in Table S2. All features either have one or two associated numbers. One number indicates that this feature is associated with that numbered atom whereas two numbers indicate a feature that is associated with both of those atoms. The exception to this is *reacting_distance_0_ts_ts* and *reacting_distance_1_ts_ts* which corresponds to the two reacting distances in the two cycloaddition datasets (ds2 and ds3).

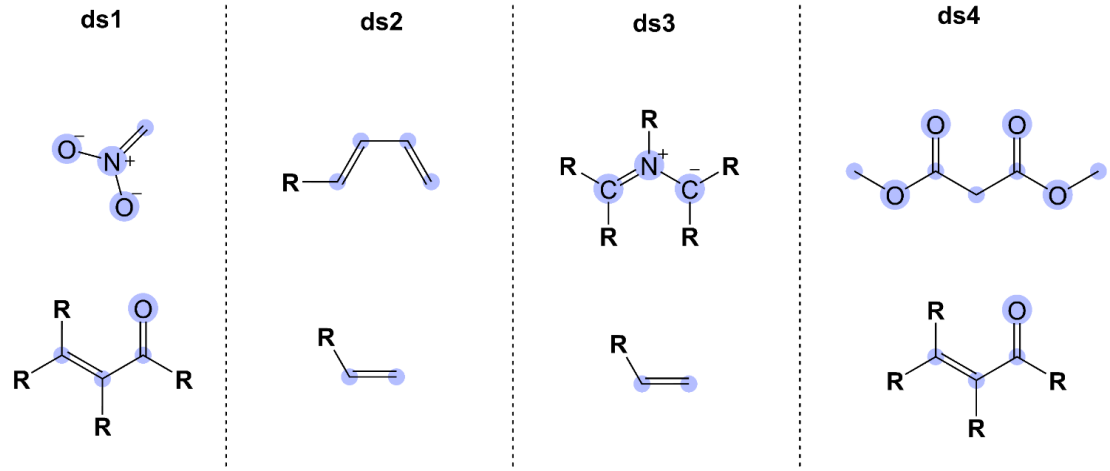

Fig. S23 - Location of the common atoms for every dataset. These locations are shown for GS/distorted GS structures however, the same positions are used for the TS geometries when extracting features.

Each dataset is unique and thus, has different common atoms to extract. As such, the features require shorthand notation to provide clarity. The characters **X** and **Y** denote the atoms that the feature are extracted from. The character **Z** denotes the chemical structure that this feature is derived from e.g., distance_**X**_**Y**_**Z** could be the feature associated with the distance between atoms **X** and **Y** for the structure **Z**. To provide a contextual example, distance_**1**_**2**_**dp_dist_gs** is the distance between atom **1** and **2** for the **dienophile distorted GS structure**.

After the feature extraction, multiple different approaches were evaluated for feature selection. These were sequential feature selection (SFS), recursive feature elimination cross validation (RFECV) and principal component analysis (PCA). After evaluating these approaches, we settled on using a manual approach, selecting specific features rather than allowing these approaches to select from a large number of features in a computationally expensive manner. The selected features can be found in Table S2.

Table S1 - List of all extracted features from AM1 calculations. Each feature has a brief description coupled with the source of how it was obtained. All features were extracted using python.

Feature	Description	Source
q_barrier_am1	Semi-empirical quasi-harmonic free energy reaction barrier.	Gaussian 16 derived[1,2], GoodVibes extracted[16].
e_barrier_am1	Activation energy of reaction ($\Delta E^{\ddagger}$).	Gaussian 16 derived[1,2], GoodVibes extracted[16].
sum_distortion_energies_am1	Summation of the distortion energies for the reaction.	Gaussian 16 derived[1,2], GoodVibes extracted[16].
distortion_energy_**Z**_am1	Distortion energy for the distorted structure **Z**. For datasets used, there are two of these features.	Gaussian 16 derived[1,2], GoodVibes extracted[16].
interaction_energies_am1	Interaction energy for the reaction. Calculated from distortion energies and $\Delta E^{\ddagger}$.	Gaussian 16 derived[1,2], GoodVibes extracted[16].
apt_charge_**X**_**Z**	APT atomic charge for atom **X** in structure **Z**.	cclib[17]
apt_sum_**Z**	APT atomic summed charge for structure **Z**.	cclib[17]
buried_volume_**X**_**Z**	Percent buried volume (3.5 Å) for atom **X** in structure **Z**.	morfeus[18]
distance_**X**_**Y**_**Z**	Calculated distance between atoms **X** and **Y** for structure **Z**.	morfeus[18]
fraction_buried_volume_**X**_**Z**	Fraction buried volume (3.5 Å) for atom **X** in structure **Z**.	morfeus[18]
free_volume_**X**_**Z**	Free volume of atom **X** in structure **Z**.	morfeus[18]
mulliken_charge_**X**_**Z**	Mulliken atomic charge for atom **X** in structure **Z**.	cclib[17]
mulliken_sum_**Z**	Mulliken atomic summed charge for structure **Z**.	cclib[17]
reacting_distance_**N**_ts	The reacting distance in Å. This feature depends upon if the TS has two (**N** = 0 or 1) or one reaction centres (**N** = 0).	Coordinates derived from Gaussian 16[1,2].
reacting_distance_diff_ts	If two reaction centres, this feature is present and is the difference between the two distances in Å.	Coordinates derived from Gaussian 16[1,2].
sasa_**X**_**Z**	Solvent accessible surface area for atom **X** in structure **Z**.	morfeus[18]
sasa_area_**Z**	Summated solvent accessible surface area for structure **Z**.	morfeus[18]
sasa_volume_**Z**	Volume inside solvent accessible surface for structure **Z**.	morfeus[18]
sterimol_B_1_value_**X**_**Y**_**Z**	Minimum rotational size of the atom **Y** relative to atom **X** in structure **Z**.	morfeus[18]
sterimol_B_5_value_**X**_**Y**_**Z**	Maximum rotational size of the atom **Y** relative to atom **X** in structure **Z**.	morfeus[18]
sterimol_L_value_**X**_**Y**_**Z**	Depth of the atom **Y** relative to atom **X** in structure **Z**.	morfeus[18]

Table S2 - List of extracted features from AM1 calculations chosen by manual feature selection. Each feature has a brief description coupled with the source of how it was obtained. All features were extracted using python.

Feature	Description	Source
q_barrier_am1	Semi-empirical quasi-harmonic free energy reaction barrier.	Gaussian 16 derived[1,2], GoodVibes extracted[16].
e_barrier_am1	Activation energy of reaction ($\Delta E^{\ddagger}$).	Gaussian 16 derived[1,2], GoodVibes extracted[16].
sum_distortion_energies_am1	Summation of the distortion energies for the reaction.	Gaussian 16 derived[1,2], GoodVibes extracted[16].
distortion_energy_**Z**_am1	Distortion energy for the distorted structure **Z**. For datasets used, there are two of these features.	Gaussian 16 derived[1,2], GoodVibes extracted[16].
interaction_energies_am1	Interaction energy for the reaction. Calculated from distortion energies and $\Delta E^{\ddagger}$.	Gaussian 16 derived[1,2], GoodVibes extracted[16].
mulliken_charge_**X**_**Z**	Mulliken atomic charge for atom **X** in structure **Z**.	cclib[17]
apt_charge_**X**_**Z**	APT atomic charge for atom **X** in structure **Z**.	cclib[17]
distance_**X**_**Y**_**Z**	Calculated distance between atoms **X** and **Y** for structure **Z**.	morfeus[18]
reacting_distance_**N**_ts	The reacting distance in Å. This feature depends upon if the TS has two (**N** = 0 or 1) or one reaction centres (**N** = 0).	Coordinates derived from Gaussian 16[1,2].
reacting_distance_diff_ts_ts	If two reaction centres, this feature is present and is the difference between the two distances in Å.	Coordinates derived from Gaussian 16[1,2].

4.2. Hyperparameter Tuning

For the models built in sklearn[19] (ridge regression, KRR, and SVR), GridSearchCV with 5-fold cross validation was utilised to search the hyperparameter space. For NNs built with TensorFlow [20], Hyperband was used. [21] For the NNs, two different architectures were used. These models were feed forward NNs that either had 2 or 4 hidden layers. The simplistic choice for the architecture was to avoid training a complex model on limited data (which would result in overfitting) with an associated increase in computational cost. The model architectures are displayed visually in Fig. S24. Full search spaces for the sklearn models and NNs can be found in Table S3 while all code associated with the hyperparameter tuning can be found on the GitHub (https://github.com/the-grayson-group/distortion-interaction_ML). A random forest model was also tested for ds2. The performance of the random forest model was similar to ridge regression but not as accurate as SVR, KRR and the 2-layer NN (Table S8). When hyperparameter tuning the models, a fixed random seed of **23** was chosen. After hyperparameter tuning with the random seed of **23**, these tuned hyperparameters were then used to train each model over 5 random seeds (**22, 23, 14, 1, and 2**).

Table S3 – Table containing the tuning method and search space for hyperparameter tuning of machine learning models. The two NNs utilise the same search space but pull from this search space multiple times depending upon how many hidden layers/dropout layers. For example, the 2-layer NN only has one dropout layer however, the 4-layer NN has three dropout layers.

Model	Hyperparameters	
	Tuning Method	**Search Space**
Ridge Regression	GridSearchCV	{'alpha':[0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0], 'tol':[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]}
Kernel Ridge Regression (Polynomial Kernel)	GridSearchCV	{'alpha':[0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0], 'gamma':[None, 0.1, 0.5, 0.9]}
Support Vector Regression (Radial Basis Function Kernel)	GridSearchCV	{'gamma':['auto', 'scale'], 'epsilon':[0.001, 0.01, 0.025, 0.05, 0.1, 0.25, 0.5, 1], 'C':[1, 30, 50], 'coef0': [0, 1], 'degree': [1, 2, 3]}
2 Layer NN	Hyperband[21]	{'dropout_rate':[0.6, 0.5, 0.4, 0.3, 0.2, 0.1], 'neurons':[32, 64, 128, 256, 512], 'reg_val':[2e-1, 1e-1, 1e-2, 1e-3], 'learning_rate':[1e-3, 1e-4, 1e-5]}
4 Layer NN	Hyperband[21]	{'dropout_rate':[0.6, 0.5, 0.4, 0.3, 0.2, 0.1], 'neurons':[32, 64, 128, 256, 512], 'reg_val':[2e-1, 1e-1, 1e-2, 1e-3], 'learning_rate':[1e-3, 1e-4, 1e-5]}
Random Forest	GridSearchCV	{ 'max_depth':[3, 5, 7], 'n_estimators':[10, 50, 100], 'max_features':[10, 20, 30], 'min_samples_leaf':[1, 2, 3]}

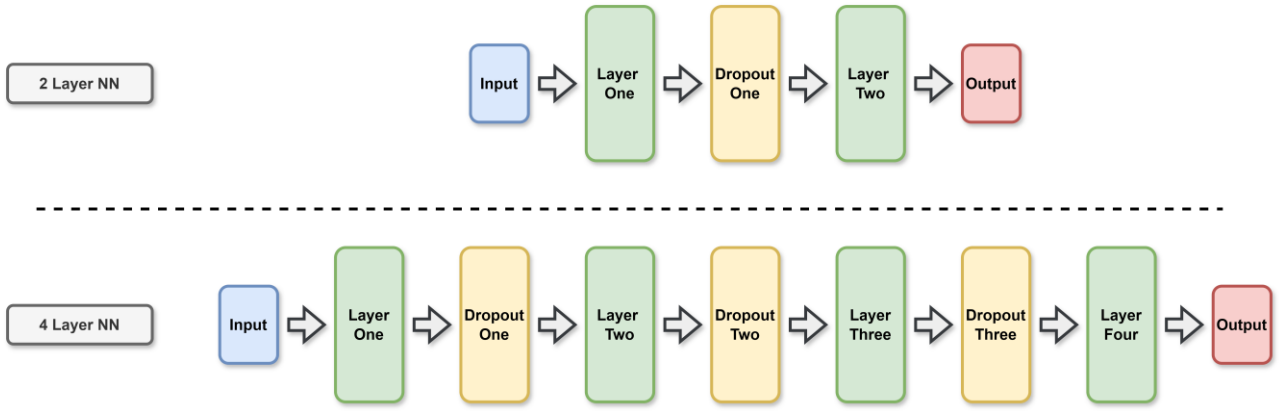

Fig. S24 - Diagram of NN architectures used in this work. Top is a feed forward NN with two hidden layers and bottom is a NN with four hidden layers.

4.3. ML Protocol and Results

For all datasets, the data was randomly split into train (80%), test (10%), and validation (10%) sets prior to training. The models were subsequently tuned and trained on the training data with the NNs using the validation set to find the best possible solution. For the models used from sklearn (ridge, SVR, and KRR), the performance was monitored using 5-fold cross validation. With the best performing hyperparameters for each model, the model was independently fit and then tested on the test set over five different random states. Full code used to perform the ML analysis can be found on the GitHub https://github.com/the-grayson-group/distortion-interaction_ML. Tables S4-S7 include pre-ML AM1-DFT MAEs, average train MAE over 5 random states, average test MAE over 5 random states, and the average MAE as a percentage of the average spread of the test set for each dataset on each task. The 5 random seeds used were 22, 23, 14, 1, 2.

Table S4 – Table showing the ds1 ML results. The table includes Pre-ML AM1-DFT MAE, average train MAEs, average test MAEs, average test set target range, and the average test MAE value as a percentage of the test set target range. All averaged results are over 5 random seeds (22, 23, 14, 1, and 2).

Model	Pre-ML AM1-DFT MAE (kcal mol^{-1})	Average Train MAE (kcal mol^{-1})	Average Test MAE (kcal mol^{-1})	Average Test Target Range (kcal mol^{-1})	Test MAE as % of Test Set Range
Nucleophile (nitromethane-derived) Distortion Energy					
Ridge	2.57	0.42	0.44 ± 0.04	0.73 to 5.36	9.5
KRR		0.38	0.38 ± 0.03		8.2
SVR		0.36	0.36 ± 0.04		7.8
2-Layer NN		0.22	0.33 ± 0.03		7.1
4-Layer NN		0.33	0.38 ± 0.04		8.2
Michael Acceptor Distortion Energy					
Ridge	4.45	1.72	1.72 ± 0.15	2.60 to 28.26	6.7
KRR		1.53	1.54 ± 0.13		6.0
SVR		1.45	1.35 ± 0.13		5.3
2-Layer NN		0.97	1.44 ± 0.13		5.6
4-Layer NN		1.24	1.43 ± 0.15		5.6
Interaction Energy					
Ridge	7.60	2.12	2.06 ± 0.18	-23.11 to -0.09	8.9
KRR		1.99	1.86 ± 0.16		8.1
SVR		1.78	1.59 ± 0.18		6.9
2-Layer NN		1.17	1.60 ± 0.17		7.0
4-Layer NN		1.59	1.73 ± 0.19		7.5
ΔG‡					
Ridge	5.71	1.42	1.44 ± 0.11	6.49 to 31.21	5.8
KRR		1.20	1.16 ± 0.10		4.7
SVR		1.05	1.00 ± 0.09		4.0
2-Layer NN		0.76	1.00 ± 0.09		4.0
4-Layer NN		0.94	1.10 ± 0.10		4.4
ΔE‡					
Ridge	5.76	1.49	1.51 ± 0.12	-5.65 to 17.80	6.4
KRR		1.17	1.10 ± 0.09		4.7
SVR		1.10	1.04 ± 0.09		4.4
2-Layer NN		0.81	1.09 ± 0.10		4.6
4-Layer NN		1.15	1.25 ± 0.11		5.3

Table S5 - Table showing the ds2 ML results. The table includes Pre-ML AM1-DFT MAE, average train MAEs, average test MAEs, average test set target range, and the average test MAE value as a percentage of the test set target range. All averaged results are over 5 random seeds (22, 23, 14, 1, and 2).

Model	Pre-ML AM1-DFT MAE (kcal mol^{-1})	Average Train MAE (kcal mol^{-1})	Average Test MAE (kcal mol^{-1})	Average Test Target Range (kcal mol^{-1})	Test MAE as % of Test Set Range
Diene Distortion Energy					
Ridge	2.80	0.76	0.61 ± 0.07	12.12 to 26.54	4.2
KRR		0.73	0.59 ± 0.08		4.1
SVR		0.39	0.33 ± 0.08		2.3
2-Layer NN		0.42	0.48 ± 0.07		3.3
4-Layer NN		0.50	0.49 ± 0.07		3.4
Dienophile Distortion Energy					
Ridge	2.04	0.80	0.77 ± 0.08	7.01 to 21.92	5.2
KRR		0.54	0.38 ± 0.05		2.5
SVR		0.44	0.34 ± 0.05		2.3
2-Layer NN		0.39	0.51 ± 0.07		3.4
4-Layer NN		0.83	0.92 ± 0.10		6.2
Interaction Energy					
Ridge	9.87	1.07	1.00 ± 0.09	-17.40 to -3.43	7.2
KRR		0.59	0.50 ± 0.06		3.6
SVR		0.54	0.50 ± 0.06		3.6
2-Layer NN		0.41	0.58 ± 0.06		4.2
4-Layer NN		0.64	0.75 ± 0.07		5.4
ΔG‡					
Ridge	9.20	1.29	1.14 ± 0.09	28.23 to 52.78	4.6
KRR		0.95	0.71 ± 0.09		2.9
SVR		0.70	0.56 ± 0.06		2.3
2-Layer NN		0.55	0.74 ± 0.07		3.0
4-Layer NN		1.03	1.06 ± 0.09		4.3
ΔE‡					
Ridge	9.13	1.35	1.20 ± 0.09	12.97 to 38.51	4.7
KRR		0.88	0.70 ± 0.08		2.7
SVR		0.78	0.61 ± 0.07		2.4
2-Layer NN		0.68	0.87 ± 0.08		3.4
4-Layer NN		0.87	0.98 ± 0.08		3.8

Table S6 - Table showing the ds3 ML results. The table includes Pre-ML AM1-DFT MAE, average train MAEs, average test MAEs, average test set target range, and the average test MAE value as a percentage of the test set target range. All averaged results are over 5 random seeds (22, 23, 14, 1, and 2).

Model	Pre-ML AM1-DFT MAE (kcal mol^{-1})	Average Train MAE (kcal mol^{-1})	Average Test MAE (kcal mol^{-1})	Average Test Target Range (kcal mol^{-1})	Test MAE as % of Test Set Range
Dipole Distortion Energy					
Ridge	3.59	3.22	3.27 ± 0.20	0.98 to 41.25	8.1
KRR		2.93	2.95 ± 0.15		7.3
SVR		2.65	2.55 ± 0.13		6.3
2-Layer NN		2.25	2.67 ± 0.13		6.6
4-Layer NN		2.76	2.91 ± 0.14		7.2
Dipolarophile Distortion Energy					
Ridge	3.81	3.17	3.08 ± 0.20	0.13 to 35.33	8.8
KRR		2.86	2.71 ± 0.13		7.7
SVR		2.54	2.37 ± 0.12		6.7
2-Layer NN		2.06	2.50 ± 0.13		7.1
4-Layer NN		2.66	2.77 ± 0.15		7.9
Interaction Energy					
Ridge	20.01	3.11	3.18 ± 0.22	-43.62 to -7.92	8.9
KRR		2.87	2.83 ± 0.14		7.9
SVR		2.47	2.46 ± 0.12		6.9
2-Layer NN		2.01	2.60 ± 0.13		7.3
4-Layer NN		2.65	2.83 ± 0.13		7.9
ΔG‡					
Ridge	22.97	3.74	3.69 ± 0.21	2.16 to 54.50	7.1
KRR		3.65	3.52 ± 0.17		6.7
SVR		3.12	3.02 ± 0.14		5.8
2-Layer NN		2.59	3.19 ± 0.15		6.1
4-Layer NN		2.95	3.20 ± 0.16		6.1
ΔE‡					
Ridge	23.07	3.86	3.81 ± 0.20	-13.46 to 37.56	7.5
KRR		3.53	3.39 ± 0.15		6.6
SVR		3.21	3.09 ± 0.14		6.1
2-Layer NN		2.72	3.34 ± 0.15		6.5
4-Layer NN		3.35	3.41 ± 0.16		6.7

Table S7 - Table showing the ds4 ML results. The table includes Pre-ML AM1-DFT MAE, average train MAEs, average test MAEs, average test set target range, and the average test MAE value as a percentage of the test set target range. All averaged results are over 5 random seeds (22, 23, 14, 1, and 2).

Model	Pre-ML AM1-DFT MAE (kcal mol^{-1})	Average Train MAE (kcal mol^{-1})	Average Test MAE (kcal mol^{-1})	Average Test Target Range (kcal mol^{-1})	Test MAE as % of Test Set Range
Nucleophile (dimethyl malonate) Distortion Energy					
Ridge	3.70	0.61	0.65 ± 0.06	3.43 to 13.98	6.2
KRR		0.68	0.69 ± 0.07		6.5
SVR		0.54	0.54 ± 0.06		5.1
2-Layer NN		0.32	0.55 ± 0.06		5.2
4-Layer NN		0.65	0.70 ± 0.07		6.6
Michael Acceptor Distortion Energy					
Ridge	4.07	1.92	1.85 ± 0.16	7.30 to 39.64	5.7
KRR		2.04	1.89 ± 0.21		5.8
SVR		1.45	1.27 ± 0.17		3.9
2-Layer NN		0.98	1.42 ± 0.16		4.4
4-Layer NN		1.76	1.75 ± 0.17		5.4
Interaction Energy					
Ridge	10.83	1.74	1.67 ± 0.17	-26.78 to -5.05	7.7
KRR		1.67	1.58 ± 0.19		7.3
SVR		1.34	1.23 ± 0.17		5.7
2-Layer NN		0.76	1.21 ± 0.17		5.6
4-Layer NN		1.12	1.30 ± 0.18		6.0
ΔG‡					
Ridge	10.46	1.45	1.30 ± 0.13	15.37 to 51.23	3.6
KRR		1.48	1.33 ± 0.16		3.7
SVR		1.09	0.97 ± 0.13		2.7
2-Layer NN		0.74	1.02 ± 0.12		2.8
4-Layer NN		0.88	1.05 ± 0.12		2.9
ΔE‡					
Ridge	10.93	1.45	1.32 ± 0.13	1.08 to 35.34	3.9
KRR		1.15	1.04 ± 0.14		3.0
SVR		1.03	0.90 ± 0.13		2.6
2-Layer NN		0.69	0.96 ± 0.11		2.8
4-Layer NN		0.82	1.01 ± 0.12		2.9

Table S8 - Table showing the average test and train MAEs for ds2. This table shows the comparison of random forest to other models presented in this work. All averaged results are over 5 random seeds (22, 23, 14, 1, and 2).

Model	Energy Predictions / kcal mol^{-1}					
	Diene Distortion		Dienophile Distortion		Interaction	
	Test MAE	Train MAE	Test MAE	Train MAE	Test MAE	Train MAE
Random Forest	0.79	0.84	0.70	0.79	0.79	0.84
Ridge	0.61	0.76	0.77	0.80	1.00	1.07
KRR	0.59	0.73	0.38	0.54	0.50	0.59
SVR	0.33	0.39	0.34	0.44	0.50	0.54
2 Layer NN	0.48	0.42	0.51	0.39	0.58	0.41
4 Layer NN	0.49	0.50	0.92	0.83	0.75	0.64

4.4. Literature ML Predictions

For ds2, we wanted to evaluate model performance on unseen examples from the literature to investigate their generalisability when predicting distortion/interaction energies. In Table S5, SVR consistently performs well across all ds2 targets therefore, these models were chosen to test performance on the literature sets.

Table S9 – Literature external test sets for SVR models trained on ds2. Predictions are on the dienophile and interaction energies.

Dataset		Pre-ML AM1-DFT MAE (kcal mol^{-1})	Average SVR Test MAE (kcal mol^{-1})
Cycloalkenones	Dienophile	1.93	1.58
	Interaction	12.71	1.52
Cyclopropene	Dienophile	4.27	1.62
	Interaction	6.73	1.42

4.5. Learning Curves

Figures S25 - S29 show the average learning curves for SVR models for ds1. All graphs show the training (red) and test (blue) metrics to help highlight any cases of overfitting. All averages are over five random states (22, 23, 14, 1, 2)

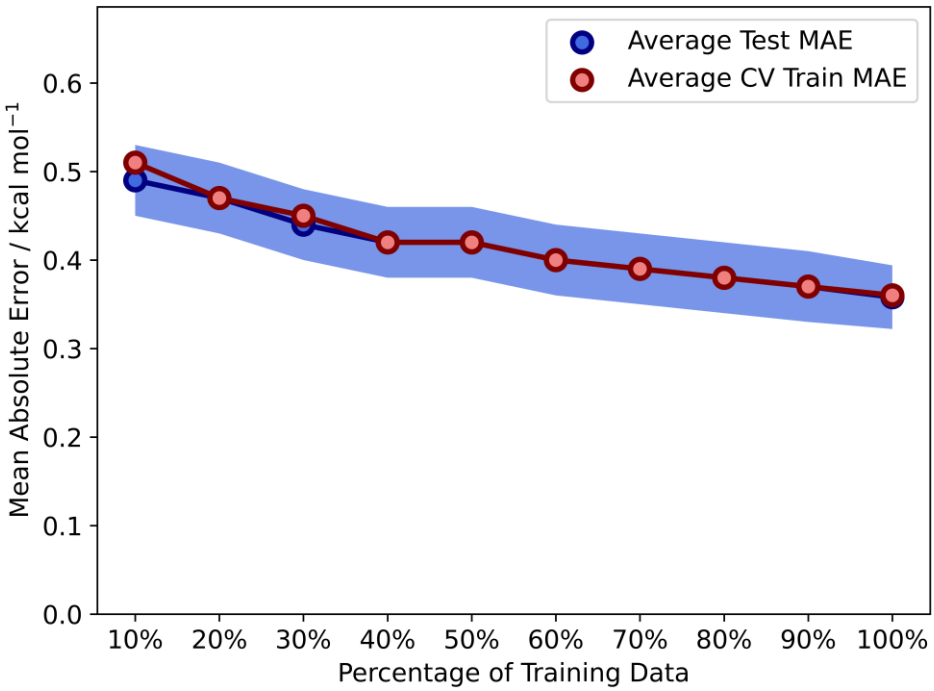

Fig. S25 – Learning curves for prediction of the nitromethane-derived nucleophile distortion energy for ds1. Blue and red indicate the average test and train MAEs, respectively. The blue region is the average standard error.

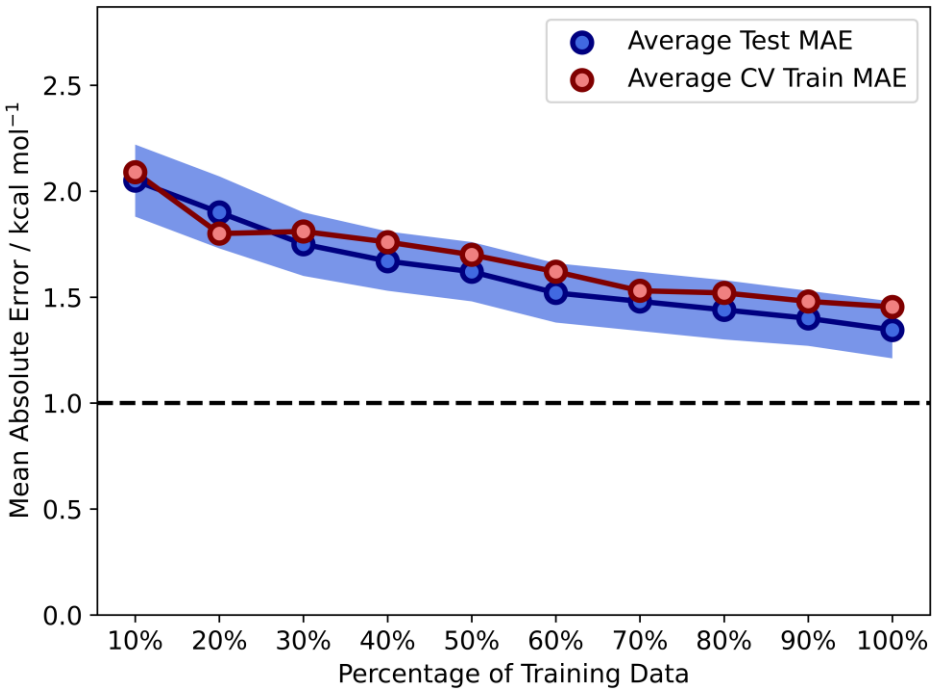

Fig. S26 - Learning curves for prediction of the Michael acceptor distortion energy for ds1. Blue and red indicate the average test and train MAEs, respectively. The blue region is the average standard error.

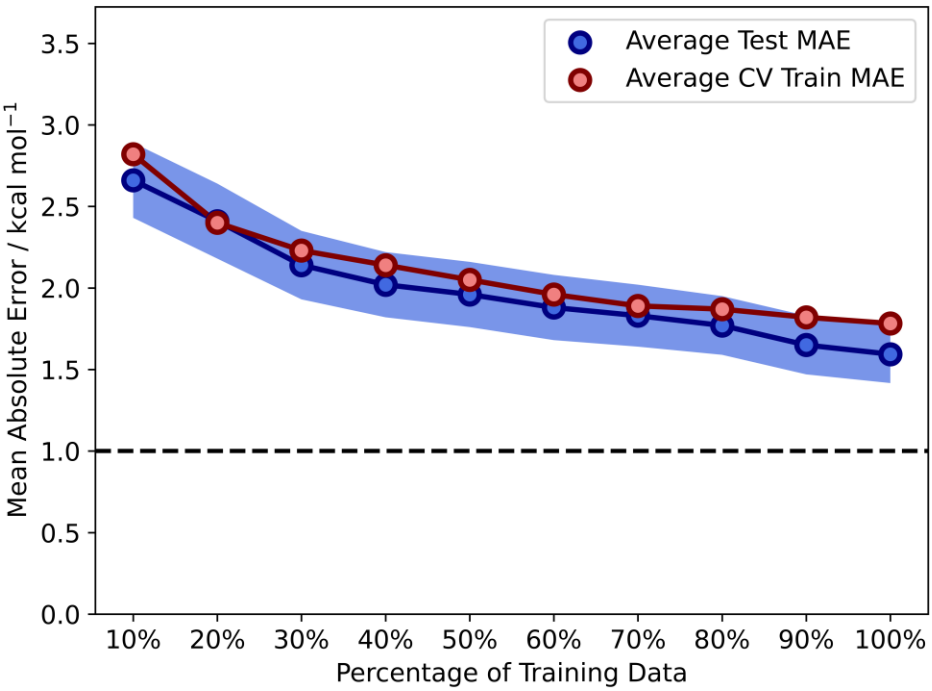

Fig. S27 - Learning curves for prediction of the interaction energy for ds1. Blue and red indicate the average test and train MAEs, respectively. The blue region is the average standard error.

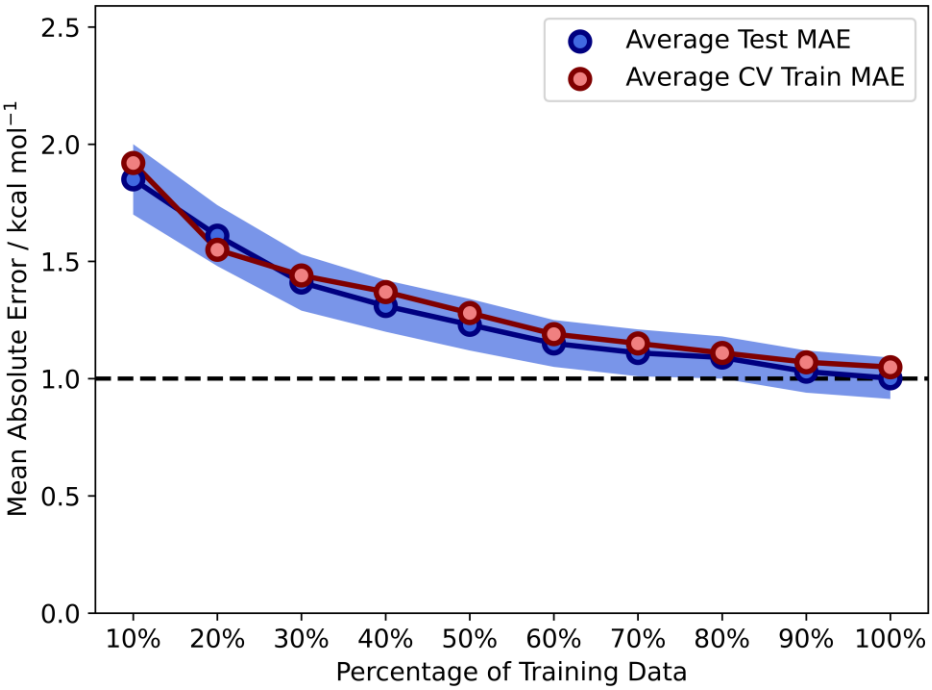

Fig. S28 - Learning curves for prediction of $\Delta G^{\ddagger}$ for ds1. Blue and red indicate the average test and train MAEs, respectively. The blue region is the average standard error.

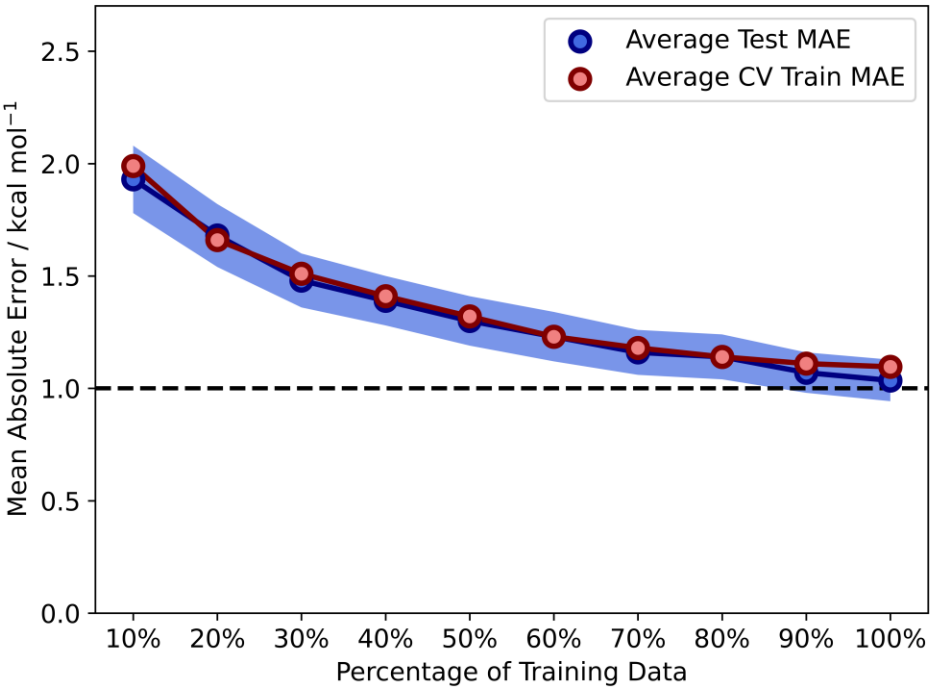

Fig. S29 - Learning curves for prediction of $\Delta E^{\ddagger}$ for ds1. Blue and red indicate the average test and train MAEs, respectively. The blue region is the average standard error.

Figures S30 - S34 show the average learning curves for SVR models for ds2.

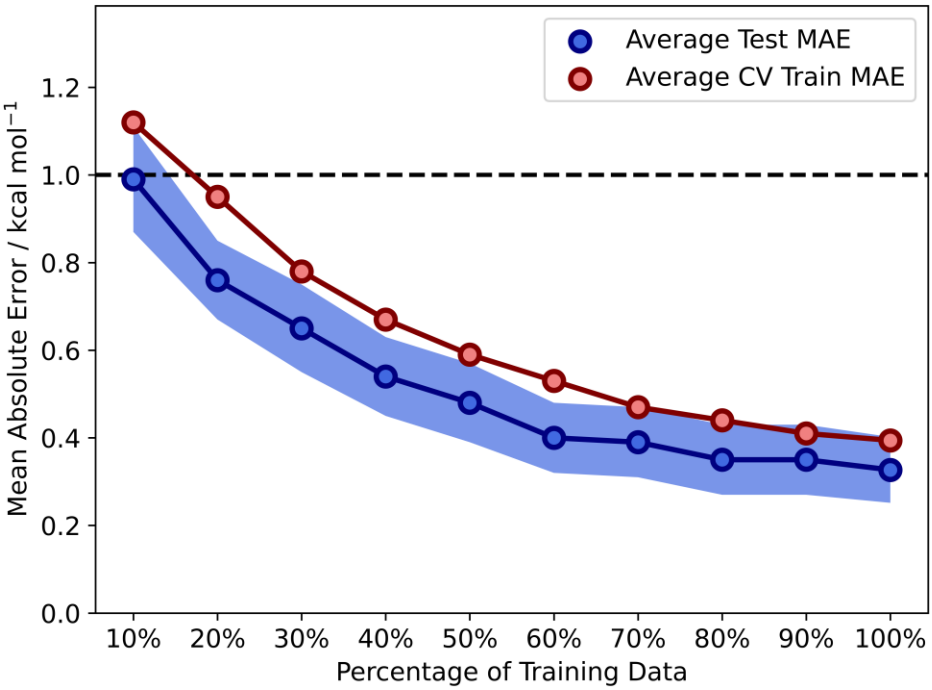

Fig. S30 - Learning curves for prediction of the diene distortion energy for ds2. Blue and red indicate the average test and train MAEs, respectively. The blue region is the average standard error.

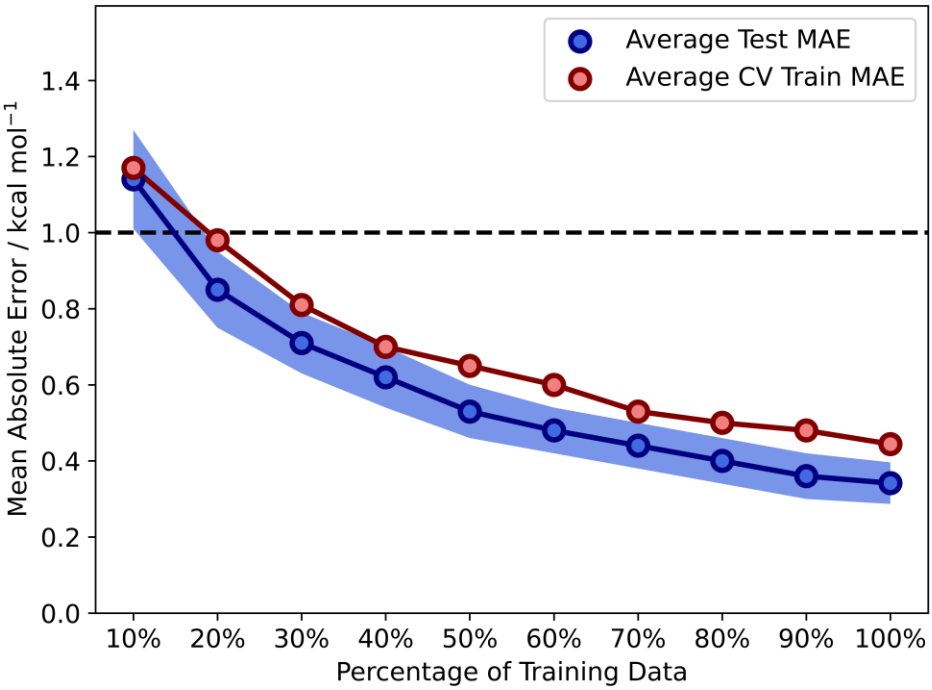

Fig. S31 - Learning curves for prediction of the dienophile distortion energy for ds2. Blue and red indicate the average test and train MAEs, respectively. The blue region is the average standard error.

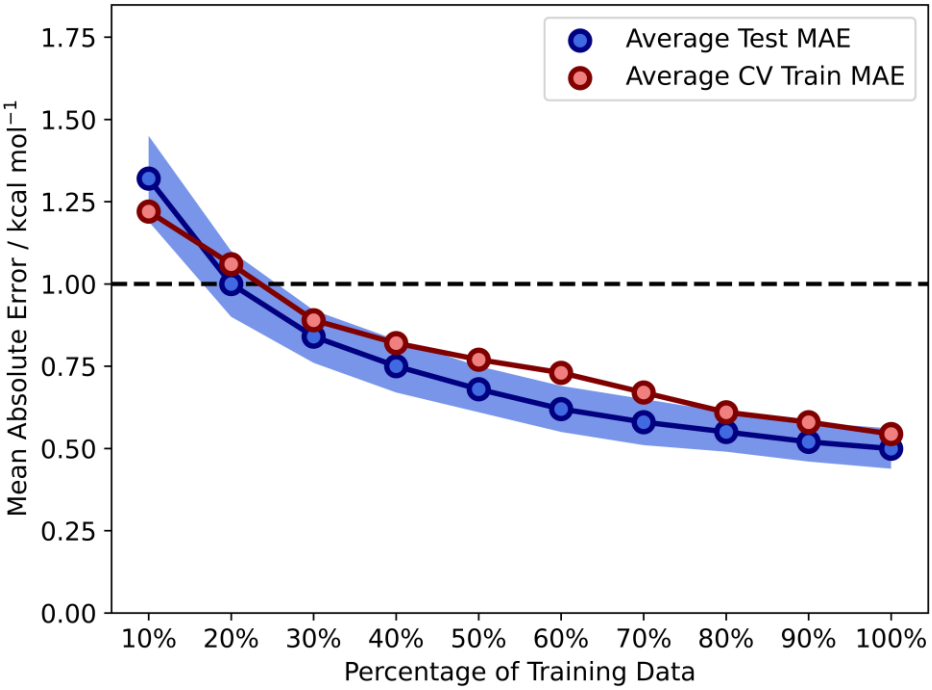

Fig. S32 - Learning curves for prediction of the interaction energy for ds2. Blue and red indicate the average test and train MAEs, respectively. The blue region is the average standard error.

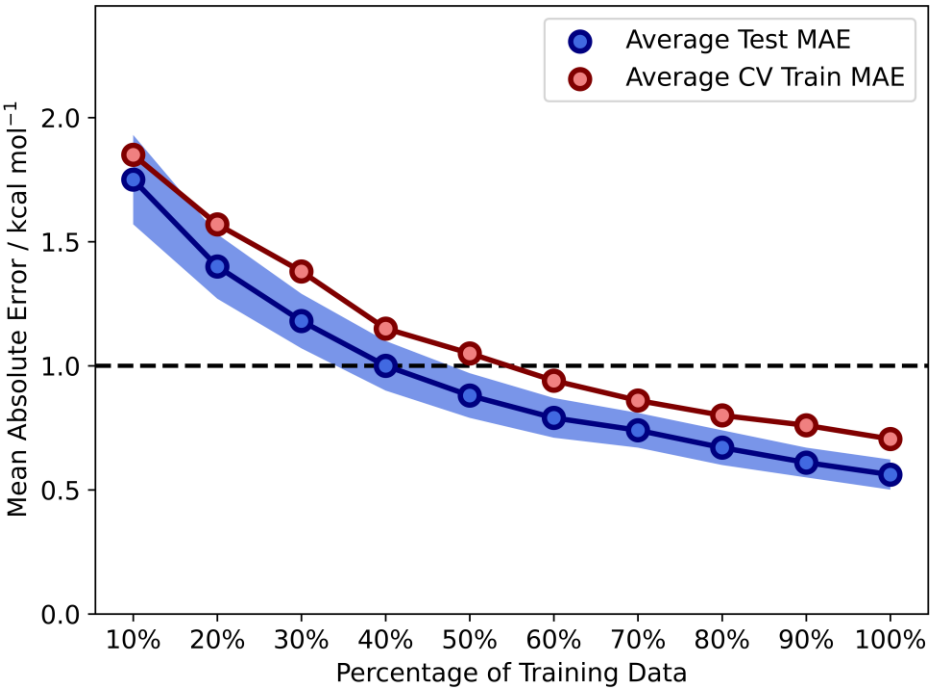

Fig. S33 - Learning curves for prediction of $\Delta G^{\ddagger}$ for ds2. Blue and red indicate the average test and train MAEs, respectively. The blue region is the average standard error.

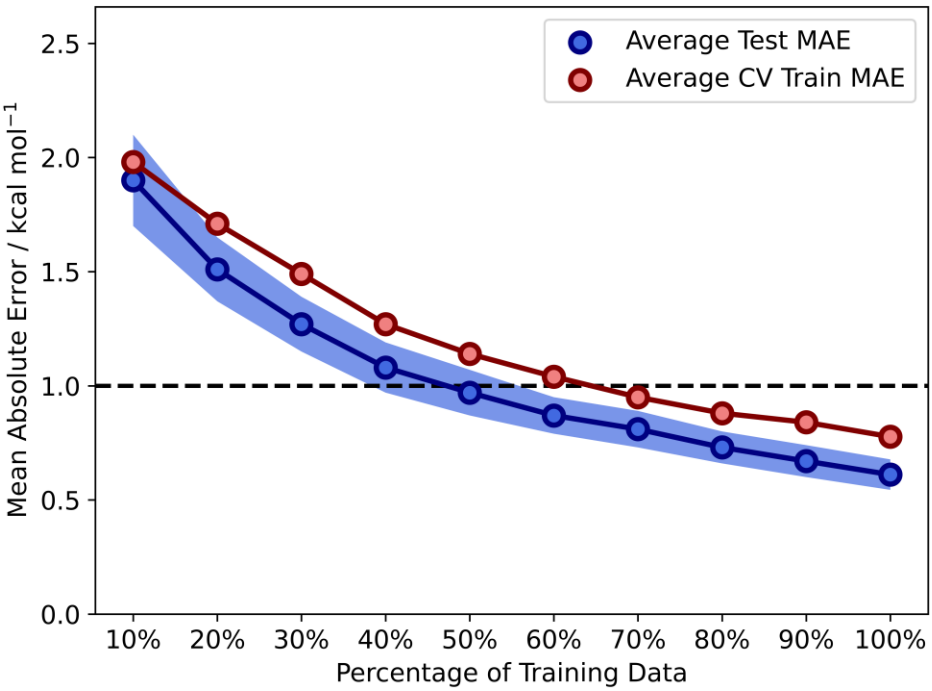

Fig. S34 - Learning curves for prediction of $\Delta E^{\ddagger}$ for ds2. Blue and red indicate the average test and train MAEs, respectively. The blue region is the average standard error.

Figures S35 - S39 show the average learning curves for SVR models for ds3.

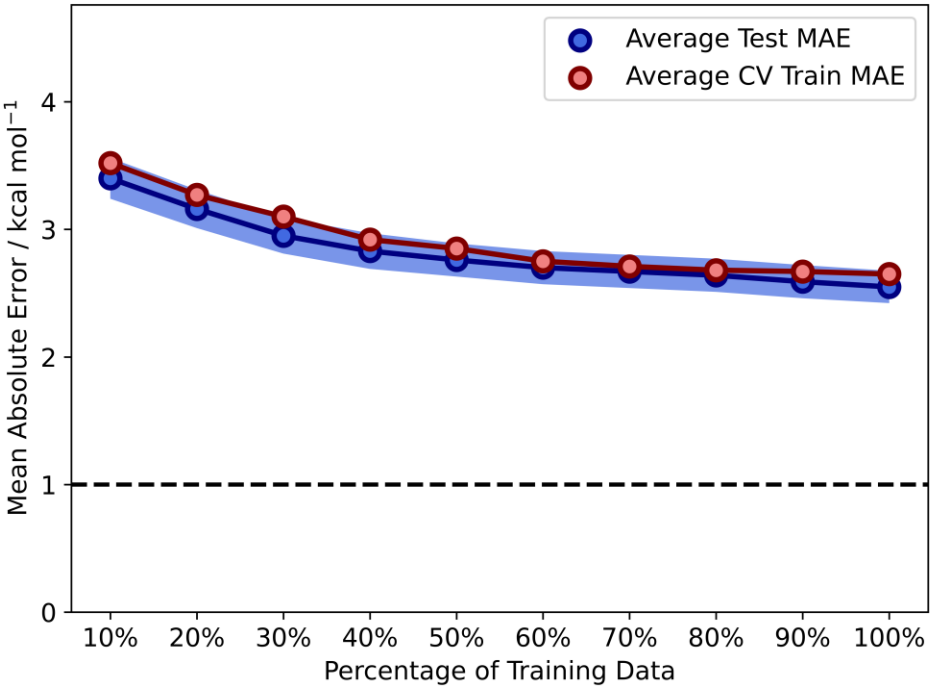

Fig. S35 - Learning curves for prediction of the dipole distortion energy for ds3. Blue and red indicate the average test and train MAEs, respectively. The blue region is the average standard error.

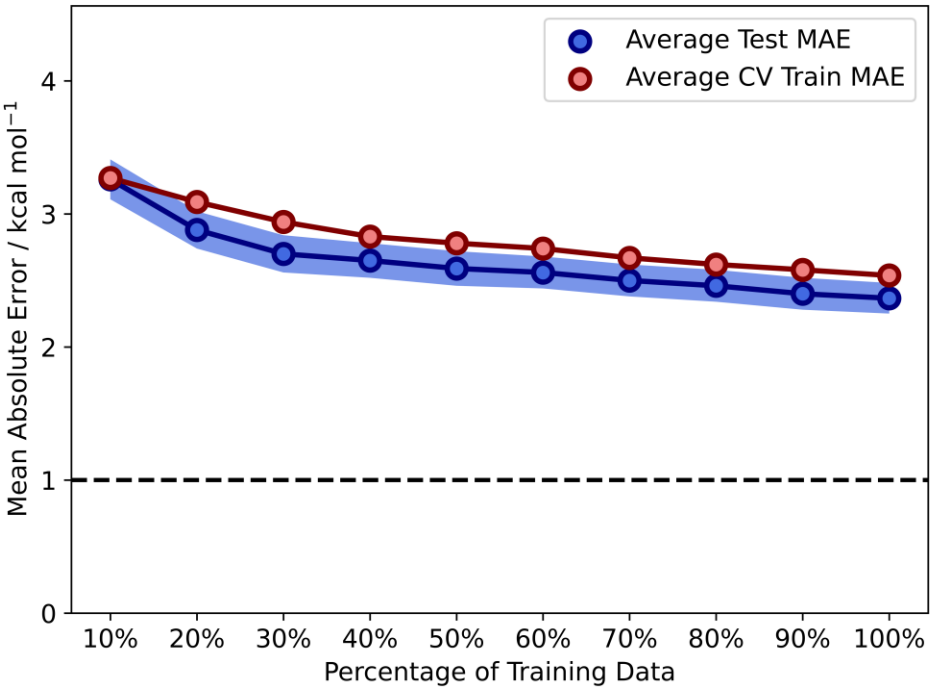

Fig. S36 - Learning curves for prediction of the dipolarophile distortion energy for ds3. Blue and red indicate the average test and train MAEs, respectively. The blue region is the average standard error.

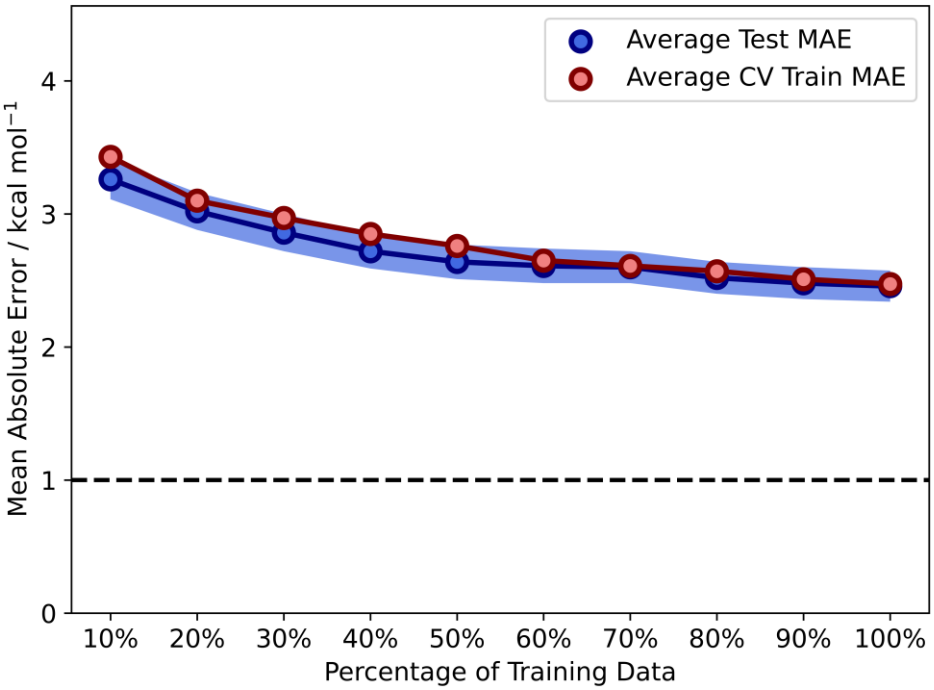

Fig. S37 - Learning curves for prediction of the interaction energy for ds3. Blue and red indicate the average test and train MAEs, respectively. The blue region is the average standard error.

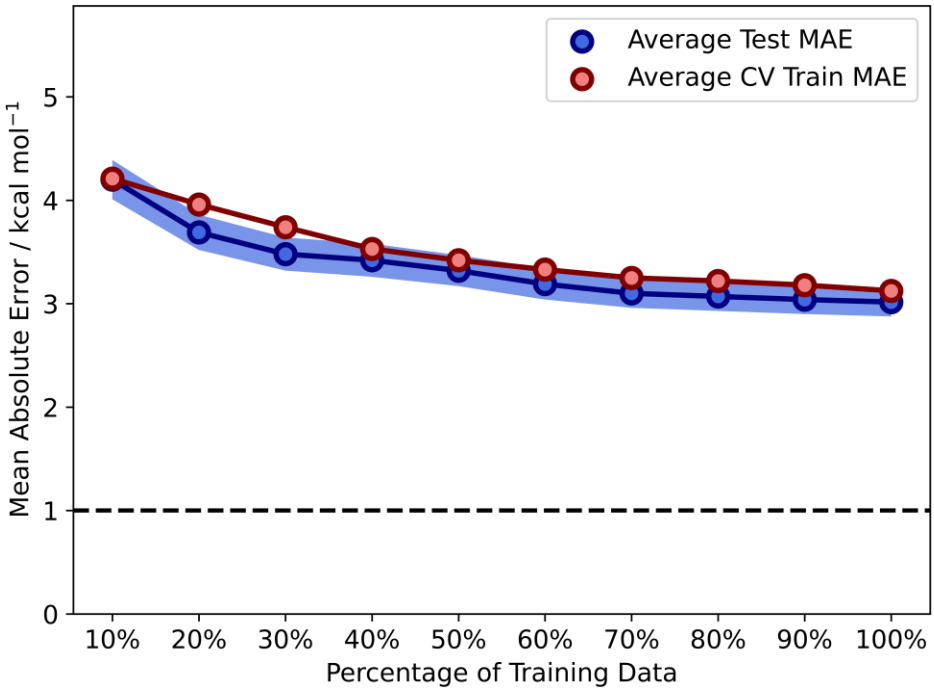

Fig. S38 - Learning curves for prediction of $\Delta G^{\ddagger}$ for ds3. Blue and red indicate the average test and train MAEs, respectively. The blue region is the average standard error.

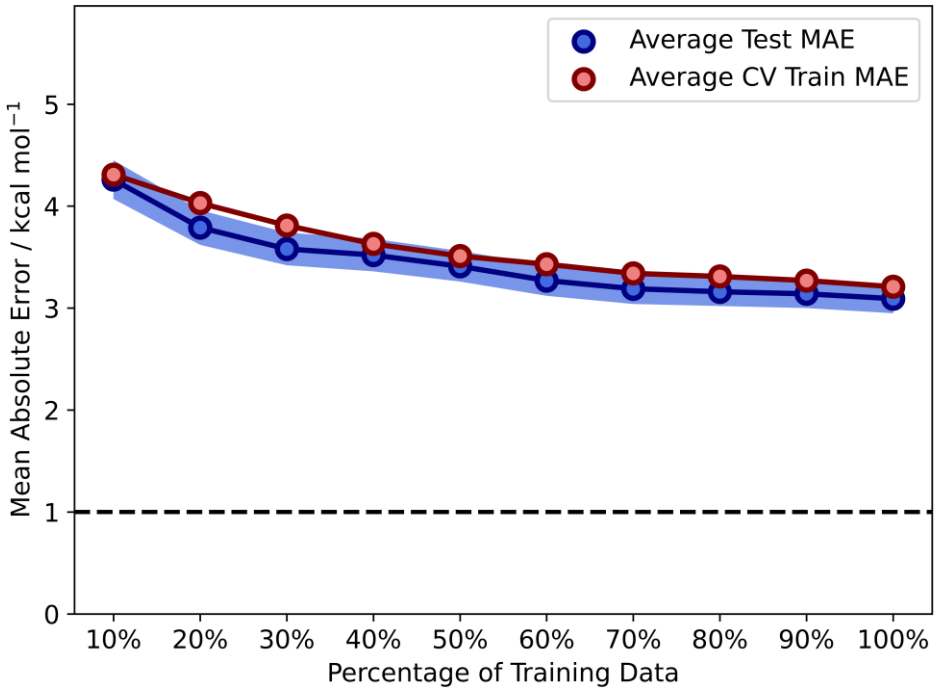

Fig. S39 - Learning curves for prediction of $\Delta E^{\ddagger}$ for ds3. Blue and red indicate the average test and train MAEs, respectively. The blue region is the average standard error.

Figures S40 - S44 show the average learning curves for SVR models for ds4.

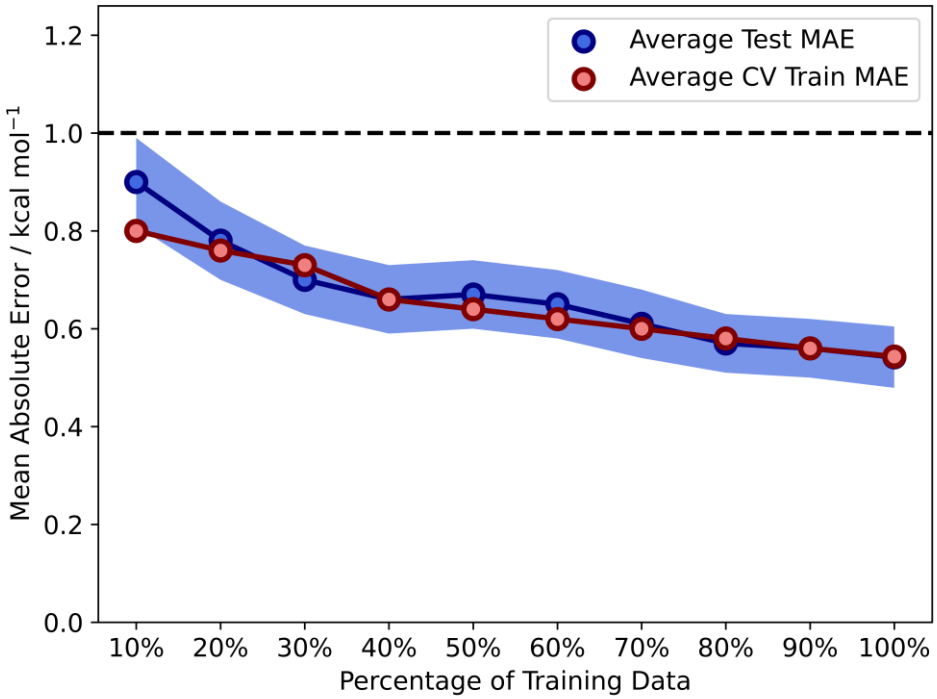

Fig. S40 - Learning curves for prediction of the dimethyl malonate nucleophile distortion energy for ds4. Blue and red indicate the average test and train MAEs, respectively. The blue region is the average standard error.

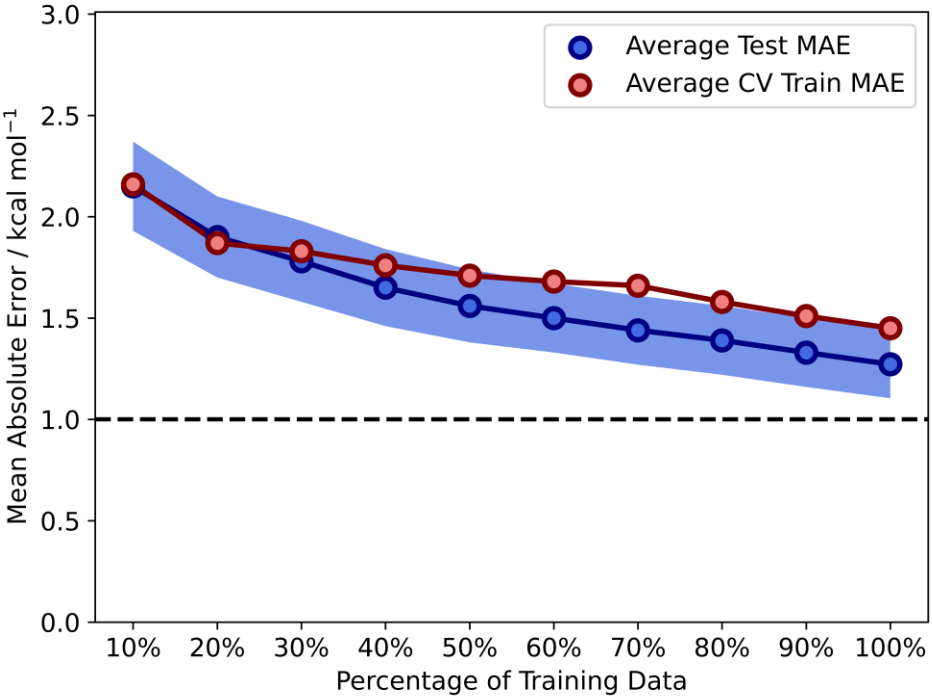

Fig. S41 - Learning curves for prediction of the Michael acceptor distortion energy for ds4. Blue and red indicate the average test and train MAEs, respectively. The blue region is the average standard error.

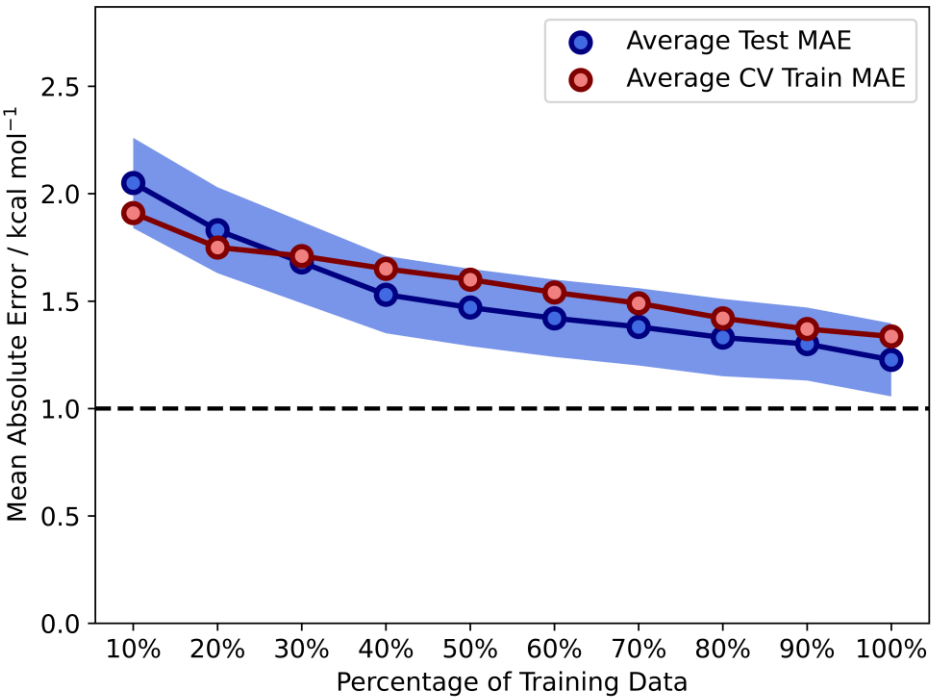

Fig. S42 - Learning curves for prediction of the interaction energy for ds4. Blue and red indicate the average test and train MAEs, respectively. The blue region is the average standard error.

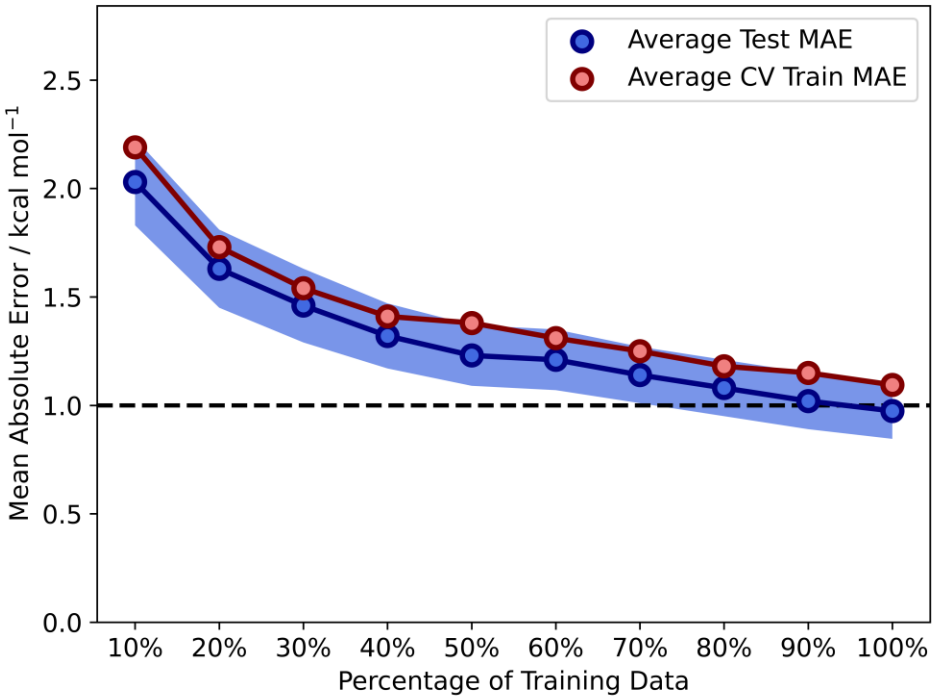

Fig. S43 - Learning curves for prediction of $\Delta G^{\ddagger}$ for ds4. Blue and red indicate the average test and train MAEs, respectively. The blue region is the average standard error.

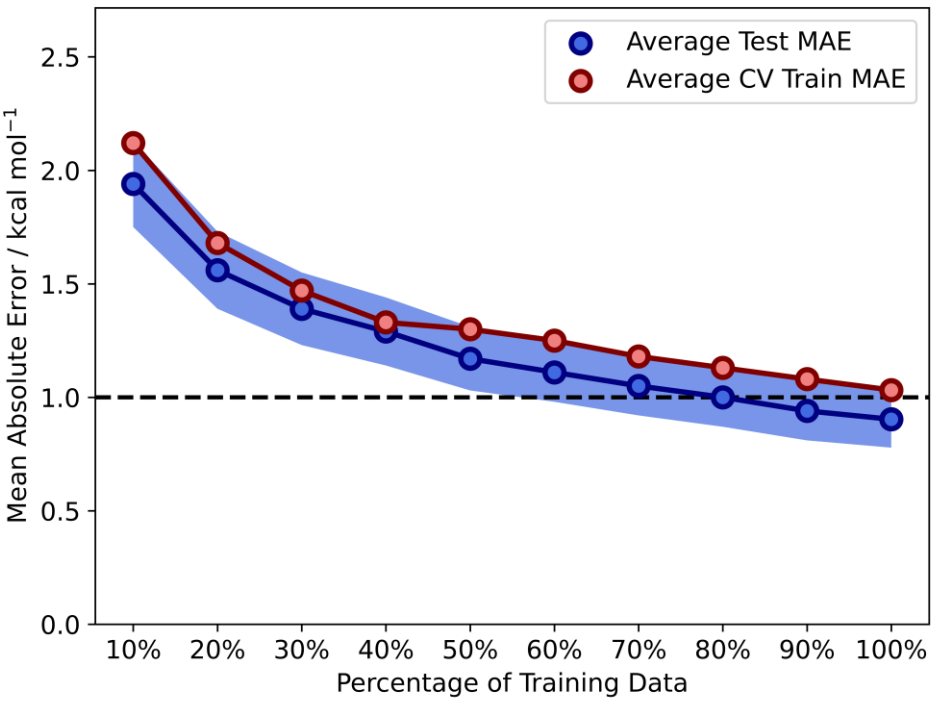

Fig. S44- Learning curves for prediction of $\Delta E^{\ddagger}$ for ds4. Blue and red indicate the average test and train MAEs, respectively. The blue region is the average standard error.

4.6. Feature Importances

In an attempt to understand model performance further, feature importances were calculated for SVR models built on ds2. The method for this was to take the tuned and trained model and randomly shuffle one feature in the test set prior to testing. This was repeated for each feature to evaluate the importance of each individual feature relative to the target. Figure S45 explains this graphically while the results for SVR models trained on ds2 are shown in Figures S46-48.

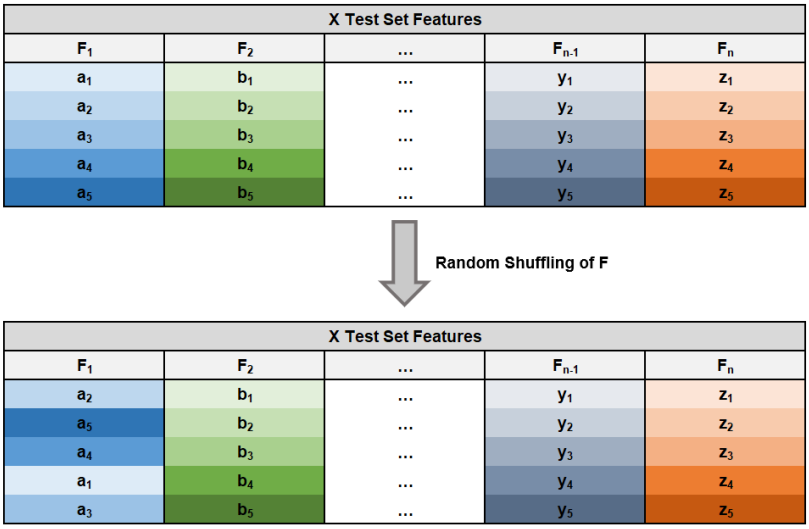

Figure S45 - Graphical explanation for how feature shuffling was performed. Each feature was individually randomly shuffled to create a new X test set that is used to evaluate how important that feature is relative to the target.

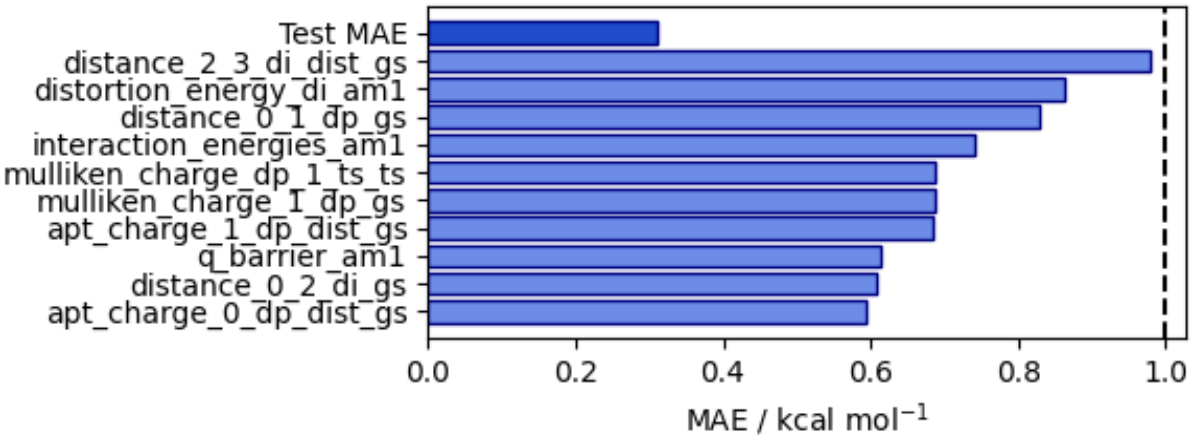

Figure S46 – Feature Importances for SVR model trained on ds2 for the prediction of DFT diene distortion energy. Displayed are the 10 features which, when randomly shuffled at inference, result in the largest test set MAE.

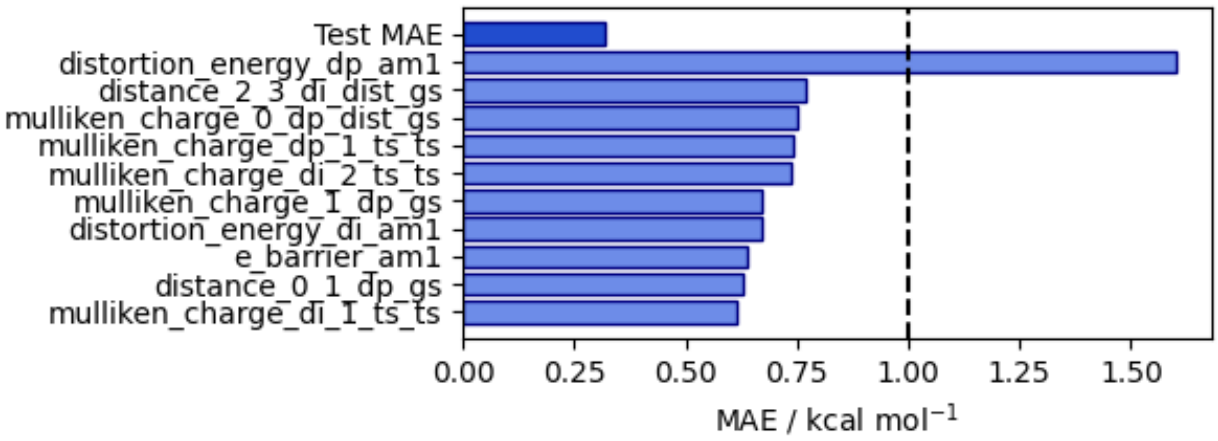

Figure S47 - Feature Importances for SVR model trained on ds2 for the prediction of DFT dienophile distortion energy. Displayed are the 10 features which, when randomly shuffled at inference, result in the largest test set MAE.

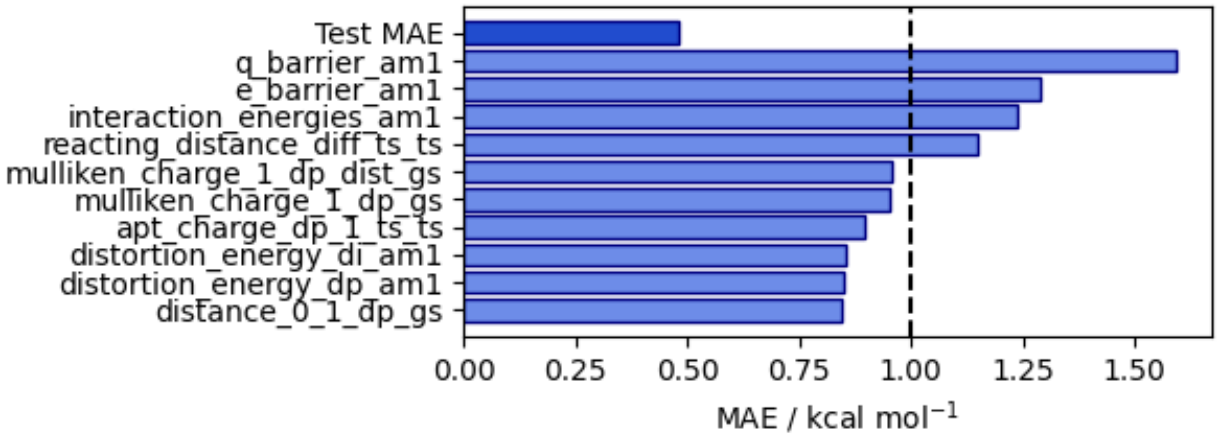

Figure S48 - Feature Importances for SVR model trained on ds2 for the prediction of DFT interaction energy. Displayed are the 10 features which, when randomly shuffled at inference, result in the largest test set MAE.

5. References

1 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, J. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, J. C. A. Rendell, S. Burant, S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, O. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski and D. J. Fox, *Gaussian 16, Revision A.03*, Gaussian, Inc., Wallingford, CT, 2016.

2 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, J. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, J. C. A. Rendell, S. Burant, S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, O. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski and D. J. Fox, *Gaussian 16, Revision C.01*, Gaussian, Inc., Wallingford, CT, 2016.

3 C. Legault, *CYLview20*, Université de Sherbrooke, 2020.

4 E. H. E. Farrar and M. N. Grayson, *Chem. Sci.*, 2022, **13**, 7594–7603.

5 B. Mennucci, R. Cammi and J. Tomasi, *J. Chem. Phys.*, 1998, **109**, 2798–2807.

6 S. G. Espley, E. H. E. Farrar, D. Buttar, S. Tomasi and M. N. Grayson, *Digit. Discov.*, 2023, **2**, 941–951.

7 T. Stuyver, K. Jorner and C. W. Coley, *Sci. Data*, 2023, **10**, 1–14.

8 Maestro Schrödinger, *Schrödinger Release 2018-2*, LLC, New York, 2018.

9 MacroModel Schrödinger, *Schrödinger Release 2018-2*, LLC, New York, 2018.

10 K. Roos, C. Wu, W. Damm, M. Reboul, J. M. Stevenson, C. Lu, M. K. Dahlgren, S. Mondal, W. Chen, L. Wang, R. Abel, R. A. Friesner and E. D. Harder, *J. Chem. Theory Comput.*, 2019, **15**, 1863–1874.

11 M. J. S. Dewar, E. G. Zoebisch, E. F. Healy and J. J. P. Stewart, *J. Am. Chem. Soc.*, 1985, **107**, 3902–3909.

12 J.-D. Chai and M. Head-Gordon, *Phys. Chem. Chem. Phys.*, 2008, **10**, 6615.

13 F. Weigend and R. Ahlrichs, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3297–3305.

14 R. S. Paton, S. Kim, A. G. Ross, S. J. Danishefsky and K. N. Houk, *Angew. Chem. Int. Ed.*, 2011, **50**, 10366–10368.

15 B. J. Levandowski and K. N. Houk, *J. Am. Chem. Soc.*, 2016, **138**, 16731–16736.

16 G. Luchini, J. V Alegre-Requena, I. Funes-Ardoiz and R. S. Paton, *F1000Research* , 2020, 9, 291.

17 N. M. O'Boyle, A. L. Tenderholt and K. M. Langner, *J. Comput. Chem.*, 2008, **29**, 839–845.

18 K. Jorner and L. Turcani, Morfeus, Zurich, 2022.

19 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

20 M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, R. Jozefowicz, Y. Jia, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, M. Schuster, R. Monga, S. Moore, D. Murray, C. Olah, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu and X. Zheng, *TensorFlow: Large-scale machine learning on heterogeneous systems*, 2022.

21 L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh and A. Talwalkar, *J. Mach. Learn. Res.*, 2018, **18**, 1–52.