Machine learning reaction barriers in low data regimes: A horizontal and diagonal transfer learning approach

Supporting Information

Samuel G. Espley,[a] Elliot H. E. Farrar, [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.
b. Data Science and Modelling, Pharmaceutical Sciences, R&D, AstraZeneca, Macclesfield, UK.
c. Chemical Development, Pharmaceutical Technology & Development, Operations, AstraZeneca, Macclesfield, UK.

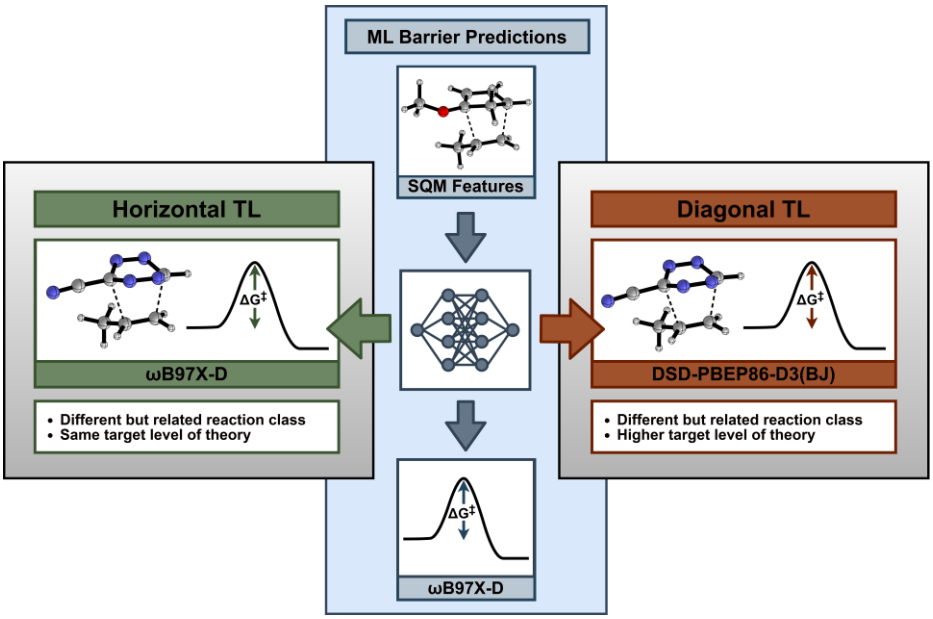

Table of Contents

1. Dataset Generation

Preliminary geometries were constructed for 1355 endo/exo and 414 tetrazine/alkyne transition states by altering chosen Diels-Alder backbones based on literature examples with a variety of functional groups.[1–13] These structures were generated using the Custom R-Group Enumeration in Schrödinger's MacroModel (version 12.7).[14]

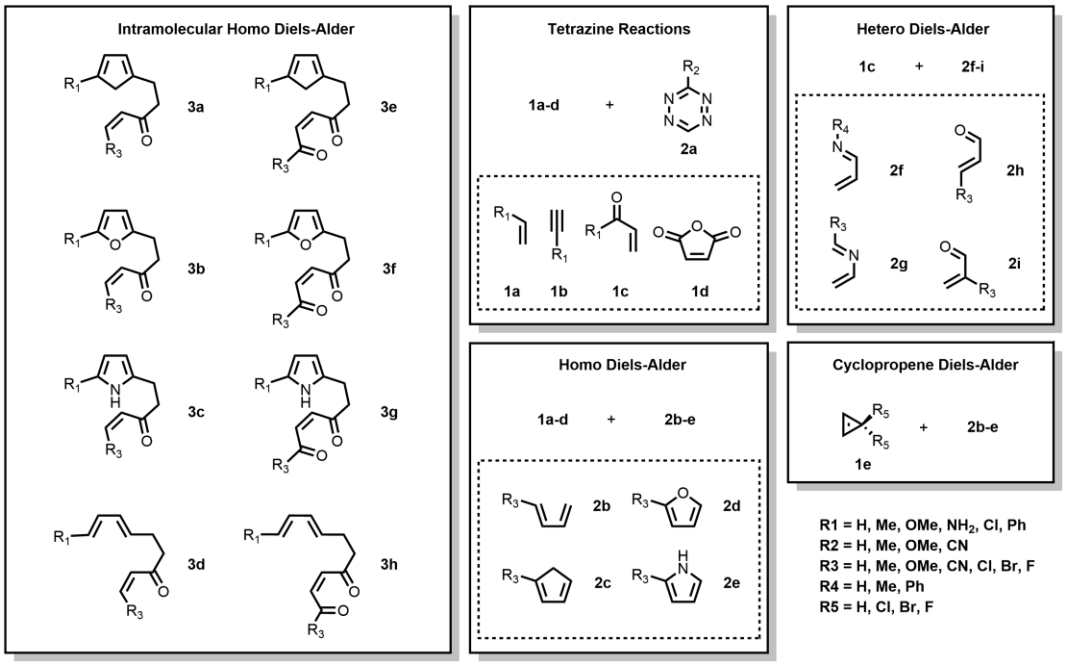

Fig. S1 - Overview of the enumerations made to create the Diels-Alder dataset.

These enumerated reactant and transition state structures (Fig. S1) were then conformationally searched using Schrödinger's MacroModel (version 12.7)[14,15] with the OPLS3e forcefield.[16] The lowest energy OPLS3e conformers for all reactant and transition state structures were then optimised with AM1[17], PM3[18], and ωB97X-D/def2-TZVP[19,20] using Gaussian 16 (Revision A.03 and C.01).[21,22] All tetrazine Diels-Alder reactions were also optimised with DSD-PBEP86-D3(BJ)/def2-TZVP[23] for diagonal transfer learning. All reactant structures optimised to minima whilst the number of optimised concerted transition state structures for each level of theory (dataset size) ranged from 1065 – 1141 structures (Table S1). Free energy reaction barriers were calculated from temperature (298.15 K) and concentration corrected (1 mol l^{-1}) quasi-harmonic free energies obtained with GoodVibes (Table S1).[24]

Dataset		Barrier Range / kcal mol^{-1}	Dataset Size (with DFT)
AM1	Endo	30.22 - 59.75	1065
	Exo	28.13 - 53.07	1109
PM3	Endo	32.3 - 56.33	1141
	Exo	31.6 - 55.99	1141
DFT	Endo	12.76 - 54.53	-
	Exo	9.4 - 50.45	-

Table S1 - Transition state barrier ranges and dataset sizes for all levels of theory when combined with DFT calculations. Dataset size is for the combined X-DFT dataset where X is either AM1 or PM3.

The baseline methods of AM1 and PM3 were chosen based on their prevalence and usage within the literature to investigate the Diels-Alder reaction.[25–28] The newer methods of PM6[29] and PM7[30] were also investigated however both exhibited issues in reaching convergence for concerted Diels-Alder reactions. For the target level of theory, the ωB97X-D[19] functional was chosen alongside the polarised triple-ζ valence quality (def2-TZVP) basis set[20] based on their performance in barrier height calculations[31,32] and previous work within this area utilising this combination.[33]

To ensure a consistent dataset of Diels-Alder transition states, only those close to a concerted mechanism were used. All transition state structures with a difference between the C1–C2 and C3–C4 bond forming distances (Fig. S2) of greater than 0.6 Å at the AM1 and PM3 levels of theory were removed; distance value calculated by equation (1).

$$\Delta Distance = |Distance_{C1-C2} - Distance_{C3-C4}| \quad (1)$$

All Gaussian 16 computed output files are publicly available in *Dataset for "Machine learning reaction barriers in low data regimes: A horizontal and diagonal transfer learning approach"* in the University of Bath Research Data Archive (accessible at: https://doi.org/10.15125/BATH-01229). All structures visualised within this work were created using CYLView.[34]

2. Feature Extraction

A number of physical organic chemical features were extracted for each Diels-Alder dataset at the AM1 and PM3 levels of theory utilising a select group of python packages (Table S2-3). Features were extracted for the core atoms in both the diene and dienophile reactant structures as well as for the associated transition state structure. Fig. S2 provides an example of an enumeration and of the common atoms (highlighted) that features were extracted for. With all models, the feature processing was consistently performed solely on the training sets before applying the same transformation to the test sets to prevent data leakage. All train-test splits were performed using the same random state (23), that was chosen at random, to ensure uniformity across splits.

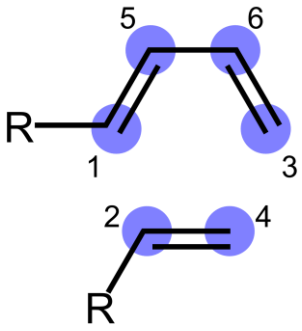

Fig. S2 - General Diels-Alder Reaction and atom numbering. Features were extracted for core atoms (highlighted blue).

Feature	Description	Source
atomcharges_apt_01_ts	APT Atomic charge for the transition state structure for each atom n (range 01-06).	CCLIB[35]
atomcharges_apt_sum_01_ts	APT Summed atomic charge for the transition state structure for each atom n (range 01-06).	CCLIB[35]
atomcharges_mulliken_01_ts	Mulliken atomic charge for the transition state structure for each atom n (range 01-06).	CCLIB[35]
atomcharges_mulliken_sum_01_ts	Mulliken summed atomic charge for the transition state structure for each atom n (range 01-06).	CCLIB[35]
homoenergies_ts	Highest occupied molecular orbital (HOMO) energy - transition state.	CCLIB[35]
lumoenergies_ts	Lowest unoccupied molecular orbital (LUMO) energy - transition state.	CCLIB[35]
hardness_ts	Global hardness - transition state.	HSAB[36]
softness_ts	Global softness - transition state.	HSAB[36]
chemicalpotential_ts	Global chemical potential - transition state.	HSAB[36]
electrophilicity_ts	Global electrophilicity - transition state.	HSAB[36]
gv_E_ts	Electronic energy.	GoodVibes[24]
gv_ZPE_ts	Zero-point energy.	GoodVibes[24]
gv_H_ts	Enthalpy.	GoodVibes[24]
gv_T.S_ts	Entropy.	GoodVibes[24]
gv_T.qh-S_ts	Quasi-harmonic entropy.	GoodVibes[24]
gv_G(T)_ts	Gibbs free energy.	GoodVibes[24]
gv_qh-G(T)_ts	Quasi-harmonic Gibbs free energy.	GoodVibes[24]
ea_ts	Semi-Empirical Quasi-harmonic free energy reaction barrier.	GoodVibes[24]

sasa_1_ts	Solvent accessible surface area for each atom n (range 1-6) - transition state.	Freesasa[37]
sasa_R1_ts	Solvent accessible surface area for R1 atom - transition state.	Freesasa[37]
sasa_total_ts	Total Solvent accessible surface area for core atoms - transition state.	Freesasa[37]
sterimol_R1_L_ts	Sterimol L parameter for R1 substituent - transition state.	DBStep[38]
sterimol_R1_Bmin_ts	Sterimol B_{min} parameter for R1 substituent - transition state.	DBStep[38]
sterimol_R1_Bmax_ts	Sterimol B_{max} parameter for R1 substituent - transition state.	DBStep[38]
PBV_1_ts	Percent buried volume (3.5 Å) for each reacting atom n (range 1-4) - transition state.	DBStep[38]
HBA2_ts	Number of hydrogen bond acceptors - transition state.	Pybel[39]
HBD_ts	Number of hydrogen bond donors - transition state.	Pybel[39]
nF_ts	Number of Fluorine atoms - transition state.	Pybel[39]
bond_forming_distance_1_ts	Transition state bond forming distance between atoms 1 and 2.	-
bond_forming_distance_2_ts	Transition state bond forming distance between atoms 3 and 4.	-
bond_forming_angle_1_ts	Transition state bond forming angle between atoms 2, 1, and 3.	-
bond_forming_angle_2_ts	Transition state bond forming angle between atoms 4, 3, and 1.	-
bond_form_diff_ts	Transition state difference between the two bond forming distances.	-
bond_ang_diff_ts	Transition state difference between the two bond forming angles.	-
atomcharges_apt_01_di	APT Atomic charge for the diene reactant structure for each atom n (01, 03, 05, 06).	CCLIB[35]
atomcharges_apt_sum_01_di	APT Summed atomic charge for the diene reactant structure for each atom n (01, 03, 05, 06).	CCLIB[35]
atomcharges_mulliken_01_di	Mulliken atomic charge for the diene reactant structure for each atom n (01, 03, 05, 06).	CCLIB[35]
atomcharges_mulliken_sum_01_di	Mulliken summed atomic charge for the diene reactant structure for each atom n (01, 03, 05, 06).	CCLIB[35]
homoenergies_di	HOMO energy - diene reactant.	CCLIB[35]
lumoenergies_di	LUMO energy - diene reactant.	CCLIB[35]
vibfreqs_01_di	Lowest vibrational frequency for the diene reactant.	CCLIB[35]
vibirs_01_di	Lowest infrared intensity for the diene reactant.	CCLIB[35]
hardness_di	Global hardness - diene reactant.	HSAB[36]
softness_di	Global softness - diene reactant.	HSAB[36]
chemicalpotential_di	Global chemical potential - diene reactant.	HSAB[36]
electrophilicity_di	Global electrophilicity - diene reactant.	HSAB[36]
sasa_1_di	Solvent accessible surface area for each atom n (1, 3, 5, 6) - diene reactant.	Freesasa[37]
sasa_R1_di	Solvent accessible surface area for R1 atom - diene reactant.	Freesasa[37]
sasa_total_di	Total solvent accessible surface area for core atoms - diene reactant.	Freesasa[37]

sterimol_R1_L_di	Sterimol L parameter for R1 substituent - diene reactant.	DBStep[38]
sterimol_R1_Bmin_di	Sterimol B_{min} parameter for R1 substituent - diene reactant.	DBStep[38]
sterimol_R1_Bmax_di	Sterimol B_{max} parameter for R1 substituent - diene reactant.	DBStep[38]
PBV_1_di	Percent buried volume (3.5 Å) for each reacting atom n (1, 3) - diene reactant.	DBStep[38]
HBA2_di	Number of hydrogen bond acceptors - diene reactant.	Pybel[39]
HBD_di	Number of hydrogen bond donors - diene reactant.	Pybel[39]
nF_di	Number of Fluorine atoms - diene reactant.	Pybel[39]
atomcharges_apt_02_dp	APT atomic charge for the dienophile reactant structure for each atom n (02, 04).	CCLIB[35]
atomcharges_apt_sum_02_dp	APT Summed Atomic charge for the dienophile reactant structure for each atom n (02, 04).	CCLIB[35]
atomcharges_mulliken_02_dp	Mulliken Atomic charge for the dienophile reactant structure for each atom n (02, 04).	CCLIB[35]
atomcharges_mulliken_sum_02_dp	Mulliken Summed Atomic charge for the dienophile reactant structure for each atom n (02, 04).	CCLIB[35]
homoenergies_dp	HOMO energy - dienophile reactant.	CCLIB[35]
lumoenergies_dp	LUMO energy - dienophile reactant.	CCLIB[35]
hardness_dp	Global hardness - dienophile reactant.	HSAB[36]
softness_dp	Global softness - dienophile reactant.	HSAB[36]
chemicalpotential_dp	Global chemical potential - dienophile reactant.	HSAB[36]
electrophilicity_dp	Global electrophilicity - dienophile reactant.	HSAB[36]
sasa_2_dp	Solvent accessible surface area for each atom n (2, 4) - dienophile reactant.	Freesasa[37]
sasa_R1_dp	Solvent accessible surface area for R1 atom - dienophile reactant.	Freesasa[37]
sasa_total_dp	Total solvent accessible surface area for core atoms - dienophile reactant.	Freesasa[37]
sterimol_R1_L_dp	Sterimol L parameter for R1 substituent - dienophile reactant.	DBStep[38]
sterimol_R1_Bmin_dp	Sterimol B_{min} parameter for R1 substituent - dienophile reactant.	DBStep[38]
sterimol_R1_Bmax_dp	Sterimol B_{max} parameter for R1 substituent - dienophile reactant.	DBStep[38]
PBV_2_dp	Percent buried volume (3.5 Å) for each reacting atom n (2, 4) - dienophile reactant.	DBStep[38]
HBA2_dp	Number of hydrogen bond acceptors - dienophile reactant.	Pybel[39]
HBD_dp	Number of hydrogen bond donors - dienophile reactant.	Pybel[39]
nF_dp	Number of Fluorine atoms - dienophile reactant.	Pybel[39]

Table S2 - AM1 extracted features with brief description and source of given feature. Information on the origin of the feature is also included (e.g., reactant or transition state species).

Feature	Description	Source
atomcharges_apt_n_ts	APT atomic charge for the transition state structure for each atom n (range 01-06).	CCLIB[35]
atomcharges_apt_sum_n_ts	APT summed atomic charge for the transition state structure for each atom n (range 01-06).	CCLIB[35]
atomcharges_mulliken_n_ts	Mulliken atomic charge for the transition state structure for each atom n (range 01-06).	CCLIB[35]
atomcharges_mulliken_sum_n_ts	Mulliken summed atomic charge for the transition state structure for each atom n (range 01-06).	CCLIB[35]
homoenergies_ts	HOMO energy - transition state.	CCLIB[35]
lumoenergies_ts	LUMO energy - transition state.	CCLIB[35]
hardness_ts	Global hardness - transition state.	HSAB[36]
softness_ts	Global softness - transition state.	HSAB[36]
chemicalpotential_ts	Global chemical potential - transition state.	HSAB[36]
electrophilicity_ts	Global electrophilicity - transition state.	HSAB[36]
gv_E_ts	Electronic energy.	GoodVibes[24]
gv_ZPE_ts	Zero-point energy.	GoodVibes[24]
gv_H_ts	Enthalpy.	GoodVibes[24]
gv_T.S_ts	Entropy.	GoodVibes[24]
gv_T.qh-S_ts	Quasi-harmonic entropy.	GoodVibes[24]
gv_G(T)_ts	Gibbs free energy.	GoodVibes[24]
gv_qh-G(T)_ts	Quasi-harmonic Gibbs free energy.	GoodVibes[24]
ea_ts	Semi-Empirical Quasi-harmonic free energy reaction barrier.	GoodVibes[24]
sasa_n_ts	Solvent accessible surface area for each atom n (range 1-6) - transition state.	Freesasa[37]
sasa_R1_ts	Solvent accessible surface area for R1 atom - transition state.	Freesasa[37]
sasa_total_ts	Total solvent accessible surface area for core atoms - transition state.	Freesasa[37]
sterimol_R1_L_ts	Sterimol L parameter for R1 substituent - transition state.	DBStep[38]
sterimol_R1_Bmin_ts	Sterimol B_{min} parameter for R1 substituent - transition state.	DBStep[38]
sterimol_R1_Bmax_ts	Sterimol B_{max} parameter for R1 substituent - transition state.	DBStep[38]
PBV_n_ts	Percent buried volume (3.5 Å) for each reacting atom n (range 1-4) - transition state.	DBStep[38]
HBA2_ts	Number of hydrogen bond acceptors - transition state.	Pybel[39]
HBD_ts	Number of hydrogen bond donors - transition state.	Pybel[39]
nF_ts	Number of Fluorine atoms - transition state.	Pybel[39]
bond_forming_distance_1_ts	Transition state bond forming distance between atoms 1 and 2.	-
bond_forming_distance_2_ts	Transition state bond forming distance between atoms 3 and 4.	-
bond_forming_angle_1_ts	Transition state bond forming angle between atoms 2, 1, and 3.	-
bond_forming_angle_2_ts	Transition state bond forming angle between atoms 4, 3, and 1.	-
bond_form_diff_ts	Transition state difference between the two bond forming distances.	-

bond_ang_diff_ts	Transition state difference between the two bond forming angles.	-
atomcharges_apt_n_di	APT atomic charge for the diene reactant structure for each atom n (01, 03, 05, 06).	CCLIB[35]
atomcharges_apt_sum_n_di	APT summed atomic charge for the diene reactant structure for each atom n (01, 03, 05, 06).	CCLIB[35]
atomcharges_mulliken_n_di	Mulliken atomic charge for the diene reactant structure for each atom n (01, 03, 05, 06).	CCLIB[35]
atomcharges_mulliken_sum_n_di	Mulliken summed atomic charge for the diene reactant structure for each atom n (01, 03, 05, 06).	CCLIB[35]
homoenergies_di	HOMO energy - diene reactant.	CCLIB[35]
lumoenergies_di	LUMO energy - diene reactant.	CCLIB[35]
vibfreqs_01_di	Lowest vibrational frequency for the diene reactant.	CCLIB[35]
vibirs_01_di	Lowest infrared intensity for the diene reactant.	CCLIB[35]
hardness_di	Global hardness - diene reactant.	HSAB[36]
softness_di	Global softness - diene reactant.	HSAB[36]
chemicalpotential_di	Global chemical potential - diene reactant.	HSAB[36]
electrophilicity_di	Global electrophilicity - diene reactant.	HSAB[36]
sasa_n_di	Solvent accessible surface area for each atom n (1, 3, 5, 6) - diene reactant.	Freesasa[37]
sasa_R1_di	Solvent accessible surface area for R1 atom - diene reactant.	Freesasa[37]
sasa_total_di	Total solvent accessible surface area for core atoms - diene reactant.	Freesasa[37]
sterimol_R1_L_di	Sterimol L parameter for R1 substituent - diene reactant.	DBStep[38]
sterimol_R1_Bmin_di	Sterimol B_{min} parameter for R1 substituent - diene reactant.	DBStep[38]
sterimol_R1_Bmax_di	Sterimol B_{max} parameter for R1 substituent - diene reactant.	DBStep[38]
PBV_n_di	Percent buried volume (3.5 Å) for each reacting atom n (1, 3) - diene reactant.	DBStep[38]
HBA2_di	Number of hydrogen bond acceptors - diene reactant.	Pybel[39]
HBD_di	Number of hydrogen bond donors - diene reactant.	Pybel[39]
nF_di	Number of Fluorine atoms - diene reactant.	Pybel[39]
homoenergies_dp	HOMO energy - dienophile reactant.	CCLIB[35]
lumoenergies_dp	LUMO energy - dienophile reactant.	CCLIB[35]
hardness_dp	Global hardness - dienophile reactant.	HSAB[36]
softness_dp	Global softness - dienophile reactant.	HSAB[36]
chemicalpotential_dp	Global chemical potential - dienophile reactant.	HSAB[36]
electrophilicity_dp	Global electrophilicity - dienophile reactant.	HSAB[36]
sasa_n_dp	Solvent accessible surface area for each atom n (2, 4) - dienophile reactant.	Freesasa[37]
sasa_R1_dp	Solvent accessible surface area for R1 atom - dienophile reactant.	Freesasa[37]
sasa_total_dp	Total solvent accessible surface area for core atoms - dienophile reactant.	Freesasa[37]
sterimol_R1_L_dp	Sterimol L parameter for R1 substituent - dienophile reactant.	DBStep[38]
sterimol_R1_Bmin_dp	Sterimol B_{min} parameter for R1 substituent - dienophile reactant.	DBStep[38]

sterimol_R1_Bmax_dp	Sterimol B_{max} parameter for R1 substituent - dienophile reactant.	DBStep[38]
PBV_n_dp	Percent buried volume (3.5 Å) for each reacting atom n (2, 4) - dienophile reactant.	DBStep[38]
HBA2_dp	Number of hydrogen bond acceptors - dienophile reactant.	Pybel[39]
HBD_dp	Number of hydrogen bond donors - dienophile reactant.	Pybel[39]
nF_dp	Number of Fluorine atoms - dienophile reactant.	Pybel[39]

Table S3 – PM3 extracted features with brief description and source of given feature. Information on the origin of the feature is also included (e.g., reactant or transition state species).

3. Machine Learning

All regression models were built with sklearn.[40] Prior to building and training models, features were standardised using sklearn's StandardScaler. For all regression models, only the X values were standardised. However, for the neural networks (NNs) built with TensorFlow[41], both the X and y were standardised with their own scalars ensuring that weights generated for the model are scaled in the same way as the input, thus reducing the computational time involved with training.

For all regression models generated using sklearn, an 80% training set was used to predict the DFT quasi-harmonic free energy barrier and the remaining 20% for testing. A similar process was used for NNs built using TensorFlow except for an additional validation set was used, thus the splitting consisted of 64% training, 16% validation, and 20% test sets. To explore the variability within algorithms, various kernels were chosen to be investigated for KRR and SVR models; radial basis function (RBF) and polynomial kernels were employed for both, whilst a Laplacian kernel was also tested for KRR models. Hyperparameter tuning was performed for all regression models built with sklearn utilising sklearn's GridSearchCV to search the hyperparameter space for the best cross validation (CV) MAE (Table S4). 5-fold CV was utilised in training to assess model and hyperparameter combinations. Upon completion of tuning, each regressor was individually fit with the associated best hyperparameters before obtaining predictions on the held-out test set.

The hyperparameter tuning for the NNs were performed using the Hyperband[42] tuner with early stopping implemented to find the best set of hyperparameters. A sequential network architecture was used with 4 hidden layers and dropout layers included after every hidden layer to help prevent overfitting. Other architectures were tested however four hidden layers was found to provide the best average performance and the simplest structure (larger networks result in an increased computational cost for model training and hyperparameter tuning). The hyperparameters tuned were learning rate, number of nodes per layer, hidden layer regularisation value, and dropout rate (for the dropout layers only). Model performance was monitored by generating loss curves on the training and validation sets by rebuilding and refitting the network at each stage with the optimised hyperparameters.

Model	Hyperparameters	
	Tuning Method	Search Space
Ridge Regression[43]	Grid Search	{'alpha': [1e-6, 1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1, 10, 100]}
Kernel Ridge Regression (RBF)[44]	Grid Search	{'alpha': [1e-6, 1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1, 10, 100], 'gamma':[None, 0.01, 0.025, 0.05, 0.1, 0.25, 0.5, 1]}
Kernel Ridge Regression (Polynomial)[44]	Grid Search	{'alpha': [1e-6, 1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1, 10, 100], 'gamma':[None, 0.01, 0.025, 0.05, 0.1, 0.25, 0.5, 1]}
Kernel Ridge Regression (Laplacian)[44]	Grid Search	{'alpha': [1e-6, 1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1, 10, 100], 'gamma':[None, 0.01, 0.025, 0.05, 0.1, 0.25, 0.5, 1]}
Support Vector Regression (RBF)[45]	Grid Search	{'gamma':['scale', 'auto'], 'epsilon':[0.001, 0.01, 0.025, 0.05, 0.1, 0.25, 0.5, 1], 'C':[0.1, 0.25, 0.5, 1, 2, 5, 10, 20, 30, 50]}
Support Vector Regression (Polynomial)[45]	Grid Search	{'gamma':['scale', 'auto'], 'epsilon':[0.001, 0.01, 0.025, 0.05, 0.1, 0.25, 0.5, 1], 'C':[1, 2, 5, 10, 20, 30, 50], 'coef0': [0, 1], 'degree': [1, 2, 3, 4, 5]}
Sequential Neural Network	Hyperband[42]	{'reg_value':[1e-2, 1e-3, 1e-4], 'learning_rate':values=[1e-2, 1e-3, 1e-4], 'dropout_rate1':values=[0.4, 0.3, 0.2, 0.1], 'dropout_rate2':values=[0.4, 0.3, 0.2, 0.1], 'dropout_rate3':values=[0.4, 0.3, 0.2, 0.1], 'dropout_rate4':values=[0.4, 0.3, 0.2, 0.1], 'node_units1':[min_value=64, max_value=512, step=32], 'node_units2':[min_value=64, max_value=512, step=32], 'node_units3':[min_value=64, max_value=512, step=32], 'node_units4':[min_value=64, max_value=512, step=32]}

Table S4 - All models built using sklearn and TensorFlow with associated tuning method and search space for hyperparameter tuning.

To generate the leave-one-out (LOO) datasets, any diene/dienophile that was enumerated to have a Cl as the R group was removed from the dataset and set as its own leave-one-out dataset (Fig. S3). This was completed for both endo and exo datasets to give certain leave-one-out datasets that are summarised in Table S5. These datasets were subsequently split into train and test sets as previously described.

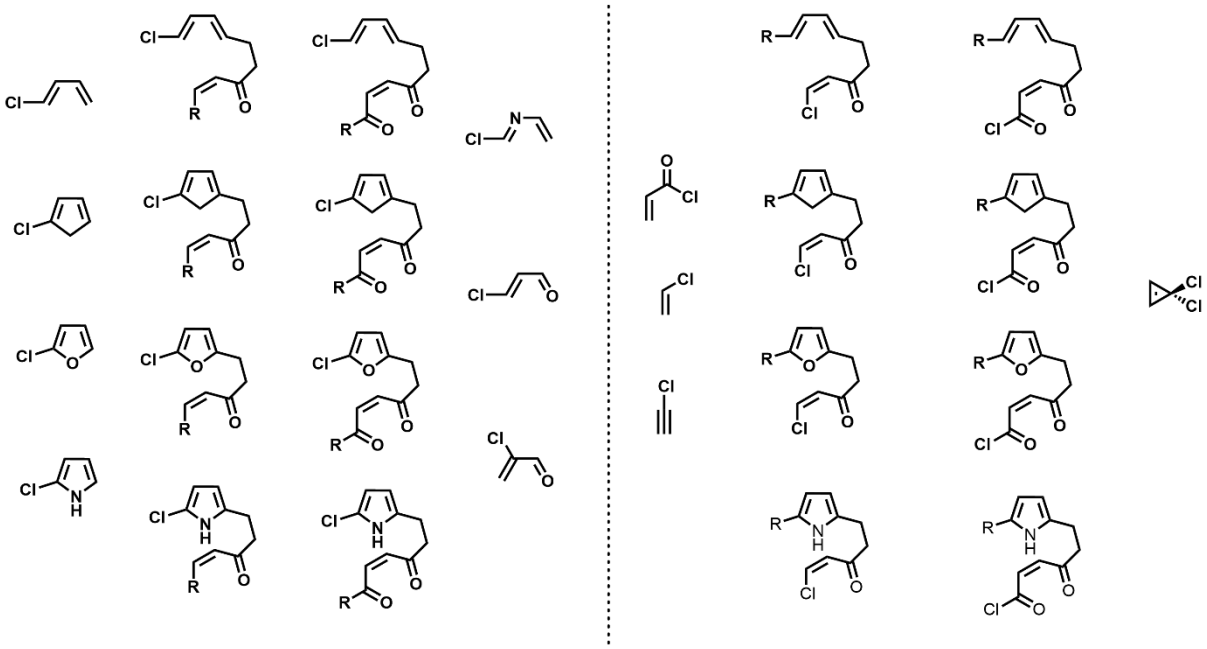

Fig. S3 - Base structures omitted in leave-one-out, dienes (left) and dienophiles (right).

Datasets	Base Dataset Size	LOO Size	LOO Pre-ML MAE / kcal mol^{-1}
AM1 Endo + LODiO	895	170	9.93
AM1 Endo + LODpO	904	161	9.63
AM1 Exo + LODiO	937	172	9.13
AM1 Exo + LODpO	932	177	9.42
PM3 Endo + LODiO	965	176	11.65
PM3 Endo + LODpO	972	169	11.33
PM3 Exo + LODiO	968	173	11.35
PM3 Exo + LODpO	961	180	11.14

Table S5 - Dataset sizes and pre-ML MAEs for all data.

To perform transfer learning (TL), a NN was first built on the appropriate selection of data for the base model. The data was then partitioned into the base model train, test, and validation sets, as well as the TL target train, test, and validation sets. The hyperparameters were tuned from the same search space using the same method as previously stated (Table S4). Once optimal hyperparameters were obtained, the model was rebuilt using these values and predictions were made on the TL target test set to provide a pre-TL prediction for comparison. The first three hidden layers of the model were then frozen (with only the final hidden layer retrainable) before retraining with the TL target train set (and using the TL target validation set to monitor model performance). Once retrained, predictions were obtained for the TL target test set and compared to the pre-TL predictions. To determine TL performance at very low data regimes, the TL target train set was partitioned into different percentages to evaluate the data limits at which TL can perform effectively.

When testing these models, test MAE values were primarily used to assess performance with comparisons drawn to train MAEs where appropriate to highlight any potential instances of overfitting. Standard errors were also calculated for each prediction by dividing the standard deviation of each individual prediction absolute errors by the square root of the number of samples.

4. Machine Learning Hyperparameters and Metrics

Tables S6 and S7 display the train MAE, test MAE, and test R^2 for each standard ML model built using both the AM1-DFT and PM3-DFT datasets. The associated test errors and hyperparameters for each model are also provided. All results are provided across endo and exo models. Leave-one-out results are displayed in Tables S8-11 with the appropriate hyperparameters, metrics, and test errors.

AM1 Endo				
Model	Train MAE / kcal mol^{-1}	Test MAE / kcal mol^{-1}	Test R^2	Hyperparameters
Ridge	0.771	0.767 ± 0.053	0.969	{'alpha': 0.01}
KRR (RBF)	0.863	0.716 ± 0.062	0.966	{'alpha': 1e-06, 'gamma': None}
KRR (Poly)	0.510	0.416 ± 0.032	0.990	{'alpha': 0.01, 'gamma': None}
KRR (Laplacian)	0.690	0.597 ± 0.047	0.979	{'alpha': 1e-06, 'gamma': None}
SVR (RBF)	0.584	0.503 ± 0.041	0.984	{'C': 50, 'epsilon': 0.001, 'gamma': 'auto'}
SVR (Poly)	0.431	0.403 ± 0.033	0.990	{'C': 50, 'coef0': 1, 'degree': 2, 'epsilon': 0.025, 'gamma': 'auto'}
NN	0.623	0.711 ± 0.046	0.977	{'reg_value': 0.0001, 'learning_rate': 0.001, 'dropout_rate1': 0.3, 'dropout_rate2': 0.2, 'dropout_rate3': 0.2, 'dropout_rate4': 0.1, 'node_units1': 480, 'node_units2': 416, 'node_units3': 352, 'node_units4': 160}
AM1 Exo				
Ridge	0.958	0.964 ± 0.055	0.972	{'alpha': 1e-05}
KRR (RBF)	0.803	0.811 ± 0.074	0.966	{'alpha': 0.0001, 'gamma': None}
KRR (Poly)	0.509	0.440 ± 0.038	0.991	{'alpha': 0.01, 'gamma': None}
KRR (Laplacian)	0.701	0.648 ± 0.050	0.983	{'alpha': 1e-06, 'gamma': None}
SVR (RBF)	0.606	0.542 ± 0.043	0.988	{'C': 50, 'epsilon': 0.001, 'gamma': 'scale'}
SVR (Poly)	0.497	0.394 ± 0.036	0.992	{'C': 10, 'coef0': 1, 'degree': 3, 'epsilon': 0.01, 'gamma': 'auto'}
NN	0.933	1.101 ± 0.079	0.980	{'reg_value': 0.0001, 'learning_rate': 0.001, 'dropout_rate1': 0.3, 'dropout_rate2': 0.2, 'dropout_rate3': 0.2, 'dropout_rate4': 0.1, 'node_units1': 480, 'node_units2': 416, 'node_units3': 352, 'node_units4': 160}

Table S6 - AM1-DFT Endo and Exo standard ML train and test set metrics with associated hyperparameters.

PM3 Endo				
Model	Train MAE / kcal mol^{-1}	Test MAE / kcal mol^{-1}	Test R^2	Hyperparameters
Ridge	1.055	0.874 ± 0.048	0.961	{'alpha': 1e-06}
KRR (RBF)	0.883	0.718 ± 0.054	0.965	{'alpha': 1e-06, 'gamma': None}
KRR (Poly)	0.561	0.463 ± 0.031	0.987	{'alpha': 0.1, 'gamma': None}
KRR (Laplacian)	0.693	0.614 ± 0.039	0.980	{'alpha': 1e-06, 'gamma': None}
SVR (RBF)	0.639	0.517 ± 0.035	0.984	{'C': 50, 'epsilon': 0.001, 'gamma': 'auto'}
SVR (Poly)	0.529	0.425 ± 0.033	0.987	{'C': 10, 'coef0': 1, 'degree': 3, 'epsilon': 0.025, 'gamma': 'scale'}
NN	0.551	0.759 ± 0.043	0.977	{'reg_value': 0.0001, 'learning_rate': 0.001, 'dropout_rate1': 0.1, 'dropout_rate2': 0.1, 'dropout_rate3': 0.1, 'dropout_rate4': 0.1, 'node_units1': 448, 'node_units2': 352, 'node_units3': 96, 'node_units4': 320}
PM3 Exo				
Ridge	1.126	1.231 ± 0.073	0.940	{'alpha': 1e-05}
KRR (RBF)	0.880	0.717 ± 0.055	0.973	{'alpha': 1e-06, 'gamma': None}
KRR (Poly)	0.598	0.459 ± 0.033	0.989	{'alpha': 0.01, 'gamma': None}
KRR (Laplacian)	0.748	0.582 ± 0.037	0.985	{'alpha': 1e-06, 'gamma': None}
SVR (RBF)	0.676	0.551 ± 0.038	0.985	{'C': 30, 'epsilon': 0.001, 'gamma': 'auto'}
SVR (Poly)	0.580	0.432 ± 0.033	0.990	{'C': 10, 'coef0': 1, 'degree': 3, 'epsilon': 0.001, 'gamma': 'auto'}
NN	0.648	0.881 ± 0.069	0.973	{'reg_value': 0.0001, 'learning_rate': 0.001, 'dropout_rate1': 0.1, 'dropout_rate2': 0.1, 'dropout_rate3': 0.1, 'dropout_rate4': 0.1, 'node_units1': 448, 'node_units2': 352, 'node_units3': 96, 'node_units4': 320}

Table S7 - PM3-DFT Endo and Exo standard ML train and test set metrics with associated hyperparameters.

AM1 LODiO Endo					
Model	Train MAE / kcal mol^{-1}	Test MAE / kcal mol^{-1}	Test R^2	LODiO MAE / kcal mol^{-1}	Hyperparameters
Ridge	0.802	0.822 ± 0.064	0.960	1.603 ± 0.18	{'alpha': 1e-06}
KRR (RBF)	0.998	0.761 ± 0.063	0.964	1.441 ± 0.236	{'alpha': 1e-06, 'gamma': None}
KRR (Poly)	0.572	0.473 ± 0.036	0.987	0.802 ± 0.129	{'alpha': 0.01, 'gamma': None}
KRR (Laplacian)	0.760	0.600 ± 0.045	0.980	0.758 ± 0.111	{'alpha': 1e-06, 'gamma': None}
SVR (RBF)	0.666	0.483 ± 0.036	0.987	0.794 ± 0.116	{'C': 30, 'epsilon': 0.001, 'gamma': 'auto'}
SVR (Poly)	0.493	0.443 ± 0.033	0.989	1.011 ± 0.119	{'C': 30, 'coef0': 1, 'degree': 2, 'epsilon': 0.05, 'gamma': 'scale'}
NN	0.533	0.705 ± 0.043	0.976	0.977 ± 0.120	{'reg_value': 0.0001, 'learning_rate': 0.001, 'dropout_rate1': 0.3, 'dropout_rate2': 0.2, 'dropout_rate3': 0.2, 'dropout_rate4': 0.1, 'node_units1': 480, 'node_units2': 416, 'node_units3': 352, 'node_units4': 160}
AM1 LODiO Exo					
Ridge	0.970	0.994 ± 0.068	0.963	1.018 ± 0.167	{'alpha': 1e-05}
KRR (RBF)	0.929	0.757 ± 0.065	0.972	1.717 ± 0.291	{'alpha': 0.001, 'gamma': None}
KRR (Poly)	0.562	0.489 ± 0.043	0.988	0.837 ± 0.123	{'alpha': 0.01, 'gamma': None}
KRR (Laplacian)	0.767	0.635 ± 0.046	0.983	0.878 ± 0.114	{'alpha': 1e-06, 'gamma': None}
SVR (RBF)	0.672	0.564 ± 0.050	0.983	0.879 ± 0.117	{'C': 50, 'epsilon': 0.001, 'gamma': 'auto'}
SVR (Poly)	0.527	0.501 ± 0.045	0.987	0.758 ± 0.114	{'C': 10, 'coef0': 1, 'degree': 3, 'epsilon': 0.01, 'gamma': 'scale'}
NN	0.792	1.059 ± 0.082	0.953	1.036 ± 0.094	{'reg_value': 0.001, 'learning_rate': 0.001, 'dropout_rate1': 0.2, 'dropout_rate2': 0.1, 'dropout_rate3': 0.1, 'dropout_rate4': 0.4, 'node_units1': 480, 'node_units2': 288, 'node_units3': 64, 'node_units4': 224}

Table S8 - AM1-DFT Endo and Exo leave-one-diene-out standard ML train and test set metrics with associated hyperparameters.

PM3 LODiO Endo					
Model	Train MAE / kcal mol^{-1}	Test MAE / kcal mol^{-1}	Test R^2	LODiO MAE / kcal mol^{-1}	Hyperparameters
Ridge	1.076	0.925 ± 0.062	0.960	2.501 ± 0.196	{'alpha': 1e-06}
KRR (RBF)	0.963	0.846 ± 0.066	0.961	2.061 ± 0.276	{'alpha': 0.0001, 'gamma': None}
KRR (Poly)	0.599	0.551 ± 0.050	0.980	1.232 ± 0.223	{'alpha': 0.1, 'gamma': None}
KRR (Laplacian)	0.779	0.683 ± 0.048	0.977	0.779 ± 0.133	{'alpha': 1e-06, 'gamma': None}
SVR (RBF)	0.670	0.596 ± 0.045	0.981	1.006 ± 0.163	{'C': 30, 'epsilon': 0.001, 'gamma': 'auto'}
SVR (Poly)	0.566	0.531 ± 0.052	0.980	1.042 ± 0.175	{'C': 5, 'coef0': 1, 'degree': 3, 'epsilon': 0.001, 'gamma': 'scale'}
NN	0.555	0.881 ± 0.074	0.963	0.960 ± 0.138	{'reg_value': 0.0001, 'learning_rate': 0.001, 'dropout_rate1': 0.1, 'dropout_rate2': 0.1, 'dropout_rate3': 0.1, 'dropout_rate4': 0.1, 'node_units1': 448, 'node_units2': 352, 'node_units3': 96, 'node_units4': 320}
PM3 LODiO Exo					
Ridge	1.166	1.151 ± 0.065	0.952	2.009 ± 0.181	{'alpha': 0.01}
KRR (RBF)	0.950	0.769 ± 0.062	0.971	1.351 ± 0.215	{'alpha': 1e-06, 'gamma': None}
KRR (Poly)	0.631	0.538 ± 0.045	0.985	1.01 ± 0.158	{'alpha': 0.01, 'gamma': None}
KRR (Laplacian)	0.779	0.654 ± 0.044	0.981	0.701 ± 0.077	{'alpha': 1e-06, 'gamma': None}
SVR (RBF)	0.730	0.582 ± 0.045	0.983	0.852 ± 0.115	{'C': 50, 'epsilon': 0.001, 'gamma': 'auto'}
SVR (Poly)	0.621	0.542 ± 0.046	0.984	0.867 ± 0.128	{'C': 20, 'coef0': 1, 'degree': 3, 'epsilon': 0.025, 'gamma': 'auto'}
NN	0.612	0.887 ± 0.063	0.971	0.793 ± 0.103	{'reg_value': 0.0001, 'learning_rate': 0.001, 'dropout_rate1': 0.3, 'dropout_rate2': 0.2, 'dropout_rate3': 0.2, 'dropout_rate4': 0.1, 'node_units1': 480, 'node_units2': 416, 'node_units3': 352, 'node_units4': 160}

Table S9 – PM3-DFT Endo and Exo leave-one-diene-out standard ML train and test set metrics with associated hyperparameters.

AM1 LODpO Endo					
Model	Train MAE / kcal mol^{-1}	Test MAE / kcal mol^{-1}	Test R^2	LODiO MAE / kcal mol^{-1}	Hyperparameters
Ridge	0.827	0.762 ± 0.045	0.975	1.616 ± 0.216	{'alpha': 0.01}
KRR (RBF)	0.969	0.791 ± 0.073	0.958	1.7 ± 0.254	{'alpha': 1e-06, 'gamma': None}
KRR (Poly)	0.564	0.521 ± 0.041	0.984	0.612 ± 0.084	{'alpha': 0.01, 'gamma': None}
KRR (Laplacian)	0.748	0.672 ± 0.048	0.977	0.935 ± 0.162	{'alpha': 1e-06, 'gamma': None}
SVR (RBF)	0.646	0.574 ± 0.053	0.977	0.766 ± 0.093	{'C': 50, 'epsilon': 0.001, 'gamma': 'auto'}
SVR (Poly)	0.499	0.463 ± 0.035	0.988	0.823 ± 0.120	{'C': 30, 'coef0': 1, 'degree': 2, 'epsilon': 0.025, 'gamma': 'scale'}
NN	0.824	0.913 ± 0.059	0.970	0.932 ± 0.104	{'reg_value': 0.0001, 'learning_rate': 0.001, 'dropout_rate1': 0.3, 'dropout_rate2': 0.2, 'dropout_rate3': 0.2, 'dropout_rate4': 0.1, 'node_units1': 480, 'node_units2': 416, 'node_units3': 352, 'node_units4': 160}
AM1 LODpO Exo					
Ridge	0.929	0.914 ± 0.061	0.971	3.325 ± 0.472	{'alpha': 0.001}
KRR (RBF)	0.923	0.688 ± 0.069	0.974	1.641 ± 0.214	{'alpha': 1e-06, 'gamma': None}
KRR (Poly)	0.527	0.513 ± 0.048	0.986	0.754 ± 0.108	{'alpha': 0.01, 'gamma': None}
KRR (Laplacian)	0.777	0.670 ± 0.051	0.982	0.868 ± 0.128	{'alpha': 1e-06, 'gamma': None}
SVR (RBF)	0.655	0.624 ± 0.054	0.981	0.669 ± 0.126	{'C': 50, 'epsilon': 0.001, 'gamma': 'scale'}
SVR (Poly)	0.499	0.474 ± 0.048	0.987	1.085 ± 0.149	{'C': 50, 'coef0': 1, 'degree': 2, 'epsilon': 0.05, 'gamma': 'auto'}
NN	0.955	1.141 ± 0.076	0.973	1.213 ± 0.137	{'reg_value': 0.0001, 'learning_rate': 0.001, 'dropout_rate1': 0.3, 'dropout_rate2': 0.2, 'dropout_rate3': 0.2, 'dropout_rate4': 0.1, 'node_units1': 480, 'node_units2': 416, 'node_units3': 352, 'node_units4': 160,

Table S10 - AM1-DFT Endo and Exo leave-one-dienophile-out standard ML train and test set metrics with associated hyperparameters.

PM3 LODpO Endo					
Model	Train MAE / kcal mol^{-1}	Test MAE / kcal mol^{-1}	Test R^2	LODiO MAE / kcal mol^{-1}	Hyperparameters
Ridge	1.007	0.968 ± 0.062	0.956	2.975 ± 0.344	{'alpha': 0.001}
KRR (RBF)	0.950	0.792 ± 0.074	0.958	1.350 ± 0.200	{'alpha': 1e-05, 'gamma': None}
KRR (Poly)	0.569	0.489 ± 0.040	0.986	0.907 ± 0.109	{'alpha': 0.1, 'gamma': None}
KRR (Laplacian)	0.743	0.628 ± 0.051	0.979	1.291 ± 0.141	{'alpha': 1e-06, 'gamma': None}
SVR (RBF)	0.669	0.575 ± 0.047	0.982	0.869 ± 0.106	{'C': 50, 'epsilon': 0.001, 'gamma': 'auto'}
SVR (Poly)	0.536	0.482 ± 0.039	0.987	1.261 ± 0.115	{'C': 30, 'coef0': 1, 'degree': 2, 'epsilon': 0.01, 'gamma': 'scale'}
NN	0.568	0.750 ± 0.047	0.978	0.991 ± 0.124	{'reg_value': 0.0001, 'learning_rate': 0.001, 'dropout_rate1': 0.3, 'dropout_rate2': 0.2, 'dropout_rate3': 0.2, 'dropout_rate4': 0.1, 'node_units1': 480, 'node_units2': 416, 'node_units3': 352, 'node_units4': 160}
PM3 LODpO Exo					
Ridge	1.121	1.078 ± 0.062	0.954	3.346 ± 0.356	{'alpha': 0.01}
KRR (RBF)	0.897	0.755 ± 0.061	0.97	1.301 ± 0.154	{'alpha': 1e-06, 'gamma': None}
KRR (Poly)	0.602	0.552 ± 0.05	0.981	0.708 ± 0.096	{'alpha': 0.01, 'gamma': None}
KRR (Laplacian)	0.806	0.655 ± 0.045	0.98	1.126 ± 0.121	{'alpha': 1e-06, 'gamma': None}
SVR (RBF)	0.692	0.588 ± 0.051	0.98	0.671 ± 0.088	{'C': 50, 'epsilon': 0.001, 'gamma': 'auto'}
SVR (Poly)	0.599	0.550 ± 0.050	0.981	0.762 ± 0.108	{'C': 30, 'coef0': 1, 'degree': 3, 'epsilon': 0.05, 'gamma': 'auto'}
NN	0.660	0.841 ± 0.064	0.967	0.891 ± 0.160	{'reg_value': 0.0001, 'learning_rate': 0.001, 'dropout_rate1': 0.3, 'dropout_rate2': 0.2, 'dropout_rate3': 0.2, 'dropout_rate4': 0.1, 'node_units1': 480, 'node_units2': 416, 'node_units3': 352, 'node_units4': 160,

Table S11 – PM3-DFT Endo and Exo leave-one-dienophile-out standard ML train and test set metrics with associated hyperparameters.

5. NN Feature Importances

Feature importances were calculated for the endo and exo NN models by taking each individual feature in the X test and randomly shuffling one feature exclusively and using this new X test to predict y test (Fig. S4). This yields an indication on the importance of a given feature on predictive ability (Fig. S5 and S6). All feature importances are for AM1-DFT endo and exo models.

X Test Features				
F_1	F_2	...	F_{n-1}	F_n
a_1	b_1	...	y_1	z_1
a_2	b_2	...	y_2	z_2
a_3	b_3	...	y_3	z_3
a_4	b_4	...	y_4	z_4
a_5	b_5	...	y_5	z_5

Random Shuffling of F_1

X Test Features				
F_1	F_2	...	F_{n-1}	F_n
a_2	b_1	...	y_1	z_1
a_5	b_2	...	y_2	z_2
a_4	b_3	...	y_3	z_3
a_1	b_4	...	y_4	z_4
a_3	b_5	...	y_5	z_5

Fig. S4 - Explanation of how feature importances were generated for NNs. Each feature (F_1 through F_2) is independently shuffled within the X test set and a prediction made on each new set of X test features.

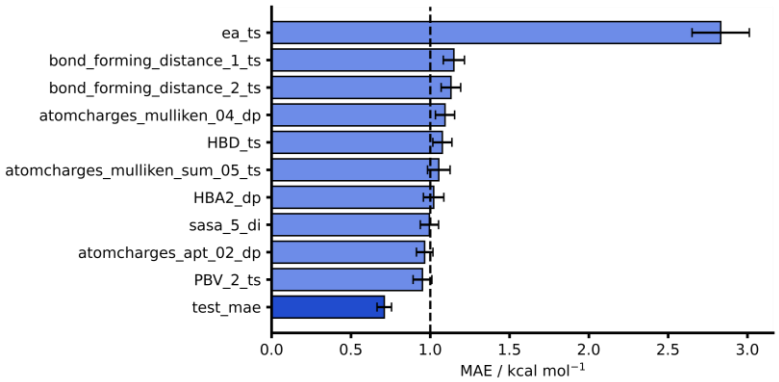

Fig. S5 - Feature Importances for AM1 endo NN. Test MAE (dark blue) plotted with 10 highest test MAEs (light blue) achieved after feature importance analysis.

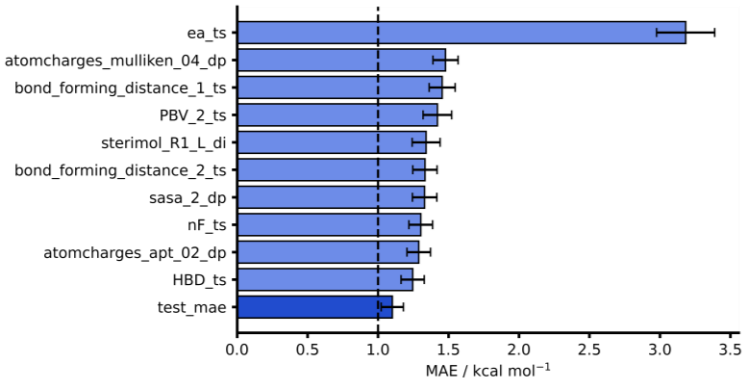

Fig. S6 - Feature Importances for AM1 exo NN. Test MAE (dark blue) plotted with 10 highest test MAEs (light blue) achieved after feature importance analysis.

6. TL Datasets

The whole dataset created was partitioned into different enumerations and Diels-Alder reaction classes to provide TL datasets. These partitions yielded source domain A and source domain B, which were paired with target domains α and β, respectively, for TL (Fig. S7 and S8). The enumeration used for generating the [3+2] cycloaddition dataset is also included here (Fig. S9). For the [3+2] cycloaddition dataset, a broadly representative sample was selected from a recently published dataset[46] and enumerated and calculated with our workflow to generate a dataset of 420 [3+2] cycloaddition reactions.

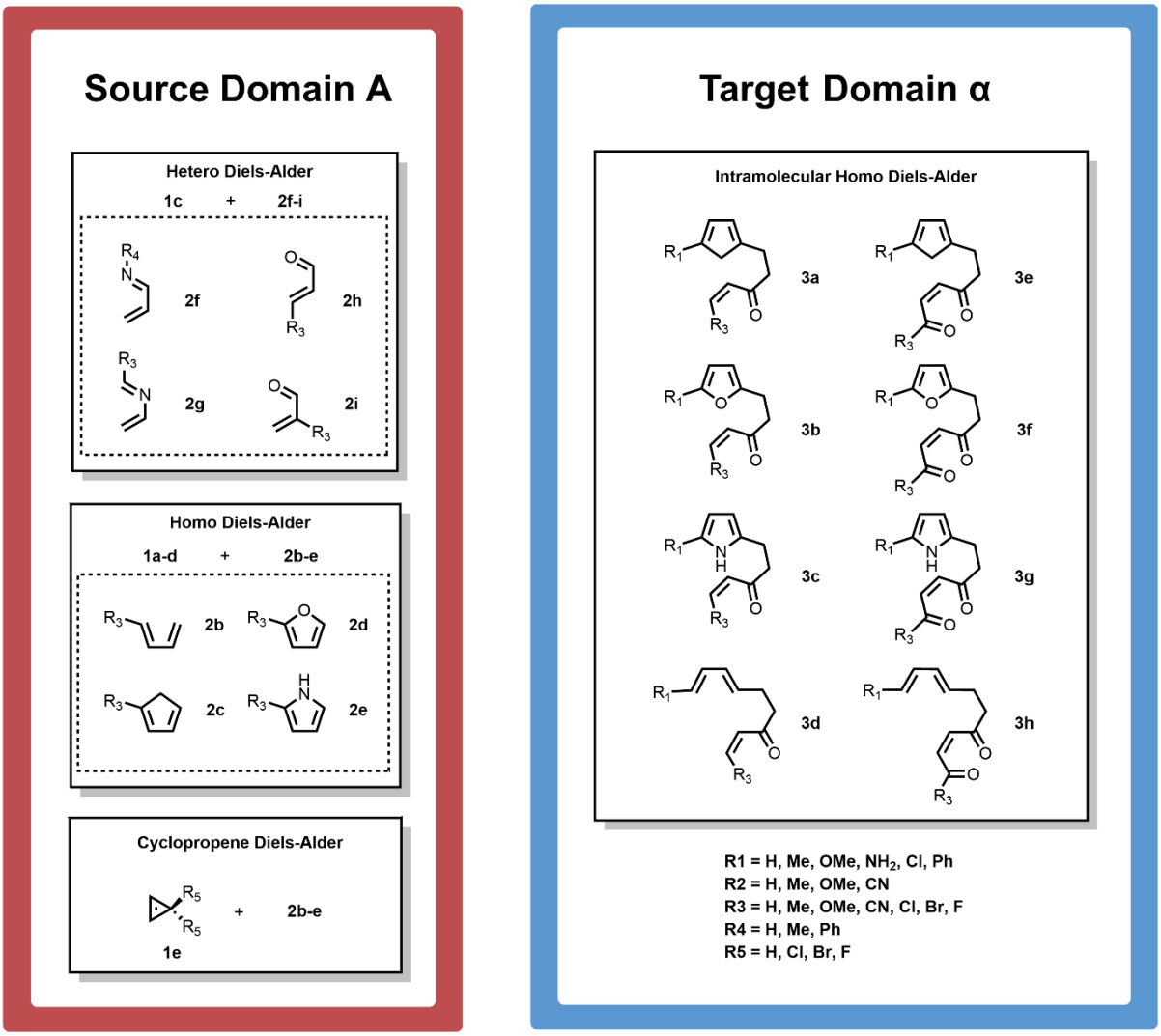

Fig. S7 - Partition of data to create source domain A and target domain α. Source domain A contains hetero/homo and cyclopropane-containing intermolecular Diels-Alder reactions. Target domain α contains intramolecular Diels-Alder reactions.

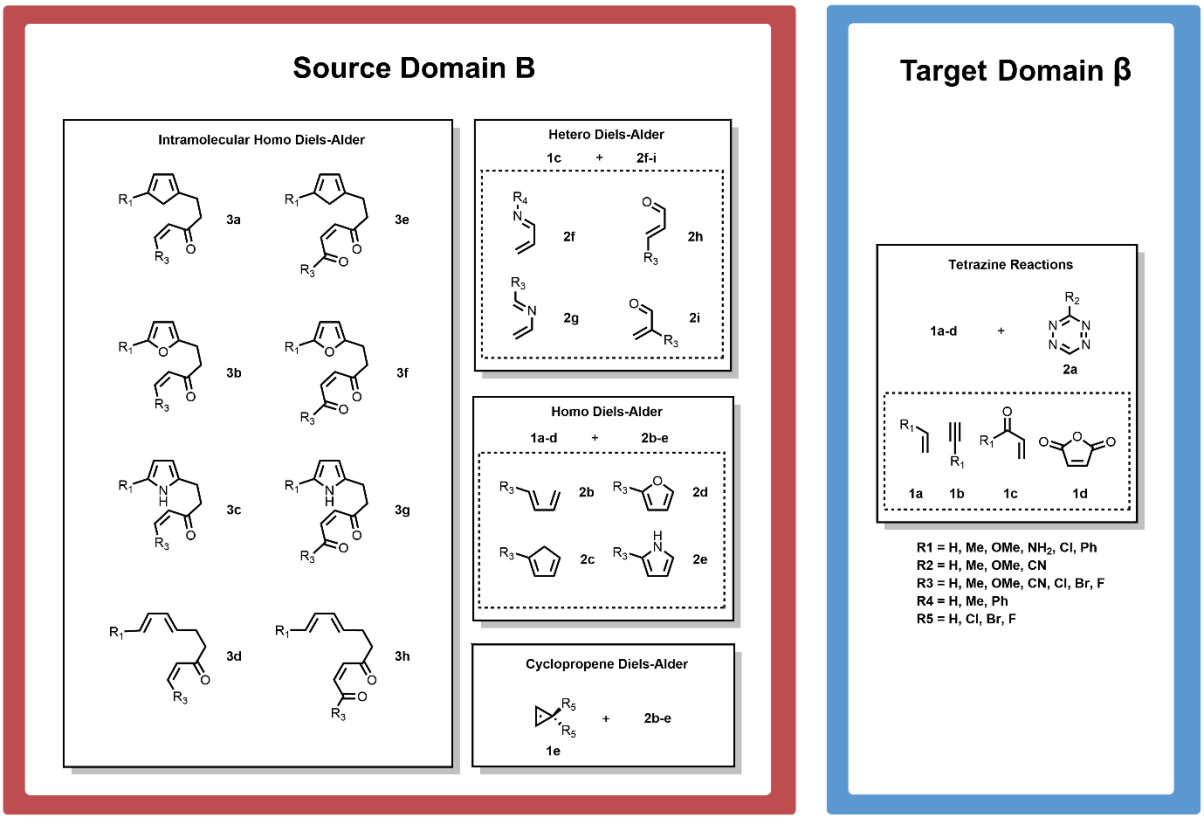

Fig. S8 - Partition of data to create source domain B and target domain β. Source domain B contains hetero/homo and cyclopropane-containing intermolecular Diels-Alder reactions along with intramolecular Diels-Alder reactions. Target domain β contains tetrazine Diels-Alder reactions.

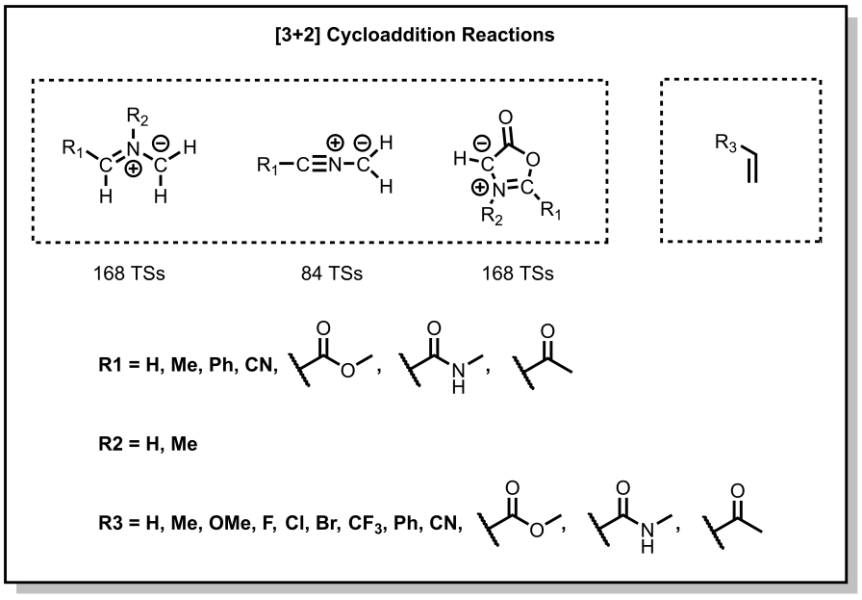

Figure S9 - Enumerated [3+2] cycloaddition dataset used as target domain hTL. The source domain for this hTL was the B source domain from Fig. S7. Out of the 420 possible enumerated reactions, 408 were successfully optimised at the DFT level of theory. Extensive attempts to optimise the TSs of the remaining 12 reactions were unsuccessful.

7. hTL Metrics

Table S12 show metrics for before and after the hTL procedure. Tables S13-S16 show the test and train MAEs for the hTL work with associated errors at three random states across different hTL training percentage splits. This work was performed on the AM1-DFT dataset. hTL was also performed for an enumerated group of [3+2] reactions to provide an extended test (Table S17 and S18 for test and train metrics respectively).

AM1 hTL Endo						
Model	Base Train MAE / kcal mol^{-1}	Base Test MAE / kcal mol^{-1}	Base Test R^2	Base Target MAE / kcal mol^{-1}	TL Target MAE / kcal mol^{-1}	TL Target R^2
A → α	0.503	0.647 ± 0.045	0.974	5.090 ± 0.443	0.946 ± 0.085	0.987
B → β	0.623	0.711 ± 0.046	0.977	1.921 ± 0.325	0.932 ± 0.191	0.783
AM1 hTL Exo						
A → α	0.464	0.680 ± 0.043	0.978	4.953 ± 0.413	1.039 ± 0.100	0.985
B → β	0.933	1.101 ± 0.079	0.980	3.318 ± 0.399	1.242 ± 0.198	0.741

Table S12 – AM1-DFT endo and exo train and base/hTL test metrics. Test metrics for the target are provided for before and after hTL was performed.

A → α Endo Test MAE / kcal mol^{-1} hTL Train Percentages				
Percent of Training Data	Random State			Average
	21	22	23	
10	2.540 ± 0.327	2.235 ± 0.256	2.839 ± 0.316	2.538 ± 0.299
20	2.361 ± 0.236	1.576 ± 0.220	1.849 ± 0.224	1.928 ± 0.227
30	1.480 ± 0.171	1.220 ± 0.136	1.629 ± 0.155	1.443 ± 0.154
40	1.395 ± 0.135	1.295 ± 0.146	1.705 ± 0.167	1.465 ± 0.149
50	1.164 ± 0.133	1.219 ± 0.139	1.159 ± 0.096	1.181 ± 0.123
60	1.160 ± 0.122	1.290 ± 0.120	1.216 ± 0.127	1.222 ± 0.123
70	0.993 ± 0.108	1.082 ± 0.116	1.051 ± 0.092	1.042 ± 0.105
80	0.947 ± 0.089	0.999 ± 0.107	0.914 ± 0.077	0.953 ± 0.091
90	0.917 ± 0.093	0.943 ± 0.105	0.790 ± 0.072	0.883 ± 0.090
100	0.835 ± 0.098	0.889 ± 0.093	0.946 ± 0.085	0.890 ± 0.092
A → α Exo Test MAE / kcal mol^{-1} hTL Train Percentages				
10	2.769 ± 0.282	2.198 ± 0.208	2.746 ± 0.267	2.571 ± 0.252
20	1.813 ± 0.207	2.891 ± 0.315	1.962 ± 0.195	2.222 ± 0.239
30	1.514 ± 0.156	1.189 ± 0.141	1.458 ± 0.159	1.387 ± 0.152
40	1.745 ± 0.210	1.362 ± 0.159	1.712 ± 0.186	1.606 ± 0.185
50	1.321 ± 0.130	1.314 ± 0.134	1.358 ± 0.155	1.331 ± 0.140
60	1.038 ± 0.114	1.155 ± 0.127	1.115 ± 0.121	1.102 ± 0.121
70	1.171 ± 0.125	1.486 ± 0.158	1.501 ± 0.131	1.386 ± 0.138
80	0.963 ± 0.100	1.028 ± 0.110	1.031 ± 0.114	1.007 ± 0.108
90	1.107 ± 0.118	1.064 ± 0.100	0.925 ± 0.107	1.032 ± 0.108
100	0.794 ± 0.078	1.016 ± 0.100	1.039 ± 0.100	0.950 ± 0.093

Table S13 - AM1-DFT A → α endo and exo hTL test MAEs across three random states (21,22,23) with averaged MAEs. Metrics provided for different percentages of hTL training data.

A → α Endo Train MAE / kcal mol^{-1} hTL Train Percentages				
Percent of Training Data	Random State			Average
	21	22	23	
10	1.073	0.859	1.078	1.003
20	1.046	0.920	0.810	0.925
30	0.571	0.600	0.772	0.648
40	0.861	0.677	0.865	0.801
50	0.598	0.433	0.802	0.611
60	0.694	0.753	0.634	0.693
70	0.489	0.415	0.475	0.460
80	0.606	0.485	0.542	0.544
90	0.470	0.409	0.393	0.424
100	0.590	0.506	0.614	0.570
A → α Exo Train MAE / kcal mol^{-1} hTL Train Percentages				
10	1.010	0.468	1.404	0.961
20	1.130	1.142	1.207	1.160
30	0.879	0.502	0.608	0.663
40	0.936	0.764	0.838	0.846
50	0.931	0.747	0.912	0.863
60	0.546	0.825	0.579	0.650
70	0.866	1.093	0.873	0.944
80	0.620	0.533	0.568	0.574
90	0.748	0.592	0.483	0.607
100	0.499	0.671	0.753	0.641

Table S14 - AM1-DFT A → α endo and exo hTL train MAEs across three random states (21,22,23) with averaged MAEs. Metrics provided for different percentages of hTL training data.

B → β Endo Test MAE / kcal mol^{-1} hTL Train Percentages				
Percent of Training Data	Random State			Average
	21	22	23	
10	1.843 ± 0.336	1.255 ± 0.224	1.698 ± 0.340	1.599 ± 0.300
20	1.470 ± 0.258	1.774 ± 0.189	1.835 ± 0.316	1.693 ± 0.254
30	1.529 ± 0.222	0.977 ± 0.154	1.315 ± 0.272	1.274 ± 0.216
40	0.997 ± 0.156	1.036 ± 0.162	1.262 ± 0.189	1.098 ± 0.169
50	1.163 ± 0.138	0.910 ± 0.161	0.896 ± 0.152	0.989 ± 0.150
60	1.241 ± 0.206	0.681 ± 0.156	0.911 ± 0.126	0.944 ± 0.163
70	1.135 ± 0.185	0.789 ± 0.150	0.675 ± 0.126	0.866 ± 0.154
80	1.166 ± 0.205	0.760 ± 0.118	0.571 ± 0.112	0.832 ± 0.145
90	0.912 ± 0.136	0.388 ± 0.080	0.658 ± 0.126	0.653 ± 0.114
100	1.031 ± 0.145	0.618 ± 0.128	0.932 ± 0.191	0.860 ± 0.155
B → β Exo Test MAE / kcal mol^{-1} hTL Train Percentages				
10	2.009 ± 0.314	2.797 ± 0.325	2.124 ± 0.354	2.310 ± 0.331
20	1.680 ± 0.203	1.554 ± 0.232	1.789 ± 0.260	1.674 ± 0.231
30	1.217 ± 0.172	1.200 ± 0.168	1.392 ± 0.213	1.270 ± 0.184
40	1.174 ± 0.202	1.197 ± 0.255	1.329 ± 0.182	1.233 ± 0.213
50	1.126 ± 0.181	1.348 ± 0.217	0.704 ± 0.146	1.059 ± 0.181
60	1.077 ± 0.179	0.993 ± 0.194	0.784 ± 0.102	0.951 ± 0.158
70	1.184 ± 0.192	0.771 ± 0.153	0.767 ± 0.128	0.907 ± 0.158
80	1.151 ± 0.190	0.559 ± 0.090	0.665 ± 0.116	0.791 ± 0.132
90	0.880 ± 0.130	0.632 ± 0.113	0.631 ± 0.150	0.715 ± 0.131
100	1.016 ± 0.134	0.814 ± 0.134	1.242 ± 0.198	1.024 ± 0.155

Table S15 - AM1-DFT B → β endo and exo hTL test MAEs across three random states (21,22,23) with averaged MAEs. Metrics provided for different percentages of hTL training data.

B → β Endo Train MAE / kcal mol^{-1} hTL Train Percentages				
Percent of Training Data	Random State			Average
	21	22	23	
10	1.595	0.984	1.039	1.206
20	0.885	1.143	0.866	0.965
30	0.938	0.605	0.616	0.720
40	0.460	0.383	0.721	0.521
50	0.704	0.798	0.496	0.666
60	0.928	0.658	0.599	0.728
70	0.387	0.560	0.427	0.458
80	0.478	0.437	0.385	0.433
90	0.379	0.438	0.278	0.365
100	0.625	0.697	0.854	0.725
B → β Exo Train MAE / kcal mol^{-1} hTL Train Percentages				
10	1.901	1.960	1.079	1.647
20	0.885	1.177	1.142	1.068
30	0.822	0.927	1.084	0.944
40	0.657	0.718	0.923	0.766
50	0.379	0.808	0.576	0.588
60	0.454	0.811	0.617	0.627
70	0.416	0.451	0.445	0.437
80	0.421	0.424	0.394	0.413
90	0.361	0.324	0.382	0.356
100	0.867	0.959	1.106	0.977

Table S16 - AM1-DFT B → β endo and exo hTL train MAEs across three random states (21,22,23) with averaged MAEs. Metrics provided for different percentages of hTL training data.

B → [3+2] Endo Test MAE / kcal mol^{-1} hTL Train Percentages				
Percent of Training Data	Random State			Average
	21	22	23	
10	2.137 ± 0.188	1.801 ± 0.148	2.126 ± 0.204	2.021 ± 0.180
20	1.976 ± 0.183	1.539 ± 0.141	1.382 ± 0.117	1.632 ± 0.147
30	1.748 ± 0.176	1.312 ± 0.124	1.098 ± 0.095	1.386 ± 0.132
40	1.388 ± 0.147	1.063 ± 0.107	1.461 ± 0.149	1.304 ± 0.134
50	1.194 ± 0.128	1.017 ± 0.090	1.366 ± 0.158	1.192 ± 0.125
60	0.912 ± 0.120	0.812 ± 0.086	0.895 ± 0.101	0.873 ± 0.102
70	1.102 ± 0.114	0.724 ± 0.073	0.752 ± 0.087	0.859 ± 0.092
80	0.734 ± 0.082	0.648 ± 0.066	0.747 ± 0.087	0.710 ± 0.078
90	0.977 ± 0.126	0.670 ± 0.079	0.772 ± 0.092	0.806 ± 0.099
100	0.822 ± 0.113	0.774 ± 0.077	0.669 ± 0.082	0.755 ± 0.091
B → [3+2] Exo Test MAE / kcal mol^{-1} hTL Train Percentages				
10	1.813 ± 0.187	1.880 ± 0.180	2.451 ± 0.215	2.048 ± 0.194
20	1.759 ± 0.178	1.380 ± 0.127	1.480 ± 0.120	1.540 ± 0.142
30	1.495 ± 0.150	1.338 ± 0.137	1.250 ± 0.122	1.361 ± 0.136
40	1.327 ± 0.133	1.085 ± 0.118	1.289 ± 0.124	1.233 ± 0.125
50	1.392 ± 0.142	0.945 ± 0.085	1.336 ± 0.149	1.224 ± 0.126
60	0.974 ± 0.120	0.689 ± 0.077	1.122 ± 0.126	0.928 ± 0.108
70	0.943 ± 0.105	0.627 ± 0.066	0.850 ± 0.108	0.807 ± 0.093
80	0.806 ± 0.091	0.736 ± 0.066	1.044 ± 0.095	0.862 ± 0.084
90	0.834 ± 0.106	0.713 ± 0.075	0.772 ± 0.080	0.773 ± 0.087
100	1.257 ± 0.142	0.904 ± 0.081	0.715 ± 0.071	0.959 ± 0.098

Table S17 - AM1-DFT B → [3+2] endo and exo hTL test MAEs across three random states (21,22,23) with averaged MAEs. Metrics provided for different percentages of hTL training data.

B → [3+2] Endo Train MAE / kcal mol^{-1} hTL Train Percentages				
Percent of Training Data	Random State			Average
	21	22	23	
10	1.071	0.723	1.432	1.075
20	0.826	0.817	0.676	0.773
30	0.821	0.663	0.529	0.671
40	0.737	0.800	0.975	0.837
50	0.630	0.689	0.843	0.720
60	0.438	0.534	0.507	0.493
70	0.670	0.473	0.531	0.558
80	0.508	0.453	0.416	0.459
90	0.509	0.450	0.624	0.528
100	0.425	0.534	0.504	0.488
B → [3+2] Exo Train MAE / kcal mol^{-1} hTL Train Percentages				
10	0.812	0.801	1.434	1.016
20	0.629	0.691	0.592	0.637
30	0.528	0.778	0.775	0.694
40	0.457	1.085	1.121	0.888
50	0.676	0.603	1.060	0.780
60	0.452	0.448	0.736	0.545
70	0.544	0.411	0.465	0.473
80	0.399	0.550	0.932	0.627
90	0.413	0.452	0.547	0.470
100	0.817	0.514	0.520	0.617

Table S18 - AM1-DFT B → [3+2] endo and exo hTL train MAEs across three random states (21,22,23) with averaged MAEs. Metrics provided for different percentages of hTL training data.

8. dTL Metrics

Table S19 displays metrics for before and after the dTL procedure. Table S20 and S21 show the test and train MAEs for the dTL work with associated errors at three random states across different dTL training percentage splits. This work was performed on the AM1-DFT dataset from source domain B to target domain β.

AM1 dTL Endo						
Model	Base Train MAE / kcal mol^{-1}	Base Test MAE / kcal mol^{-1}	Base Test R^2	Base Target MAE / kcal mol^{-1}	TL Target MAE / kcal mol^{-1}	TL Target R^2
B → β	0.623	0.711 ± 0.046	0.977	10.920 ± 0.492	1.584 ± 0.284	0.557
AM1 dTL Exo						
B → β	0.933	1.101 ± 0.079	0.980	10.155 ± 0.721	0.781 ± 0.150	0.854

Table S19 - AM1-DFT endo and exo train and base/dTL test metrics. Test metrics for the target are provided for before and after dTL was performed.

B → β Endo Test MAE / kcal mol^{-1} dTL Train Percentages				
Percent of Training Data	Random State			Average
	21	22	23	
10	2.368 ± 0.418	2.630 ± 0.352	2.290 ± 0.378	2.430 ± 0.383
20	1.748 ± 0.263	1.900 ± 0.286	1.999 ± 0.319	1.883 ± 0.289
30	1.428 ± 0.231	1.508 ± 0.212	1.746 ± 0.381	1.561 ± 0.275
40	1.262 ± 0.166	1.020 ± 0.187	1.375 ± 0.161	1.219 0.171
50	1.301 ± 0.194	1.232 ± 0.210	0.839 ± 0.153	1.124 ± 0.182
60	1.274 ± 0.213	0.996 ± 0.158	1.036 ± 0.156	1.102 ± 0.176
70	1.476 ± 0.209	0.797 ± 0.184	0.759 ± 0.129	1.011 ± 0.174
80	1.299 ± 0.221	0.811 ± 0.156	0.862 ± 0.141	0.991 ± 0.173
90	1.167 ± 0.170	0.883 ± 0.161	0.647 ± 0.148	0.899 ± 0.160
100	1.354 ± 0.233	1.146 ± 0.199	1.584 ± 0.284	1.361 ± 0.239
B → β Exo Test MAE / kcal mol^{-1} dTL Train Percentages				
10	2.562 ± 0.427	2.597 ± 0.366	2.486 ± 0.340	2.548 ± 0.378
20	1.802 ± 0.271	2.165 ± 0.268	2.776 ± 0.350	2.248 ± 0.296
30	1.809 ± 0.307	1.288 ± 0.227	1.824 ± 0.310	1.640 ± 0.282
40	1.193 ± 0.291	1.194 ± 0.232	1.372 ± 0.186	1.253 ± 0.236
50	1.229 ± 0.182	1.463 ± 0.179	0.851 ± 0.165	1.18 ± 0.175
60	1.112 ± 0.184	1.104 ± 0.160	0.700 ± 0.122	0.972 ± 0.155
70	1.261 ± 0.213	1.023 ± 0.182	0.817 ± 0.124	1.033 ± 0.173
80	1.502 ± 0.249	1.146 ± 0.187	0.790 ± 0.134	1.146 ± 0.190
90	1.376 ± 0.203	0.970 ± 0.171	0.762 ± 0.142	1.036 ± 0.172
100	1.821 ± 0.310	1.076 ± 0.171	0.781 ± 0.150	1.226 ± 0.210

Table S20 - AM1-DFT B → β endo and exo dTL test MAEs across three random states (21,22,23) with averaged MAEs. Metrics provided for different percentages of dTL training data.

B → β Endo Train MAE / kcal mol^{-1} dTL Train Percentages				
Percent of Training Data	Random State			Average
	21	22	23	
10	1.206	2.035	1.145	1.462
20	0.833	1.716	0.966	1.172
30	0.887	0.908	1.166	0.987
40	0.700	1.054	0.809	0.854
50	0.741	0.808	0.527	0.692
60	0.549	0.629	0.855	0.678
70	0.796	0.523	0.423	0.581
80	0.690	0.523	0.595	0.603
90	0.430	0.633	0.469	0.511
100	0.759	1.130	1.167	1.019
B → β Exo Train MAE / kcal mol^{-1} dTL Train Percentages				
10	1.953	2.497	1.057	1.836
20	0.570	1.634	1.305	1.170
30	1.331	1.035	1.386	1.251
40	0.886	0.565	0.726	0.726
50	0.603	0.710	0.486	0.600
60	0.838	0.602	0.594	0.678
70	0.854	0.624	0.488	0.655
80	0.582	0.427	0.757	0.589
90	0.537	0.628	0.605	0.590
100	1.204	1.070	0.928	1.067

Table S21 - AM1-DFT B → β endo and exo dTL train MAEs across three random states (21,22,23) with averaged MAEs. Metrics provided for different percentages of dTL training data.

9. Dataset Plots

Plots displaying SQM vs. DFT barriers with the chemical accuracy threshold displayed (grey zones). Plots show spread of both training and test sets for the ML models along with pre-ML MAE and R^2 (Fig. S10 – S13).

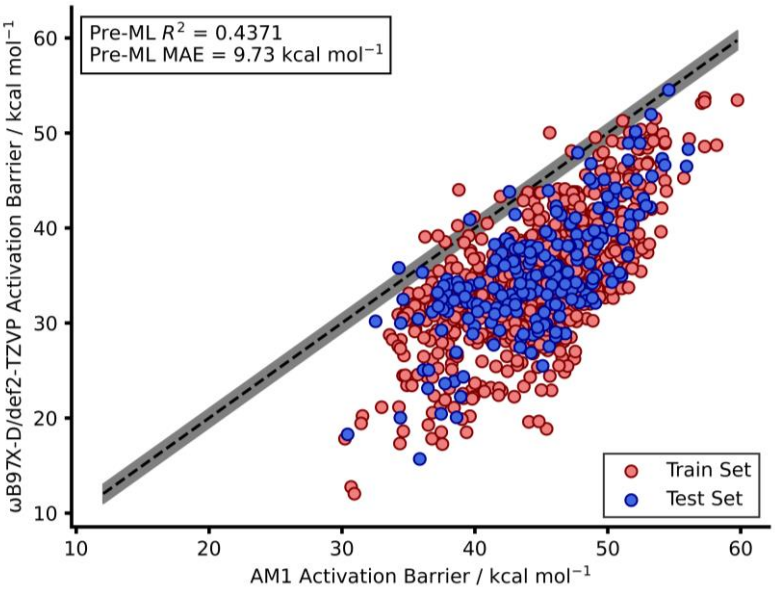

Fig. S10 - AM1 vs DFT endo activation barrier plot with pre-ML R^2 and MAE.

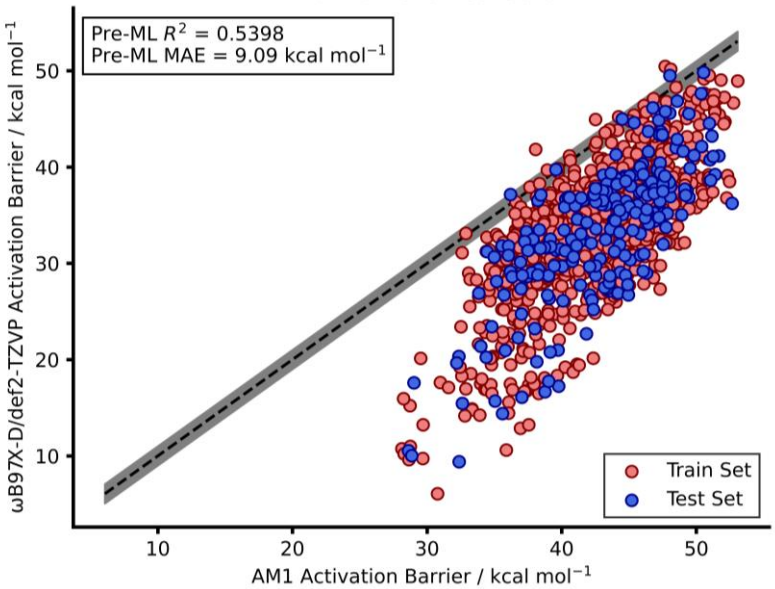

Fig. S11 - AM1 vs DFT exo activation barrier plot with pre-ML R^2 and MAE.

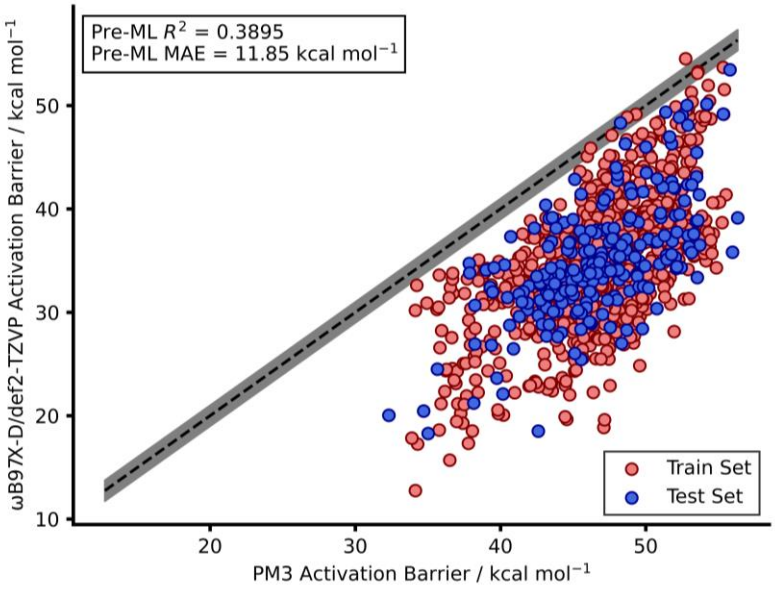

Fig. S12 – PM3 vs DFT endo activation barrier plot with pre-ML R^2 and MAE.

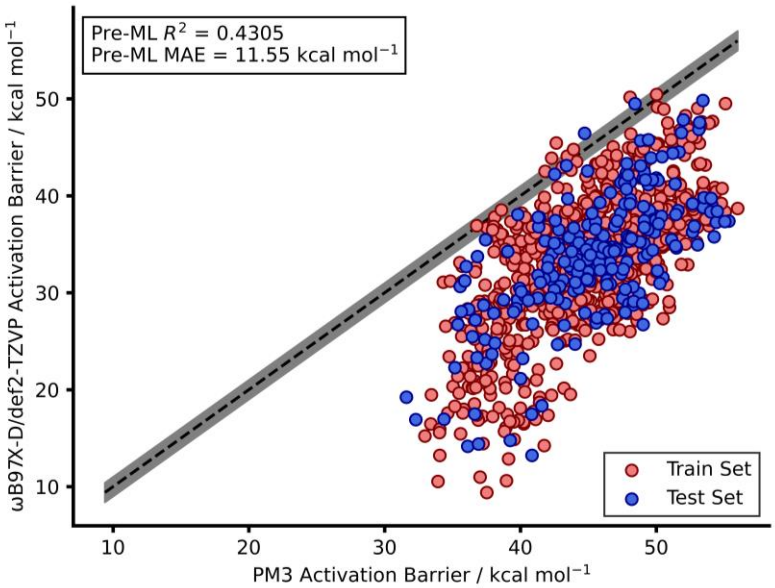

Fig. S13 – PM3 vs DFT exo activation barrier plot with pre-ML R^2 and MAE.

10. Learning Curves

The learning curves for the standard ML approaches on both the endo and exo datasets are provided in Fig. S14-25. In each case, the train and test MAEs match well suggesting that no significant overfitting has occurred. All curves are for AM1 models.

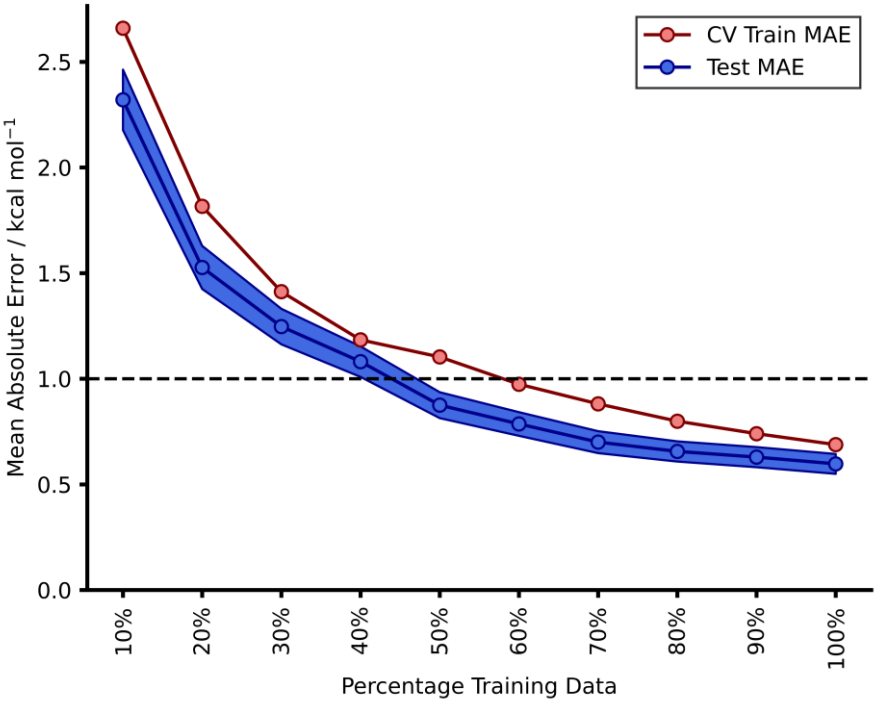

Fig. S14 - AM1-DFT endo KRR (Laplacian) learning curve.

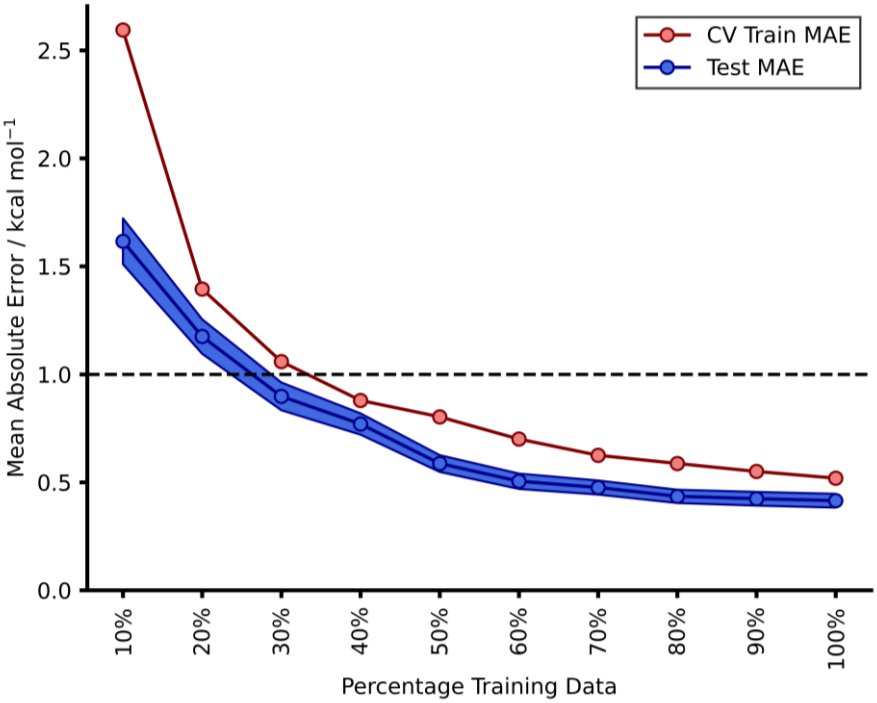

Fig. S15 - AM1-DFT endo KRR (Polynomial) learning curve.

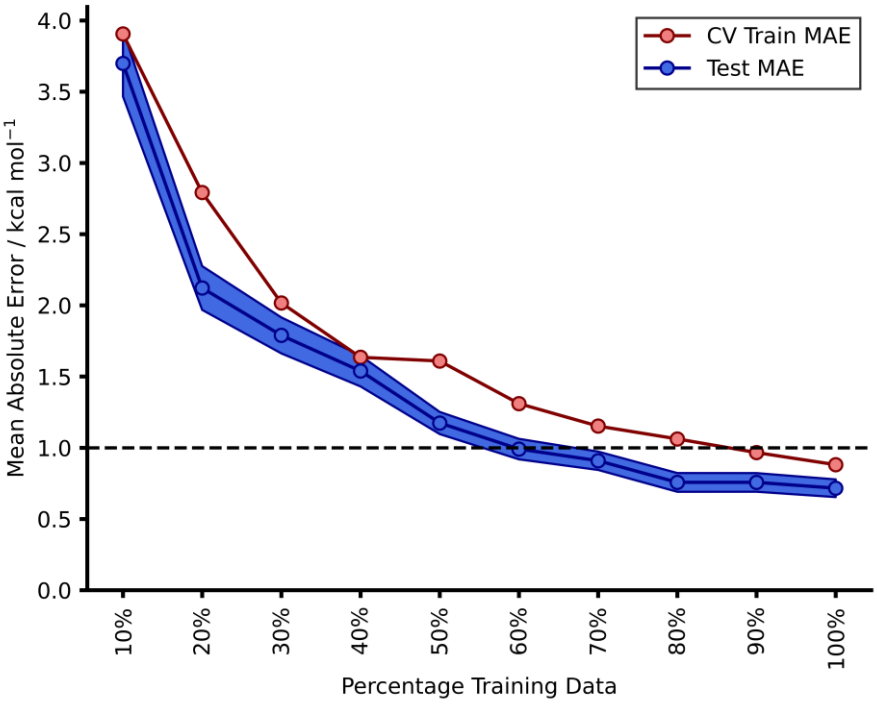

Fig. S16 - AM1-DFT endo KRR (RBF) learning curve.

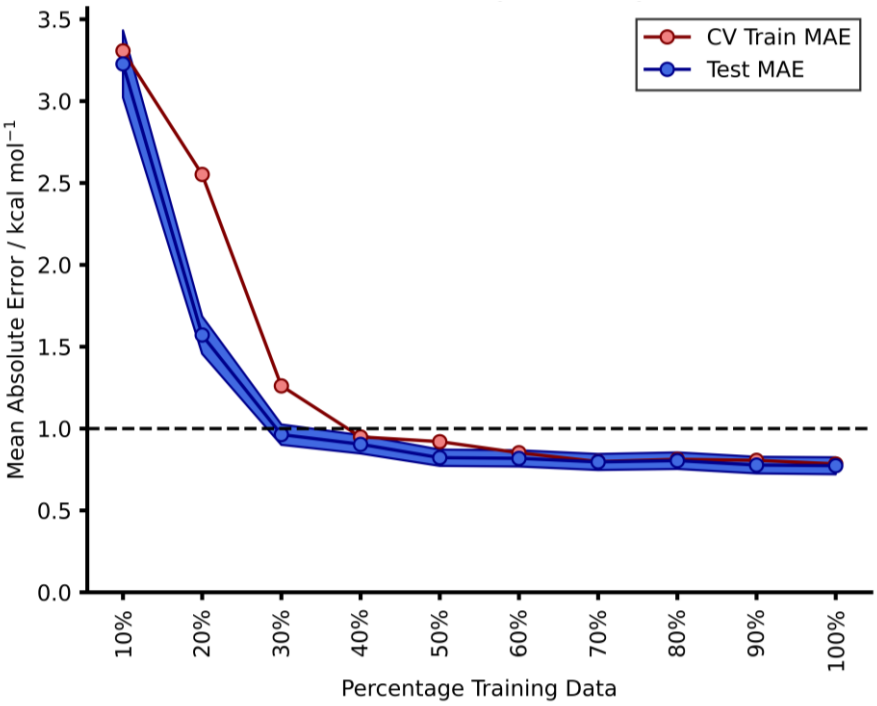

Fig. S17 - AM1-DFT endo Ridge Regression learning curve.

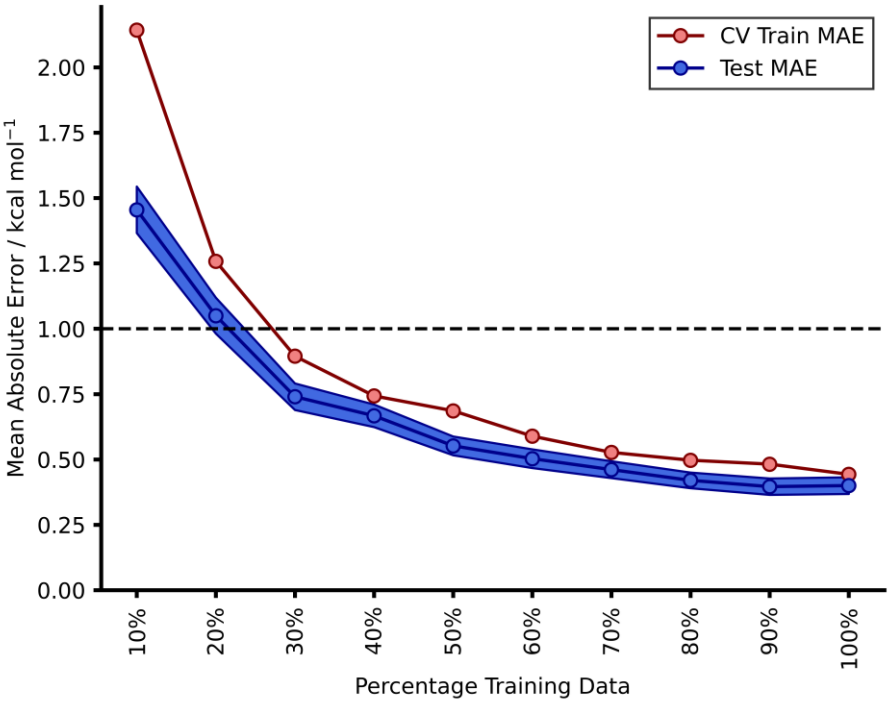

Fig. S18 - AM1-DFT endo SVR (Polynomial) learning curve.

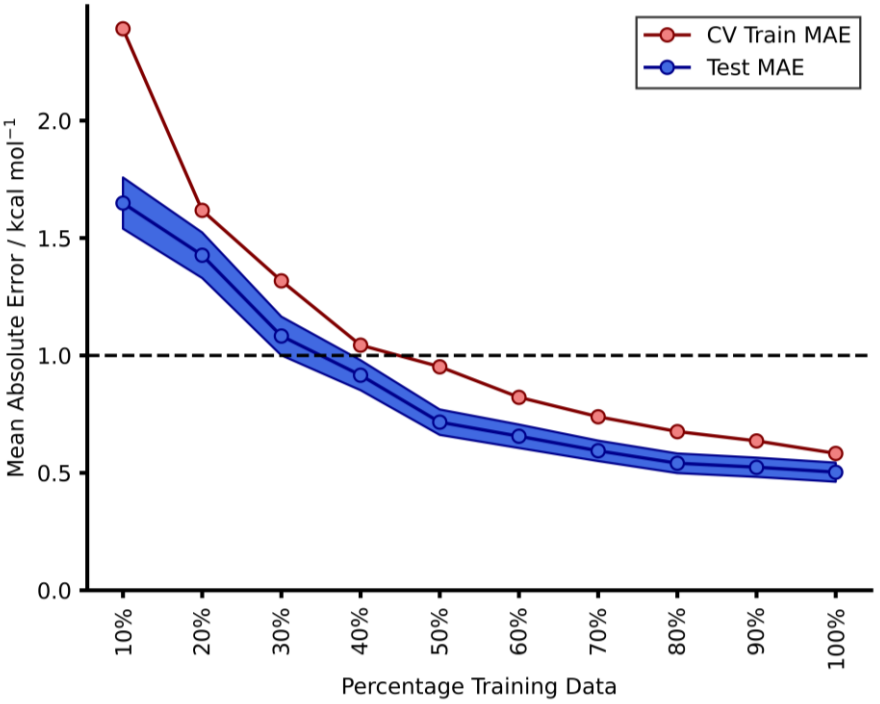

Fig. S19 - AM1-DFT endo SVR (RBF) learning curve.

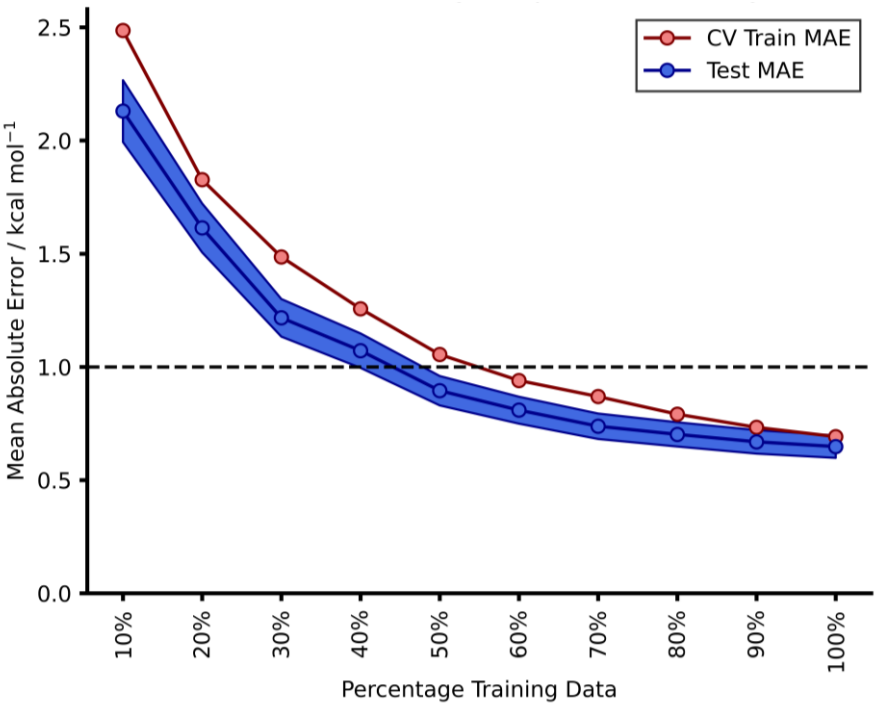

Fig. S20 - AM1-DFT exo KRR (Laplacian) learning curve.

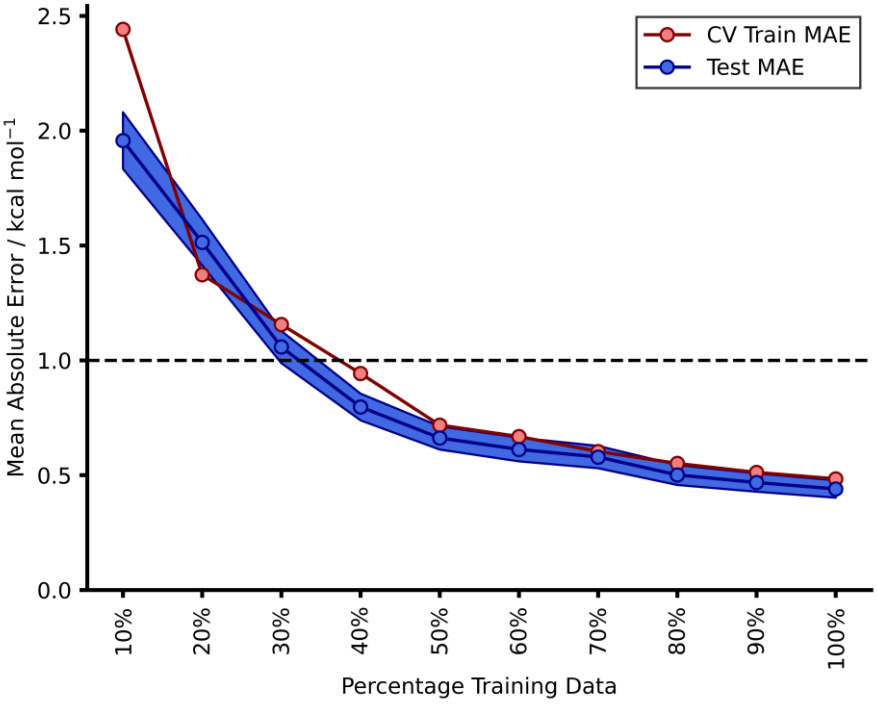

Fig. S21- AM1-DFT exo KRR (Polynomial) learning curve.

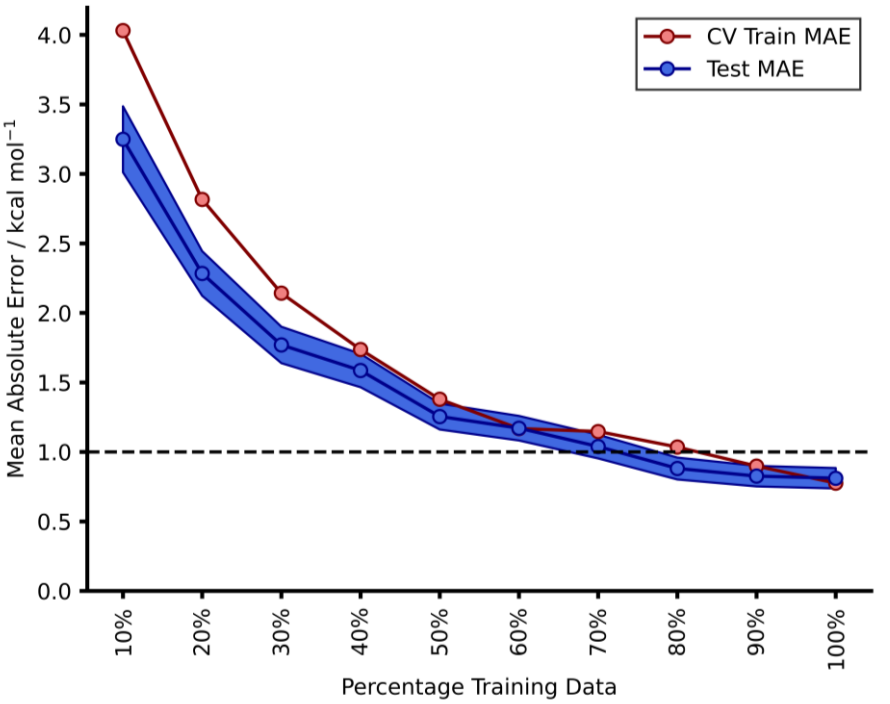

Fig. S22 - AM1-DFT exo KRR (RBF) learning curve.

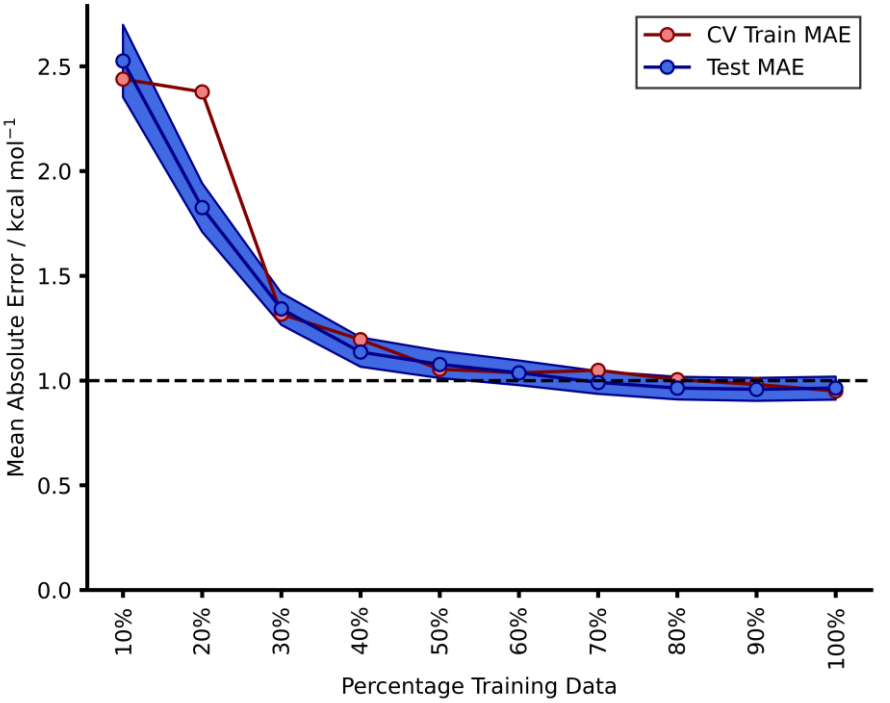

Fig. S23 - AM1-DFT exo Ridge Regression learning curve.

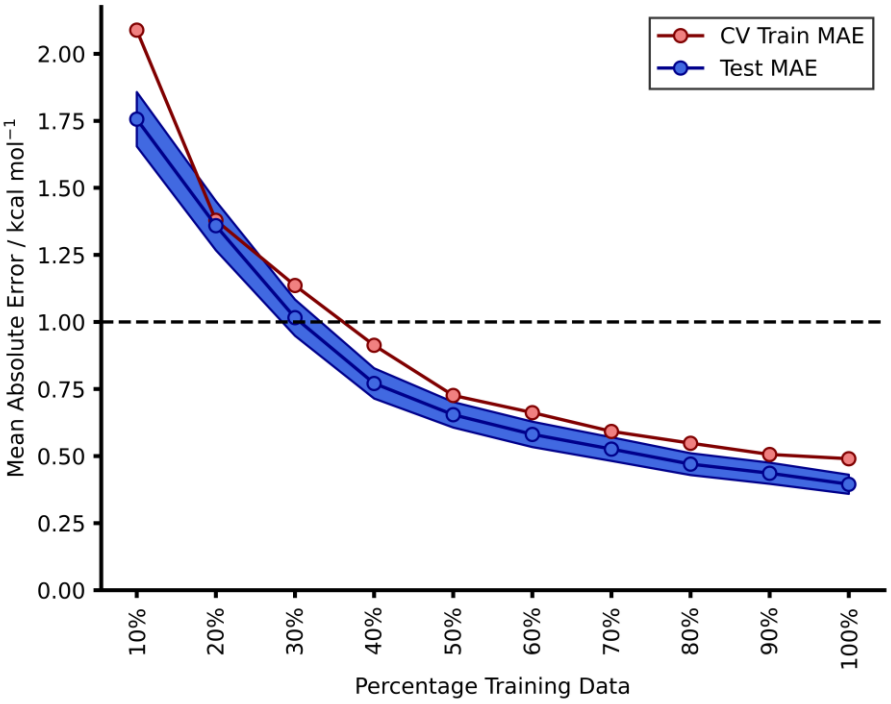

Fig. S24 - AM1-DFT exo SVR (Polynomial) learning curve.

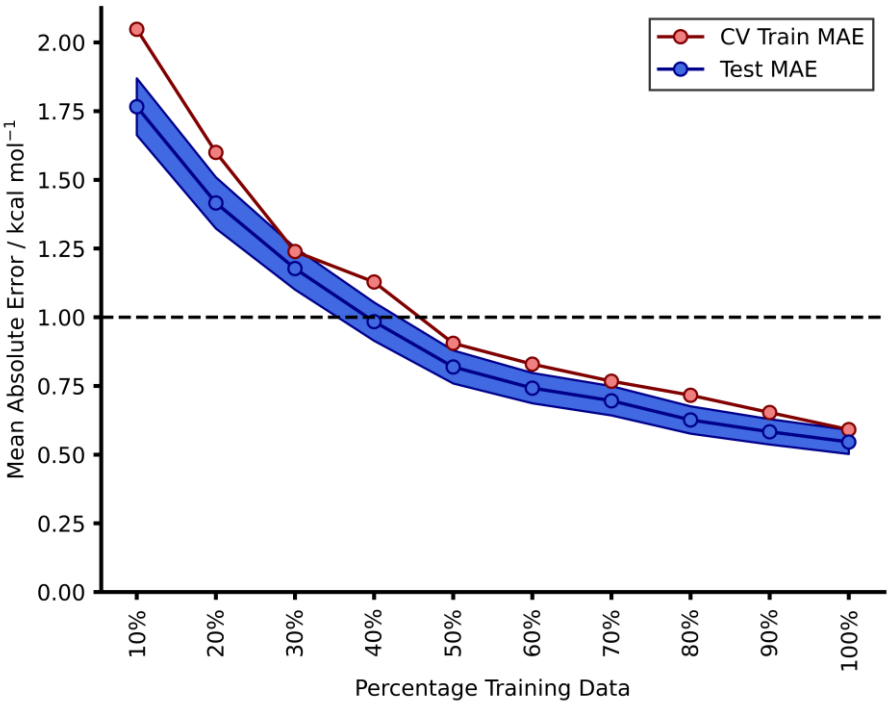

Fig. S25 - AM1-DFT exo SVR (RBF) learning curve.

The learning curves for both hTL and dTL were calculated by taking percentage splits of the target domain training set and performing the hTL/dTL with these splits. The metrics were obtained across three different random states and mean values taken and plotted. These curves look at the lower limits of training data required for obtaining chemically accurate free energy barrier target test set predictions on a different reaction class at the same or at a higher level of theory (LoT) relative to the base model. All curves are for AM1 data. Figures S26-S29 are for hTL and Figures S30 and S31 are for dTL.

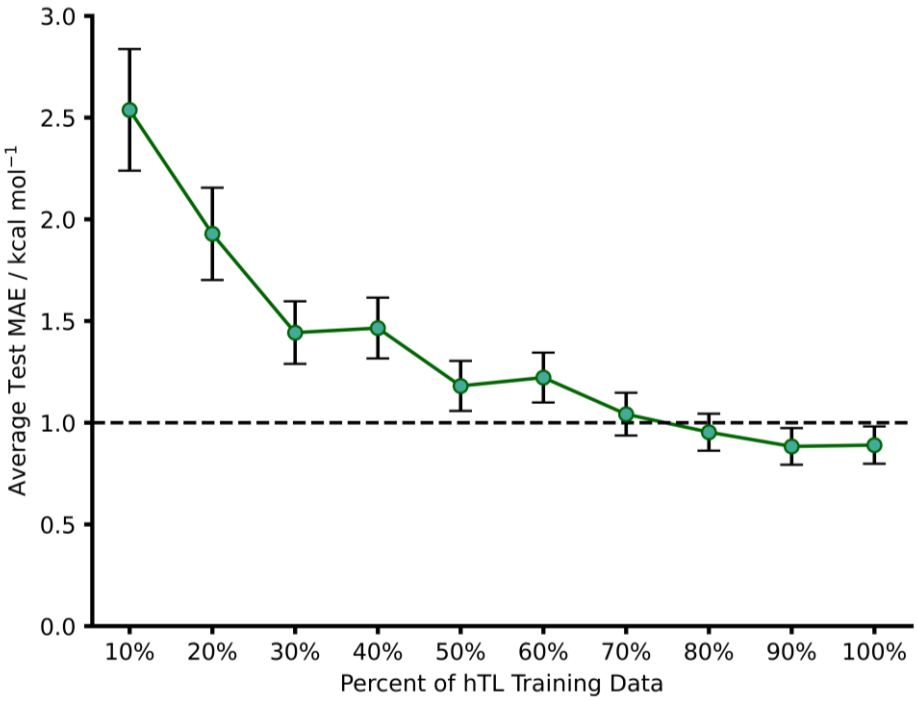

Fig. S26 - AM1-DFT endo A → α hTL target test set learning curve.

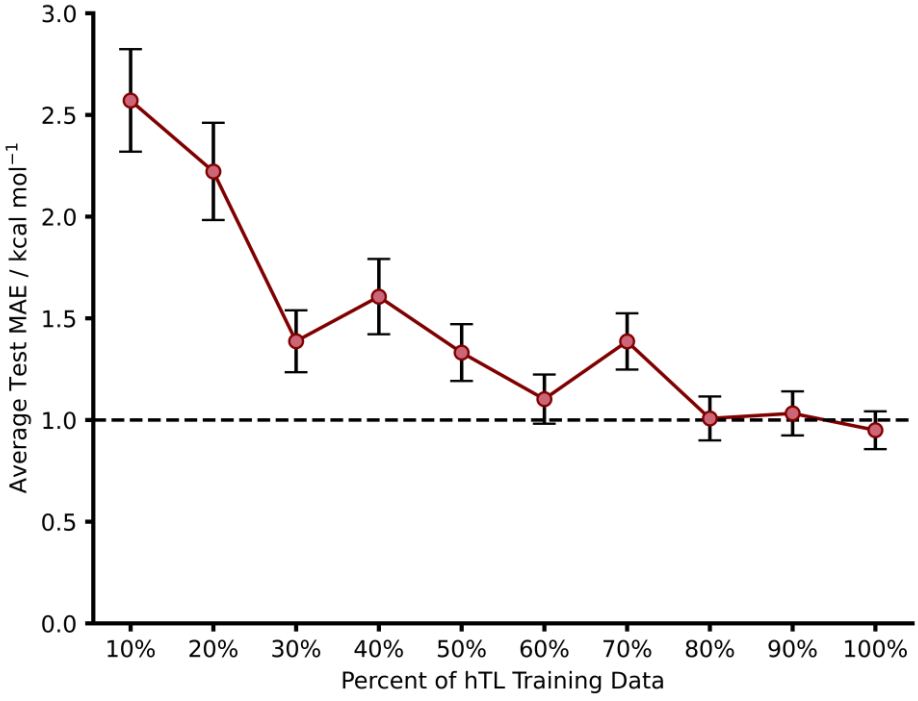

Fig. S27 - AM1-DFT exo A → α hTL target test set learning curve.

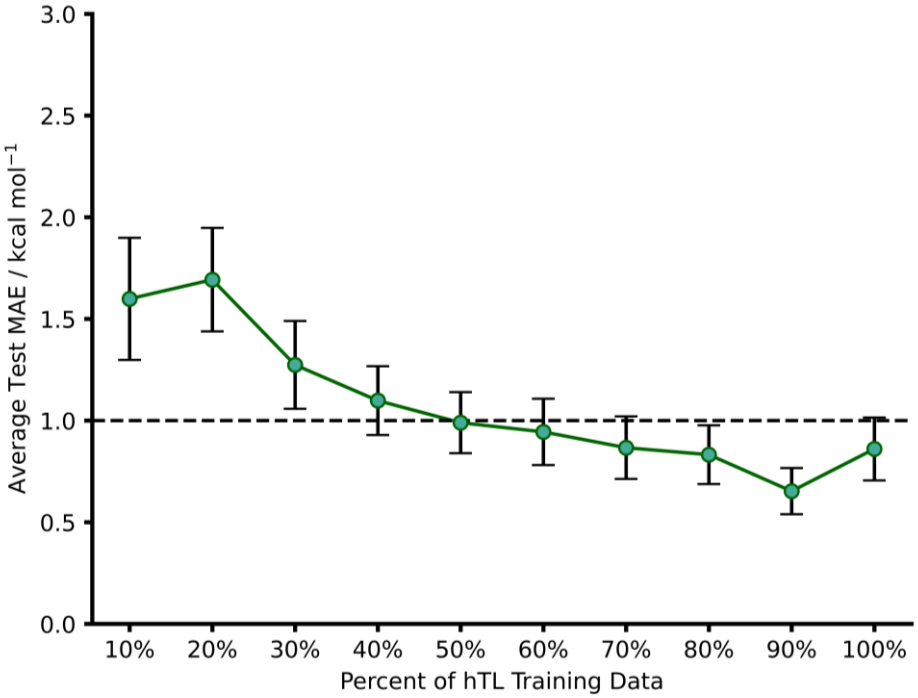

Fig. S28 - AM1-DFT endo B → βhTL target test set learning curve.

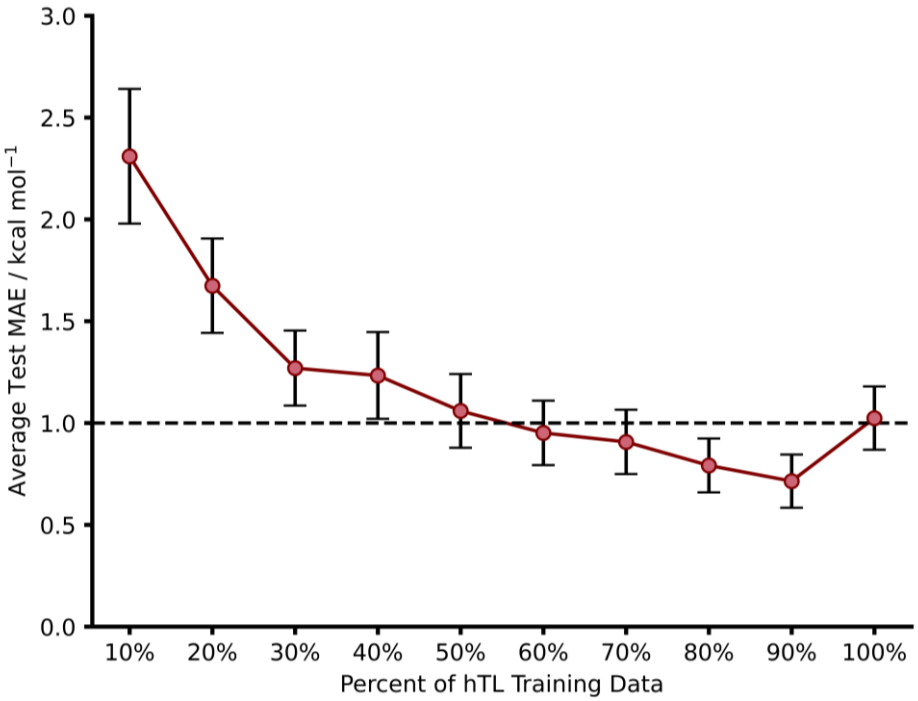

Fig. S29 - AM1-DFT exo B → βhTL target test set learning curve.

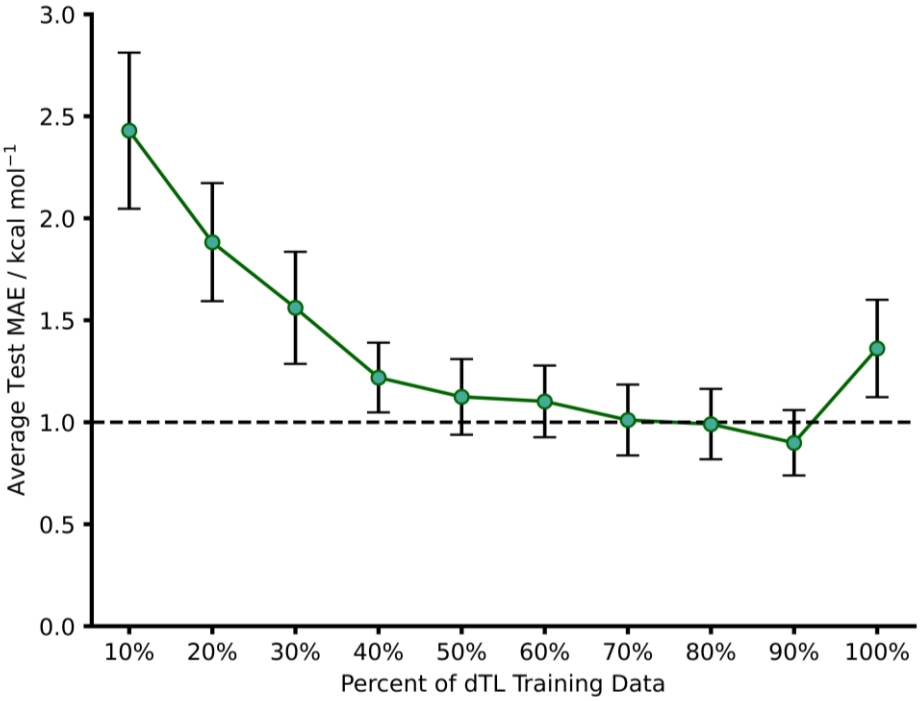

Fig. S30 - AM1-higher LoT DFT endo B → ϐ dTL target test set learning curve.

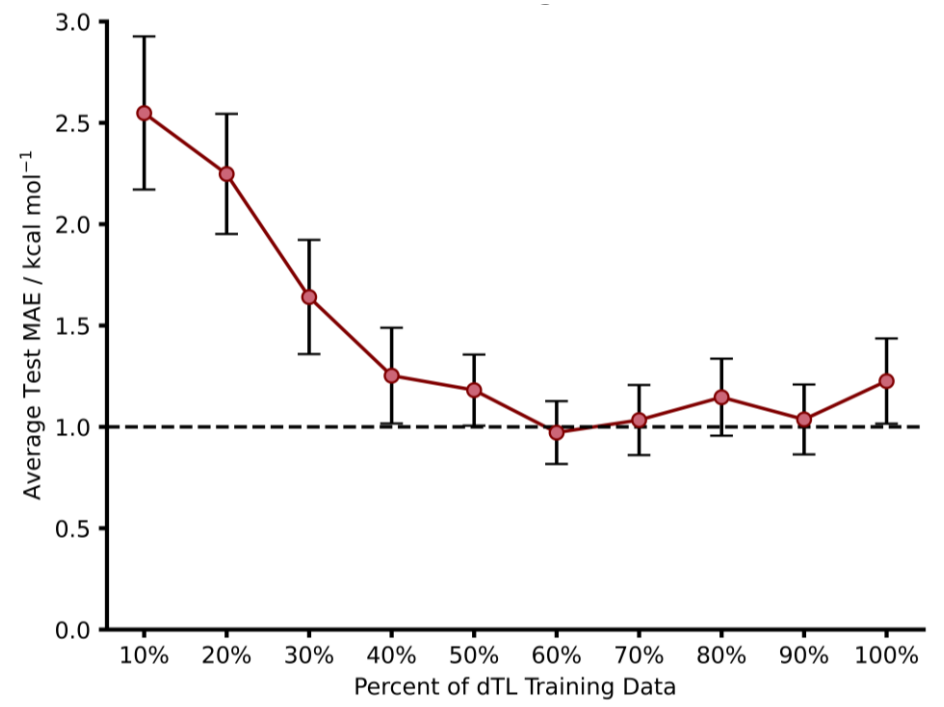

Fig. S31 - AM1-higher LoT DFT exo B → ϐ dTL target test set learning curve.

11. Transition State Structural Analysis

Root-mean-squared deviations of atomic positions (RMSDs) were calculated for both AM1 and PM3 transition state (TS) structures against their respective DFT TS structures. RMSDs were calculated using a quaternion-based characteristic polynomial method[47] with the spyrmsd Python package (Fig. S32 – S34).[48]

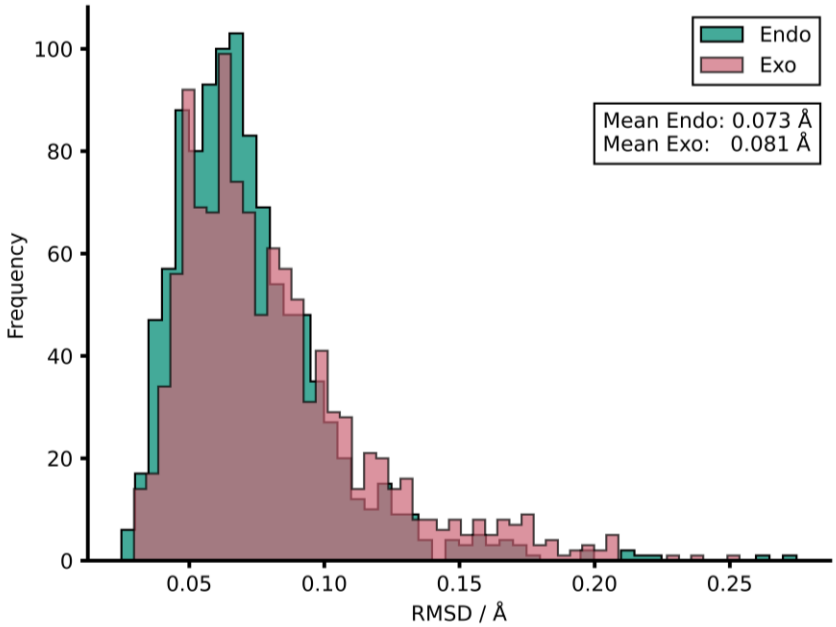

Fig. S32 - AM1-DFT TS RMSDs for both endo and exo structures.

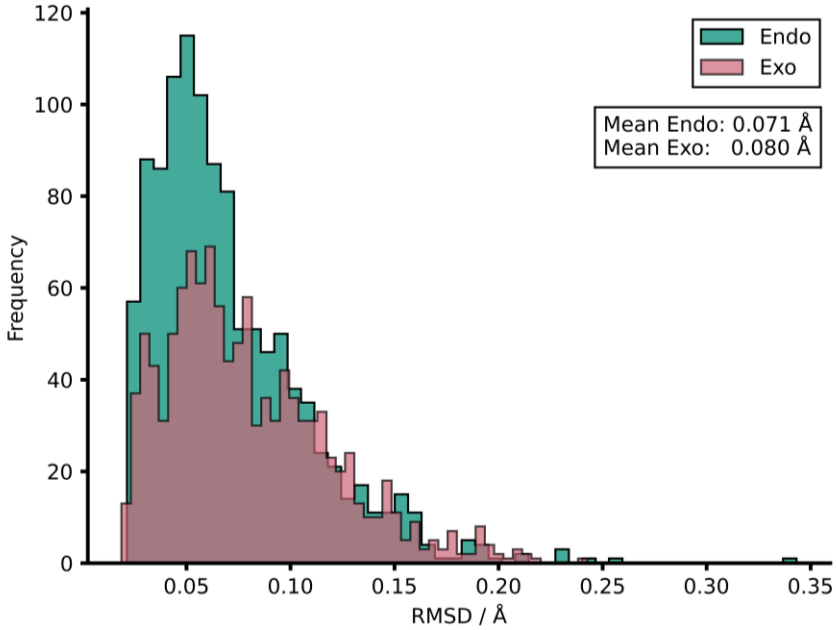

Fig. S33 – PM3-DFT TS RMSDs for both endo and exo structures.

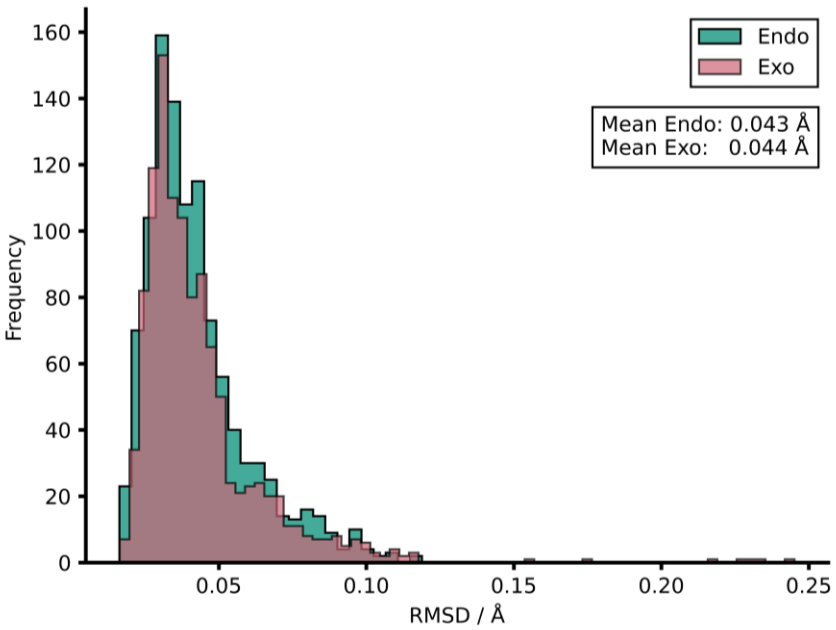

Fig. S34 – PM3-AM1 TS RMSDs for both endo and exo structures.

For TL similarity metric analysis, RDKit was used to generate Morgan fingerprints for each AM1 TS structure in the source and target domain. Tanimoto and Dice similarities were then calculated for every structure in the source domain against every structure in the target domain. Fig. S35 - S38 show A → α and B → β Tanimoto and Dice similarity frequencies, respectively, along with corresponding mean values.

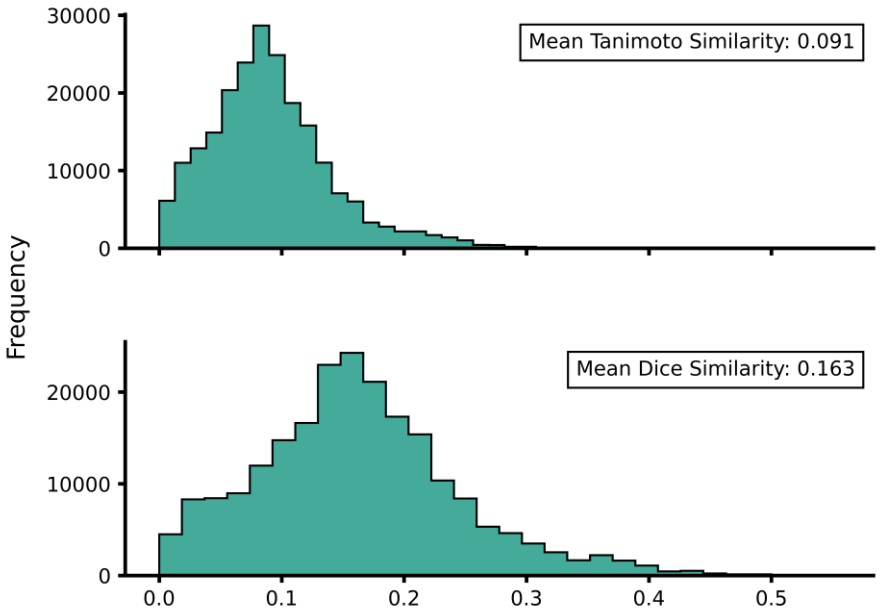

Fig. S35 – Endo AM1 A → α source and target domain Tanimoto and Dice similarities obtained from Morgan fingerprints.

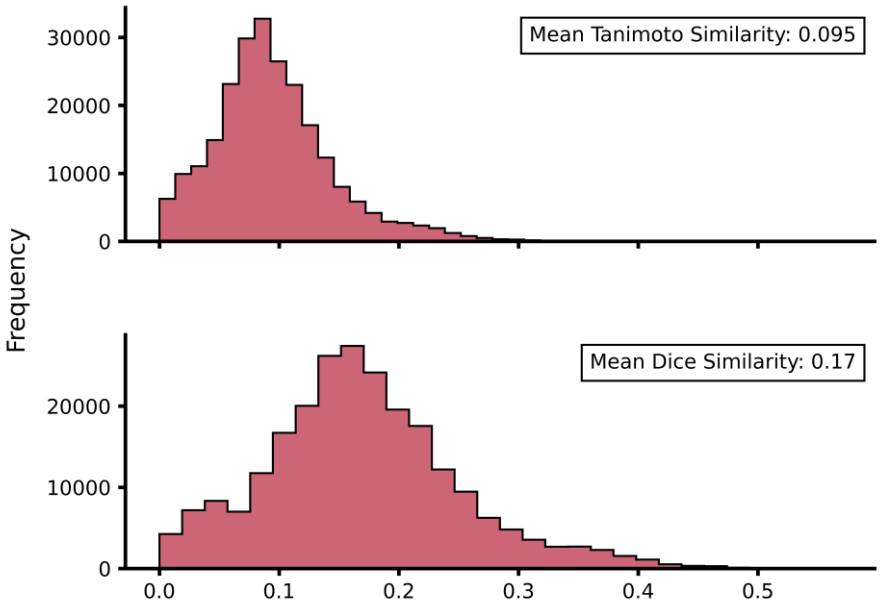

Fig. S36 - Exo AM1 A → α source and target domain Tanimoto and Dice similarities obtained from Morgan fingerprints.

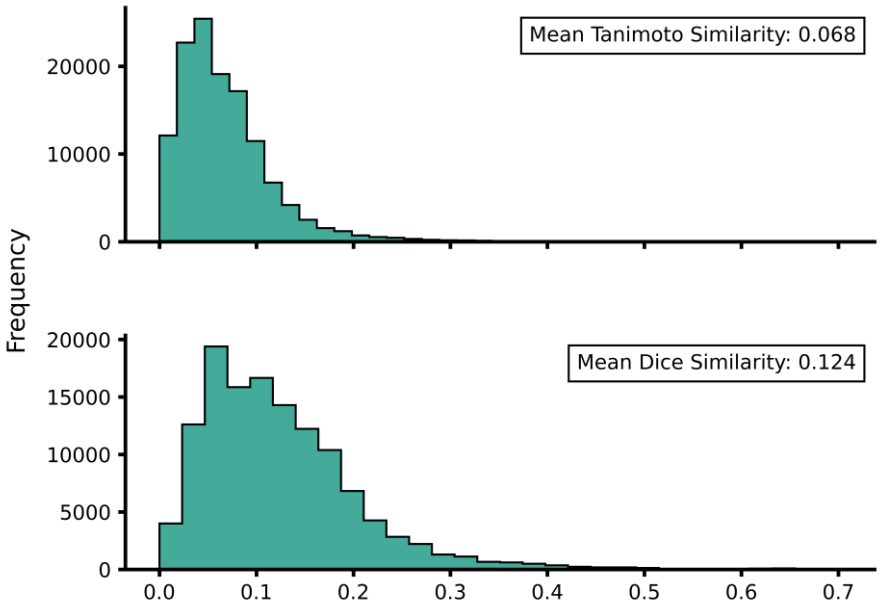

Fig. S37 - Endo AM1 B → β source and target domain Tanimoto and Dice similarities obtained from Morgan fingerprints.

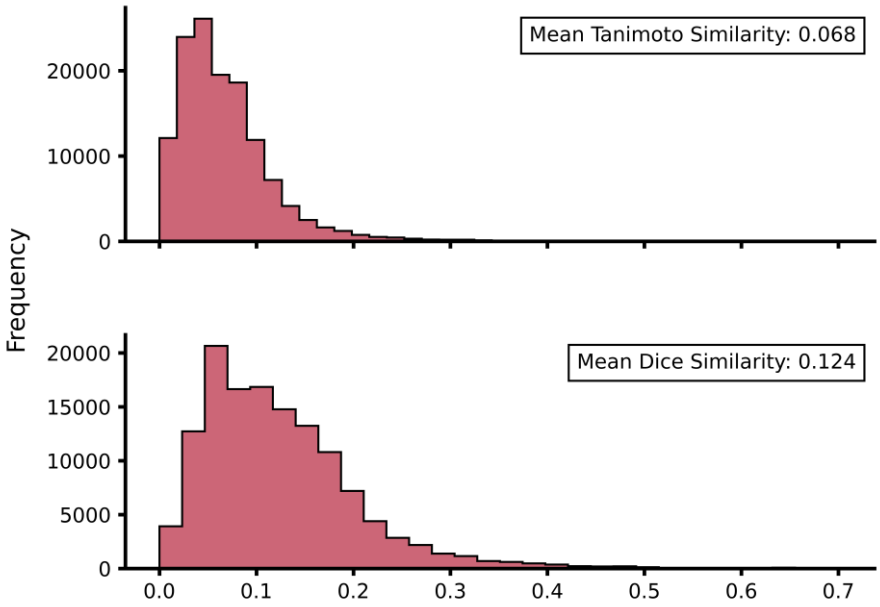

Fig. S38 - Exo AM1 B → β source and target domain Tanimoto and Dice similarities obtained from Morgan fingerprints.

A feature vector similarity approach was also performed on the normalised feature vectors. Below is a step by step of how this was performed:

1. Normalise the feature vectors from both the source and target domains.
2. Calculate the mean for each normalised feature vector in both the source and target domains.
3. Calculate the absolute difference between each normalised feature vector mean from the source and the target domain to give a difference value for every feature vector.
4. Calculate the mean value for all normalised feature vector differences to obtain a singular similarity metric between a given source and target domain.

Table S22 shows the results from this. For comparative purposes, the metric was calculated for the case in which the endo dataset is both the source and target domain. This would yield a metric of 0 as the values are identical thus, values closer to 0 should indicate higher similarity between a given source and target domain.

Source and Target Domain	Similarity Metric
Endo - Endo	0
Exo - Exo	0
Endo A - α	0.0024043
Exo A - α	0.0024942
Endo B - β	0.0041303
Exo B - β	0.0040968
Endo – [3+2]	0.0042639
Exo – [3+2]	0.0043456

Table S22 – Normalised feature vector similarity metrics for combinations of source and target domain.

12. Direct Training

To evaluate overfitting in the extreme low data regimes, the target datasets for hTL and dTL (α and β) as well as the [3+2] dataset were used to train NNs directly. Tables S23-S26 show the results from this direct training.

α Endo Test MAE / kcal mol^{-1} Direct Training Train Percentages				
Percent of Training Data	Random State			Average
	21	22	23	
10	3.122 ± 0.257	2.924 ± 0.315	3.449 ± 0.323	3.165 ± 0.298
20	2.341 ± 0.230	1.935 ± 0.207	2.613 ± 0.294	2.296 ± 0.243
30	1.994 ± 0.219	1.849 ± 0.219	1.363 ± 0.123	1.736 ± 0.187
40	1.608 ± 0.190	1.476 ± 0.161	1.546 ± 0.142	1.544 ± 0.164
50	1.306 ± 0.154	1.617 ± 0.172	1.349 ± 0.125	1.424 ± 0.15
60	1.362 ± 0.159	1.566 ± 0.16	1.264 ± 0.13	1.397 ± 0.149
70	1.208 ± 0.161	1.227 ± 0.132	1.106 ± 0.089	1.18 ± 0.127
80	1.328 ± 0.170	1.194 ± 0.129	1.032 ± 0.109	1.185 ± 0.136
90	1.111 ± 0.149	1.121 ± 0.098	0.928 ± 0.102	1.053 ± 0.117
100	0.994 ± 0.112	0.998 ± 0.087	0.975 ± 0.08	0.989 ± 0.093
α Endo Train MAE / kcal mol^{-1} Direct Training Train Percentages				
10	1.001	1.273	0.907	1.06
20	1.045	1.315	1.242	1.2
30	0.852	0.958	0.684	0.831
40	1.004	0.914	0.814	0.911
50	0.503	0.700	0.710	0.638
60	0.802	0.705	0.868	0.791
70	0.558	0.479	0.566	0.534
80	0.791	0.569	0.770	0.71
90	0.683	0.537	0.647	0.622
100	0.609	0.484	0.503	0.532

Table S23 - AM1-DFT α endo direct training test and train MAEs across three random states (21,22,23) with averaged MAEs. Metrics provided for different percentages of direct training data.

α Exo Test MAE / kcal mol^{-1} Direct Training Train Percentages				
Percent of Training Data	Random State			Average
	21	22	23	
10	4.643 ± 0.417	4.607 ± 0.417	2.328 ± 0.206	3.859 ± 0.347
20	1.856 ± 0.176	2.953 ± 0.386	2.948 ± 0.318	2.585 ± 0.293
30	1.825 ± 0.172	1.314 ± 0.155	1.787 ± 0.228	1.642 ± 0.185
40	1.483 ± 0.178	1.493 ± 0.172	1.626 ± 0.197	1.534 ± 0.182
50	1.065 ± 0.112	1.343 ± 0.148	1.286 ± 0.161	1.231 ± 0.140
60	0.975 ± 0.118	1.040 ± 0.123	1.687 ± 0.189	1.234 ± 0.143
70	1.602 ± 0.129	1.551 ± 0.205	2.442 ± 0.242	1.865 ± 0.192
80	0.866 ± 0.088	1.322 ± 0.141	1.207 ± 0.115	1.132 ± 0.115
90	1.000 ± 0.093	1.240 ± 0.127	1.129 ± 0.131	1.123 ± 0.117
100	0.846 ± 0.116	1.120 ± 0.126	1.100 ± 0.111	1.022 ± 0.118
α Exo Train MAE / kcal mol^{-1} Direct Training Train Percentages				
10	1.753	1.577	1.877	1.736
20	1.327	0.925	1.691	1.314
30	0.881	0.511	0.848	0.747
40	0.787	0.604	0.747	0.713
50	0.613	0.626	0.588	0.609
60	0.520	0.524	1.384	0.809
70	1.204	1.169	2.082	1.485
80	0.536	0.632	0.717	0.628
90	0.721	0.810	0.639	0.724
100	0.407	0.722	0.668	0.599

Table S24 - AM1-DFT α exo direct training test and train MAEs across three random states (21,22,23) with averaged MAEs. Metrics provided for different percentages of direct training data.

β Test MAE / kcal mol^{-1} Direct Training Train Percentages				
Percent of Training Data	Random State			Average
	21	22	23	
10	2.054 ± 0.410	1.958 ± 0.177	1.629 ± 0.321	1.880 ± 0.303
20	1.318 ± 0.268	1.997 ± 0.249	1.343 ± 0.324	1.553 ± 0.280
30	1.080 ± 0.184	0.896 ± 0.141	1.476 ± 0.242	1.151 ± 0.189
40	1.049 ± 0.160	1.003 ± 0.180	1.315 ± 0.172	1.122 ± 0.171
50	1.032 ± 0.187	0.686 ± 0.123	0.959 ± 0.158	0.892 ± 0.156
60	1.003 ± 0.206	0.628 ± 0.120	0.453 ± 0.079	0.694 ± 0.135
70	1.108 ± 0.203	0.659 ± 0.131	0.612 ± 0.106	0.793 ± 0.147
80	1.035 ± 0.184	0.404 ± 0.088	0.503 ± 0.076	0.647 ± 0.116
90	0.927 ± 0.142	0.529 ± 0.095	0.475 ± 0.082	0.644 ± 0.107
100	0.857 ± 0.139	0.533 ± 0.098	0.496 ± 0.076	0.629 ± 0.104
β Train MAE / kcal mol^{-1} Direct Training Train Percentages				
10	1.161	1.038	0.425	0.875
20	0.483	1.356	0.394	0.744
30	0.326	0.375	0.579	0.427
40	0.368	0.392	0.387	0.382
50	0.393	0.342	0.388	0.374
60	0.349	0.387	0.408	0.381
70	0.351	0.501	0.466	0.44
80	0.412	0.368	0.409	0.396
90	0.445	0.445	0.366	0.419
100	0.517	0.509	0.513	0.513

Table S25 - AM1-DFT β direct training test and train MAEs across three random states (21,22,23) with averaged MAEs. Metrics provided for different percentages of direct training data.

[3+2] Test MAE / kcal mol^{-1} Direct Training Train Percentages				
Percent of Training Data	Random State			Average
	21	22	23	
10	2.116 ± 0.182	2.116 ± 0.189	2.716 ± 0.203	2.316 ± 0.191
20	2.108 ± 0.178	1.377 ± 0.119	1.373 ± 0.114	1.619 ± 0.137
30	1.660 ± 0.157	1.284 ± 0.127	1.176 ± 0.099	1.373 ± 0.128
40	1.273 ± 0.149	1.014 ± 0.101	1.118 ± 0.086	1.135 ± 0.112
50	1.310 ± 0.138	1.052 ± 0.094	1.088 ± 0.079	1.150 ± 0.104
60	0.947 ± 0.124	0.718 ± 0.081	0.886 ± 0.096	0.851 ± 0.100
70	0.880 ± 0.098	0.850 ± 0.087	0.689 ± 0.072	0.806 ± 0.085
80	0.812 ± 0.090	0.559 ± 0.061	0.783 ± 0.077	0.718 ± 0.076
90	0.789 ± 0.077	0.632 ± 0.068	0.620 ± 0.079	0.680 ± 0.075
100	0.749 ± 0.080	0.626 ± 0.060	0.616 ± 0.060	0.664 ± 0.067
[3+2] Train MAE / kcal mol^{-1} Direct Training Train Percentages				
10	0.966	0.417	0.775	0.720
20	0.737	0.645	0.520	0.634
30	0.430	0.757	0.575	0.587
40	0.488	0.720	0.738	0.649
50	0.665	0.785	0.614	0.688
60	0.434	0.438	0.495	0.456
70	0.442	0.613	0.413	0.489
80	0.502	0.359	0.572	0.478
90	0.553	0.423	0.450	0.475
100	0.429	0.433	0.374	0.412

Table S26 - AM1-DFT [3+2] direct training test and train MAEs across three random states (21,22,23) with averaged MAEs. Metrics provided for different percentages of direct training data.

13. References

1 L. M. Harwood, G. Jones, J. Pickard, R. M. Thomas and D. Watkin, *Tetrahedron Lett.*, 1988, **29**, 5825–5828.

2 L. M. Harwood, S. A. Leeming, N. S. Isaacs, G. Jones, J. Pickard, R. M. Thomas and D. Watkin, *Tetrahedron Lett.*, 1988, **29**, 5017–5020.

3 R. Gordillo and K. N. Houk, *J. Am. Chem. Soc.*, 2006, **128**, 3543–3553.

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

5 P. Binger, P. Wedemann, R. Goddard and U. H. Brinker, *J. Org. Chem.*, 1996, **61**, 6462–6464.

6 L. A. Fisher, N. J. Smith and J. M. Fox, *J. Org. Chem.*, 2013, **78**, 3342–3348.

7 F. Liu, R. S. Paton, S. Kim, Y. Liang and K. N. Houk, *J. Am. Chem. Soc.*, 2013, **135**, 15642–15649.

8 R. Ukis and C. Schneider, *J. Org. Chem.*, 2019, **84**, 7175–7188.

9 V. Eschenbrenner-Lux, K. Kumar and H. Waldmann, *Angew. Chem. Int. Ed.*, 2014, **53**, 11146–11157.

10 D. v. Osipov, V. A. Osyanin, G. D. Khaysanova, E. R. Masterova, P. E. Krasnikov and Y. N. Klimochkin, *J. Org. Chem.*, 2018, **83**, 4775–4785.

11 S. N. Pieniazek and K. N. Houk, *Angew. Chem. Int. Ed.*, 2006, **45**, 1442–1445.

12 N. K. Devaraj, R. Weissleder and S. A. Hilderbrand, *Bioconjug. Chem.*, 2008, **19**, 2297–2299.

13 F. Liu, Y. Liang and K. N. Houk, *J. Am. Chem. Soc*, 2014, **136**, 11483–11493.

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

15 F. Mohamadi, N. G. J. Richards, W. C. Guida, R. Liskamp, M. Lipton, C. Caufield, G. Chang, T. Hendrickson and W. C. Still, *J. Comput. Chem.*, 1990, **11**, 440–467.

16 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.

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

18 J. J. P. Stewart, *J. Comput. Chem.*, 1989, **10**, 221–264.

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

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

21 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.

22 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.

23 S. Kozuch and J. M. L. Martin, *Phys. Chem. Chem. Phys.*, 2011, **13**, 20104–20107.

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

25 D. Margetic and R. N. Warrener, *Croat. Chem. Acta*, 2003, **76**, 357–363.

26 C. Cativiela, V. Dillet, J. I. García, J. A. Mayoral, M. F. Ruiz-López and L. Salvatella, *J. Mol. Struct. (Theochem)*, 1995, **331**, 37–50.

27 B. S. Jursic and Z. Zdravkovski, *Tetrahedron*, 1994, **50**, 10379–10390.

28 T. H. Musslimani and H. Mettee, *J. Mol. Struct. (Theochem),* 2004, **672**, 35–43.

29 J. J. P. Stewart, *J. Mol. Model.*, 2007, **13**, 1173–1213.

30 J. J. P. Stewart, *J. Mol. Model.*, 2013, **19**, 1–32.

31 N. Mardirossian and M. Head-Gordon, *Mol. Phys.*, 2017, **115**, 2315–2372.

32 L. Goerigk and S. Grimme, *WIREs Comput. Mol. Sci.*, 2014, **4**, 576–600.

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

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

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

36 R. M. LoPachin, T. Gavin, A. DeCaprio and D. S. Barber, *Chem. Res. Toxicol.*, 2012, **25**, 239–251.

37 S. Mitternacht, S. J. Hubbard and Y. Zhou, *F1000Research*, **5**, 189.

38 G. Luchini and R. Paton, *DBSTEP: 1.2-alpha Release*, 2021.

39 N. M. O'Boyle, C. Morley and G. R. Hutchison, *Chem. Cent. J.*, 2008, 2.

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

41 Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S.Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Rafal Jozefowicz, Yangqing Jia, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Mike Schuster, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu and Xiaoqiang Zheng, *TensorFlow: Large-scale machine learning on heterogeneous systems*, 2015.

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

43 A. E. Hoerl and R. W. Kennard, *Technometrics*, 1970, **12**, 55.

44 V. Vovk, *Empirical Inference*, Springer Berlin Heidelberg, 2013.

45 C. Cortes, V. Vapnik and L. Saitta, *Mach. Learn.*, 1995, **20**, 273–297.

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

47 D. L. Theobald, *Acta Crystallogr., Sect. A: Found. Crystallogr.*, 2005, **61**, 478–480.

48 R. Meli and P. C. Biggin, *J. Cheminform.*, 2020, **12**, 1–7.