Learning Activation Energies of Enantioselective Catalytic Reactions

A Thesis

submitted to

Indian Institute of Science Education and Research Pune

in partial fulfilment of the requirements for the

BS-MS Dual Degree Programme

by

Sinjini Bhattacharjee

Indian Institute of Science Education and Research Pune

Dr. Homi Bhabha Road,

Pashan, Pune 411008, INDIA.

April, 2020

Project Supervisor

Prof. Clémence Corminboeuf

Department of Chemistry and Chemical Engineering

École Polytechnique Fédérale de Lausanne (EPFL)

Switzerland

“Blessed are the curious for they shall have adventures…..”

-Lovelle Dracman

“This dissertation is dedicated to my loving parents and grandparents for all their love and encouragement.

Also, to my dear teachers, for helping me achieve all the success I have.”

Certificate

*This is to certify that this dissertation entitled "***Learning Activation Energies of Enantioselective Catalytic Reactions***" towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents study/work carried out by* ***Miss Sinjini Bhattacharjee*** *at* ***École Polytechnique Fédérale de Lausanne (EPFL) Switzerland*** *under the supervision of* ***Prof. Clemence Corminboeuf****, LCMD,* ***Department of Chemistry and Chemical Engineering****, during the academic year* ***2019-2020****.*

Sinjini Bhattacharjee

Signature of the Student

Signature of the Project Supervisor

Declaration

*I hereby declare that the matter embodied in the report entitled "**Learning Activation Energies of Enantioselective Catalytic Reactions**" are the results of the work carried out by me at the Department of Chemistry and Chemical Engineering, at **École Polytechnique Fédérale de Lausanne (EPFL) Switzerland**, under the supervision of **Prof. Clemence Corminboeuf** and the same has not been submitted elsewhere for any other degree.*

Sinjini Bhattacharjee

Signature of the Student

Signature of the Project Supervisor

Acknowledgements

The National Centre of Competence in Research (NCCR) “Materials’ Revolution: Computational Design and Discovery of Novel Materials (MARVEL)” of the Swiss National Science Foundation (SNSF) and EPFL are acknowledged for financial support. This work was partially supported by the ***MARVEL INSPIRE Potentials Master's Fellowship****.*

I would like to acknowledge Dr. Matthew Wodrich, Dr. Benjamin Meyer for their supervision along with many helpful discussions and critical reading of the report and Dr. Alberto Fabrizio for all the technical guidance throughout this project.

I am grateful to Prof. Clemence Corminboeuf for her supervision and for giving me the opportunity to work on such an innovative project. I would like to thank all the LCMD members for providing me with the perfect work environment.

I would like to thank the Faculty and entire community at ***IISER Pune*** *for their unconditional support throughout my studies. I am especially thankful to my friends for creating all the much-cherished memories at IISER, Pune.*

Contents

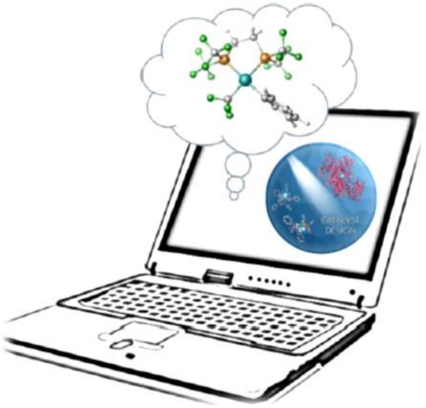
List of Figures

Abstract

The movement toward large-scale screening studies aimed at understanding and predicting the behaviour of organocatalysts poses significant challenges in computational chemistry. The primary bottleneck in studying these systems using traditional techniques rooted in density functional theory is the effort required to locate the computationally expensive transition states (TS). As such, it would be ideal to establish a suitable theoretical model capable of quickly and accurately predicting this critically important data with a minimal computational cost. Historically, concepts based on Linear Scaling Relationships (LSRs), such as the Bell-Evans-Polanyi (BEP) principle that relates the activation barrier and enthalpy of analogous reactions, provided practical, simple to use guidelines for estimating transition states. Here, we seek to establish a quantitatively more accurate relationship beyond simple linear regressions and leverage machine learning to estimate the TS activation barriers. To accomplish this, we directly optimize geometries and establish the energies associated with key intermediates using a variety of inexpensive theoretical levels, such as semiempirical methods. The energies are then used to train machine learning (ML) models by applying a non-linear regression, which provides an approximation of the TS energies at the target DFT level directly from the energies of intermediates computed using the aforementioned methods. In essence, this procedure is an analogue to the BEP principle, which, rather than relying on LSRs, uses non-linear regression and machine learning to draw connections between the structures and energies of intermediates with the associated activation barriers. The energetic data obtained using this ML framework also extends beyond simple BEP type relationships and could be used to accurately predict targeted chemical properties (e.g., stereoselectivity) with minimal computation cost.

1. Introduction

A significant portion of chemical reactions involved in industry and academia rely on catalysts in order to ensure better yield and selectivity of products. Several tools have been developed to improve on random search procedures, including combinatorial chemistry[1-3], high-throughput screening[4-6] and computational methods[7-10] have recently been developed to accelerate the identification of efficient catalysts.



Traditional approaches to computational-based catalyst screening generally involved generation of free energy profiles or, more recently, microkinetic reaction modelling using transition-state theory.[11-12] In principle, all of the necessary information for such detailed modelling, including the energies of all catalytic cycle intermediates and transition states, can be obtained from density functional theory (DFT) calculations. Furthermore**,** computational quantum chemistry has seen considerable developments over the past and one can now obtain crucial mechanistic insights into a multitude of organocatalyzed reactions through applications of modern density functional theory (DFT) methods[13-15]. However, the rational design of organo-catalysts poses a significant challenge and potential catalyst design still relies extensively upon experimental screening techniques. Despite this, density functional theory (DFT) has been successfully applied to identify reactivity patterns for several catalytic reactions. The primary bottleneck that prevented its widespread use was the significant amount of time, effort and expertise needed to accurately compute the transition state structures using DFT. Thus, to date experimentally testing potential catalysts has proved to be the more efficient strategy over large scale computational screening. Following Moore’s law, which defines the exponential increase in computational processor speed over time[16], *in silico* approaches have become increasingly popular and accessible and are now routinely employed in catalysis. The rational design of organo-catalysts has subsequently emerged as an important area of research with the aim of identifying novel catalysts with minimal computational resources. A number of strategies such as Linear Scaling Relationships (LSRs) can aid in reducing the computational burden and bring about a paradigm change. These models assume linear correlations between two reaction properties such as activation energies, relative free energies, reaction rate, equilibrium constant and bond distance, including the Sabatier's rule[17], **Bell-Evans-Polanyi (BEP)** principle[18a-d], and the Hammett equation[19a-c].

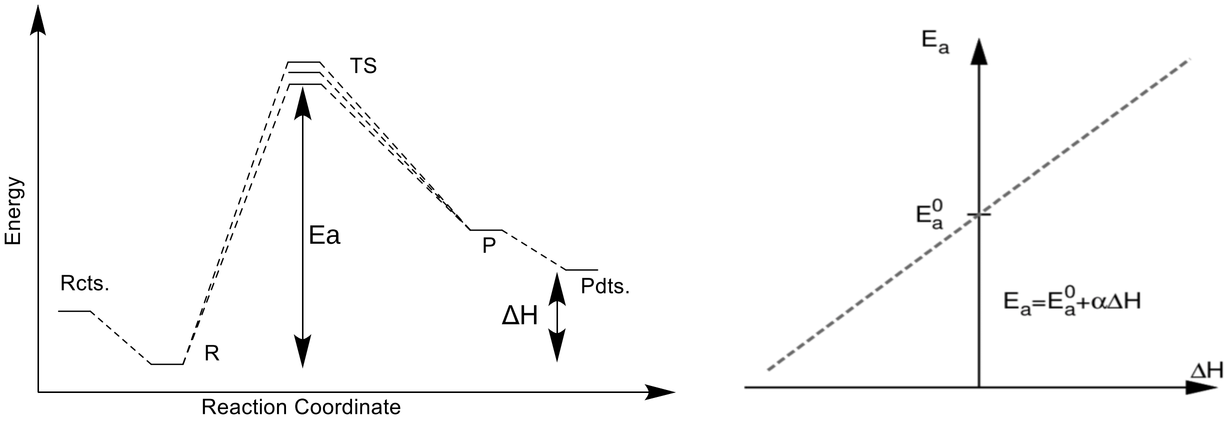


Figure 1. Schematic Representation highlighting the linear relationship for the **BEP** Principle (**TS**- Transition State, **R**- Reactant complex, **P**- Product complex)

The BEP correlations were initially formulated by Brønsted, Bell, Evans and Polanyi based on experimental observations and theoretical studies in the context of homogeneous systems[18a-d].

$$E_a = E_0 + \alpha\Delta H \tag{1}$$

The BEP equation (**Eqn.1**) states that based on the particular type of catalytic reaction the change in activation energy of the reaction, $\boldsymbol{E_a}$, can be expressed as a linear function of the corresponding change of reaction energy, $\boldsymbol{\Delta H}$, for different reaction intermediates (**Fig. 1**). The activation energy which is a kinetic parameter can be determined directly from the reaction energy, which is a thermodynamic parameter[20-21]. For several years, the primary use of these correlations was to compare the reactivities of molecules in a homologous series. It was not until much later with the work of Klein and co-workers[22], among others, that these correlations were successfully applied in the kinetic modelling of homogeneous chemistries. Following this ground-breaking work, BEP type correlations were also extensively applied to model heterogeneous catalytic reactions[20]. The simple yet elegant concept of the BEP principle also extends to the qualitative understanding of Linear Free Energy Relationships (LFERs), volcano plots[23], and Transition State Scaling (TSS)[11] relations among others. Such tools have been widely applied to investigate catalytic activities across homogeneous systems.

1.1 Machine Learning in Catalysis

In Chemistry, patterns are seen everywhere spanning from solid crystal structures to phospholipid chains or even complex combinations of functional groups[24]. These patterns largely govern the underlying properties of molecules and materials. Machine Learning (ML) has inevitably proven to be one of the most powerful strategies when it comes to big data analytics and data mining approaches across industry and academic research. Until a decade ago, hardly a few hundred studies on the applications of ML in Chemistry were reported. In the review by Cova and Pais et al.[25] It has been reported that in 2018, about 8000 articles in the Web of Science database comprised of ML keywords, which implies an exponential increase of 35% within a decade[26]. The quality and quantity of data generated from experiments and simulations encompass a lot of unstructured information yet to be explored (**Fig. 3**). This has been the primary backbone of the new data-driven paradigm, developing a bridge between theory, experiment, computation, and simulation.

Machine Learning broadly describes a set of algorithms that are committed to identify and learn patterns directly from data and are capable of making fast and accurate predictions without being given explicit instructions[31,34]. There are three different types of ML namely, supervised, unsupervised and reinforcement learning[25]. Supervised learning aims at identifying relations between the data and a target variable (e.g. chemical property) that we want to predict. In this algorithm, the model is constructed from 'training' molecules with known chemical properties that allow us to make predictions on an unseen dataset. The regression model is used in predicting continuous properties and establishes a relationship between a dependent variable (the target chemical property) and one or more independent variables (molecular descriptors)[33].

In the realm of homogeneous catalysis, the linear regression methods are widely used to establish a quantitative relation between the structural descriptors and catalytic activities and other properties[27-31]. Most often, a linear fitting model is characterized by a linear relationship between the descriptor and target properties (**Fig. 2**)[33]. For instance, multiple linear regression algorithms have been used in the prediction of catalytic activity of several analogues of pyridine metal complexes.

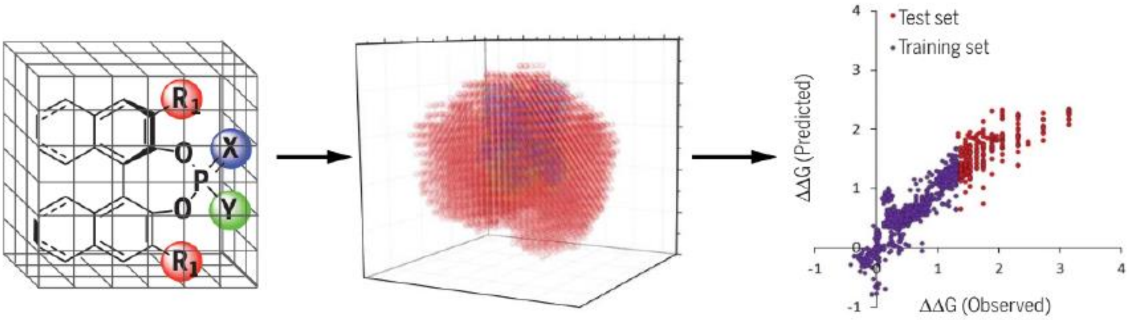


Figure 2. Denmark *et. al.* demonstrated the application of ML in predicting high-selectivity reactions from moderate- to low-selectivity reactions using an *in silico* library of catalysts[27]**.**

There also have been studies on cross-coupling reactions. Lilenfield and Corminboeuf have predicted the energy of the oxidative addition step for organometallic complexes using kernel ridge regression algorithms[32]. Recently, Sunoj et al. accurately predicted the products from regioselective difluorination of alkenes using neural networks[35]. Sigman and co-workers have established a data-driven linear regression protocol in a set of enantioselective catalytic reactions[36].

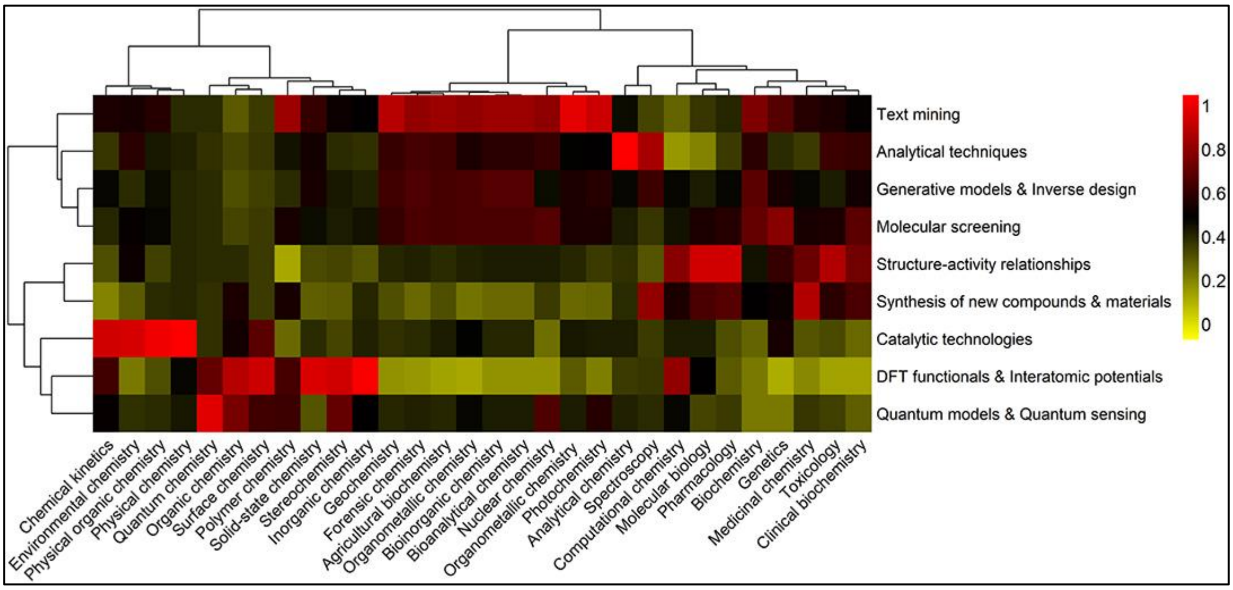


Figure 3. The clustering heatmap depicts the relative counts of ML outcomes in each area of Chemistry **(2008-2019).** The colour scheme represents co-occurrences with **1(red**) being the highest and **0(yellow)** being the lowest relative contribution[24]**.**

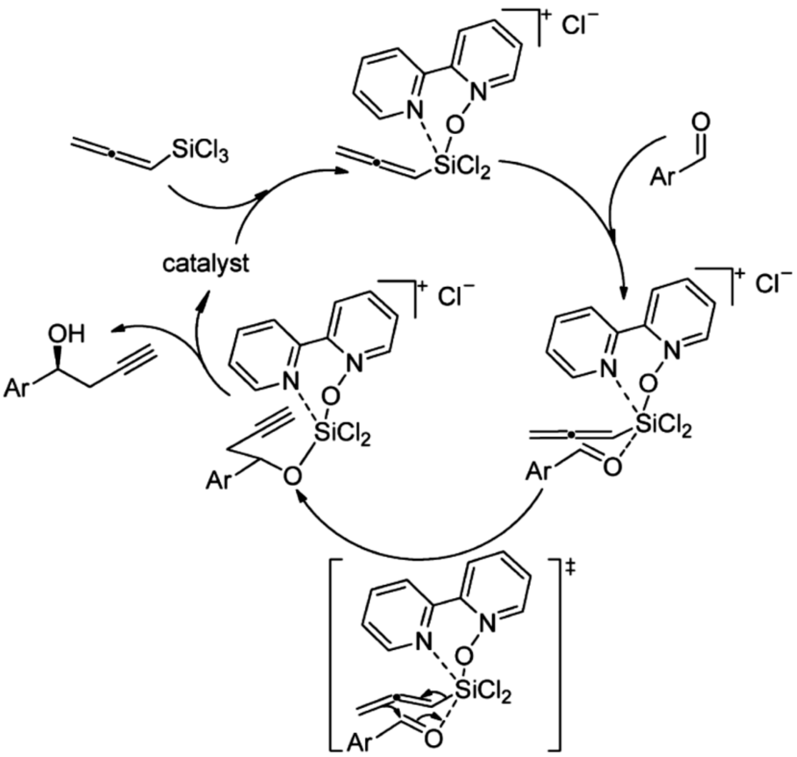
Despite the considerable amount of progress in applying ML models to chemical problems, the majority of the aforementioned contributions tackled issues surrounding homogeneous and heterogeneous catalysis, while ML applications to organocatalytic systems remain quite unexplored.

1.2 The Chemical System

The immense practical applicability of synthetic chiral molecules in single-enantiomer pharmaceutical compounds, optoelectronic devices, as polymeric components with novel properties and as probes to study biological systems, has made asymmetric catalysis a prominent area of investigation. It was generally accepted that transition metal complexes and enzymes were the two main classes of very efficient asymmetric catalysts. Synthetic chemists have rarely used small organic molecules as catalysts throughout the last century, even though some of the very first asymmetric catalysts synthesized were purely organic molecules[14].

A transition occurred during the last decade when several studies confirmed that relatively simple organic molecules can be highly efficient and remarkable enantioselective catalysts for diverse fundamentally important chemical transformations[37]. This rediscovery had subsequently led to an explosive scientific advancement in organo-catalysis. As the realization dawned that organic molecules not only have the flexibility of manipulation and a “green” advantage but also could be very efficient catalysts, asymmetric organo-catalysis began to parallel the enormous advancements of enantioselective transition metal catalysis. Additionally, catalytic asymmetric reactions play an integral role in modern organic synthesis. They allow efficient access to a variety of important enantiomerically rich molecules relevant to both industry and academia. This class of reactions can potentially yield large quantities of optically active products with a very high efficiency using meagre amounts of chiral catalysts. Consequently, this area of research has great economic potential and is becoming increasingly attractive.

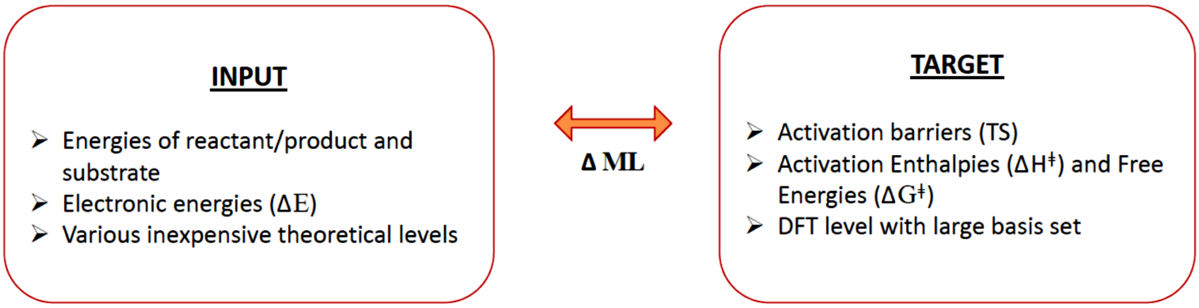
In the realm of organo-catalysis asymmetric allylations has received sufficient attention from the chemical community in recent decades, only limited attention had been paid to catalytic asymmetric propargylation compared to the tremendous advances in asymmetric catalysis. Optically active homopropargylic alcohols are crucial chiral building blocks in organic synthesis due to the versatility of the acetylene unit. The asymmetric propargylation of aldehydes provides direct access to this class of compounds. However, these reactions often encounter difficulties associated with low regioselectivity and/or reactivity[40-41]. Allenyltrichlorosilane is a potential candidate as a nucleophile partner in such reactions because of its mildness, regiospecificity and low toxicity.



Scheme 1. Catalytic Cycle for the Bipyridine N-Oxide Catalysed Propargylation of Aromatic Aldehydes[39]

The asymmetric propargylation reaction here typically constitutes the conversion of an aromatic aldehyde (e.g. substituted benzaldehydes) to a chiral homopropargylic alcohol[13,40]. Several reports exist with experimental studies based on the use of axially chiral N, N'-dioxides in Lewis base promoted allylations, however, for propargylations it proves to be much more challenging. It was only in 2013, that Takaneka and co-workers developed a helical bipyridine N-oxide catalyst that yields the alcohol with sufficient enantioselectivity (**Scheme 1**)[38]. There has been significant work illustrating the reaction mechanism and origin of stereoselectivity of these reactions[39-41]. When the reaction is carried out in solvents like dichloromethane (DCM), the stereo controlling step proceeds with a closed, chair-like transition state hexacoordinating a silicon (Si) centered intermediate[13].

2. Objectives



INPUT

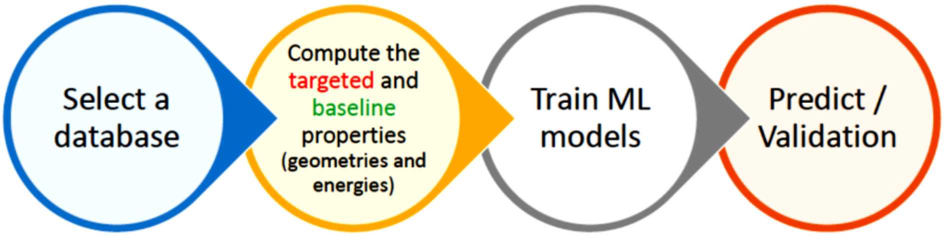
- Energies of reactant/product and substrate
- Electronic energies (ΔE)
- Various inexpensive theoretical levels

Δ ML

TARGET

- Activation barriers (TS)
- Activation Enthalpies ($\Delta H^{‡}$) and Free Energies ($\Delta G^{‡}$)
- DFT level with large basis set

The purpose of this work is to demonstrate how ML models can be used to estimate the activation barriers of an organocatalytic reaction. The barrier heights are key to determining the product selectivity in a catalytic reaction. The database is usually inspired by experimental investigations and to this end, we selected the catalysts for the asymmetric propargylation reaction. Specifically, we trained and applied the Δ-ML approach[59] using the relative electronic energy and molecular geometries associated with reactant and product side intermediates corresponding to each catalyst in the database. The structures and relative electronic energies (***ΔE***) of all the reactant and product intermediates were computed using various inexpensive theoretical methods. The activation barriers in terms of relative Enthalpies ($\boldsymbol{\Delta H^{‡}}$) and Free energies ($\boldsymbol{\Delta G^{‡}}$) were to be obtained from a pre-compiled database computed at the DFT level with a large basis set. The differences between the Input and Target energetic values were then fed into the Δ-ML algorithm[59]. The overall machine learning (ML) workflow is summarized in **Scheme 2**. Even though kinetic profiles are crucial to obtaining a complete understanding of catalytic performance, here we rely on a simplified thermodynamic picture coupled with the concept of Bell-Evans-Polanyi Principle. Precisely, the ultimate goal is to move to very inexpensive methods without a total loss of accuracy and identify the best-suited framework to accurately predict the targeted chemical properties.

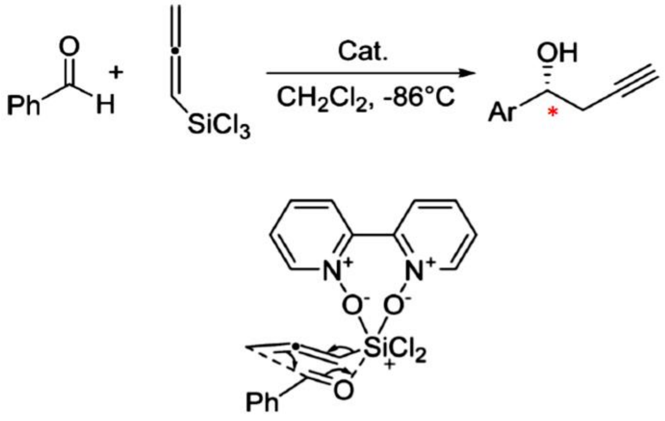

Select a database
Compute the targeted and baseline properties (geometries and energies)
Train ML models
Predict / Validation

Scheme 2. The Machine Learning Workflow

3. Methods

3.1 Database Construction

Inspired from experimental results on asymmetric allylations, Wheeler and co-workers[13] established a computational screening method to the design asymmetric propargylation catalysts[40-41].



Scheme 3. Stereocontrolling step in the Asymmetric Propargylation of Benzaldehyde using Allenyltrichlorosilane [13]

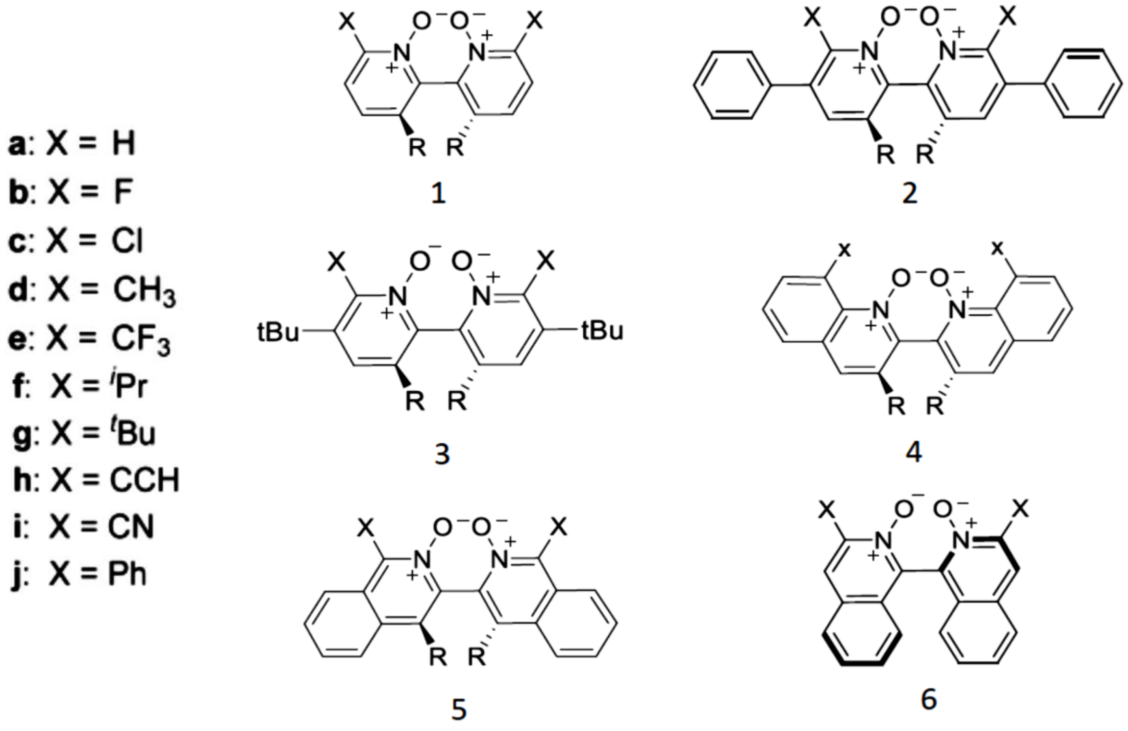


Figure 4. Set of Catalysts used for Constructing the Database[13]

An asymmetric propargylation reaction scheme was studied (**Scheme 3**), using a virtual library of 59 catalysts based on the bipyridine N, N'- dioxide scaffold (**Fig.4**) using allenyltrichlorosilane[41]. In this paper, using DFT based methods, enantiomeric excess (ee) values were predicted for all the 59 potential catalysts[13].

The six backbones represent catalysts based on different classes of bipyridine-N, N-dioxide derivatives. The parent scaffold (**1**), is the (S)-2,2'-bipyridine N, N'-dioxide with substituents (X) at 6,6'-positions. Catalysts **2** and **3** consist of Ph and ^{t}Bu substituents at the 5,5-positions respectively. Scaffold **4** is an (S)-8,8'-disubstituted 2,2'-biquinoline N, N' -dioxide, **5** is an (S)-1,1' -disubstituted 3,3' -biisoquinoline N, N-dioxide and **6** is an (S)-3,3' -disubstituted 1,1' -biisoquinoline N,N' -dioxide. Most of these catalysts were predicted to be synthetically viable, however only certain catalysts based on backbones 1,4 and 6 had been previously used for asymmetric allylations and **4a** for propargylation reactions in practice. Thus, it was an interesting case to look into the catalytic activity of these catalyst derivatives.

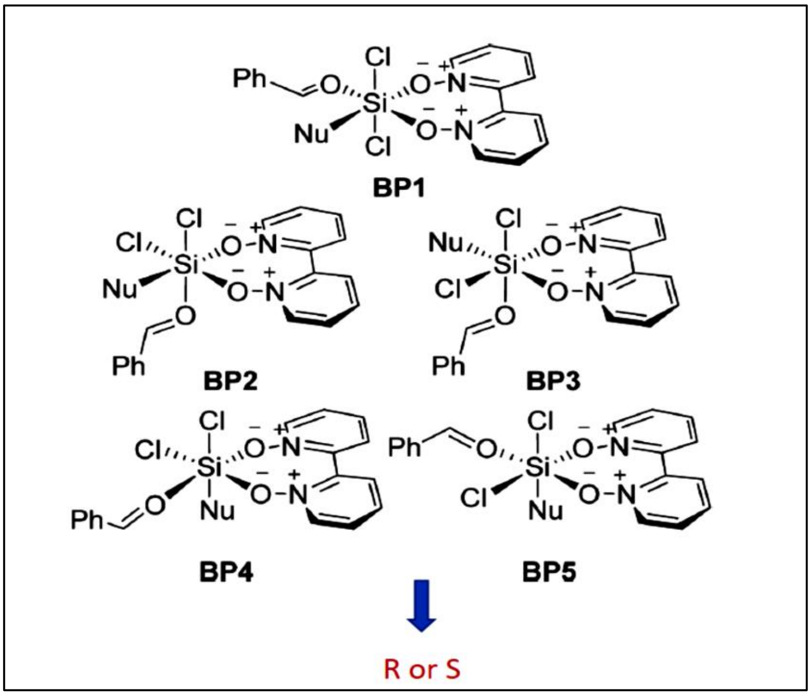


Figure 5. The five ligand arrangements for C2-symmetric bidentate Lewis base catalyzed alkylation reactions (**Nu** is the alkyl nucleophile)[13]

Subsequently, we shifted to computing intermediates directly as it is relatively much simpler and computationally less expensive. We constructed a final database spanning across **62** catalysts based on the bipyridine N, N'-dioxide scaffold. The combination of **10** substituents

(**Fig.4**), **5** ligand configurations and enantiomers (**Fig.5**) gave a total of **576** structures to construct the entire database. The delta learning approach simplifies the training process and involves lower computational cost. So instead of just the absolute values, all the energies of TS and intermediates respectively (**ΔE/ ΔH‡/ ΔG‡**) were computed concerning those of the starting reactants (**Scheme 3**). The relative energies at both the baseline (DFT, semi-empirical) and target level (DFT) were fed as inputs to the ML algorithm.

3.2 Computational Details

3.2.1 Computation of Target Properties

The stereoselectivity of asymmetric reactions arises from the difference in relative rates of product formation, and the number of accessible TS structures is often huge. So, to aid in the construction and optimization of all TS geometries, we used **AARON**, which is an automated TS search procedure, developed by the Wheeler group[42].

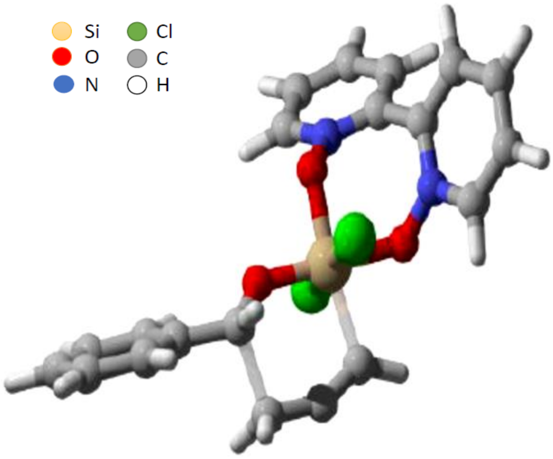
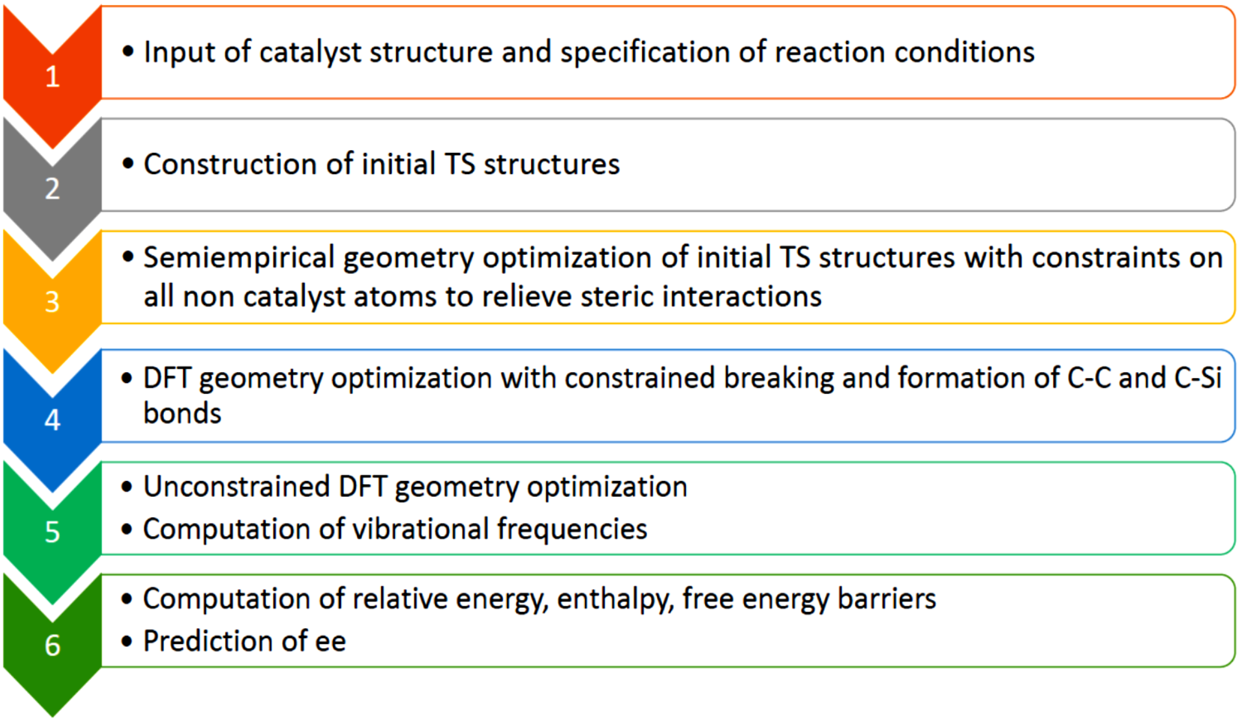


Figure 6. Schematic representation of a TS structure computed at **B97D/3-2**

AARON (**A**n **A**utomated **R**eaction **O**ptimizer for **N**ew catalysts), is a computational toolkit, that can locate multiple conformations and configurations of TS structures, and simultaneously screen potential catalysts and substrates for organocatalytic as well as organometallic systems[42].

AARON works on a text-based interface with **Gaussian 09**[43], and performs a tiered series of constrained and unconstrained TS optimizations, based on a user-defined template (**Scheme 4**). A representative TS structure computed at B97D/3-21G using AARON, is shown in **Fig. 6**.



1

- Input of catalyst structure and specification of reaction conditions

2

- Construction of initial TS structures

3

- Semiempirical geometry optimization of initial TS structures with constraints on all non catalyst atoms to relieve steric interactions

4

- DFT geometry optimization with constrained breaking and formation of C-C and C-Si bonds

5

- Unconstrained DFT geometry optimization
- Computation of vibrational frequencies

6

- Computation of relative energy, enthalpy, free energy barriers
- Prediction of ee

Scheme 4. The overall six-step **AARON** workflow

The AARON toolkit extracts information from a text-based input file, which contains information about the location of the template library and keywords specifying the reaction conditions (temperature, solvent, etc.) as well as the level of theory. In the input file, specific ligands/ catalysts/ substrates may also be specified[42].

AARON constructs the initial structures corresponding to each catalyst/substrate combination and locates all possible TS structures. The overall protocol is summarized in **Scheme 4**.

All TS computations were carried out at the **B97D/ TZV (2p,2d)** level of theory[44-47] and density fitting techniques. The solvent included was dichloromethane (DCM) using the polarizable continuum model[48] (**PCM**) and harmonic vibrational frequency analysis was performed to confirm the transition states. The enantiomeric excess (**ee**) values for each catalyst were computed using the Boltzmann weighted average of relative enthalpy barriers (**$\Delta H^{\ddagger}$**) and free energies ($\Delta G^{\ddagger}$) of thermodynamically accessible TS at **195 K** temperature (***Ei*** represents the relative energies of the R and S conformers respectively; ***R*** is the universal gas constant; ***T*** is the absolute temperature in Kelvin).

$$E_{eff} = -RTln\left(\sum_{i}^{conformers} e^{-\frac{E_i}{RT}}\right) \tag{2}$$

$$ee(\%) = \frac{\sum_i^{TS} e^{-\Delta E_{eff}(R_i)/RT} - \sum_i^{TS} e^{-\Delta E_{eff}(S_i)/RT}}{\sum_i^{TS} e^{-\Delta E_{eff}(R_i)/RT} + \sum_i^{TS} e^{-\Delta E_{eff}(S_i)/RT}} \tag{3}$$

3.2.2 Computation of Baseline Properties

Corresponding to each catalyst structure both the reactant and product side intermediate complexes were computed from the respective transition states by considering relative displacements along the **C$_3$-C$_{15/16}$** bond between the attacking nucleophile and carbonyl centre of benzaldehyde.

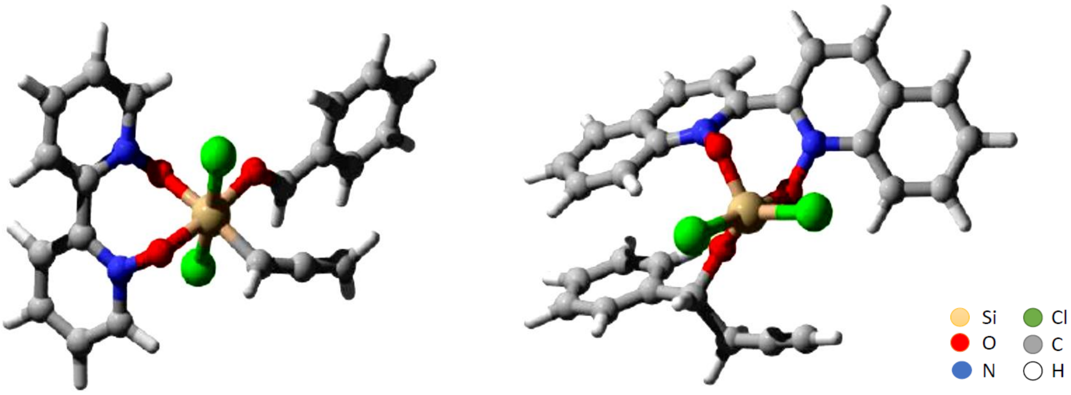


Figure 7. Schematic representation of reactant and product side intermediates computed at **B97D/3-21G,** characterized by a hexacoordinate and pentacoordinate **Si** centre respectively**.**

All geometry optimizations and energy computations for the intermediates were done using the baseline theoretical methods at **B97D/ 3-21G (**DFT**)**[44-47]**, HF-3c**[49] and **PM6-D3 (**semi empirical**)**[50-51]. The electronic energies of all 576 reactant/product intermediates were computed in the gas phase with respect to that of the separate reactants (ΔE). The DFT computations were performed in **Gaussian**[43] program package, HF-3c and PM6-D3 were done in **Orca**[52] and **Mopac**[53] respectively.

3.3 Training Set Selection

In order to ensure efficient training of the machine learning model, a representative subset of the database needs to be chosen to compute the target property. A widely used approach for selecting the training set is to perform farthest point sampling which ensures the selected data points are as diverse as possible. However, this method requires collective variables that are not straightforward to obtain especially when the dataset constitutes many different types of molecules. Also, farthest point sampling is a computationally demanding step. Screening 505 optimal molecules within **576** potential candidates typically would require $^{576}C_{505} = 1.27 \times 10^{92}$ operations.

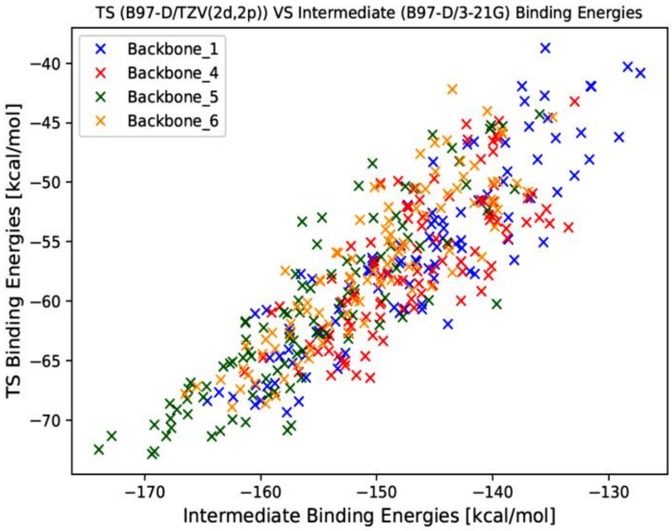
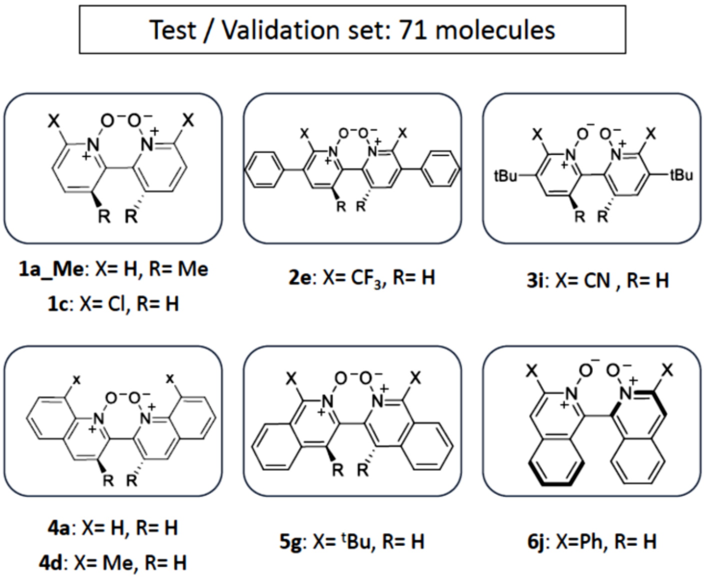


Figure 8. Precomputed distribution of Relative Energies for selected Backbones among the Training Set

Another small subset of the database has to be categorized as the test set, which was used to validate the trained learning models. So, we employed a more intuitive approach to select the training and test sets based on the specific target property. A total of 576 molecules in the database was divided into the training and test set.

Based on earlier results involving transition states, it was seen that including at least one conformer/stereoisomer corresponding to the 62 different catalyst/backbone combinations, for the training, the mean absolute error (**MAE**) on the test set was significantly reduced compared to those when the ML model is made to predict on a completely new set of species. The latter would be more like extrapolation for which regression models do not perform satisfactorily. Preferably, all substituents and possible conformers for each type of the 6 catalyst backbones, should appear uniformly in the training set (**Fig. 8**). Also, stereoisomers corresponding to each TS conformer were accounted for while selecting the test set. These principles were applied to selectively choose the training and test sets and subsequently **505** molecules were used to train the machine learning models.



1a_Me: X= H, R= Me
1c: X= Cl, R= H

2e: X= CF_3, R= H

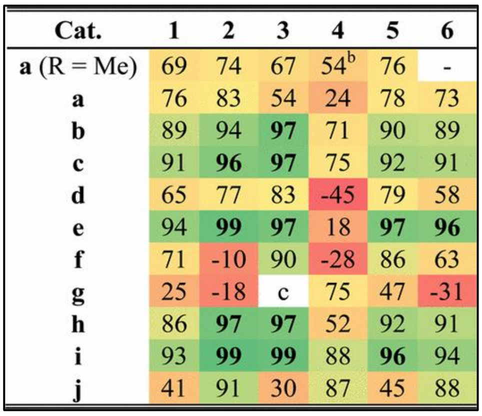
3i: X= CN , R= H

4a: X= H, R= H
4d: X= Me, R= H

5g: X= ^{t}Bu, R= H

6j: X=Ph, R= H

Figure 9. Test set selected to validate the ML models towards the prediction of **stereoselectivity** (ee values). Eight possible substituent/backbone combinations were included.



Cat.	1	2	3	4	5	6
a (R = Me)	69	74	67	54[b]	76	-
a	76	83	54	24	78	73
b	89	94	**97**	71	90	89
c	91	**96**	**97**	75	92	91
d	65	77	83	-45	79	58
e	94	**99**	**97**	18	**97**	**96**
f	71	-10	90	-28	86	63
g	25	-18	c	75	47	-31
h	86	**97**	**97**	52	92	91
i	93	**99**	**99**	88	**96**	94
j	41	91	30	87	45	88

Figure 10. Previously reported **ee values** based on relative electronic energies[13]

In order to validate the ML model towards the estimation of stereoselectivity, separate training, and test sets were chosen. A total of **71** structures spanning across 8 different catalyst/backbone/substituent conformations and well-distributed values of reported enantiomeric excess[13] values (**Fig. 10**) were included in the Test set.

Here, the complete set of conformers were included for each catalyst (in contrast to the previous case) such that the ML model does not see a similarly structured catalyst in the training dataset. The Δ-learning was performed on the differences between computed reactant intermediates (Baseline: B97D/3-21G) and the transition states (Target: B97D/TZV). The enantiomeric excess (ee) values were calculated using **Eqn.3**, via the Boltzmann weighted average at T=195 K, based on the relative enthalpy and free energy barriers (**$\Delta H^{‡}$/ $\Delta G^{‡}$**) of thermodynamically accessible TS structures.

3.4 Theoretical Details

3.4.1 Kernel Ridge Regression (KRR)

In many real situations, the correlation among constituting data points cannot be described by a linear function in the input space. Learning non-linear relationships between data points is a fundamental problem in machine learning.

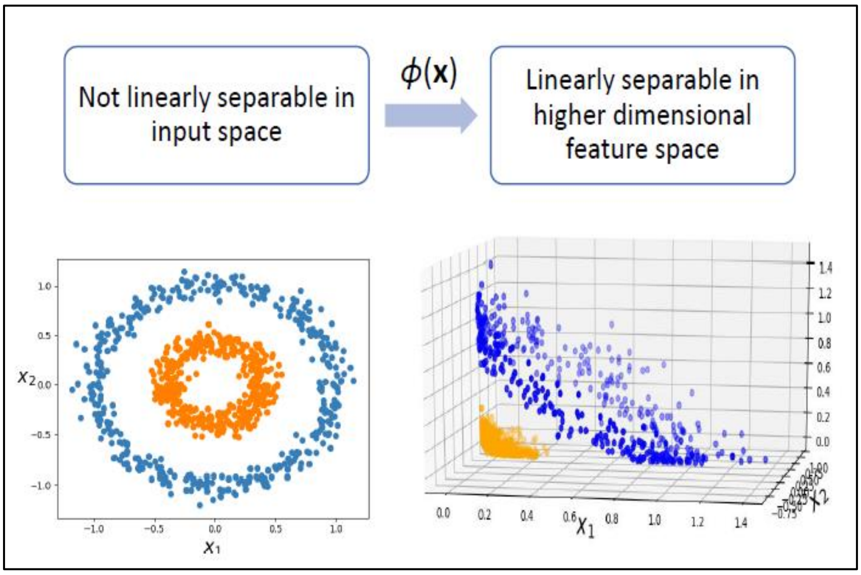


Figure 11. A non-linear polynomial transformation yields the optimal separating hyperplane

For such cases, a linear ridge regression model may lead to a poor prediction and a common approach is to map samples from this space to a higher dimensional space using a nonlinear transformation (**Fig.11**), and then learn the model in the higher dimensional space where the problem becomes linearly separable[54]. However, explicitly calculating each of the polynomial combinations in each space coordinate may incur a very high and impractical computation cost. The Kernel trick is a widely used state-of-the-art approach to conduct this learning procedure implicitly by defining a kernel function which represents the similarity of samples in the high dimensional space, through a scalar dot product[54] (where, ***k(x, x')*** is the kernel function, ***I*** is the identity matrix, ***λ*** is the regression term, ***α*** is the co-efficient matrix and ***ΔE*** is the relative energy difference, respectively).

$$k(x, x') = \phi^T(x)\phi(x') =< \phi(x)\phi(x') > \tag{4}$$

$$\alpha = (K + \lambda I)^{-1}\Delta E_{ref} \tag{5}$$

$$\Delta E(x) = \sum_i^N a_i k(x, x_i) \tag{6}$$

This method combined with normal ridge regression yields a simplified approach of finding an optimal separating hyperplane in the higher dimensional space, without any explicit calculation or even knowing anything about the actual transformation.

3.4.2 Molecular Representations

To establish an effective machine learning framework to learn the energetic data from the structure of the molecular species, we numerically represent the relative energies (ΔE/ ΔH‡ / ΔG‡) by vectors of constant size. These vector representations should in principle, encode the atomic composition and structural information of a given molecule. Herein, the machine learning models were trained using three different representations: Coulomb Matrix (CM)[60], Bag of Bonds (BoB)[57] and Bags of London and Axillrod-Teller-Muto potentials (SLATM)[55].

The **Coulomb Matrix (CM)** is one of the simplest molecular representations. First proposed in the seminal work[60] by Von Lilienfeld et al., this representation includes information about the constituent atoms as well as their connectivity. This representation has been widely used in several QM/ML models[56-58] for gas-phase molecules. The elements of this square atom-by-atom matrix are computed using the following expression:

$$M_{IJ} = \begin{cases} 0.5 Z_I^{2.4}, & I = J \\ \frac{Z_I Z_J}{|R_I - R_J|}, & I \neq J \end{cases} \tag{7}$$

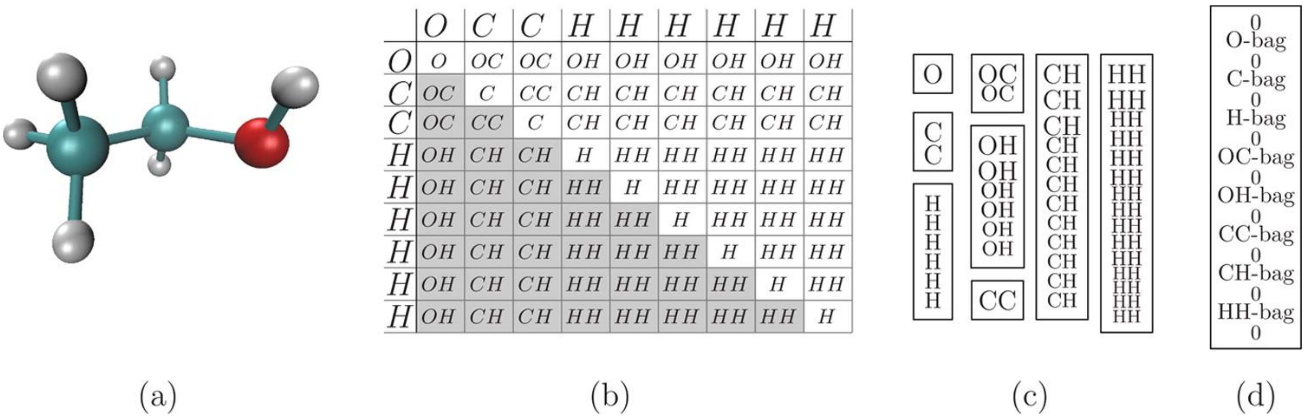
The CM is inspired by the fact that in principle, molecular properties can be estimated from the Schrödinger equation, taking the Hamiltonian operator as its input. The off-diagonal elements correspond to the Coulomb repulsion between each pair of atoms in a given molecule while the diagonal elements approximate the electronic potential energy of free atoms through a polynomial fit (***Z_I*, *Z_J*** are the nuclear charges; ***R_I*, *R_J*** are the distance vectors). The molecular representation should be unique and the CM is invariant to rotation and translation but not atomic permutations. There are several different ways to sort the order of atoms in a coulomb matrix. One common way is by using the eigenvalue spectrum of the CM and permuting the matrix to compute the norm of each row and column and further reordering the matrix of eigenvalues in descending order.

The **Bag of Bonds (BoB)** representation was formulated by Hansen *et al*. in 2015[57]. It originates from the “bag of words” featurization commonly used in natural language processing. Bag of bonds follows an approach by having “bags” that are grouped based on different types of bonds (e.g. Si-C, N-O, C-O, C-H, etc).

The chemical bonds are uniquely represented by the atoms involved and the order of the bond (single, double, triple). Moreover, each “bag” is essentially a vector where each element is computed as

$$\frac{Z_I Z_J}{|R_I - R_J|} \tag{8}$$

These bag vectors between molecules are constrained to have a fixed length by padding them with zeros. The entries in each bag vector are sorted in a descending order based on magnitudes to ensure a unique representation (**Fig.12**). Even though BoB accounts for the collective effects beyond pairwise potentials, important higher-order information (e.g. angular terms) is missing.



	O	C	C	H	H	H	H	H	H
O	O	OC	OC	OH	OH	OH	OH	OH	OH
C	OC	C	CC	CH	CH	CH	CH	CH	CH
C	OC	CC	C	CH	CH	CH	CH	CH	CH
H	OH	CH	CH	H	HH	HH	HH	HH	HH
H	OH	CH	CH	HH	H	HH	HH	HH	HH
H	OH	CH	CH	HH	HH	H	HH	HH	HH
H	OH	CH	CH	HH	HH	HH	H	HH	HH
H	OH	CH	CH	HH	HH	HH	HH	H	HH
H	OH	CH	CH	HH	HH	HH	HH	HH	H

Figure 12. BoB representation scheme (a) 3D structure of ethanol (CH_3CH_2OH) (b) nuclear charges for each CM entry. (c) Different CM elements grouped into bags and (d) BoB vector obtained by concatenating these bags [57]

For both the CM and BoB representations, the matrices were set to a fixed size of 89 x 89, corresponding to the largest number of atoms in the database. Zero-padding is generally employed to fill the matrices for molecules containing fewer atoms. The matrices were then linearized by joining their rows into a one-dimensional vector that was then used as input for the machine learning models.

A more sophisticated molecular representation includes all possible interactions between atoms through many-body potential terms multiplied by a normalized Gaussian distribution. The **Spectrum of London and Axilrod-Teller-Muto potentials (SLATM)** representation[55,61] has been found to outperform the Coulomb matrix[60] and Bag of Bonds

model for computing quantum mechanical properties of small organic molecules and thermodynamic properties of organometallic compounds.

The one-body term simply consists of the nuclear charge (***Z_I***). The two-body part is expressed as

$$\frac{1}{2}\sum_{J\neq I} = Z_J\delta(r - R_{IJ})g(r) \tag{9}$$

where ***$\delta(.)$*** is set to normalized Gaussian function and ***$g(r)$*** is a distance-dependent

scaling function corresponding to the leading order term in the dissociative tail of London potential. The three-body part is represented by

$$\frac{1}{3}\sum_{J\neq K\neq I} = Z_J Z_K\delta(\theta - \theta_{IJK})h(\theta, R_{IJ}, R_{IK}) \tag{10}$$

where θ is the angle spanned by vector ***R_{IJ}*** and ***R_{IK}*** and ***$h(.)$*** is the three-body contribution chosen in the form to model the Axilrod-Teller-Muto vdW potential[61].

However, computation of these three-body interactions incurs a higher-order complexity as a function of the number of atoms in the molecules considered. Thus, the construction of this representation incurs a significantly higher computational cost compared to CM and BoB.

3.5 Machine Learning Models

Construction of the representations for the compiled database as well as machine learning model optimization, training and predictions were performed using **Python 3.7.1**, using the Quantum Machine Learning (QML) package[58,62] along with NumPy[63a] implementation of arrays. The SciPy implementation[63b] of Nelder-Mead optimization method was used for the hyperparameters and Matplotlib to visualize the results.

We used Kernel Ridge Regression (KRR) for the machine learning framework to map the molecular geometries to their corresponding energy descriptors. For this work, we considered the Gaussian (**Eqn.11**) and Laplacian (**Eqn.12**) kernels as they have been widely used for applications of KRR in Chemistry. The Machine Learning protocol using KRR primarily consists of three main steps namely, parameter optimization, learning, and validation.

The inputs for the models were chosen to be representations corresponding to the reactant intermediates at various inexpensive theoretical levels. The models were trained on the optimized molecular structures using the above baseline methods and corresponding predictions were performed using the optimized geometries from the same theoretical level.

$$k(x, x') = \exp\left(-\frac{||x - x'||^2}{2\sigma^2}\right) \tag{11}$$

$$k(x, x') = \exp\left(-\frac{|x - x'|}{\sigma}\right) \tag{12}$$

Based on the results obtained from similar studies in the past, we chose to use the Laplacian kernel function for the Coulomb matrix and BoB representations and the Gaussian kernel for SLATM representation[32,57,60]. Altogether 30 models have been trained and validated for the compiled database of 62 catalysts. For each resulting model, the two hyperparameters namely the width of the kernel function $\boldsymbol{\sigma}$ and the regression term $\boldsymbol{\lambda}$, must be optimized to minimize the prediction error for the target property of the unseen dataset. In our case, this target property is the energetic data ($\Delta H^{‡}$, $\Delta G^{‡}$) of the TS structures obtained at a higher level of theory (B97D/ TZV).

Subsequently, ten-fold cross-validation has been incorporated into the model to assess the prediction accuracy. The Nelder-Mead optimization scheme was applied to find the set of hyperparameters that minimizes the average of the mean absolute error (**Eqn.13**) on ten iterations of the ten-fold cross-validation, starting from initial values of

σ = 0.1, 0.2, 0.5, 1 ,5, 10, 50, 100, 200, 500, 1k, 2k, 5k, 10k, 20k, 30k, 50k,100k

$\lambda = 10^i$, i = -10, -8, -6, -4

Following the optimization of the hyperparameters, learning curves were obtained for each model by varying the size of the training set and computing the mean absolute errors (MAE) on the test set of **71** data points. The different sizes of the training data chosen were **10, 50, 100, 350** and **500**. Ten iterations were performed by a random selection of data points from the training set.

Validation of each ML model was performed on the pre-selected test set of **71** data points, and training the model on the remaining dataset of **505** points. The predictions of the descriptor on this validation set were correlated to the actual known reference values. The standard error of estimation (σ) for each model was calculated based on the mean squared deviations from the actual expected linear fit, where N is the total number of data points in the validation set. (**Eqn.14**)

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_{predicted} - y_{actual}\right| \tag{13}$$

$$RMSE(\sigma) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_{predicted} - y_{actual}\right)^2} \tag{14}$$

4. Results and Discussion

4.1 Estimation of Activation Barriers

4.1.1 Machine Learning: Training

Each of the machine learning models were tested with different molecular representations and kernel functions and varying the respective input parameters/ geometries. For each resulting model, the pair of hyperparameters (***σ, λ***) which yielded the lowest average mean absolute error (MAE), are tabulated (**Table 1**)

Input Structure (B97D/ 3-21G)	**Target Property (B97D/TZV)**	**CM (Laplacian kernel)**		**BoB (Laplacian kernel)**		**SLATM (Gaussian kernel)**	
		σ	***λ***	***σ***	***λ***	***σ***	***λ***
Reactant (R)	$\Delta H^{\ddagger}$	1.12E+05	9.01E-07	2.09E+04	1.02E-08	2.13E+04	1.02E-10
	$\Delta G^{\ddagger}$	1.10E+05	1.15E-10	5.09E+03	1.01E-08	1.99E+03	9.69E-09
Product (P)	$\Delta H^{\ddagger}$	1.01E+05	1.06E-08	2.19E+03	9.47E-11	4.82E+03	8.74E-11
	$\Delta G^{\ddagger}$	2.85E+04	1.10E-10	9.57E+02	1.06E-08	3.15E+04	1.05E-10

(a)

Input Structure (PM6-D3)	**Target Property (B97D/TZV)**	**CM (Laplacian kernel)**		**BoB (Laplacian kernel)**		**SLATM (Gaussian kernel)**	
		σ	***λ***	***σ***	***λ***	***σ***	***λ***
Reactant (R)	$\Delta H^{\ddagger}$	1.01E+05	1.03E-06	2.00E+04	1.05E-08	1.01E+03	1.01E-08
	$\Delta G^{\ddagger}$	5.37E+04	9.56E-11	1.95E+04	1.04E-08	5.03E+02	1.02E-10

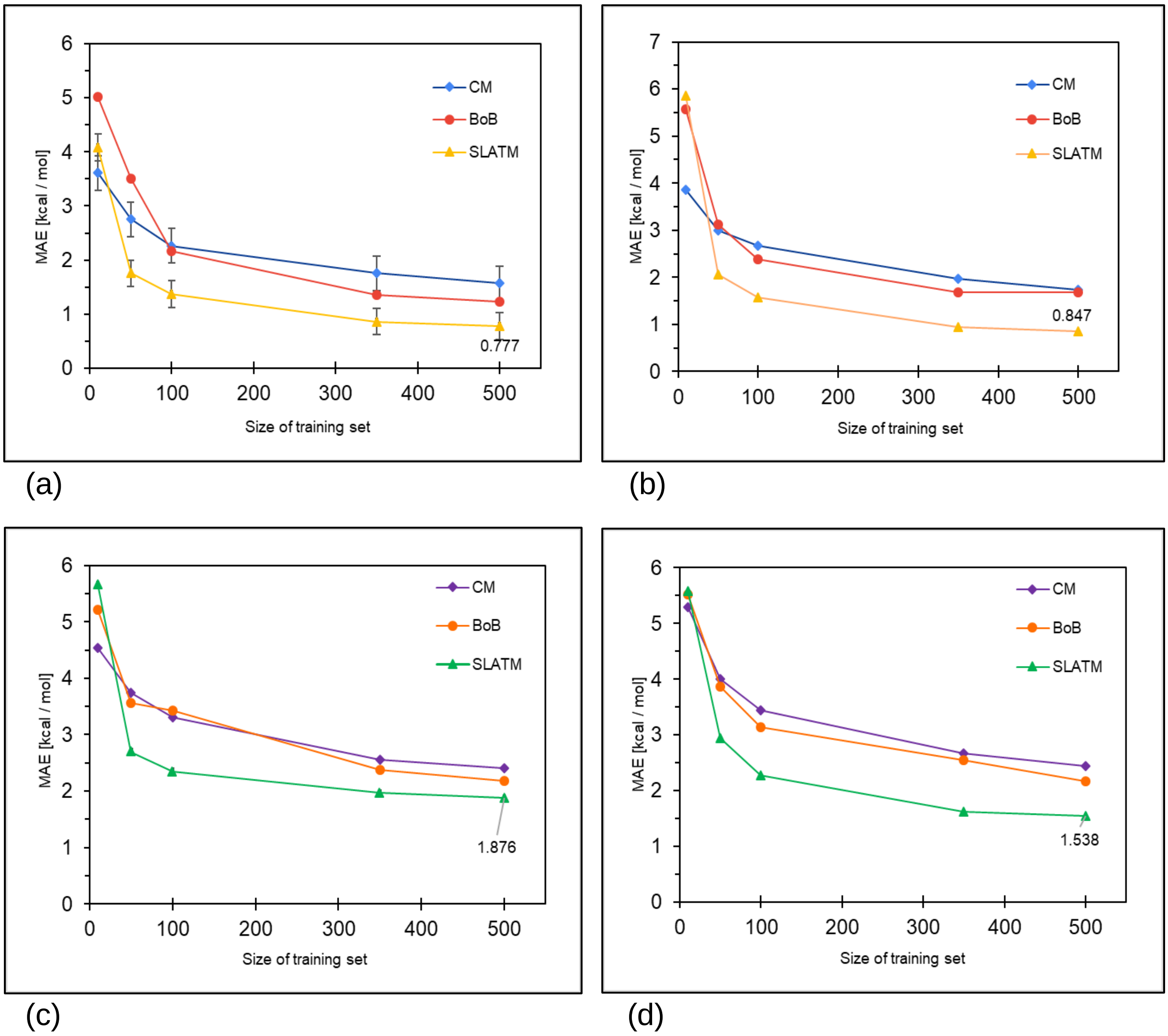
(b)

Input Structure (HF-3c)	**Target Property (B97D/TZV)**	**CM (Laplacian kernel)**		**BoB (Laplacian kernel)**		**SLATM (Gaussian kernel)**	
		σ	***λ***	***σ***	***λ***	***σ***	***λ***
Reactant (R)	$\Delta H^{\ddagger}$	5.38E+04	9.04E-11	9.90E+04	1.01E-08	3.11E+04	1.07E-10
	$\Delta G^{\ddagger}$	2.10E+04	9.48E-09	1.90E+04	1.05E-10	5.24E+04	1.00E-10

(c)

Table 1. The pair of Hyperparameters (***σ,λ***) that yielded the smallest error for each model; *σ* represents the width of the Kernel Function and λ is the Regression term. The values correspond to three different theoretical methods (a,b,c) used to compute the Baseline Properties

The learning step is generally represented through plots of MAE on a test set as a function of the training set size. The resulting mean absolute error (MAE) on the estimation of the activation energies on the Test set showed a reduction with an increase in the size of the Training set. The decrease of MAE values as a function of the training set size validated successful learning for all the ML models. This is evident from the obtained saturation curves corresponding to each of the models. The learning curves are primarily constructed to demonstrate the efficiency of the ML models. (**Fig.13**)



(a)

(b)

(c)

(d)

Figure 13. Saturation curves (**MAE** on a test set as a function of the number of training data) constructed for all representations considered, with Intermediates computed using **B97D/3-21G.**
(a) Reactants/**ΔH‡** (b) Reactants/**ΔG‡** (c) Products/**ΔH‡** and (d) Products/**ΔG‡**

In this work, we analyze two major aspects, one being the approximately accurate estimation of activation barriers ($\Delta H^{\ddagger}$, $\Delta G^{\ddagger}$) directly from the intermediates and the other to identify the least expensive method to achieve comparable level of accuracy. For the intermediates computed at the baseline DFT level (B97D / 3-21G), learning the activation enthalpies (**$\Delta H^{\ddagger}$**) with reactant side intermediates was found to be most efficient (**Fig.14**). Overall, the SLATM representation yielded the best final test error for the Gaussian based KRR (**Fig.14**). The learning was found to be inefficient when the model was trained using SLATM for the Laplacian kernel, and this observation is consistent with previous studies on similar chemical systems. However, when estimating the free energy barriers (**$\Delta G^{\ddagger}$**) for small training data SLATM does not perform very well relative to the Coulomb matrix or Bag of Bonds representations with Laplacian kernels. This behaviour can be ascribed to the higher complexity of SLATM compared to CM and BoB, making learning more difficult with sparse training data, but resulting in a more powerful estimation when trained on a sufficiently large number of data points. Also, even though BoB and SLATM representations begin with a relatively high MAE for smaller training data, they saturate much faster than CM. The rate of saturation with an increasing number of training points is highest for SLATM compared to the other two representations.

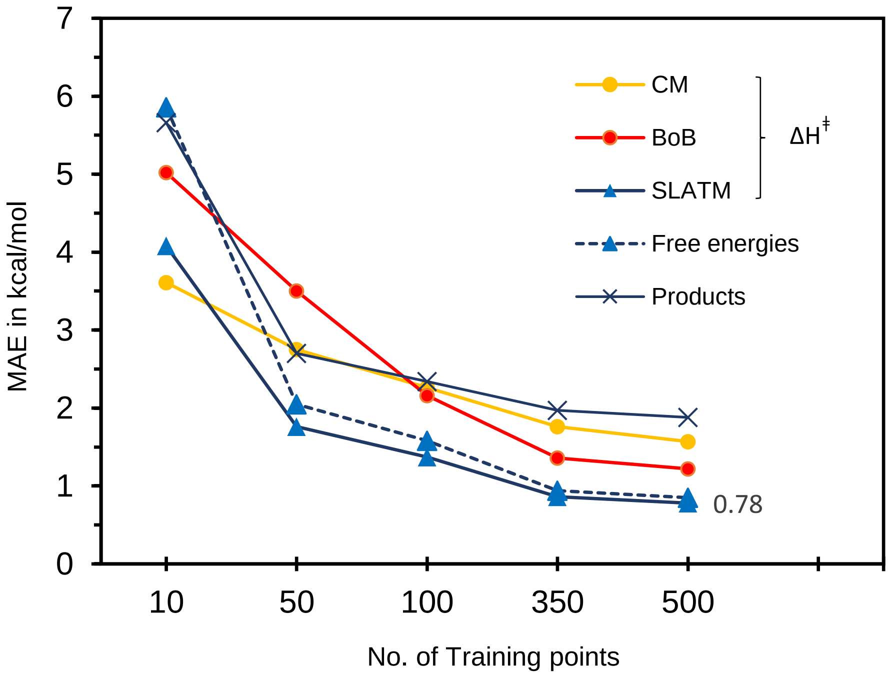


Figure 14. Comparison of Learning trends among all three ML representations, Input structure types and Target properties. (**CM, BoB**: Laplacian kernel, **SLATM**: Gaussian kernel)

Compared to CM, BoB was seen to perform better, thereby highlighting the importance of ‘bagging’ in this form of molecular representation.

As for the product side intermediates, the reactant side intermediates led to a more efficient learning, as evident from the MAE values corresponding to all three representations (**Fig.14**). There can be several factors contributing to this difference, one being that the reactant intermediates supposedly resemble the TS structures more than that of the products (**Fig.7**). Secondly, the hexacoordinate reactant complex has a much more rigid molecular structure than the product complex.

4.1.2 Machine Learning: Validation

The computing of the targeted energy values is crucial to validate the efficiency of the ML algorithm. In **Fig.15**, we depict the relationship between the relative electronic energies of intermediates computed at the baseline level (B97D / 3-21G) and the energetics of the transition states at the target DFT level (B97D / TZV) for the 71 data points in the validation set. All the relative energy values are normalized with respect to the mean value of each dataset in kcal/mol. The plot between the computed reactant intermediates and the reference TS barrier enthalpies (**Fig.15**) shows a correlation coefficient of R^2 = 0.79 compared to a value of 0.45 corresponding to the free energy barriers (**Fig.15**).

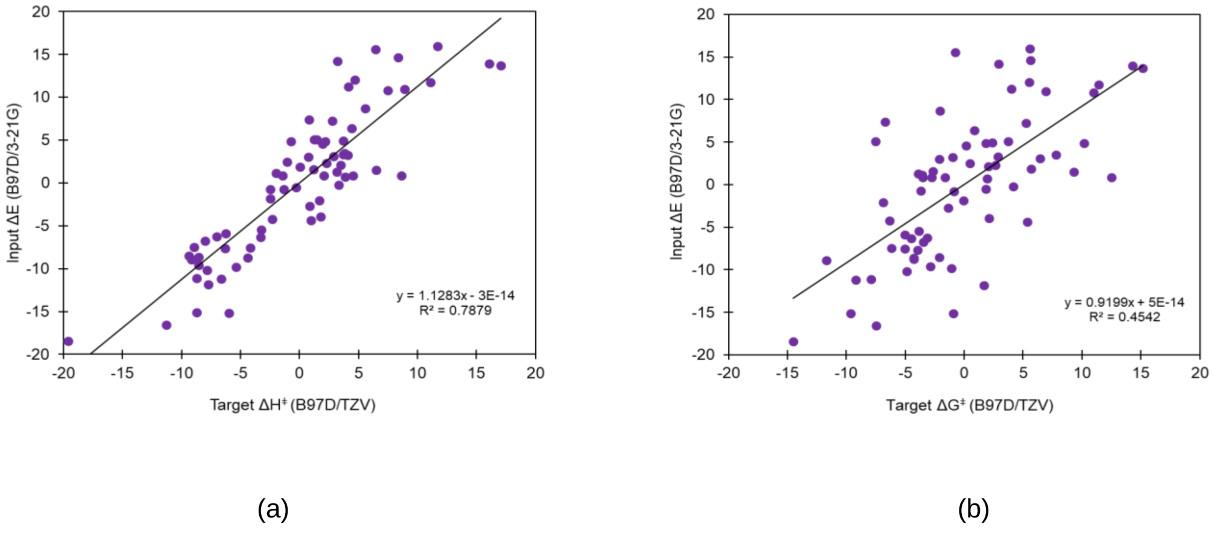


Figure 15. Relative energies (ΔE) of Reactant intermediates computed at **B97D/3-21G** (y-axis) versus Activation Barriers (x-axis), (a) ΔH‡ and (b) ΔG‡ respectively, computed at **B97D/TZV (2p,2d)** on a validation set of **71** datapoints.

There is not much-ordered correlation as expected, arising from the multiple parametric differences in the computation of the baseline and targeted properties (**see Methods**).

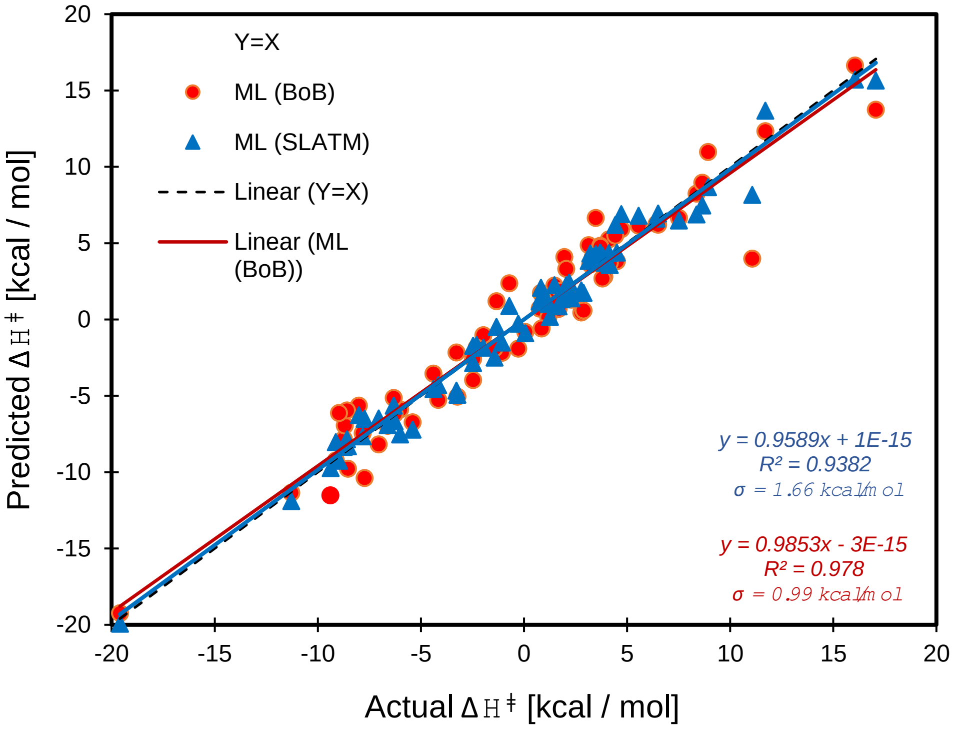


Figure 16. ML predicted (y-axis) and actual (x-axis) values of the descriptor($\Delta H^{\ddagger}$) for **B97D/3-21G Reactant intermediates** compared on a validation set of **71** points after training 505 data points. The identity line (y = x, in black), corresponds to perfect predictions of the descriptor value.

On training the Δ-learning model on the selected training set data using the kernel ridge regression (KRR) algorithm followed by computing the enthalpy and free energy barriers for the test set, the corresponding correlations obtained are illustrated. We compare three different parameters for the validation namely the type of input structures (Reactant versus Product side intermediates), the type of activation barrier ($\Delta H^{\ddagger}$ vs $\Delta G^{\ddagger}$) and the best molecular representation (CM, BoB or SLATM). From the saturation curves (**Fig.14**), we obtained an intuitive idea that the ML model proves to be more efficient in learning the reactant side intermediates compared to the products. We began our analysis by investigating the linear correlation plots between the estimated activation enthalpy barriers and the actual reference barriers. The kernel ridge regression model accounts for the non-linearity of data computed at the baseline level (B97D /3-21G) and using the Coulomb matrix (CM) representation corrects for the BEP correlation between energies of the intermediates and corresponding TS barriers. The correlation coefficient (**R^2**) between the actual and estimated enthalpy barriers for the 71 data points obtained by ML was 0.89 using the CM representation. As expected, the correlation improved significantly on subsequently training the KRR model using BoB ($R^2 = 0.94$) and SLATM ($R^2 = 0.98$) respectively (**Fig.16**). Also, the improvement over different representations have been quantified by the Standard Errors of Estimation (**σ**) from the actual ideal fit (y=x line). Considering the relative computational costs associated, BoB performs well in estimating the expensive DFT level energetic data, with a standard estimation error of 1.66 kcal/mol.

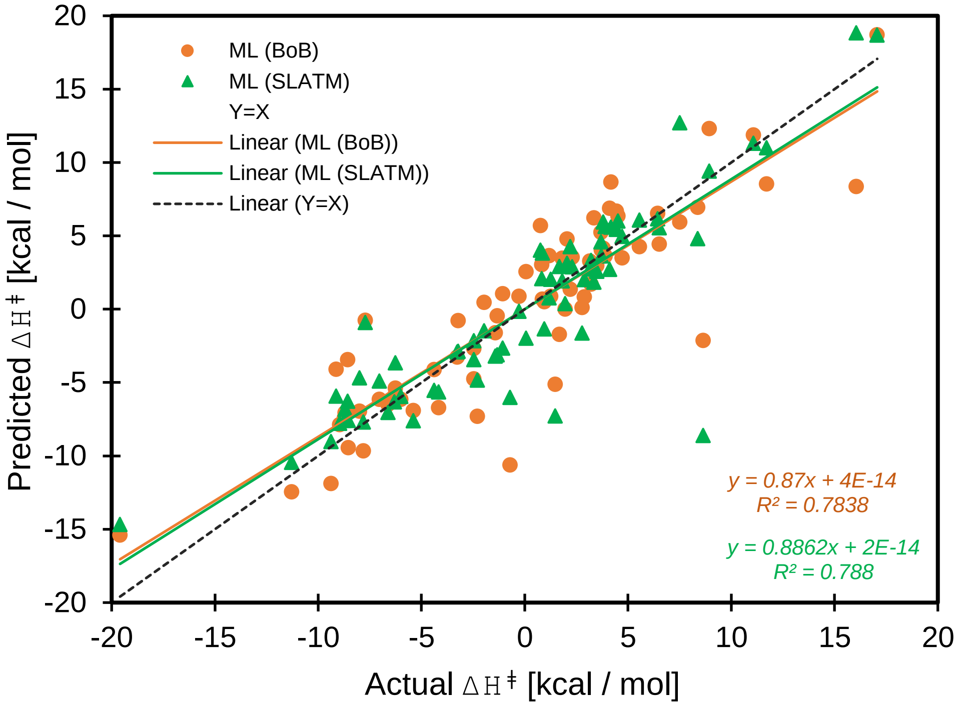


Figure 17. ML predicted (y-axis) and actual (x-axis) values of the descriptor($\Delta H^{\ddagger}$) for **B97D/3-21G** Product intermediates compared on a validation set of **71** points after training 505 data points. The identity line (y = x, in black), corresponds to perfect predictions of the descriptor value.

We shifted our analysis to the learning of the free energy barriers (***ΔG‡***) with reactant side intermediates as the input structures. The obtained correlation between the actual and estimated barriers is depicted in **Fig.18**, using the BoB and SLATM molecular representations respectively. Again, SLATM was observed to yield a better correlation with σ = 1.04 kcal/mol, $R^2 = 0.97$.

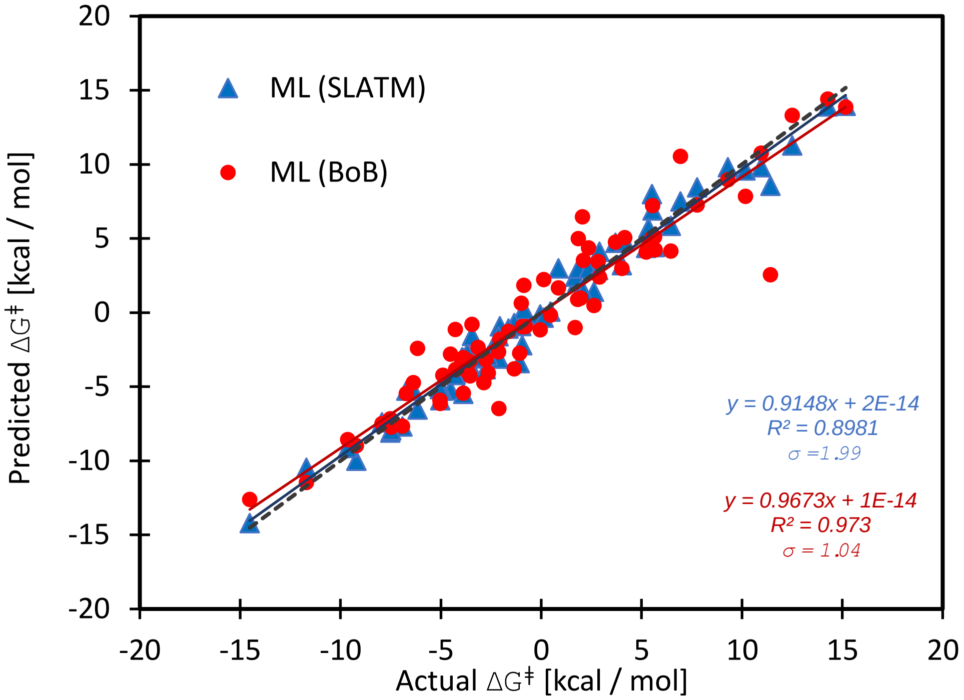


Figure 18. ML predicted (y-axis) and actual (x-axis) values of the descriptor(ΔG‡) for **B97D/3-21G** Reactant intermediates compared on a validation set of **71** points after training 505 data points. The identity line (y = x, in black), corresponds to perfect predictions of the descriptor value

Training the Δ-learning models subsequently with the product side intermediates and relative enthalpy barriers as the reference did not produce improved correlations as compared to that of the reactants. The correlation coefficient for SLATM was found to be 0.79 compared to an R^2 = 0.98 in the case of the reactants (**Fig.17**). The standard estimation errors also quantified the lack of accuracy of the ML model, which was previously evident from the respective learning trends. The reactant complex being structurally closer to the TS than the product complex, the ML model was able to estimate the energetic data much more accurately in the former case. Overall, for all the models trained with B97D/3-21G intermediates as a baseline and B97D/TZV (2p,2d) computed transition states as the target, SLATM depicted the best correlations among the three representations.

4.2 Machine Learning using other Baseline Methods:

Motivated by the results using DFT as the baseline method, we further extended the ML framework to learn the activation barriers at the target level (B97D/TZV) from the reactant intermediates using computationally inexpensive semi-empirical methods (PM6-D3 and HF-3c). These baseline methods have proven to work efficiently for several Δ-learning algorithms in the past.

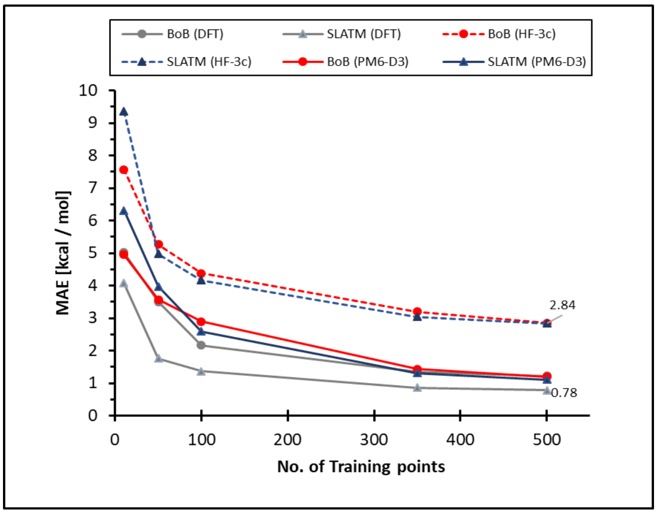


Figure 19. Learning trends from data computed using Semiempirical methods **(BoB: Laplacian kernel, SLATM: Gaussian kernel)**

We represented the mean absolute errors of prediction on the Test set (**Fig.19**) by varying the size of the training set. The saturation curves show a final MAE of 2.84 kcal/mol for HF-3c and 0.78 kcal/mol for PM6-D3. The SLATM representation worked best in both cases, compared to BoB and CM.

The learning trends of the different models were also evident from the validation step where the input energies and geometries computed at PM6-D3 demonstrated a much-improved correlation compared to those at HF-3c **(Fig.20)**. The correlation coefficient for HF-3c was found to be 0.81 **(Fig.21)**, compared to R^2 = 0.96 for PM6-D3 (as the baseline level). The definite reason behind this finding is still to be investigated and the outliers need to be analysed further. It could be hypothesized that the presence of 'bad' molecular geometries in the training set or the underlying potential energy surface might have contributed to such a trend in the learning approach. We computed the weighted root mean squared deviations (RMSD) of each of the geometries for the DFT computes structures. The molecules which depicted huge deviations were removed from the training set and the ML model was retrained. However, it yielded the same MAEs as before and no significant change was observed in the learning. Nevertheless, it would be interesting to look at how the ML framework can be improved to take this into account and estimate the activation barriers directly from intermediates computed using other semiempirical methods.

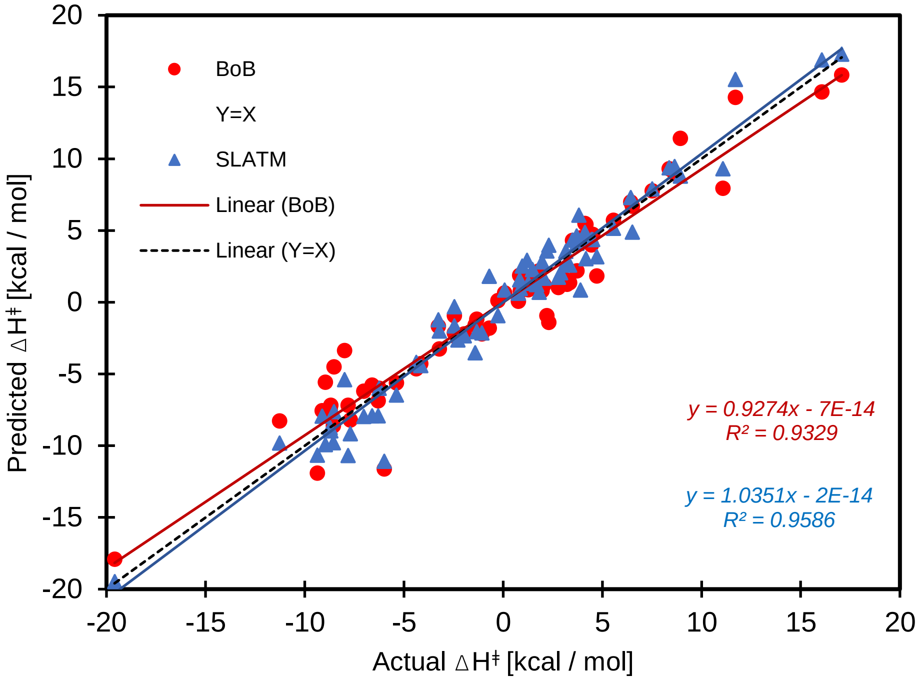


Figure 20. ML predicted (y-axis) and actual (x-axis) values of the descriptor(ΔH‡) for **PM6-D3** Reactant intermediates compared on a validation set of **71** points after training 505 data points. The identity line (y = x, in black), corresponds to perfect predictions of the descriptor value

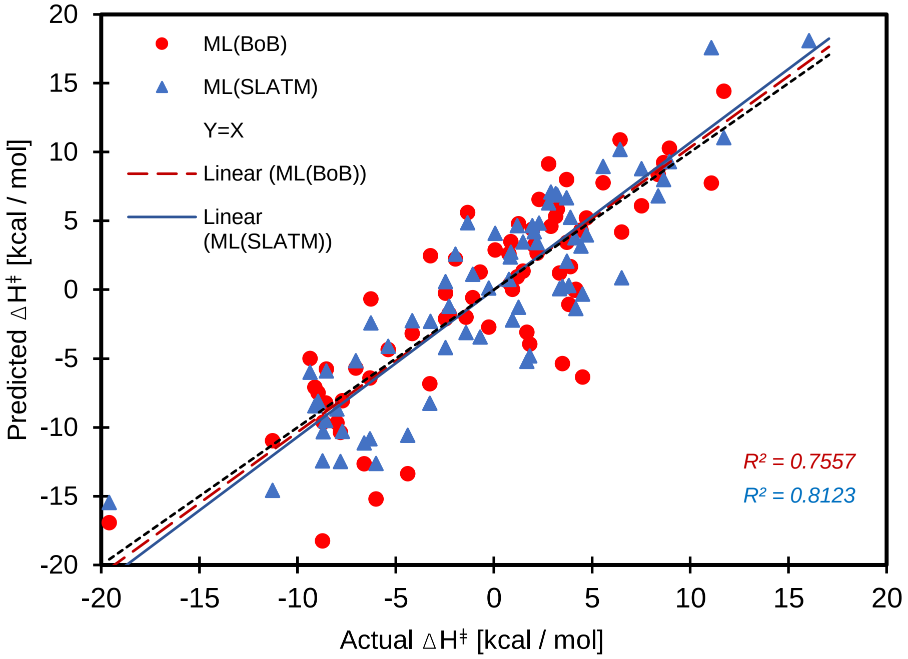


Figure 21. ML predicted (y-axis) and actual (x-axis) values of the descriptor(ΔH‡) for **HF-3c.** Reactant intermediates compared on a validation set of **71** points after training 505 data points. The identity line (y = x, in black), corresponds to perfect predictions of the descriptor value.

4.3 Prediction of Stereoselectivities:

The selectivity of products is a crucial issue, especially in pharmaceutical and drug discovery, that cannot be easily tuned. Here, we wanted to assess if learning BEP correlations could also be used to predict the enantio- or regioselectivity of catalytic processes from a quickly computable descriptor variable, namely energy of the intermediates. Naturally, the product selectivity depends upon the barrier height of the key step that determines the selectivity. Corresponding to the second set of 505 Training and 71 Test data points (**Fig.9**), the ML models were trained on all three molecular representations (CM, BoB and SLATM).

The learning was more efficient for the activation enthalpies (**ΔH‡**) as compared to the free energies (**ΔG‡**) as can be seen from the reported MAE values in kcal/mol. Also, the absolute errors are relatively higher than that of the previous Training/test set.

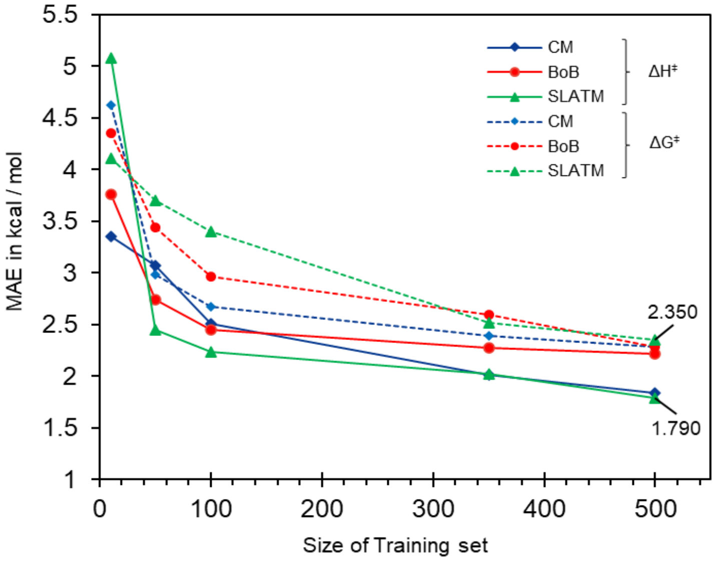


Figure 22. Learning trends for all three ML representations **(CM, BoB: Laplacian kernel, SLATM: Gaussian kernel)**

These findings can be clearly attributed to an out of sample prediction that is in the latter case the ML model does not see similar chemistry in the Training set. Consequently, this makes learning more difficult (**Fig. 22**). The solid lines represent enthalpy barriers while the dashed ones correspond to relative free energy barriers. The saturation curves depicted analogous trends in learning as before. For a small amount of Training data **(<50)**, SLATM is seen to be performing the worse than the other two representations. However, the MAE's saturate faster for larger training data and decreases almost linearly for SLATM, yielding a final validation error of 1.78±0.01 kcal/mol on the completely unseen test set. The CM representation also seems to work pretty well and this is clearly an advantage, keeping in mind that it is the least complex and most economic of all the representations.

Moreover, the plot of learned versus computed activation energy values show a good correlation demonstrating the efficiency of the ML model. Following obtaining the correlations for computed activation barriers ($\Delta H^{\ddagger}$) with respect to the reference values **(Fig. 23**), the efficiency of the ML model was validated by computing the targeted chemical property.

The chemical system under consideration being an asymmetric propargylation reaction, accurate estimation of enantiomeric excess values (ee) is of primary importance to quantify the stereoselectivity of the catalyst. We compared the ee values obtained from each of the machine learning models with those predicted at the target level (B97D/ TZV(2d,2p)).

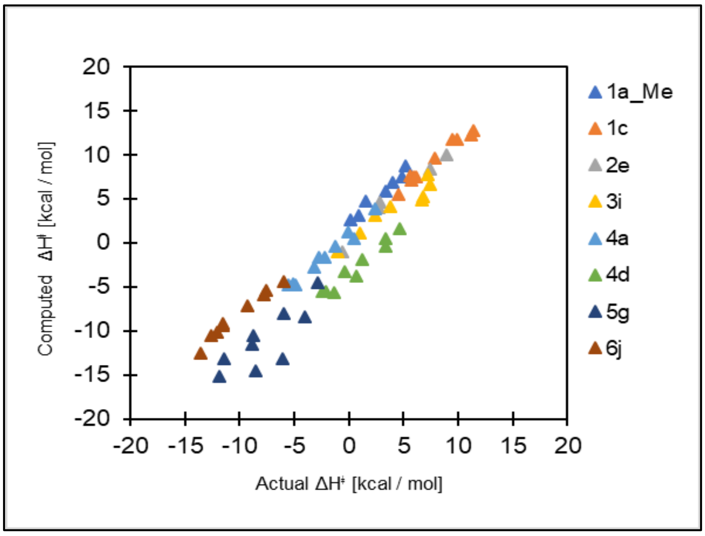


Figure 23. Distribution of Estimated Enthalpy Barriers among the Test Set (**71 data points**)

The enantiomeric excess values for all the **8** different scaffolds (**Fig.9**) in the test set have been computed based on relative enthalpies and free energies of the thermodynamically accessible TS structures (Appendix, Table **S10**). We limit our discussion to the activation enthalpies (**$\Delta H^{\ddagger}$**). As seen from the learning trends, the estimated stereoselectivities for SLATM are closest to the reference values (**Fig.24**). This ML model trained on SLATM predicted well for 6 of the 8 scaffolds as shown.

Based on previous computational work on similar propargylation and allylation catalysts, the reference enantiomeric excess(ee) values were predicted to be within 10-20% of the experiment. However, considering that we focus on only the enantioselectivity without taking into account the overall mechanism of catalytic activity, the ML models are seen to yield satisfactory estimations. Also, a mere difference of even 0.5 kcal/mol in the relative TS energies can lead to a complete inversion in configuration for the products. Consequently, the absolute enantiomeric excess values are extremely sensitive to the models tested.

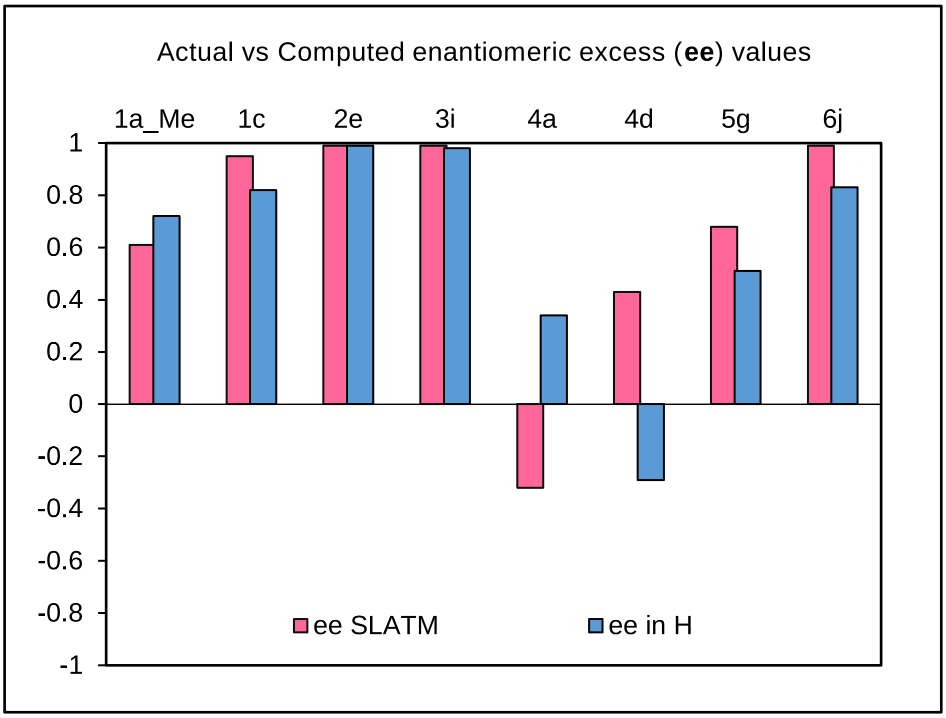


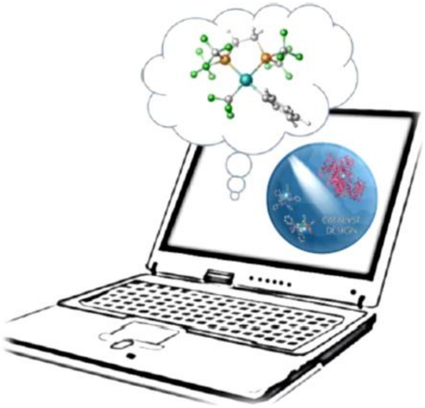
Figure 24: Histograms depicting the enantiomeric excess values. Positive ee values correspond to excess (**R**)-alcohol formation, while negative values represent excess (**S**)-alcohol. These ee values are based on relative enthalpy barriers.

(Values based on other representations and relative free energy barriers are provided in Appendix **S10, S11**)

Another significant trend is observed from the results, which is consistent with previously reported data. The catalysts build on scaffold **4** (**Fig.4)** specifically tend to be outliers and exhibit ee values quite different from those of the rest of the scaffolds even for the same substituent. This can be clearly seen from the fact that even using the SLATM representation, the ML model predicts an opposite stereoselectivity for the pair of catalysts based on scaffold **4**. This finding, however, is not unexpected given the different placement of the substituent **X** relative to the reaction centre on this scaffold when compared to the other backbones. As the ML model sees a consistent trend with the location of the substituents for a majority of the molecules in the training set, which is structurally different from that of this scaffold, it is unable to estimate the accurate stereoselectivity in this particular case. Even though we learn from previous reports that none of the catalysts build on backbone **4** has been predicted to yield high stereoselectivities, it would be an interesting case to investigate through subsequent improvements in the ML framework. This is mainly because to date only catalyst **4a** has been experimentally tested[41] for reaction **1**, with a reported ee of 52%. Also, considering that there are still several caveats on the exact mechanism, the present results demonstrate the feasibility of using machine learning approaches to estimate selectivity.

5. Conclusion

We have trained and used machine learning models to accurately estimate the activation barriers of 62 promising organocatalysts for the asymmetric propargylation reaction. We sought to improve on the accuracy of BEP relationships by moving beyond standard linear relationships using ML. The models were based on the capability of the Bell-Evans-Polanyi principle to correlate the thermodynamics and kinetics of a catalytic reaction. Overall, we have studied a database of 576 complexes based on the bipyridine N, N'- dioxide scaffold. Our findings indicate that machine learning representations can be successfully applied on data computed at inexpensive semiempirical levels and ultimately used to predict the energetics at a higher level of theory, retaining a sufficient degree of accuracy. This work demonstrated the applicability of DFT coupled with machine learning models to quantitatively estimate characteristic chemical properties of a reaction, bypassing the detailed kinetic mechanism.



The feasibility of the Δ-machine learning approach paves way for an appealing future improvement of the proposed ML framework further using semi-empirical methods (e.g. DFTB+, GFN$_2$-xtb)[64-65]. However, such frameworks would require tuning certain theoretical parameters and further be optimized according to the selected database. Also, it would be imperative to modify the algorithm to estimate the transition state energies with respect to both reactant and products simultaneously, using an active learning approach. Such a framework effectively establishes a novel design paradigm in which the database can be extended to more potential organo-catalysts with experimentally testable predictions and ultimately lead to large scale screening of promising candidates using minimal computational resources.

6. References

1. Fagan, Paul J., Elisabeth Hauptman, Rafael Shapiro, and Albert Casalnuovo. "Using Intelligent/Random Library Screening To Design Focused Libraries for the Optimization of Homogeneous Catalysts: Ullmann Ether Formation." *Journal of the American Chemical Society* 122, no. 21 (May 1, **2000**): 5043–51. https://doi.org/10.1021/ja000094c.
2. Reetz, Manfred T. "Combinatorial and Evolution-Based Methods in the Creation of Enantioselective Catalysts." *Angewandte Chemie International Edition* 40, no. 2 (**2001**): 284–310. https://doi.org/10.1002/1521-3773(20010119)40:2<284::AID-ANIE284>3.0.CO;2-N.
3. Weis, Martine, Christoph Waloch, Wolfgang Seiche, and Bernhard Breit. "Self-Assembly of Bidentate Ligands for Combinatorial Homogeneous Catalysis: Asymmetric Rhodium-Catalyzed Hydrogenation." *Journal of the American Chemical Society* 128, no. 13 (April 5, **2006**): 4188–89. https://doi.org/10.1021/ja058202o.
4. Stambuli, James P., Shaun R. Stauffer, Kevin H. Shaughnessy, and John F. Hartwig. "Screening of Homogeneous Catalysts by Fluorescence Resonance Energy Transfer. Identification of Catalysts for Room-Temperature Heck Reactions." *Journal of the American Chemical Society* 123, no. 11 (March 1, **2001**): 2677–78. https://doi.org/10.1021/ja0058435.
5. Senkan, Selim M. "High-Throughput Screening of Solid-State Catalyst Libraries." *Nature* 394, no. 6691 (July **1998**): 350–53. https://doi.org/10.1038/28575.
6. Cong, P., A. Dehestani, R. Doolen, D. M. Giaquinta, S. Guan, V. Markov, D. Poojary, K. Self, H. Turner, and W. H. Weinberg. "Combinatorial Discovery of Oxidative Dehydrogenation Catalysts within the Mo-V-Nb-O System." *Proceedings of the National Academy of Sciences* 96, no. 20 (September 28, **1999**): 11077–80. https://doi.org/10.1073/pnas.96.20.11077.
7. Fey, Natalie. "The Contribution of Computational Studies to Organometallic Catalysis: Descriptors, Mechanisms and Models." *Dalton Transactions* 39, no. 2 (December 15, **2009**): 296–310. https://doi.org/10.1039/B913356A.
8. Siegel, Justin B., Alexandre Zanghellini, Helena M. Lovick, Gert Kiss, Abigail R. Lambert, Jennifer L. St.Clair, Jasmine L. Gallaher, et al. "Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction." *Science* 329, no. 5989 (July 16, **2010**): 309–13. https://doi.org/10.1126/science.1190239.
9. Nørskov, J. K., T. Bligaard, J. Rossmeisl, and C. H. Christensen. "Towards the Computational Design of Solid Catalysts." *Nature Chemistry* 1, no. 1 (April **2009**): 37–46. https://doi.org/10.1038/nchem.121.
10. Greeley, Jeffrey. "Theoretical Heterogeneous Catalysis: Scaling Relationships and Computational Catalyst Design." *Annual Review of Chemical and Biomolecular Engineering* 7, no. 1 (**2016**): 605–35. https://doi.org/10.1146/annurev-chembioeng-080615-034413.
11. Sutton, Jonathan E., and Dionisios G. Vlachos. "A Theoretical and Computational Analysis of Linear Free Energy Relations for the Estimation of Activation Energies." *ACS Catalysis* 2, no. 8 (August 3, **2012**): 1624–34. https://doi.org/10.1021/cs3003269.
12. T. Bligaard, J.K. Nørskov, S. Dahl, J. Matthiesen, C.H. Christensen, J. Sehested. "The Brønsted–Evans–Polanyi relation and the volcano curve in heterogeneous catalysis." *Journal of Catalysis*, Volume 224, Issue 1, **2004**, 206-217. https://doi.org/10.1016/j.jcat.2004.02.034

13. Doney, Analise C., Benjamin J. Rooks, Tongxiang Lu, and Steven E. Wheeler. "Design of Organocatalysts for Asymmetric Propargylations through Computational Screening." *ACS Catalysis* 6, no. 11 (November 4, **2016**): 7948–55. https://doi.org/10.1021/acscatal.6b02366.
14. Houk, K. N., and Paul Ha-Yeon Cheong. "Computational Prediction of Small-Molecule Catalysts." *Nature* 455, no. 7211 (September **2008**): 309–13. https://doi.org/10.1038/nature07368.
15. Reid, Jolene P., Luis Simón, and Jonathan M. Goodman. "A Practical Guide for Predicting the Stereochemistry of Bifunctional Phosphoric Acid Catalyzed Reactions of Imines." *Accounts of Chemical Research* 49, no. 5 (May 17, **2016**): 1029–41. https://doi.org/10.1021/acs.accounts.6b00052.
16. Schaller, R.R. "Moore's Law: Past, Present and Future." *IEEE Spectrum* 34, no. 6 (June **1997**): 52–59. https://doi.org/10.1109/6.591665.
17. Sabatier, P., *La Catalyse En Chimie Organique*, Librairie polytechnique: **1913**
18. (a) Evans, M. G.; Polanyi, M., "Inertia and Driving Force of Chemical Reactions". *Trans. Faraday Soc.* **1938,** *34* (0), 11-24. (b) Bell, R. P., "The Theory of Reactions Involving Proton Transfers". *Proc. Royal Soc. A.* **1936,** *154* (882), 414-429. (c) Brönsted, J. N.; Pedersen, K., "Die Katalytische Zersetzung Des Nitramids Und Ihre Physikalisch-Chemische Bedeutung". *Z. Phys. Chem.* **1924,** *108*, 185-235 (d) Jensen, Frank. *Introduction to Computational Chemistry*. 2nd ed. Chichester, England; Hoboken, NJ: John Wiley & Sons, **2007**
19. (a) Hammett, Louis P. "The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives." *Journal of the American Chemical Society* 59, no. 1 (January **1937**): 96–103. https://doi.org/10.1021/ja01280a022 (b) Hammett, Louis P. "Some Relations between Reaction Rates and Equilibrium Constants." *Chemical Reviews* 17, no. 1 (August **1935**): 125–36. https://doi.org/10.1021/cr60056a010. (c) Hammett, Louis P. "Linear Free Energy Relationships in Rate and Equilibrium Phenomena." *Transactions of the Faraday Society* 34 (**1938**): 156. https://doi.org/10.1039/tf9383400156.
20. Wang, Shengguang, Vassili Vorotnikov, Jonathan E. Sutton, and Dionisios G. Vlachos. "Brønsted–Evans–Polanyi and Transition State Scaling Relations of Furan Derivatives on Pd(111) and Their Relation to Those of Small Molecules." *ACS Catalysis* 4, no. 2 (February 7, **2014**): 604–12. https://doi.org/10.1021/cs400942u.
21. Santen, Rutger A. van, Matthew Neurock, and Sharan G. Shetty. "Reactivity Theory of Transition-Metal Surfaces: A Brønsted−Evans−Polanyi Linear Activation Energy−Free-Energy Analysis." *Chemical Reviews* 110, no. 4 (April 14, **2010**): 2005–48. https://doi.org/10.1021/cr9001808.
22. Nigam, Abhash, and Michael T. Klein. "A Mechanism-Oriented Lumping Strategy for Heavy Hydrocarbon Pyrolysis: Imposition of Quantitative Structure-Reactivity Relationships for Pure Components." *Industrial & Engineering Chemistry Research* 32, no. 7 (July 1, **1993**): 1297–1303. https://doi.org/10.1021/ie00019a003.
23. Man, Isabela C., Hai-Yan Su, Federico Calle-Vallejo, Heine A. Hansen, José I. Martínez, Nilay G. Inoglu, John Kitchin, Thomas F. Jaramillo, Jens K. Nørskov, and Jan Rossmeisl. "Universality in Oxygen Evolution Electrocatalysis on Oxide Surfaces." *ChemCatChem* 3, no. 7 (July 11, **2011**): 1159–65. https://doi.org/10.1002/cctc.201000397.
24. Cova, Tânia F. G. G., and Alberto A. C. C. Pais. "Deep Learning for Deep Chemistry: Optimizing the Prediction of Chemical Patterns." *Frontiers in Chemistry* 7 (**2019**): 809. https://doi.org/10.3389/fchem.2019.00809.
25. Yang, Wenhong, Timothy Tizhe Fidelis, and Wen-Hua Sun. "Machine Learning in Catalysis, From Proposal to Practicing." *ACS Omega* 5, no. 1 (January 14, **2020**): 83–88. https://doi.org/10.1021/acsomega.9b03673.

26. Aspuru-Guzik, Alán, Mu-Hyun Baik, Shankar Balasubramanian, Rahul Banerjee, Suzanne Bart, Nadine Borduas-Dedekind, Sukbok Chang, et al. "Charting a Course for Chemistry." *Nature Chemistry* 11, no. 4 (April **2019**): 286–94. https://doi.org/10.1038/s41557-019-0236-7.
27. Denmark *et. al., Science,* 18 Jan **2019**: Vol. 363, Issue 6424
28. Jordan, M. I., and T. M. Mitchell. "Machine Learning: Trends, Perspectives, and Prospects." *Science* 349, no. 6245 (July 17, **2015**): 255–60. https://doi.org/10.1126/science.aaa8415.
29. Nayak, Sanjay, Satadeep Bhattacharjee, Jung-Hae Choi, and Seung Cheol Lee. "Machine Learning and Scaling Laws for Prediction of Accurate Adsorption Energy." *The Journal of Physical Chemistry A* 124, no. 1 (January 9, **2020**): 247–54. https://doi.org/10.1021/acs.jpca.9b07569.
30. Abdelfatah, Kareem, Wenqiang Yang, Rajadurai Vijay Solomon, Biplab Rajbanshi, Asif Chowdhury, Mehdi Zare, Subrata Kumar Kundu, Adam Yonge, Andreas Heyden, and Gabriel Terejanu. "Prediction of Transition-State Energies of Hydrodeoxygenation Reactions on Transition-Metal Surfaces Based on Machine Learning." *The Journal of Physical Chemistry C* 123, no. 49 (December 12, **2019**): 29804–10. https://doi.org/10.1021/acs.jpcc.9b10507.
31. "Machine Learning in Catalysis", *Nature Catalysis*. https://www.nature.com/articles/s41929-018-0056-y.
32. Meyer, Benjamin, Boodsarin Sawatlon, Stefan Heinen, O. Anatole von Lilienfeld, and Clémence Corminboeuf. "Machine Learning Meets Volcano Plots: Computational Discovery of Cross-Coupling Catalysts." *Chemical Science* 9, no. 35 (**2018**): 7069–77. https://doi.org/10.1039/C8SC01949E.
33. Durand, Derek J., and Natalie Fey. "Computational Ligand Descriptors for Catalyst Design." *Chemical Reviews* 119, no. 11 (June 12, 2019): 6561–94. https://doi.org/10.1021/acs.chemrev.8b00588.
34. (a) Sanchez-Lengeling, Benjamin, and Alán Aspuru-Guzik. "Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering." *Science* 361, no. 6400 (July 27, **2018**): 360–65. https://doi.org/10.1126/science.aat2663. (b) Sánchez-Lengeling, Benjamín, and Alán Aspuru-Guzik. "Learning More, with Less." *ACS Central Science* 3, no. 4 (April 26, **2017**): 275–77. https://doi.org/10.1021/acscentsci.7b00153.
35. Banerjee, Sayan, A. Sreenithya, and Raghavan B. Sunoj. "Machine Learning for Predicting Product Distributions in Catalytic Regioselective Reactions." *Physical Chemistry Chemical Physics* 20, no. 27 (**2018**): 18311–18. https://doi.org/10.1039/C8CP03141J.
36. Reid, J.P., Sigman, M.S. Holistic prediction of enantioselectivity in asymmetric catalysis. *Nature* **571,** 343–348 (**2019**). https://doi.org/10.1038/s41586-019-1384-z
37. Lam, Yu-hong, Matthew N. Grayson, Mareike C. Holland, Adam Simon, and K. N. Houk. "Theory and Modeling of Asymmetric Catalytic Reactions." *Accounts of Chemical Research* 49, no. 4 (April 19, **2016**): 750–62. https://doi.org/10.1021/acs.accounts.6b00006.
38. Chen, J. S.; Captain, B.; Takenaka, N. *Org. Lett*. **2011**, 13, 1654−1657.
39. Lu, Tongxiang, Rongxiu Zhu, Yi An, and Steven E. Wheeler. "Origin of Enantioselectivity in the Propargylation of Aromatic Aldehydes Catalyzed by Helical N-Oxides." *Journal of the American Chemical Society* 134, no. 6 (February 15, **2012**): 3095–3102. https://doi.org/10.1021/ja209241n.
40. Rooks, Benjamin J., Madison R. Haas, Diana Sepúlveda, Tongxiang Lu, and Steven E. Wheeler. "Prospects for the Computational Design of Bipyridine *N* , *N* '-Dioxide Catalysts for Asymmetric Propargylation Reactions." *ACS Catalysis* 5, no. 1 (January 2, **2015**): 272–80. https://doi.org/10.1021/cs5012553.

41. Nakajima, Makoto, Makoto Saito, Motoo Shiro, and Shun-ichi Hashimoto. “(S)-3,3‘-Dimethyl-2,2‘-Biquinoline N,N‘-Dioxide as an Efficient Catalyst for Enantioselective Addition of Allyltrichlorosilanes to Aldehydes.” *Journal of the American Chemical Society* 120, no. 25 (July **1998**): 6419–20. https://doi.org/10.1021/ja981091r.

42. Guan, Yanfei, Victoria M. Ingman, Benjamin J. Rooks, and Steven E. Wheeler. “AARON: An Automated Reaction Optimizer for New Catalysts.” *Journal of Chemical Theory and Computation* 14, no. 10 (October 9, **2018**): 5249–61. https://doi.org/10.1021/acs.jctc.8b00578

43. *Gaussian 09* (Gaussian, Inc., Wallingford CT, **2009**)

44. Schäfer, Ansgar, Christian Huber, and Reinhart Ahlrichs. “Fully Optimized Contracted Gaussian Basis Sets of Triple Zeta Valence Quality for Atoms Li to Kr.” *The Journal of Chemical Physics* 100, no. 8 (April 15, **1994**): 5829–35. https://doi.org/10.1063/1.467146.

45. Gordon, Mark S., J. Stephen Binkley, John A. Pople, William J. Pietro, and Warren J. Hehre. “Self-Consistent Molecular-Orbital Methods. 22. Small Split-Valence Basis Sets for Second-Row Elements.” *Journal of the American Chemical Society* 104, no. 10 (May **1982**): 2797–2803. https://doi.org/10.1021/ja00374a017

46. Frisch, Michael J., John A. Pople, and J. Stephen Binkley. “Self-consistent Molecular Orbital Methods 25. Supplementary Functions for Gaussian Basis Sets.” *The Journal of Chemical Physics* 80, no. 7 (April **1984**): 3265–69. https://doi.org/10.1063/1.447079.

47. Binkley, J. Stephen, John A. Pople, and Warren J. Hehre. “Self-Consistent Molecular Orbital Methods. 21. Small Split-Valence Basis Sets for First-Row Elements.” *Journal of the American Chemical Society* 102, no. 3 (January **1980**): 939–47. https://doi.org/10.1021/ja00523a008.

48. (a) Tomasi, Jacopo, Benedetta Mennucci, and Roberto Cammi. “Quantum Mechanical Continuum Solvation Models.” *Chemical Reviews* 105, no. 8 (August **2005**): 2999–3094. https://doi.org/10.1021/cr9904009. (b) Nottoli, Michele, Benjamin Stamm, Giovanni Scalmani, and Filippo Lipparini. “Quantum Calculations in Solution of Energies, Structures, and Properties with a Domain Decomposition Polarizable Continuum Model.” *Journal of Chemical Theory and Computation* 15, no. 11 (November 12, **2019**): 6061–73. https://doi.org/10.1021/acs.jctc.9b00640.

49. Sure, Rebecca, and Stefan Grimme. “Corrected Small Basis Set Hartree-Fock Method for Large Systems.” *Journal of Computational Chemistry* 34, no. 19 (July 15, **2013**): 1672–85. https://doi.org/10.1002/jcc.23317.

50. (a) Grimme, Stefan, Stephan Ehrlich, and Lars Goerigk. “Effect of the Damping Function in Dispersion Corrected Density Functional Theory.” *Journal of Computational Chemistry* 32, no. 7 (May **2011**): 1456–65. https://doi.org/10.1002/jcc.21759. (b) Grimme, Stefan, Jens Antony, Stephan Ehrlich, and Helge Krieg. “A Consistent and Accurate *Ab Initio* Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu.” *The Journal of Chemical Physics* 132, no. 15 (April 21, **2010**): 154104. https://doi.org/10.1063/1.3382344.

51. Stewart, James J. P. “Optimization of Parameters for Semiempirical Methods V: Modification of NDDO Approximations and Application to 70 Elements.” *Journal of Molecular Modeling* 13, no. 12 (October 20, **2007**): 1173–1213. https://doi.org/10.1007/s00894-007-0233-4.

52. (a) Neese, Frank. “Software Update: The ORCA Program System, Version 4.0.” *WIREs Computational Molecular Science* 8, no. 1 (January **2018**). https://doi.org/10.1002/wcms.1327. (b) Neese, Frank. “The ORCA Program System.” *WIREs Computational Molecular Science* 2, no. 1 (January **2012**): 73–78. https://doi.org/10.1002/wcms.81

53. MOPAC2016, James J. P. Stewart, Stewart Computational Chemistry, Colorado Springs, CO, USA, HTTP://OpenMOPAC.net (**2016**).

54. K. Vu, J. C. Snyder, L. Li, M. Rupp, B. F. Chen, T. Khelif, K.-R. Muller, K. Burke. "Understanding Kernel Ridge Regression: Common Behaviors from Simple Functions to Density Functionals" - *International Journal of Quantum Chemistry* **2015**, 115, 1115-1128 https://onlinelibrary.wiley.com/doi/pdf/10.1002/qua.24939.

55. Bereau, Tristan, Denis Andrienko, and O. Anatole von Lilienfeld. "Transferable Atomic Multipole Machine Learning Models for Small Organic Molecules." *Journal of Chemical Theory and Computation* 11, no. 7 (July 14, **2015**): 3225–33. https://doi.org/10.1021/acs.jctc.5b00301.

56. Faber, Felix A., Anders S. Christensen, Bing Huang, and O. Anatole von Lilienfeld. "Alchemical and Structural Distribution Based Representation for Universal Quantum Machine Learning." *The Journal of Chemical Physics* 148, no. 24 (June 28, **2018**): 241717. https://doi.org/10.1063/1.5020710.

57. Hansen, Katja, Franziska Biegler, Raghunathan Ramakrishnan, Wiktor Pronobis, O. Anatole von Lilienfeld, Klaus-Robert Müller, and Alexandre Tkatchenko. "Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space." *The Journal of Physical Chemistry Letters* 6, no. 12 (June 18, **2015**): 2326–31. https://doi.org/10.1021/acs.jpclett.5b00831.

58. "Quantum Machine Learning in Chemical Compound Space." https://onlinelibrary.wiley.com/doi/epdf/10.1002/anie.201709686.

59. Ramakrishnan, Raghunathan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. "Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach." *Journal of Chemical Theory and Computation* 11, no. 5 (May 12, **2015**): 2087–96. https://doi.org/10.1021/acs.jctc.5b00099.

60. Rupp, Matthias, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. "Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning." *Physical Review Letters* 108, no. 5 (January 31, **2012**): 058301. https://doi.org/10.1103/PhysRevLett.108.058301.

61. Axilrod, B. M., and E. Teller. "Interaction of the van Der Waals Type Between Three Atoms." *The Journal of Chemical Physics* 11, no. 6 (June **1943**): 299–300. https://doi.org/10.1063/1.1723844.

62. A. Christensen, F. Faber, B. Huang, L. Bratholm, A. Tkatchenko, K. Muller and O. von Lilienfeld, "Qml: A python toolkit for quantum machine learning", **2017**

63. (a) T. E. Oliphant, "A guide to NumPy", Trelgol Publishing USA, **2006**, vol. 1. (b) J. D. Hunter, "Computing in science & engineering", **2007**, 9, 90

64. Aradi, B., B. Hourahine, and Th. Frauenheim. "DFTB+, a Sparse Matrix-Based Implementation of the DFTB Method †." *The Journal of Physical Chemistry A* 111, no. 26 (July **2007**): 5678–84. https://doi.org/10.1021/jp070186p.

65. Bannwarth, Christoph, Sebastian Ehlert, and Stefan Grimme. "GFN2-XTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions." *Journal of Chemical Theory and Computation* 15, no. 3 (March 12, **2019**): 1652–71. https://doi.org/10.1021/acs.jctc.8b01176.

Appendix

1. Preliminary Results: Δ-ML on Activation Barriers

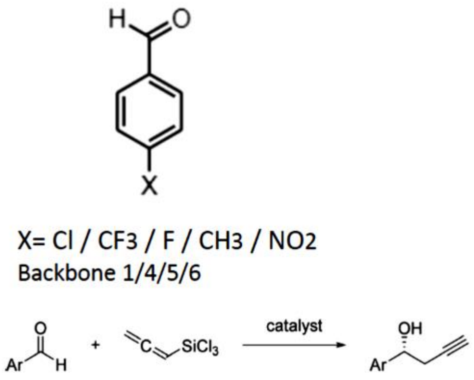


Figure S1. The Library of Substrates included in the Initial Database

We began our initial study by taking the previously computed database of 62 catalysts by Wheeler and co-workers, and then further expanded it by including more potential substrates. For each catalyst/substrate combination we optimized all possible TS structures which yielded a virtual library of 539 TS structures.

We tested the KRR method using each of the three representations to validate the ML models on a test set of 40 datapoints. The input structures were computed at **B97D / 3-21G** and Δ-ML was performed by learning the energy differences with respect to the target **B97D / TZV (2p,2d)** level.

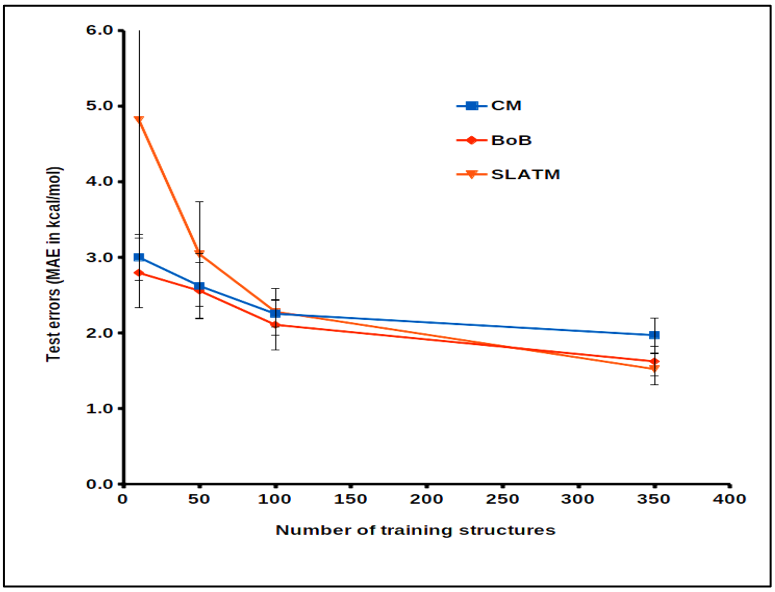


Figure S2. Preliminary Saturation curves depicting the Test errors on 40 TS molecules using three different representations

2. Preliminary Results: Learning using DFT Baseline Method

Table S3: Mean Absolute Errors (MAE) using KRR: Input Properties computed at B97D/3-21G; Target values computed at B97D/TZV (2p,2d). CM, BoB- Laplacian kernel, SLATM- Gaussian kernel

	Training set size	CM	Std. dev.	BoB	Std. dev.	SLATM	Std. dev.
		[kcal/mol]	[kcal/mol]	[kcal/mol]	[kcal/mol]	[kcal/mol]	[kcal/mol]
R/H	10	3.613	1.131	5.020	0.975	4.084	0.887
	50	2.751	0.244	3.498	1.483	1.756	0.108
	100	2.261	0.109	2.161	0.558	1.368	0.156
	350	1.755	0.094	1.355	0.191	0.862	0.062
	500	1.572	0.020	1.224	0.014	0.777	0.008
R/G	10	3.860	0.516	5.568	2.231	5.857	3.884
	50	3.004	0.169	3.117	0.474	2.055	0.293
	100	2.671	0.109	2.380	0.690	1.580	0.152
	350	1.967	0.114	1.675	0.208	0.937	0.041
	500	1.744	0.009	1.675	0.028	0.847	0.007
P/H	10	4.546	0.211	5.215	1.423	5.659	2.797
	50	3.743	0.235	3.556	0.801	2.699	0.464
	100	3.304	0.202	3.430	0.732	2.345	0.580
	350	2.561	0.112	2.369	0.395	1.966	0.230
	500	2.409	0.016	2.187	0.066	1.876	0.019
P/G	10	5.298	0.559	5.520	5.520	5.585	1.670
	50	4.002	0.159	3.861	3.861	2.947	0.693
	100	3.446	0.187	3.132	3.132	2.276	0.246
	350	2.672	0.116	2.543	2.543	1.625	0.096
	500	2.442	0.024	2.168	2.168	1.538	0.010

R- Reactant intermediates P- Product intermediates H- Enthalpy barriers G- Free Energy barriers

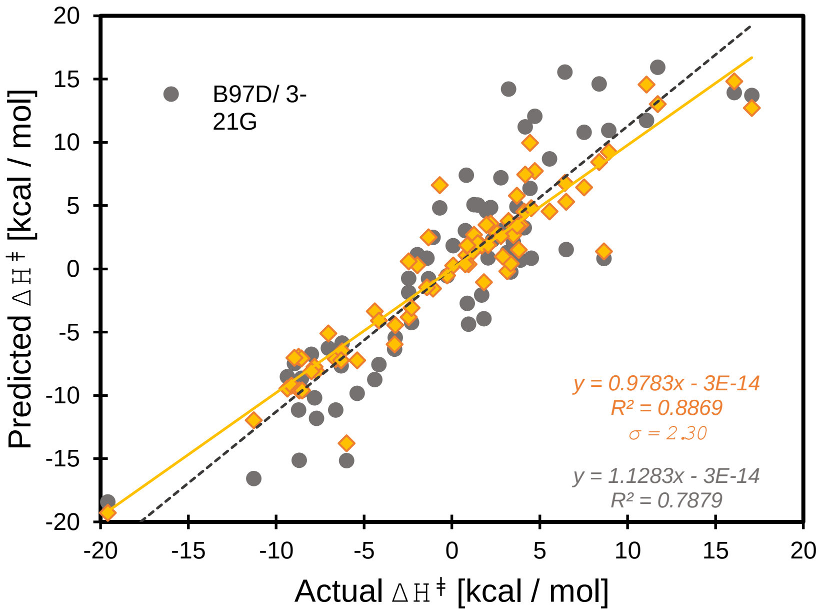
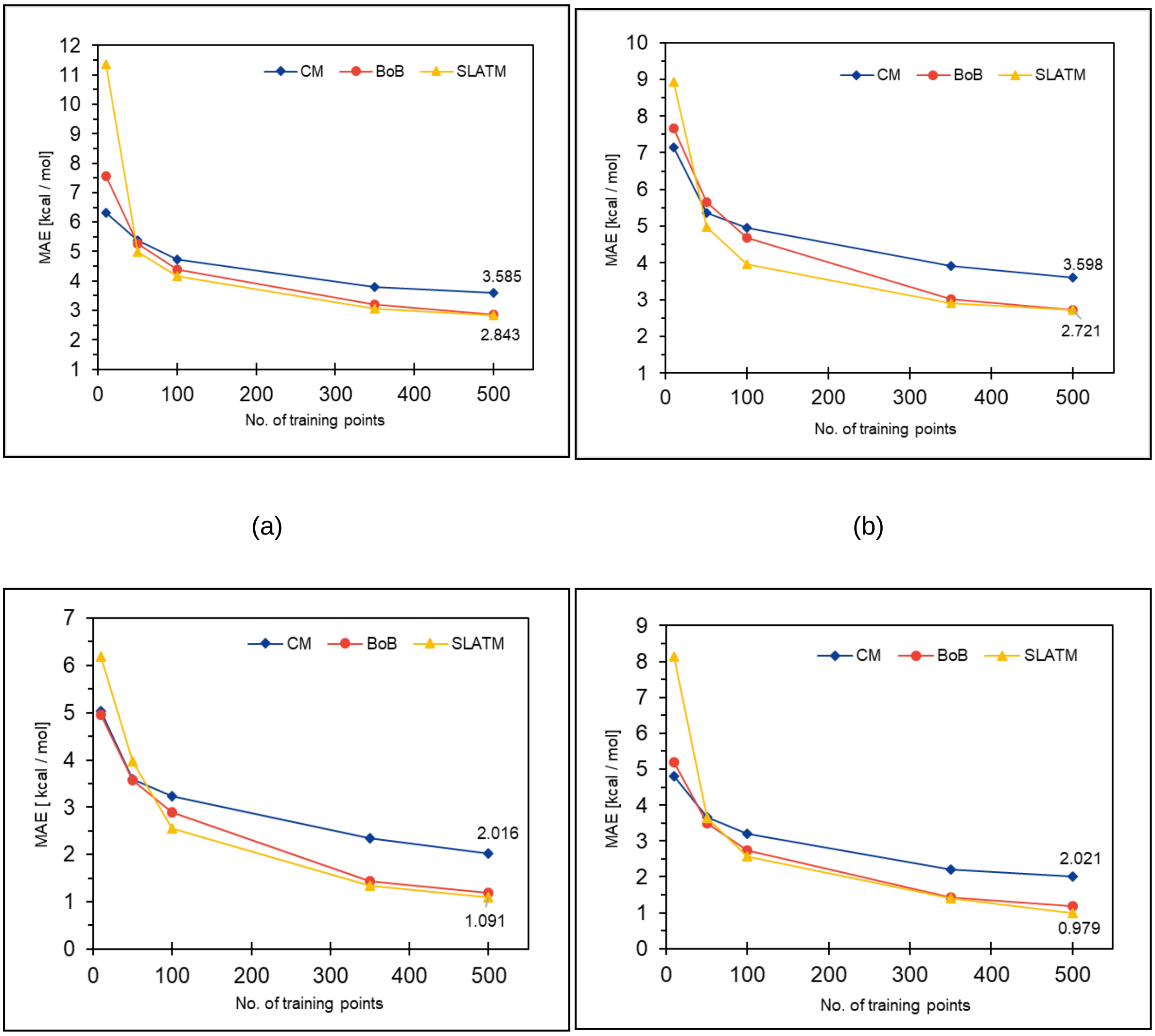


Figure S4. ML predicted (y-axis) and actual (x-axis) values of the descriptor for **B97D/3-21G** Reactant intermediates compared on a validation set of **71** points after training 505 datapoints. The identity line (y = x, in black), corresponds to perfect predictions of the descriptor value.

3. Learning trends for semi-empirical baseline methods:



(a)

(b)

(c)

(d)

Figure S5. Saturation curves (MAE on a test set as a function of the number of training data) constructed for all representations considered, with Reactant intermediates computed using **HF-3c** targeting (a) ΔH‡ (b) ΔG‡ and **PM6-D3** targeting (c) ΔH‡ (d) ΔG‡, respectively

4. Learning using semi-empirical baseline methods:

Table S6: Mean Absolute Errors (MAE) using KRR: Input Properties computed at PM6-D3; Target values computed at B97D/TZV (2p,2d). CM, BoB- Laplacian kernel, SLATM- Gaussian kernel

R/H	Training set size	CM	Std. dev.	BoB	Std. dev.	SLATM	Std. dev.
		[kcal/mol]	[kcal/mol]	[kcal/mol]	[kcal/mol]	[kcal/mol]	[kcal/mol]
	10	5.026	0.429	4.951	0.820	6.178	1.480
	50	3.599	0.130	3.570	0.300	3.979	0.518
	100	3.230	0.213	2.894	0.209	2.555	0.306
	350	2.340	0.077	1.440	0.064	1.346	0.082
	500	2.016	0.014	1.197	0.014	1.091	0.011

R/G	Training set size	CM	Std. dev.	BoB	Std. dev.	SLATM	Std. dev.
		[kcal/mol]	[kcal/mol]	[kcal/mol]	[kcal/mol]	[kcal/mol]	[kcal/mol]
	10	4.802	0.548	5.186	0.743	8.134	2.737
	50	3.676	0.264	3.506	0.353	3.633	0.493
	100	3.210	0.225	2.751	0.291	2.574	0.188
	350	2.218	0.071	1.427	0.076	1.395	0.118
	500	2.021	0.020	1.182	0.011	0.979	0.014

Table S7: Mean Absolute Errors (MAE) using KRR: Input Properties computed at HF-3c; Target values computed at B97D/TZV (2p,2d). CM, BoB- Laplacian kernel, SLATM- Gaussian kernel

R/H	Training set size	CM	Std. dev.	BoB	Std. dev.	SLATM	Std. dev.
		[kcal/mol]	[kcal/mol]	[kcal/mol]	[kcal/mol]	[kcal/mol]	[kcal/mol]
	10	6.304	0.602	7.569	1.771	11.348	5.458
	50	5.380	0.353	5.269	0.350	4.970	0.933
	100	4.737	0.399	4.382	0.460	4.169	0.583
	350	3.780	0.185	3.205	0.160	3.045	0.175
	500	3.585	0.039	2.848	0.040	2.843	0.050

R/G	Training set size	CM	Std. dev.	BoB	Std. dev.	SLATM	Std. dev.
		[kcal/mol]	[kcal/mol]	[kcal/mol]	[kcal/mol]	[kcal/mol]	[kcal/mol]
	10	7.158	0.496	7.658	0.876	8.937	3.013
	50	5.366	0.246	5.658	0.707	4.980	0.707
	100	4.967	0.388	4.676	0.420	3.973	0.786
	350	3.907	0.133	3.016	0.236	2.898	0.239
	500	3.598	0.023	2.717	0.032	2.721	0.020

R- Reactant intermediates P- Product intermediates H- Enthalpy barriers G- Free Energy barriers

5. Estimation of Stereoselectivities:

Table S8: Computed pair of Hyperparameters for KRR

Input Structure	Target Property	CM (Laplacian kernel)		BoB (Laplacian kernel)		SLATM (Gaussian kernel)	
B97D/3-21G	B97D/TZV	σ	λ	σ	λ	σ	λ
reactant (R)	enthalpies (H)	5.11E+04	1.02E-10	9.64E+04	1.04E-10	2.85E+04	1.05E-10
	free energies (G)	1.05E+05	9.50E-11	1.05E+05	9.82E-11	1.00E+04	1.01E-10

Table S9: Mean Absolute Errors (MAE) using KRR: Input Properties computed at B97D/3-21G; Target values computed at B97D/TZV (2p,2d). CM, BoB- Laplacian kernel, SLATM- Gaussian kernel

R/H	Training set size	CM	Std. dev.	BoB	Std. dev.	SLATM	Std. dev.
		[kcal/mol]	[kcal/mol]	[kcal/mol]	[kcal/mol]	[kcal/mol]	[kcal/mol]
	10	3.350	0.402	3.76	0.455	5.080	1.432
	50	3.070	0.296	2.74	0.401	2.450	0.360
	100	2.510	0.167	2.45	0.238	2.240	0.096
	350	2.010	0.125	2.28	0.100	2.020	0.244
	500	1.840	0.014	2.22	0.014	1.790	0.011

R/G	Training set size	CM	Std. dev.	BoB	Std. dev.	SLATM	Std. dev.
		[kcal/mol]	[kcal/mol]	[kcal/mol]	[kcal/mol]	[kcal/mol]	[kcal/mol]
	10	4.623	0.939	4.354	1.102	4.115	1.133
	50	2.981	0.328	3.443	0.736	3.699	0.716
	100	2.674	0.214	2.967	0.477	3.401	0.521
	350	2.391	0.106	2.596	0.220	2.515	0.369
	500	2.281	0.030	2.284	0.021	2.350	0.052

R- Reactant intermediates **P-** Product intermediates **H-** Enthalpy barriers **G-** Free Energy barriers

Table S10: Enantiomeric Excess values (ee); Reference values computed using B97D/TZV (2p,2d) considering only thermodynamically accessible TS**.**

	Calculated ee values (Enthalpies)			**Reference ee values**			**Calculated ee values (Free energies)**		
Substituents (X)	**SLATM**	**BoB**	**CM**	**in H**	**in E**	**in G**	**SLATM**	**BoB**	**CM**
H	0.61	**0.66**	0.89	0.72	0.69	0.83	-0.91	0.11	**0.97**
Cl	**0.95**	0.44	0.62	0.82	0.91	0.6	**0.56**	0.88	0.53
CF3	**0.99**	**0.99**	**0.99**	0.99	0.99	0.97	**0.99**	0.79	0.71
CN	**0.99**	0.35	**0.99**	0.98	0.99	0.84	**0.99**	-0.74	**0.9**
H	-0.32	-0.82	**0.32**	0.34	0.24	0.31	-0.94	-0.99	**0.15**
Me	0.43	**-0.79**	0.55	-0.29	-0.45	0.08	**-0.43**	-0.72	0.99
tBu	**0.68**	-0.62	-0.94	0.51	-0.47	0.51	-0.17	-0.58	-0.98
Ph	**0.99**	**0.98**	**0.95**	0.83	0.88	0.63	0.99	0.99	0.12

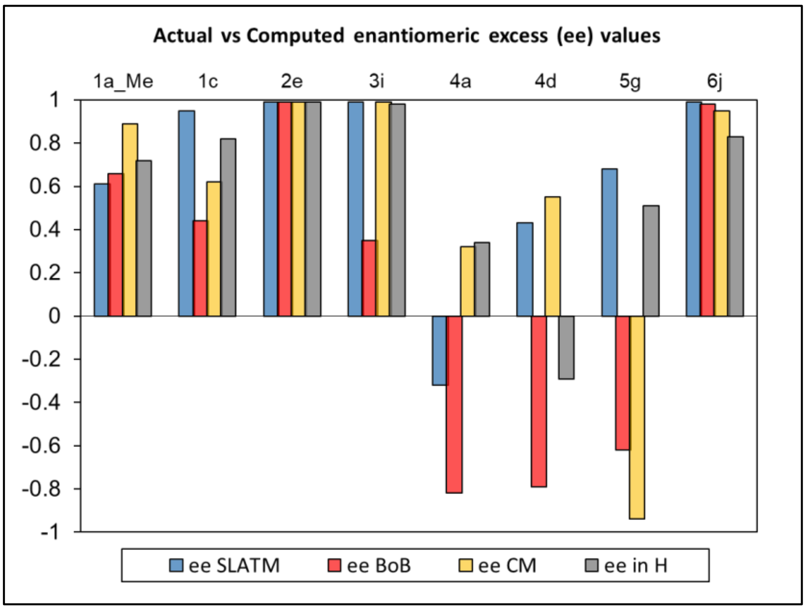


Figure S11: Histograms depicting the enantiomeric excess values. Positive ee values correspond to excess (**R**)-alcohol formation, while negative values represent excess (**S**)-alcohol. These ee values are based on relative enthalpy barriers.