EDGE ARTICLE
Clemence Corminboeuf et al.
Reaction-based machine learning representations for predicting the enantioselectivity of organocatalysts
Reaction-based machine learning representations for predicting the enantioselectivity of organocatalysts†

Simone Gallarati, Raimon Fabregat, Rubén Laplaza, Sinjini Bhattacharjee, Matthew D. Wodrich and Clemence Corminboeuf

Hundreds of catalytic methods are developed each year to meet the demand for high-purity chiral compounds. The computational design of enantioselective organocatalysts remains a significant challenge, as catalysts are typically discovered through experimental screening. Recent advances in combining quantum chemical computations and machine learning (ML) hold great potential to propel the next leap forward in asymmetric catalysis. Within the context of quantum chemical machine learning (QML, or atomistic ML), the ML representations used to encode the three-dimensional structure of molecules and evaluate their similarity cannot easily capture the subtle energy differences that govern enantioselectivity. Here, we present a general strategy for improving molecular representations within an atomistic machine learning model to predict the DFT-computed enantiomeric excess of asymmetric propargylation organocatalysts solely from the structure of catalytic cycle intermediates. Mean absolute errors as low as 0.25 kcal mol$^{-1}$ were achieved in predictions of the activation energy with respect to DFT computations. By virtue of its design, this strategy is generalisable to other ML models, to experimental data and to any catalytic asymmetric reaction, enabling the rapid screening of structurally diverse organocatalysts from available structural information.

Introduction

Society’s growing need for pharmaceuticals, agricultural chemicals, and materials requires a continuous push in the development of asymmetric catalytic methods.$^{1,11}$ In particular, enantioselective organocatalysis has emerged as a powerful strategy for the stereocontrolled assembly of structurally diverse molecules$^{3-9}$ with constant effort placed in making chemical transformations more selective, efficient, or generally applicable.$^{8}$ Although the computational design of highly selective catalysts has long been viewed as a “Holy Grail” in chemistry,$^{7,8}$ it is generally still more efficient to experimentally screen a range of potential organocatalysts for a given reaction than to assess their performance in silico.$^{9}$ That is because e.e. (enantioselectivity) values, estimated as the ratio between the competitive reaction rates leading to the two enantiomeric products,$^{10}$ are relatively computationally expensive and challenging to predict accurately with standard electronic structure computations. The energy difference between the transition states (TSs) leading to the major and minor enantiomers can be quite small ($\leq 2$ kcal mol$^{-1}$) and multiple diastereomeric transition states, stemming from the large conformational space of flexible organocatalysts, can yield the same enantiomer.$^{11,12}$ As the relation between rate constants and computed selectivity is exponential, minor errors in computed energies can lead to major errors in stereoochemical outcome prediction. These factors pose a monumental challenge for traditional quantum mechanical (QM) methods, in terms of both accuracy and cost.$^{12,13}$ especially if many conformers and substrate-catalyst combinations have to be computed. While the intrinsic error of the quantum chemical level is often addressed in comprehensive benchmark studies,$^{10,14-17}$ automated toolkits,$^{18,19}$ such as AARON$^{20}$ and CatVS,$^{21}$ have been developed to streamline the tedious and error-prone task of optimising hundreds of thermodynamically accessible stereocontrolling transition states. Starting from user-defined libraries, multiple conformations and configurations of TS structures are located and optimised.

$^1$Laboratory for Computational Molecular Design, Institute of Chemical Sciences and Engineering, Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland. E-mail: clemence.corminboeuf@epfl.ch
$^2$National Center for Competence in Research-Catalysis (NCCR-Catalysis), Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland
$^3$Indian Institute of Science Education and Research, Dr Homi Bhabha Rd, Ward No. 8, NCL Colony, Pashan, Pune, Maharashtra 411008, India
$^4$National Center for Computational Design and Discovery of Novel Materials (MARVEL), Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland
† Electronic supplementary information (ESI) available. See DOI: 10.1039/d1sc00482d
‡ These authors contributed equally to this work.
Although such accelerated prediction of selectivity is enticing for the prospect of computational catalyst design, the applicability of QM-based tools such as AARON remains limited either by the cost of the quantum mechanical computations, which quickly become prohibitive, or by the inherent difficulty of locating all transition state structures. On the other hand, tools using QM-derived molecular mechanics force fields (Q2MM), like CatVS, require the development of an MM force field for each new reaction type considered, a major limitation to their widespread application.

An alternative approach pioneered by Norrby and Pradhan and popularised by Sigman and co-workers is to fit experimental reaction outcomes to computationally- and experimentally-derived physical organic molecular descriptors. The difference in free energies at the stereocontrolling transition states can be expressed as a polynomial function of global or local steric and electronic parameters, such as Sterimol values, natural bond orbital charges, IR frequencies, HOMO/LUMO energies, and polarisabilities. In principle, the resulting statistical model allows for extrapolation to out-of-sample examples, however, like all QSSR-type methods, such multivariate linear regressions are not easily transferable and most suitable only for closely related analogues of the training set, given that a set of appropriate molecular descriptors must be redefined for every new regression.

Nonlinear regression models (e.g., artificial neural networks, random forest, kernel methods) have demonstrated the potential to overcome some of the previous limits in catalyst screening and constitute an alternative to multilinear regression models with parameters derived from chemical knowledge and mechanistic hypotheses (e.g., Hammett constants, Tolman cone angles, percent buried volume, vibrational frequencies, pKa values). Recently, the organic synthetic community has exploited these artificial intelligence-based approaches for predicting $\Delta \Delta G^\ddagger$ e.e., the activation energy, the product distribution, or the yield of (asymmetric) catalytic reactions. These models rely on the identification of a large set of system-specific molecular descriptors (e.g., physical organic descriptors like Charton or Sterimol values, NBO charges, NMR chemical shifts, bond distances and angles, HOMO–LUMO gaps, local electron/nucleophilicity, or RDKit descriptors) used as the input from which an algorithm can “learn” while being “supervised” by the reaction outcome (output, $\Delta \Delta G^\ddagger$, e.e., or yield). While the reaction outcome is often obtained from experiment (i.e., phenomenological models), alternatives based on computed data are highly valuable as well. Indeed, so-called quantum (or atomistic) ML models, which map a three-dimensional molecular structure (called molecular representations, e.g., CM, SLATM, SOAP) to a representative target computed quantum chemically, constitute an appealing complementary strategy owing to its broad applicability and dependence on the laws of physics. While these approaches provide a favourable combination of efficiency, scalability, accuracy, and transferability for predicting energetic and more complex molecular properties, identifying enantioselective organocatalysts requires precise predictions of the relative energy barriers for the stereocontrolling transition states, a target currently beyond their accuracy. Recently, SOAP features of isolated reactants were used to train a machine learning classifier and predict transition state barriers of regioselective arene C–H functionalization. In this work, a large number of molecular fingerprints were combined with the SOAP features to improve the regression, and the resulting model was outperformed in out-of-sample predictions by a random forest model using chemical descriptors with physical organic basis (PhysOrg).

Here, we provide a stepwise route to improve such QML approaches to reach sufficient accuracy for subtle properties such as those associated with asymmetric catalysis (i.e., e.e.). This objective is achieved by rationally designing a reaction-based representation (vide infra) that is a more faithful fingerprint of the enantiodetermining TS energy. The performance of the approach is demonstrated through accurately predicting the DFT-computed enantiomeric excess of Lewis base-catalysed propargylation reactions directly from the structure of the catalytic cycle intermediates. Unlike other ML models trained on (absolute) experimental e.e.’s, our model is able to predict the absolute configuration of the excess product, because it is trained on the activation energy of the enantiodetermining step for each pair of enantiomers (pro-R and pro-S) independently.

### Methods

#### Reaction and organocatalysts database

Asymmetric allylations and propargylations of aromatic aldehydes are key C–C bond forming transformations, providing access to optically enriched homoallylic and homopropargylic alcohols, respectively, which serve as valuable building blocks for the synthesis of complex chiral molecules. Catalysts that are selective for allylations are generally not highly stereoselective for propargylations, which has led to a dearth of stereoselective propargylation catalysts. Tools to screen dozens of allylation catalysts to find promising...
candidates for propargylation reactions are therefore highly valuable.\textsuperscript{9} To this end, Wheeler and co-workers have investigated 76 Lewis base organocatalysts (Scheme 1)\textsuperscript{74} and used the computational toolkit AARON\textsuperscript{20} to build a database of 760 stereocontrolling transition states to predict their enantioselectivity in the propargylation of benzaldehyde (Scheme 2).\textsuperscript{14,74,86} Large databases of kinetic data for asymmetric catalysis generated \textit{in silico} are scarce.\textsuperscript{63} Therefore, this library constitutes an ideal training and validation set for the development of an atomistic ML model with reaction-based representations capable of predicting the e.e. of organocatalysts readily from the structures of intermediates. Note that the workflow presented below would improve the ML performance independently of the size of the training data. The target of the ML model is the DFT-computed relative forward activation energy ($E_a$, i.e., the energy difference between the TS and the preceding intermediate) associated with each of the 10 ($R$)- or ($S$)-ligand arrangements (see Fig. S1$^\dagger$) of the enantiodetermining TS in Scheme 2 for the 76 catalysts in Scheme 1 (11 catalysts of type 1, 16 of type 2, 15 3, 11 4, 13 5, and 10 catalysts of type 6), yielding a total of 754 $E_a$ values.\textsuperscript{87} e.e. values are computed from $E_a$ (vide infra), thus accurate predictions of $E_a$ lead to accurate e.e. predictions.

**General ML workflow**

The general workflow exploited and improved herein relies on a physics-based ML model for the prediction of the e.e. of the asymmetric catalytic reactions, as illustrated in Scheme 3 and described hereafter. It comprises two parts: part (1) is a training procedure that relies on the following steps:

1. Database construction: a library of 3D geometries and energies of catalytic cycle intermediates is curated. Here, the structures of 754 pairs of intermediates 2 and 3 are optimised with DFT (see the next section) and used to train the ML model. As shown in our previous work,\textsuperscript{43} accurate geometries are not necessarily needed as inputs for atomistic ML models; thus, rough-coordinate estimates (e.g., obtained directly from SMILES...

![Scheme 2](image_url)  
**Scheme 2** Catalytic cycle for the propargylation of benzaldehyde with allenyltrichlorosilane, showing the rate-limiting and stereocontrolling transition state. Adapted from ref. 85.

![Scheme 3](image_url)  
**Scheme 3** Graphical overview of the workflow used to build an atomistic ML model for e.e. prediction.
strings) or low-cost DFTB structures could potentially be used to
generate suitable molecular representations.

(2) Generation of molecular representations: information
intrinsically contained within the 3D structure of each inter-
mediate is transformed into a suitable molecular representa-
tion. Here we build different variants based on the Spectral
London and Axilrod-Teller-Muto (SLATM)\textsuperscript{96}\textsuperscript{96} representation.
SLATM is composed of two- and three-body potentials, which are
derived from the atomic coordinates and contain most of the
relevant information to predict molecular properties.\textsuperscript{70,86–94}

(3) Training of the model: input representations are mapped
onto the corresponding target values ($E_a$, computed at the DFT
level, see the next section) using Kernel Ridge Regression
(KRR)\textsuperscript{95} with a Gaussian kernel. Note that even if target values
based on DFT are used here to train the ML model, the strategy
proposed hereafter is expected to perform equally well on experi-
mental or more accurate quantum chemical data.

(4) Hyperparameter optimisation and cross-validation: the
full dataset is split randomly 100 times into 90/10 training/test
sets (678/76 datapoints) to optimise the KRR hyperparameters
and obtain the learning curves.

In part (2), the trained ML model is used to predict the
activation energy of out-of-sample organocatalysts. The model
requires as input the 3D structures of 2 and 3 and delivers the
corresponding $E_a$ value. Using the energy of 2 as reference, the
relative energies of the enantiodetermining ($R$) - and ($S$)-TSs can be
calculated, and the e.e. of the catalyst under investigation computed (vide infra).

### Computational details

#### Quantum chemistry

Catalytic cycle intermediates 2 and 3 were optimised at the B97-
D/TZV(2p,2d) level of theory,\textsuperscript{96–99} accounting for solvent effects
dichloromethane, $\varepsilon = 8.93$) using the polarizable continuum
model (PCM)\textsuperscript{99–101} with Gaussian16,\textsuperscript{102,103} in analogy with the study by Wheeler and co-workers.\textsuperscript{74} Density fitting techniques
were used throughout. The structures of 1508 intermediates were
obtained via intrinsic reaction coordinate calculations
(IRC)\textsuperscript{104} from the TS database curated by Wheeler et al.\textsuperscript{74} 754
target $E_a$ values (11 catalysts of type 1, 16 type 2, 15 3, 11 4, 13 5,
and 10 of type 6, each in 5 distinct pro-($R$) and pro-($S$) ligand
arrangements)\textsuperscript{97} were computed (relative to the lowest-lying
intermediate 2 ligand arrangement) at the same level, which was
shown to provide the best compromise between accurate
predictions of low-lying TS energies and stereoselectivities for
allylation and propargylation reactions.\textsuperscript{14} e.e. values were not
predicted from Gibbs free energy barriers, but rather from
relative energy barriers (i.e., electronic energies plus solvation
free energies), since they have been found to be more reliable
than those based on either relative enthalpies or free energy
barriers for this reaction.\textsuperscript{14} The symbol $E_a$ was therefore used to
indicate the energy (electronic plus solvation) difference be-
tween the TS and the preceding intermediate. For each $C_2$-
symmetric catalyst (Scheme 1), 10 distinct ligand arrangements
around a hexacoordinate Si centre are possible (BP1–5, ($R$)- and
($S$)-, Fig. S1†).\textsuperscript{84–86} Since each of these can lead to
thermodynamically accessible reaction pathways, and the ster-
oselectivity is largely a consequence of which ligand arrange-
ment is low-lying for a particular catalyst, all diastereomeric TSs
were considered viable and the e.e. calculated from a Boltz-
mann weighting of the relative energy barriers.\textsuperscript{74} In eqn (1)–(3),
$\Delta E_{a,eff}$ is the relative Boltzmann-weighted activation energy of
each ($R$)- or ($S$)-species, $\Delta \Delta E^\ddagger$ is the difference between the ($R$)-
and ($S$)-Boltzmann-weighted activation energies, $R$ is the ideal
gas constant, and $T$ is the propargylation reaction temperature
(195 K).

\[
\Delta E_{a,eff} = -RT \ln \left( \sum_{i} e^{-\frac{E_{a,i}}{k_B T}} \right)
\]

\[
\Delta \Delta E^\ddagger = \Delta E_{a,eff}^{(R)} - \Delta E_{a,eff}^{(S)}
\]

\[
e.e. (\%) = 100 \times \frac{1 - e^{\Delta \Delta E^\ddagger / RT}}{1 + e^{\Delta \Delta E^\ddagger / RT}}
\]

#### Machine learning

The Python package QML\textsuperscript{95} was used to construct standard
SLATM representations. Feature selection and the construction of the reaction-based representations SLATMDIFF and
SLATMDIFF$^+$ were done using the Python package Scikit-learn.\textsuperscript{100}
To generate the learning curves and the e.e. predictions, a cross-
validation scheme was used with 100 different 90/10 training/
test sets (678/76). The KRR hyperparameters (the width of the
Gaussian kernel $\sigma$ and the ridge regularization $\lambda$) were opti-
mised for each train/test split, systematically obtaining essen-
tially the same results for each split (see the ESI†). From the 100
train/test splits, the $E_a$ of each intermediate pair (2 and 3) was
predicted approximately 10 times; these test predictions were
then averaged to obtain one final prediction. The standard
deviation from the $\sim$10 test predictions were used to generate
the error bars. The final average prediction of the $E_a$ value was
used to calculate the Boltzmann-weighted $\Delta E\text{e}_{a,eff}$ values (eqn (1))
and the $\Delta \Delta E^\ddagger$ of each ($R$)- and ($S$)-pair (eqn (2)), and so the e.e.
value of each organocatalyst (eqn (3)). The out-of-sample
predictions were done with the same SLATMDIFF$^+$ models
trained in the cross-validation scheme. Additionally, out-of
sample predictions were done re-training the model on the
entire dataset (see Fig. S6†), although this did not lead to
noticeable improvement. While simpler representations (e.g.,
CM,\textsuperscript{84} BoB\textsuperscript{87,88}) were tested, SLATM performs significantly better (see Fig. S2†).

#### Results and discussion

#### Molecular representations

The key step of the workflow presented above is generating a
molecular representation, which is mapped onto the target
value (i.e., the activation energy $E_a$) and used as a fingerprint of
the enantiodetermining TS. Representations can be constructed
from single molecules and more recently as “ensemble repre-
sentations”: instead of associating one fixed configuration of
atoms to a single-point geometry energetic target value, information from multiple structures can be combined to generate a representation for an ensemble property, such as the free energy of solvation ($\Delta G_{\text{sol}}$). This has recently been achieved by calculating the ensemble average of the FCHL19 representations of a set of configurational snapshots obtained through MD sampling. Here, we propose an alternative approach that goes beyond standard QML representations (i.e., KRR using one given gas-phase geometry) by describing the chemical transformation occurring during the enantiodetermining step of an asymmetric reaction through the comparison of the representations of the two catalytic cycle intermediates preceding and following the stereocontrolling TS.

This allows us to generate a “reaction-based” representation, which can be closely mapped to the activation energy of the enantiodetermining step, as discussed later. We rely on “dissimilarity” plots as a diagnostic tool to determine whether a particular representation can adequately characterize the stereocontrolling step. By dissimilarity plots, we refer to histograms of the Euclidean distance between any two representations $E_a$ vs. the difference in their target property, which in this case is $E_a$. For a particular representation to be effective, small distances between structures must correspond to small differences between target properties, as the Euclidean distance is used to measure the similarity of two molecular representations. Similar plots have previously been exploited to analyse the behaviour of molecular representations, but only parenthetically. Here, we highlight their importance as a fundamental analytical tool to understand the performance of molecular representations in kernel methods for asymmetric catalysis and demonstrate their utility for constructing reliable ML models.

Before discussing our proposed representation variants, we report in Fig. 1a the performance of the standard SLATM representation using the structure of a single intermediate (e.g., 2). Due to the structural similarities between 2 and the enantiodetermining TS (in both, the Si atom has 6 coordination sites occupied, whereas the coordination number is only 5 or 4 in intermediate 3), intermediate 2 was first chosen to construct the input representation. The learning curve for the prediction of $E_a$ using SLATM (blue) of intermediate 2 (denoted SLATM$_2$) reaches a Mean Absolute Error (MAE) of $0.54 \pm 0.06$ kcal mol$^{-1}$ for the prediction of $E_a$ with 90% of the data used for training (i.e., 680 structures). Considering the exponential relationship between relative activation energies and e.e. values, which implies a dramatic propagation of errors, the accuracy of this approach is insufficient. This is further demonstrated in Fig. 2, which shows the correlation between the predicted and reference $\Delta \Delta E^f$ values (MAE = 0.96 kcal mol$^{-1}$), and in Fig. 3, where the e.e. values obtained from SLATM$_2$ are compared to the
reference quantities: the large number of red-coloured cells indicates large deviations between ML-predicted and DFT-computed e.e. values. The rather poor mapping between SLATM$_2$ and the $E_a$ of the stereocontrolling step (associated with the key 2 $\rightarrow$ 3 transition state) is evident from the visual inspection of Fig. 3, where the large number of red-coloured cells associated with catalysts bearing substituents a, d, e, g, f and j indicates inaccurate predictions of e.e. values, and from the analysis of the corresponding dissimilarity plot in Fig. 1b (left). In the latter, the large scattering of points lying outside the area delimited by the dotted lines, particularly when the Euclidean distance tends to zero, means that two different structures might be considered equal by the kernel (distance $=0$) albeit leading to very different $E_a$ values. Thus, the shape of the dissimilarity plot of SLATM$_2$ deviates considerably from ideal one, indicated by the dotted straight lines. Note that the MAE for $E_a$ increases up to 0.77 $\pm$ 0.05 kcal mol$^{-1}$ (see Fig. S2) if starting from the SLATM representation of 3, the intermediate following the enantiodetermining step in the catalytic cycle (Scheme 2). The higher accuracy achieved using the representation of 2 vs. 3 could be attributed to the reaction step being exergonic and, according to the Hammond Postulate, the enantiodetermining TS resembling 2 more closely. In any case, neither the structure of 2 or 3 provides sufficiently good fingerprints of $E_a$ on their own.

Unlike other intrinsic molecular properties that depend on the structure of a single molecule, enantioselectivity is determined by electronic and/or steric effects stabilising or destabilising one enantiomeric TS to a greater or lesser degree than the other. In that sense, it is to be expected that our target accuracy for $E_a$, well below 1 kcal mol$^{-1}$, cannot be reached using only one structure that does not adequately describe the stereocontrolling transition state as an input. To improve the model performance, an alternative representation is
constructed by comparing the representations of both intermediates. Knowing that neither the structure of 2 or 3 are uniquely related to the corresponding activation energies, we can generate such a “reaction-based” representation that draws information from both structures, subtracting the global SLATM of 2 from 3. This is reminiscent of binary reaction fingerprints (obtained by subtracting the products’ from the reactants’ RDKit fingerprints), which reflect changes in molecular features over reaction processes. The resulting representation (denoted SLATM_diff) accounts for the differences between the two intermediates and is thus more sensitive to the structural changes occurring during the enantio-determining step. By subtracting “reactant” from “product”, the global features that do not change during the catalytic cycle step are eliminated from the representation, and the structural changes between intermediates are highlighted. In this way, we obtain a more faithful representation of the reaction step, which corresponds to a more unique fingerprint of $E_a$. Although the construction of SLATM_diff requires the SLATM representations of both intermediates (2 and 3), the computational cost associated with its generation is negligible.

As depicted in the dissimilarity plot (Fig. 1b, middle), the reaction-based representation (SLATM_diff) is significantly better than SLATM: the difference in $E_a$ values tends to zero as the Euclidean distance between their representations tends to zero. In line with this observation, the learning curve (shown by the orange line in Fig. 1a) is significantly improved. The MAE of SLATM_diff is reduced to 0.31 ± 0.2 kcal mol$^{-1}$, roughly 50% better than SLATM and up to 60% better than that of SLATM using 90% of the data for training (i.e., 680 structures) in the train/test splits of the cross-validation scheme. Given the rationality of the approach leading to the construction of SLATM_diff, its gain in accuracy is encouraging. As shown in Fig. 2 and 3, the halved MAE leads to a very notable improvement in the prediction of e.e. values. Nevertheless, we note again that very small errors in $E_a$ are amplified when e.e. values are calculated, and therefore even a small accuracy gain can be significant.

The high probability density of normalised Euclidean distances between 0.5 and 0.75 seen in Fig. 1b (middle, SLATM_diff) indicates that the shape adopted by the dissimilarity histogram of SLATM_diff is not yet ideal, and that further improvement is possible. To achieve higher accuracy, we focus on improving the shape of this dissimilarity plot. Notice that in our ML model, the Euclidean distance is used as a measure of similarity between representations. This means that features with high variance (i.e., that change the most between molecules) dominate the notion of similarity, as they contribute the most to the Euclidean distance between representations. By feature, we mean each of the terms in the molecular representation, which, for SLATM, consist of two- (London dispersion) and three- (Axilrod–Teller–Muto) body potentials computed on groups of atoms closer than a certain cut-off (here, 4.8 Å). The results of these potentials are averaged over their atom-type sets (e.g., all C–C interactions for the two-body terms, all the C–C–C for the three-body terms), which are then concatenated to generate the SLATM vector. The size of the SLATM representation depends on the existing atom-type sets in the database. Given that our dataset contains the elements C, H, O, N, F, Cl and Si, the total number of features of the SLATM representations is 27,827.

In SLATM_diff, features with high variance dominate the notion of similarity, measured through the Euclidean distance. However, we are using SLATM to predict a property that is very different from the single-molecule properties for which it was originally designed. Consequently, features with high variance in SLATM are not necessarily the most important fingerprints of $E_a$. In pursuit of the best possible fingerprint of the activation energy, we assign importance scores to each feature and attempt to focus on the most relevant ones. The linear correlation coefficient ($r^2$) between each feature and the target property is used as an estimate of the importance of the different terms in the representation. The results, presented in Fig. 4, show that in SLATM_diff there are only a few high-variance features, while the computed importance scores are spread over many other features that have relatively small variances. Simply put, the variances in the features of the SLATM_diff representation are not well correlated with their real importance in this application.

Based on this observation, an improved representation, labelled SLATM_diff++, is generated by selecting only the $N_I$ most important features of SLATM_diff (specifically, $N_I = 500$) and discarding the rest. This feature selection was done using only the training data at each train/test split of the cross-validation step, as otherwise it could lead to severe overfitting. Nevertheless, the importance scores were consistent across the cross-validation splits thanks to the robustness of the linear regressions. An improved relationship between representation and target distances (Fig. 1b, right) is obtained with the SLATM_diff++ representation, in spite of its reduced size. This simple feature selection leads to a noticeable improvement in accuracy, with a cross-validated MAE of 0.25 ± 0.4 kcal mol$^{-1}$ (see the green curve in Fig. 1a). Using the SLATM_diff++ representation, the resulting cross-validated correlation coefficients for the
The hybrid ML model presented herein is able to accurately predict enantioselectivity of the 10 distinct ligand arrangements afforded by 7j and 7k. The out-of-sample results are shown in Fig. 5. Despite scaffold 7 and substituent k not being in the original training set, excellent correlation between predicted and reference $E_a$ values is obtained ($r^2 = 0.97$). Thus, the enantioselectivity of these out-of-sample catalysts is qualitatively reproduced, despite not achieving exact quantitative agreement between DFT and ML predicted $\Delta \Delta E^2$ values (1.2 and 1.3 for 7j and 7k, respectively, vs. 0.2 and 0.5 kcal mol$^{-1}$).

In summary, we provide a logical route to improve atomistic ML methods for enantioselectivity prediction of asymmetric catalytic reactions, which are limited by both the required accuracy and the small amount of data generally available. Firstly, the intermediates associated with the enantioselectivity-determining step (2 and 3 in Scheme 2) must be identified, and their SLATM representations generated. Secondly, using the difference between the two SLATM representations (SLATMDiff) as input, a set of features that map the activation energy accurately can be obtained. Finally, feature engineering can be used to improve SLATMDiff, keeping only the most relevant features that relate to the target property. The results show that the ML workflow presented herein is able to accurately predict enantioselectivity from the molecular structures of catalytic cycle intermediates.

**Conclusions**

In this work, we have developed an atomistic machine learning model to predict the DFT-computed e.e. of Lewis base-catalysed propargylation reactions (Scheme 2). The use of dissimilarity

---

difference between (R)- and (S)-activation energies ($\Delta \Delta E^2$, Fig. 2) in the test set are greatly improved ($r^2 > 0.95$). The quality of our fitted model far supersedes previously reported approaches. Good qualitative and even quantitative agreement is achieved between predicted and reference e.e. values computed using the test data splits from the cross-validation runs (Fig. 3).

Since linear correlation constitutes a very limited notion of relevance, other methods, such as nonlinear mutual information criteria, were tested as feature importance estimators, but the resulting models showed similar or even worse performance (see the ESIT). Similarly, methods based on metric learning did not lead to any improvement, as the high dimensionality of the problem led to severe overfitting. Ceriotti et al. suggested the use of principal covariates regression (PCovR) to solve similar issues. PCovR is a supervised feature selection method that interpolates between principal component analysis (PCA) and linear regression. Herein, because the variance of the features is completely unrelated to the importance scores, the addition of PCA would not offer any advantage. Nevertheless, these findings highlight the importance of adapting molecular representations to the application at hand, while still preserving the overall generality of the approach.

**Chemical insight on asymmetric propargylation catalysts**

The ML model is able to reproduce the main trends in e.e. observed across the different catalysts from the 100 different random train/test splits (Fig. 3, top left table). For example, using SLATMDiff (Fig. 3, bottom right table), which gives the best predictions with respect to the reference data, catalysts built on scaffold 4 (Scheme 1) are revealed to be outliers, yielding e.e.’s that are significantly different to those obtained with other scaffolds, for a given substituent a-j. This is due to the different placement of the substituent X on the organocatalysts’ scaffold. Excluding 4, the effect of different substituents on the e.e. is qualitatively the same across all scaffolds, with the exception of f (Pr) and j (Ph). The introduction of a phenyl group on the organocatalysts’ scaffold leads to highly varied e.e. values, from –97 (S) to 91 (R). This variation, which is due to the presence of favourable π-stacking and CH/π interactions stabilising some (S)-TSs and degrading the enantioselectivity, is nicely captured by SLATMDiff. Overall, the high enantioselectivity displayed by (most) catalysts in the library can be attributed to the favourable electrostatic interaction between the formyl C=H of benzaldehyde and one of the chlorines bound to Si, which is present in the lowest-lying (R)-ligand arrangement, and absent in the (S)-structures.

In their computational screening with AARON, Wheeler and co-workers identified derivatives of 6 as promising candidates for propargylation reactions. However, these catalysts are difficult to synthesize stereoselectively. Recently, Malkov et al. reported the synthesis of a set of terpene-derived atropisomeric bipyridine $N,N’$-dioxides 7 (Fig. 5) as easily-separated diastereoisomers. Aromatically-substituted catalysts 7j and 7k were shown to be highly active and selective (e.e. of 96 and 97, respectively); additionally, the TS structures for 7 were computationally shown to be nearly identical to the corresponding substituted forms of 6. Prompted by these results, we decided to test the ML model with SLATMDiff, to predict the activation energy of the 10 distinct ligand arrangements afforded by 7j and 7k. The out-of-sample results are shown in Fig. 5. Despite scaffold 7 and substituent k not being in the original training set, excellent correlation between predicted and reference $E_a$ values is obtained ($r^2 = 0.97$). Thus, the enantioselectivity of these out-of-sample catalysts is qualitatively reproduced, despite not achieving exact quantitative agreement between DFT and ML predicted $\Delta \Delta E^2$ values (1.2 and 1.3 for 7j and 7k, respectively, vs. 0.2 and 0.5 kcal mol$^{-1}$).

In this work, we have developed an atomistic machine learning model to predict the DFT-computed e.e. of Lewis base-catalysed propargylation reactions (Scheme 2). The use of dissimilarity
plots allowed us to rationally develop and progressively improve a reaction-based representation that can be adequately mapped onto the activation energy of the stereocontrolling step. We identified two fundamental limitations of many standard physics-based molecular representations for subtle catalytic properties. First, we have shown that neither the structure of the preceding nor that of the following catalytic cycle intermediate is a fine fingerprint of the energy of the stereocontrolling transition state. This issue can be circumvented by using a reaction-based molecular representation derived from both structures. Finally, we have demonstrated how feature selection can be used to fine-tune this representation.

The resulting model can accurately predict the DFT-computed enantioselectivity of asymmetric propargylations from the structure of catalytic cycle intermediates. Thus, it constitutes a valuable tool to quickly identify potentially selective propargylation organocatalysts. By design, the model is well-balanced between computational cost, generality and accuracy. It is easy to implement for a wide range of chemical space and seamlessly compatible with experimental (e.g., X-ray structures of stable intermediates) and computational data alike. Our results prove that semi-quantitative predictions of e.e. values in asymmetric catalysis can be achieved by accurately predicting $E_e$.

We conclude that atomistic ML models with adequately tailored molecular representations can be a practical and accurate alternative to both traditional quantum chemical computations of relative rate constants and multivariate linear regression with physical organic molecular descriptors. The stepwise improvement to the model described in this work opens the door to more complex reaction-based and catalytic cycle-based representations. Indeed, ensemble representations, which were recently introduced for properties very sensitive to conformational freedom, such as the free energy of solvation $\Delta G_{\text{sol}}$, are a promising path to go beyond the single structure-to-property paradigm and allow for further generalisation, once combined with the approach discussed herein. Such methodologies will be explored in future work for the accurate screening of enantioselective catalysts in asymmetric reactions.

Author contributions
S. G. performed DFT computations and analyzed the results. R. F. trained and improved the ML models. S. G. and R. F. jointly wrote the manuscript with help from R. L. M. D. W. and R. L. provided feedback on the DFT and ML components, respectively. S. B. and M. D. W. ran preliminary computations initiating this work. All authors discussed the results and commented on the manuscript. C. C. conceived the project with M. D. W., provided supervision and wrote the final version of the manuscript.

Conflicts of interest
There are no conflicts to declare.

Acknowledgements
The authors are grateful to the EPFL for financial support. S. G., R. F. and C. C. acknowledge funding from the European Research Council (ERC, Grant Agreement No. 817977) within the framework of European Union’s H2020. The National Centre of Competence in Research (NCCR) “Sustainable chemical process through catalysis (Catalysis)” of the Swiss National Science Foundation (SNSF) is acknowledged for financial support of R. L. S. B. acknowledges the NCCR “Materials’ Revolution: Computational Design and Discovery of Novel Materials (MARVEL)” for providing a “INSPIRE” master fellowship. The authors thank Benjamin Meyer for helpful discussions at the origin of the project.

References
21 A. R. Rosales, J. Wahlers, E. Limé, R. E. Meadows, K. W. Leslie, R. Savin, F. Bell, E. Hansen, P. Helquist,
Because the original TS database was computed with Gaussian09, the fine (75 302) integration grid (default of Gaussian09) was used instead of the ultrafine (99 590) grid (default of Gaussian16).

References

103. Because the original TS database was computed with Gaussian09, the fine (75 302) integration grid (default of Gaussian09) was used instead of the ultrafine (99 590) grid (default of Gaussian16).


