Describing Chiral Ligands in Palladium-catalyzed Decarboxylative Asymmetric Allylic Allylation: A Critical Comparison of Three Machine Learning Approaches

Declan Galvin,[1‡] Eduardo Aguilar,[2,3,4‡] David Rogers,[2] Simon Woodward,[2,3] Ender Özcan,[4] Patrick J. Guiry,[1] Grazziela Figueredo[5*]*

[1]School of Chemistry, Centre for Synthesis and Chemical Biology, University College Dublin

[2]School of Chemistry, University of Nottingham, University Park, Nottingham, NG7 2RD, United Kingdom.

[3]GSK Carbon Neutral Laboratories for Sustainable Chemistry, University of Nottingham, Jubilee Campus, Triumph Road, Nottingham, NG7 2TU, United Kingdom.

[4]School of Computer Science, University of Nottingham, Jubilee Campus, Nottingham, NG8 1BB, United Kingdom

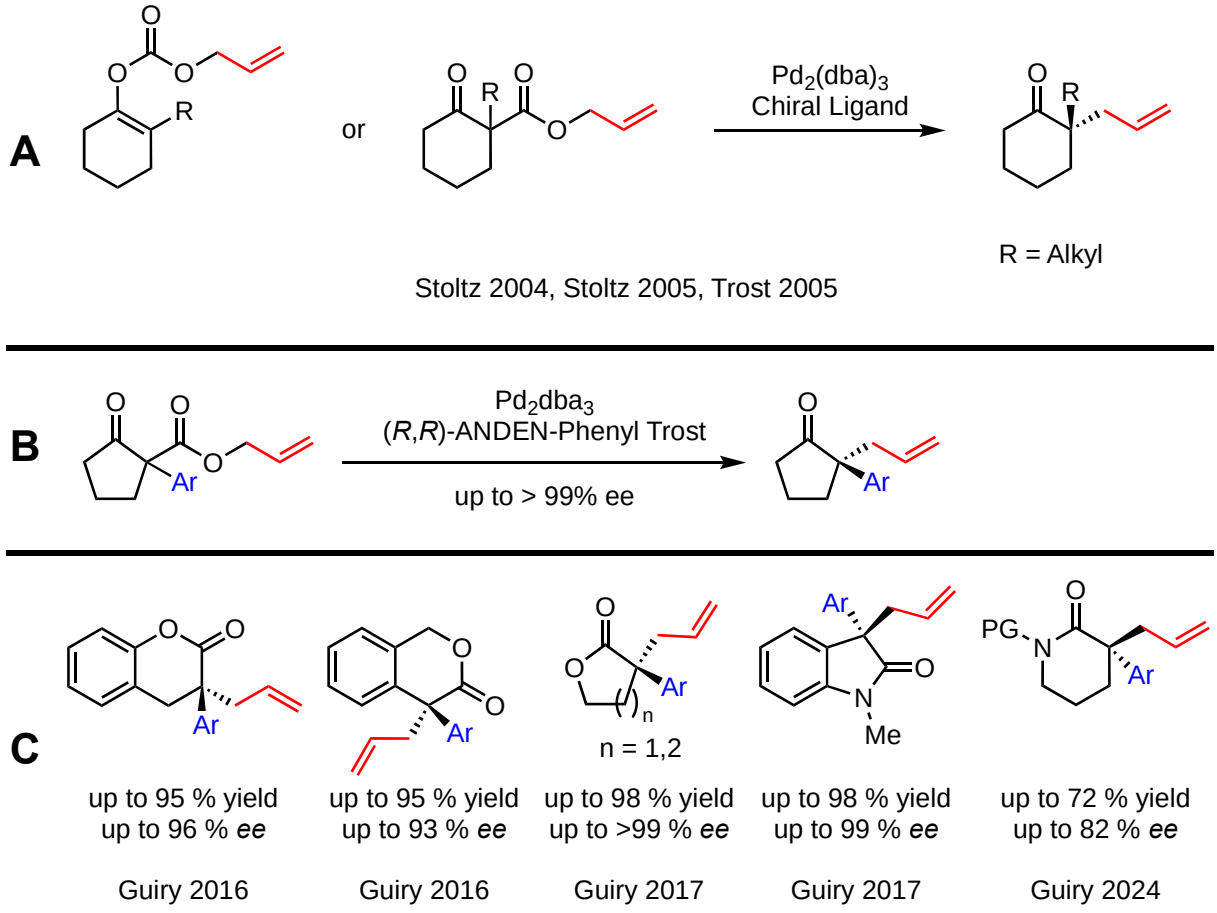
[5]School of Medicine, University of Nottingham, University Park, Nottingham, NG7 2RD, Nottinghamshire, United Kingdom.

‡ These authors contributed equally.

Despite its current popularity, machine learning (ML) applied to asymmetric catalysis remains underexplored. Present strategies include direct use of existing descriptors (e.g. those originally formulated for medicinal chemistry), the development of new bespoke steric and electronic descriptors, or the use of molecular graphs. This method diversity, in the absence of user guidelines, makes selecting an optimal ML algorithm unclear. The fact that asymmetric catalysis data sets are frequently small also make interpretable ML of chiral ligand understanding difficult to realize. Herein, we present an exhaustive evaluation of reaction representations in combination with different machine learning algorithms (including linear regression, random forests, gradient boosting, and graph neural networks) using a realistic-size database compromising 103 palladium-catalyzed decarboxylative asymmetric allylic alkylation (DAAA). This database consists of the combination of three different Trost-type ligands with 54 different substrates. It is concluded that our new bespoke steric and electronic descriptors offer the best performance, while overcoming the problem of interpretability of using existing topo-electronic descriptors, and the problem of data requirements of Graph Neural Networks.

Introduction

Palladium-catalyzed decarboxylative asymmetric allylic alkylation (DAAA, Scheme 1) is a mild, but powerful tool for the synthesis of compounds possessing α-quaternary stereocentres.[1,2] The DAAA reaction has been developed to give high levels of both yields and enantioselectivities using allyl enol carbonate and β-keto allyl ester substrates (Scheme 1A).[3–5] While the substrate class and allyl substitution pattern has been extensively varied, the identity of the α-substituent (R in Scheme 1A) has been much less studied. Traditionally, this is a small alkyl group such as a methyl or ethyl, or a substituted methyl such as a benzyl group. In 2016 Guiry and co-workers expanded the scope of DAAA reactions to include bulky α-aryl cyclopentanones (Scheme 1B) with high yields and excellent enantioselectivities (>99% *ee*).[6] These studies could be generalized to other α-aryl containing related substrates (Scheme 1C). [7–10]



Scheme 1. (A) Traditional α-alkyl DAAA reactions (2004-05). (B) Initial α-aryl DAAA reaction reported by Guiry.[6] (C) Recent motifs accessible by α-aryl DAAA reactions.[7-10]

The catalyst for the DAAA comprises two parts: a Pd(0) source, commonly $Pd_2(dba)_3$, and a suitable chiral ligand. In the case of α-aryl containing substrates the optimized choice of ligand is typically a Trost-type ligand (e.g. Figure 1). The high levels of enantioselectivity seen when using Trost-type ligands is attributed to the H-bonding capabilities of the amide proton.[11] As the enolate approaches the Pd-π-allyl complex, the steric clash between the bulky aryl group and the backbone

of the ligand causes the enolate to orient itself with the aryl group pointing away from the steric clash (Figure 1, pathway A). It has been observed that di-*ortho* aryl substitution and electron-donating groups on the aryl group leads to the highest levels of enantioselectivity. It is proposed that this di-*ortho* substitution pattern breaks the co-planarity of the enolate-aryl group, leading to an unstabilized enolate that reacts more selectively. It should be noted from that the outset that results in the literature result from the use of a range of (*R*, *R*)- and (*S*,*S*)-Trost ligands and this needs to be accounted for when applying any stereochemistry defining (ML) mechanistic model.

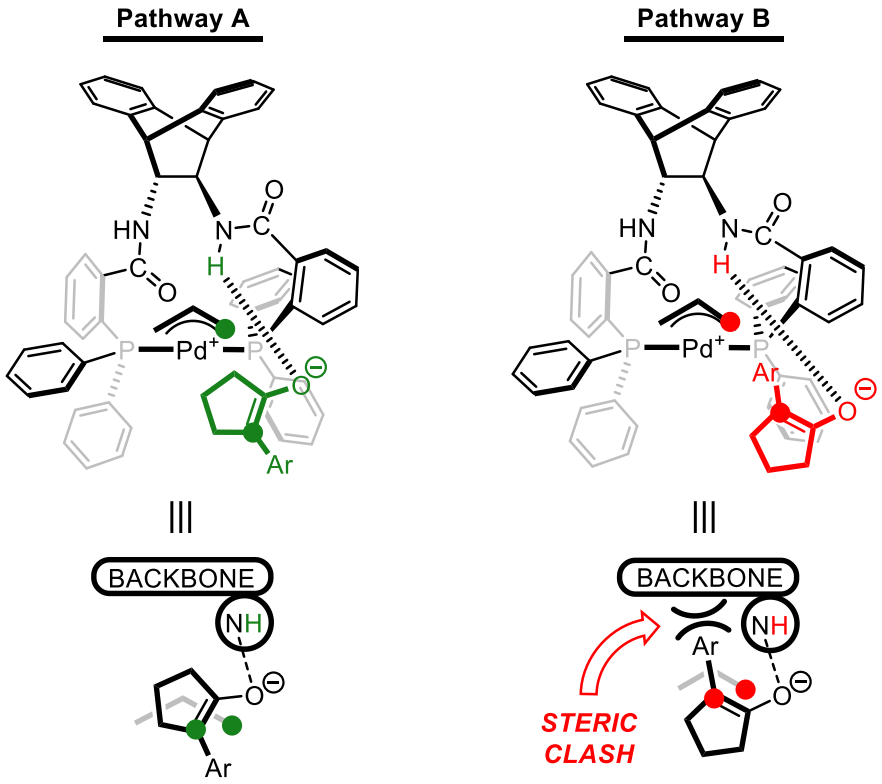


Figure 1. Origin of stereochemical selectivity using a (*R,R*)-ANDEN-phenyl Trost ligand and cyclopentanone substrates shown in Scheme 1C.

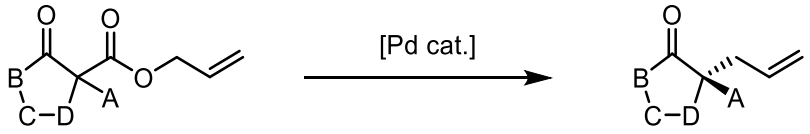
Ligand controlled asymmetric methodologies, such as the DAAA reaction, can be treated as pattern recognition problems. Machine learning (ML) is a useful tool for interpreting and solving such problems. ML has found applications in predicting the bioactivities of new drugs,[12,13] optimizing reaction conditions,[14] predicting chemical yields,[15,16] and in the development of new synthetic routes to desired molecules.[17] ML has also found notable success in the area of asymmetric catalysis.[18–20] Furthermore, the first Graph Neural Network (GNN) capable of understanding relations between a reaction graph representation and stereoselectivity outcome has been recently reported by *Aguilar et al.*[21]

Despite the insights that ML can provide to the area of catalytic chemistry, its uptake by synthetic chemists has been low due, in part, to limited awareness/examples of ML in the field. Herein we compare simple, but effective, methodologies for supervised ML modelling of the DAAA reaction. Our aim was to determine systems where easily understood chemical features show which substituents within the ligand(s) used are likely to lead to selectivity improvement. We chose to

compare, a bespoke steric/electronic feature set against one developed for medicinal chemistry and against our recently developed Graph Neural Network approach (*HCat-GNet*).[21] All these ML models were asked to predict the product enantioselectivities of the same data set of Pd-catalyzed α-aryl DAAA reactions and asked to highlight the areas of the substrate/ligand that are key to the success of the reaction.

Results and Discussion

A total of 103 DAAA reactions of α-aryl containing substrates were gathered from published work from within the Guiry group.[6–10] Initially, we sought a very straightforward approach to featurization of both the substrate and the catalyst to maximize the subsequent human interpretability resulting from the ML outcomes based on our previous work.[22] Each substrate was described using a common core structure broken down into the structural elements A-D (Scheme 2). To minimize the descriptors needed for featurization, the structural units B-D were approximated just by the van der Waals volume of each unit (as calculated by the method of Zhao[23]) and electronically by the Hammett parameter of the nearest published analogue. The latter being obtained from a wide-ranging review (e.g., methyl used for CH_2, ethyl for CH_2CH_2, phenyl for the *o*-phenylenes).[24] Our aim was to compare the utility of this bespoke explainable feature approach to both traditional molecular featurization methods (i.e. via RDKit),[25] and GNN approaches (e.g. *HCat-GNet*).[21]



Scheme 2. A-D fragment approach. For example, the left most structure in Scheme 1C is defined as containing the fragments: A = Ar, B = O, C = *ortho*-C_6H_4, D = CH_2.

The structural unit A denotes the α-aryl substituent, where the electronic and steric of this ring are known to have a significant impact on the enantioselectivity of the DAAA reaction. We chose to represent the bias in the steric demand of this type of fragment by its 'stoutness' (van der Waals volume over longest perpendicular chain length, as proposed by *Owen et al.*)[22] and to represent their electronics by their Hammett parameter. The 14 aryl fragments (**a-m**) used in our study are shown in **Figure 2**. While the Hammett parameter[24] of 'a' is +0.00 by definition (that derived from benzoic acid), only two other substitution patterns have experimentally derived Hammett parameters (fragments 'b + c'). Therefore, the 10 remaining examples were calculated or approximated by other methods.

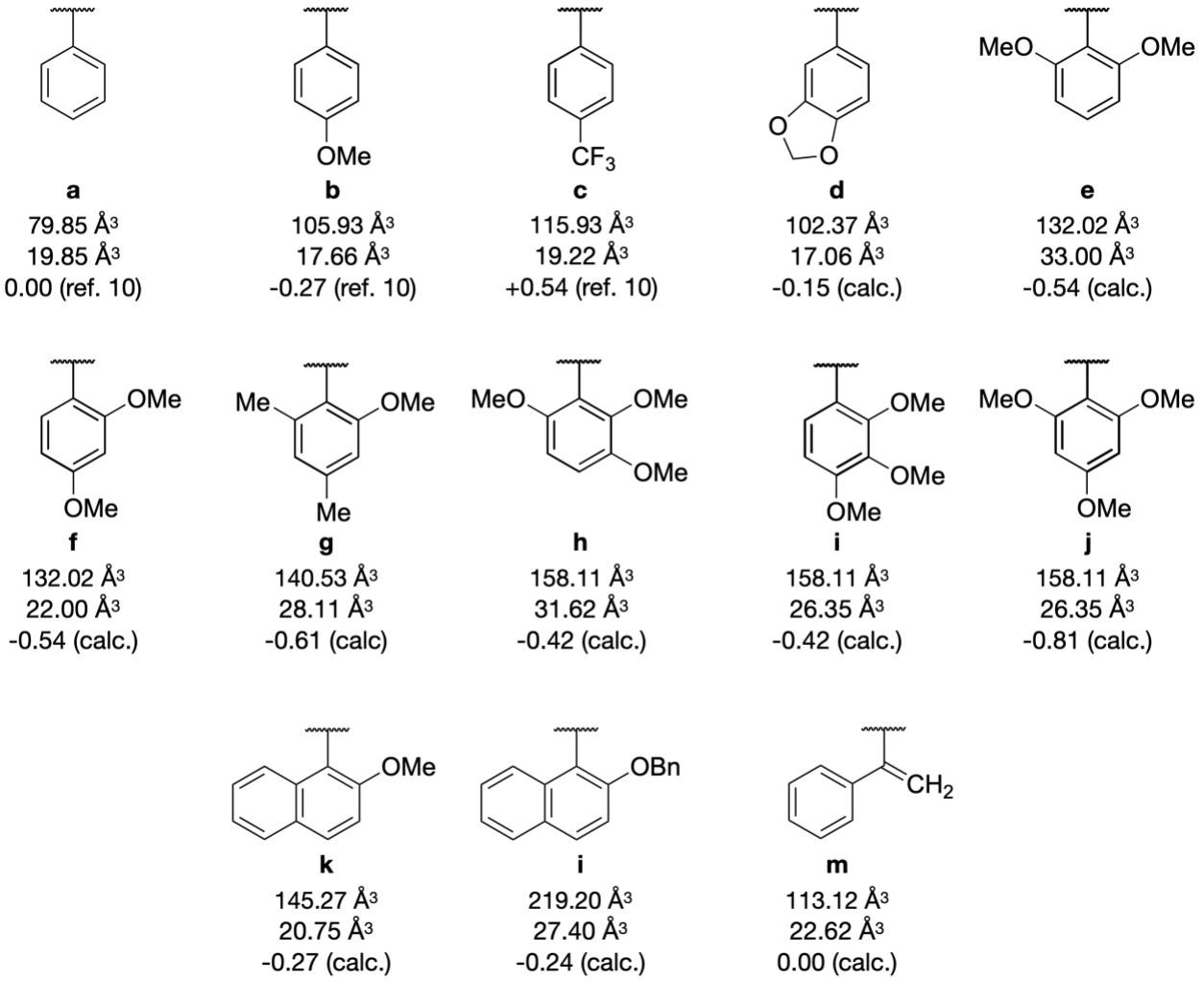


Figure 2. Structures of subunits 'A' present in our dataset. The calculated[9] fragment volume used is given in Å^3 first, then the 'stoutness' (in Å), and finally the actual, or calculated using the method **V2** (see text for details) Hammett parameter used.

Which enantiomer of the DAAA product is formed is controlled by the absolute chirality (*R*,*R*) vs. (*S*,*S*) of the Trost ligand used. For the compiled curated dataset of DAAA transformations, three different catalysts are pertinent, namely complexes of Pd with the three Trost-type ligands (cat **I-III**), **Figure 3**. When this work was being undertaken, crystallographic data for these palladium complexes was sparse and their solution behavior is complicated by association phenomena.[26] Due the absence of usable mononuclear crystallographic data at the time, catalyst fragments **I-III** were subjected to repeated MM2 force field dynamics in Chem3D until they converged to their lowest global strain energy. The obtained structures were checked against proposed and generally accepted transition-state models and each catalyst structure agreed with what would be predicted. In 2023, Arseniyadis and Leitch reported an excellent paper on the preparation of chiral, air stable, and reliable Pd(0) precatalysts for use in asymmetric allylic alkylation chemistry, such as DAAA.[27] As part of their work, they reported Pd(0) complexes with Trost-type ligands. Significantly for our work, their paper contained X-ray crystallographic data that was previously not available. We were pleased to find that the structures we obtained after MM2 force field dynamics showed significant

similarities to the crystallographic data in their report, suggesting that we had a good starting point for breaking down our three catalysts for featurization.

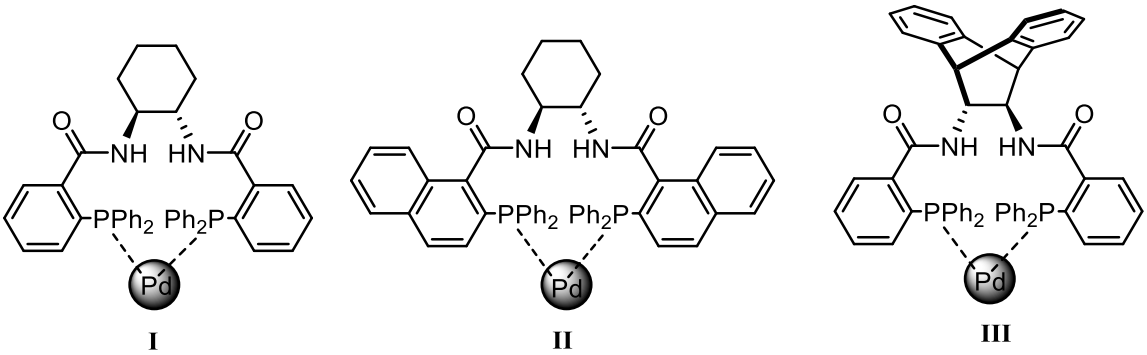


Figure 3. Structure of the three Trost ligands present in our dataset.

The three minimized conformations, **Figure 4**, show the structures divided into a quadrant model, giving the four sub-units UL, LL, UR and LR (where L = Left, R = Right, U = Upper, L = Lower), that are all forward of an arbitrary P-P baseline. Inspecting the MM2 minimized structures reveals the Fragments shown in **Figure 4** to be projecting into the four regions for each of catalysts **I-III**. For example, catalyst I shows a 'HC=CH-CH=CH' fragment is in the LR quadrant. By using Zhao's[23] approach these can approximated into an equivalent total steric burden without the need for complex molecular dynamics to be employed. This is shown in **Figure 4** for all catalyst families. Due to low exemplar numbers we only featurized the steric profile of the chiral space around catalysts **I-III** and did not include any electronic description, as these Trost-type ligands all likely to have similar electronic properties.

The standard Cahn-Ingold-Prelog (CIP) method for stereoisomer assignment of *R* or *S* is not self-consistent for the library of resultant products from the DAAA reaction, as this assignment is dependent on the identity of the structural units A-D and the added allyl fragment. Therefore, the facial selectivity driven by the Trost ligands would not be accounted for. The enantioselectivity target of this ML analysis was therefore defined as '%topA' or the percentage of the 'top' isomer, as proposed and used elsewhere.[21,22] This is defined as placing the α-carbonyl to the left of the site of C-allylation and with the aryl unit A pointing up or 'on top'. Therefore a 50 % *ee* would have a 75% or 25% '%topA', depending on which face the aryl unit is situated on.

The %topA in the dataset we are working with is somewhat imbalanced as we mostly see very high levels of enantioselectivity in the DAAA products. This leads to a bimodal distribution in the dataset. Generally, this would lead to a categorisation approach being favoured over regression. However, a longer-term aim of this ML workflow is *de novo* ligand design. Therefore, the regression approach was maintained here. However, we also present the results of the models on predicting the face of the addition for a more insightful study on the capability of these approaches to understand the reaction representation and predict stereoselectivity outcomes. This way, we also target the variable 'face of addition', which determines whether the addition is done on the 'top' side (%top > 50%), or not (%top < 50%).

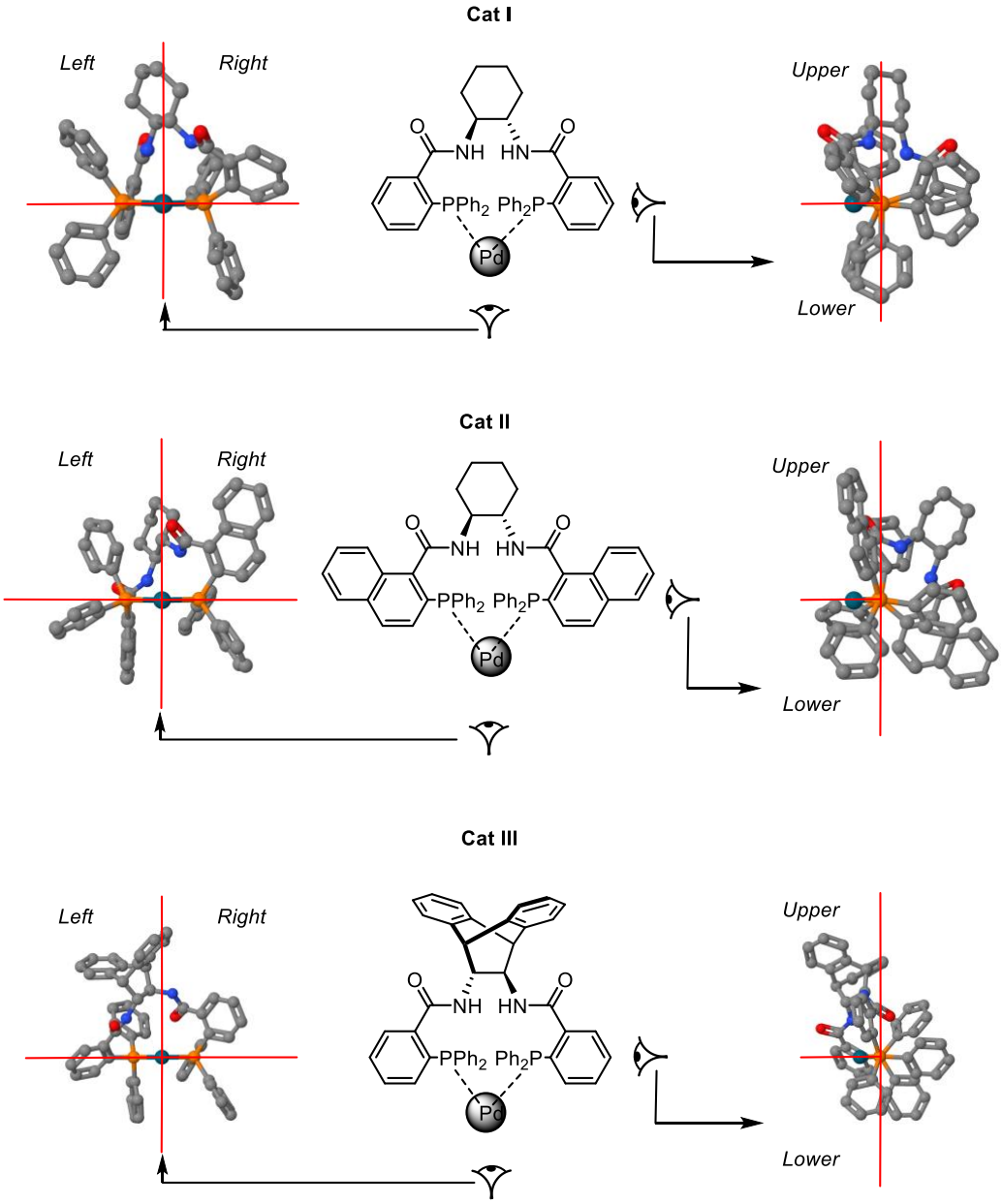


Figure 4. Quadrant model applied to Chem3D optimized ligand structures to generate steric features for the ML model input.

To analyse the selected descriptors, a correlation matrix was generated from the above bespoke features is generated (**Figure 5**). This heatmap allows for visualisation of the relationship between features. Isolation of the %topA row (boxed in **Figure 5**) highlights which descriptors are likely to be impactful in the ML process. For example, the correlation for %topA is strongly positive for UL and LR volumes of the catalyst and strongly negative for LL and UR volumes. These correlations are expected, as the volumes of these quadrants align with either the (*S*,*S*)- or (*R*,*R*)-hand of the Trost ligands, which ultimately control whether the products of the DAAA are high or low %topA. However, other descriptors such as electronics and sterics of certain fragments and solvent effects did show strong correlation with the %topA outcome.

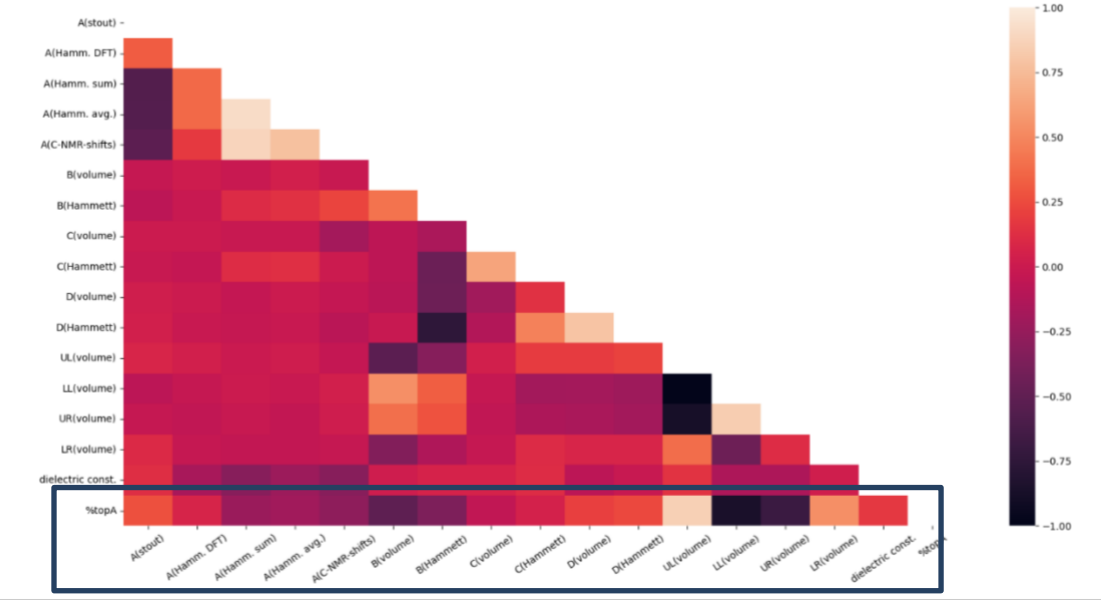


Figure 5. Heatmap feature correlation to induced stereoselectivity (%top). Lighter colors indicate strong positive effects to %top, darker colors the reverse.

The reactions were represented into four different versions (**V1-V4**), Table 1. Each version contained identical entries, except for the values of the electronics of the aryl unit 'A', which was defined by different methods. **V1** contains values obtained from the DFT method developed by Ziegler, **V2** contains values defined by the summed Hammett value method, **V3** contains results by the averaged Hammett value method, and **V4** contains values defined from the ^{13}C NMR spectroscopic method. As the number of exemplars in the collated dataset was relatively small, it was decided to look at the dataset as a whole and not break it down into smaller subsets for analysis. The four representations **V1-V4** were tested under a variety of ML algorithms. Simple **Linear Regression** (**LR**), which attempts to model the relationship between two variables by fitting a linear equation to the data, was initially employed. Ensemble methods such as **Gradient Boosted Regression** (**GBR**) and **Random Forrest Regression** (**RFR**) were also examined. These methods combine the outputs of several decision trees looking at subsets of the data into a single result. This reduces variance in the results.

The performance of the ML algorithms outlined above were assessed using a nested-cross validation approach. This means that we created ten-folds, which led to a total of ten test sets, each being evaluated by nine different training and validation sets. This generated a total of 90 training-testing processes. This allows assessment of robustness and stability of the different methods evaluated. For evaluation of the prediction of the '%topA' variable, we report the RMSE (%), while for the face of addition, accuracy is used. The finalized performance of each of the ML models (**LR**, **RFR** and **GBR**) using each of the different electronic descriptors evaluated herein are given in Table 1 for all the testing points gathered from the 90 training-testing processes, leading to a total of 927 testing reactions.

Table 1. Overall ML algorithm predictive performance of the enantioselectivity variable '%topA' and 'face of addition' compiling all the testing data-points from the 90 train-testing trials.

Data Set	**GBR** *RMSE (%)* *Accuracy*	**LR** *RMSE (%)* *Accuracy*	**RFR** *RMSE (%)* *Accuracy*
V1	10.17 **0.997**	14.18 0.940	13.33 0.961
V2	10.06 0.986	14.14 0.940	**9.87** 0.967
V3	11.58 0.996	14.24 0.940	12.50 0.961
V4	11.66 0.989	13.89 0.940	12.05 0.961

The best prediction of '%topA' was obtained using the **RFR** algorithm using the **V2** electronic descriptor, with a RMSE value of 9.87%. This low RMSE value shows that the ML algorithm was able to successfully predict selectivity for unseen substrates using easily interpretable chemical features. Results using the **GBR** algorithm obtained similarly low RMSE values, while the **LR** algorithm performs the weakest of the three tested algorithms. The representation **V2** attained the lowest RMSE values, although the other descriptors performed quite similarly. The small difference between the results from the electronic descriptors **V1-V4** suggests that the methods used to estimate the electronics of the aryl unit were consistent to each other.

The best prediction of the enantioface of the substrate attacked by the allyl is from the **GB** algorithm, with a precision of 0.997. We have noticed the classification algorithms make no major distinction between electronic descriptors. The reason of this is that all the electronic descriptors have very high correlation between them, which in a classification algorithm, makes little impact on the final training and testing of the model.

To confirm our new proposed methodology is effective compared to classical RDKit featurization processes, we have evaluated the performance of such ML models trained to predict on the same variable but using descriptors calculated from the SMILES string of substrate, ligand, and solvent using RDKit. In the case of the ligands, since the whole dataset only has 3 different ligands, we have decided to calculate descriptors related to lipophilicity, polarity, topology, electronics, and electro-topology, which are calculated from the chemical structure (operations on the molecular graph). In the case of the substrates, we have decided to use fragment-count based descriptors since

the chemical diversity of these molecules will allow the ML algorithms to learn the contribution of the fragments in terms of electronics and topology separately, rather than calculating a descriptor based on the whole molecule, that might or not capture the effect of specific functional groups. Lastly, for the solvent the Hall-Kier Alpha descriptor was calculated. A hyperparameter optimization process for Random Forest and Gradient Boosting was performed and then a Feature Selection Process was applied. The results obtained are shown in Table 4.

Table 2. Overall ML algorithm predictive performance of the enantioselectivity variable '%topA' and 'face of addition' compiling all the testing data-points from the 90 train-testing trials.

Metric	GBR	LR	RFR
RMSE / %	**11.967**	13.981	13.472
Accuracy	**0.990**	0.947	0.961

In the case of the results obtained using the RDKit descriptors, GBR is the best performing algorithm for both the regression and classification tasks. The RFR and LR, although perform worse than the GBR, both attain satisfactory results. This means that the RDKit descriptors used do correlate to the selectivity outcome of the reaction to the ligand.

Finally and GNN (HCat-GNet) approach has also been evaluated. To do this, we took the architecture proposed by *Aguilar et al.*, while the graph representation has been slightly changed.[21] The reason for this is that our dataset is significantly smaller than that used in the original study (about 85% smaller), which makes the learning of the Graph Neural Network significantly more challenging and more likely to overfit on the training samples. To avoid this, we have used the most minimal graph representation that allows correlation between the reaction representation and the selectivity outcome. This way, our graph representation consists of the concatenation of the graph representation of the three participant molecules (substrate, ligand, and solvent). The graph representation of the molecules consisted of a graph where the nodes represent the atoms, while edges represent covalent bonds. We further used node features of atom identity and chirality of the atom, while no edge features were used. By using this simple representation and the original architecture, *HCat-GNet* attained an RMSE of 20.046% and 0.941 accuracy.

As our objective is to compare these three approaches (manual chemical-informed features, RDKit features and *HCat-GNet*) in performance and interpretability for ligand optimization, we compared the best performing models using the different descriptors, which are **RFR** using handcrafted features with the **V2** electronic descriptor, and the **GBR** using RDKit features. The three methods have been trained and evaluated using the exact same sets of reactions, which means that the results obtained per test fold can be compared directly. We present the results of the regression task using the RMSE (%) and of the classification task using the accuracy per test fold in a bar plot, where the value of the bar represents the mean and the error bars the standard deviation of the population of metrics obtained in each of the 9 times that each test set was used. We also show a strip plot to visualize the distribution of errors generated by each method. These results are shown in **Figure 6**.

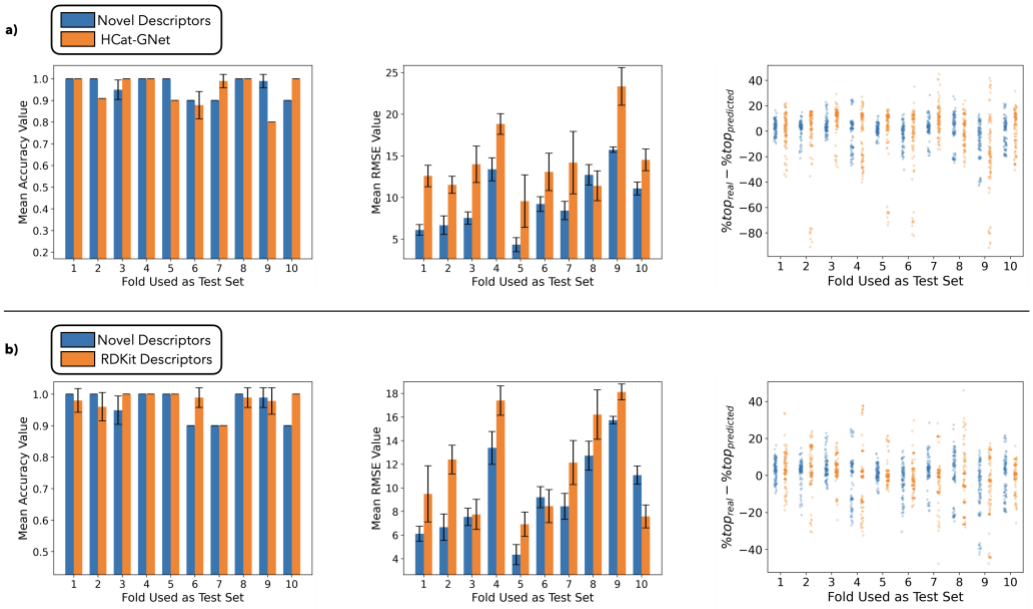


Figure 6. Summary of results obtained by the novel descriptors, *HCat-GNet*, and the RDKit descriptors on prediction of selectivity of the DAAA reaction. A) Shows the comparison of results between the novel approach and *HCat-GNet*. B) Shows the comparison of results between the novel approach and the RDKit descriptors.

Figure 6a shows the comparison between the handcrafted features proposed herein and *HCat-GNet.* In general, the performance of *HCat-GNet* was the poorest of all the methods evaluated. However, the results per test fold show that HCat-*GNet* can attain similar results than those obtained by the other methods. In the case of Accuracy, *HCat-GNet* is less accurate in folds 2, 5, 9, while in the other folds the performance is undistinguishable or better than the novel approach. On the other hand, the RMSE values show that *HCat-GNet* delivers less accurate predictions for most folds, where only folds 6, and 8 show similar errors between methods. From the strip plot, it can be noticed that *HCat-GNet* generated higher errors, particularly on folds 2, 5, 6, and 9 where the error was more than 50% (meaning that the prediction of facial selectivity was incorrect). We hypothesize that the poor performance *HCat-GNet* is due to the requirement for larger quantities of data in deep learning methods to perform well compared to RFR and GBR. Although the performance is not as good as the handcrafted features, it is remarkable that despite the size of the database, HCat-GNet was able to accurately predict with significant accuracy some of the data points and obtain comparable results to the other approaches for certain folds, while maintaining the advantage of not requiring human input to generate high level features. This can overcome modelling problems in asymmetric catalysis, particularly when transition states are completely unknown, and the design of handcrafted features is more challenging.

When comparing the handcrafted and the (medicinal chemistry derived) RDKit features in **Figure 6b**, no major differences are found in the accuracy when prediction the face of addition. In the case of RMSE, in general, the novel features provide lower errors, where only folds 3, 6, and 10 show similarities between methods. Also, no major differences were found in the distribution of errors in the strip plot. The results obtained show that there is not much difference between using the handcrafted features and the RDKit features. However, the latter makes use of descriptors that, for the most part, lack of meaning to human chemists, while the handcrafted features are highly

interpretable and meaningful. This small difference is of importance, as *de novo* design can be inspired from the interpretation that can be potentially applied to the model, which ultimately will explain the effect of each feature in the model's final prediction. To demonstrate this, we have used SHAP analysis to understand the impact of each variable to the model's final prediction to the RFR and GBR models using the RDKit descriptors and to the RFR using **V2** handcrafted features. The results are shown in **Figure 7**.

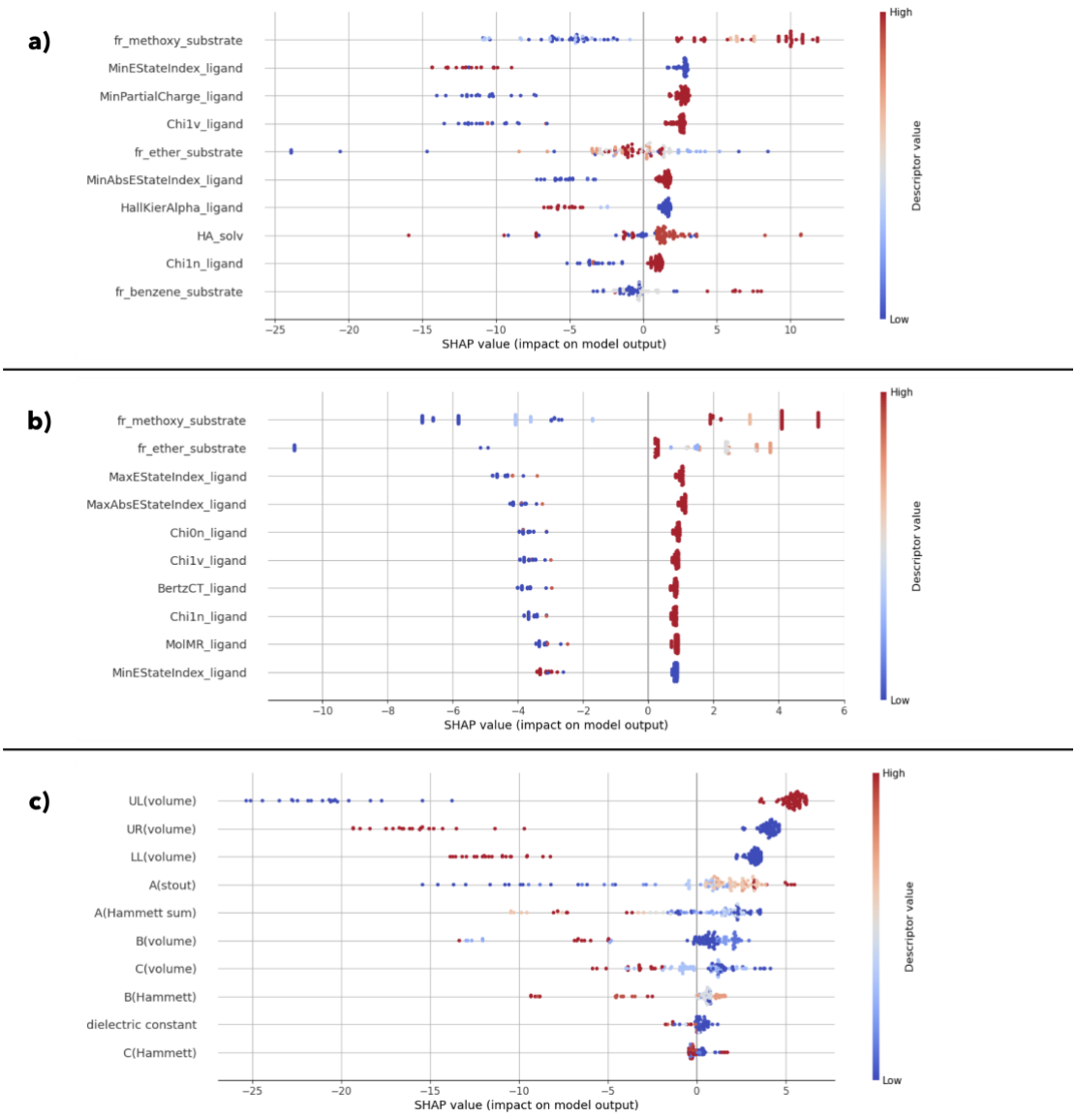


Figure 7. SHAP analysis applied to the models trained using the RDKit and the handcrafted features. A) Shows the SHAP analysis of the model **GBR** using the RDKit descriptors. B) Shows the SHAP analysis of the model **RFR** using the RDKit descriptors. C) Shows the SHAP analysis of the model **RFR** using the handcrafted descriptors and the **V2** electronic descriptor.

Figure 7a shows the SHAP analysis applied to the **GBR** model using RDKit descriptors. Although the plot shows clearly how low or high values of each feature affect the models' prediction, some of these features (for example min energy state index ligand, Chi1v ligand, min absolute energy state index ligand) are not interpretable to scientist, which prevents human design in further chiral ligand optimization. Similar effects occur in the **RFR** model using RDKit features, shown in **Figure 7b**. For this model, more features were used, but there is still a lack of chemical interpretability, as the case of Min and Max Energy State Index, Chi0n, and BertzCT (all describing the ligand). Although in both these models there are variables that are human interpretable, these consist of fragment counts in the substrate, which is not useful as usually the substrate to convert is not a variable of the reaction, where ligand optimization is the pre-eminent goal.

Alternatively, **Figure 7c** shows the SHAP analysis of the **RFR** model using the handcrafted features with the **V2** electronic descriptor. Our proposed new features also show a clear trend in how they impact the models' final prediction. For the case of the UL volume of the ligand, high values in general increase the model's prediction (larger %topA or prediction of addition on the top side), while UR volume and LL volume decrease the model prediction (lower %topA or prediction of addition on the bottom side) when the values are bigger. This result agrees with the proposed selectivity model in **Figure 1**, which means that the model is effectively learning the relation between the steric of the ligand and the selectivity outcome. Unlike the RDKit descriptors, these ligand descriptors are highly interpretable and ultimately can help chemists to design *de novo* ligands to optimize this type of conversions. Other properties are found to impact the stereochemical outcome of the reaction, however, these properties depend on the substrate, which is not a molecule to be optimized.

HCat-GNet also allows interpretability by applying SHAP analysis to each node feature vector. We applied this tool as it was implemented originally to analyze a subset of three reactions that shared same substrate and solvent, and the only difference between them was the ligand used. The results are shown in **Figure 8**.

From **Figure 1** and the analysis presented in **Figure 7c**, it is expected that the interpretations from SHAP in HCat-GNet are related to the disposition of atoms in space resulting from the chiral carbons in the structure. For the case of ligand (*R,R*)-**L3**, the addition is done in the top face, while for (*S,S*)-**L1** and (*S,S*)-**L2** is done in the bottom face. Remarkably, the SHAP analysis demonstrates that HCat-GNet is effectively learning that the main drivers of the facial selectivity in the reaction are the asymmetric carbons in the structure. As shown in **Figure 8a**, the asymmetric carbons in the structure are found to contribute positively to the outcome variable (larger %topA), while from **Figure 8b** and **8c** the asymmetric carbons are found to contribute negatively (lower %topA). This is reassuring, as it indicates that the model is being able to learn the correlation between the configuration of the ligands and the selectivity outcome, and that the predictions done make sense from a chemical perspective. Due to the limited diversity of ligands in our dataset, no further information, such as the impact of substituents in the structure of the ligand, is available.

When comparing the three approaches, the novel proposed featurization process outperforms the RDKit featurization and HCat-GNet approaches for this sparse data set. In the case of HCat-GNet, the biggest limitation found is its poor performance due to the lack of data, while for the RDKit descriptors, although attain high accuracies, are harder to interpret, which makes *de novo* ligand design from them more challenging. Our novel features are demonstrated to competently represent

the reaction and are highly interpretable. This means that the models are not only useful to predict selectivity of new reactions, but also the conclusions from SHAP analysis can be used to ultimately optimize a ligand structure to obtain more efficient transformations.

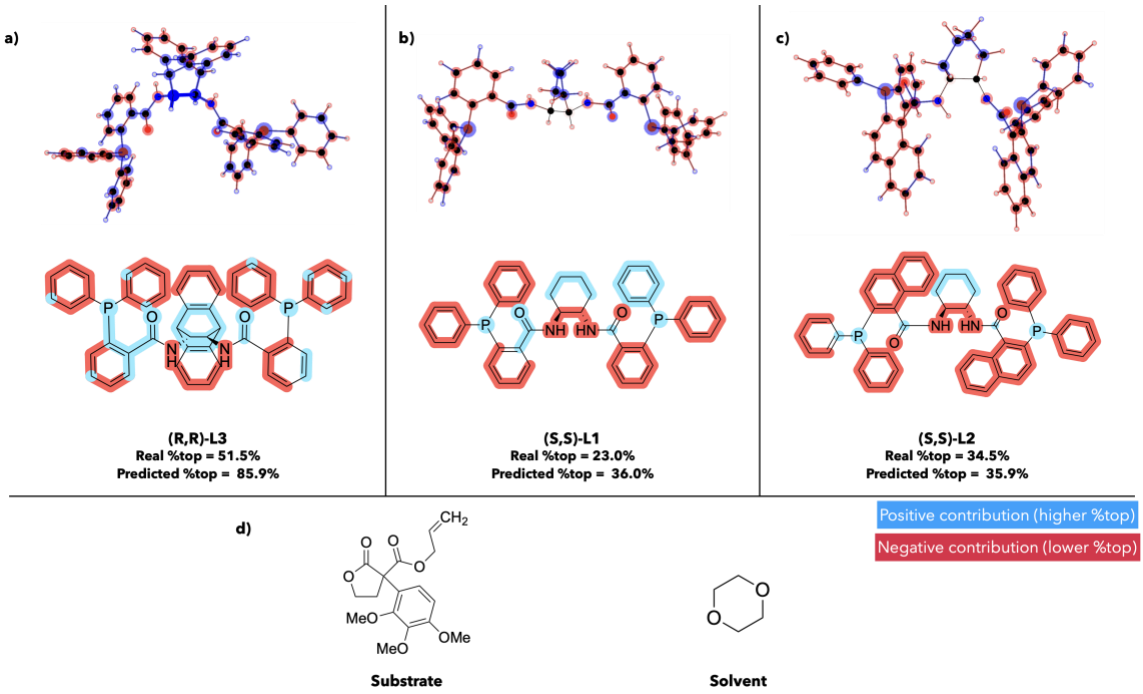


Figure 8. SHAP analysis applied to the ligand structure used in HCat-GNet for three reactions that only differed in the ligand used. a) Shows the SHAP analysis for ligand (*R*,*R*)-**L3**, b) for (*S*,*S*)-**L1**, and c) for (*S*,*S*)-**L2**. d) Shows the structure of the substrate and solvent.

Conclusions

We proposed a novel feature generation procedure for the representation of DAAA reactions and the training of ML models. This representation is based on simple molecular mechanics calculations of optimized ligand structures, where simple descriptors of steric and electronic factors are used. We combined this reaction representation with different ML algorithms to identify the combination that generated the most accurate predictions. To benchmark our proposal, other two competitor ML methods were compared: RDKit descriptors and the GNN *HCat-GNet*. We found that the RDKit descriptors perform almost as well as our novel descriptors, while we found that *HCat-GNet,* although giving satisfactory results, it performed the poorest of the three methods. SHAP analysis was performed to understand the impact of the variables to the models' final predictions. We found that the RDKit descriptors generated a distinct separation in variables that clearly show the direction of impact of each variable. However, these descriptors were less understandable and therefore less useful for *de novo* design. For our novel features, a clear

separation of variables showed the direction of impact to the final prediction, but maintaining high interpretability, which can be used by chemists to ultimately create novel ligands to obtain more efficient transformations. We analyzed the interpretability of *HCat-GNet* by applying SHAP analysis in terms of its node feature vectors. We found that the models were effectively learning the relation between the configuration of the ligands and the selectivity outcome they induced, which demonstrated the ability of this method to understand to a certain level the relation between reactants and selectivity outcome. Our proposed methodology has shown to be effective to model the reaction, while overcoming limitations of former methods including interpretability and data requirements.

Corresponding Authors Contact Information

Patrick Guiry - School of Chemistry, Centre for Synthesis and Chemical Biology, University College Dublin. Email: patrick.guiry@ucd.ie

Grazziela Figueredo - School of Medicine, University of Nottingham, University Park, Nottingham, NG7 2RD, Nottinghamshire, United Kingdom. Email: grazziela.figueredo@nottingham.ac.uk

Author Contributions

D.G. created and calculated the novel features. E. A. carried out the experiments using graph neural networks. D.G. and E.A. jointly performed the ML experiments and wrote the original draft. P.G., G.F., E.Ö., and S.W. jointly supervised, directed, and commented the research. The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Acknowledgements

We acknowledge the AI CDT of the School of Computer Science of the University of Nottingham for funds for this research. D. G. is grateful for the financial support from BiOrbic for his PhD programme "Atoms 2 Products", a CDT supported by Science Foundation Ireland (SFI) and the Engineering and Physical Sciences Research Council (EPSRC) under Grant No. 18/EPSRC-CDT/3582. BiOrbic is funded by Ireland's European Structural and Investment Programmes, Science Foundation Ireland (16/RC/3889) and the European Regional Development Fund.

ABBREVIATIONS

GNN, Graph Neural Network; ML, Machine Learning; DAAA, decarboxylative asymmetric allylic alkylation.

REFERENCES

(1) James, J.; Jackson, M.; Guiry, P. J. Palladium-Catalyzed Decarboxylative Asymmetric Allylic Alkylation: Development, Mechanistic Understanding and Recent Advances. *Adv. Synth. Cat.* **2019**, *363* (13), 3016-3049. https://doi.org/10.1002/adsc.201801575.

(2) Pàmies, O.; Margalef, J.; Cañellas, S.; James, J.; Judge, E.; Guiry, P. J.; Moberg, C.; Bäckvall, J.-E.; Pfaltz, A.; Pericàs, M. A.; Diéguez, M. Recent Advances in Enantioselective Pd-Catalyzed Allylic Substitution: From Design to Applications. *Chem. Rev.* **2021**, *121* (8), 4373-4505. https://doi.org/10.1021/acs.chemrev.0c00736.

(3) Behenna, D. C.; Stoltz, B. M. The Enantioselective Tsuji Allylation. *J. Am. Chem. Soc.* **2004**, *126* (46), 15044–15045. https://doi.org/10.1021/ja044812x.
(4) Mohr, J. T.; Behenna, D. C.; Harned, A. M.; Stoltz, B. M. Deracemization of Quaternary Stereocenters by Pd-Catalyzed Enantioconvergent Decarboxylative Allylation of Racemic β-Ketoesters. *Angew. Chem. Int. Ed.* **2005**, *44* (42), 6924–6927. https://doi.org/10.1002/anie.200502018.
(5) Trost, B. M.; Xu, J. Regio- and Enantioselective Pd-Catalyzed Allylic Alkylation of Ketones through Allyl Enol Carbonates. *J. Am. Chem. Soc.* **2005**, *127* (9), 2846–2847. https://doi.org/10.1021/ja043472c.
(6) Akula, R.; Doran, R.; Guiry, P. J. Highly Enantioselective Formation of α-Allyl-α-Arylcyclopentanones via Pd-Catalysed Decarboxylative Asymmetric Allylic Alkylation. *Chem. Eur. J.* **2016**, *22* (29), 9938–9942. https://doi.org/10.1002/chem.201602250.
(7) Akula, R.; Guiry, P. J. Enantioselective Synthesis of α-Allyl-α-Aryldihydrocoumarins and 3-Isochromanones via Pd-Catalyzed Decarboxylative Asymmetric Allylic Alkylation. *Org. Lett.* **2016**, *18* (21), 5472–5475. https://doi.org/10.1021/acs.orglett.6b02584.
(8) James, J.; Guiry, P. J. Highly Enantioselective Construction of Sterically Hindered α-Allyl-α-Aryl Lactones via Palladium-Catalyzed Decarboxylative Asymmetric Allylic Alkylation. *ACS Catal.* **2017**, *7* (2), 1397–1402. https://doi.org/10.1021/acscatal.6b03355.
(9) Jackson, M.; O'Broin, C. Q.; Müller-Bunz, H.; Guiry, P. J. Enantioselective Synthesis of Sterically Hindered α-Allyl-α-Aryl Oxindoles: Via Palladium-Catalysed Decarboxylative Asymmetric Allylic Alkylation. *Org. Biomol. Chem.* **2017**, *15* (38), 8166–8178. https://doi.org/10.1039/c7ob02161e.
(10) Galvin, D.; Guiry, P. Enantioselective Synthesis of Sterically Hindered α-Allyl-α-Aryl Lactams via Palladium-Catalysed Decarboxylative Asymmetric Allylic Alkylation. *Eur. J. Org. Chem.* **2024**, e202400314. https://doi.org/10.1002/ejoc.202400314.
(11) Butts, C. P.; Filali, E.; Lloyd-Jones, G. C.; Norrby, P.-O.; Sale, D. A.; Schramm, Y. Structure-Based Rationale for Selectivity in the Asymmetric Allylic Alkylation of Cycloalkenyl Esters Employing the Trost 'Standard Ligand' (TSL): Isolation, Analysis and Alkylation of the Monomeric Form of the Cationic η3 -Cyclohexenyl Complex. *J. Am. Chem. Soc.* **2009**, *131* (29), 9945–9957. https://doi.org/10.1021/ja8099757.
(12) Rohall, S. L.; Auch, L.; Gable, J.; Gora, J.; Jansen, J.; Lu, Y.; Martin, E.; Pancost-Heidebrecht, M.; Shirley, B.; Stiefl, N.; Lindvall, M. An Artificial Intelligence Approach to Proactively Inspire Drug Discovery with Recommendations. *J. Med. Chem* **2020**, *63*, 40. https://doi.org/10.1021/acs.jmedchem.9b02130.
(13) Muratov, E. N.; Bajorath, J.; Sheridan, R. P.; Tetko, I. V.; Filimonov, D.; Poroikov, V.; Oprea, T. I.; Baskin, I. I.; Varnek, A.; Roitberg, A.; Isayev, O.; Curtalolo, S.; Fourches, D.; Cohen, Y.; Aspuru-Guzik, A.; Winkler, D. A.; Agrafiotis, D.; Cherkasov, A.; Tropsha, A. QSAR without Borders. *Chem. Soc. Rev.* **2020**, *49* (11), 3525–3564. https://doi.org/10.1039/D0CS00098A.
(14) Strieth-Kalthoff, F.; Sandfort, F.; Segler, M. H. S.; Glorius, F. Machine Learning the Ropes: Principles, Applications and Directions in Synthetic Chemistry. *Chem. Soc. Rev.* **2020**, *49* (17), 6154–6168. https://doi.org/10.1039/C9CS00786E.
(15) Davies, J. C.; Pattison, D.; Hirst, J. D. Machine Learning for Yield Prediction for Chemical Reactions Using in Situ Sensors. *J. Mol. Graph. Model.* **2023**, *118*, 108356. https://doi.org/10.1016/J.JMGM.2022.108356.

(16) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting Reaction Performance in C–N Cross-Coupling Using Machine Learning. *Science* **2018**, *360* (6385), 186–190. https://doi.org/10.1126/SCIENCE.AAR5169/SUPPL_FILE/AAR5169-AHENMAN-SM_REVISION_1.PDF.
(17) Struble, T. J.; Alvarez, J. C.; Brown, S. P.; Chytil, M.; Cisar, J.; Desjarlais, R. L.; Engkvist, O.; Frank, S. A.; Greve, D. R.; Griffin, D. J.; Hou, X.; Johannes, J. W.; Kreatsoulas, C.; Lahue, B.; Mathea, M.; Mogk, G.; Nicolaou, C. A.; Palmer, A. D.; Price, D. J.; Robinson, R. I.; Salentin, S.; Xing, L.; Jaakkola, T.; Green, W. H.; Barzilay, R.; Coley, C. W.; Jensen, K. F. Current and Future Roles of Artificial Intelligence in Medicinal Chemistry Synthesis. *J. Med. Chem.* **2020**, *63* (16), 8667–8682. https://doi.org/10.1021/ACS.JMEDCHEM.9B02120.
(18) Tsuji, N.; Sidorov, P.; Zhu, C.; Nagata, Y.; Gimadiev, T.; Varnek, A.; List, B. Predicting Highly Enantioselective Catalysts Using Tunable Fragment Descriptors. *Angew. Chem. Int. Ed.* **2023**, *62* (11), e202218659. https://doi.org/10.1002/ANIE.202218659.
(19) Hoque, A.; Sunoj, R. B. Deep Learning for Enantioselectivity Predictions in Catalytic Asymmetric β-C–H Bond Activation Reactions. *Digit. Discov.* **2022**, *1* (6), 926–940. https://doi.org/10.1039/D2DD00084A.
(20) Hirst, J. D.; Boobier, S.; Coughlan, J.; Streets, J.; Jacob, P. L.; Pugh, O.; Özcan, E.; Woodward, S. ML Meets MLn: Machine Learning in Ligand Promoted Homogeneous Catalysis. *Artif. Intell. Chem.* **2023**, *1* (2), 100006. https://doi.org/10.1016/j.aichem.2023.100006.
(21) Aguilar Bejarano, E. A.; Figueredo, G.; Woodward, S.; Lam, H. W.; Özcan, E.; Rit, R. HCat-GNet: An Interpretable Graph Neural Network for Catalysis Optimization. February 22, 2024. https://doi.org/10.26434/chemrxiv-2024-zjnkd.
(22) Owen, B.; Wheelhouse, K.; Figueredo, G.; Özcan, E.; Woodward, S. Machine Learnt Patterns in Rhodium-Catalysed Asymmetric Michael Addition Using Chiral Diene Ligands. *Results Chem.* **2022**, *4*, 100379. https://doi.org/10.1016/j.rechem.2022.100379.
(23) Zhao, Y. H.; Abraham, M. H.; Zissimos, A. M. Fast Calculation of van Der Waals Volume as a Sum of Atomic and Bond Contributions and Its Application to Drug Compounds. *J. Org. Chem.* **2003**, *68* (19), 7368–7373. https://doi.org/10.1021/JO034808O/SUPPL_FILE/JO034808OSI20030611_100954.XLS.
(24) Hansch, C.; Leo, A.; Taft, R. W. A Survey of Hammett Substituent Constants and Resonance and Field Parameters. *Chem. Rev.* **1991**, *91* (2), 165–195. https://doi.org/10.1021/CR00002A004/ASSET/CR00002A004.FP.PNG_V03.
(25) *The RDKit Documentation — The RDKit 2022.03.1 documentation*. https://www.rdkit.org/docs/ (accessed 2022-05-30).
(26) Eastoe, J.; Fairlamb, I. J. S.; Fernández-Hernández, J. M.; Filali, E.; Jeffery, J. C.; Lloyd-Jones, G. C.; Martorell, A.; Meadowcroft, A.; Norrby, P.-O.; Riis-Johannessen, T.; Sale, D. A.; Tomlin, P. M. Interrogation of a Dynamic Multi-Catalyst Ensemble in Asymmetric Catalysis. *Faraday Discuss.* **2010**, *145* (0), 27–47. https://doi.org/10.1039/B910022A.
(27) Huang, J.; Keenan, T.; Richard, F.; Lu, J.; Jenny, S. E.; Jean, A.; Arseniyadis, S.; Leitch, D. C. Chiral, Air Stable, and Reliable Pd(0) Precatalysts Applicable to Asymmetric Allylic Alkylation Chemistry. *Nat. Commun.* **2023**, *14* (1), 8058. https://doi.org/10.1038/s41467-023-43512-8.

(28) Ziegler, B. E.; McMahon, T. B. Computational Analysis of Substituent Effects and Hammett Constants for the Ionization of Gas Phase Acids. *Comput. Theor. Chem.* **2013**, *1008*, 46–51. https://doi.org/10.1016/J.COMPTC.2012.12.015.