HCat-GNet: An Interpretable Graph Neural Network for Catalysis Optimization

Eduardo Aguilar Bejarano,[a,b,c] Ender Özcan,[c] Raja K. Rit,[a,b] Hon Wai Lam,[a,b] Simon Woodward [a,b] and Grazziela Figueredo*[d]

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

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

[c] *School of Computer Science, University of Nottingham, Jubilee Campus, Wollaton Road, Nottingham, Nottingham, NG8 1BB, United Kingdom.*

[d] *School of Medicine, University of Nottingham, Medical School, Nottingham, NG7 2UH, United Kingdom. E-mail: grazziela.figueredo@nottingham.ac.uk*

Homogeneous catalysts enable faster conversions of molecules with higher selectivities (stereo- and regioselectivity) in chemical reactions. Traditionally, catalyst improvements are made through empirical trials, where the catalyst is functionalised by adding, removing or modifying groups within its structure and, subsequently, reevaluating the new catalytic activity. This procedure is not efficient and leads to unsuccessful trials that waste resources. Machine learning (ML) approaches have been proposed to accelerate homogeneous asymmetric catalyst optimization. However, these often lack a general descriptor generation procedure to allow encoding of molecules from a broad region of chemical space. To overcome this, we propose a homogeneous catalyst graph neural network (HCat-GNet) for the prediction of selectivity of catalysts given the SMILES of participant molecules. We demonstrate its use in rhodium-catalyzed asymmetric 1,4-addition (RhCAA), a reaction of major importance in organic synthesis. We benchmark HCat-GNet against traditional ML methods for its ability to predict RhCAA stereoselectivity from two chiral diene ligand two datasets; one for learning and one for final testing. For the learning dataset, both traditional ML and HCat-GNet methods give comparable results. However, when presented with the new unseen test dataset, traditional ML models perform poorly, while HCat-GNet retains a general ability to accurately predict product absolute stereochemistry and reaction stereoselectivity. Furthermore, HCat-GNet allows model interpretability, permitting analysis of the effect of ligand substituents in determining reaction selectivity. HCat-GNet shows greater potential for catalyst optimization than traditional ML, as it allows the use of a non-fixed number of participant molecules to train the model, only requiring the SMILES of the molecules to create graph representations. HCat-GNet allows more general models that accurately extrapolate into unseen regions of chemical space.

Introduction

Catalysts are widely used in industry, driving around 85% of industrial chemical reactions.[1] While only 15% of these processes involve homogeneous catalysts, they typically allow higher selectivity compared with heterogeneous catalysts.[1,2] Examples of industrially important process that use homogeneous catalysts include hydroformylation (for detergents),[3] Suzuki-Miyaura cross-coupling (for pharmaceuticals),[4] and Ziegler-Natta reactions (for polymers).[5] Adding a ligand to a homogeneous metal-catalyzed reaction often increases its rate and selectivity, by an effect known as ligand-accelerated catalysis (LAC), which allows the discovery of novel catalysts for more efficient transformations and improved synthetic routes to target compounds.[6]

The process of homogeneous catalyst optimization from initial conception to final industrial application can take years. The reason for this is that classical catalyst development is typically done by changing one variable in the system (such as the core structure of a catalyst or its ligands) and re-evaluating the new catalytic activity, which requires arduous synthetic work with no guarantee of a positive outcome.[7,8]

Machine learning (ML) has gained popularity in chemical science as it allows quantitative structure-property relationship (QSAR) studies in an efficient, relatively inexpensive, yet accurate way.[9–11] In these approaches, variables that depend on molecular structures, known as descriptors, are used to represent the molecule computationally.[12] Dos Passos Gomes *et al.* highlighted the paucity of ML studies on asymmetric homogeneous catalysts and the need for more research in this area,[13] while Hirst *et al.* and Kalikadien *et al.* encourage the use of these techniques for novel catalyst discovery.[14,15]

Some research papers have been published in the area of machine learning for homogeneous catalysts;[16–21] however, just a handful study asymmetric catalysis. An example of application of ML for asymmetric catalyst optimization is the strategy of Bretholomé *et al.*,[22] where the stereoselectivity of a Michael addition to a benzylideneacetone substrate

was optimized. A lead ligand was identified and other ligands that only differed in one substituent were generated. A linear regression between the selectivity of the reaction and the calculated *logP* of the changed substituent was created, which allowed successful identification of a ligand that maximized reaction selectivity. Although this approach was successful, the explored variables were limited to only one substituent within the structure of the catalysts, thus, limiting the variable optimization space.

In a second example, Zahrt *et al*. proposed a general strategy to generate descriptors of participant molecules in an catalytic asymmetric reaction.[23] This strategy is reaction agnostic, allowing encoding of any participant molecule in any asymmetric reaction accelerated by homogeneous catalysts. However, their methodology does not allow interpretation, which in ML is desired. Their descriptors are not rotational nor translational invariant, which ultimately can lead to wrong predictions if a molecule is not well-oriented in 3D space.

In another study, Owen *et al*. proposed a descriptor generation procedure for rhodium-catalyzed asymmetric 1,4-additions (RhCAA) (**Fig. 1a**) using bridged bicyclic chiral 1,4 diene ligands (**Fig. 1b**).[24] Inspired by the empirical Hayashi selectivity model (**Fig. 1c**),[25] steric and electronic descriptors of substituents in the core structure of the substrate, chiral diene ligand, and organoboron reagent were used (**Fig. 1d**). Although their approach was simple and accurate for standard chiral diene ligands, it only allows encoding of molecules that share a core structure with those of the molecules in the training data, thus limiting its applicability to structurally different ligands.

Other approaches are also found in the literature.[26–28] However, these methods have common shortcomings, including:

1. Most methods work for a very limited chemical space, where the exploration space is limited to only one set of ligands, substrate, or reagents that are all very similar in structure.[24,26–28]
2. The ML algorithm is trained to only explore one variable at the time, such as just one ligand substituent.[22,26]
3. Some methods lack interpretability, which is desired in machine learning to understand which chemical patterns the algorithm correlates with selectivity outcomes, to inspire chemists in the *de novo* design of catalysts.[17,23]

Considering these points, it is highly desirable to develop a universal, easy-to-use, and highly interpretable method that allows accurate prediction of the stereochemical outcomes of chiral catalysts, which can potentially be applied to any reaction.

Graph neural networks (GNNs) are a deep learning approach that has gained popularity in quantitative structure-activity relationship (QSAR) studies.[29–32] These algorithms are able to discover high-level descriptors within molecular graphical representations of data to make chemical predictions.[33,34] Advances in this area enabled the availability of code to transform digital representations of molecules (.cif files, SMILES strings, .vasp files) into graphs, allowing the training of models directly from simple input data. Methods to explain GNN predictions have also been used in chemistry,[35,36] allowing understanding of those molecular fragments that contribute the most to the sought outcome.[37–40]

Garrison *et al*. demonstrated the ability of GNN algorithms to predict DFT-calculated ground state energies of transition metal complexes.[41] We concluded that GNN algorithms are suitable to identify both chemical patterns and interactions between catalysts, reagents, and substrates that correlate with reaction selectivity. To our knowledge, no one has tested this idea in asymmetric catalysis. The ability of GNNs to generate high-level features that correlate with a property automatically has potential to generate descriptors that are more general than those in prior ML strategies. GNN used in combination with explanation algorithms would overcome the limitations of current ML methods.

Herein, we describe the Homogeneous-Catalyst Graph-Network (HCat-GNet), the first interpretable GNN able to predict both the absolute stereochemistry of the product (*R/S*) and the enantiomeric ratio (er) of asymmetric reactions involving homogeneous chiral catalysts given only SMILES representations of the molecules involved. We tested HCat-GNet in a case study of rhodium-catalyzed asymmetric 1,4-additions (RhCAA) of organoboron reagents to prochiral Michael acceptors (**Fig. 1a**).[25,42–44] We compared our approach to the methods described by Owen *et al.*,[24] using three different traditional ML (TML) algorithms: linear regression, random forest, and gradient boosting (see **Fig. 2a**). RhCAA was selected because: 1) many examples have been described in the literature,[42-44,] which leads to a richer dataset, 2)

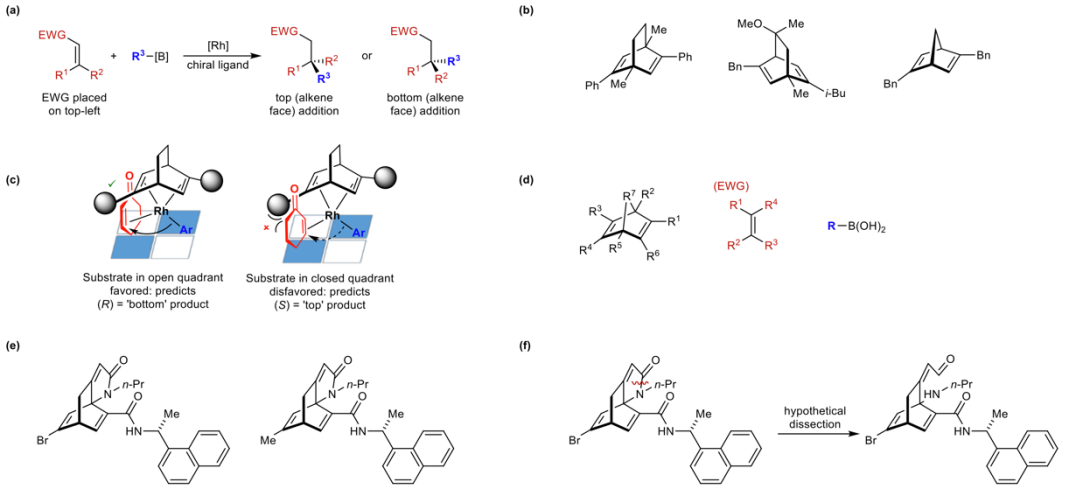

Fig. 1 The RhCAA reaction and representative chiral ligands. (a) General RhCAA, where the structure of the chiral ligand determines the absolute stereochemistry of the product. (b) Representative bridged bicyclic chiral 1,4-diene ligands. (c) The steric selectivity model proposed by Hayashi.[25] (d) Substituent nomenclature used by Owen et al. in their descriptor generation procedure.[24] (e) Examples of the new (genuinely unseen) chiral ligands of Rit et al.[49] (f) The methodology used here to 'cut' the Rit ligands allowing the feature generation procedure proposed by Owen et al.

pre-existing TML models are available to use as benchmarks,[24] and 3) a historical simple steric stereochemical model for the process existed (**Fig. 1c**),[25] allowing comparison of our GNN approach against a human-derived explanation. With our experiments, we aimed to answer three research questions: 1) are GNNs suitable for predicting the stereoselectivity of asymmetric reactions accelerated by ligated homogeneous catalysts? If so, 2) how do the GNN models perform in predicting the stereoselectivity of genuinely unseen ligands (structurally distinct from the training data) compared to TML approaches? 3) Can explainability algorithms be applied to the GNN models to obtain information about chemical variables that can increase or decrease the stereoselectivity?

Methods

HCat-GNet graph representation

HCat-GNet is designed to predict the (stereo) selectivity of reactions accelerated by homogeneous metal-ligand complexes. To be able to encode *any* reaction with *any* number of reacting molecules, we propose creating a graph representation of each molecule (this can be a solvent, ligand, substrate, or reagent) in the reaction and further to create a reaction-level (containing all reactants) graph representation.

In the first step, the algorithm iterates over all the atoms in one of the participating molecules. For each atom, the algorithm retains information on which other atoms share a bond with it, as well as its atomic properties, including its identity, degree (number of non-hydrogen atoms are attached to it), hybridization, whether or not it is part of an aromatic system, whether or not it is part of a ring, and the absolute configuration of the stereocenter at that atom (either *R*, *S* or none). Connectivity information is transcribed into an adjacency list, and atomic property information is retained in a node feature matrix, one-hot-encoded. The adjacency list along with the node feature matrix are the graph representation of the molecule. This is done automatically for all participant molecules, so that each molecule has a graph representation associated.

When all graphs have been generated, the algorithm concatenates the molecular graphical representations into a single disconnected graph, so that a reaction graph consists of *n* graphs (where *n* is the number of participant molecules) that are disconnected between them. The node feature matrix is concatenated as well, and this along with the adjacency list, is the reaction-graph representation. This algorithm is ideal for asymmetric homogeneous catalysis reaction graph representation, as it allows it to include *n* different molecules in the model that correlate with a given reaction outcome.

For our case study, the three molecules in the system (ligand, substrate, and organoboron reagent) are automatically converted into molecular graphs from their SMILES strings. Initially, the algorithm takes the SMILES representation of the ligand and creates a

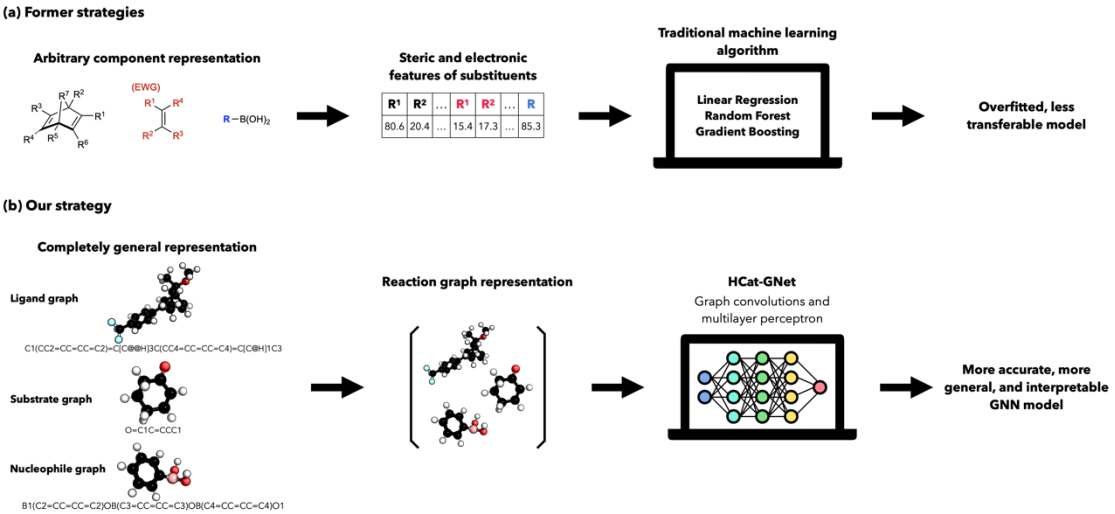

Fig. 2 Comparison of strategies analysed in this study for prediction of stereoselectivity of RhCAA reactions. (a) Shows the approach of Owen *et al.*[24] (b) Shows the HCat-GNet Strategy described herein.

graph (see above). Next, it does the same for the substrate, and then for the organoboron reagent. We added a reaction-level feature to all three graphs) named as "configuration", which identifies whether the chiral ligand has an *(R,R)* or *(S,S)* configuration (**Fig. 1b**). We do not use the edge features, as node level representation encompasses information about the electronic effects connecting neighbouring atoms. Lastly, the algorithm created the reaction-graph representation by concatenating the three molecular graph representation (see **Fig. 2b**).

HCat-GNet architecture

HCat-GNet consists of two phases: (i) message passing and (ii) readout. The first phase is a node-level operation block, which explores the topology of the graph to capture the complex relations between neighbouring nodes. This operation is known as convolution. We have used the Graph Convolutional Operator,[45] inspired from the usual convolutional operator applied to images, which can also be generalised to the graph structures where each node can have different numbers of neighbours. This operator is defined as shown in Eq. 1, where i is the central node that needs to be updated, $h_i^{(l)}$ is the node i current state, $j \in N(i)$ represents all the neighbours j of i, $\hat{d}_x$ represents the number of neighbours of node x, $W^{(l)}$ is a learnable matrix, and $h_i^{(l+1)}$ is the new state of the central node i. This process is run in parallel for all nodes within the system, so that each node receives new node features after each convolution, and it can run as many times as the programmer states.

$$h^{(l+1)} = \sum_{j \in N(i)} \frac{1}{\sqrt{\hat{d}_i \hat{d}_j}} h_j^{(l)} W^{(l)} \qquad [1]$$

The second phase is the predictive phase occurring at the graph level. The first step is to summarise the information from all the nodes contained within the graph into a single graph-level feature vector. This is achieved by an operation called pooling. This operation is usually an element-wise operation that runs for all the node's feature vectors contained in the system. This allows encoding of molecules within a database that have different numbers of atoms. Each molecule (ligand, substrate, and organoboron reagent) has a different number of nodes, and consequently, a different number of total features. Pooling ensures that all the systems within the database attain a single fixed size vector representing it. Finally, this latter vector serves as an input into a multilayer perceptron (MLP) that outputs a final prediction. This way, HCat-GNet steps from graph input to prediction are:

- The node features are taken (length of 24) and these are expanded to a final length of 64 by a graph convolutional operator with leaky ReLU activation function, $R^{nodes \times 24} \rightarrow R^{nodes \times 64}$.
- The graph convolution operator updates all the node states in parallel, updating the nodes features to graph aware features once with leaky ReLU activation function, $R^{nodes \times 64} \rightarrow R^{nodes \times 64}$.
- Mean and max pooling is applied elementwise to all the node feature vectors to get a graph-level feature vector, $R^{nodes \times 64} \rightarrow R^{1 \times 128}$.
- A fully connected layer with leaky ReLU activation function takes the graph-level

vector and maps it to half of its length, $R^{1\times128} \rightarrow R^{1\times64}$.

- A last fully connected layer with sigmoid activation function transforms the feature vector into a scalar number, this being the prediction of the model, $R^{1\times64} \rightarrow R^{1}$.
- The prediction is multiplied by 100 to give a percentage.

Interpretability of HCat-GNet

Explainable Artificial Intelligence (XAI) algorithms can be applied to HCat-GNet to get insights into those chemical patterns that contribute the most in making a catalyst more or less selective. The open-source package PyTorch Geometric allows for the implementation of numerous explainable approaches for graph neural networks. We use GNNExplainer,[46] implemented in PyTorch Geometric 2.4.0, to understand the relevance of each node feature within the graph representation of the reaction and Shapley Value Sampling[47,48] (implemented in PyTorch Geometric 2.4.0) to understand the contribution of each node (atom) within the graph to the selectivity of the reaction.

GNNExplainer identifies subgraphs and subsets of the node features that are more influential for the model's prediction.[46] The mathematical definition of this approach is given in Eq. 2, where MI quantifies the change in the probability of prediction $\hat{y} = \Theta(G_c, X_c)$ when the computation graph is limited to the subgraph G_s and node features limited to X_s, and H is the entropy (uncertainty) of the model's prediction Y. Effectively, GNNExplainer aims to 'denoise' the computation graph and remove edges, nodes, and node features that add to the uncertainty of the prediction. We used this algorithm to explore which node-level properties mostly drive the model to predict a certain outcome in each molecule separately (ligand, substrate, and organoboron reagent).

$$\max MI\big(Y, (G_s, A_s)\big) = H(Y) - H(Y|G_s, A_s) \quad [2]$$

Shapley Value Sampling is a perturbation method that computes an attribution score for the node features within the graph using cooperative game theory. This method takes a random permutation of the input graph node features and sequentially adds the real value to the given feature baseline. The output difference after adding each feature corresponds to its attribution.[47,48] This importance approximation method provides understanding of positive or negative impact of nodes to the prediction of the GNN. Thus, the most important fragments within the molecules can be identified with this method.

The Rh-CAA case study

Datasets. We tested our approach using two RhCAA datasets: classical chiral diene ligands (which we will refer to as the seen dataset or learning dataset), and new Rit *et al.*[49] chiral diene ligands (which we will refer to as the unseen dataset or final testing dataset). The seen dataset was taken from directly from *Owen et al.*[24] and a data augmentation strategy was applied. The unseen database was manually built from the new exemplars reported by Rit *et al.*[49] The ligands in the unseen dataset differ structurally from those in the seen dataset (**Fig. 1b** vs **Fig. 1e**). With the seen dataset, we aimed to answer whether or not GNNs can be used for asymmetric catalysis prediction problems, while the unseen dataset would answer how do GNNs performs predicting on truly unseen ligand exemplars, by predicting stereoselectivity of those reactions when only training with samples in the seen dataset.

Data labelling. As the Cahn-Ingold-Prelog (CIP) *R/S* nomenclature depends on the identity of the R^1/R^2 groups within the substrate (**Fig. 1a**), we use the '%top' target variable adopted by Owen *et al.*, which consists of the percentage of the addition of the nucleophile to the 'top' face of the substrate, as defined by the face of the substrate seen when placing the electron-withdrawing group (EWG) on the top left corner of the alkene, as shown in **Fig. 1a**.[24]

Prediction task. We aimed to predict two target variables: face of addition ('top' or 'bottom') and %top. The former is a classification problem, where '1' represents those reactions where the major product is the 'top' isomer (%top>50%), and '0' otherwise. To convert the continuous regression predictions of the ML models into discrete values, predicted values greater than 50% are considered as prediction of 'top' isomer, and 'bottom' otherwise. For the %top regression task, the predicted values were taken as delivered by the algorithm.

The benchmark descriptors. We used the Owen *et al.* descriptors to train the TML algorithms.[24] These consist of electronic and steric parameters of substituents in the different positions of ligands, substrate, and organoboron reagent (see **Fig.** 1d and **Fig.** 2a).[24,50,51] This molecule encoding process leads to a total of 19 descriptors. For the unseen dataset, the ligand structure had to be adapted to be compatible with this descriptor

generation process. As the R^2 and R^7 substituents are joint by a lactam bond, we cut this bond, which leads to a representation as an aldehyde and amine structure (**Fig.** 1f). Although this does not represent the real structure of the original molecule, it enables matching with the Owen *et al.* procedure.

Model training and evaluation. To evaluate the transferability and robustness of both methods, we used a stratified nested cross validation approach, as proposed by García *et al.*[52] for the 'seen' dataset. We created ten folds, which led to a total of ten test sets, each one being evaluated by nine different training and validation sets. This raised to a total of 90 training-testing processes. The folds used in the TML and HCat-GNet are the same, meaning that both methods were trained and evaluated using the same sets of points. For the TML approach, linear regression, random forest, and gradient boosting are used as ML algorithms. For the face of addition task, we evaluate the models using the metrics of Precision, Recall, and Accuracy. In the case of the %top task, we report the metrics of mean absolute error (MAE), root mean squared error (RMSE) and determination coefficient (R^2) for those datapoints that the face of addition was predicted correctly, as otherwise those datapoints would significantly increase the mean errors, not representing the real performance of the model in the predictive task. We present the results based on the mean of the metrics for each test set separately, and standard deviation as error bars. For the case of prediction selectivity of the 'unseen' dataset reactions, we took the 90 models trained with the 'seen' dataset and, with no further training, predicted their selectivity.

Results and discussion

Gradient boosting regression attained the best results in the classical 'seen' dataset, and therefore we report the results of this algorithm for both datasets.

HCat-GNet applicability to homogeneous catalysis analysis

Fig. 3a shows the results for predicting the face of addition. In the case of precision, test folds 1, 2, 3, 4, 6, and 7 show both methods have scores within the same range of values. HCat-GNet shows superior performance for test folds 5, 8, and 10 while only for fold 9 the Gradient Boosting performs better.

In the case of recall, only test folds 3 and 10 show differences between the methods, with HCat-GNet showing better results. Lastly, in accuracy, only fold 10 shows differences, with Gradient Boosting showing worse results.

Although slight differences were identified between the methods, the scores obtained for both methods are above 0.90 for all metrics. This demonstrates that both methods are suitable for predicting the face of addition of a reaction.

For the TML method, the chirality of the ligand was implicitly encoded by the order given to the substituents within the chiral ligand (**Fig. 1d**). This differentiation makes possible the mapping of the relative positions of substituents (overall ligand stereoconfiguration) to the absolute stereoconfiguration of the product. For our HCat-GNet, we have explicitly encoded the configuration of the ligand more generally by both atom stereoconfiguration and ligand configuration variables. The results herein show the chirality descriptors in our method are excellent predictors of the absolute stereoconfiguration of the product, allowing accuracies ranging from 0.92 up to 1.00.

Fig. 3b shows the results obtained for the '%top' task. In the case of mean absolute error, only folds 3 and 6 show a difference, with HCat-GNet obtaining larger errors. For the RMSE and R^2, no differences were noticed between the methods, which indicate that both methods are suitable for the predictive task.

The violin plot in **Fig. 3c** shows that the distribution of errors generated by both methods is very similar. The strip plot in **Fig. 3c** shows that the error values of both methods occur in similar locations. These two last plots confirm that both the TML and HCat-GNet methods generate very similar predictions of the '%top' variable.

The results show that the performance of both methods is almost indistinguishable, meaning that either can be used for the predictive task. This is a positive result for HCat-GNet, as the features generated by Owen *et al.* were specifically designed for this reaction. For HCat-GNet, the encoding was simply automatically obtained from the SMILES strings of the participating molecules. All high-level

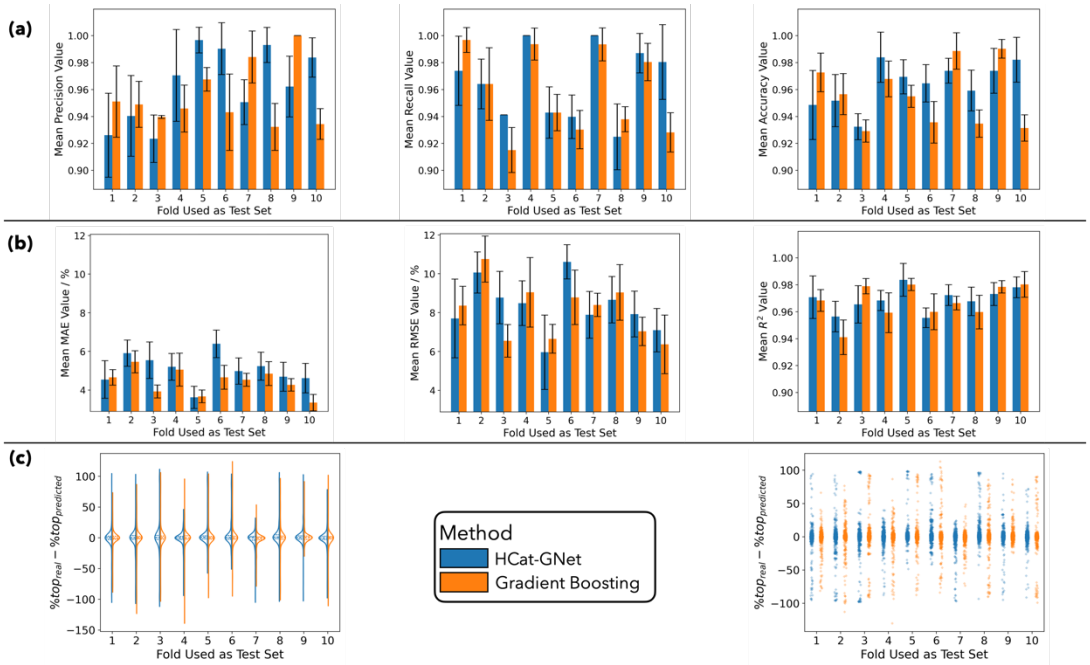

Fig. 3 Summary of results obtained predicting selectivity of test set reactions in the seen dataset. The reported values are shown for each test fold separately, using the mean of the metrics for the value in the bar and the standard deviation as error bars. (a) Shows the results for the 'face of addition, classification task, (b) shows the results of '%top' regression task, and (c) shows the error distribution for all the predicted points in the nine train-testing procedures for each test fold. The colour key in (c) applies for all the metrics shown.

features are only generated in a data-driven way. Although we explored the capabilities of the proposed method with the RhCAA study, the results demonstrate that HCat-GNet has the potential to generate predictions at least as good as TML approaches, without requiring bespoke optimized descriptors. The advantage of using SMILES representations is that formally, *any* molecule that can be represented by these types of strings can be used as an input into the model for learning and predicting structure-selectivity relationships. TML methods are usually limited by their fragment characterization processes that often cause the machine to become overly dependent on overly small regions of chemical space when predicting catalytic reactions. We hypothesize that this causes many TML methods to overfit and fail to predict closely related, but truly unseen catalysts. To test this latter point, we sought to predict the selectivities of reactions catalyzed by ligands never seen by either TML or HCat-GNet.

Model transferability evaluation

Fig. 4a shows the results for face of addition prediction for reactions with unseen, structurally distinct ligands, described by Rit *et al.*[49] HCat-GNet is 100% accurate in this task. It consistently predicts the correct product stereoconfiguration even though it has never been trained on structurally similar ligands. In comparison, TML Gradient Boosting struggles to deliver against the same challenge. In the case of precision, its values fall as low as 0.60 for test fold 4. However, for fold 5 the score obtained is 1.00. In the case of recall, the worst value obtained is in fold 6, while fold 3 scores 1.00. Lastly, accuracy shows ranges of values from 0.50 for fold 4 and 0.90 for fold 3.

The results obtained show that although the models performed similarly for the seen dataset, TML is distinctly inferior at predicting genuinely unseen reactions. Thus, the chemical patterns found by HCat-GNet on the seen dataset are similar to those found by the Gradient Boosting but are more transferable to new chemistries.

Similar results are obtained for '%top' prediction, shown in **Fig. 4b**. HCat-GNet shows an MAE below 5%, RMSE below 15%, and R^2 close to 1.00 with low variance. In the case of Gradient Boosting, the MAE values range from 15 to 20%, RMSE from 15 to 25%, and R^2 from 0.50 to 0.90. The violin plot in **Fig. 4c** shows that the error distribution in TML is significantly broader than the HCat-GNet. From the strip plot it is also evident that HCat-GNet never produces errors greater than 50%, while the TML generated errors from -100 to 100%.

We hypothesize that the seen dataset was useful for both methods to learn structure-

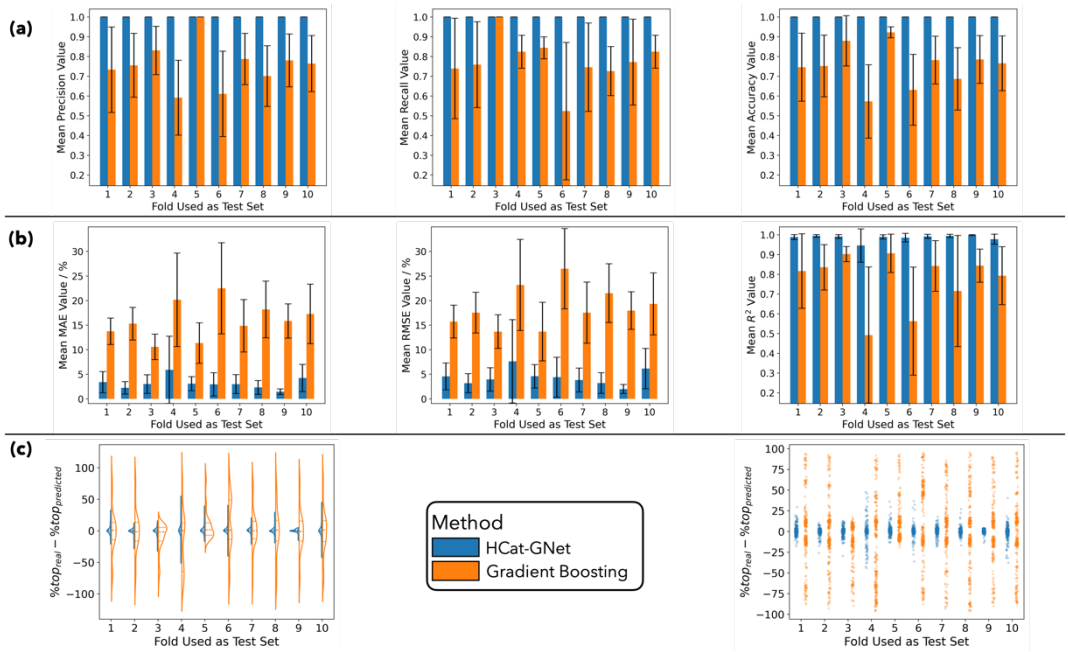

Fig. 4 Summary of results obtained predicting selectivity of test set reactions in the unseen dataset. The reported values are shown for each test fold separately, using the mean of the metrics for the value in the bar and the standard deviation as error bars. (a) Shows the results for the 'face of addition, classification task, (b) shows the results of '%top' regression task, and (c) shows the error distribution for all the predicted points in the nine train-testing procedures for each test fold. The colour key in (c) applies for all the metrics shown.

selectivity relationships close to the real ground-truth relationships. However, the nature of GNNs and the representation used provide greater transferable learning to new (genuinely unseen) ligands, whereas the TML is limited by the arbitrary component representation chosen.

To analyse the overall performance of both models with both datasets, we created a parity plot of all predicted points within the 90 training-evaluation processes. We grouped the plots by method and by dataset. Since all testing points are now present in a single plot, we calculated the metrics for all those points and present this instead of a mean metric (Table 1). In addition, **Fig. 5** plots the parities of the data predictions for both seen and unseen data using both HCat-GNet and Gradient Boosting.

Table 1 Summary of metrics obtained for the total of 6174 test datapoints from the Seen Dataset and 3060 test datapoints from the Unseen Dataset. Precision, recall, and accuracy metrics correspond to the classification task, while MAE, RMSE, and R^2 correspond to the regression task.

Metric	HCat-GNet		Gradient Boosting	
	Seen Dataset	Unseen dataset	Seen Dataset	Unseen dataset
Precision	0.963	1.000	0.954	0.740
Recall	0.965	1.000	0.958	0.775
Accuracy	0.964	1.000	0.956	0.752
MAE / %	7.627	3.140	7.275	28.945
RMSE/ %	17.410	5.912	17.393	39.538
R^2	0.852	0.985	0.852	0.342

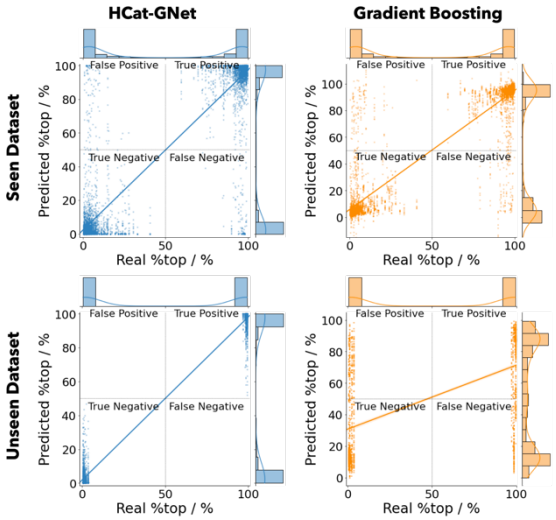

Fig. 5 Summary of results obtained by both approaches on both datasets. In the case of the seen dataset, the total number of test datapoints is 6174, and in the case of the unseen dataset the total number of test points is 3060. All the plots show the metrics obtained for both regression task and classification task. For this last report, the regression task metrics includes all datapoints (including the points with mis-predicted face of addition).

The plots of the seen dataset demonstrate that both TML and HCat-GNet methods generate similar distributions to those actually obtained from real experiments, indicating that both methods usefully predict the reaction stereoselectivities given access to the identical seen training set. As it is not trained on the unseen data, TML Gradient Boosting is unable to predict the distribution of real (experimental)

'%top' values. Although the classification metrics obtained by the TML are acceptable (0.752 accuracy), the values of predicted selectivity differ highly from the real lab experimental values. This means that although the TML could predict the face of addition, quantitative prediction of selectivity is not reliable, limiting its use for real-world new ligand discovery.

On the other hand, HCat-GNet is able to predict the stereoselectivities from both its training set ligands and truly unseen (but related) ligands. We hypothesize that the GNN has an ability to learn high-level features that approximate to the important factors of a chemical reaction that determine the experimental selectivity. Another possibility is that the HCat-GNet representation is more universal due to SMILES encoding of all molecular components. Both options deliver the same conclusion: HCat-GNet is more robust and general, while the TML method(s) evaluated herein are more limited because of their representation or overfitting issues. This answers our second research question (how TML methods compare to GNNs in predicting truly unseen exemplar reactions), confirming the limitations of TML approaches when trying to extrapolate to truly unknown regions of chemical space.

Even though the GNN models developed could be improved by using more sophisticated convolutional operators or by adding edge features to the graph, we argue that the success of our models in predicting unseen ligands is due to its simplicity. The minimal graph representation combined with a simple GNN architecture allow the models to gain more general insight about the chemistry of the process, making the GNN to be more transferable than the TML approach. The GNN's ability to predict the behaviour of unseen ligands give it an 'intuitive' character that allows patterns found by the model to be correlated with real world chemistries meaningful to human chemists. We predict that these traits will be useful for *de novo* design of (chiral) ligands.

Structure-selectivity relationships from HCat-GNet

Fig. 6 shows the results of the GNNExplainer tool applied to HCat-GNet. This reveals that ligand node features have higher attribution scores than either boron reagent and substrate node features. The node feature with the highest scores for the ligand are atom degree (i.e. primary, secondary, tertiary substitution), followed by atom identity, atom hybridization, atom stereoconfiguration, and then overall ligand chirality. We argue that the first node feature is related to the steric factor of the process, since atoms with a higher degree of substitution correspond to groups with higher volumes. Such steric analysis is also transferable to hybridization of the ligand atoms, where atoms with sp^2 and sp^3 hybridization frequently correspond to high-volume groups, while sp hybridization indicate lower volume alkynes. Atom identity has a lower importance because the majority of atoms in the ligands are carbon atoms and the effect of any heteroatoms present is minimal. The data suggests the GNN attributes greater importance to the electronegativity of each atom in determining the selectivity of the process. Lastly, both atom absolute stereoconfiguration and ligand overall configuration importance shows that the models are able to learn the correlation between the chirality of the ligand and the stereochemical outcome.

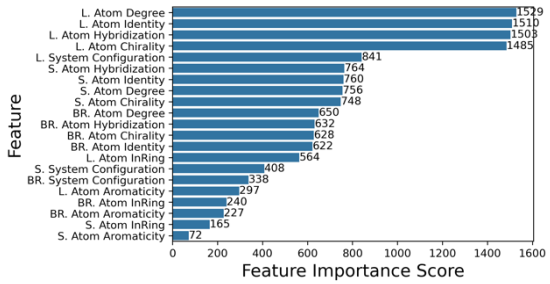

Fig. 6 Attribution score of node features separated by molecule in the system approximated by GNNExplainer XAI algorithm, where L symbolises the atoms of the catalyst, S to the atoms of the substrate and BR to the atoms of the organoboron reagent. Further information about these node features is given in Methods Hcat-GNet graph representation section.

Hybridization, atom identity, degree, and chirality of the substrate also rank highly while the boron reagent properties show less importance in the stereochemical outcome of the reaction. This makes sense as the organoboron reagent causes smaller changes in selectivity (~5%) in comparison to the ligand, where a change in configuration leads to a completely different enantiomeric product. As would be expected, the system level overall chirality descriptor (where the ligand configuration is tagged to all the components reacting) showed less relevance for the substrate and the boron reagent than for the ligand. We argue that this means that the GNN is learning that this descriptor is not relevant for the non-ligand components. Finally, the aromaticity of the atoms (nodes) and whether

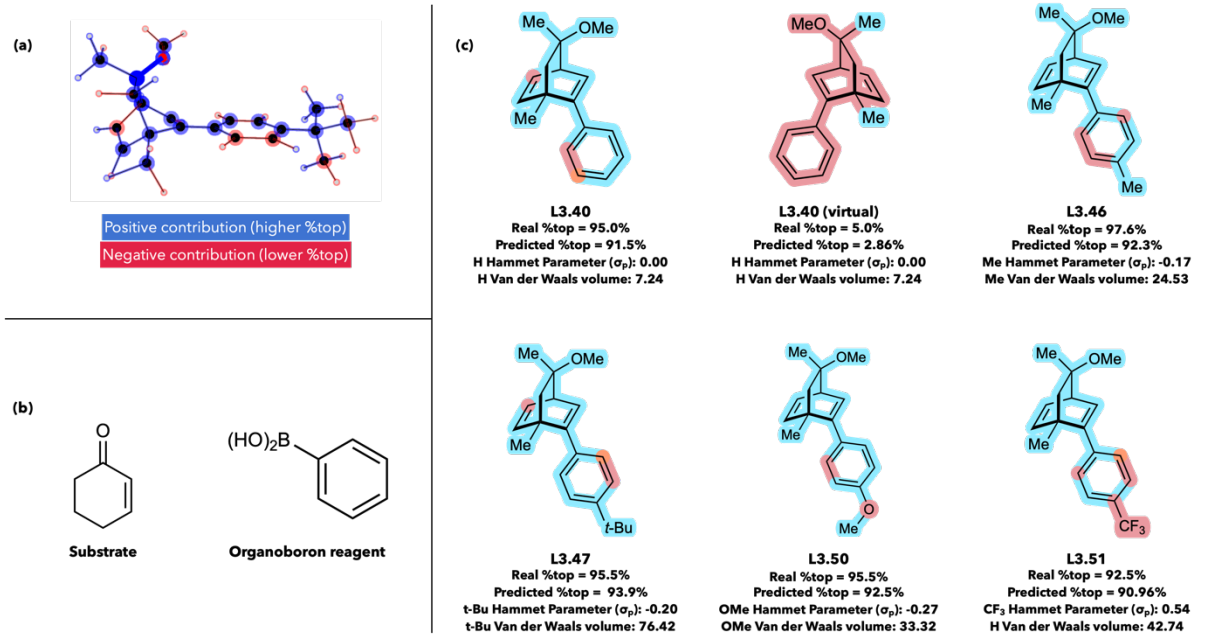

Fig. 4 Results of Shapley Values Sampling Analysis to different reactions that only change on the structure of the ligand. (a) Representative image of the output of the software developed to visualise the importance of each edge and node within a molecule to the final decision of the GNN, (b) shows the structure of the substrate and organoboron reagent of the reaction analysed, and (c) shows the structure of the ligands with highlighted edges and nodes that show their impact in the final decision of the GNN. Van der Waals volumes were approximated using the Zhao method,[51] and Hammett parameters were extracted from an extensive review.[50]

they belong to a ring system or not are found to be of less relevance for most predictions. We hypothesize this is the case because such features can be derived from pre-existing low-level node features, such as atom identity and degree. Thus, the electronic and steric effects of these high-level features are already learnt in the training process.

The agreement of the (ranked) important GNN node features with experimental (human) learnt RhCAA characteristics shows the GNN can independently learn and identify which structural variables within the reacting molecules correlate with the stereoselectivity of the process. Although for RhCAA a previous human-derived (Hayashi[24]) stereochemical model is available, this is *not* used in HCat-GNet, which independently derived correlations that make sense from the chemical point of view. Importantly, HCat-GNet can in principle be used for other processes where the reaction mechanism is unknown.

To understand the electronic effects of a given substituent within the ligand, we use Shapley Value Sampling. While the steric selection factors in RhCAA are well-understood, this is not the case for electronic effects. We decided to probe this using six different reactions, where five of them differ only in the R^4 substituent in the ligand (see **Fig. 1d**): *para*-substituted C_6H_4R (R = H, Me, *t*-Bu, OMe, or CF_3) group, together with the enantiomer of one ligand (R = H) as a control. We aimed to understand how the electronics of the substituent affects the GNN facial selectivity decision. To do this, we created code that plots the ligand in 3D with colors depicting the contribution of each edge and node to the final decision of the GNN (**Fig. 7a**). For these experiments, we chose identical reaction conditions (organoboron reagent, substrate, temperature, and solvent, **Fig. 7b**), with only the ligand differing in each case. The results are shown in **Fig. 7c**.

The ligands **L3.40** and *ent*-**L3.40** (virtual) are enantiomers. Ligand *ent*-**L3.40** (virtual) corresponds to a data (inversion) augmentation strategy used to train the GNN in enantiomeric behaviour**. L3.40** and *ent*-**L3.40** (virtual) are, by definition, expected to give opposite product enantiomers but with the same degree of stereoselectivity. As shown in **Fig. 7c**, the GNN successfully identifies **L3.40** as a positive contributor (higher %top or top face addition), while *ent*-**L3.40** (virtual) is identified as negative contributor (lower %top or bottom face addition). This agrees with the Hayashi stereochemical model,[25] and shows the GNN is able to effectively model the effect of both relative and absolute ligand stereochemistry on reaction enantiofacial selectivity.

For the case of ligands **L3.40**, **L3.46**, **L3.47**, **L3.50**, and **L3.51**, we manually correlated the GNN-derived and experimental %top of these

reactions with the Hammett parameters[50] and calculated van der Waals volume[51] of the different R substituents in the *para*-substituted C_6H_4R ligand group. Negative Hammett values (electron-donating groups) engender an increase in %top compared to **L3.40**, while positive Hammett parameters correspond to lower %top. Even though the GNN does not exactly predict the selectivity for these ligands, the trend of impact of the substituent in the reaction outcome is the same as that experimentally observed (Me, *t*-Bu, and OMe increase %top, and CF_3 decreases it).

Shapley Value Sampling analysis suggests HCat-GNet can augment these electronic deductions from low-level node-features. In the case of ligand **L3.40**, no specific negative contributions are found. For inductive electron-donating R groups (**L3.46**, Me and **L3.47** *t*-Bu), the GNN successfully identifies them as positive contributors to the %top. In the case of **L3.50**, the GNN identifies an electron-withdrawing inductive effect (–I) from the electronegative oxygen atom as a negative contributor and its methyl substituent as a positive contributor. This is different to just considering OMe as an electron-donating (+M) substituent. Potentially, the GNN is reasoning is that inductive effects are more important in the reaction, and therefore classifying the oxygen as negative contributor towards the selectivity, while the methyl group is considered a positive contributor. This could be because of two reasons: 1) the GNN is capturing that the real importance of inductive effects are higher than mesomeric effects, or 2) the lack of samples with mesomeric effects creates a bias in the models that makes them overestimate inductive effects. Lastly, for **L3.51**, the Shapley analysis confirms the CF_3 group as a negative contributor due to its inductive electron-withdrawing properties.

Comparing the Shapley Value Sampling method and the Hayashi stereochemical model,[25] the former is more information-rich compared with the latter, purely empirical model. Thus, the GNNExplainer along with Shapley Value sampling can present a fuller picture of asymmetric catalytic reactions.

The agreement between the real RhCAA selectivity data and the GNN-determined values suggests that valuable information regarding the factors that determine that stereoselectivities of reactions may well be possible for a wide range of asymmetric catalytic reactions using the GNNExplainer, complemented by Shapley Value Sampling. Thus, HCat-GNet is not only useful in making accurate predictions of selectivity, but also in understanding underlying phenomena that human scientists have not yet deciphered or observed. This opens a wide range of opportunities for asymmetric catalysis optimization and *de novo* ligand design.

Conclusions

We have developed a homogeneous catalyst graph network (HCat-GNet), which consists of a Graph Neural Network model predicting the stereoselectivity and absolute stereoconfiguration of the major product of an asymmetric reaction accelerated by a metal-ligand complex ML* (L* = any chiral ligand). We demonstrated the applicability of HCat-GNet by predicting the face of addition and the quantitative selectivity of >500 experimental RhCAA reactions using only SMILES inputs. Our method had a comparable performance to previously published approaches,[26] but without the need for bespoke descriptors/curation. Our GNN was able to find human-understandable high-level features demonstrating that HCat-GNet can be successfully used to create machine learning models of asymmetric homogeneous asymmetric catalysts with little or no human intervention. We further tested the robustness of HCat-GNet by predicting the selectivity associated with 30 new completely unseen ligands[49] without further chemical space training, showing that HCat-GNet is much better at extrapolation tasks than traditional machine learning. Our results suggest that overfitting in current TML approaches might be more common than is presently appreciated. Conversely, HCat-GNet's use of molecular graphs makes it more robust in providing insights into (at least 'nearby') unknown chemical space. Lastly, the use of GNNExplainer and Shapley Value Sampling allows the unmasking of both ligand steric and electronic effects making HCat-GNet a potentially useful tool for probing catalytic reaction mechanisms where the underlying phenomena have not yet been fully elucidated, and for optimal ligand design.

Authors Contributions

E. A. B. performed the ML studies and wrote the original paper draft. E. Ö. and G. F. commented and augmented the draft. R. R. provided the BD data set prior to its publication. H. W. L. commented on the draft. S.W. Review and editing. E. A. B., G. F., E. Ö. and S. W. jointly conceived, designed directed the research.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

We acknowledge funding from the Doctoral Training Centre in Artificial Intelligence of the School of Computer Science of the University of Nottingham.

References

1 P. G. Levi and J. M. Cullen, *Environmental Science & Technology*, 2018, **52**, 1725–1734.
2 J. Hagen, *Industrial Catalysis: A Practical Approach, 3rd Edition*, Wiley, 2015.
3 R. Franke, D. Selent and A. Börner, *Chemical Reviews*, 2012, **112**, 5675–5732.
4 Norio. Miyaura and Akira. Suzuki, *Chemical Reviews*, 1995, **95**, 2457–2483.
5 A. Vittoria, A. Meppelder, N. Friederichs, V. Busico and R. Cipullo, *ACS Catalysis*, 2017, **7**, 4509–4518.
6 D. J. Berrisford, C. Bolm and K. B. Sharpless, *Angewandte Chemie International Edition in English*, 1995, **34**, 1059–1070.
7 D. W. Robbins and J. F. Hartwig, *Science*, 2011, **333**, 1423–1427.
8 A. McNally, C. K. Prier and D. W. C. MacMillan, *Science*, 2011, **334**, 1114–1117.
9 Y.-F. Shi, Z.-X. Yang, S. Ma, P.-L. Kang, C. Shang, P. Hu and Z.-P. Liu, *Engineering*, 2023, S2095809923002813.
10 E. N. Muratov, J. Bajorath, R. P. Sheridan, I. V. Tetko, D. Filimonov, V. Poroikov, T. I. Oprea, I. I. Baskin, A. Varnek, A. Roitberg, O. Isayev, S. Curtalolo, D. Fourches, Y. Cohen, A. Aspuru-Guzik, D. A. Winkler, D. Agrafiotis, A. Cherkasov and A. Tropsha, *Chem. Soc. Rev.*, 2020, **49**, 3525–3564.
11 K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, *Nature*, 2018, **559**, 547–555.
12 D. S. Wigh, J. M. Goodman and A. A. Lapkin, *WIREs Computational Molecular Science*, 2022, **12**, e1603.
13 G. dos Passos Gomes, R. Pollice and A. Aspuru-Guzik, *Trends in Chemistry*, 2021, **3**, 96–110.
14 J. D. Hirst, S. Boobier, J. Coughlan, J. Streets, P. L. Jacob, O. Pugh, E. Özcan and S. Woodward, *Artificial Intelligence Chemistry*, 2023, **1**, 100006.
15 A. V. Kalikadien, A. Mirza, A. N. Hossaini, A. Sreenithya and E. A. Pidko, *ChemPlusChem*, 2024, **n/a**, e202300702.
16 B. Meyer, B. Sawatlon, S. Heinen, O. A. von Lilienfeld and C. Corminboeuf, *Chem. Sci.*, 2018, **9**, 7069–7077.
17 J. A. Hueffel, T. Sperger, I. Funes-Ardoiz, J. S. Ward, K. Rissanen and F. Schoenebeck, *Science*, 2021, **374**, 1134–1140.
18 M. Cordova, M. D. Wodrich, B. Meyer, B. Sawatlon and C. Corminboeuf, *ACS Catalysis*, 2020, **10**, 7021–7031.
19 P. Friederich, G. dos Passos Gomes, R. De Bin, A. Aspuru-Guzik and D. Balcells, *Chem. Sci.*, 2020, **11**, 4584–4601.
20 O. Schilter, A. Vaucher, P. Schwaller and T. Laino, *Digital Discovery*, 2023, **2**, 728–735.
21 S. M. Maley, D.-H. Kwon, N. Rollins, J. C. Stanley, O. L. Sydora, S. M. Bischof and D. H. Ess, *Chem. Sci.*, 2020, **11**, 9665–9674.
22 A. V. Brethomee, R. S. Paton and S. P. Fletcher, *ACS catalysis*, 2019, **9**, 7179–7187.
23 A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow and S. E. Denmark, *Science*, 2019, **363**, eaau5631.
24 B. Owen, K. Wheelhouse, G. Figueredo, E. Özcan and S. Woodward, *Results in Chemistry*, 2022, **4**, 100379.
25 T. Hayashi, K. Ueyama, N. Tokunaga and K. Yoshida, *Journal of the American Chemical Society*, 2003, **125**, 11508–11509.
26 R. Ardkhean, M. Mortimore, R. S. Paton and S. P. Fletcher, *Chem. Sci.*, 2018, **9**, 2628–2632.
27 S. Singh, M. Pareek, A. Changotra, S. Banerjee, B. Bhaskararao, P. Balamurugan and R. B. Sunoj, *Proceedings of the National Academy of Sciences*, 2020, **117**, 1339–1345.
28 J. P. Reid and M. S. Sigman, *Nature*, 2019, **571**, 343–348.
29 J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals and G. E. Dahl, .
30 P. Reiser, M. Neubert, A. Eberhard, L. Torresi, C. Zhou, C. Shao, H. Metni, C. van Hoesel, H. Schopmans, T. Sommer and P. Friederich, *Commun Mater*, 2022, **3**, 1–18.
31 K. Atz, F. Grisoni and G. Schneider, *Nat Mach Intell*, 2021, **3**, 1023–1032.
32 X. Fang, L. Liu, J. Lei, D. He, S. Zhang, J. Zhou, F. Wang, H. Wu and H. Wang, *Nat Mach Intell*, 2022, **4**, 127–134.
33 J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li and M. Sun, *AI Open*, 2020, **1**, 57–81.
34 S. Zhang, H. Tong, J. Xu and R. Maciejewski, *Computational Social Networks*, 2019, **6**, 11.
35 T. Tian, S. Li, M. Fang, D. Zhao and J. Zeng, *Journal of Chemical Information and Modeling*, **0**, null.
36 Z. Wu, J. Wang, H. Du, D. Jiang, Y. Kang, D. Li, P. Pan, Y. Deng, D. Cao, C.-Y. Hsieh and et al, *Nature News*, 2023.

37 A. Kotobi, K. Singh, D. Höche, S. Bari, R. H. Meißner and A. Bande, *Journal of the American Chemical Society*, 2023, **145**, 22584–22598.
38 Y. Jian, Y. Wang and A. Barati Farimani, *ACS Sustainable Chemistry & Engineering*, 2022, **10**, 16681–16691.
39 Q. Yuan, F. T. Szczypiński and K. E. Jelfs, *Digital Discovery*, 2022, **1**, 127–138.
40 A. R. N. Aouichaoui, F. Fan, S. S. Mansouri, J. Abildskov and G. Sin, *Journal of Chemical Information and Modeling*, 2023, **63**, 725–744.
41 A. G. Garrison, J. Heras-Domingo, J. R. Kitchin, G. dos Passos Gomes, Z. W. Ulissi and S. M. Blau, *Journal of Chemical Information and Modeling*, 2023, **63**, 7642–7654.
42 L. Zoute, G. Kociok-Köhn and C. G. Frost, *Org. Lett.*, 2009, **11**, 2491–2494.
43 S. Tan, J.-G. Liu and M.-H. Xu, *Org. Lett.*, 2022, **24**, 9349–9354.
44 A. R. Burns, H. W. Lam and I. D. Roy, in *Organic Reactions*, John Wiley & Sons, Ltd, 2017, pp. 1–415.
45 T. N. Kipf and M. Welling, 2017.
46 R. Ying, D. Bourgeois, J. You, M. Zitnik and J. Leskovec, *CoRR*, 2019, abs/1903.03894.
47 J. Castro, D. Gómez and J. Tejada, *Computers and Operations Research*, 2009, **36**, 1726–1730.
48 Erik \vStrumbelj and I. Kononenko, *J. Mach. Learn. Res.*, 2010, **11**, 1–18.
49 R. K. Rit, H. Li, S. P. Argent, K. M. Wheelhouse, S. Woodward and H. W. Lam, *Advanced Synthesis & Catalysis*, 2023, **365**, 1629–1639.
50 Corwin. Hansch, A. Leo and R. W. Taft, *Chemical Reviews*, 1991, **91**, 165–195.
51 Y. H. Zhao, M. H. Abraham and A. M. Zissimos, *The Journal of Organic Chemistry*, 2003, **68**, 7368–7373.
52 S. Pablo-García, S. Morandi, R. A. Vargas-Hernández, K. Jorner, Z. Ivković, N. López and A. Aspuru-Guzik, *Nature Computational Science*, 2023, **3**, 433–442.