Reaction-Agnostic Featurization of Bidentate Ligands for Bayesian Ridge Regression of Enantioselectivity

Alexandre A. Schoepfer,[†,‡,¶] Ruben Laplaza,[†,¶] Matthew D. Wodrich,[†,¶] Jerome Waser,[*,‡,¶] and Clemence Corminboeuf[*,†,¶]

†*Laboratory for Computational Molecular Design, Institute of Chemical Sciences and Engineering, École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland*

‡*Laboratory of Catalysis and Organic Synthesis, Institute of Chemical Sciences and Engineering, École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland*

¶*National Center for Competence in Research-Catalysis (NCCR-Catalysis), École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland*

E-mail: jerome.waser@epfl.ch; clemence.corminboeuf@epfl.ch

Abstract

Chiral ligands are important components in asymmetric homogeneous catalysis, but their synthesis and screening can be both time-consuming and resource-intensive. Data-driven approaches, in contrast to screening procedures based on intuition, have the potential to reduce the time and resources needed for reaction optimization by more rapidly identifying an ideal catalyst. These approaches, however, are often non-transferable and cannot be applied across different reactions. To overcome this drawback, we introduce a general featurization strategy for bidentate ligands that is coupled with an automated feature selection pipeline and Bayesian ridge regression to perform multivariate linear regression modeling. This approach, which is applicable to any reaction, incorporates electronic, steric, and topological features (rigidity/flexibility, branching, geometry, constitution) and is well-suited for early-stage ligand optimization. Using only small datasets, our workflow capably predicts the enantioselectivity of four metal-catalyzed asymmetric reactions. Uncertainty estimates provided by Bayesian ridge regression permit the use of Bayesian optimization to efficiently explore pools of prospective new ligands. Using this procedure, a new library of 312 chiral bidentate ligands was screened to identify promising ligand candidates for a challenging asymmetric oxy-alkynylation reaction.

1 Introduction

Statistical methods accelerate the discovery and optimization of chemical reactions in homogeneous catalysis.[1–17] Employing these "data-driven" approaches requires abundant, high-quality data,[18–20] that is often scarce. Ligand optimization, in particular, suffers from this problem, since most experimental datasets tend to be size limited as a result of ligand screening campaigns that often consist of fewer than a dozen experiments. In such "low data" regimes, nonlinear statistical models perform poorly due to overfitting. On the other hand, multivariate linear regression (MLR) models

offer data-efficient alternatives that can be developed from only a few samples, yet, are robust, interpretative, and extrapolative to unseen ligands.

To develop MLR models, catalysts are usually first optimized using DFT and then featurized.[12,15,16,21–26] The resulting molecular features (*e.g.*, atomic charges, local stretching frequencies, cone/bite angles) are low-dimensional and highly interpretable,[16,27–29] which allows design principles and hypothesized reaction pathways to be derived from the fitted models.[17,21,23–26,30] This established approach, however, suffers from two significant drawbacks: first, specific features for the chemical problem of interest must be selected for the MLR model and second, only the most relevant of these features are used in developing the final model. As a result, MLRs are often not transferable to different settings (*e.g.*, a similar reaction incorporating a different family of ligands) as those features previously selected may not be defined. To overcome this problem Gensch et al.[31] recently introduced a comprehensive featurization strategy for monodentate organophosphorus ligands that facilitates the creation of MLRs for any possible reaction class. Establishing this paradigm for more complex ligand types is of great interest for developing transferable predictive models across catalyst families and reaction classes.[16]

To this end, here we present a reaction-agnostic workflow applied to bidentate ligands that employs, amongst others, seldom-used topological-based features. Coupling this featurization strategy with an automated feature selection pipeline using Bayesian ridge regression (BRR)[32] allows development of models that capably predict the enantioselectivity of four different reaction classes while highlighting the importance of using topological-based features. Moreover, by leveraging the calibrated uncertainty estimations from BRR we demonstrate Bayesian optimization (BO) for optimal ligand screening[33–39] by examining an original set of 312 ligands extracted from the Cambridge Structural Database (CSD). Overall, this work demonstrates that linear models are indeed more accurate and data-efficient than nonlinear methods in the "extremely low" data regime.

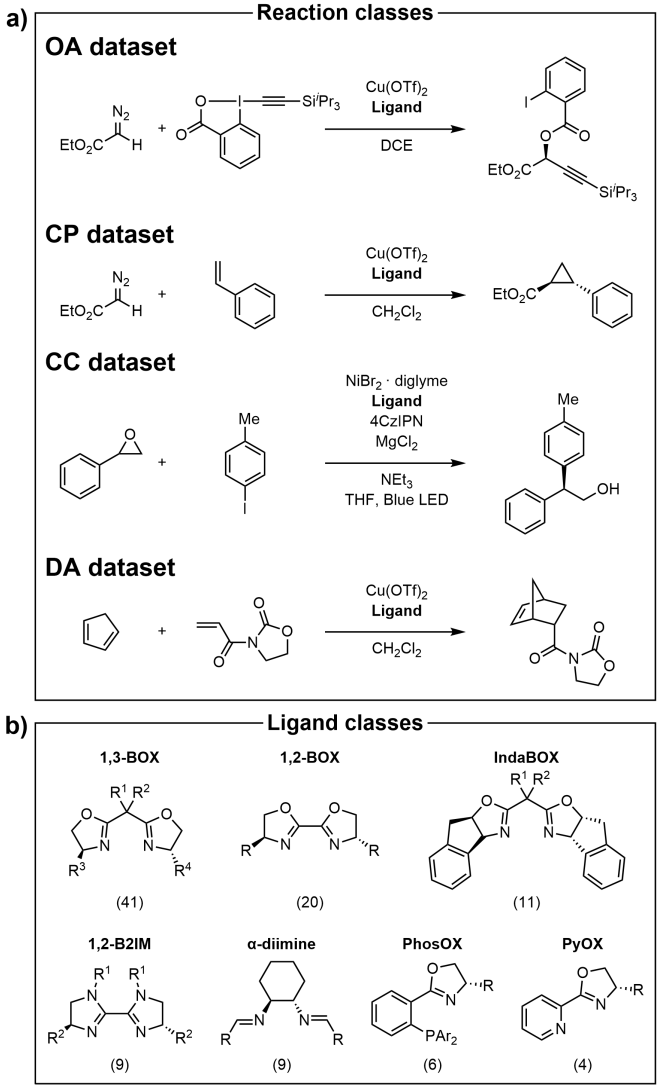
2 Methods

2.1 Datasets

To develop, train, and test our pipeline, four asymmetric reaction classes that previously underwent extensive experimental ligand screening were selected for examination (Scheme 1, top): copper-catalyzed oxy-alkynylation of diazo compounds with hypervalent iodine reagents (**OA**),[40] copper-catalyzed cyclopropanation of styrene with diazo esters (**CP**),[41] nickel/photoredox-catalyzed cross-electrophile coupling of styrene oxides with aryl iodides (**CC**),[21] and a copper-catalyzed Diels-Alder ligand benchmark reaction with cyclobutadiene and an imide (**DA**).[42–48] Table 1 gives an overview of these datasets. For each reaction, the reactants, reagents (except the ligand), and solvent were kept constant, while reaction conditions (metal loading, time, temperature, *etc.*) varied among experiments. For this reason, the datasets are small (ranging from 19 to 30 data points) as they include only ligand screening experiments (see SI Section S2 for details). For three of the four curated datasets, all ligands originated from a single publication while the fourth dataset (**DA**) contains ligands taken from seven different publications, which introduces additional noise in the data resulting from different experimental setups. For the **OA** dataset, additional ligands with enantiomeric excesses not part of the original publication,[40] were included from electronic laboratory notebook entries (see SI Section S2.1).

Combining all four datasets gives a total of 100 unique bidentate ligands, which were curated as a ligand pool for exploration (see Section 3.2). Most reactions (across all datasets) used bis-oxazoline (BOX) type ligands (see Table 1), but other ligand classes [bi-2-imidazolines (B2IM), α-diimines, phosphorous-oxazolines (PhosOX),

and pyridine-oxazolines (PyOX, Scheme 1, bottom)] are also present in smaller numbers. In addition, bidentate ligands bound to Cu(I) or Cu(II) found in the Cambridge structural database (CSD)[49,50] from an unbiased curated subset were extracted with Cell2mol,[51] which yielded an additional 312 new ligands possessing at least one chiral atom. These ligands were used as a separate pool for further exploration (*vide infra*).



Scheme 1: (a) Candidate asymmetric reactions used to develop and test the presented pipeline. (b) Ligand families used in these reaction classes. The number of unique ligands per class is shown in parentheses.

2.2 Bidentate ligand featurization

Molecular features (*i.e.*, molecular descriptors) were split into three categories: electronic, steric, and topological (Figure 1a) and further categorized based on their intensive or extensive nature. To maximize generality, the only local features used describe the ligand's two complexing atoms (that bind the metal), which are present in all bidentate ligands. Steric features were computed using a consistent alignment for all ligands that allows ligand's molecular volume to reproducibly split into octants, quadrants, and halves (see SI Section S4.1). Full and buried volumes were computed following the recommendations of Cavallo et al.[52]

Table 1: Overview of the datasets used in this work.

Dataset	OA	CP	CC	DA
# of datapoints	19	30	29	30
# of publications	1	1	1	7
Oxazoline ligands	16	30	20	21

Topological features, seldom exploited in multi-linear regression models for homogeneous catalysis, were determined from vertex and edge information of the ligand's molecular graph that were generated using covalent radii to assign bonds based on the DFT optimized geometry. This category includes global topological feature (*e.g.* Wiener, Hosoya Z, and Balaban J indices),[53–55] bond-fragment based descriptors (*e.g.*, the indices introduced by Kier and Hall),[56–63] and bond-order quantities (*e.g.*, local and global simple indices and the CREST flexibility index).[64,65] Variants of existing topological descriptors, originally used for drug design, were also developed and included to capture catalyst rigidity/flexibility.

Summed together, this strategy yields a total of 232 features for each ligand, which constitutes a representation of bidentate ligand space. The tSNE plot of the featurized ligands in Figure 1b showcases how these features successfully group ligands belonging to the same family (see Scheme 1b) while keeping related ligand families adjacent, in agreement with chemical intuition. Note that, by design, all features can be obtained for any possible bidentate ligand using the *Moltop* package and associated scripts available at `https://github.com/lcmd-epfl/rafbl`. Further

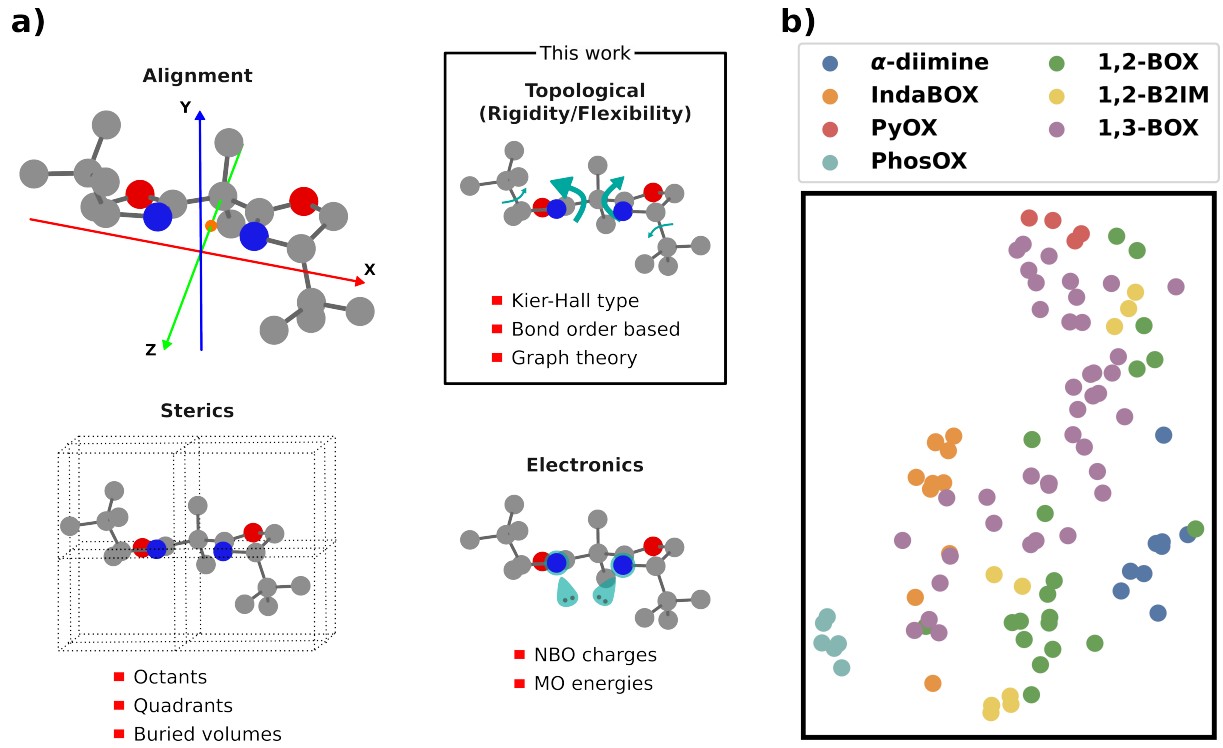


Figure 1: Featurization of bidentate ligands. (a) Alignment of the bidentate ligands in space and features classes. (b) T-distributed stochastic neighbor embedding (tSNE) map of the feature space for all ligands with all available features.

details and a complete description of all 232 features are given in the SI (Section S4).

2.3 Regression & optimization methods

In this work, Bayesian ridge regression (BRR, a regularized variation of least-squares fitting) was used to fit the MLR models whose parameters were estimated using Bayesian inference, which provides a calibrated uncertainty for each prediction. To avoid overfitting, model complexity was limited to one feature from each class (electronic, steric or topological) with an additional requirement that, at minimum, at least one of these features must be extensive. Employing these constraints reduced (by several orders of magnitude) the possible number of combinations with respect to brute-force exploration, which leads to a manageable screening to determine the best features (see SI Section S5.1 for details). We find that this approach leads to highly interpretable, robust models that outperform non-linear models (see SI Section S5.2 for a detailed comparison).

To guide ligand screening a pool-based Bayesian Optimization (BO), in which prospective ligands are run through the BRR-fitted model, was employed. For each ligand $\mathbf{x}$ in the pool the Expected Improvement (EI)[66] defined as

$$\mathrm{EI}(\mathbf{x}) = \sigma(\mathbf{x})(\mathbf{Z}\Phi(\mathbf{Z}) + \phi(\mathbf{Z})) \tag{1}$$

$$\mathbf{Z} = \left(\mu(\mathbf{x}) - \mu^{+}(\mathbf{x})\right)/\sigma(\mathbf{x}) \tag{2}$$

was computed. Here, μ represents the predicted value, μ^{+}, the current best value, σ the standard deviation, Φ the cumulative distribution function, and ϕ the probability density function. New results are subsequently incorporated into the training set and the process repeated until unexplored ligands each have EI scores lower than those already seen by the model. This implies that no further improvement (*i.e.*, no better ligand) is expected within the pool.

Figure 2: General workflow for model selection. (1) Ligands are optimized with a metal center to obtain the desired geometry. (2) 232 features are extracted from the metal-free structure. (3) The most promising feature combinations are identified through testing. (4) The best features undergo cross-validation and the best Bayesian ridge regression (BRR) model is obtained. (5) This resulting model is used for ligand screening with BO.

2.4 General workflow

Figure 2 overviews our proposed workflow. In step 1, $CuCl_2L$ structures are optimized at the PBE0-D3(BJ)/def2-SVP level (see Computational details) followed by automatic feature extraction for each structure (Step 2). The Eyring equation then is used to convert enantiomeric excesses (for experimentally available data) to energies ($\Delta\Delta G^{\ddagger}$) with the corresponding reaction temperature. Pre-selected feature combinations (see Section 2.3) are then evaluated with ordinary least-squares, first with no cross-validation (CV), to reduce computational time (Step 3). Those combinations with the best R^2 are then re-evaluated with BRR and the leave-one-out (LOO) CV scheme (Step 4). R^2_{LOO} is then calculated for all left-out points and used to score the final models. With the final model a pool of ligands is screened and BO used to identify the most promising candidates, which should next be tested (Step 5).

3 Results & Discussion

3.1 Generation of MLR models

Using the above pipeline, an interpretable MLR model was generated for each of the four Scheme 1 datasets (Figure 3). Recall that the MLR expressions contain, by design, one feature from each of the different classes (electronic, steric, and topological) and that no additional human input was required for featurization or feature selection (*i.e.*, all reactions used the same initial features and were run through the pipeline in an automated fashion). In general, the models perform well, with MAE_{LOO} not being higher than 0.29 kcal/mol. As in standard MLR, examining the normalized weights of these models reveals insight into the key aspects of the ligand that lead to high enantioselectivity (Figure 3).

For the oxy-alkynylation of diazo compounds model (**OA**, Figure 3a) the selected features are: the lone-pair NBO energies of the smaller ligand half ($-x$, electronic), the hydrogen-free volume of the other larger half (x, steric), and the normalized atom Kier ${}^2\kappa_\alpha$ index (lower values-more rigid, topological). The large topological feature weight indicates that a rigid catalyst structure is the most crucial element in determining enantioselectivity with sterics and electronics relegated to smaller roles. Taken together, these design principles indicate that Indane-derived BOX (IndaBOX) ligands are well suited for this reaction, as they are simultaneously both bulky and rigid. We hypothesize that catalyst rigidity favors more selective transition states.

For the cyclopropanation of styrene with diazo esters model (**CP**, Figure 3b) the selected features are: the lone-pair NBO energies (electronic), a buried octant ($-x, y, z$, steric) and the normalized Hall ${}^2\chi$ index (higher values-more rigid, topological). Here, sterics play a more important role than either electronics and rigidity, in agreement with previously proposed models.[41] A closer examination of the model reveals that the

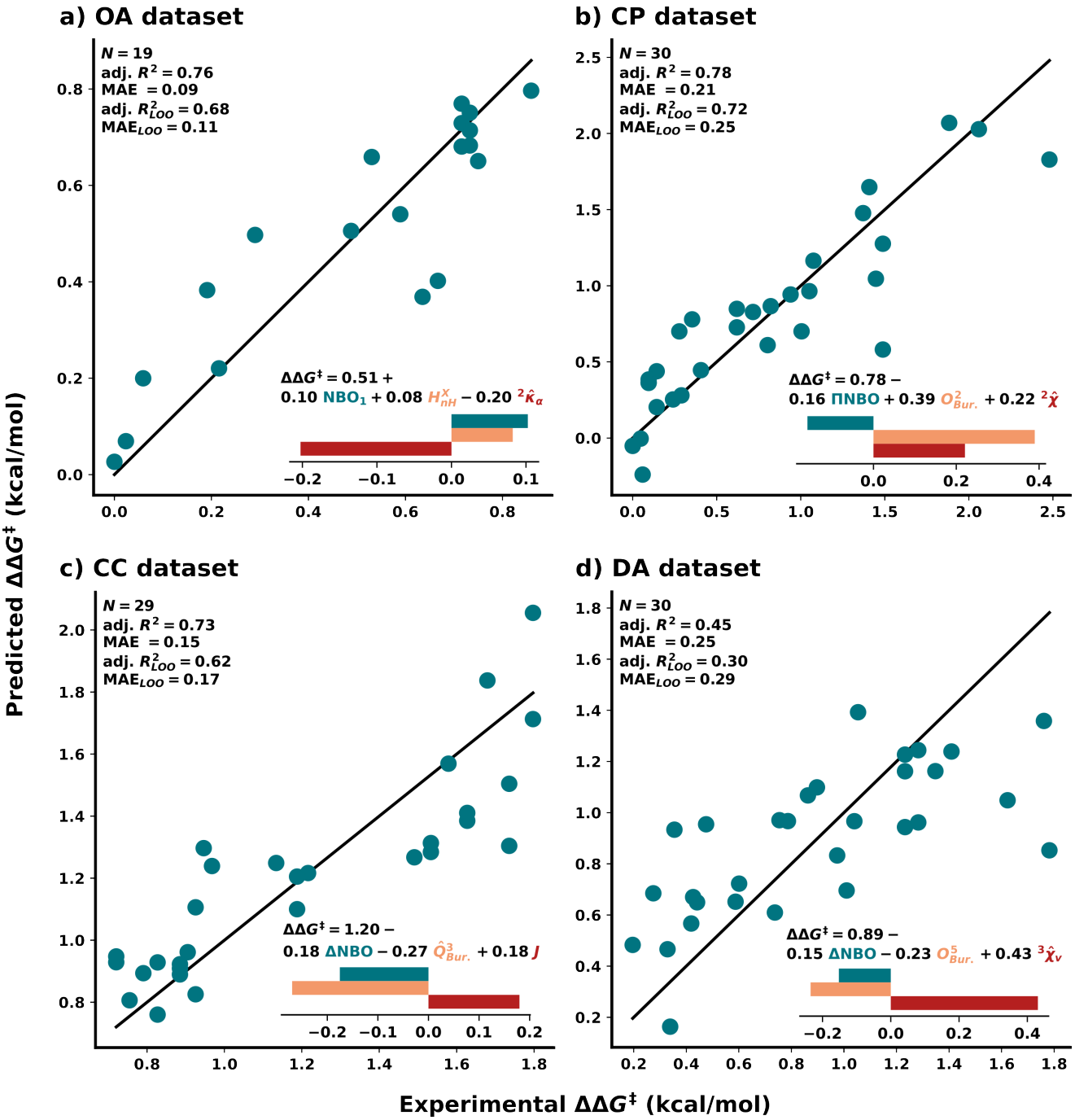


Figure 3: Fitted models for the four datasets, where N is the number of points, adj. R^2 the adjusted coefficient of determination, MAE the mean absolute error and the LOO subscript indicates the scores for the leave one out cross-validation. The model equation is represented in each plot with a depiction of the normalized weights. Electronic features are represented in teal, steric in orange, and topological in red. (a, **OA** dataset) Lone-pair NBO energies of the smaller half, hydrogen-free volume of larger half (along positive x-axis), and relative atom Kier ${}^2\kappa_\alpha$ index. (b, **CP** dataset) Product of lone-pair NBO energies, volume of buried octant 2 $(-x, y, z)$, and relative ${}^2\chi$. (c, **CC** dataset) Absolute difference of lone-pair NBO energies, relative buried volume of the south-west quadrant Q^3 $(-x, -y)$, and Balaban J index. (d, **DA** dataset) Absolute difference of lone-pair NBO energies, buried volume of octant 5 $(x, -y, -z)$, and relative ${}^3\chi_v$ index. See SI Section S4 for a complete description of all features.

$-x, y, z$ octant ($O^2_{Bur.}$) is important for enantioselectivity. This fact, in addition to the benefits of having lower electron occupancy near the chelating atoms, indicate that aza-BOX ligands should be good for this reaction.

For the cross-electrophile coupling of styrene oxides with aryl iodides model (**CC**, Figure

3c) the selected features are: the difference of lone-pair NBO energies (electronic), the normalized south-west quadrant ($-x, -y$, steric), and the Balaban-J index (higher values-more branched, topological). Here, the most important feature is steric ($\hat{Q}^3_{Bur.}$) and indicates that the southwest quadrant should be kept free, which aligns with previous postulations regarding the origin of stereoselectivity.[21] Overall, the family of B2IM ligands with closed, more rigid, backbones match well with the features of the model. Note that the adjusted R^2 of 0.73 of our model is similar to the adjusted R^2 of 0.74 previously reported for **CC**,[21] which shows that our pipeline yields similar predictivity and analogous interpretation without requiring information about reaction intermediates.

Finally, for the Diels-Alder reaction of cyclobutadiene an imide model (**DA**, Figure 3d) the selected features are: the NBO energies (electronic), a buried hydrogen-free octant ($x, -y, -z$, steric), and the normalized ${}^3\chi_v$ Hall index (higher values-more rigid, topological). Here, as in the **OA** dataset, the topological feature is found to be dominant. Being built from seven different publications, this reaction is particularly challenging. Nevertheless, our model has cross-validated errors of less than 0.3 kcal/mol.

The MLR models obtained from our pipeline for each of the four reaction datasets discussed above are both simple and interpretable owing to their limited number of features and selected composition. The importance of topological features that describe catalyst rigidity/flexibility (which, recall are typically absent in multi-linear regression models for homogeneous catalysis) across all four reactions is noteworthy, as in all cases these factors play, at minimum, an equally important role to electronic features (as seen through examination of the normalized weights).

3.2 Pool-based ligand optimization with BRR and BO

As illustrated above, the fitted BRR models can be used to elucidate design principles by analyzing the selected features/weights and interpreting the trends. Additionally, they may also be directly employed for ligand optimization (*e.g.*, to predict ligands that will impart higher selectivity). In this context, the ability to accurately extrapolate to unseen samples is crucial, particularly for cases where the training set contains only ligands with low enantioselectivities (*e.g.*, an initial batch based on similar reactions). To simulate this situation, we performed a 80/20 train/test split on the **OA** dataset to test the model on out-of-range predictions. As shown in Figure 4a, the test set includes the four best experimental ligands (red points), while the training set contains ligands with similar or worse performance (teal points). The complete pipeline was then rerun using the reduced training data (teal points only), which produced a similar (but not identical) model to that shown in Figure 3. Overall, this re-fit model shows low errors (MAE of 0.15 kcal/mol) for unseen samples and well-calibrated uncertainties that are nearly within 1σ from the reported experimental values. The enhanced uncertainty estimation, powered by BRR, coupled with the low prediction error on the test set, demonstrates that our pipeline yields models capable of extrapolating towards unknown ligands, including those anticipated to have greater selectivity.

Ideally, the newly developed BRR model can be used to more rapidly identify an ideal catalyst, which would avoid performing experiments that yield no improvement past the optimum. To assess this, we constructed a timeline depicting the original experimental reporting of each ligand (teal, Figure 4b). Here, the best performing ligand was found during the seventh experiment; all subsequent attempts did not yield any further improvement. Having established the model's ability to reliably predict out-of-range ligands with calibrated uncertainties (*vide supra*), we conducted BO to efficiently find the optimal ligand from a pool of candidates using the same initial three ligands as the training set. As shown in red (Figure 4b), the acquisition function identifies the best

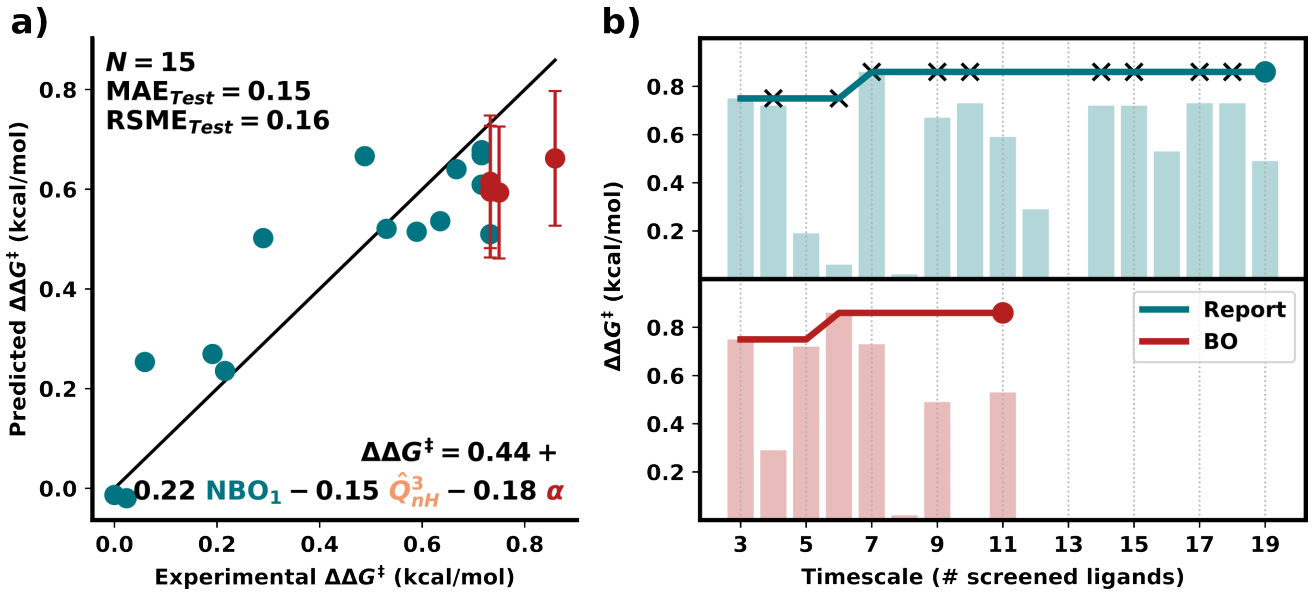


Figure 4: Retrospective and prospective experiments for **OA** dataset. (a) 80/20 train (teal)/test (red) split. Error bars correspond to 1σ. The selected features are: lone-pair NBO energies of the smaller half, hydrogen-free volume south-west quadrant Q^3 $(-x,-y)$, and constitutional Kier α. (b) Retrospective BO analysis. The "Report" line follows the original optimization timeline. New batches of ligands are represented as black crosses. Bar plots represent the $\Delta\Delta G^{\ddagger}$ values of each ligand in the batches. The BO starts with the three points from the first batch.

ligand as the third candidate to be tested (sixth total ligand, including the three included in the initial training set), faster than in the original experimental optimization procedure. From that point forward, five additional ligands were (incorrectly) predicted to bring potential improvement, however, given the uncertainty of these predictions, these additional species did not ultimately demonstrate improved selectivity. At this point, the stop criterion was met as no other ligand in the pool was predicted to provide further improvement, in agreement with experimental observations.

Compared to the original purely experimental screening, using the BO pipeline reduced the amount of required resources for the ligand screening by roughly a factor of two. Thus, BO was successful at rapidly finding the best ligand from the candidate pool while avoiding wasteful experiments. All of this while operating in the low data regime.

With these promising results in hand, we screened those ligands reported in the **CP**, **CC**, **DA** reactions (Scheme 1) to test their enantioselectivity for the **OA** reaction. Figure 5a (top left) shows the best ligand found experimentally as well as the top three "not-yet-sampled" ligand candidates derived from the other three reaction classes that were predicted by the BRR. Unfortunately, none of these new ligands (**2-4**) significantly improves the results for **OA** according to predictions from the model. Undeterred, we searched for more promising candidates by considerably expanding the pool to include 312 chiral bidentate ligands extracted from copper complexes in the Cambridge Structural Database (CSD, see Section 2.1). Interestingly, the most promising candidates include simple molecules such as tartaric acid (**5**), as well as amino acid derivatives from phenylalanine (**6**). Other amino acids (tyrosine, alanine, and tryptophan) also received good EI scores, as did N,N-ligands (**7**,**8**) such as BOX ligands with a sp^2 carbon bridge. On one hand, considerable differences in both the coordinating properties and charge of these CSD predicted ligands from those previously shown to be experimentally viable might lead to the conclusion that these new ligand would be incompatible with the reaction of interest. On the other hand, more "out-of-the-box" proposals may lead to significant advances by revealing new regions of ligand space to examine that had not been

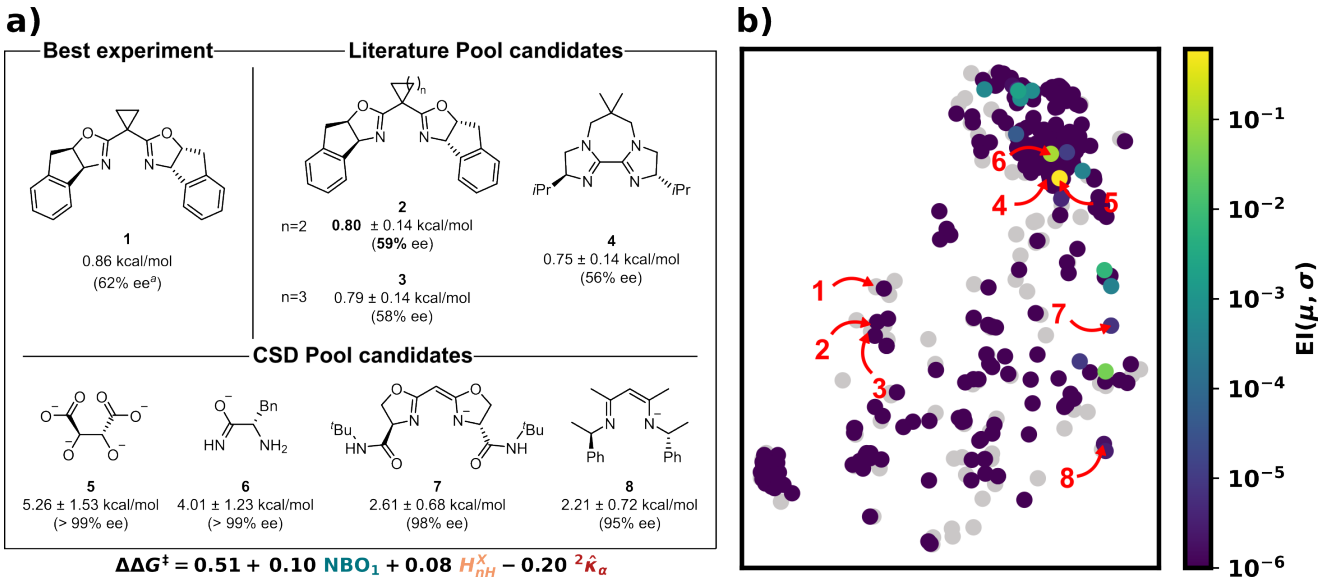


Figure 5: Predictions and analysis of literature and CSD extracted ligands for the **OA** reaction. (a) Top left: Best ligand found experimentally. Top right: top three EI candidates from pool-based predictions for the reaction, using the ligands from the three remaining literature datasets. Bottom: four diverse candidate ligands from the CSD which exhibit a high EI score. The predicted $\Delta\Delta G^{\ddagger}$ and its uncertainty are given. The used model is given at the bottom. (b) tSNE map of the CSD extracted ligands using the same embedding as in Figure 1b (gray points). The coloring corresponds to EI scores of each remaining CSD ligand (truncated to 10^{-6} for clarity). Red numbers correspond to structures in panel (a). [a]Optimized reaction conditions with this ligand yield 90% ee.

previously considered.

The four newly identified ligand classes (**5-8**) each possess distinct electronic properties (*i.e.*, negative charges) from the original sets of tested ligands, as well as high rigidities for **7** and **8**. The presence of oxygen and electron rich nitrogen atoms imparts substantially higher NBO charges than the previously tested ligands, which leads to higher predicted enantioselectivities according to the linear model. On the other hand, these ligands also have higher prediction uncertainties (*e.g.*, the uncertainty is an order of magnitude higher for tartaric acid than for ligands taken from the literature set) which arise because the CSD contains ligands that have not been experimentally tested for the **OA** reaction. To assess this increase in diversity, the CSD ligands were plotted with the previously discussed dimensionality reduction embedding (Figure 5b). In general, the previously known space is well covered by the new dataset, meaning that the ligands reported in the original four reactions are all well represented within the larger CSD set (but not *vice versa*). The emergence of new clusters (*e.g.*, in the top right and bottom left) indicates the presence of novel ligand families within the CSD set that were unexplored in the experimental ligand screenings. The most promising candidates (with high EI scores) are found in the top right cluster or along the right border, corresponding to ligands with sp^2 carbon bridges as well as amino acid derivatives. Based on these findings, we propose a series of follow-up ligand optimization reactions (**5,6,7,8**) for the **OA** dataset. While the uncertainties for these predictions are high and the probability to gain any additional improvements for the **OA** reaction is low, the use of amino acid derivatives[67] could open the door to inexpensive and sustainable ligands. In the future, we intend to make use of the proposed pipeline for ligand optimization of other **OA** related reactions.

4 Conclusions

In this work, we introduced a general workflow for constructing linear models from small numbers of screening experiments that predict enantioselectivity in reactions involving

bidentate ligands. Datasets comprising four different reaction classes (totaling 100 bidentate ligands belonging to seven ligand families) were curated to validate this approach and supplemented with an additional 312 ligands taken from the CSD that were used as a pool for further ligand optimization. Our workflow retrieves the best possible linear model established from a combination of electronic, steric, and (critically important but frequently overlooked) topological features that were determined using Bayesian Ridge Regression (BRR). By coupling BRR with Bayesian Optimization (BO) we were able to efficiently screen ligands, even in limited data scenarios, which allowed design principles to be extracted and new ligands to be proposed for the oxy-alkynylation reaction. Overall, the approach presented here enables researchers to optimize ligand selection and design at any stage of experimentation, resulting in more efficient and cost-effective enantioselective reaction development.

5 Computational details

DFT computations of ligands were done at the PBE0-D3(BJ)/def2-SVP level using Gaussian16.[68–70] For ligands extracted from the literature, 3D coordinates were generated using Openbabel, then optimized at the GFN2-xTB level[71] before final optimization with DFT.[72] The desired structure (chelating groups oriented towards the metal atom) was obtained by adding $CuCl_2$ to the molecules before the optimization, as previously reported (see Table S1 for comparison with CuCl geometries).[22,41] Ligands used in the Ni-catalyzed reactions were optimized with $CuCl_2$ and NiF_2 for comparison (see Table S2). Similar to Cu(I) and Cu(II), the RMSD for the tested structures was found to be lower than 1 Å on average. All electronic features, including NBO analyses,[73] were performed on the metal-free ligand structures. Atoms for the different NBO charges (atom itself and lone-pair) were defined based on distance to the metal center. The optimized or crystal structure coordinates were used to compute the steric features and build the molecular graphs. For the steric features, both libarvo and Morfeus were used.[74–76] Features derived from the molecular graph were generated using the newly developed *Moltop* Python package. Whenever bond orders are required for a specific feature (such as the Crest flexibility index), these have to be defined explicitly. Supported bond orders currently include ones from NBO analyses, xTB, and RDkit. All *Moltop* instructions and scripts used in this study are available on Github at `https://github.com/lcmd-epfl/rafbl`. The Sklearn package was used for linear models.[77]

6 Acknowledgements

The authors thank EPFL for computational resources. This publication was created as part of NCCR Catalysis (grant number 180544), a National Centre of Competence in Research funded by the Swiss National Science Foundation. We also thank Simone Gallarati and Dr. Nieves P. Ramirez for fruitful discussions, and Prof. Durga Hari for the unreported experiments.

References

(1) 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*, 186.

(2) Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science* **2019**, *363*.

(3) Smith, A.; Keane, A.; Dumesic, J. A.; Huber, G. W.; Zavala, V. M. A machine learning framework for the analysis and prediction of catalytic activity from experimental data. *Appl. Catal., B* **2020**, *263*, 118257.

(4) Kulik, H. J.; Sigman, M. S. Advancing Discovery in Chemistry with Artificial Intelligence: From Reaction Outcomes to New Materials and Catalysts. *Acc. Chem. Res.* **2021**, *54*, 2335.

(5) Crawford, J. M.; Kingston, C.; Toste, F. D.; Sigman, M. S. Data Science Meets Physical Organic Chemistry. *Acc. Chem. Res.* **2021**, *54*, 3136.

(6) Williams, W. L.; Zeng, L.; Gensch, T.; Sigman, M. S.; Doyle, A. G.; Anslyn, E. V. The Evolution of Data-Driven Modeling in Organic Chemistry. *ACS Cent. Sci.* **2021**, *7*, 1622.
(7) Gallarati, S.; Fabregat, R.; Laplaza, R.; Bhattacharjee, S.; Wodrich, M. D.; Corminboeuf, C. Reaction-based machine learning representations for predicting the enantioselectivity of organocatalysts. *Chem. Sci.* **2021**, *12*, 6879.
(8) Żurański, A. M.; Martinez Alvarado, J. I.; Shields, B. J.; Doyle, A. G. Predicting Reaction Yields via Supervised Learning. *Acc. Chem. Res.* **2021**, *54*, 1856.
(9) Hueffel, J. A.; Sperger, T.; Funes-Ardoiz, I.; Ward, J. S.; Rissanen, K.; Schoenebeck, F. Accelerated dinuclear palladium catalyst identification through unsupervised machine learning. *Science* **2021**, *374*, 1134.
(10) Rose, B. T.; Timmerman, J. C.; Bawel, S. A.; Chin, S.; Zhang, H.; Denmark, S. E. High-Level Data Fusion Enables the Chemoinformatically Guided Discovery of Chiral Disulfonimide Catalysts for Atropselective Iodination of 2-Amino-6-arylpyridines. *J. Am. Chem. Soc.* **2022**, *144*, 22950.
(11) Laplaza, R.; Gallarati, S.; Corminboeuf, C. Genetic Optimization of Homogeneous Catalysts. *Chemistry–Methods* **2022**, *2*.
(12) Gallarati, S.; Laplaza, R.; Corminboeuf, C. Harvesting the fragment-based nature of bifunctional organocatalysts to enhance their activity. *Org. Chem. Front.* **2022**, *9*, 4041.
(13) Schwaller, P.; Vaucher, A. C.; Laplaza, R.; Bunne, C.; Krause, A.; Corminboeuf, C.; Laino, T. Machine intelligence for chemical reaction space. *WIREs Comput. Mol. Sci.* **2022**, *12*.
(14) Torres, J. A. G.; Lau, S. H.; Anchuri, P.; Stevens, J. M.; Tabora, J. E.; Li, J.; Borovika, A.; Adams, R. P.; Doyle, A. G. A Multi-Objective Active Learning Platform and Web App for Reaction Optimization. *J. Am. Chem. Soc.* **2022**, *144*, 19999.
(15) Betinol, I. O.; Kuang, Y.; Reid, J. P. Guiding Target Synthesis with Statistical Modeling Tools: A Case Study in Organocatalysis. *Org. Lett.* **2022**, *24*, 1429–1433.
(16) Dotson, J. J.; van Dijk, L.; Timmerman, J. C.; Grosslight, S.; Walroth, R. C.; Gosselin, F.; Püntener, K.; Mack, K. A.; Sigman, M. S. Data-Driven Multi-Objective Optimization Tactics for Catalytic Asymmetric Reactions Using Bisphosphine Ligands. *J. Am. Chem. Soc.* **2023**, *145*, 110.
(17) Wang, Y.-Z.; Wang, Z.-H.; Eshel, I. L.; Sun, B.; Liu, D.; Gu, Y.-C.; Milo, A.; Mei, T.-S. Nickel/biimidazole-catalyzed electrochemical enantioselective reductive cross-coupling of aryl aziridines with aryl iodides. *Nat. Commun.* **2023**, *14*, 2322.
(18) Strieth-Kalthoff, F.; Sandfort, F.; Kühnemund, M.; Schäfer, F. R.; Kuchen, H.; Glorius, F. Machine Learning for Chemical Reactivity: The Importance of Failed Experiments. *Angew. Chem. Int. Ed.* **2022**, *61*.
(19) Schleinitz, J.; Langevin, M.; Smail, Y.; Wehnert, B.; Grimaud, L.; Vuilleumier, R. Machine Learning Yield Prediction from NiCOlit, a Small-Size Literature Data Set of Nickel Catalyzed C–O Couplings. *J. Am. Chem. Soc.* **2022**, *144*, 14722.
(20) Maloney, M. P.; Coley, C. W.; Genheden, S.; Carson, N.; Helquist, P.; Norrby, P.-O.; Wiest, O. Negative Data in Data Sets for Machine Learning Training. *Org. Lett.* **2023**,
(21) Lau, S. H.; Borden, M. A.; Steiman, T. J.; Wang, L. S.; Parasram, M.; Doyle, A. G. Ni/Photoredox-Catalyzed Enantioselective Cross-Electrophile Coupling of Styrene Oxides with Aryl Iodides. *J. Am. Chem. Soc.* **2021**, *143*, 15873.
(22) Werth, J.; Sigman, M. S. Linear Regression Model Development for Analysis of Asymmetric Copper-Bisoxazoline Catalysis. *ACS Catal.* **2021**, *11*, 3916.
(23) Haas, B. C.; Goetz, A. E.; Bahamonde, A.; McWilliams, J. C.; Sigman, M. S. Predicting relative efficiency of amide bond formation using multivariate linear regression. *Proc. Natl. Acad. Sci.* **2022**, *119*.
(24) Cammarota, R. C.; Liu, W.; Bacsa, J.; Davies, H. M. L.; Sigman, M. S. Mechanistically Guided Workflow for Relating Complex Reactive Site Topologies to Catalyst Performance in C–H Functionalization Reactions. *J. Am. Chem. Soc.* **2022**, *144*, 1881.
(25) Lustosa, D. M.; Milo, A. Mechanistic Inference from Statistical Models at Different Data-Size Regimes. *ACS Catal.* **2022**, *12*, 7886.
(26) Lustosa, D. M.; Barkai, S.; Domb, I.; Milo, A. Effect of Solvents on Proline Modified at the Secondary Sphere: A Multivariate Exploration. *J. Org. Chem.* **2022**, *87*, 1850.
(27) Durand, D. J.; Fey, N. Computational Ligand Descriptors for Catalyst Design. *Chem. Rev.* **2019**, *119*, 6561.
(28) Durand, D. J.; Fey, N. Building a Toolbox for the Analysis and Prediction of Ligand and Catalyst Effects in Organometallic Catalysis. *Acc. Chem. Res.* **2021**, *54*, 837.
(29) Żurański, A. M.; Wang, J. Y.; Shields, B. J.; Doyle, A. G. Auto-QChem: an automated workflow for the generation and storage of DFT calculations for organic molecules. *React. Chem. Eng.* **2022**, *7*, 1276.

(30) Santiago, C. B.; Guo, J.-Y.; Sigman, M. S. Predictive and mechanistic multivariate linear regression models for reaction development. *Chem. Sci.* **2018**, *9*, 2398.
(31) Gensch, T.; dos Passos Gomes, G.; Friederich, P.; Peters, E.; Gaudin, T.; Pollice, R.; Jorner, K.; Nigam, A.; Lindner-D'Addario, M.; Sigman, M. S.; Aspuru-Guzik, A. A Comprehensive Discovery Platform for Organophosphorus Ligands for Catalysis. *J. Am. Chem. Soc.* **2022**, *144*, 1205.
(32) Tipping, M. E. Sparse Bayesian Learning and the Relevance Vector Machine. *J. Mach. Learn. Res.* **2001**, *1*, 211.
(33) Häse, F.; Roch, L. M.; Kreisbeck, C.; Aspuru-Guzik, A. Phoenics: A Bayesian Optimizer for Chemistry. *ACS Cent. Sci.* **2018**, *4*, 1134.
(34) Brethomé, A. V.; Paton, R. S.; Fletcher, S. P. Retooling Asymmetric Conjugate Additions for Sterically Demanding Substrates with an Iterative Data-Driven Approach. *ACS Catal.* **2019**, *9*, 7179–7187.
(35) Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G. Bayesian reaction optimization as a tool for chemical synthesis. *Nature* **2021**, *590*, 89.
(36) Jorayev, P.; Russo, D.; Tibbetts, J. D.; Schweidtmann, A. M.; Deutsch, P.; Bull, S. D.; Lapkin, A. A. Multi-objective Bayesian optimisation of a two-step synthesis of p-cymene from crude sulphate turpentine. *Chem. Eng. Sci.* **2022**, *247*, 116938.
(37) Hickman, R. J.; Aldeghi, M.; Häse, F.; Aspuru-Guzik, A. Bayesian optimization with known experimental and design constraints for chemistry applications. *Digital Discovery* **2022**, *1*, 732.
(38) McCullough, K. E.; King, D. S.; Chheda, S. P.; Ferrandon, M. S.; Goetjen, T. A.; Syed, Z. H.; Graham, T. R.; Washton, N. M.; Farha, O. K.; Gagliardi, L.; Delferro, M. High-Throughput Experimentation, Theoretical Modeling, and Human Intuition: Lessons Learned in Metal–Organic-Framework-Supported Catalyst Design. *ACS Cent. Sci.* **2023**,
(39) Tom, G.; Hickman, R. J.; Zinzuwadia, A.; Mohajeri, A.; Sanchez-Lengeling, B.; Aspuru-Guzik, A. Calibration and generalizability of probabilistic models on low-data chemical datasets with DIONYSUS. *Digital Discovery* **2023**,
(40) Hari, D. P.; Waser, J. Enantioselective Copper-Catalyzed Oxy-Alkynylation of Diazo Compounds. *J. Am. Chem. Soc.* **2017**, *139*, 8420.
(41) Aguado-Ullate, S.; Urbano-Cuadrado, M.; Villalba, I.; Pires, E.; García, J. I.; Bo, C.; Carbó, J. J. Predicting the Enantioselectivity of the Copper-Catalysed Cyclopropanation of Alkenes by Using Quantitative Quadrant-Diagram Representations of the Catalysts. *Chem. Eur. J.* **2012**, *18*, 14026.
(42) Evans, D. A.; Lectka, T.; Miller, S. J. Bis(imine)-copper(II) complexes as chiral lewis acid catalysts for the Diels-Alder reaction. *Tetrahedron Lett.* **1993**, *34*, 7027.
(43) Davies, I. W.; Gerena, L.; Castonguay, L.; Senanayake, C. H.; Larsen, R. D.; Verhoeven, T. R.; Reider, P. J. The influence of ligand bite angle on the enantioselectivity of copper(II)-catalysed Diels-Alder reactions. *Chem. Commun.* **1996**, 1753.
(44) Ghosh, A. K.; Mathivanan, P.; Cappiello, J. Conformationally constrained bis(oxazoline) derived chiral catalyst: A highly effective enantioselective Diels-Alder reaction. *Tetrahedron Lett.* **1996**, *37*, 3815.
(45) Davies, I. W.; Gerena, L.; Cai, D.; Larsen, R. D.; Verhoeven, T. R.; Reider, P. J. A conformational toolbox of oxazoline ligands. *Tetrahedron Lett.* **1997**, *38*, 1145.
(46) Evans, D. A.; Miller, S. J.; Lectka, T.; von Matt, P. Chiral Bis(oxazoline)copper(II) Complexes as Lewis Acid Catalysts for the Enantioselective Diels-Alder Reaction. *J. Am. Chem. Soc.* **1999**, *121*, 7559.
(47) Kanemasa, S.; Adachi, K.; Yamamoto, H.; Wada, E. Bisoxazoline and Bioxazoline Chiral Ligands Bearing 4-Diphenylmethyl Shielding Substituents. Diels-Alder Reaction of Cyclopentadiene with 3-Acryloyl-2-oxazolidinone Catalyzed by the Aqua Nickel(II) Complex. *Bull. Chem. Soc. Jpn.* **2000**, *73*, 681.
(48) O'Leary, P.; Krosveld, N. P.; De Jong, K. P.; van Koten, G.; Klein Gebbink, R. J. Facile and rapid immobilization of copper(II) bis(oxazoline) catalysts on silica: application to Diels-Alder reactions, recycling, and unexpected effects on enantioselectivity. *Tetrahedron Lett.* **2004**, *45*, 3177.
(49) Groom, C. R.; Allen, F. H. The Cambridge Structural Database in Retrospect and Prospect. *Angew. Chem. Int. Ed.* **2014**, *53*, 662.
(50) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta. Crystallogr. B. Struct. Sci. Cryst. Eng. Mater.* **2016**, *72*, 171.
(51) Vela, S.; Laplaza, R.; Cho, Y.; Corminboeuf, C. cell2mol: encoding chemistry to interpret crystallographic data. *npj Comput. Mater.* **2022**, *8*, 188.
(52) Poater, A.; Ragone, F.; Mariz, R.; Dorta, R.; Cavallo, L. Comparing the Enantioselective Power of Steric and Electrostatic Effects

in Transition-Metal-Catalyzed Asymmetric Synthesis. *Chem. Eur. J.* **2010**, *16*, 14348.
(53) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17.
(54) Hosoya, H. Topological Index. A Newly Proposed Quantity Characterizing the Topological Nature of Structural Isomers of Saturated Hydrocarbons. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332.
(55) Balaban, A. T. Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **1982**, *89*, 399.
(56) Kier, L. B. A Shape Index from Molecular Graphs. *Quant. Struct.-Act. Relat.* **1985**, *4*, 109.
(57) Kier, L. B. Shape Indexes of Orders One and Three from Molecular Graphs. *Quant. Struct.-Act. Relat.* **1986**, *5*, 1.
(58) Kier, L. B. Distinguishing Atom Differences in a Molecular Graph Shape Index. *Quant. Struct.-Act. Relat.* **1986**, *5*, 7.
(59) Kier, L. B. An Index of Molecular Flexibility from Kappa Shape Attributes. *Quant. Struct.-Act. Relat.* **1989**, *8*, 221.
(60) Hall, L. H.; Kier, L. B. Determination of Topological Equivalence in Molecular Graphs from the Topological State. *Quant. Struct.-Act. Relat.* **1990**, *9*, 115.
(61) Kier, L. B.; Hall, L. H. A Differential Molecular Connectivity Index. *Quant. Struct.-Act. Relat.* **1991**, *10*, 134.
(62) Hall, L. H.; Mohney, B.; Kier, L. B. The Electrotopological State: An Atom Index for QSAR. *Quant. Struct.-Act. Relat.* **1991**, *10*, 43.
(63) Caron, G.; Digiesi, V.; Solaro, S.; Ermondi, G. Flexibility in early drug discovery: focus on the beyond-Rule-of-5 chemical space. *Drug Discov. Today* **2020**, *25*, 621.
(64) Fisanick, W.; Cross, K. P.; Rusinko, A. Characteristics of computer-generated 3D and related molecular property data for CAS registry substances. *Tetrahedron Comput. Methodol.* **1990**, *3*, 635.
(65) Pracht, P.; Bohle, F.; Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169.
(66) Jones, D. R.; Schonlau, M.; Welch, W. J. Efficient Global Optimization of Expensive Black-Box Functions. *J. Glob. Optim.* **1998**, *13*, 455.
(67) Shao, Q.; Wu, K.; Zhuang, Z.; Qian, S.; Yu, J.-Q. From Pd(OAc) 2 to Chiral Catalysts: The Discovery and Development of Bifunctional Mono-N-Protected Amino Acid Ligands for Diverse C–H Functionalization Reactions. *Acc. Chem. Res.* **2020**, *53*, 833.
(68) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **1999**, *110*, 6158.
(69) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297.
(70) Frisch, M. J. et al. Gaussian 16 Revision C.01. 2016; Gaussian Inc. Wallingford CT.
(71) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652.
(72) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 33.
(73) Glendening, E. D.; Reed, A. E.; Carpenter, J. E.; Weinhold., F. NBO Version 3.1. 2001.
(74) Buša, J.; Džurina, J.; Hayryan, E.; Hayryan, S.; Hu, C.-K.; Plavka, J.; Pokorný, I.; Skřivánek, J.; Wu, M.-C. ARVO: A Fortran package for computing the solvent accessible surface area and the excluded volume of overlapping spheres via analytic equations. *Comput. Phys. Commun.* **2005**, *165*, 59.
(75) Jorner, K. MORFEUS: molecular features for machine learning. https://kjelljorner.github.io/morfeus/, (accessed 2023-05-15).
(76) Laplaza, R. libarvo: library to compute molecular surfaces and volumes. https://github.com/rlaplaza/libarvo, (accessed 2023-05-15).
(77) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825.