From Data to Chemistry: Revealing Causality and Reaction Coordinates through Interpretable Machine Learning in Supramolecular Transition Metal Catalysis

R. A. Talmazan[1], J. Gamper[2], I. Castillo[3], T.S. Hofer*[2], M. Podewitz*[1]

[1] Institute of Materials Chemistry, TU Wien, Getreidemarkt 9, A-1060 Wien (Austria), maren.podewitz@tuwien.ac.at

[2] Institute of Inorganic and Theoretical Chemistry, Leopold Franzens University of Innsbruck, Innrain 80/82, 6020, Innsbruck (Austria), t.hofer@uibk.ac.at

[3] Instituto de Química, Universidad Nacional Autónoma de México, Ciudad Universitaria, 04510, Ciudad de México (México)

Abstract

Supramolecular transition metal catalysts with tailored reaction environments allow for the usage of abundant 3d metals as catalytic centres, leading to more sustainable chemical processes. However, such catalysts are large and flexible systems with intricate interactions, resulting in complex reaction coordinates. To capture their dynamic nature, we developed a broadly applicable, high-throughput workflow, leveraging quantum mechanics/molecular mechanics (QM/MM) molecular dynamics in explicit solvent, to investigate a Cu(I)-calix[8]arene catalysed C-N coupling reaction. The system complexity and high amount of data generated from sampling the reaction require automated analyses. To identify and quantify the reaction coordinate from noisy simulation trajectories, we applied interpretable machine learning techniques (Lasso, Random Forest, Logistic Regression) in a consensus model, alongside dimensionality reduction methods (PCA, LDA, tICA). Leveraging a Granger Causality model, we go beyond the traditional view of a reaction coordinate, by defining it as a sequence of molecular motions that led up to the reaction.

Introduction

Catalysis plays a pivotal role in chemistry. By lowering reaction barriers, catalysts facilitate chemical transformation under mild conditions, contributing to sustainable and economic processes. Nature has perfected this principle in enzymes, through precise control of the environment surrounding the catalytic centre and substrate.[1] In an effort to mimic the tight control over the environment, the field of supramolecular catalysis chemistry has emerged.[2–6] For example, the use of a macrocycle allows the substitution of rare earth metals with more abundant counterparts, while maintaining high catalytic performance.[7,8]

Complimentary to experimental advances, computational chemistry has played a key role in the design and understanding of catalytic systems, by elucidating reaction mechanisms.[9–12] Despite considerable efforts, quantum chemistry is limited in its predictive abilities.[13] As these systems grow in complexity, there is a need to improve not only on the underlying electronic structure theory used to study them, but also the chemical model that describes the catalytic system in its environment.[14] Often, the errors introduced by a too simplistic chemical model exceed those arising from the use of an approximate theoretical methodology such as DFT.[13,15] Consequently, the goal is to create a chemical model which is a "digital twin" of the reaction flask, where the in-silico procedure can fully replicate experimental conditions, meaning a catalyst in explicit solvent, at finite temperature and pressure. As a full ab-initio quantum chemical (QC) level is not feasible, a tailored multiscale strategy has to be developed, accounting for conformational flexibility, explicit solvation effects and the dynamic nature of chemical reactions.

While the majority of computational mechanistic studies are performed with a single structure, with conformer search recently gaining popularity[14,16,17] thanks to easily accessible tools,[18,19] a transition to structure ensembles provides a more complete picture.[14,17]

Another crucial aspect which needs to be considered is solvation. While widely used implicit solvation models provide remarkable performance,[20–22] they cannot describe explicit interactions between solute and solvent. In combination with dispersion corrections and a small basis set, they often favour very compact structures with many intramolecular bonds[7,19] – a poor representation of a solvated system. It is therefore crucial to include explicit solvation in any realistic model, ideally via a full condensed phase calculation or through microsolvation.[23,24]

Describing catalyst ensembles in explicit solvent is a significant improvement of the chemical model, yet it is important to recognize that catalysts are, by nature, dynamic entities. By tracking the motion of the nuclei during a reaction, we may observe a different reaction pathway,[25,26] The dynamic effects are well documented for organic molecules in implicit solvents,[27] yet rarely investigated in transition metal catalysis.[25,27,28]

While the inclusion of dynamics in mechanistic studies makes a model more realistic, it adds complexity and significantly increases computational cost. This limitation can be addressed by relying on multiscale methods such as quantum mechanics/molecular mechanics (QM/MM) which describe the catalytic centre and substrates at a QM level, while the surrounding environment is treated with MM.[29–31] This hybrid methodology can be combined with MD to investigate the dynamic behaviour. A transition to QM/MM MD in explicit solvent allows for sampling timescales to the order of nanoseconds, magnitudes

higher than in a pure QM approach. Due to the rare nature of the reactions, repeated sampling is a requirement for statistically relevant information regarding energy barriers and structural information.

While the setup of such multiscale methods is a challenge in itself, a vast amount of data is generated from the simulations. As it is not feasible to interpret them by observation only, an automatic way of processing reaction trajectories is needed. Machine learning approaches can be used to evaluate the data and extract condensed results which contain insight into the reaction coordinate.[28] While neural network approaches[32] have garnered a lot of attention recently, they are not very well suited for data analysis approaches where understanding of the process is required, due to their black box nature, which makes them very difficult to evaluate.[33] Instead, interpretable machine learning techniques, such as decision trees, random forests or logistic regression, offer good performance in extracting relevant information from large datasets and presenting them in easily understandable ways. In addition, dimensionality reduction techniques—Principal Component Analysis[34] (PCA) or time-lagged Independent Component Analysis[35] (tICA) effectively detect combined coordinates from the trajectories, revealing the key motions of a system. Yet these dimensionality reduction techniques have almost exclusively been applied to biomolecules[36–38] with few exceptions.[39,40] A combination of aforementioned methods offers great promise to detect a cumulative reaction coordinate from a multitude of independent trajectories, providing chemical insight into the mechanism and reactivity of the system. To the best of our knowledge these combined methods have not been applied to study reaction mechanisms in explicit solvent, let alone large supramolecular transition metal catalysts.

Another aspect that has until now been neglected in chemistry is causality. While the concept is widespread across various scientific domains[41]—ranging from economics[42–46] and climate research[47–51] to biology[52,53] and medical studies[54–56]—it remains surprisingly absent in the field of chemistry. Although a handful of precedents in biomolecular simulations exists,[57–59] it has not been explored to study chemical reactions, not to mention transition-metal catalysis. As trajectories are essentially discrete time series, containing the various degrees of freedom of the system, causality can be statistically inferred from the analysis of these trajectories. Consequently, the reaction coordinate can be decomposed into a sequence of motions leading up to the reaction, exposing the intricate interplay of functional groups of the system, offering an unprecedented view of reactivity.

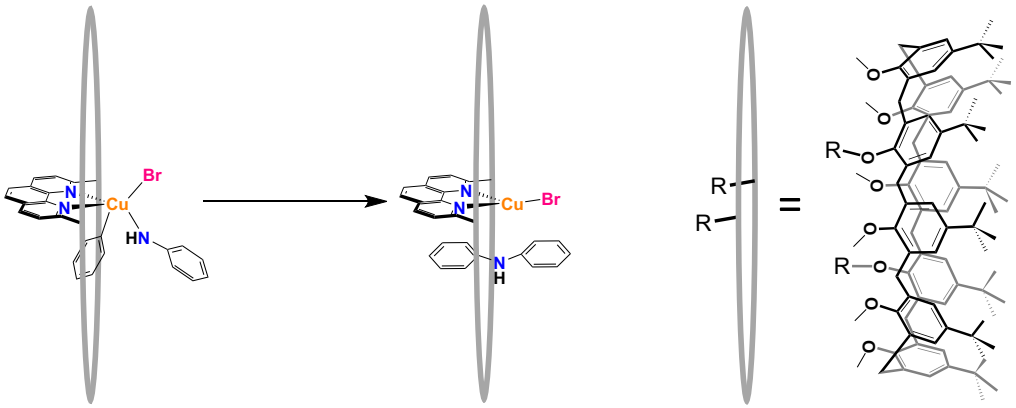

Figure 1. The supramolecular system calix[8]arene-based [Cu(C_8PhenMe$_6$)I] catalyses the C-N coupling reaction of phenyl bromide and aniline.

A supramolecular catalyst that has shown remarkable catalytic efficacy for C-N coupling is the the Cu(I)-1,5-(2,9-dimethyl-1,10-phenanthroyl)-2,3,4,6,7,8-hexamethyl-*p*-tert-butylcalix[8]arene system.[7] The macrocyclic ligand allows for usage of earth abundant metals, such as Cu, an essential step towards more sustainable chemical processes.[60–65] The investigated system necessitates explicit solvation for accurate results, as implicit solvation models lead to a collapse of the macrocyclic cage, which compromises catalytic activity.[7] To account for the conformational flexibility of the supramolecular cage[7,66] and the dynamic nature of the system as a whole, a dynamic ensemble-based approach is required to study the reaction. While the mechanism of this catalysts was established to be a sequence of oxidative addition / reductive elimination,[7] the dynamic effects of the system, in particular the contribution of the cage, and the influence of explicit solvent molecules are completely unknown. Consequently, we developed a multiscale QM/MM MD approach to understand the bond formation dynamics of the C-N coupling step with the Cu-calix[8]arene catalyst in explicit chloroform. By relying on the GFN2-xTB[67] method to describe the QM part, we could achieve massive sampling, resulting in 152 individual reaction trajectories. To extract chemically relevant information from these data, we employed supervised and unsupervised interpretable machine learning dimensionality reduction models, in order to identify the cumulative reaction coordinate and to detect critical movements in the structure. A consensus approach combining individual machine learning techniques improved the performance, while random forest models and decision rules allowed us to quantify the reaction coordinate. Finally, we employed the statistical Granger Causality analysis model[68,69] to decompose the reaction coordinate into a sequence of individual consecutive movements. This works serves as a widely applicable template for any mechanistic investigations, revealing and quantifying complex reaction coordinates, alongside causal effects derived from the individual movements leading up to the reaction.

Results

We obtained 152 QM/MM MD reaction trajectories for the C-N coupling step. Out of these, 142 reacted within 20 ps of simulation time. From these trajectories we evaluated the reaction energy and labelled the structural data accordingly as educt, transition state, or product – resulting in three ensembles. We then used this information to identify the reaction coordinate, quantify it and identify a sequence of movements leading to the reaction.

Reaction Energetics Analysis

We analysed 142 simulations where a reaction occurred to extract insights about the C-N coupling process. As the reaction happened spontaneously during the simulations, the energy profile can be obtained directly (SI Figure S2 A) and served as indicator to label the three distinct states, educt, transition state, and product. The reaction energy was obtained by averaging the ensembles of the educt and product states and amounted to -212 ± 25 kJ $mol^{-1.}$ A sigmoid fit through the smoothened energy profile of each simulation (see SI, Figure S2 B and C) allowed identification and the calculation of the energy barrier to be 13 $\pm$ 9 kJ mol^{-1}. These GFN2-xTB reaction energies are in excellent agreement with full DFT data, obtained with PBE0/def2-SVP/D3 (see SI, Table S2).

Extracting Chemical Information from Structural Data

To obtain information about the changes in chemical structure from the reaction trajectories with a total of over 1.5 million frames, we resorted to interpretable machine learning approaches.

Determination of a suitable coordinate system

A standard method to extract reaction coordinates from trajectories, either in biomolecular or reaction dynamics studies, is principal component analysis (PCA) in cartesian coordinate space. However, this approach proved unsuccessful for the Cu-calix[8]arene catalyst due to the difficult to properly align this highly flexible system not showing any separation between the three states (Figure S3).

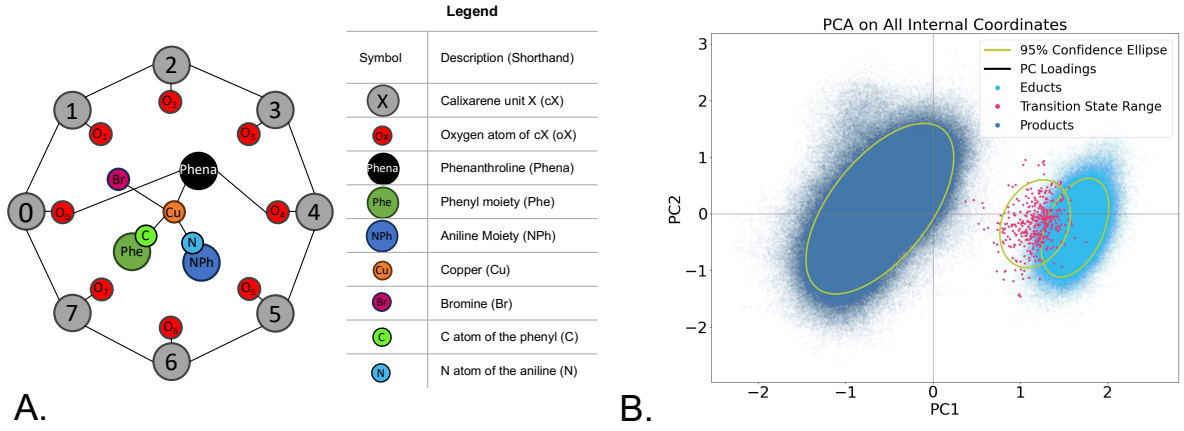

Figure 2. A. Schematic depiction of the calixarene, showing the centres of mass used for the calculation of the reduced set of internal coordinates. B. PCA analysis performed on the reduced internal coordinates.

To achieve good separation between the three states, educt, TS, and product, we developed a reduced internal coordinate description of the system (see Figure 2A) to minimize the noise from highly correlated coordinates.[82] Hence, we described rigid fragments, such as the individual calixarene units (cX), the phenanthroline bridge (Phena), the phenyl (Phe) and aniline (NPh) moieties by their respective centres of mass (Figure 2A). An overview over the distribution of the internal coordinates can be found in This internal coordinate set nicely separates educts and products in the PCA space (Figure 2B), but still shows overlap between educts and transition states. Analysing the loadings of the principal components (see SI Figure S5), we can see that the main contributions belong to the C-N, NPh-Phe, Cu-N distances, as well as to the Phena-Cu-Br angle, describing changes in the coordination at the Cu centre as the product is formed.

A second less common method, to reveal the reaction coordinates in biomolecular studies, is the time-lagged independent Component Analysis (tICA). While PCA focuses on the largest variance in the dataset, which generally corresponds to fast molecular motion, tICA can be used to separate and extract the internal coordinates which exhibit the strongest time-correlations for a chosen lag time, thus revealing slow movements in the system.

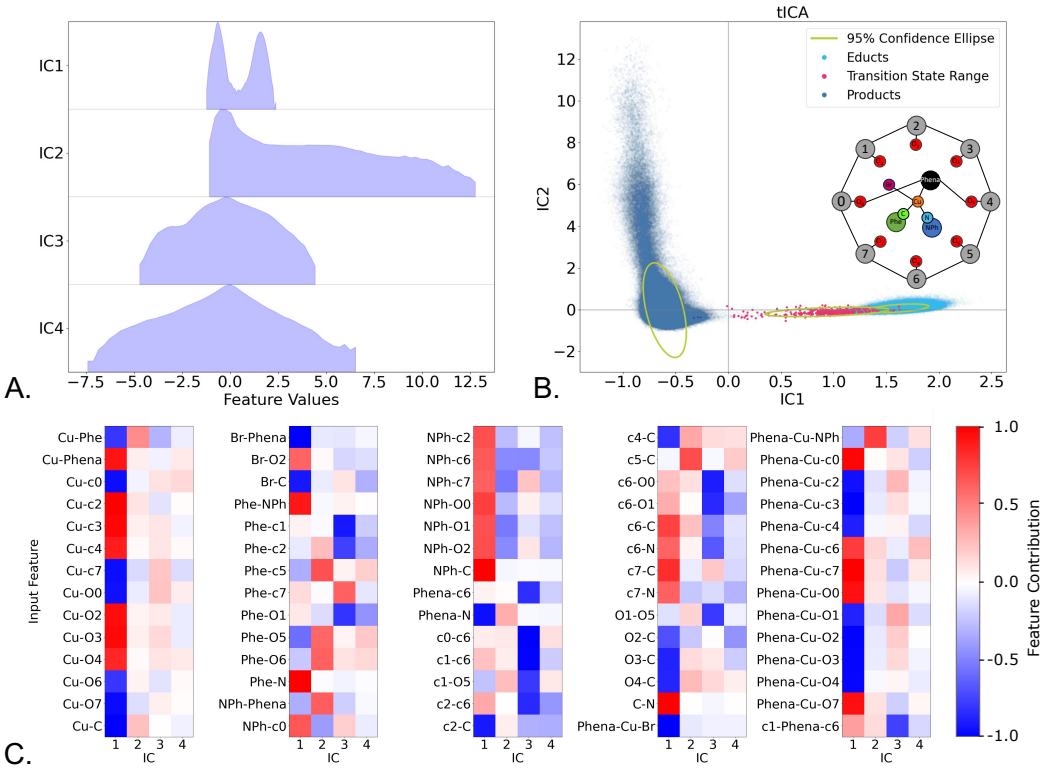

Figure 3. Time-Lagged Independent Component Analysis of the MD trajectories, using a lag time of 20 fs. A. Distribution of the structural ensembles over each of the four independent components. B. Projection of the ensembles in the space of independent components 1 and 2, with the ensemble colouring performed a posteriori. C. Normalized internal coordinate (feature) contribution to each independent component.

The tICA performed on the reduced internal coordinates set, resulted in a good separation of the product and educt states, mainly across the first independent component (IC1), as shown in Figure 3A, yet the transition state ensemble cannot be fully separated from the educts. When taking into consideration further ICs (Figure 3A and 3B), we observe a broad distribution of the product ensemble, indicating significant conformational flexibility. The contributions to ICs can be traced down, by relating the contribution strength (Figure 3C) to the degree of freedom it corresponds to (Figure 3B insert). A positive contribution (coloured in red) means that the respective feature increases as the values of the IC increases, while a negative contribution (coloured in blue) means the feature increases as the IC values decrease. The absolute value of a contribution (colour intensity) represents the importance of the feature in defining the IC.

IC1 reveals the changes at the reaction centre (see 1st column, Figure 3C), related the Cu adopting a planar configuration upon product formation. Of particular interest is the strong contributions of the distance of NPh to the c0, c2, c6 and c7 (2nd column bottom/3rd column top) and their similarity to the C-N distance (4th column bottom). Additionally, a translation of the calixarene units can be inferred, when looking at the changes in the Phena-Cu-cX and Phena-Cu-oX angles (5th column). Their similar contribution values indicate a translation, rather than a rotation. This indicates that the distance between the phenyl part of the product and the calixarene units decrease as the reaction proceeds,

which is indicative of π-π stacking between the product and calixarene cage. IC2 acts to separate various conformers within the product ensemble.

Improved Reaction Coordinate Detection Through Supervised Methods

We intended to further improve the separation of the three states in the PCA by utilizing the labelling of the data, indicating each structure as educt, product or transition state. Using this information, we trained a model that maximized the separation between the three ensembles and simultaneously reduced the number of internal coordinates (features) to those that contribute the most to the separation. The latter is referred to as feature elimination. There are various ways of achieving this goal and we tested a selection of them in PCA and Linear Discriminant Analysis (LDA) dimensionality reduction approaches (see Figure 4). While PCAs depend on the technique used to eliminate the features (Figure 4A-C and SI Figure S4), LDA is independent on the feature selection and consistently yields excellent separation (Figure 4D-F and SI Figure S4). While LDA results in compact distinct ensembles, PCA also achieves separation within the ensembles depending on the feature elimination approach that was used.

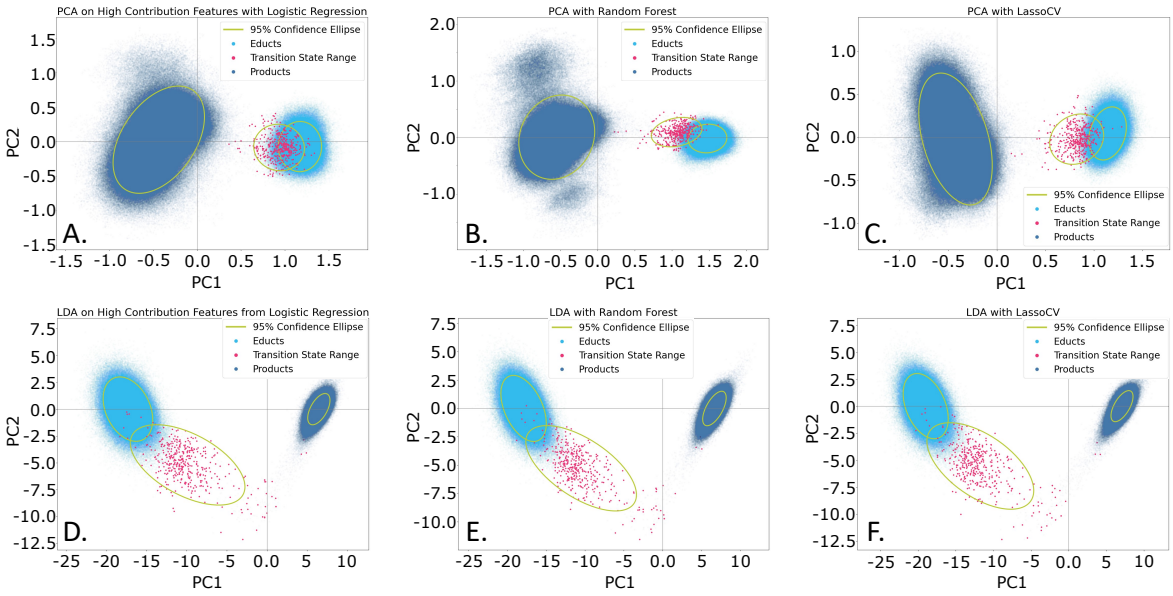

Figure 4. Dimensionality reduction performance with various feature reduction methods: A. PCA with Logistic Regression with mean-based cutoff; B. PCA with Random Forest with mean-based cutoff; C. PCA with LassoCV; D. LDA with Recursive feature elimination with cross validation using a logistic regression classifier; E. LDA with Recursive feature elimination with cross validation using a random forest classifier; F. LDA with LassoCV.

Amongst the top contributors are the change in the Phena-Cu-Br angle, as well as the C-N distance. Secondary features such as the distance between the product and phenanthroline bridge also play an important role for the LR classifier. The RF identifies the distances between the Cu and several calixarene units as being important, as well as the distance between the aniline and phenanthroline. The separation with the LassoCV approach is superior to that obtained with the logistic regression, as the educt and transition state ensembles are better separated. However, the there is only one ensemble visible for the product state. The principal contributors are the Cu-N, Cu-Phe, Cu-C, Phe-O3 and Phe-O2 distances.

It can be observed that the various feature selection methods show variability and differences in the selected internal coordinates and separation performance.

Revealing Causality in the Reaction Coordinate – Information from Consensus Model

In order to combat the variability associated with the various feature selection methods, we switched to a consensus model, which integrates the features highlighted by all supervised ML methods. The model identifies 49 features (internal coordinates) (Figure 5A) of high importance. We performed an additional PCA (Figure 5B) on the consensus features. Additional insight can be obtained by performing hierarchical clustering on the consensus features (see SI) and calculating the correlations between them, resulting in a clustermap (Figure 5C), which reorders the internal coordinates according to their correlation to each other.

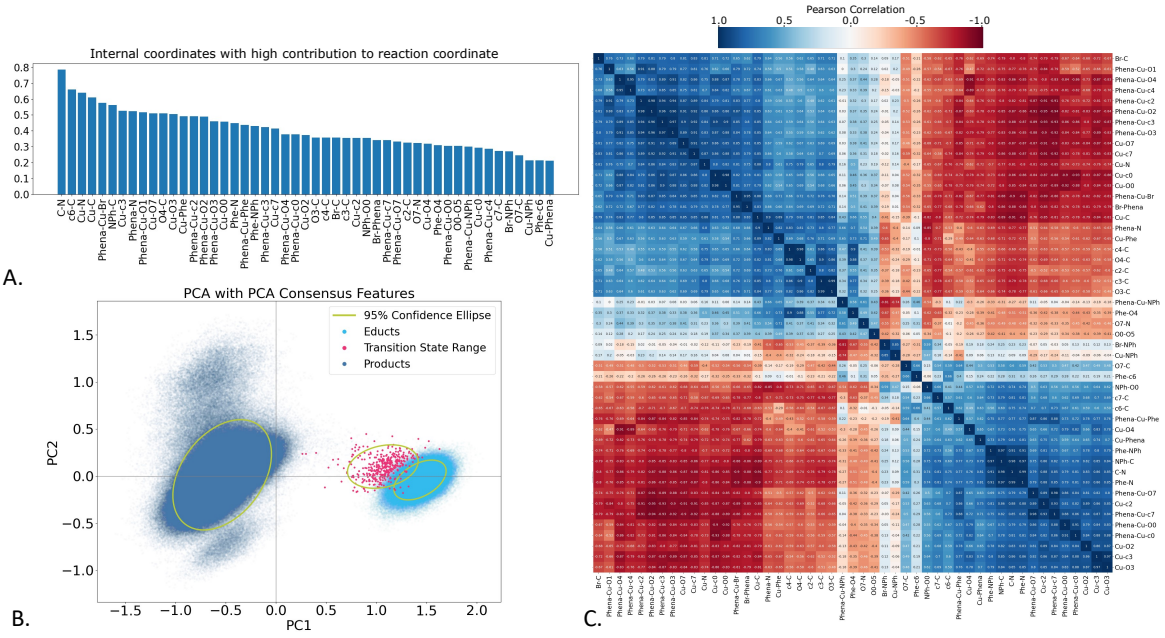

Figure 5. Analysis of internal coordinates deemed as high importance by the consensus model. A. Internal coordinates and their contribution; B. PCA of the consensus features; C. Clustermap of the internal coordinates Pearson correlations.

Judging the importance of the consensus model features (Figure 5A), we see that the C-N bond distance is most important (in agreement with chemical intuition), alongside several distances and angles corresponding to the reaction centre. Notably, the c6-C distance is also deemed highly important, which indicates that the cage indeed plays a role in defining the reaction coordinate. This small set of internal coordinates yields almost perfect separation of the three states as well as distinguishing the product conformations (Figure 5B), thereby outperforming any of the individual feature selection methods.

When looking at the clustermap plot (Figure 5C), we see two distinct regions of the complex with opposite correlation. Upon closer inspection of the individual features and their clustering, we can see the internal coordinates corresponding to the each calixarene unit group together. Coordinates that show positive correlation with the C-N bond (depicted in blue) decrease in value as the reaction proceeds, while those showing a negative correlation (depicted in red) increase during the reaction. Hence, the distances between Phe and NPh (comprising the product) and calixarene units 6 and 7

decrease, indicating π-π stacking interactions (depicted in blue), whereas C moves away from calixarene units 2-4 (coloured in red).

While the correlation analysis shows which movements take place in a correlated fashion, it is also interesting to evaluate the causality of these movements and how they propagate through the system. As tICA, a time-lag based method, revealed new information regarding the cage movement and product interaction, we decided to apply the Granger causality model, also time-lag based, to evaluate the reaction trajectories.

Using the coordinates resulting from the consensus model, we can construct a propagation cluster map of our system (Figure 6), which allows us to observe how the internal coordinates influence the C-N bond formation. This combines the hierarchical clustering of the consensus features with the results of the Granger causality model. Thus, we can infer the causality between coordinates and groups of coordinates.

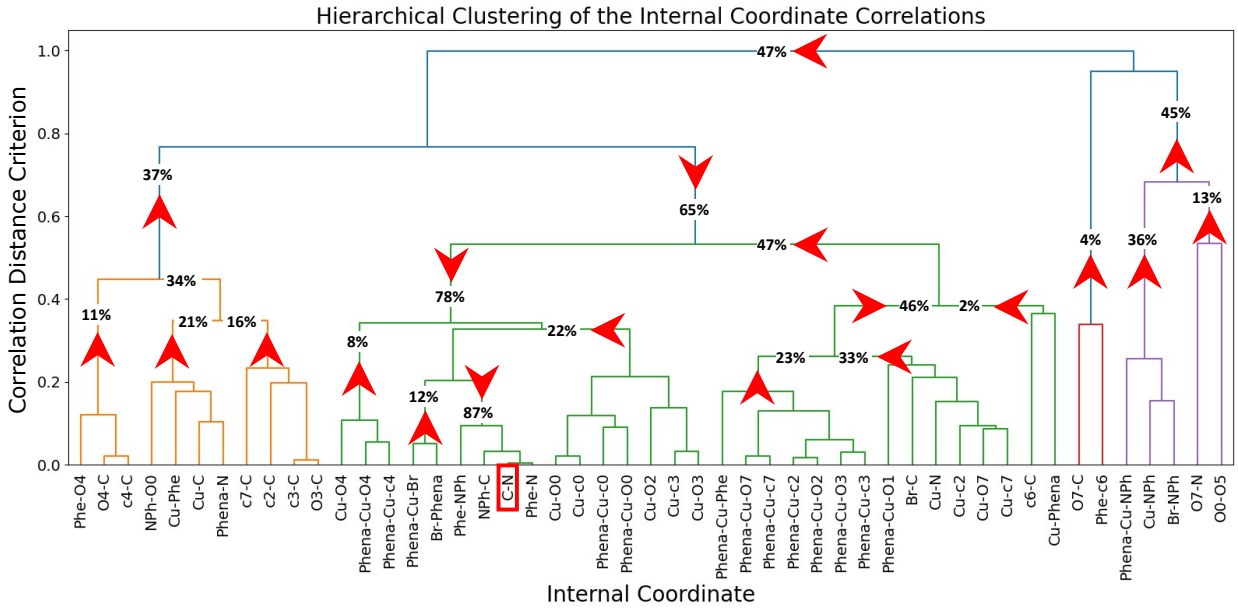

Figure 6. Propagation cluster map of the system, leading to the C-N bond formation. The numbers represent the total causality effects uncovered by the respective coordinate(s) and the red arrows indicate the direction of causality.

It allows us to examine the effects leading to the change in the C-N distance, identified as the most important feature in the consensus model. Causality was evaluated between each of the 49 consensus features and then examined with regards to the C-N bond distance. Thus, Figure 6 can be read as a map of movements leading to the coupling reaction, by choosing a starting point and following the arrows towards the highlighted C-N feature. The numbers next to the arrows indicate the number of trajectories where the C-N bond formation can be attributed to the respective features. In 52 trajectories out of 139, the coupling is triggered by a move in the orange cluster, consisting of the calixarene 2,3,4 and 7 to C distances as well as the proximity of Phe to the Cu. When we include information about the Phena-Cu-Br angle, alongside Cu and Br distances to NPh, from the purple cluster, as well as information from the red cluster, we can infer causality in 90 of the trajectories. The green cluster consists mostly of information related to the position of the calixarenes around the reaction centre. When we combine all information together, we can infer that the C-N coupling is caused by

movements in the calixarene cage, alongside the Cu coordination change in 121 trajectories. Naturally the causality analysis can be expanded to include the sources of changes in each feature, however this quickly becomes a very complex multidimensional problem.

Quantification of Reaction Coordinates

While the previous approaches identified the relevant internal degrees of freedom which determine the reaction coordinate, as a next step, we sought to quantify it, that is, to define ranges which separate the data into the three ensembles. A method to achieve this, is the decision tree, which splits ensembles by setting cut-offs based on the coordinates that show the largest distribution differences between classes.

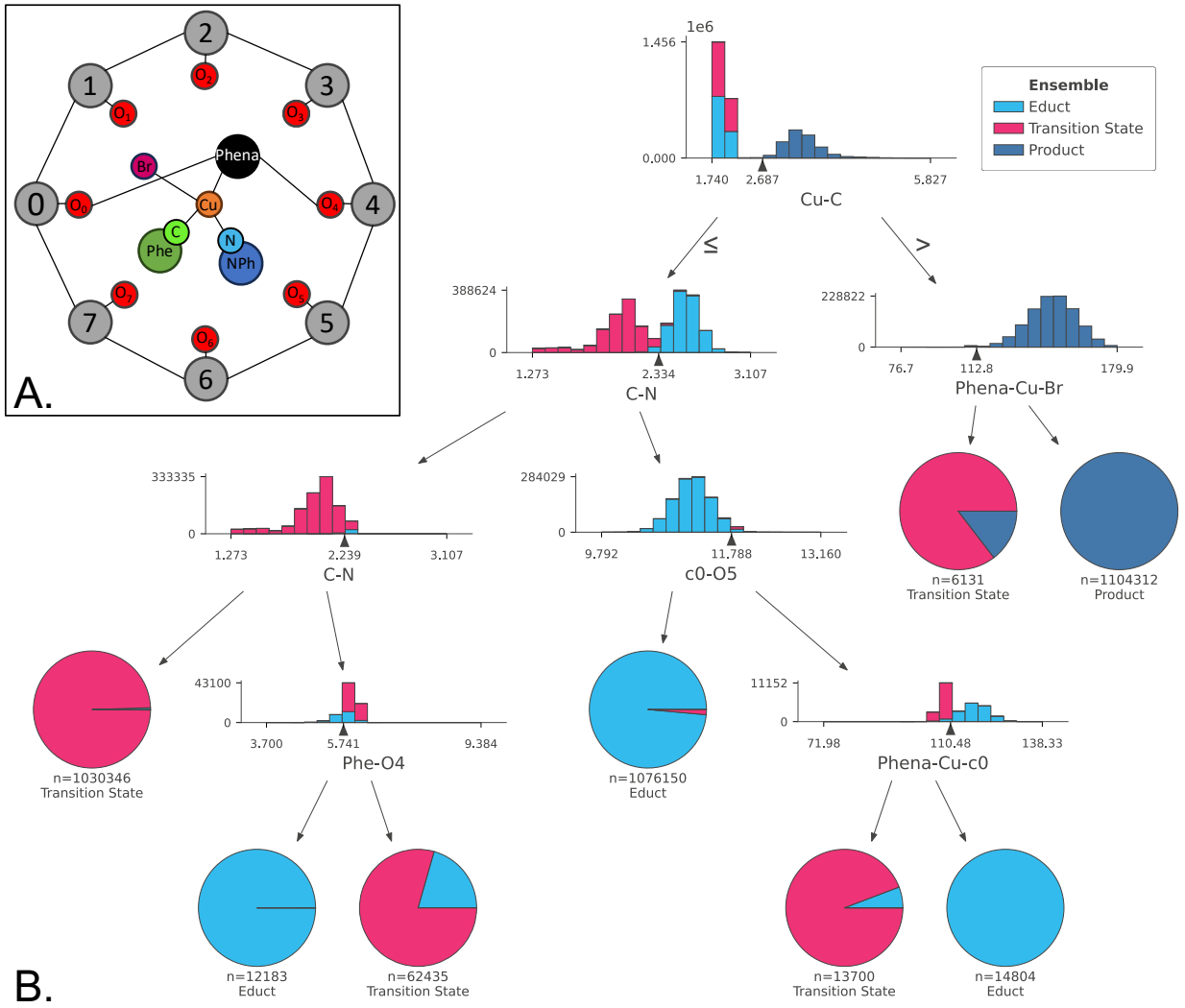

Figure 7. Decision Tree classifier used to interpret the differences between the 3 structure groups. A. Model of the calixarene with the reduced centres of mass. B. Decision tree trained on the balanced classes.

The decision tree in Figure 7 has been trained on the whole dataset, with balanced weighting given to the groups via oversampling, in order to remove biasing against the transition state ensemble, which contains significantly fewer structures. The results of an unbalanced tree can be found in Figure S7.

Analysing the tree, we can see that the Cu-C distance plays a key role in the splitting of the educts and products, with the majority of transition states being grouped with the educt class. The remaining transition state contamination of the product ensemble can be separated by taking the angle determined by the Phenanthroline bridge, Cu and Br atoms into account (Phena-Cu-Br), where values below 112.8° are indicative of a transition state.

To differentiate between the transition states combined with educts, the C-N bond represents an effective metric, where values higher than 2.33 Å indicate an educt, while distances below 2.24 Å indicate a transition state. The educt states that do exhibit a C-N bond distance similar to that of the TS can be identified by a smaller phenyl to calixarene unit 4 oxygen atom distance (Phe-O4). When transition states exhibit a C-N bond distance over 2.33 Å, we can perform a selection based on the calixarene c0 and calixarene O5 distance (which should be greater than 11.79 Å) and the angle defined by the Phenanthroline bridge, Cu and c0, where values below 110.5 indicate a TS.

The reliability of a single decision tree-based approach can often be improved by utilizing a random forest classifier and averaging the results. In general, this improves accuracy, but reduces the interpretability. To overcome this limitation decision rules can be used, providing a semantic understanding of the RF classifier. We used 30 decision trees, each trained on a subset of the data, to yield the random forest. When applied to our dataset, this method provides rules for each of the three classes, as seen in Figure 8, below. A complete diagram of the decision rules can be found in the SI, Figure S8.

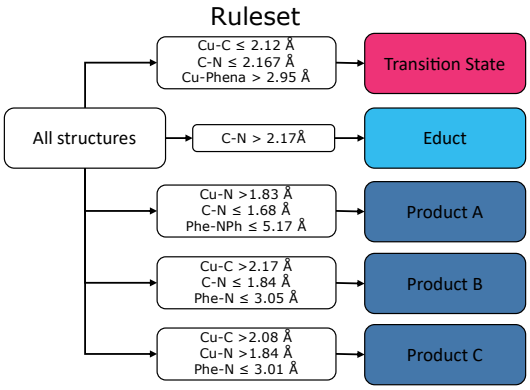

Figure 8. Graphical depiction of the decision rules derived from the random forest approach, Product A, B, C refers to different conformer ensembles within the product category.

Notably, the decision rules approach identifies three distinct rulesets for defining a product. This correlates with the findings from the PCA consensus analysis and indicates the existence of various conformers in the product ensemble.

Discussion

Taking advantage of semi-empirical quantum chemistry methods allowed us to obtain a large sample set of the reductive elimination step of the C-N coupling reaction with a Cu-calix[8]arene catalyst, using a hybrid QM/MM molecular dynamics approach. Since the amount of generated data, alongside the very high dimensionality, makes the reaction difficult to interpret by visual analysis, statistical methods and machine learning techniques were used to extract chemically relevant information from the dataset.

To investigate the C-N coupling step we generated 152 reaction trajectories, amounting to more than 3 ns of total sampling time. We observed a convergence of the PCA analysis with increased sampling (see SI, Figure S9), we assumed the total simulation time to be sufficient. As we observed spontaneous C-N coupling in 142 of the 152 trajectories, we could directly analyse these unbiased simulations.

By analysing the reaction energy profiles, we were able to quantify the reaction energy, as well as the reaction barrier, including uncertainty values. While the reaction barrier computed with a static modelling approach is within 10 kJ mol^{-1} (13 ± 9 kJ mol^{-1} vs 23 kJ mol^{-1}), the reaction energy exhibits some differences, -212 ± 25 kJ mol^{-1} vs -255 kJ mol^{-1}.[7] The remaining differences in the reaction energy can be attributed to the different chemical models that were used: in the static model, the investigated structure was obtained at 0 K, representing the bottom of the potential energy surface, while the dynamic study not only averages over all conformations but also considers thermal energy, hence, the structures are *not* 0 K structures and hence no minima on the potential energy surface. Alongside energetics, the reaction profile allowed for the categorization of the structures into three ensembles, namely educt, transition state and product. This was a key step in improving the reaction coordinate detection, as it allowed for the use of supervised learning methods to reduce the coordinate space.

To yield any separation between the three states in the PCA, we had to transform cartesian coordinates to a reduced set of internal coordinates, to minimize the number of highly correlated coordinates in the data set, thus reducing noise. While product states could be separated from educts and transition states, the latter two still showed overlap. In contrast, standard PCA on the cartesian coordinates resulted in no separation of the three states. We suspect the poor performance stems from failure to fully eliminate rotational and translational degrees of freedom from the system. Complementary to PCA tICA was used to identify slow movements of the system. While the analysis revealed some structural insights into the product conformer ensemble – revealing π-π stacking interactions between the coupling product, phenanthroline, and calixarene – it was not able to separate educts from transition states. Hence, tICA could not fully identify the reaction coordinate.

Utilizing the labelled data in PCA and LDA combined with supervised ML approaches resulted in a much better separation of the three ensembles. While the performance of PCA was highly dependent on the internal coordinate set, LDA showed remarkable separation between the three ensembles, highlighting the robustness of the method. A consensus model developed to combine the performance of the various dimensionality and feature reduction methods, identified 49 internal coordinates to be relevant, with the C-N distance being the most prominent one, which is in agreement with chemical intuition.

By applying the Granger causality model, pioneered in econometrics and widely applied in other scientific fields, [41–56] we were able to decompose the reaction coordinate into a sequence of molecular motions. The four main components of the reaction coordinate, separated in time and happening as a domino effect, are depicted schematically in Figure 9: i) the calixarene cage tilts perpendicularly to the phenanthroline; ii) the Cu changes coordination to become planar; iii) the C-N bond distance shortens; iv) π-π stacking effects drive the movement of the product inside the cage and below the phenanthroline. It appears that in the majority of the sampled data, the C-N coupling distance is largely influenced by a movement in the upper cage of the calixarene, namely units 2,3 and 4, which can be seen as the trigger for the reaction. This works presents a causal analysis, based on the robust Granger model for the C-N bond formation as a proof of principle. However, the model relies on the assumptions of stationary time-series and provides limited insight into the scale of the relationship between cause

and effect. A transition to more advanced causal discovery models, such as PC Momentary Conditional Independence (PCMCI)[112] based methods would allow for analysis of non-stationary time series and better false positive control but is beyond the scope of this work.

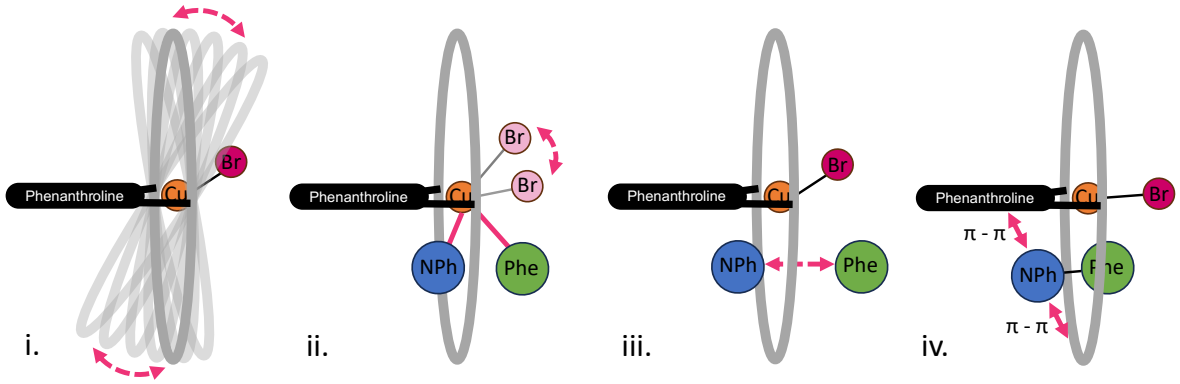

Figure 9. Schematic representation of the movements corresponding to the reaction coordinate of the C-N coupling reaction. i) Tilting movement of the calixarene cage; ii) Change in coordination of the copper centre; iii) Shortening of the C-N bond; iv) π-π stacking effects stabilise the product in the cavity.

To quantify the changes in internal coordinates, we employed a decision tree trained on the three ensembles, which highlighted the Cu-C and C-N distances as being the main feature for separating products from educts and transition states, respectively. However, decision trees are dependent on the initialisation condition and must be limited in their size to maintain interpretability. A transition to the decision rule approach overcomes some of the limitations by assigning a semantic interpretation to a random forest classifier, thus, clearly separating the three ensembles. In addition, three distinct product conformers are identified, each with a different set of rules defining them. These conformations will likely converge as the product diffuses out of the cavity.

To the best of our knowledge there is only a single study of the reaction dynamics of a transition-metal complex with explicit solvent with repeated sampling of the reaction step.[28] This study on Fe-oxo-mediated C–H functionalization reactions by Joy et. al used kinetic energies, quantum numbers and velocities to distinguish between two different dynamic reaction pathways.[28] While they also used ML for feature selection, their focus is on physical chemical factors that impact reactivity. Our focus lies on the analysis of structural changes to ease interpretation and to facilitate causality analysis, which has never been attempted for chemical reactions but opens a completely new angle on how to understand reactivity.

Conclusion

We developed a workflow to identify and quantify the reaction coordinate from a set of trajectories, detecting chemically intuitive and less intuitive contributions.

By devising a high throughput QM/MM MD workflow, we were able to study the C-N coupling reaction dynamics of a supramolecular Cu-calix[8]arene catalyst under experimental conditions. This development is a crucial step towards a predictive operando model for complex catalytic reactions. It

allowed us to extract not only reaction energies and barriers with uncertainties but also provide insights into the intricate dynamic nature of the macrocyclic transition metal catalyst in explicit solvent.

To process the vast amount of data, interpretable machine learning techniques proved to be invaluable, thanks to the ability to map results back to structural changes. However, a consensus model is needed to eliminate the inherent variability/instability of the individual ML approaches.

By performing a causal analysis of the internal coordinates of a system that contribute to the categorisation of the ensembles, an extra temporal dimension can be added to the reaction coordinate, allowing us to explain the chemical reaction as a sequence of movements leading up to C-N bond formation. One can then utilize this information to pinpoint the exact source (group of atoms) that triggers the reaction and by suitable chemical mutation tweak the system in such way that the reactivity is enhanced. By checking the outcome of decision trees and rules run on this modified system, one can gauge the impact of a specific change on the reaction coordinate. Through the implementation of the methodology in a straight forward workflow and the restriction to analysis of structural parameters, this technique is accessible to the non-expert user, while the obtained results, that is changes in coordinates during the reaction, can easily be understood by a general chemist.

Our methodology was demonstrated on a highly flexible Cu-calix[8]arene catalyst, but serves as a blueprint to identify and quantify the reaction coordinate in any dynamic chemical system, from small (in)organic complexes to large (bio-)molecules.

Methods

Workflow for Determination of Complex Reaction Coordinates

The multistep protocol developed to investigate the C-N coupling dynamics with a Cu-calix[8]arene catalyst is highlighted in Figure 10. It involves high throughput explicit solvent MD sampling of the reaction step, followed by machine learning analysis, where consensus features are extracted. These are utilized for qualitative and quantitative analysis of the reaction coordinate. As a last step, the time evolution of the system is considered by applying a causality model that allows to redefine the reaction coordinate as a sequence of individual movements of groups of atoms.

Simulation Protocol

For the QM/MM MD simulation, we chose the in-house developed ab-initio quantum mechanical charge field (QMCF) molecular dynamics approach[70] using the link-bond method to describe the bonds crossing between QM and MM.[71,72] The simulation parameters were set up using the GAFF[73] force-field[74], using the PyConSolv[19] 1.0.0 tool, with default settings. For the geometry optimization, the PBE0 functional[75] was used with the def2-SVP basis set[76] and D3 dispersion corrections[77] in implicit chloroform, using CPCM.[20] The system was solvated in a cubic periodic box with 1708 chloroform molecules. For detailed information regarding the simulation parameters, see the supporting information. The 54 atoms at the centre of the calixarene cage were included in the QM zone (see SI Figure S1). The quantum mechanical calculations were performed at two distinct levels. The semi-empirical method GFN2-xTB was utilized, providing a vast speedup of the QM calculation. As semi-empirical methods require benchmarking,[78–80] a full DFT reference was used, with PBE0/def2-SVP/D3 using Turbomole.[81]

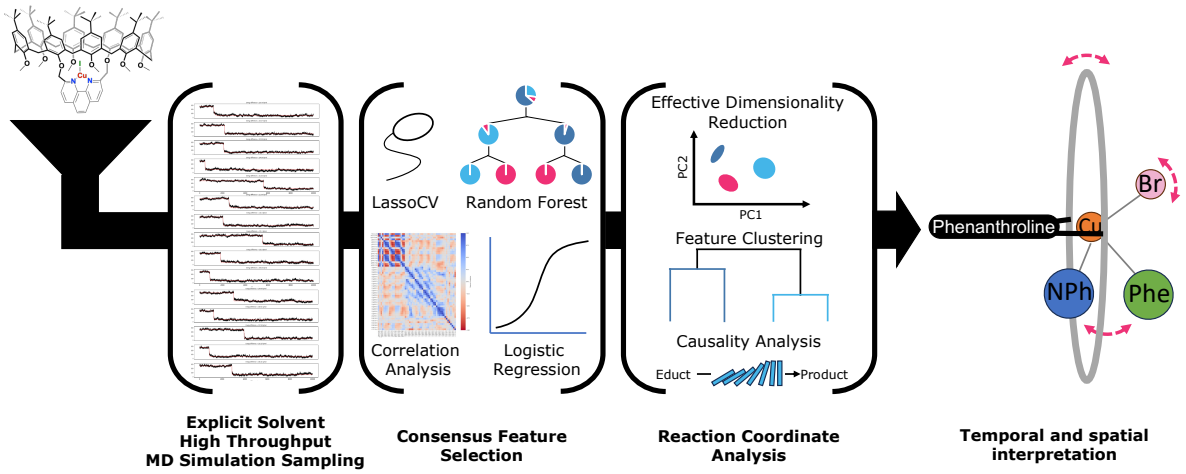

Figure 10. General workflow for extracting highly detailed information about the reaction coordinate from both a correlational and causal point of view.

To generate an appropriate starting structure the system was equilibrated and a 100ns MM/MD simulation was carried out, using the multistep protocol implemented in PyConSolv.[19] We conducted 152 simulation runs using the QM/MM MD protocol of the reductive elimination step, employing GFN2-xTB as QM method. Among these runs, 142 simulations captured the reductive elimination step and were utilized for energy and structural analysis.

Ensemble Labelling and Reaction Energy

From the MD simulation trajectories, cartesian coordinates and QM energies of the catalyst were extracted. The QM energy was used to label the individual frames as educt, transition state and product, while employing a filter function to minimize random fluctuations. The transition states were identified as the frames that define the last energy maximum before the large drop in energy associated with the formation of the product (See SI for details). The reaction barrier was defined as the difference between the transition state structure energy and the maximum value of a sigmoid function fitted through the reaction profile. The reaction energy, with uncertainty, was calculated based on the energy difference between the average energy of the product and educt ensembles, respectively. As DFT sampling on a similar scale as GFN2-xTB is not achievable, we cannot compare the reaction energies using the DFT educts and products. Thus, we used the fitted sigmoids to calculate the energy difference between the educts and product, by subtracting the highest sigmoid value from the lowest.

Identification of the Reaction Coordinate

We resorted to three coordinate systems to describe the reaction coordinate and significant secondary movements. Firstly, a cartesian coordinate system, obtained by combining all simulations and aligning them on the rigid phenanthroline bridge. This helps mitigate the impact of rotational and translational movements of the system during the simulations.[82] Secondly, we generated a set of internal coordinates for all trajectories, using the MDanalysis package.[83] This has the advantage of removing the issue of noise due to alignment artefacts, yet introduces more correlational effects.[82] Thirdly, we generated a

reduced set of internal coordinates, describing highly rigid chemical moieties by their centre of mass (see SI for details)

To determine the importance of each feature in distinguishing between the three ensembles (educt, TS, product), and therefore its contribution to the reaction coordinate, we performed dimensionality reduction on the coordinate sets, aiming to increase separation between ensembles.

To this end, we utilized Principal Component Analysis (PCA)[34] and Time-lagged Independent Component Analysis (tICA),[35] which complement each other in regards to addressing the variances present in the dataset,[84,85] as unsupervised methods. For supervised dimensionality reduction, we opt for Linear Discriminant Analysis (LDA)[86] due to its efficacy in separating distinct classes within a given dataset. To further enhance the separation capability of PCA and LDA, several methods of automated feature selection were chosen and implemented, namely Recursive Feature Elimination with Cross Validation[87] (RFECV) using both Random Forest [88] (RF) and Logistic Regression[89] as classifier models, and Lasso[90] with Cross Validation (LassoCV), all using fivefold stratified cross-validation.[91] The RF and Logistic Regression classifiers were also evaluated manually for performance, eliminating features under a certain threshold (see SI for details). Additionally, a simple Pearson correlation evaluation between the energy and each feature was performed, removing any coordinate which exhibited low correlation (higher than -0.75 and lower than 0.75). For the final feature selection, we employed a consensus model that extracted high-contributing features. These features were identified by performing a PCA and LDA based on the Recursive Feature Elimination with Cross-Validation (RFECV), LassoCV, and manual assessments approaches. Subsequently, we computed the average contributions to the first three principal components, as well as the two LDA components. The selected features needed to appear in at least 70% of the elimination models to be deemed significant. This process yielded a set of 49 internal coordinates for the PCA and 46 for the LDA. These coordinates were further subjected to causal inference (as detailed in the Statistical Methods section) and correlation interpretation. To enhance comprehensibility, we applied a hierarchical clustering algorithm to group together similar movements within the consensus features.[92]

In tandem with automated feature reduction methodologies, we leveraged decision trees to highlight features from our dataset. These decision trees underwent training under two distinct conditions: one with class balancing and another without. Furthermore, we adopted a decision rule framework, employing the skoperules[93] library. Within this framework, a random forest bagging classifier, consisting of 30 decision trees, was used to provide a semantically quantitative characterization of the three ensembles.

Statistical Model for Causality Inference

We aimed to infer the cause of the onset of the chemical reaction. To this end, we focused our investigation on the educt and TS ensembles of each trajectory. We eliminated trajectories with few educt structures (less than 100), thus having a total of 139 reaction sampling events. To infer Granger Causality[68] (GC) we followed the protocol outlined by Toda and Yamamoto.[69] This involved performing the augmented Dickey-Fuller[94] and Kwiatkowski–Phillips–Schmidt–Shin[95] tests to ascertain the stationarity of the various time series. Most of the time series were deemed stationary, with only a couple of the features presenting non-stationarity, which were rendered stationary through differencing (see SI for the critical and test statistics for each test and trajectory).[96] A multivariate vector autoregression model[97,98] was constructed and fitted with lag times varying from 0 to 50, for each time

series. The appropriate lag time was chosen for each trajectory, based on the Akaike information criterion[97,99] (see SI for additional information). The correlated time series were checked for cointegration using the Johansen test.[100] Finally, we calculated a GC matrix for each feature, for every simulation, resulting in a total of 139 matrices, using a VAM trained at the appropriate lag time. To account for the occurrence of false positives with repeated sampling of the reaction, we applied the false discovery rate (FDR) correction proposed by Benjamini and Hochberg,[101] with a threshold alpha of 0.1. The threshold for the GC test was set to $p < 0.05$. The full p-values for the causality matrices, non-FDR corrected, can be found in the SI, along with the results of all statistical tests.

Technical Note

Our implementation leveraged the following Python packages for data processing, statistical analysis and machine learning: pandas 2.2.1,[102] numpy 1.26.4,[103] scikit-learn 1.4.1.post1,[104] statsmodels 0.14.1,[105] scipy 1.12.0,[106] mdanalysis 2.7.0,[83,107] skoperules 1.0.1,[93] and pyemma 2.5.12.[108] For visualizations, we utilized dtreeviz 2.2.217,[109] seaborn 0.13.2,[110] and matplotlib 3.8.318.[111] The code used for analysis is made available as a jupyter notebook on Github (https://github.com/PodewitzLab/MLReactCoord) and will be integrated in the analysis suite of a future version of PyConSolv (https://github.com/PodewitzLab/PyConSolv).

Data availability

The p-values for the statistical analysis are available on Github (https://github.com/PodewitzLab/MLReactCoord), alongside two example trajectories. Due to the size of the simulation data, it can be made available upon request from the corresponding authors.

Code availability

The python code used for analysis is made available on Github (https://github.com/PodewitzLab/MLReactCoord) and the respective functions will be implemented in a future version of PyConSolv (https://github.com/PodewitzLab/PyConSolv) to facilitate a broad applicability.

Author contributions

The project was conceived by R.A.T. and M.P., while I.C provided chemical input for the system, required to devise the project. R.A.T. performed the QM/MM/MD calculations together with T.S.H, while J.G. implemented the rescaling barostat into the QMCFC package, specifically for this project. R.A.T. devised the Jupyter Notebook to conduct the analyses. Analyses were edited by M.P. R.A.T wrote the original draft, M.P. T.S.H and I.C. edited the draft. All authors agree with the final version of the draft. M. P. acquired the funding and supervised this project. The computational resources were provided by T.S.H and M.P.

Acknowledgment

The authors could like to thank Jonny Proppe for fruitful discussion. M. P. would like to thank the Austrian Science Fund (FWF) for financial support (P-33528).

Author Information

Corresponding Authors

T.S. Hofer - Institute of Inorganic and Theoretical Chemistry, Leopold Franzens University of Innsbruck, Innrain 80/82, 6020, Innsbruck (Austria); https://orcid.org/0000-0002-6559-1513
Email: t.hofer@uibk.ac.at

M. Podewitz- Institute of Materials Chemistry, TU Wien, Getreidemarkt 9, A-1060 Wien (Austria); https://orcid.org/0000-0001-7256-1219
Email: maren.podewitz@tuwien.ac.at

Authors

R. A. Talmazan - Institute of Materials Chemistry, TU Wien, Getreidemarkt 9, A-1060 Wien (Austria); https://orcid.org/0000-0001-6678-7801

J. Gamper - Institute of Inorganic and Theoretical Chemistry, Leopold Franzens University of Innsbruck, Innrain 80/82, 6020, Innsbruck (Austria); https://orcid.org/0000-0003-1136-2536

I. Castillo - Instituto de Química, Universidad Nacional Autónoma de México, Ciudad Universitaria, 04510, Ciudad de México (México); https://orcid.org/0000-0002-4876-4339

References

(1) Ringe, D.; Petsko, G. A. How Enzymes Work. *Science* **2008**, *320* (5882), 1428–1429. https://doi.org/10.1126/science.1159747.
(2) Koblenz, T. S.; Wassenaar, J.; Reek, J. N. H. Reactivity within a Confined Self-Assembled Nanospace. *Chem. Soc. Rev.* **2008**, *37* (2), 247–262. https://doi.org/10.1039/B614961H.
(3) Raynal, M.; Ballester, P.; Vidal-Ferran, A.; Leeuwen, P. W. N. M. van. Supramolecular Catalysis. Part 1: Non-Covalent Interactions as a Tool for Building and Modifying Homogeneous Catalysts. *Chem. Soc. Rev.* **2014**, *43* (5), 1660–1733. https://doi.org/10.1039/C3CS60027K.
(4) Raynal, M.; Ballester, P.; Vidal-Ferran, A.; Leeuwen, P. W. N. M. van. Supramolecular Catalysis. Part 2: Artificial Enzyme Mimics. *Chem. Soc. Rev.* **2014**, *43* (5), 1734–1787. https://doi.org/10.1039/C3CS60037H.
(5) Pachisia, S.; Gupta, R. Supramolecular Catalysis: The Role of H-Bonding Interactions in Substrate Orientation and Activation. *Dalton Trans.* **2021**, *50* (42), 14951–14966. https://doi.org/10.1039/D1DT02131A.
(6) Olivo, G.; Capocasa, G.; Giudice, D. D.; Lanzalunga, O.; Stefano, S. D. New Horizons for Catalysis Disclosed by Supramolecular Chemistry. *Chem. Soc. Rev.* **2021**, *50* (13), 7681–7724. https://doi.org/10.1039/D1CS00175B.
(7) Talmazan, R. A.; Refugio Monroy, J.; del Río-Portilla, F.; Castillo, I.; Podewitz, M. Encapsulation Enhances the Catalytic Activity of C-N Coupling: Reaction Mechanism of a Cu(I)/Calix[8]Arene Supramolecular Catalyst. *ChemCatChem* **2022**, *14* (20). https://doi.org/10.1002/cctc.202200662.
(8) Guzmán-Percástegui, E.; J. Hernández, D.; Castillo, I. Calix[8]Arene Nanoreactor for Cu(i)-Catalysed C–S Coupling. *Chemical Communications* **2016**, *52* (15), 3111–3114. https://doi.org/10.1039/C5CC09232A.

(9) Sciortino, G.; Maseras, F. Computational Study of Homogeneous Multimetallic Cooperative Catalysis. *Top Catal* **2022**, *65* (1), 105–117. https://doi.org/10.1007/s11244-021-01493-2.
(10) Artús Suàrez, L.; Balcells, D.; Nova, A. Computational Studies on the Mechanisms for Deaminative Amide Hydrogenation by Homogeneous Bifunctional Catalysts. *Top Catal* **2022**, *65* (1), 82–95. https://doi.org/10.1007/s11244-021-01542-w.
(11) Jones, G. O.; Liu, P.; Houk, K. N.; Buchwald, S. L. Computational Explorations of Mechanisms and Ligand-Directed Selectivities of Copper-Catalyzed Ullmann-Type Reactions. *J. Am. Chem. Soc.* **2010**, *132* (17), 6205–6213. https://doi.org/10.1021/ja100739h.
(12) Larsson, P.-F.; Wallentin, C.-J.; Norrby, P.-O. Mechanistic Aspects of Submol % Copper-Catalyzed C□N Cross-Coupling. *ChemCatChem* **2014**, *6* (5), 1277–1282. https://doi.org/10.1002/cctc.201301088.
(13) Fey, N.; Lynam, J. M. Computational Mechanistic Study in Organometallic Catalysis: Why Prediction Is Still a Challenge. *WIREs Computational Molecular Science* **2022**, *12* (4), e1590. https://doi.org/10.1002/wcms.1590.
(14) Harvey, J. N.; Himo, F.; Maseras, F.; Perrin, L. Scope and Challenge of Computational Methods for Studying Mechanism and Reactivity in Homogeneous Catalysis. *ACS Catal.* **2019**, *9* (8), 6803–6813. https://doi.org/10.1021/acscatal.9b01537.
(15) Besora, M.; Braga, A. A. C.; Ujaque, G.; Maseras, F.; Lledós, A. The Importance of Conformational Search: A Test Case on the Catalytic Cycle of the Suzuki–Miyaura Cross-Coupling. *Theor Chem Acc* **2011**, *128* (4–6), 639–646. https://doi.org/10.1007/s00214-010-0823-6.
(16) Podewitz, M.; Sen, S.; Buchmeiser, M. R. On the Origin of E-Selectivity in the Ring-Opening Metathesis Polymerization with Molybdenum Imido Alkylidene N-Heterocyclic Carbene Complexes. *Organometallics* **2021**, *40* (15), 2478–2488. https://doi.org/10.1021/acs.organomet.1c00229.
(17) Eisenstein, O.; Ujaque, G.; Lledós, A. What Makes a Good (Computed) Energy Profile? In *New Directions in the Modeling of Organometallic Reactions*; Lledós, A., Ujaque, G., Eds.; Topics in Organometallic Chemistry; Springer International Publishing: Cham, 2020; pp 1–38. https://doi.org/10.1007/3418_2020_57.
(18) Pracht, P.; Grimme, S.; Bannwarth, C.; Bohle, F.; Ehlert, S.; Feldmann, G.; Gorges, J.; Müller, M.; Neudecker, T.; Plett, C.; Spicher, S.; Steinbach, P.; Wesołowski, P. A.; Zeller, F. CREST—A Program for the Exploration of Low-Energy Molecular Chemical Space. *The Journal of Chemical Physics* **2024**, *160* (11), 114110. https://doi.org/10.1063/5.0197592.
(19) Talmazan, R. A.; Podewitz, M. PyConSolv: A Python Package for Conformer Generation of (Metal-Containing) Systems in Explicit Solvent. *J. Chem. Inf. Model.* **2023**, *63* (17), 5400–5407. https://doi.org/10.1021/acs.jcim.3c00798.
(20) Barone, V.; Cossi, M. Quantum Calculation of Molecular Energies and Energy Gradients in Solution by a Conductor Solvent Model. *J. Phys. Chem. A* **1998**, *102* (11), 1995–2001. https://doi.org/10.1021/jp9716997.
(21) Klamt, A.; Schüürmann, G. COSMO: A New Approach to Dielectric Screening in Solvents with Explicit Expressions for the Screening Energy and Its Gradient. *J. Chem. Soc., Perkin Trans. 2* **1993**, No. 5, 799–805. https://doi.org/10.1039/P29930000799.
(22) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **2009**, *113* (18), 6378–6396. https://doi.org/10.1021/jp810292n.
(23) Simm, G. N.; Türtscher, P. L.; Reiher, M. Systematic Microsolvation Approach with a Cluster-Continuum Scheme and Conformational Sampling. *Journal of Computational Chemistry* **2020**, *41* (12), 1144–1155. https://doi.org/10.1002/jcc.26161.

(24) Steiner, M.; Holzknecht, T.; Schauperl, M.; Podewitz, M. Quantum Chemical Microsolvation by Automated Water Placement. *Molecules* **2021**, *26* (6), 1793. https://doi.org/10.3390/molecules26061793.
(25) Joy, J.; Ess, D. H. Direct Dynamics Trajectories Demonstrate Dynamic Matching and Nonstatistical Radical Pair Intermediates during Fe-Oxo-Mediated C–H Functionalization Reactions. *J. Am. Chem. Soc.* **2023**, *145* (13), 7628–7637. https://doi.org/10.1021/jacs.3c01196.
(26) Yang, Z.; Jamieson, C. S.; Xue, X.-S.; Garcia-Borràs, M.; Benton, T.; Dong, X.; Liu, F.; Houk, K. N. Mechanisms and Dynamics of Reactions Involving Entropic Intermediates. *TRECHEM* **2019**, *1* (1), 22–34. https://doi.org/10.1016/j.trechm.2019.01.009.
(27) Ess, D. H. Quasiclassical Direct Dynamics Trajectory Simulations of Organometallic Reactions. *Acc. Chem. Res.* **2021**, *54* (23), 4410–4422. https://doi.org/10.1021/acs.accounts.1c00575.
(28) Joy, J.; Schaefer, A. J.; Teynor, M. S.; Ess, D. H. Dynamical Origin of Rebound versus Dissociation Selectivity during Fe-Oxo-Mediated C–H Functionalization Reactions. *J. Am. Chem. Soc.* **2024**, *146* (4), 2452–2464. https://doi.org/10.1021/jacs.3c09891.
(29) Warshel, A.; Karplus, M. Calculation of Ground and Excited State Potential Surfaces of Conjugated Molecules. I. Formulation and Parametrization. *J. Am. Chem. Soc.* **1972**, *94* (16), 5612–5625. https://doi.org/10.1021/ja00771a014.
(30) Warshel, A.; Levitt, M. Theoretical Studies of Enzymic Reactions: Dielectric, Electrostatic and Steric Stabilization of the Carbonium Ion in the Reaction of Lysozyme. *Journal of Molecular Biology* **1976**, *103* (2), 227–249. https://doi.org/10.1016/0022-2836(76)90311-9.
(31) Gao, J. Hybrid Quantum and Molecular Mechanical Simulations: An Alternative Avenue to Solvent Effects in Organic Chemistry. *Acc. Chem. Res.* **1996**, *29* (6), 298–305. https://doi.org/10.1021/ar950140r.
(32) Nandy, A.; Duan, C.; Taylor, M. G.; Liu, F.; Steeves, A. H.; Kulik, H. J. Computational Discovery of Transition-Metal Complexes: From High-Throughput Screening to Machine Learning. *Chem. Rev.* **2021**, *121* (16), 9927–10000. https://doi.org/10.1021/acs.chemrev.1c00347.
(33) Montavon, G.; Samek, W.; Müller, K.-R. Methods for Interpreting and Understanding Deep Neural Networks. *Digital Signal Processing* **2018**, *73*, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011.
(34) Wold, S.; Esbensen, K.; Geladi, P. Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems* **1987**, *2* (1), 37–52. https://doi.org/10.1016/0169-7439(87)80084-9.
(35) Molgedey, L.; Schuster, H. G. Separation of a Mixture of Independent Signals Using Time Delayed Correlations. *Phys. Rev. Lett.* **1994**, *72* (23), 3634–3637. https://doi.org/10.1103/PhysRevLett.72.3634.
(36) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. Essential Dynamics of Proteins. *Proteins: Structure, Function, and Bioinformatics* **1993**, *17* (4), 412–425. https://doi.org/10.1002/prot.340170408.
(37) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of Slow Molecular Order Parameters for Markov Model Construction. *The Journal of Chemical Physics* **2013**, *139* (1), 015102. https://doi.org/10.1063/1.4811489.
(38) Schwantes, C. R.; Pande, V. S. Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *J Chem Theory Comput* **2013**, *9* (4), 2000–2009. https://doi.org/10.1021/ct300878a.
(39) Hare, S. R.; Bratholm, L. A.; Glowacki, D. R.; Carpenter, B. K. Low Dimensional Representations along Intrinsic Reaction Coordinates and Molecular Dynamics Trajectories Using Interatomic Distance Matrices. *Chem. Sci.* **2019**, *10* (43), 9954–9968. https://doi.org/10.1039/C9SC02742D.
(40) *Reaction Space Projector (ReSPer) for Visualizing Dynamic Reaction Routes Based on Reduced-Dimension Space | Topics in Current Chemistry*. https://link.springer.com/article/10.1007/s41061-022-00377-7 (accessed 2024-06-04).

(41) Imbens, G. W.; Rubin, D. B. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*; Cambridge University Press: Cambridge, 2015. https://doi.org/10.1017/CBO9781139025751.
(42) Xing, Q.; Wu, C.; Chen, F.; Liu, J.; Pradhan, P.; Bryan, B. A.; Schaubroeck, T.; Carrasco, L. R.; Gonsamo, A.; Li, Y.; Chen, X.; Deng, X.; Albanese, A.; Li, Y.; Xu, Z. Intranational Synergies and Trade-Offs Reveal Common and Differentiated Priorities of Sustainable Development Goals in China. *Nat Commun* **2024**, *15* (1), 2251. https://doi.org/10.1038/s41467-024-46491-6.
(43) Cárdenas-García, P. J.; Brida, J. G.; Segarra, V. Modeling the Link between Tourism and Economic Development: Evidence from Homogeneous Panels of Countries. *Humanit Soc Sci Commun* **2024**, *11* (1), 1–12. https://doi.org/10.1057/s41599-024-02826-8.
(44) Kaufmann, R. K.; Newberry, D.; Xin, C.; Gopal, S. Feedbacks among Electric Vehicle Adoption, Charging, and the Cost and Installation of Rooftop Solar Photovoltaics. *Nat Energy* **2021**, *6* (2), 143–149. https://doi.org/10.1038/s41560-020-00746-w.
(45) Tanaka, T.; Guo, J. International Price Volatility Transmission and Structural Change: A Market Connectivity Analysis in the Beef Sector. *Humanit Soc Sci Commun* **2020**, *7* (1), 1–13. https://doi.org/10.1057/s41599-020-00657-x.
(46) Ouyang, T.; Liu, F.; Huang, B. Dynamic Econometric Analysis on Influencing Factors of Production Efficiency in Construction Industry of Guangxi Province in China. *Sci Rep* **2022**, *12* (1), 17509. https://doi.org/10.1038/s41598-022-22374-y.
(47) Le, T. Increased Impact of the El Niño–Southern Oscillation on Global Vegetation under Future Warming Environment. *Sci Rep* **2023**, *13* (1), 14459. https://doi.org/10.1038/s41598-023-41590-8.
(48) Runge, J.; Bathiany, S.; Bollt, E.; Camps-Valls, G.; Coumou, D.; Deyle, E.; Glymour, C.; Kretschmer, M.; Mahecha, M. D.; Muñoz-Marí, J.; van Nes, E. H.; Peters, J.; Quax, R.; Reichstein, M.; Scheffer, M.; Schölkopf, B.; Spirtes, P.; Sugihara, G.; Sun, J.; Zhang, K.; Zscheischler, J. Inferring Causation from Time Series in Earth System Sciences. *Nat Commun* **2019**, *10* (1), 2553. https://doi.org/10.1038/s41467-019-10105-3.
(49) Kretschmer, M.; Coumou, D.; Donges, J. F.; Runge, J. Using Causal Effect Networks to Analyze Different Arctic Drivers of Midlatitude Winter Circulation. *Journal of Climate* **2016**, *29* (11), 4069–4081. https://doi.org/10.1175/JCLI-D-15-0654.1.
(50) Triacca, U. Is Granger Causality Analysis Appropriate to Investigate the Relationship between Atmospheric Concentration of Carbon Dioxide and Global Surface Air Temperature? *Theor. Appl. Climatol.* **2005**, *81* (3), 133–135. https://doi.org/10.1007/s00704-004-0112-1.
(51) McGraw, M. C.; Barnes, E. A. Memory Matters: A Case for Granger Causality in Climate Variability Studies. *Journal of Climate* **2018**, *31* (8), 3289–3300. https://doi.org/10.1175/JCLI-D-17-0334.1.
(52) Hill, S. M.; Heiser, L. M.; Cokelaer, T.; Unger, M.; Nesser, N. K.; Carlin, D. E.; Zhang, Y.; Sokolov, A.; Paull, E. O.; Wong, C. K.; Graim, K.; Bivol, A.; Wang, H.; Zhu, F.; Afsari, B.; Danilova, L. V.; Favorov, A. V.; Lee, W. S.; Taylor, D.; Hu, C. W.; Long, B. L.; Noren, D. P.; Bisberg, A. J.; Mills, G. B.; Gray, J. W.; Kellen, M.; Norman, T.; Friend, S.; Qutub, A. A.; Fertig, E. J.; Guan, Y.; Song, M.; Stuart, J. M.; Spellman, P. T.; Koeppl, H.; Stolovitzky, G.; Saez-Rodriguez, J.; Mukherjee, S. Inferring Causal Molecular Networks: Empirical Assessment through a Community-Based Effort. *Nat Methods* **2016**, *13* (4), 310–318. https://doi.org/10.1038/nmeth.3773.
(53) Walter, J. Establishing Microbiome Causality to Tackle Malnutrition. *Nat Microbiol* **2024**, *9* (4), 884–885. https://doi.org/10.1038/s41564-024-01653-6.
(54) Friston, K. J.; Harrison, L.; Penny, W. Dynamic Causal Modelling. *NeuroImage* **2003**, *19* (4), 1273–1302. https://doi.org/10.1016/S1053-8119(03)00202-7.
(55) Duggento, A.; Passamonti, L.; Valenza, G.; Barbieri, R.; Guerrisi, M.; Toschi, N. Multivariate Granger Causality Unveils Directed Parietal to Prefrontal Cortex Connectivity during Task-Free MRI. *Sci Rep* **2018**, *8* (1), 5571. https://doi.org/10.1038/s41598-018-23996-x.

(56) Zheng, J.; Anderson, K. L.; Leal, S. L.; Shestyuk, A.; Gulsen, G.; Mnatsakanyan, L.; Vadera, S.; Hsu, F. P. K.; Yassa, M. A.; Knight, R. T.; Lin, J. J. Amygdala-Hippocampal Dynamics during Salient Information Processing. *Nat Commun* **2017**, *8* (1), 14413. https://doi.org/10.1038/ncomms14413.
(57) Kamberaj, H.; van der Vaart, A. Correlated Motions and Interactions at the Onset of the DNA-Induced Partial Unfolding of Ets-1. *Biophys J* **2009**, *96* (4), 1307–1317. https://doi.org/10.1016/j.bpj.2008.11.019.
(58) Kamberaj, H.; van der Vaart, A. Extracting the Causality of Correlated Motions from Molecular Dynamics Simulations. *Biophys J* **2009**, *97* (6), 1747–1755. https://doi.org/10.1016/j.bpj.2009.07.019.
(59) Sobieraj, M.; Setny, P. Granger Causality Analysis of Chignolin Folding. *J. Chem. Theory Comput.* **2022**, *18* (3), 1936–1944. https://doi.org/10.1021/acs.jctc.1c00945.
(60) Buskes, M. J.; Blanco, M.-J. Impact of Cross-Coupling Reactions in Drug Discovery and Development. *Molecules* **2020**, *25* (15), 3493. https://doi.org/10.3390/molecules25153493.
(61) Gay, M.; Carato, P.; Coevoet, M.; Renault, N.; Larchanché, P.-E.; Barczyk, A.; Yous, S.; Buée, L.; Sergeant, N.; Melnyk, P. New Phenylaniline Derivatives as Modulators of Amyloid Protein Precursor Metabolism. *Bioorganic & Medicinal Chemistry* **2018**, *26* (8), 2151–2164. https://doi.org/10.1016/j.bmc.2018.03.016.
(62) Jiang, R.; Li, L.; Sheng, T.; Hu, G.; Chen, Y.; Wang, L. Edge-Site Engineering of Atomically Dispersed Fe–N4 by Selective C–N Bond Cleavage for Enhanced Oxygen Reduction Reaction Activities. *J. Am. Chem. Soc.* **2018**, *140* (37), 11594–11598. https://doi.org/10.1021/jacs.8b07294.
(63) Hayakawa, S.; Kawasaki, A.; Hong, Y.; Uraguchi, D.; Ooi, T.; Kim, D.; Akutagawa, T.; Fukui, N.; Shinokubo, H. Inserting Nitrogen: An Effective Concept To Create Nonplanar and Stimuli-Responsive Perylene Bisimide Analogues. *J. Am. Chem. Soc.* **2019**, *141* (50), 19807–19816. https://doi.org/10.1021/jacs.9b09556.
(64) Izumi, S.; Higginbotham, H. F.; Nyga, A.; Stachelek, P.; Tohnai, N.; Silva, P. de; Data, P.; Takeda, Y.; Minakata, S. Thermally Activated Delayed Fluorescent Donor–Acceptor–Donor–Acceptor π-Conjugated Macrocycle for Organic Light-Emitting Diodes. *J. Am. Chem. Soc.* **2020**, *142* (3), 1482–1491. https://doi.org/10.1021/jacs.9b11578.
(65) Picini, F.; Schneider, S.; Gavat, O.; Vargas Jentzsch, A.; Tan, J.; Maaloum, M.; Strub, J.-M.; Tokunaga, S.; Lehn, J.-M.; Moulin, E.; Giuseppone, N. Supramolecular Polymerization of Triarylamine-Based Macrocycles into Electroactive Nanotubes. *J. Am. Chem. Soc.* **2021**, *143* (17), 6498–6504. https://doi.org/10.1021/jacs.1c00623.
(66) Berlanga-Vázquez, A.; Talmazan, R. A.; Reyes-Mata, C. A.; Percástegui, E. G.; Flores-Alamo, M.; Podewitz, M.; Castillo, I. Conformational Effects of Regioisomeric Substitution on the Catalytic Activity of Copper/Calix[8]Arene C−S Coupling. *Eur J Inorg Chem* **2022**. https://doi.org/10.1002/ejic.202200596.
(67) 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* (3), 1652–1671. https://doi.org/10.1021/acs.jctc.8b01176.
(68) Granger, C. W. J. Investigating Causal Relations by Econometric Models and Cross-Spectral Methods. *Econometrica* **1969**, *37* (3), 424–438. https://doi.org/10.2307/1912791.
(69) Toda, H. Y.; Yamamoto, T. Statistical Inference in Vector Autoregressions with Possibly Integrated Processes. *Journal of Econometrics* **1995**, *66* (1), 225–250. https://doi.org/10.1016/0304-4076(94)01616-8.
(70) Rode, B. M.; Hofer, T. S.; Randolf, B. R.; Schwenk, C. F.; Xenides, D.; Vchirawongkwin, V. Ab Initio Quantum Mechanical Charge Field (QMCF) Molecular Dynamics: A QM/MM – MD Procedure for

Accurate Simulations of Ions and Complexes. *Theor Chem Acc* **2006**, *115* (2), 77–85. https://doi.org/10.1007/s00214-005-0049-1.
(71) Amara, P.; Field, M. J. Evaluation of an Ab Initio Quantum Mechanical/Molecular Mechanical Hybrid-Potential Link-Atom Method. *Theor Chem Acc* **2003**, *109* (1), 43–52. https://doi.org/10.1007/s00214-002-0413-3.
(72) Singh, U. C.; Kollman, P. A. A Combined Ab Initio Quantum Mechanical and Molecular Mechanical Method for Carrying out Simulations on Complex Molecular Systems: Applications to the CH3Cl + Cl− Exchange Reaction and Gas Phase Protonation of Polyethers. *Journal of Computational Chemistry* **1986**, *7* (6), 718–730. https://doi.org/10.1002/jcc.540070604.
(73) Case, D. A.; Ben-Shalom, I. Y.; Brozell, S. R.; Cerutti, D. S.; Cheatham III, T. E.; Cruzeiro, V. W. D.; Darden, T. A.; Duke, R. E.; Ghoreishi, D.; Gilson, M. K. AMBER20. *University of California, San Francisco* **2020**.
(74) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25* (9), 1157–1174. https://doi.org/10.1002/jcc.20035.
(75) Adamo, C.; Barone, V. Toward Reliable Density Functional Methods without Adjustable Parameters: The PBE0 Model. *The Journal of Chemical Physics* **1999**, *110* (13), 6158–6170. https://doi.org/10.1063/1.478522.
(76) 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* (18), 3297–3305. https://doi.org/10.1039/B508541A.
(77) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J Chem Phys* **2010**, *132* (15), 154104. https://doi.org/10.1063/1.3382344.
(78) Bursch, M.; Hansen, A.; Pracht, P.; Kohn, J. T.; Grimme, S. Theoretical Study on Conformational Energies of Transition Metal Complexes. *Phys. Chem. Chem. Phys.* **2021**, *23* (1), 287–299. https://doi.org/10.1039/D0CP04696E.
(79) Husch, T.; Vaucher, A. C.; Reiher, M. Semiempirical Molecular Orbital Models Based on the Neglect of Diatomic Differential Overlap Approximation. *International Journal of Quantum Chemistry* **2018**, *118* (24), e25799. https://doi.org/10.1002/qua.25799.
(80) Podewitz, M. Towards Predictive Computational Catalysis – a Case Study of Olefin Metathesis with Mo Imido Alkylidene N-Heterocyclic Carbene Catalysts. In *Chemical Modelling: Volume 17*; Bahmann, H., Tremblay, J. C., Eds.; The Royal Society of Chemistry, 2022; Vol. 17, p 0. https://doi.org/10.1039/9781839169342-00001.
(81) Balasubramani, S. G.; Chen, G. P.; Coriani, S.; Diedenhofen, M.; Frank, M. S.; Franzke, Y. J.; Furche, F.; Grotjahn, R.; Harding, M. E.; Hättig, C.; Hellweg, A.; Helmich-Paris, B.; Holzer, C.; Huniar, U.; Kaupp, M.; Marefat Khah, A.; Karbalaei Khani, S.; Müller, T.; Mack, F.; Nguyen, B. D.; Parker, S. M.; Perlt, E.; Rappoport, D.; Reiter, K.; Roy, S.; Rückert, M.; Schmitz, G.; Sierka, M.; Tapavicza, E.; Tew, D. P.; van Wüllen, C.; Voora, V. K.; Weigend, F.; Wodyński, A.; Yu, J. M. TURBOMOLE: Modular Program Suite for Ab Initio Quantum-Chemical and Condensed-Matter Simulations. *The Journal of Chemical Physics* **2020**, *152* (18), 184107. https://doi.org/10.1063/5.0004635.
(82) Sittel, F.; Jain, A.; Stock, G. Principal Component Analysis of Molecular Dynamics: On the Use of Cartesian vs. Internal Coordinates. *The Journal of Chemical Physics* **2014**, *141* (1), 014111. https://doi.org/10.1063/1.4885338.
(83) Gowers, R. J.; Linke, M.; Barnoud, J.; Reddy, T. J. E.; Melo, M. N.; Seyler, S. L.; Domański, J.; Dotson, D. L.; Buchoux, S.; Kenney, I. M.; Beckstein, O. MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. *Proceedings of the 15th Python in Science Conference* **2016**, 98–105. https://doi.org/10.25080/Majora-629e541a-00e.

(84) Schultze, S.; Grubmüller, H. Time-Lagged Independent Component Analysis of Random Walks and Protein Dynamics. *J. Chem. Theory Comput.* **2021**, *17* (9), 5766–5776. https://doi.org/10.1021/acs.jctc.1c00273.
(85) Wu, H.; Nüske, F.; Paul, F.; Klus, S.; Koltai, P.; Noé, F. Variational Koopman Models: Slow Collective Variables and Molecular Kinetics from Short off-Equilibrium Simulations. *The Journal of Chemical Physics* **2017**, *146* (15), 154104. https://doi.org/10.1063/1.4979344.
(86) Mika, S.; Ratsch, G.; Weston, J.; Scholkopf, B.; Mullers, K. R. Fisher Discriminant Analysis with Kernels. In *Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop (Cat. No.98TH8468)*; 1999; pp 41–48. https://doi.org/10.1109/NNSP.1999.788121.
(87) Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene Selection for Cancer Classification Using Support Vector Machines. *Machine Learning* **2002**, *46* (1), 389–422. https://doi.org/10.1023/A:1012487302797.
(88) Breiman, L. Random Forests. *Machine Learning* **2001**, *45* (1), 5–32. https://doi.org/10.1023/A:1010933404324.
(89) Cox, D. R. The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society: Series B (Methodological)* **1958**, *20* (2), 215–232. https://doi.org/10.1111/j.2517-6161.1958.tb00292.x.
(90) Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **1996**, *58* (1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
(91) Zeng, X.; Martinez, T. R. Distribution-Balanced Stratified Cross-Validation for Accuracy Estimation. *Journal of Experimental & Theoretical Artificial Intelligence* **2000**, *12* (1), 1–12. https://doi.org/10.1080/095281300146272.
(92) Chormunge, S.; Jena, S. Correlation Based Feature Selection with Clustering for High Dimensional Data. *Journal of Electrical Systems and Information Technology* **2018**, *5* (3), 542–549. https://doi.org/10.1016/j.jesit.2017.06.004.
(93) Scikit-Learn-Contrib/Skope-Rules, 2024. https://github.com/scikit-learn-contrib/skope-rules (accessed 2024-04-15).
(94) Dickey, D. A.; Fuller, W. A. Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *Journal of the American Statistical Association* **1979**, *74* (366), 427–431. https://doi.org/10.2307/2286348.
(95) Kwiatkowski, D.; Phillips, P. C. B.; Schmidt, P.; Shin, Y. Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root: How Sure Are We That Economic Time Series Have a Unit Root? *Journal of Econometrics* **1992**, *54* (1), 159–178. https://doi.org/10.1016/0304-4076(92)90104-Y.
(96) Hyndman, R. J.; Athanasopoulos, G. Forecasting: Principles and Practice. **2018**.
(97) Lütkepohl, H. *Introduction to Multiple Time Series Analysis*; Springer: Berlin, Heidelberg, 1991. https://doi.org/10.1007/978-3-662-02691-5.
(98) Sims, C. A. Macroeconomics and Reality. *Econometrica* **1980**, *48* (1), 1–48. https://doi.org/10.2307/1912017.
(99) Akaike, H. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control* **1974**, *19* (6), 716–723. https://doi.org/10.1109/TAC.1974.1100705.
(100) Johansen, S. Statistical Analysis of Cointegration Vectors. *Journal of Economic Dynamics and Control* **1988**, *12* (2), 231–254. https://doi.org/10.1016/0165-1889(88)90041-3.
(101) Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **1995**, *57* (1), 289–300.

(102) The pandas development team. Pandas-Dev/Pandas: Pandas, 2024. https://doi.org/10.5281/zenodo.10957263.
(103) Harris, C. R.; Millman, K. J.; van der Walt, S. J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N. J.; Kern, R.; Picus, M.; Hoyer, S.; van Kerkwijk, M. H.; Brett, M.; Haldane, A.; del Río, J. F.; Wiebe, M.; Peterson, P.; Gérard-Marchant, P.; Sheppard, K.; Reddy, T.; Weckesser, W.; Abbasi, H.; Gohlke, C.; Oliphant, T. E. Array Programming with NumPy. *Nature* **2020**, *585* (7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2.
(104) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12* (85), 2825–2830.
(105) Seabold, S.; Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. In *Proceedings of the 9th Python in Science Conference*; Walt, S. van der, Millman, J., Eds.; 2010; pp 92–96. https://doi.org/10.25080/Majora-92bf1922-011.
(106) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, İ.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat Methods* **2020**, *17* (3), 261–272. https://doi.org/10.1038/s41592-019-0686-2.
(107) Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. MDAnalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations. *Journal of Computational Chemistry* **2011**, *32* (10), 2319–2327. https://doi.org/10.1002/jcc.21787.
(108) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noé, F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* **2015**, *11* (11), 5525–5542. https://doi.org/10.1021/acs.jctc.5b00743.
(109) Parr, T. Parrt/Dtreeviz, 2024. https://github.com/parrt/dtreeviz (accessed 2024-04-15).
(110) Waskom, M. L. Seaborn: Statistical Data Visualization. *Journal of Open Source Software* **2021**, *6* (60), 3021. https://doi.org/10.21105/joss.03021.
(111) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* **2007**, *9* (3), 90–95. https://doi.org/10.1109/MCSE.2007.55.
(112) Runge, J.; Nowack, P.; Kretschmer, M.; Flaxman, S.; Sejdinovic, D. Detecting and Quantifying Causal Associations in Large Nonlinear Time Series Datasets. *Science Advances* **2019**, *5* (11), eaau4996. https://doi.org/10.1126/sciadv.aau4996.