Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Digital Discovery*, 2025, **4**, 2954

Coupling causality and interpretable machine learning to reveal the reaction coordinate of C–N coupling with a supramolecular Cu-calix[8]arene catalyst

R. A. Talmazan,[a] J. Gamper,[b] I. Castillo,[c] T. S. Hofer *[b] and M. Podewitz *[a]

Received 21st May 2025
Accepted 29th August 2025

DOI: 10.1039/d5dd00216h

rsc.li/digitaldiscovery

Supramolecular 3d transition-metal catalysts are large, flexible systems with intricate interactions, resulting in complex reaction coordinates. To capture their dynamic nature, we developed a broadly applicable, high-throughput workflow, that leverages quantum mechanics/molecular mechanics molecular dynamics (QM/MM MD) 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 requires 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). By employing a Granger Causality model, we move beyond the traditional view of a reaction coordinate, by defining it instead as a sequence of molecular motions leading up to the reaction.

Introduction

Nature has perfected the principle of catalysis in enzymes, where a precise control of the environment surrounding the catalytic centre and substrate lowers the energy barrier.[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, such as calix[*n*]arene, allows the substitution of precious 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 the systems grow in complexity, there is a need to improve 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 Density Functional Theory (DFT).[13,15] Consequently, the goal is to create a “digital twin” of the reaction flask that is an *in silico* model, which fully replicates experimental conditions: a catalyst in explicit solvent, at finite temperature and pressure. As a description with full *ab initio* quantum chemistry is not feasible, a tailored multiscale strategy is required, accounting for conformational flexibility, explicit solvation, and the dynamic nature of reactions.

Although many computational mechanistic studies are still performed on a single structure, conformer searches recently gained popularity[14,16,17] thanks to easily accessible tools,[18,19] facilitating a transition to structure ensembles, which provide a more complete picture of the reactivity.[14,17] The standard implicit solvation model can favour unrealistic, very compact structures with many intramolecular hydrogen bonds.[7,19] Ideally, a more realistic model would account for explicit solvent molecules either through a full condensed phase calculation or through a microsolvation approach.[20,21]

Molecular dynamics simulations can describe the dynamic nature of the catalysts bringing its description closer to *operando* conditions, which can reveal new insights into the reaction mechanism.[22–25] While this treatment increases computational demands, by relying on multiscale methods, this hurdle can be greatly diminished. Quantum mechanics/molecular mechanics (QM/MM) models describe the catalytic centre and substrates at a QM level, while the surrounding environment is treated with MM.[26–28] A transition to QM/MM molecular dynamics (MD) in explicit solvent allows for sampling timescales magnitudes higher than in a pure QM approach. Due to thermal fluctuations, repeated sampling is a requirement for statistically

[a] *Institute of Materials Chemistry, TU Wien, Getreidemarkt 9, A-1060 Wien, Austria. E-mail: maren.podewitz@tuwien.ac.at*

[b] *Institute of General, Inorganic and Theoretical Chemistry, Leopold Franzens University of Innsbruck, Innrain 80/82, 6020 Innsbruck, Austria. E-mail: t.hofer@uibk.ac.at*

[c] *Instituto de Química, Universidad Nacional Autónoma de México, Ciudad Universitaria, Ciudad de México, 04510, Mexico*

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. Chemical knowledge can be extracted from these data, for example in the form of the reaction coordinate *i.e.*, the movements of atoms which take place as the chemical reaction proceeds. Knowledge of the reaction coordinate allows for further reduction in computational costs, by enabling the use of enhanced sampling methods geared towards overcoming large energy barriers, such as those observed in chemical reactions.[29] The identification of the reaction coordinate from simulation data can be performed in a variety of ways.[30] While machine learning (ML) approaches can be used to evaluate the data and extract condensed results from simulations, such as the committor function,[24,31] complex neural network approaches generally are not directly interpretable.[32] Whether through the use of convolutional neural networks or graph neural networks,[33,34] a subsequent analysis is still needed to render the reaction coordinate interpretable. Instead, interpretable, or explainable, machine learning techniques, such as Decision Trees, Random Forests or Logistic Regression, as well as path sampling-based methods, such as predictive power analysis,[35,36] offer good performance in extracting relevant information from large datasets and presenting them in easily understandable ways. In addition, dimensionality reduction techniques, such as Principal Component Analysis[37] (PCA) or time-lagged Independent Component Analysis[38] (tICA) effectively detect combined coordinates from the trajectories, revealing the key motions of a system. Furthermore, methods such as *k*-nearest neighbour and *t*-distributed stochastic neighbour embedding have shown some promise in dealing with complex datasets.[39–41] Yet these dimensionality reduction techniques have almost exclusively been applied to biomolecules[42–44] with few exceptions.[45,46] 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. However, 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[47] – ranging from economics[48–52] and climate research[53–57] to biology[58,59] and medical studies[60–62] – it remains surprisingly absent in the field of chemistry. Although a handful of precedents in biomolecular simulations exists,[63–65] it has not been explored to study chemical reactions, not to mention transition-metal catalysis. As MD simulation 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.

A supramolecular catalyst that has shown remarkable catalytic efficacy for C–N coupling is the Cu(I)-1,5-(2,9-dimethyl-1,10-phenanthroyl)-2,3,4,6,7,8-hexamethyl-*p*-*tert*-butylcalix[8]arene, short noted as [$Cu(C_8PhenMe_6)I$].[7] The macrocyclic ligand, sketched in Fig. 1, allows for usage of earth abundant metals, here Cu, which is an essential step towards more sustainable chemical processes.[66–71] 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 (see also Fig. S14–S15, Table S5 in the SI).[7] To account for the conformational flexibility of the supramolecular cage[7,72] 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 catalyst 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, are unknown. From previous studies, we know that explicit solvent molecules are crucial to maintain the shape and functionality of the cavity, however, we see no experimental or theoretical evidence of their participation in the reaction itself.[7,72]

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[73] method to describe the QM part, we achieved massive sampling, resulting in 152 individual unbiased 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. The statistical Granger Causality analysis model[74,75] was employed to decompose the reaction coordinate into a sequence of individual consecutive movements. Finally, Random Forest models and Decision Rules allowed us to quantify the reaction coordinate. This work serves as a broadly applicable template for any mechanistic investigation, revealing and quantifying complex reaction coordinates, along with causal effects derived from the individual movements leading up to the reaction.

Results

We set-up a QM/MM model [$Cu(C_8PhenMe_6)I$] catalyst, where the reaction centre is modelled by QM and the macrocyclic ligand, as well as the solvent, by MM (see methods and SI Fig. S1 and Table S1 for details). We obtained 152 unbiased QM/MM MD reaction trajectories for the C–N coupling step O7 with a total of over 3 ns of simulation time. Out of these, 142 reacted spontaneously within 20 ps of simulation time without the need to bias. 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, identify a sequence of movements leading to the reaction, and quantify it.

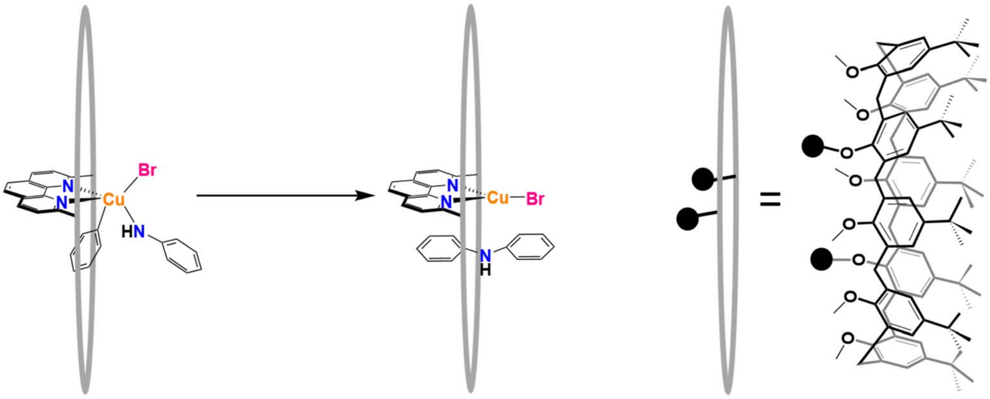

Fig. 1 The supramolecular calix[8]arene-based [Cu(C_8PhenMe$_6$)I] system that catalyses the C–N coupling reaction of phenyl bromide and aniline. The black spheres represent the connections of the calixarene ring to the phenanthroline moiety.

Reaction energetics analysis and ensemble labelling

We analysed 142 simulations, in which a reaction occurred, to gain insights about the C–N coupling process. As the reaction happened spontaneously during the simulations, the energy profile could be obtained directly (see SI Fig. S2) and was used to identify the three states, educt, transition state, and product. These states were labelled based on the energies, with the educts considerably higher in energy than the products. The transition states were labelled as the highest energy point, before the energy drop associated with C–N coupling occurred. For technical details see methods section and SI Section 2.

The reaction energy was obtained by averaging the ensembles of the educt and product states and calculating the difference; it amounts to -212 ± 25 kJ mol^{-1} (See SI Sections 3 for details regarding uncertainty estimation). A sigmoid fit through the smoothened energy profile of each simulation (see SI, Fig. S2B and S2C) allowed identification and the calculation of the energy barrier to be 13 ± 9 kJ mol^{-1}. These GFN2-xTB reaction energies and structures are in excellent agreement with full DFT data, obtained with PBE0/def2-SVP/D3 (see SI Section 4, Fig. S3 and 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)[37] in cartesian coordinate space. However, this approach proved unsuccessful for the [Cu(C_8PhenMe$_6$)I] catalyst due to the difficulty in properly aligning this highly flexible system. The corresponding PCA does not show any separation between the three states (Fig. S4).

To achieve good separation between the three states, educt, transition state, and product, we developed a reduced internal coordinate description of the system (see Fig. 2A) to minimize the noise from highly correlated coordinates.[76] We used bond, angles, and torsion coordinates and 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 (Fig. 2A). An overview over the distribution of the internal coordinates which define the reduced model can be found in the SI, Fig. S5. This internal coordinate set nicely separates educts and products in the PCA space (Fig. 2B), but still shows overlap between educts and transition states. Analysing the loadings of the principal components (see SI Tables S3 and S4), we can see that the main contributions belong to the coordinates defining the reaction centre (C–N, NPh–C, NPh–Phena distances), as well as to the distances between the product and the cage, describing changes in the coordination at the Cu centre as the product is formed and the settling of the product in the cavity.

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 (See SI, Section 2), 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. This process is known as feature elimination. There are several methods to achieve this, and we tested a few of them using PCA-based dimensionality reduction approaches (see Fig. 3). The results show that the performance of PCA varies depending on the feature elimination technique used (Fig. 3A–C and SI Fig. S6). However, good separation, even within the ensembles, can still be achieved, depending on the method of feature elimination applied.

A second method, Linear Discriminant Analysis (LDA),[77] was also used for comparison. LDA is independent of feature selection and consistently yields excellent separation between groups (Fig. 3D and SI Fig. S6). While LDA produces compact,

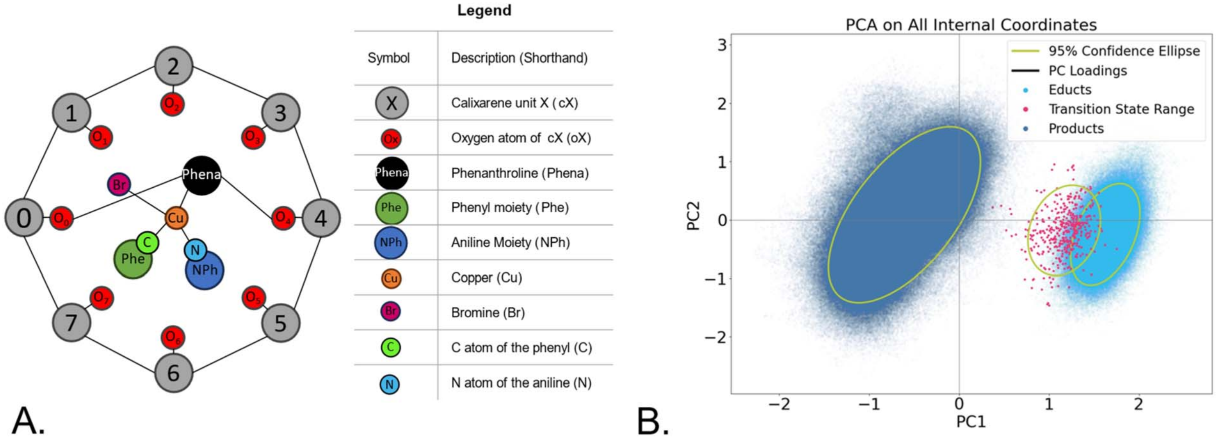

Fig. 2 (A). Schematic depiction of the [Cu(C$_8$PhenMe$_6$)I] intermediate that we denoted as "educt" here, showing the centres of mass used for the calculation of the reduced set of internal coordinates; (B). PCA performed on the reduced internal coordinates.

distinct ensembles, it does not provide separation within the ensembles themselves.

Analysis of the PCA loadings (see SI Table S3) revealed that amongst the top contributors to separating the three ensembles are the change in the distances defining the reaction centre (C–N, Phe–N, NPh–N, NPh-C and Phena-Cu-Br). Secondary features, such as the distance between the copper and O7, also plays an important role for the Logistic Regression (LR) classifier (Fig. 3A). The Random Forest (RF) identifies the C–N distance alongside Phe–N and NPh-C as important (Fig. 3B) with distinct product ensembles emerging. While in RF PC1 has the highest loading of all methods (0.68, see Table S3), the separation between the states is best for the LassoCV approach.

As the feature selection methods differ in the selected internal coordinates and performance (see also SI Fig. S7 for a visual display of features across the methods), we switched to a consensus model (see methods for details on the creation of the consensus model). This consensus approach retains smaller number of features (49 internal coordinates), namely only those that were found to be of high importance in 75% of all previously used ML methods (Fig. 4A), that is in PCA and LDA models. For sake of comparison, the highly important features

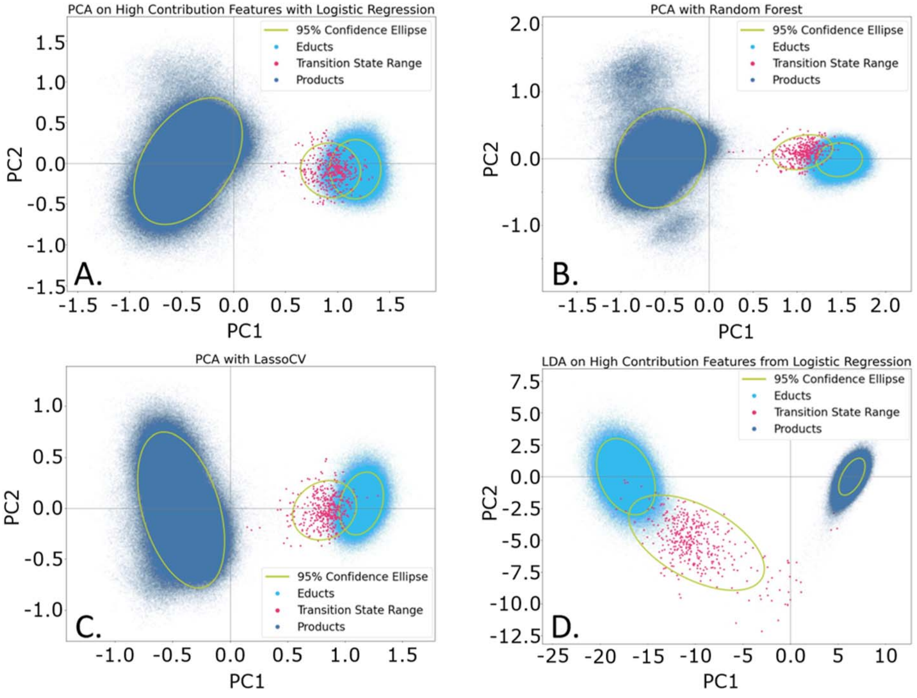

Fig. 3 Dimensionality reduction performance with various feature reduction methods: (A). PCA with Logistic Regression with mean-based cutoff; (B). PCA with Random Forest with mean-based cutoff; (C). PCA with LassoCV; (D). LDA with Recursive Feature Elimination with cross validation using a Logistic Regression classifier.

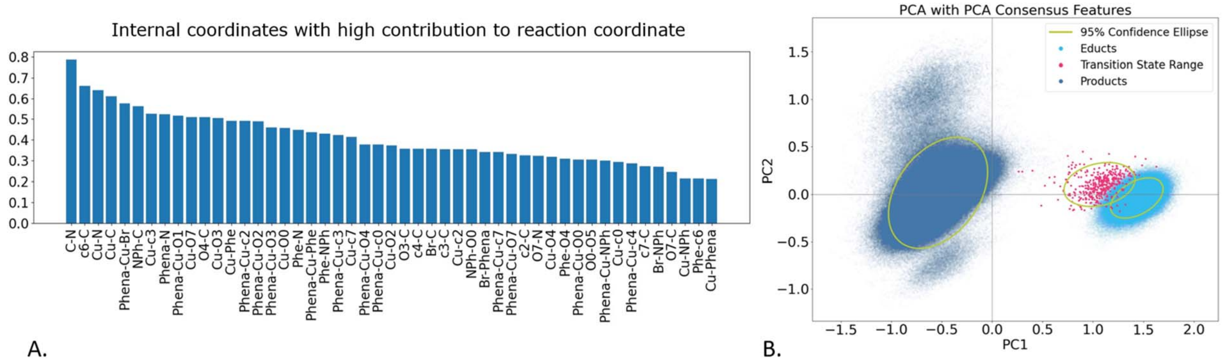

Fig. 4 Analysis of internal coordinates deemed as highly important by the consensus model. (A) Internal coordinates and their contribution; (B) PCA of the consensus features.

of a PCA-only and LDA-only consensus model are depicted in the SI in Fig. S8. It is important to note that individual models assign varying importance scores to each internal coordinate. As a result, the feature importance ranking derived from the overall consensus model is likely to differ from those of the individual models. We performed an additional PCA on the 49 consensus features (Fig. 4B). As evident from Table S3, in this model PC1 shows the highest retained variance (0.76) of all PCAs.

Judging the importance of the consensus model features (Fig. 4A), we see that the C–N bond distance is most important (in agreement with chemical intuition), alongside several distances and angles describing 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 (Fig. 4B), thereby outperforming any of the individual feature selection methods.

Time correlation-guided identification of slow system movements

While PCA focuses on the largest variance in the dataset, 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.

Employing tICA on the initial reduced internal coordinate set (Fig. 2A), resulted in a good separation of the product and educt states, mainly across the first independent component (IC1), as shown in Fig. 5A, where the values of IC1 and IC2 are plotted separately, and Fig. 5B, where they are plotted against each other and the structures are color-coded according to their labels. Yet the transition state ensemble cannot be fully separated from the educts. When taking into consideration IC2 (Fig. 5A and B), 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 (Fig. 5C) to the degree of freedom it corresponds to (Fig. 5B insert). A positive contribution (coloured in red) means that the respective feature values increase as the values of the IC increases, while a negative contribution (coloured in blue) means the feature values decreases as the IC values increase. The absolute value of a contribution (colour intensity) represents the importance of the feature in defining the IC. Please note that going from the educt to the product corresponds to a decrease in IC1.

IC1 reveals the changes at the reaction centre related the Cu adopting a planar configuration upon product formation, as evident for example by the Phena-Cu-Br angle (1st column) that increases with decreasing IC1 (1st column). Of particular interest are the increase in Cu–C and the Cu–N bond distances (1st column) when going to the products, accompanied to a shortening of C–N (5th column), as indicative of the reductive elimination step, as well as the strong contributions of the distances between the aniline product moiety (NPh) and the c0, c2, c6 and c7 calixarene units (4th column). The later hint at the formation of π–π interactions between the cage and the product. Additionally, a tilting of the calixarene cage can be inferred, when looking at the changes in the Phena-Cu-cX and Phena-Cu-oX angles: For example, Phena-Cu-O2, Phena-Cu-O3, and Phena-Cu-O4 (1st column), all located at the lower rim of the calixarene cage (Fig. 5B (insert)), anticorrelate with IC1, hence, they increase when the product is formed (decreasing IC1), whereas those on the lower rim of the cage, such as Phena-Cu-c6 (4th column) as well as Phena-Cu-O7 and Phena-Cu-c7 (5th column) decrease. This finding indicates the movement in opposing directions, when looking at units on opposite sides of the cage. IC2 acts to separate various conformers within the product ensemble, where we can see a difference in the position of the product in the calixarene cage, inferred from the strength of the contributions of the product-calixarene unit distances. To further corroborate π–π interactions, we clustered the product ensemble and performed an NCIPlot analysis[78] of the non-covalent interactions, which revealed weakly attracting van-der Waals interactions between the formed coupling product and the cage (see SI Fig. S16–S20).

Revealing causality in the reaction coordinate

While the correlation analysis shows which movements take place in a correlated fashion, it is also interesting to evaluate the

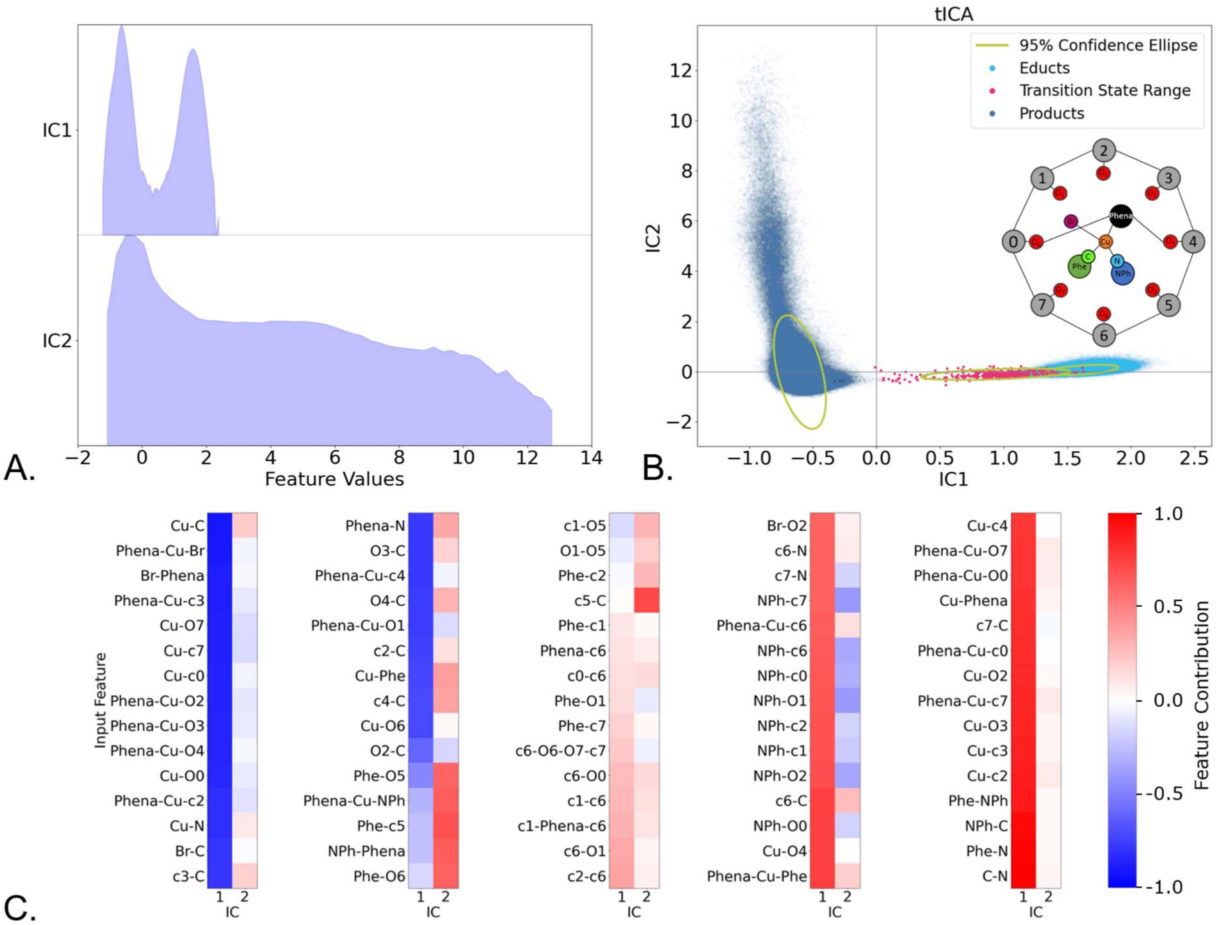

Fig. 5 Time-lagged independent component analysis of the MD trajectories, using a lag time of 20 fs. (A) Distribution of the structural ensembles (feature values) over the first two independent components; (B). Projection of the ensembles in the space of independent components 1 and 2 (IC1 and IC2), with the ensemble colouring performed *a posteriori*; (C). Normalized internal coordinate (feature) contribution to each independent component.

causality of these movements and how they propagate through the system. Given that tICA, a time-lag-based method, provided new insights into cage movement and product interaction, we applied the Granger Causality (GC) model, another time-lag-based approach, to assess the reaction trajectories for causal relationships. GC determines whether the prediction of a variable X in the future improves by including past events of Y. If this is the case, Y is found to Granger cause X.

With this causality model, we utilized the 49 consensus features (Fig. 4) and analysed each trajectory separately. Therefore, we can assess which each feature is influenced by which of the other 48 features; however, a complete analysis of all features quickly becomes impractical. Having identified C–N bond formation and changes in Cu-coordination as the most critical factors during product formation, we focused our GC analysis on the C–N feature. We analysed which of the remaining 48 features influenced the C–N feature and evaluated how often each of the 48 variables was found to Granger-cause C–N across all trajectories. For an overview over the 49 × 49 causality matrices generated for each reaction trajectory, the reader is referred to the GitHub.

In addition, we performed a hierarchical clustering of the consensus features (branches in Fig. 6), which tells us, what features are correlated. Further insight can be obtained by quantifying the correlation with a Pearson score, resulting in a clustermap (Fig. S9 SI), which also reorders the internal coordinates according to their correlation to each other. From this we can observe two distinct regions in the catalyst, corresponding largely to the upper and lower rim of the calixarene cage.

Combing the information of the correlation with Granger Causality, we analysed the causality for C–N coupling, *i.e.*, C–N bond shortening, not only for the individual features but also for the combined branches (Fig. 6). While the red arrows in Fig. 6 show the direction of causality the percentages indicate in how many trajectories C–N bond formation was caused by the respective feature or group of features. Fig. 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.

For example, in 11% of all trajectories C–N coupling is caused by a change in Phe-O4, O4-C, and c4-C distances (most

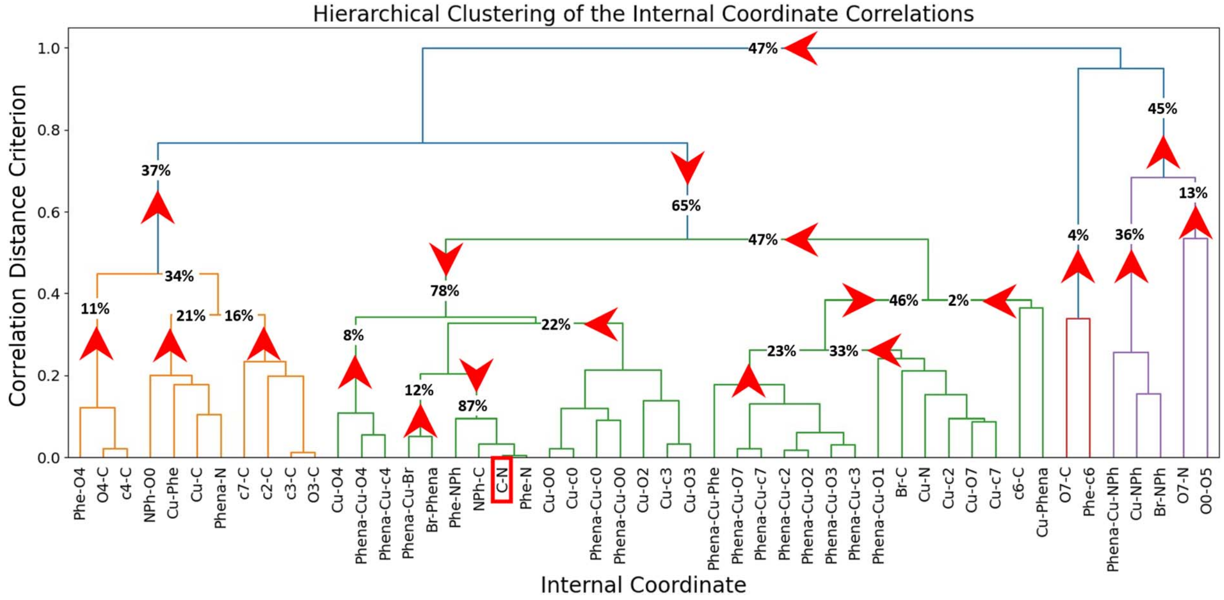

Fig. 6 Propagation cluster map of the system, leading to the C–N bond formation. The numbers represent the total causality effects uncovered by the respective coordinate(s) and the red arrows indicate the direction of causality.

left part of orange cluster, Fig. 6). Combining all features in the orange cluster, in 37% of all trajectories, the coupling is triggered by a move of features summarized in this cluster, consisting of the calixarene 2,3,4 and 7 to C distances as well as Cu-Phe distance. When we include information about the Phena-Cu-Br angle, alongside Cu and Br distances to NPh, from the purple cluster (right hand side of Fig. 6), as well as information from the red cluster, we can infer causality in 65% of all trajectories. The green cluster consists mostly of information related to the position of the calixarenes around the reaction centre. We see that changes in the Phena-Cu-calixarene angles (*e.g.*, Phena-Cu-O7, Phena-Cu-O2, …) on the upper and lower rim are correlated. Combined with the information from the clustermap (Fig. S9), we see that the angle with the upper rim decreases (positive Pearson correlation with C–N coupling, coloured in red), whereas those on the lower rim increases (negative Pearson correlation, coloured in blue), indicative of a tilt of the calixarene cage. 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 87% of the trajectories. Please note that because of the correlation of the individual features, percentages of the combined features do not necessarily equal the sum of the individual contribution.

In addition to C–N coupling, the Phena-Cu-Br angle was selected as a high importance feature by the consensus model (Fig. 4A), it's the 5th in the ranking and the first angle, and we subjected it to GC analysis. It also corresponds to an intuitive view of the change in coordination around the copper. From the correlation analysis and clustering, it relates closely to the C–N bond shortening and it directly Granger-causes C–N coupling 12% of the time. Conversely, we found that the C–N distance shortening does not cause the change in the Phena-Cu-Br angles, allowing us to deduce that the two movements happen either simultaneously or, more likely, the change in the angle precedes the C–N bond shortening. Furthermore, we observed that movements of the calixarene cage Granger-cause changes in the Phena-Cu-Br angle in 72% of all trajectories, establishing the sequence of cage movement – Phena-Cu-Br change – C–N-shortening. Due to the increasing complexity, we opted not to conduct further analysis of the GC matrices.

Quantification of reaction coordinates

While the consensus approach revealed the relevant internal degrees of freedom that define the reaction coordinate, as a next step, we sought to quantify it, by identifying ranges of individual features that separate the data into three ensembles. To achieve this, we used Decision Trees, which split ensembles by applying cut-offs to those coordinates that show the largest distribution differences between classes.

The Decision Tree in Fig. 7 has been trained on the whole dataset. To avoid biasing against the transition state ensemble, which contains significantly fewer structures, balanced weight is given to all classes *via* oversampling. For comparison, the results of an unbalanced tree can be found in Fig. S10.

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, indicative of an early transition state. 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 and educts, the C–N bond represents an effective metric, where values higher than 2.33 Å indicate an educt, while distances below indicate

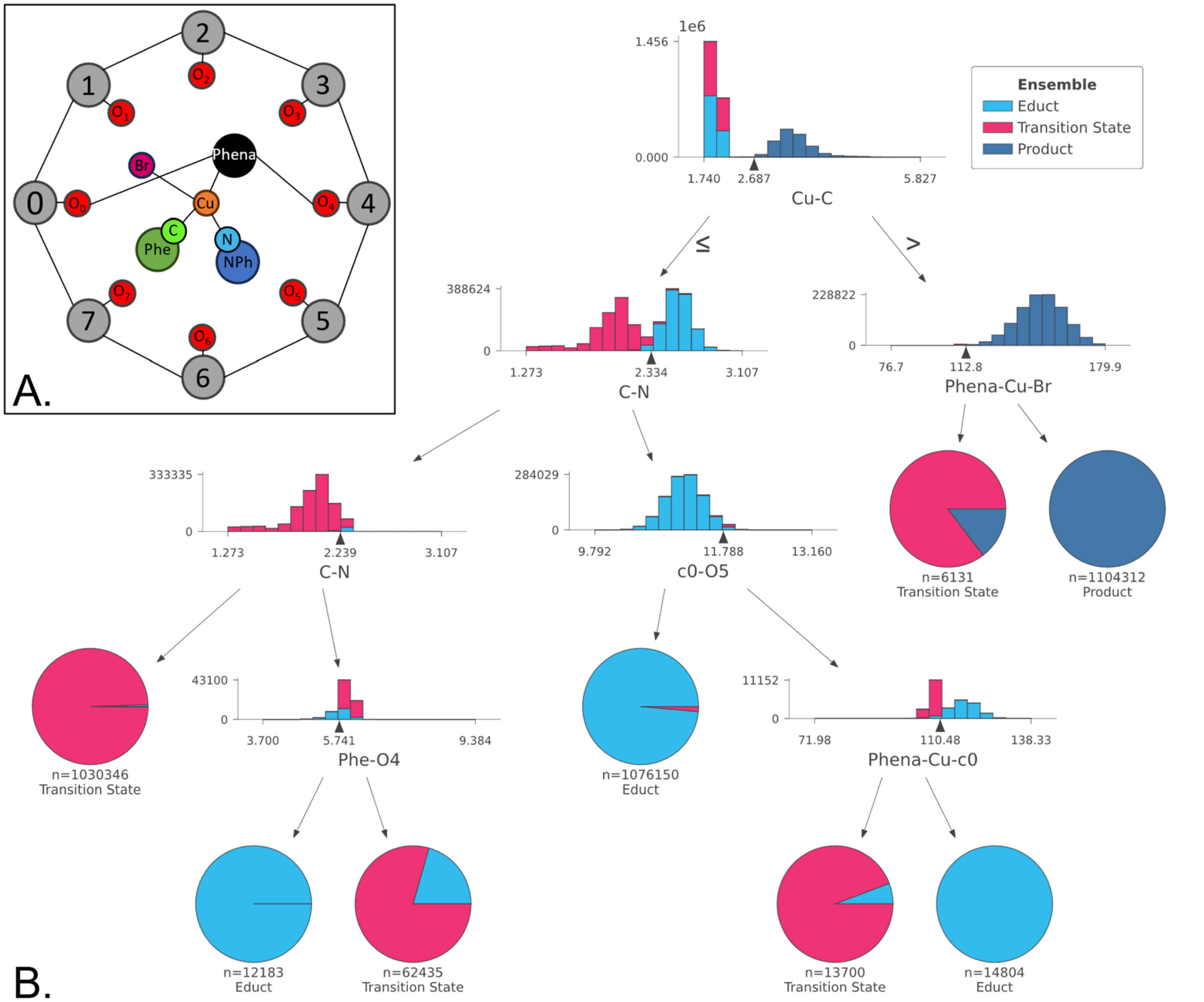

Fig. 7 (A). Model of the Cu-calix[8]arene catalyst with the reduced centres of mass; (B). Decision Tree classifier used to interpret the differences between the 3 structure ensembles, trained on the balanced classes. As indicated by the inequality operators, structures with values smaller or equal to the threshold are summarized in the left branch, those with larger values on in the right branch.

a transition state. The educt states that do exhibit a C–N bond distance similar to that of the transition state 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, they can be distinguished from the educts by the calixarene c0 and calixarene O5 distance greater than 11.79 Å and the angle defined by the phenanthroline bridge, Cu and c0 (Phena-Cu-c0) smaller than 110.5°.

A major shortcoming of Decision Trees is that their results depend on the initiation conditions. However, their reliability can be improved by utilizing many Decision Trees in a Random Forest (RF) classifier and averaging the results. In general, this improves accuracy, but reduces the interpretability. To overcome this limitation Decision Rules can be deduced from the results, providing a semantic understanding of the RF classifier. We used 30 Decision Trees, each trained on a subset of the data, to yield the RF. When applied to our dataset, this method provides rules for each of the three classes, as seen in Fig. 8, below. A complete diagram of the Decision Rules can be found in the SI, Fig. S11.

Notably, the Decision Rules approach identifies three distinct rule sets for defining a product. Color-coding these three product states in the PCA with consensus showed clear separation (see SI, Fig. S12), which was further confirmed by kmeans clustering, where only minor overlap occurred. Hence, these three states are distinguishable to some degree.

Discussion

Using semi-empirical quantum chemistry methods, we generated massive sampling of the reductive elimination step of the C–N coupling reaction with the [Cu(C_8PhenMe$_6$)I] catalyst,

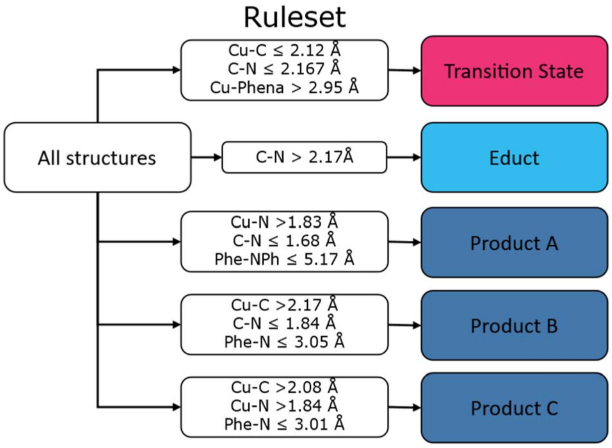

Fig. 8 Graphical depiction of the Decision Rules derived from the RF approach, Product A, B, C refers to different conformer ensembles within the product category.

through 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.

As we observed a convergence of the PCA with increased sampling (see SI, Fig. S13), we assumed the total simulation time to be sufficient. In addition, we observed spontaneous C–N coupling in 142 of the 152 trajectories, hence, 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. The computed reaction barrier amounted to 13 ± 9 kJ mol^{-1}, which is within 6 kJ mol^{-1} of that obtained from the full DFT trajectories (18 kJ mol^{-1}). Notably, the static quantum chemical (DFT) approach yielded a slightly higher barrier of 23 kJ mol^{-1},[7] indicating that dynamics lower the barrier. For the reaction energy, differences are more pronounced with −212 ± 25 kJ mol^{-1} obtained with QM/MM MD *vs.* −255 kJ mol^{-1} with static DFT.[7] However, these differences are due to the different conditions modelled: In the static DFT model, the studied structure was obtained at 0 K, which represents the bottom of the potential energy surface, while in this QM/MM MD study not only the average over all conformations is taken into account, but also the thermal energy, so the structures are not 0 K structures and therefore no minima on the potential energy surface. In addition, we assume that the product can undergo slow relaxation to lower energy conformations, which are too slow to observe in the timescales investigated here and which further contribute to the difference in reaction energies decreasing the variance between the static and the dynamic approach. Besides energetics, the reaction profile allowed for the categorization of the structures into three ensembles, namely educt, transition state, and product. This step was a key in improving the reaction coordinate recognition, as it allowed for the use of supervised learning methods to reduce the coordinate space. Due to the large energy difference between reactants and products, which renders the transformation irreversible in the context of unbiased MD at 300 K, we sampled a single chemical reaction per trajectory. Thus, there is one transition state structure before the large drop in energy and, as we have not biased our simulation, this structure should be on the dividing surface. Naturally, structures very similar to the transition state are likely to be found in the sampling leading up to the identified transition state due to recrossing events. Such transition state and transition state-like structures could probably be identified using a structural criterion and, in an iterative process, be removed from the educt ensemble. However, in the present case, we observe a good separation between transition state and reactant ensembles, also visible in the analysis of the internal coordinate distribution (SI Fig. S5), where for example, the C–N or the Phe–N coordinate show excellent separation of the three states. This finding indicates that there is no large transition state structure contamination within the reactant ensemble.

While PCA has been used in the past to determine the reaction coordinates, we have demonstrated that it is insufficient for a highly complex system, with many degrees of freedom. To yield any separation between the three states in the PCA, we had to transform cartesian coordinates to a set of internal coordinates, which we further reduced to minimize the number of highly correlated coordinates in the data set, thus reducing noise. Although this single step of our workflow is not fully automated at the moment, clear guidelines to obtain the reduced coordinate set can be applied, (i) usage of internal coordinates and (ii) representation of rigid groups such as phenyl by their centre of mass, which are applicable to describe any chemical system. 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.

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.

From looking at the cumulative results of all analyses performed on the reaction trajectories we deduced which movements took place during the reaction. From the PCA, LDA and the correlation analysis (clustermap), we saw a tilt in the calixarene cage. Likewise, we observed changes in the coordination around the Cu centre, as indicated for example by Phena-Cu-Br angle changes and the C–N bond shortening upon product formation.

Complementary to PCA, tICA was used to identify slow movements of the system. While tICA provided structural insights into the product conformer ensemble, it failed to distinguish educts from transition states. Thus, tICA was insufficient fully identify the reaction coordinate Finally, the tICA results suggest a settling of the product in the calixarene cage, governed by π–π stacking interactions, which was further corroborated by analysis of non-covalent interactions of representative structures (SI Section 12 Fig. S16–S20). These correlations can be then temporally defined by interpreting the Granger causality analysis. Through the causality analysis of changes in the C–N bond, we deduce that the reaction is Granger-caused by a movement in the reaction centre (such as a change in the Phena-Cu-Br angle) or a change in the calixarene cavity conformation, particularly calixarene units 2,3 and 4, whereas in turn changes in the Phena-Cu-Br angle are Granger-caused by changes in the calixarene conformation. Finally, the π–π stacking increased once the formed product reoriented itself inside the cavity. Therefore, we may create a sequential image of the reaction, schematically depicted in Fig. 9 as follows: (i). The calixarene cage tilts perpendicularly to the phenanthroline; (ii). The change in Cu-coordination takes place before the C–N distance shortening, but after the tilt of the cage; (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.

To the best of our knowledge, this study represents the first application of Granger Causality analysis to a complex chemical reaction – specifically, the C–N coupling reaction step catalysed by [Cu(C_8PhenMe$_6$)I] – and serves as a proof of principle. However, the model provides a simple interpretation of causality, as it analyses pairs of variables and not more complex composite variables. Transitioning to more advanced causal discovery methods, such as PC Momentary Conditional Independence (PCMCI)-based approaches,[79] would enable better interpretation of nonlinearity and the presence of hidden variables within the dataset. Nonetheless, such advancements are beyond the scope of this work.

To quantify the changes in internal coordinates, we employed a Decision Tree trained on the three ensembles, identifying the Cu–C and C–N distances as main feature for separating products from educts and transition states, respectively. While Decision Trees are sensitive on the initialisation condition and require size limitations to maintain interpretability, a Random Forest (RF) approach mitigates these limitations by aggregating multiple trees. Applying the Decision Rule method then allows for the semantic interpretation to a RF classifier, effectively separating the three ensembles. In addition, three distinct product conformers are identified, each defined by a unique set of decision rules and distinguishable in the PCA consensus features plot (Fig. S12) as well as in IC2 of the tICA analysis. Structural differences are related to changes in cage conformations. 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.[24] 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.[24] While they also used ML for feature selection, their focus is on physical chemical factors that impact reactivity. Our focus lies on the investigation 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. From all statistical analyses, we can see that the transition state structures tend to be close to the educt or even overlay with the educt space: for example, this is visible in the unsupervised PCA (Fig. 2), supervised PCA (Fig. 3) or tICA (Fig. 5), as well as in the histograms of all internal coordinates (SI Fig. S5), where educt and transition state distributions overlap for many internal coordinates, while being clearly separated from the product. We surmise that this is an indication that we have an early transition state here – information that can be utilised in further optimisation of the catalyst. While the reaction step under investigation does not represent the rate-determining step, as previously reported, the energy for release of the product from the catalyst is notably reduced by the presence of the calixarene cage, allowing for improved catalytic performance.[7] Our

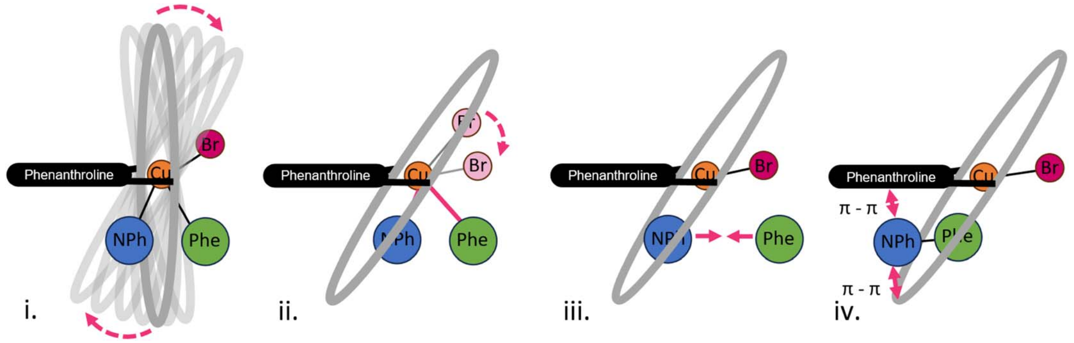

Fig. 9 Schematic representation of the movements corresponding to the reaction coordinate of the C–N coupling reaction. (i) Tilting movement of the calixarene cage; (ii) Change in coordination of the copper centre; (iii) Shortening of the C–N bond; (iv) π–π stacking effects stabilise the product in the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

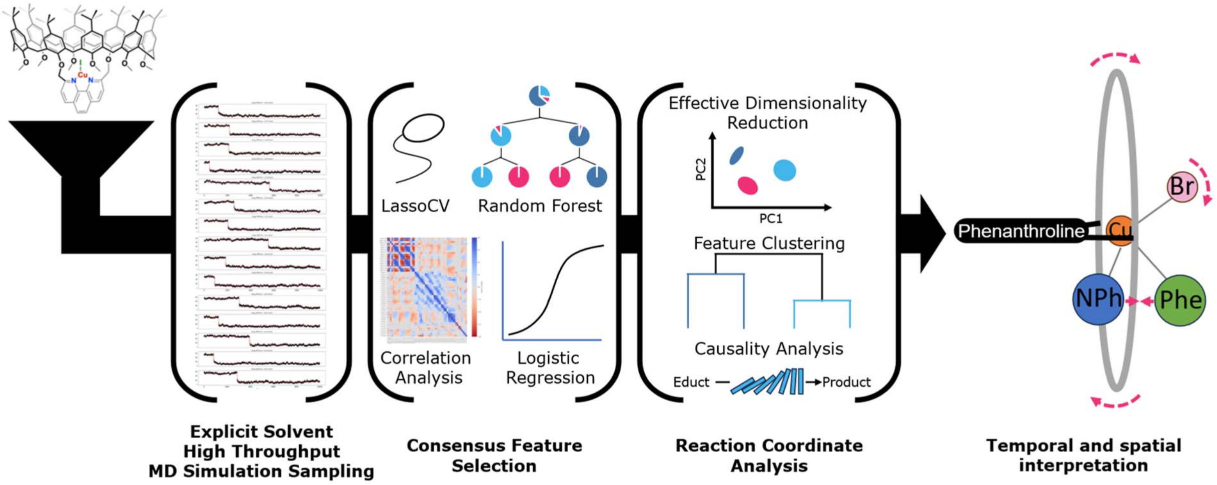
LassoCV
Random Forest
Correlation Analysis
Logistic Regression
Effective Dimensionality Reduction
PC2
PC1
Feature Clustering
Causality Analysis
Educt
Product
Phenanthroline
Cu
Br
NPh
Phe
Explicit Solvent High Throughput MD Simulation Sampling
Consensus Feature Selection
Reaction Coordinate Analysis
Temporal and spatial interpretation

Fig. 10 General workflow to identify and quantify the reaction coordinate by analyzing correlations and causality in an ensemble of reaction trajectories.

analysis revealed that the reaction starts with a tilt in the cage. Therefore, modification of this moiety distal to the reaction centre can affect reactivity and even selectivity, similar to the allosteric effects seen in natural enzymes. Experimentally, modifications could include changing the size and rigidity of the cage, or introducing asymmetric modifications to one side of the cage. All of these efforts can result in greater control over substrate selectivity. Identifying possible allosteric sites in other systems could help advance supramolecular chemistry in rational design, which is of particular interest for homogeneous catalysis, where such insights may lead to greater substrate selectivity or more efficient catalysts.

Our study presents a proof-of-concept approach to the discovery of complex reaction coordinates that is transferable to any chemical reaction. The presented methodology is not only applicable to the study of single reaction steps, but can be extended to reactions with competing reaction pathways that determine selectivity: If the reaction energy profile is different, then the energy can be used as a simple criterion to group the trajectories and separate the pathways. In case, the energy differences are similar and only the product structures are different, this will be highlighted by PCA or tICA, allowing separation using a clustering approach. This information can then be leveraged to either separate the sampling data and categorize the reaction trajectories or to perform new, biased or restrained simulations which direct the reaction mechanism through a single pathway.

Our protocol can be integrated into computational chemistry workflows to investigate the reactivity of complex systems under *operando* conditions. By identifying secondary contributions to the reaction coordinate, such as remote conformational changes, it can guide further studies and inform the design of reactivities and selectivity by tracing the impact of modifications on the reaction profile under *operando* conditions. Consequently, additional insights can be obtained from reaction trajectories with minimal additional computational effort.

Conclusion

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

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 provides insights into the intricate dynamic nature of the macrocyclic transition-metal catalyst in explicit solvent.

Interpretable machine learning techniques have proven invaluable in dealing with the vast amount of data, because of their ability to trace results back to structural changes. However, a consensus model is needed to identify the internal coordinates with the highest contribution to separating educt, transition state and product, and to eliminate the inherent variability and instability of individual ML approaches.

By performing a causality analysis of the internal coordinates of the consensus model, 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. This information, pinpointing the exact source (group of atoms) that triggers the reaction, allows the experimental chemist to generate testable hypotheses to enhance reactivity, for example by suitable chemical modifications. By checking the outcome of Decision Trees and Decision Rules run on this modified system, we can gauge the impact of a specific change on the reaction coordinate. The methodology is implemented in a python script and presented in the Jupyter notebook provided on github (see Data availability Section). The analysis is largely automated, requiring only a few inputs from the user, namely: (i) system trajectories; (ii) coordinates which should be considered for dimensionality reduction *e.g.* centres of mass for rings; (iii) labelling criteria for the various states of the system; (iv) cutoff values for the consensus model. Notably, the labelling of the states can be done on any descriptor of the system, including energetic criteria, as well as partial charges or other molecular properties. As a result, it can be readily applied to other systems as a digital tool. By limiting the analysis to structural changes in the system, we believe this technique is accessible to non-expert users. The results reflect coordinate changes during the reaction, which makes them easy to understand for a general chemist.

Our methodology was demonstrated on a highly flexible Cu-calix[8]arene catalyst. However, the approach supports the exploration of both covalent and non-covalent interactions, allowing for the applicability to the investigation of other phenomena such as cluster formation and aggregation. Nevertheless, a requirement is that structures can be labelled, using either energetic criteria or any molecular feature. If that is given, it offers a broadly applicable framework for probing reaction coordinates in a variety of dynamic chemical systems, ranging from small (in)organic complexes to more intricate biomolecular structures.

Methods

Workflow for determination of complex reaction coordinates

The multistep protocol developed to investigate the C–N coupling dynamics with the [Cu(C_8PhenMe$_6$)I] catalyst is highlighted in Fig. 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[80] using the link-bond method to describe the bonds crossing between QM and MM.[81,82] The simulation parameters were set up using the GAFF[83] forcefield,[84] using the PyConSolv[19] 1.0.0 tool, with default settings. For the geometry optimization, the PBE0 functional[85] was used with the def2-SVP basis set[86] and D3 dispersion corrections[87] in implicit chloroform, using CPCM.[88] The system was solvated in a cubic periodic box with 1708 chloroform molecules. For detailed information regarding the simulation parameters, see the SI. The 54 atoms at the centre of the calixarene cage were included in the QM zone (see SI Fig. S1). The quantum mechanical calculations were performed at two distinct levels. The semi-empirical method GFN2-xTB was utilized, providing a vast speed-up of the QM calculation. As semi-empirical methods require benchmarking,[89–91] a full DFT reference was used, with PBE0/def2-SVP/D3, using Turbomole, showing very good agreement.[92] A detailed comparison can be found in the SI (Section 4).

We used chloroform in our simulations instead of toluene, which was chosen experimentally. Chloroform and toluene have comparable permittivities (ε_r 4.8 and 2.4),[93] both typical of apolar solvents. Thus, both solvents provide similar apolar environments, consistent with the observation that the catalyst is active in different media. Experimentally, the reaction also works in tetrachloroethane but toluene was selected as a more benign and environmentally safe option. As there is no evidence for strong solvent effects on reactivity, the choice of solvent is unlikely to significantly alter the system. In simulations, chloroform was preferred because its smaller size enables faster equilibration than toluene.

To generate an appropriate starting structure the system was equilibrated and a 100 ns of a 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 the QM method. Among these runs, 142 simulations captured the reductive elimination step and were utilized for evaluating the reaction barrier and structural analysis of the reaction. All 152 trajectories were used for labelling and evaluation of the reaction energy (see below).

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 Section 2 for details), that is, all frames before the TS region were considered as educt, all frames following the TS region as product. To assess whether this is a viable labelling of the states, we analysed the distribution of internal coordinates in Fig. S5, where we can see a clear separation of educt, transition state, and product for example in the decisive C–N coordinate, but also in the Phe–N bond length distribution.

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 (see Section 2 of the SI). As DFT sampling on a similar scale as GFN2-xTB was not achievable, the ensembles were too small to obtain the reaction energy by averaging over them. Thus, to evaluate the reaction energy from the DFT simulations, we fitted sigmoids between the educts and product. By subtracting the highest sigmoid value from the lowest we were able to calculate a reaction energy and compared it to that obtained from GFN2-xTB using the same approach. As further confirmation of the accuracy of the transition state labelling, we confirmed that the structures in the ensemble resemble that of a transition state optimized using eigenvector following in a static approach (see SI Section 4 and Fig. S3).

Identification of the reaction coordinate

We resorted to different coordinate systems to describe the reaction coordinate. We generated a set of fully redundant internal coordinates for all trajectories, using the MDanalysis package.[94] This has the advantage of removing the issue of noise due to alignment artefacts, yet introduces more correlational effects.[76] We also generated a reduced set of internal coordinates, describing highly rigid chemical moieties by their centre of mass (see SI Section 5 for details).

We utilized Principal Component Analysis (PCA)[37] and Time-lagged Independent Component Analysis (tICA)[38] as dimensionality reduction techniques to identify coordinates that separate the states. The two unsupervised methods complement each other in regards to addressing the variances present in the dataset.[95,96] For supervised dimensionality reduction, we opted for Linear Discriminant Analysis (LDA)[77] 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[97] (RFECV) using both Random Forest[98] (RF) and Logistic Regression (LR)[99] as classifier models, and Lasso[100] with Cross Validation (LassoCV), all using fivefold stratified cross-validation.[101] The RF and Logistic Regression classifiers were additionally evaluated by selecting only the top 10% of features. When subjecting the highlighted features to dimensionality reduction using PCA or LDA, we observed varying importance of the features to the principal components (SI, Tables S3 and S4). To alleviate this discrepancy between the results and increase separation performance, we defined a consensus model, which combines the results of all previously mentioned approaches and highlights features that are consistently found across all models. These features were identified by performing a PCA and LDA on each of the models and examining the loadings. 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 75% of the elimination models to be deemed significant. This process yielded a set of 49 internal coordinates for the PCA. These coordinates were hierarchically clustered to group together correlated movements[102] and then subjected to causal inference analysis (see statistical model for causality inference).[102]

In tandem with automated feature reduction methodologies, we leveraged Decision Trees to quantify the most relevant features from our dataset. These Decision Trees underwent training with class balancing that is augmenting the ensembles *via* oversampling. Furthermore, we adopted a Decision Rule framework, employing the skoperules[103] 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. We eliminated trajectories with few educt structures (less than 100), thus having a total of 139 reaction sampling events. To infer Granger Causality[74] (GC) we followed the protocol outlined by Toda and Yamamoto.[75] This involved performing the augmented Dickey–Fuller[104] and Kwiatkowski–Phillips–Schmidt–Shin[105] 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 GitHub for the critical and test statistics for each test and trajectory).[106] A multivariate vector autoregression model (VAR)[107,108] 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.[107,109] The correlated time series were checked for cointegration using the Johansen test.[110] Finally, we calculated a GC matrix for each feature, for every simulation, resulting in a total of 139 matrices, using VAR models 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,[111] 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 on GitHub, along with the results of all statistical tests.

Limitations of Granger causality statistical model

While the GC model has been shown to be a reliable approach for identifying causality, it suffers from some inherent limitations. The model requires the dataset to be presented in the form of time-series with constant variance and mean (stationary). To this end, the time-series stationarity must be verified and, depending on the results, rendered stationary through common statistical approaches such as differentiation. Another limitation in the nature of the GC analysis is the influence of hidden variables that may not have been taken into account in the initial dataset. This implies that the dataset must be carefully chosen, either manually curated or through automated means, such as feature selection. Moreover, the GC test also looks at first degree causality between the selected features, meaning that while we observe one variable influencing another, we cannot rule out that the influencing variable was not provoked by a 2nd variable beforehand. This can be overcome by performing a GC analysis on each possible pair of features and interpreting the results; however, this becomes difficult as the number of important features increases.

Code availability

The python code used for analysis is made available on Github (**https://github.com/PodewitzLab/MLReactCoord/releases/tag/v1.0.0**) 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 QMCF 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.

Conflicts of interest

There are no conflicts of interest to declare.

Data availability

The *p*-values for the statistical analysis are provided on the supplies Github repository, alongside two example trajectories (**https://github.com/PodewitzLab/MLReactCoord/releases/tag/v1.0.0**). The full set of dry trajectory data used for analysis is provided on Zenodo under **https://doi.org/10.5281/zenodo.14886709**.

Supplementary information: it contains the detailed simulation protocol, information on the labeling of data as well as statistics and performance of the methodology, robustness and performance of sampling, an overview over the distribution of all internal coordinates, details about the statistical analyses, detailed analysis of the product (distribution) and solvation effects. See DOI: **https://doi.org/10.1039/d5dd00216h**.

Acknowledgements

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

References

1 D. Ringe and G. A. Petsko, How Enzymes Work, *Science*, 2008, **320**(5882), 1428–1429, DOI: **10.1126/science.1159747**.

2 T. S. Koblenz, J. Wassenaar and J. N. H. Reek, Reactivity within a Confined Self-Assembled Nanospace, *Chem. Soc. Rev.*, 2008, **37**(2), 247–262, DOI: **10.1039/B614961H**.

3 M. Raynal, P. Ballester, A. Vidal-Ferran and P. W. N. M. van Leeuwen, Supramolecular Catalysis. Part 1: Non-Covalent Interactions as a Tool for Building and Modifying Homogeneous Catalysts, *Chem. Soc. Rev.*, 2014, **43**(5), 1660–1733, DOI: **10.1039/C3CS60027K**.

4 M. Raynal, P. Ballester, A. Vidal-Ferran and P. W. N. M. van Leeuwen, Supramolecular Catalysis. Part 2: Artificial Enzyme Mimics, *Chem. Soc. Rev.*, 2014, **43**(5), 1734–1787, DOI: **10.1039/C3CS60037H**.

5 S. Pachisia and R. Gupta, Supramolecular Catalysis: The Role of H-Bonding Interactions in Substrate Orientation and Activation, *Dalton Trans.*, 2021, **50**(42), 14951–14966, DOI: **10.1039/D1DT02131A**.

6 G. Olivo, G. Capocasa, D. D. Giudice, O. Lanzalunga and S. D. Stefano, New Horizons for Catalysis Disclosed by Supramolecular Chemistry, *Chem. Soc. Rev.*, 2021, **50**(13), 7681–7724, DOI: **10.1039/D1CS00175B**.

7 R. A. Talmazan, J. Refugio Monroy, F. del Río-Portilla, I. Castillo and M. Podewitz, Encapsulation Enhances the Catalytic Activity of C–N Coupling: Reaction Mechanism of a Cu(I)/Calix[8]Arene Supramolecular Catalyst, *ChemCatChem*, 2022, **14**(20), e202200662, DOI: **10.1002/cctc.202200662**.

8 E. Guzmán-Percástegui, J. Hernández D. and I. Castillo, Calix[8]Arene Nanoreactor for Cu(ɪ)-Catalysed C–S Coupling, *Chem. Commun.*, 2016, **52**(15), 3111–3114, DOI: **10.1039/C5CC09232A**.

9 G. Sciortino and F. Maseras, Computational Study of Homogeneous Multimetallic Cooperative Catalysis, *Top. Catal.*, 2022, **65**(1), 105–117, DOI: **10.1007/s11244-021-01493-2**.

10 L. Artús Suàrez, D. Balcells and A. Nova, Computational Studies on the Mechanisms for Deaminative Amide Hydrogenation by Homogeneous Bifunctional Catalysts, *Top. Catal.*, 2022, **65**(1), 82–95, DOI: **10.1007/s11244-021-01542-w**.

11 G. O. Jones, P. Liu, K. N. Houk and S. L. Buchwald, Computational Explorations of Mechanisms and Ligand-Directed Selectivities of Copper-Catalyzed Ullmann-Type Reactions, *J. Am. Chem. Soc.*, 2010, **132**(17), 6205–6213, DOI: **10.1021/ja100739h**.

12 P.-F. Larsson, C.-J. Wallentin and P.-O. Norrby, Mechanistic Aspects of Submol% Copper-Catalyzed C–N Cross-Coupling, *ChemCatChem*, 2014, **6**(5), 1277–1282, DOI: **10.1002/cctc.201301088**.

13 N. Fey and J. M. Lynam, Computational Mechanistic Study in Organometallic Catalysis: Why Prediction Is Still a Challenge, *WIREs Comput. Mol. Sci.*, 2022, **12**(4), e1590, DOI: **10.1002/wcms.1590**.

14 J. N. Harvey, F. Himo, F. Maseras and L. Perrin, Scope and Challenge of Computational Methods for Studying Mechanism and Reactivity in Homogeneous Catalysis, *ACS Catal.*, 2019, **9**(8), 6803–6813, DOI: **10.1021/acscatal.9b01537**.

15 M. Besora, A. A. C. Braga, G. Ujaque, F. Maseras and A. Lledós, 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, DOI: **10.1007/s00214-010-0823-6**.

16 M. Podewitz, S. Sen and M. R. Buchmeiser, 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, DOI: **10.1021/acs.organomet.1c00229**.

17 O. Eisenstein, G. Ujaque and A. Lledós, What Makes a Good (Computed) Energy Profile? In New Directions in the

Modeling of Organometallic Reactions, *Topics in Organometallic Chemistry*, ed. Lledós, A. and Ujaque, G., Springer International Publishing, Cham, 2020, pp. 1–38, DOI: 10.1007/3418_2020_57.

18 P. Pracht, S. Grimme, C. Bannwarth, F. Bohle, S. Ehlert, G. Feldmann, J. Gorges, M. Müller, T. Neudecker, C. Plett, S. Spicher, P. Steinbach, P. A. Wesołowski and F. Zeller, CREST—A Program for the Exploration of Low-Energy Molecular Chemical Space, *J. Chem. Phys.*, 2024, **160**(11), 114110, DOI: 10.1063/5.0197592.

19 R. A. Talmazan and M. Podewitz, PyConSolv: A Python Package for Conformer Generation of (Metal-Containing) Systems in Explicit Solvent, *J. Chem. Inf. Model.*, 2023, **63**(17), 5400–5407, DOI: 10.1021/acs.jcim.3c00798.

20 G. N. Simm, P. L. Türtscher and M. Reiher, Systematic Microsolvation Approach with a Cluster-Continuum Scheme and Conformational Sampling, *J. Comput. Chem.*, 2020, **41**(12), 1144–1155, DOI: 10.1002/jcc.26161.

21 M. Steiner, T. Holzknecht, M. Schauperl and M. Podewitz, Quantum Chemical Microsolvation by Automated Water Placement, *Molecules*, 2021, **26**(6), 1793, DOI: 10.3390/molecules26061793.

22 J. Joy and D. H. Ess, 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, DOI: 10.1021/jacs.3c01196.

23 D. H. Ess, Quasiclassical Direct Dynamics Trajectory Simulations of Organometallic Reactions, *Acc. Chem. Res.*, 2021, **54**(23), 4410–4422, DOI: 10.1021/acs.accounts.1c00575.

24 J. Joy, A. J. Schaefer, M. S. Teynor and D. H. Ess, Dynamical Origin of Rebound *versus* Dissociation Selectivity during Fe-Oxo-Mediated C–H Functionalization Reactions, *J. Am. Chem. Soc.*, 2024, **146**(4), 2452–2464, DOI: 10.1021/jacs.3c09891.

25 Z. Yang, C. S. Jamieson, X.-S. Xue, M. Garcia-Borràs, T. Benton, X. Dong, F. Liu and K. N. Houk, Mechanisms and Dynamics of Reactions Involving Entropic Intermediates, *Trends Chem.*, 2019, **1**(1), 22–34, DOI: 10.1016/j.trechm.2019.01.009.

26 A. Warshel and M. Karplus, Calculation of Ground and Excited State Potential Surfaces of Conjugated Molecules. I. Formulation and Parametrization, *J. Am. Chem. Soc.*, 1972, **94**(16), 5612–5625, DOI: 10.1021/ja00771a014.

27 A. Warshel and M. Levitt, Theoretical Studies of Enzymic Reactions: Dielectric, Electrostatic and Steric Stabilization of the Carbonium Ion in the Reaction of Lysozyme, *J. Mol. Biol.*, 1976, **103**(2), 227–249, DOI: 10.1016/0022-2836(76)90311-9.

28 J. Gao, Hybrid Quantum and Molecular Mechanical Simulations: An Alternative Avenue to Solvent Effects in Organic Chemistry, *Acc. Chem. Res.*, 1996, **29**(6), 298–305, DOI: 10.1021/ar950140r.

29 H. Fu, H. Bian, X. Shao and W. Cai, Collective Variable-Based Enhanced Sampling: From Human Learning to Machine Learning, *J. Phys. Chem. Lett.*, 2024, **15**(6), 1774–1783, DOI: 10.1021/acs.jpclett.3c03542.

30 S. Bhakat, Collective Variable Discovery in the Age of Machine Learning: Reality, Hype and Everything in Between, *RSC Adv.*, 2022, **12**(38), 25010–25024, DOI: 10.1039/D2RA03660F.

31 A. Nandy, C. Duan, M. G. Taylor, F. Liu, A. H. Steeves and H. J. Kulik, Computational Discovery of Transition-Metal Complexes: From High-Throughput Screening to Machine Learning, *Chem. Rev.*, 2021, **121**(16), 9927–10000, DOI: 10.1021/acs.chemrev.1c00347.

32 G. Montavon, W. Samek and K.-R. Müller, Methods for Interpreting and Understanding Deep Neural Networks, *Digit. Signal Process.*, 2018, **73**, 1–15, DOI: 10.1016/j.dsp.2017.10.011.

33 H. Chen, B. Roux and C. Chipot, Discovering Reaction Pathways, Slow Variables, and Committor Probabilities with Machine Learning, *J. Chem. Theory Comput.*, 2023, **19**(14), 4414–4426, DOI: 10.1021/acs.jctc.3c00028.

34 A. Ma and A. R. Dinner, Automatic Method for Identifying Reaction Coordinates in Complex Systems, *J. Phys. Chem. B*, 2005, **109**(14), 6769–6779, DOI: 10.1021/jp045546c.

35 T. S. van Erp, M. Moqadam, E. Riccardi and A. Lervik, Analyzing Complex Reaction Mechanisms Using Path Sampling, *J. Chem. Theory Comput.*, 2016, **12**(11), 5398–5410, DOI: 10.1021/acs.jctc.6b00642.

36 K. Wilke, S. Tao, S. Calero, A. Lervik and T. S. Van Erp, NaCl Dissociation Explored Through Predictive Power Path Sampling Analysis, *J. Chem. Theory Comput.*, 2025, **21**(9), 4604–4614, DOI: 10.1021/acs.jctc.5c00054.

37 S. Wold, K. Esbensen and P. Geladi, Principal Component Analysis, *Chemom. Intell. Lab. Syst.*, 1987, **2**(1), 37–52, DOI: 10.1016/0169-7439(87)80084-9.

38 L. Molgedey and H. G. Schuster, Separation of a Mixture of Independent Signals Using Time Delayed Correlations, *Phys. Rev. Lett.*, 1994, **72**(23), 3634–3637, DOI: 10.1103/PhysRevLett.72.3634.

39 H. Kwon, Z. A. Ali and B. M. Wong, Harnessing Semi-Supervised Machine Learning to Automatically Predict Bioactivities of Per- and Polyfluoroalkyl Substances (PFASs), *Environ. Sci. Technol. Lett.*, 2023, **10**(11), 1017–1022, DOI: 10.1021/acs.estlett.2c00530.

40 B. Ozgode Yigin and G. Saygili, Effect of Distance Measures on Confidences of T-SNE Embeddings and Its Implications on Clustering for scRNA-Seq Data, *Sci. Rep.*, 2023, **13**(1), 6567, DOI: 10.1038/s41598-023-32966-x.

41 E. Frasnetti, I. Cucchi, S. Pavoni, F. Frigerio, F. Cinquini, S. A. Serapian, L. F. Pavarino and G. Colombo, Integrating Molecular Dynamics and Machine Learning Algorithms to Predict the Functional Profile of Kinase Ligands, *J. Chem. Theory Comput.*, 2024, **20**(20), 9209–9229, DOI: 10.1021/acs.jctc.4c01097.

42 A. Amadei, A. B. M. Linssen and H. J. C. Berendsen, Essential Dynamics of Proteins, *Proteins: Struct., Funct., Bioinf.*, 1993, **17**(4), 412–425, DOI: 10.1002/prot.340170408.

43 G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis and F. Noé, Identification of Slow Molecular Order

Parameters for Markov Model Construction, *J. Chem. Phys.*, 2013, **139**(1), 015102, DOI: **10.1063/1.4811489**.

44 C. R. Schwantes and V. S. Pande, 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, DOI: **10.1021/ct300878a**.

45 S. R. Hare, L. A. Bratholm, D. R. Glowacki and B. K. Carpenter, Low Dimensional Representations along Intrinsic Reaction Coordinates and Molecular Dynamics Trajectories Using Interatomic Distance Matrices, *Chem. Sci.*, 2019, **10**(43), 9954–9968, DOI: **10.1039/C9SC02742D**.

46 Reaction Space Projector (ReSPer) for Visualizing Dynamic Reaction Routes Based on Reduced-Dimension Space, *Topics in Current Chemistry*, **https://doi.org/10.1007/s41061-022-00377-7** accessed 2024-06-04.

47 G. W. Imbens and D. B. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press, Cambridge, 2015, DOI: **10.1017/CBO9781139025751**.

48 Q. Xing, C. Wu, F. Chen, J. Liu, P. Pradhan, B. A. Bryan, T. Schaubroeck, L. R. Carrasco, A. Gonsamo, Y. Li, X. Chen, X. Deng, A. Albanese, Y. Li and Z. Xu, Intranational Synergies and Trade-Offs Reveal Common and Differentiated Priorities of Sustainable Development Goals in China, *Nat. Commun.*, 2024, **15**(1), 2251, DOI: **10.1038/s41467-024-46491-6**.

49 P. J. Cárdenas-García, J. G. Brida and V. Segarra, Modeling the Link between Tourism and Economic Development: Evidence from Homogeneous Panels of Countries, *Humanit. Soc. Sci. Commun.*, 2024, **11**(1), 1–12, DOI: **10.1057/s41599-024-02826-8**.

50 R. K. Kaufmann, D. Newberry, C. Xin and S. Gopal, Feedbacks among Electric Vehicle Adoption, Charging, and the Cost and Installation of Rooftop Solar Photovoltaics, *Nat. Energy*, 2021, **6**(2), 143–149, DOI: **10.1038/s41560-020-00746-w**.

51 T. Tanaka and J. Guo, International Price Volatility Transmission and Structural Change: A Market Connectivity Analysis in the Beef Sector, *Humanit. Soc. Sci. Commun.*, 2020, **7**(1), 1–13, DOI: **10.1057/s41599-020-00657-x**.

52 T. Ouyang, F. Liu and B. Huang, Dynamic Econometric Analysis on Influencing Factors of Production Efficiency in Construction Industry of Guangxi Province in China, *Sci. Rep.*, 2022, **12**(1), 17509, DOI: **10.1038/s41598-022-22374-y**.

53 T. Le, Increased Impact of the El Niño–Southern Oscillation on Global Vegetation under Future Warming Environment, *Sci. Rep.*, 2023, **13**(1), 14459, DOI: **10.1038/s41598-023-41590-8**.

54 J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. D. Mahecha, J. Muñoz-Marí, E. H. van Nes, J. Peters, R. Quax, M. Reichstein, M. Scheffer, B. Schölkopf, P. Spirtes, G. Sugihara, J. Sun, K. Zhang and J. Zscheischler, Inferring Causation from Time Series in Earth System Sciences, *Nat. Commun.*, 2019, **10**(1), 2553, DOI: **10.1038/s41467-019-10105-3**.

55 M. Kretschmer, D. Coumou, J. F. Donges and J. Runge, Using Causal Effect Networks to Analyze Different Arctic Drivers of Midlatitude Winter Circulation, *J. Clim.*, 2016, **29**(11), 4069–4081, DOI: **10.1175/JCLI-D-15-0654.1**.

56 U. Triacca, 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, DOI: **10.1007/s00704-004-0112-1**.

57 M. C. McGraw and E. A. Barnes, Memory Matters: A Case for Granger Causality in Climate Variability Studies, *J. Clim.*, 2018, **31**(8), 3289–3300, DOI: **10.1175/JCLI-D-17-0334.1**.

58 S. M. Hill, L. M. Heiser, T. Cokelaer, M. Unger, N. K. Nesser, D. E. Carlin, Y. Zhang, A. Sokolov, E. O. Paull, C. K. Wong, K. Graim, A. Bivol, H. Wang, F. Zhu, B. Afsari, L. V. Danilova, A. V. Favorov, W. S. Lee, D. Taylor, C. W. Hu, B. L. Long, D. P. Noren, A. J. Bisberg, G. B. Mills, J. W. Gray, M. Kellen, T. Norman, S. Friend, A. A. Qutub, E. J. Fertig, Y. Guan, M. Song, J. M. Stuart, P. T. Spellman, H. Koeppl, G. Stolovitzky, J. Saez-Rodriguez and S. Mukherjee, Inferring Causal Molecular Networks: Empirical Assessment through a Community-Based Effort, *Nat. Methods*, 2016, **13**(4), 310–318, DOI: **10.1038/nmeth.3773**.

59 J. Walter, Establishing Microbiome Causality to Tackle Malnutrition, *Nat. Microbiol.*, 2024, **9**(4), 884–885, DOI: **10.1038/s41564-024-01653-6**.

60 K. J. Friston, L. Harrison and W. Penny, Dynamic Causal Modelling, *Neuroimage*, 2003, **19**(4), 1273–1302, DOI: **10.1016/S1053-8119(03)00202-7**.

61 A. Duggento, L. Passamonti, G. Valenza, R. Barbieri, M. Guerrisi and N. Toschi, Multivariate Granger Causality Unveils Directed Parietal to Prefrontal Cortex Connectivity during Task-Free MRI, *Sci. Rep.*, 2018, **8**(1), 5571, DOI: **10.1038/s41598-018-23996-x**.

62 J. Zheng, K. L. Anderson, S. L. Leal, A. Shestyuk, G. Gulsen, L. Mnatsakanyan, S. Vadera, F. P. K. Hsu, M. A. Yassa, R. T. Knight and J. J. Lin, Amygdala-Hippocampal Dynamics during Salient Information Processing, *Nat. Commun.*, 2017, **8**(1), 14413, DOI: **10.1038/ncomms14413**.

63 H. Kamberaj and A. van der Vaart, Correlated Motions and Interactions at the Onset of the DNA-Induced Partial Unfolding of Ets-1, *Biophys. J.*, 2009, **96**(4), 1307–1317, DOI: **10.1016/j.bpj.2008.11.019**.

64 H. Kamberaj and A. van der Vaart, Extracting the Causality of Correlated Motions from Molecular Dynamics Simulations, *Biophys. J.*, 2009, **97**(6), 1747–1755, DOI: **10.1016/j.bpj.2009.07.019**.

65 M. Sobieraj and P. Setny, Granger Causality Analysis of Chignolin Folding, *J. Chem. Theory Comput.*, 2022, **18**(3), 1936–1944, DOI: **10.1021/acs.jctc.1c00945**.

66 M. J. Buskes and M.-J. Blanco, Impact of Cross-Coupling Reactions in Drug Discovery and Development, *Molecules*, 2020, **25**(15), 3493, DOI: **10.3390/molecules25153493**.

67 M. Gay, P. Carato, M. Coevoet, N. Renault, P.-E. Larchanché, A. Barczyk, S. Yous, L. Buée, N. Sergeant and P. Melnyk, New Phenylaniline Derivatives as Modulators of Amyloid Protein Precursor Metabolism, *Bioorg. Med. Chem.*, 2018, **26**(8), 2151–2164, DOI: 10.1016/j.bmc.2018.03.016.

68 R. Jiang, L. Li, T. Sheng, G. Hu, Y. Chen and L. Wang, 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, DOI: 10.1021/jacs.8b07294.

69 S. Hayakawa, A. Kawasaki, Y. Hong, D. Uraguchi, T. Ooi, D. Kim, T. Akutagawa, N. Fukui and H. Shinokubo, Inserting Nitrogen: An Effective Concept To Create Nonplanar and Stimuli-Responsive Perylene Bisimide Analogues, *J. Am. Chem. Soc.*, 2019, **141**(50), 19807–19816, DOI: 10.1021/jacs.9b09556.

70 S. Izumi, H. F. Higginbotham, A. Nyga, P. Stachelek, N. Tohnai, P. D. Silva, P. Data, Y. Takeda and S. Minakata, 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, DOI: 10.1021/jacs.9b11578.

71 F. Picini, S. Schneider, O. Gavat, A. Vargas Jentzsch, J. Tan, M. Maaloum, J.-M. Strub, S. Tokunaga, J.-M. Lehn, E. Moulin and N. Giuseppone, Supramolecular Polymerization of Triarylamine-Based Macrocycles into Electroactive Nanotubes, *J. Am. Chem. Soc.*, 2021, **143**(17), 6498–6504, DOI: 10.1021/jacs.1c00623.

72 A. Berlanga-Vázquez, R. A. Talmazan, C. A. Reyes-Mata, E. G. Percástegui, M. Flores-Alamo, M. Podewitz and I. Castillo, Conformational Effects of Regioisomeric Substitution on the Catalytic Activity of Copper/Calix[8] Arene C–S Coupling, *Eur. J. Inorg. Chem.*, 2023, **26**(6), e202200596, DOI: 10.1002/ejic.202200596.

73 C. Bannwarth, S. Ehlert and S. Grimme, 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, DOI: 10.1021/acs.jctc.8b01176.

74 C. W. J. Granger, Investigating Causal Relations by Econometric Models and Cross-Spectral Methods, *Econometrica*, 1969, **37**(3), 424–438, DOI: 10.2307/1912791.

75 H. Y. Toda and T. Yamamoto, Statistical Inference in Vector Autoregressions with Possibly Integrated Processes, *J. Econom.*, 1995, **66**(1), 225–250, DOI: 10.1016/0304-4076(94)01616-8.

76 F. Sittel, A. Jain and G. Stock, Principal Component Analysis of Molecular Dynamics: On the Use of Cartesian *vs.* Internal Coordinates, *J. Chem. Phys.*, 2014, **141**(1), 014111, DOI: 10.1063/1.4885338.

77 S. Mika, G. Ratsch, J. Weston, B. Scholkopf and K. R. Mullers, 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, DOI: 10.1109/NNSP.1999.788121.

78 J. Contreras-Garcia, E. R. Johnson, S. Keinan, R. Chaudret, J. P. Piquemal, D. N. Beratan and W. T. Yang, NCIPLOT: A Program for Plotting Noncovalent Interaction Regions, *J. Chem. Theory Comput.*, 2011, **7**(3), 625–632, DOI: 10.1021/ct100641a.

79 J. Runge, P. Nowack, M. Kretschmer, S. Flaxman and D. Sejdinovic, Detecting and Quantifying Causal Associations in Large Nonlinear Time Series Datasets, *Sci. Adv.*, 2019, **5**(11), 4996, DOI: 10.1126/sciadv.aau4996.

80 B. M. Rode, T. S. Hofer, B. R. Randolf, C. F. Schwenk, D. Xenides and V. Vchirawongkwin, 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, DOI: 10.1007/s00214-005-0049-1.

81 P. Amara and M. J. Field, Evaluation of an *Ab Initio* Quantum Mechanical/Molecular Mechanical Hybrid-Potential Link-Atom Method, *Theor. Chem. Acc.*, 2003, **109**(1), 43–52, DOI: 10.1007/s00214-002-0413-3.

82 U. C. Singh and P. A. Kollman, 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, *J. Comput. Chem.*, 1986, **7**(6), 718–730, DOI: 10.1002/jcc.540070604.

83 D. A. Case, I. Y. Ben-Shalom, S. R. Brozell, D. S. Cerutti, T. E. Cheatham III, V. W. D. Cruzeiro, T. A. Darden, R. E. Duke, D. Ghoreishi and M. K. Gilson, *Amber20*, University of California, San Francisco, 2020.

84 J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, Development and Testing of a General Amber Force Field, *J. Comput. Chem.*, 2004, **25**(9), 1157–1174, DOI: 10.1002/jcc.20035.

85 C. Adamo and V. Barone, Toward Reliable Density Functional Methods without Adjustable Parameters: The PBE0 Model, *J. Chem. Phys.*, 1999, **110**(13), 6158–6170, DOI: 10.1063/1.478522.

86 F. Weigend and R. Ahlrichs, 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, DOI: 10.1039/B508541A.

87 S. Grimme, J. Antony, S. Ehrlich and H. Krieg, 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, DOI: 10.1063/1.3382344.

88 V. Barone and M. Cossi, Quantum Calculation of Molecular Energies and Energy Gradients in Solution by a Conductor Solvent Model, *J. Phys. Chem. A*, 1998, **102**(11), 1995–2001, DOI: 10.1021/jp9716997.

89 M. Bursch, A. Hansen, P. Pracht, J. T. Kohn and S. Grimme, Theoretical Study on Conformational Energies of

Transition Metal Complexes, *Phys. Chem. Chem. Phys.*, 2021, **23**(1), 287–299, DOI: 10.1039/D0CP04696E.

90 T. Husch, A. C. Vaucher and M. Reiher, Semiempirical Molecular Orbital Models Based on the Neglect of Diatomic Differential Overlap Approximation, *Int. J. Quantum Chem.*, 2018, **118**(24), e25799, DOI: 10.1002/qua.25799.

91 M. Podewitz, Towards Predictive Computational Catalysis – a Case Study of Olefin Metathesis with Mo Imido Alkylidene N-Heterocyclic Carbene Catalysts, in. *Chemical Modelling: Volume 17*, ed. Bahmann, H. and Tremblay, J. C., The Royal Society of Chemistry, 2022, vol. 17, DOI: 10.1039/9781839169342-00001.

92 S. G. Balasubramani, G. P. Chen, S. Coriani, M. Diedenhofen, M. S. Frank, Y. J. Franzke, F. Furche, R. Grotjahn, M. E. Harding, C. Hättig, A. Hellweg, B. Helmich-Paris, C. Holzer, U. Huniar, M. Kaupp, A. Marefat Khah, S. Karbalaei Khani, T. Müller, F. Mack, B. D. Nguyen, S. M. Parker, E. Perlt, D. Rappoport, K. Reiter, S. Roy, M. Rückert, G. Schmitz, M. Sierka, E. Tapavicza, D. P. Tew, C. van Wüllen, V. K. Voora, F. Weigend, A. Wodyński and J. M. Yu, TURBOMOLE: Modular Program Suite for *Ab Initio* Quantum-Chemical and Condensed-Matter Simulations, *J. Chem. Phys.*, 2020, **152**(18), 184107, DOI: 10.1063/5.0004635.

93 A. A. Maryott and E. R. Smith, *Table of Dielectric Constants of Pure Liquids*, National Bureau of Standards, Washington, D.C., 1951, vol. 514, https://nvlpubs.nist.gov/nistpubs/Legacy/circ/nbscircular514.pdf.

94 R. J. Gowers, M. Linke, J. Barnoud, T. J. E. Reddy, M. N. Melo, S. L. Seyler, J. Domański, D. L. Dotson, S. Buchoux, I. M. Kenney and O. Beckstein, MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations, *Proceedings of the 15th Python in Science Conference*, 2016, pp. 98–105, DOI: 10.25080/Majora-629e541a-00e.

95 S. Schultze and H. Grubmüller, Time-Lagged Independent Component Analysis of Random Walks and Protein Dynamics, *J. Chem. Theory Comput.*, 2021, **17**(9), 5766–5776, DOI: 10.1021/acs.jctc.1c00273.

96 H. Wu, F. Nüske, F. Paul, S. Klus, P. Koltai and F. Noé, Variational Koopman Models: Slow Collective Variables and Molecular Kinetics from Short off-Equilibrium Simulations, *J. Chem. Phys.*, 2017, **146**(15), 154104, DOI: 10.1063/1.4979344.

97 I. Guyon, J. Weston, S. Barnhill and V. Vapnik, Gene Selection for Cancer Classification Using Support Vector Machines, *Mach. Learn.*, 2002, **46**(1), 389–422, DOI: 10.1023/A:1012487302797.

98 L. Breiman, Random Forests, *Mach. Learn.*, 2001, **45**(1), 5–32, DOI: 10.1023/A:1010933404324.

99 D. R. Cox, The Regression Analysis of Binary Sequences, *J. Roy. Stat. Soc. B*, 1958, **20**(2), 215–232, DOI: 10.1111/j.2517-6161.1958.tb00292.x.

100 R. Tibshirani, Regression Shrinkage and Selection *Via* the Lasso, *J. Roy. Stat. Soc. B*, 1996, **58**(1), 267–288, DOI: 10.1111/j.2517-6161.1996.tb02080.x.

101 X. Zeng and T. R. Martinez, Distribution-Balanced Stratified Cross-Validation for Accuracy Estimation, *J. Exp. Theor. Artif. Intell.*, 2000, **12**(1), 1–12, DOI: 10.1080/095281300146272.

102 S. Chormunge and S. Jena, Correlation Based Feature Selection with Clustering for High Dimensional Data, *J. Electr. Syst. Inf. Technol.*, 2018, **5**(3), 542–549, DOI: 10.1016/j.jesit.2017.06.004.

103 *Scikit-Learn-Contrib/Skope-Rules*, 2024, https://github.com/scikit-learn-contrib/skope-rules, accessed 2024-04-15.

104 D. A. Dickey and W. A. Fuller, Distribution of the Estimators for Autoregressive Time Series With a Unit Root, *J. Am. Stat. Assoc.*, 1979, **74**(366), 427–431, DOI: 10.2307/2286348.

105 D. Kwiatkowski, P. C. B. Phillips, P. Schmidt and Y. Shin, 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?, *J. Econom.*, 1992, **54**(1), 159–178, DOI: 10.1016/0304-4076(92)90104-Y.

106 R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 2018.

107 H. Lütkepohl, *Introduction to Multiple Time Series Analysis*, Springer, Berlin, Heidelberg, 1991, DOI: 10.1007/978-3-662-02691-5.

108 C. A. Sims, Macroeconomics and Reality, *Econometrica*, 1980, **48**(1), 1–48, DOI: 10.2307/1912017.

109 H. Akaike, A New Look at the Statistical Model Identification, *IEEE Trans. Autom. Control*, 1974, **19**(6), 716–723, DOI: 10.1109/TAC.1974.1100705.

110 S. Johansen, Statistical Analysis of Cointegration Vectors, *J. Econ. Dynam. Control*, 1988, **12**(2), 231–254, DOI: 10.1016/0165-1889(88)90041-3.

111 Y. Benjamini and Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, *J. Roy. Stat. Soc. B*, 1995, **57**(1), 289–300.