

Chemography-guided analysis of a reaction path network for ethylene hydrogenation with a model Wilkinson's catalyst

Philippe Gantzer¹, Ruben Staub¹, Yu Harabuchi¹, Satoshi Maeda¹ and Alexandre Varnek*^{1,2}

¹ Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University, Kita 21, Nishi 10, Kita-ku, Sapporo, Hokkaido 001-0021, Japan

² Laboratory of Chemoinformatics, UMR 7140, CNRS, University of Strasbourg, 67081 Strasbourg, France

* varnek@unistra.fr

Abstract.

Visualization and analysis of large chemical reaction networks become rather challenging when conventional graph-based approaches are used. As an alternative, we propose to use the chemical cartography (“chemography”) approach, describing the data distribution on a 2-dimensional map. Here, the Generative Topographic Mapping (GTM) algorithm – an advanced chemography approach – has been applied to visualize the reaction path network of a simplified Wilkinson's catalyst-catalyzed hydrogenation containing some 10⁵ structures generated with the help of the Artificial Force Induced Reaction (AFIR) method using either Density Functional Theory or Neural Network Potential (NNP) for potential energy surface calculations. Using new atoms permutation invariant 3D descriptors for structure encoding, we've demonstrated that GTM possesses the abilities to cluster structures that share the same 2D representation, to visualize potential energy surface, to provide an insight on the reaction path exploration as a function of time and to compare reaction path networks obtained with different methods of energy assessment.

Keywords: Generative Topographic Mapping, Artificial Force Induced Reaction, Neural Network Potential

Introduction

Ab initio kinetics studies play a crucial role in deepening our understanding of reaction mechanisms.^[1-4] The most reliable but still computationally costly approach considers a systematical exploration of reaction pathways between equilibrium states, resulting in the creation of a reaction path network. Within this network, individual nodes correspond to equilibrium states (EQ), while their connecting edges represent elementary chemical

transformations.^[5-7] Each EQ represents a local minimum on the Potential Energy Surface (PES) whereas each Transition State (TS) represents a saddle point along a reaction path between two EQs. As such, each edge of the reaction path network is associated with a TS (Figure 1).

The Artificial Force Induced Reaction (AFIR) algorithm applies artificial molecular forces to overcome energy barriers, providing a powerful tool to systematically explore reaction pathways.^[8-10] Usually, the compilation of all reaction paths generated by AFIR leads to particularly large reaction path networks. For example, a recent AFIR-based reaction path search on a 20-atoms system led to the generation of 1.2×10^4 equilibrium states (EQ) and 4.5×10^5 distinct geometries.^[11] As more complex reactions are being considered, the size of networks is anticipated to grow even larger. Usually, to obtain accurate results, such a PES exploration is ideally performed using Density Functional Theory (DFT) to assess the structures' energy and forces. This represents the main bottleneck of the approach, as nearly all the computation time is dedicated to DFT calculations. In order to accelerate these calculations and to treat large-size systems, Neural Network Potentials (NNPs) has recently been employed as a rapid and cost-effective alternative to the traditionally expensive DFT calculations, see Figure 1. Recently, we reported an AFIR-based path search for the hydrogenation of ethylene catalyzed by a model Wilkinson's catalyst, using DFT, xTB and NNP methods, and discussed various aspects of PES machine-learning.^[12] In particular, we have demonstrated that a pure NNP architecture should be coupled to a physics-based potential, such as a semi-empirical xTB method, in order to greatly reduce the generation of broken 3D structures.

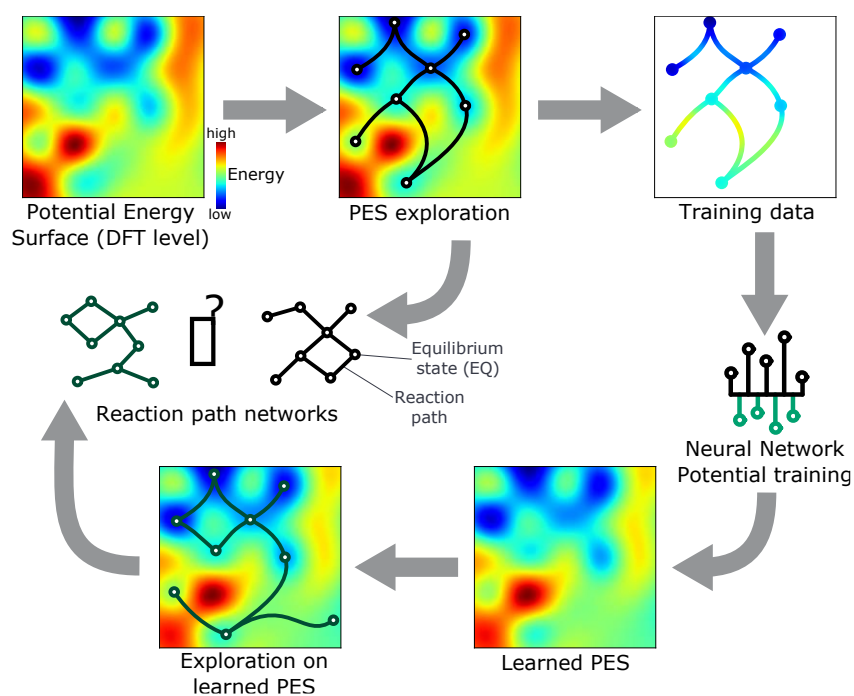


Figure 1. Reaction path network generation workflow: path geometries discovered in the AFIR-based reaction path search at a DFT level are used to train a Neural Network Potential (NNP). An AFIR-based search performed with the trained NNP, leads to an alternative reaction path network. Comparison of the DFT and NNP-based results allows to elucidate the impact of the underlying potential energy surfaces on the reaction path exploration.

AFIR-based kinetics studies automatically explore many reaction paths, and therefore, typically yield large reaction path networks. Traditionally, a reaction path network is visualized by a connected graph in which each node represents a 2D or 3D molecular structure and each edge connecting the nodes corresponds to a chemical transformation. Each 3D structure can be characterized by its atomic coordinates, chemical bonding topology and some properties (potential energy, charge on selected atom, etc.). However, a graph-based network containing several hundred or thousand nodes hardly allows to analyze such Structure-Property Relationships (SPR). Moreover, direct comparisons of reaction path networks calculated with the help of different theoretical methods are challenging.

The disconnectivity graph is one of conventional approach to visualize a complex energy landscape represented by many elementary steps, and has been utilized in cluster and biomolecular systems.^[13–15] As an alternative method of reaction data visualization and analysis, we have suggested to use the Generative Topographic Mapping (GTM), describing the data distribution on a 2-dimensional map.^[12] For the reaction path network of hydrogenation using Wilkinson’s catalyst, it has been shown that GTM possesses the abilities to cluster structures that share the same 2D representation, to visualize potential energy surface

and to provide an insight on the reaction path exploration as a function of time. However, many aspects of GTM application were out of the scope of that previous study.

In this paper, we demonstrate how GTM can be used for detailed SPR analysis and investigate its capability to compare reaction path networks obtained from different potentials. We also introduce new Distance Distribution Descriptors which have some advantage over previously used Pairwise Sorted Distances-Based descriptors to encode 3D structures.

METHOD.

Data

We used four datasets generated in our previous study using DFT and NNP(+xTB) for potential energy assessment.^[12] The DFT dataset contains 118240 geometries, including 6298 approximate TSs and 2049 EQs, and their associated potential energy and gradients, computed at the R ω B97X-D/Def2-SVP level of theory, see Table 1. These geometries correspond to the reaction paths for the 6298 elementary processes explored with the AFIR method. These AFIR explorations were performed using the kinetics-based navigation,^[16] which controls the automated pathway search based on an index called traffic volume. The traffic volume represents the amount of reaction flux through each EQ, and thus preferentially finds kinetically important EQs and elementary steps.

The conventional alkenes hydrogenation by H₂, catalyzed by the original Wilkinson's catalyst (i.e., RhCl(PPh₃)₃), involves several steps subsequent to the initial PPh₃ elimination: oxidative addition of H₂ to the metal complex; alkene coordination; alkene insertion; and reductive elimination of alkane.^[17] Earlier^[12], we have examined the reaction using the simplified catalyst RhCl(PH₃)₃. We found that at the first step, the ethylene coordinates to the catalyst producing RhCl(PH₃)₃(C₂H₄); followed by the elimination of a PH₃ ligand. Then the oxidative addition of H₂, ethylene insertion and reductive elimination of ethane proceed with two PH₃ ligands. Finally, the initial RhCl(PH₃)₃ catalyst is regenerated, completing the catalytic cycle. Figure 2 (bottom) shows the kinetically important 2D structures of the reaction path network at a DFT level. In this Figure, the leftmost 2D motif represents the reactants (**R**), whereas the product (**P**) - the 2D motif with the highest yield at 300 K - is represented in the bottom right-hand corner. Only the lowest reaction barriers between 2D motifs are shown, assuming that the reaction barriers between conformers are sufficiently low.

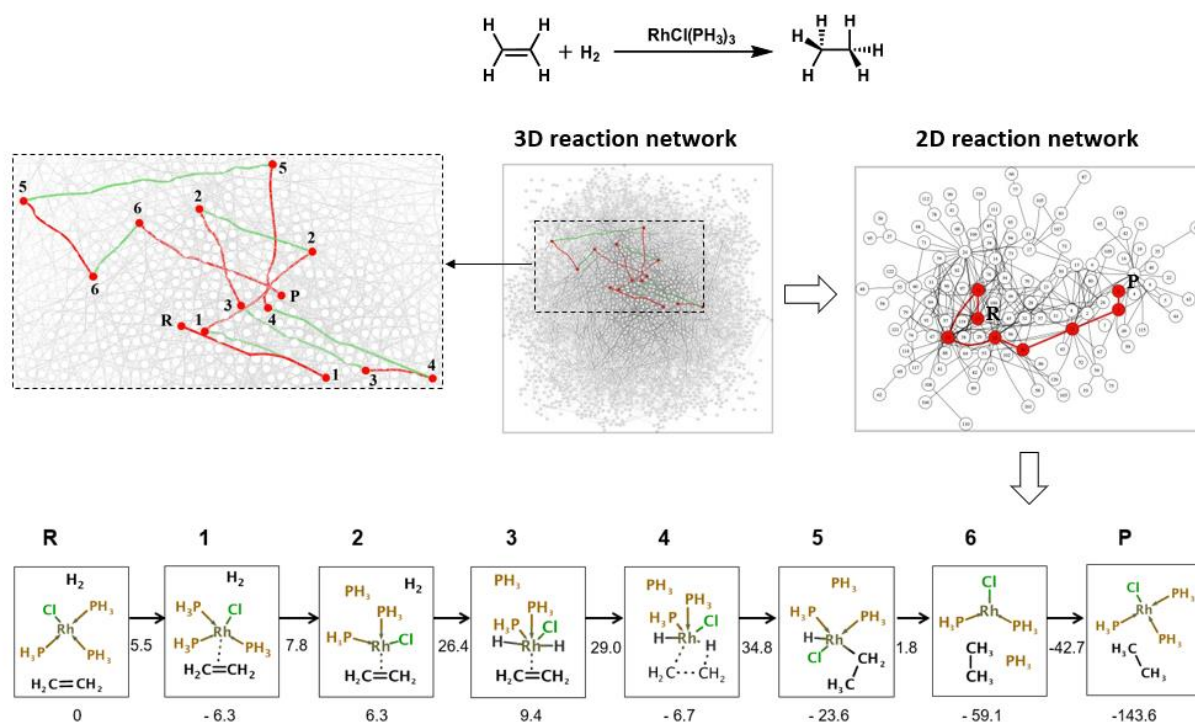


Figure 2. (top) Reaction scheme of hydrogenation using a simplified Wilkinson's catalyst $\text{RhCl}(\text{PH}_3)_3$. (middle) Reaction path networks represented by connected graphs; each node corresponds to either a 3D equilibrium structure (3D network) or to its related 2D structural motif (2D network). The sequence of reaction path from reactants (**R**) to product (**P**) is shown in red. The green lines in the 3D network correspond to conformational transitions. (bottom) A part of the 2D network showing a sequence of reaction paths from **R** to **P**. The numbers above the boxes correspond to the reaction steps **1-6**, whereas those below the boxes and edges correspond, respectively, to the relative free energies of equilibrium structures and transition states in kJ/mol, compared to the reactant structure, at 300K.

In order to accelerate the network expansion, a Neural Network Potential (NNP) associated with the semi-empirical xTB method was suggested to approximate the DFT potential energy surface at a much lower cost.^[12] Three NNP(+xTB) models (simply called NNP-xx in the following) were trained on the first 20%, 50% and 80% of the DFT generated dataset, respectively (Table 1).

Table 1: The number of 3D structures including equilibrium (EQ) and transition (TS) structures, and of related 2D structures, obtained during an AFIR-based reaction path search, depending on the level of theory used. NNP(+xTB) models were trained on the first 20%, 50% and 80% data of the DFT network.

Level of theory used for search	# structures	# EQ	# 2D motifs	# TS
DFT	118 240	2 049	122	6 298

NNP-20	103 679	1 557	190	4 837
NNP-50	106 714	1 691	184	5 135
NNP-80	107 651	1 745	165	5 268

Generative Topographic Mapping.

GTM is a method of non-linear mapping of data points from a multi-dimensional chemical space to two-dimensional space.^[18] The probabilistic topology-preserving characteristic of GTM has made it a popular tool for data analysis and chemical visualization. The algorithm inserts a two-dimensional “rubber sheet”-like manifold into the initial descriptor space in order to reproduce the best data by a simulated probability distribution function. The latter is represented by an ensemble of Gaussian functions located at the nodes of two-dimensional grid related to the manifold. Distortion of the manifold is controlled within the limits of a predefined set of parameters. Finally, the molecules are projected with a given probability onto each node of the manifold, which then is then projected onto a two-dimensional latent space in which a molecular structure is associated with one or more nodes.

For each molecular structure M mapped onto GTM, a probability matrix $R(M, K)$ is calculated which gives the probability of M residing in node K , i.e., the responsibility of node K related to structure M . Generally, the responsibilities related to a molecular structure may be distributed across several nodes. The overall probability to see a structure anywhere on the map, i.e., $\sum_K R(M, K)$ is always equal to 1.0. The set of structures S residing in a node K are represented by cumulated responsibilities of K towards all of its members, $\rho(S, K) = \sum_{M \in S} R(M, K)$. It represents the density distribution or fuzzy membership of structures in a set in a particular node of GTM. $\rho(S, K)$ defines the node-bound density distribution of the compound set S .

In order to visualize a given property distribution on GTM, the weighted average of properties of all structures associated with any particular node is used to “color” the manifold, resulting into a fuzzy property landscape, where the projected property can be energy, error of energy prediction or relative population of the structures of the given type. The responsibilities related to molecular structures are used as weights. A property landscape can be used to build classification and regression models.^[19]

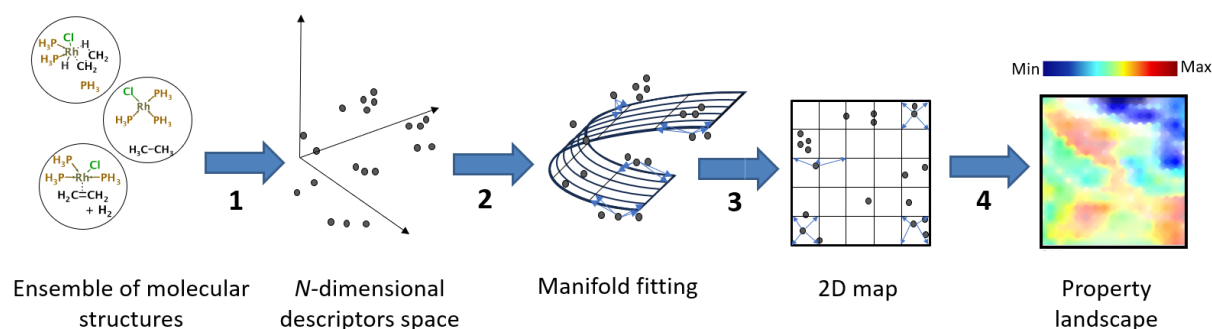


Figure 3: GTM preparation workflow. Some studied molecular systems are encoded by 3D descriptors and represented as objects in N -dimensional space. A flexible manifold fits the data following by fuzzy projection of the datapoints onto the manifold. For each datapoint, the mapping algorithm calculates the probabilities (responsibilities) to its projection into the nodes of rectangular grid superposed with the manifold. Unbending the manifold results in a 2D map. Location of a datapoint on the map is calculated as a gravity center of the responsibilities' distribution.

In this work, the manifold was developed on the frame set of 10^4 randomly selected structures. Either Pairwise Sorted Distances-Based (PSDB) descriptors or Distance Distribution descriptors (D^3) were used to encode the 3D structures. An evolutionary algorithm-based approach was implemented in order to optimize the parameters required for GTM setup: number of nodes, number of radial basis functions (RBF) defining the manifold and their width, the regularization coefficient. The optimal set of GTM parameters defined a Pareto front of locally best solutions corresponding to (i) minimal energy prediction error and (ii) maximal informational entropy.

To compute energy prediction errors, a 3-folds cross validation procedure was employed. Sequentially, two-third of all the structures was used to build the energy landscape whereas the remaining one-third was then used as test set for which the predicted energy values were compared to related DFT values. The process was repeated three times, so that each molecular structure gets a predicted energy value.

Descriptors

Two types of descriptors were tested to encode 3D molecular structures - Pairwise Sorted Distances-Based descriptors and Distance Distribution Descriptors. Both types of descriptors are alignment-free and invariant by rotation, translation, and permutation of atoms with the same atomic number.

Pairwise sorted distances-based descriptors

In the Pairwise-Sorted Distance-Based Descriptors (PSDB), the retrieved interatomic distances are first grouped according to their corresponding atomic element pair and then sorted within each group in ascending order.^[20,21] The sorted distances of each group are then concatenated to form the descriptor vector of each 3D structure (Figure 4A). The molecular system RhClP₃C₂H₁₅ considered in the simplified Wilkinson's catalyst-catalyzed hydrogenation reaction contains 22 atoms. Thus, the number of interatomic distances and, hence, the length of the PSDB descriptor vector is 231.

Distance distribution descriptors

Distance Distribution Descriptors (D³) are derived from atoms encoding used by SchNet.^[22] Like most end-to-end neural networks, SchNet uses a highly sophisticated representation of atoms, accounting their local environment through pairwise distances.

For each possible atom pair (a_i, a_j), the distance d_{ij} is embedded in a M -dimensional space using Gaussian smearing. The k -th component of the resulting vector V_{ij} is then:

$$V_{ij}[k] = \exp\left(-\frac{(d_{ij}-\mu_k)^2}{\sigma^2}\right), \quad (1)$$

Where $\sigma = \frac{cutoff}{M}$ is the bin interval and μ_k is the k -th Gaussian center, with the μ centers being equidistributed along the $[0 \text{ \AA}, cutoff]$ segment. Embeddings of individual distances belonging to the same atomic pair type (z, z') are then aggregated according to eq. 2 (see Figure 4B). The resulting vector is invariant by permutation of equivalent atoms.

$$V_{zz'} = \sum_{i,j|\{z_i,z_j\}=\{z,z'\}} V_{ij}/2 \quad (2)$$

This aggregation allows to reduce the dimensionality of the input in a chemically meaningful manner compared to the concatenation of vectors V_{ij} . Compared to the PSDB descriptors, the D³ descriptors are (i) tunable because of the customizability of both the number of Gaussian functions (bins) and the cutoff distance in eq. 1 and (ii) have the same size regardless of the number of atoms for each atom type. In the case of the simplified Wilkinson's catalyst-catalyzed hydrogenation reaction, the descriptor vector consists of 25 blocks, each corresponding to a pair of elements (H-H, H-C, H-Cl, H-P, H-Rh, C-C, C-P, ...). Thus, as an example, for 30 bins per interatomic distance, the resulting D³ descriptor is of 750 (= 25 * 30) components length.

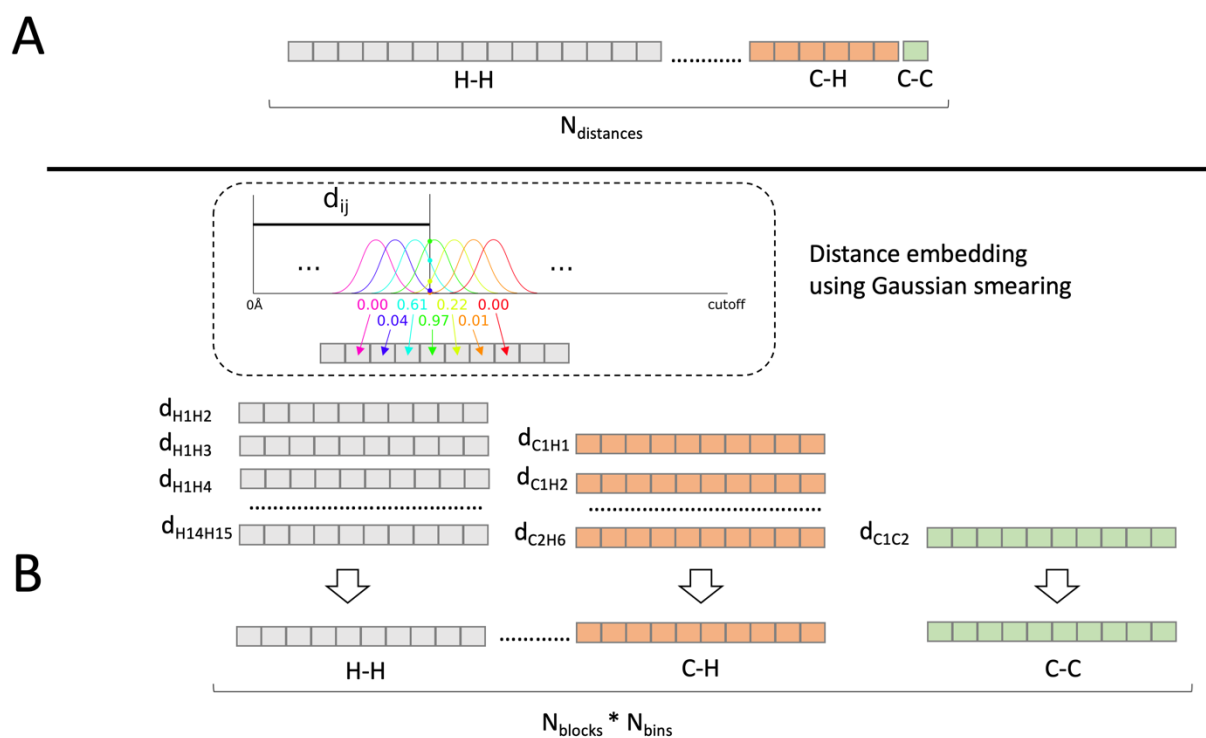


Figure 4: Construction of a descriptor vector for the studied molecular system $\text{RhClP}_3\text{C}_2\text{H}_{15}$. (A) Pairwise Sorted Distances-Based descriptor resulted from concatenation of groups containing sorted interatomic distances of given pairs of atoms. (B) Distance Distribution Descriptor results in concatenation of blocks, each corresponding to the summed embeddings of distances for particular atom types.

Projection of representative 2D structures

If several 3D structures (e.g., conformers of a given complex) share the same 2D motif, the latter can be associated with a single “representative” 3D structure, for which descriptors $\{X_p\}$ are calculated as the Boltzmann-weighted sum of descriptors for real 3D structures in the pool

$$X_p = \sum w_i X_{i,p}, \quad w_i = \frac{e^{(E_i/(k_B * T))}}{\sum e^{(E_j/(k_B * T))}} \quad (3)$$

where X_p and $X_{i,p}$ are the p -th term of the descriptor characterizing, respectively, the entire subset and the i -th structure in the pool, E_i is the relative potential energy of the i -th structure with respect to the lowest energy observed within the reaction path network. The projection of such representative structure on the map can be associated with the “representative position” of the related 2D motif.

RESULTS AND DISCUSSION

DFT-based reaction path network

During the GTM manifold training, the D^3 parameters were systematically varied (*cutoff* from 2 to 10 Å and number of bins k from 5 to 30). The parameters values *cutoff*=3 Å and k =30 found in the grid search were retained for further calculations. For each set of D^3 parameters, GTM parameters were optimized by a dedicated Genetic Algorithm^[23] and the one corresponding to minimal energy prediction error and maximum entropy at the Pareto front were selected.

The resulting density and energy landscapes obtained with both PSDB and D^3 descriptors look similar (Figure 5). The PSDB-based energy landscape provides with the slightly better accuracy in cross-validation in energy predictions compared to the D^3 -based landscape: RMSE = 44.2 kJ/mol (PSDB) and 48.9 kJ/mol (D^3). On the other hand, the density landscapes show that the data are distributed more homogeneously on the D^3 -based map, therefore, only landscapes built on D^3 descriptors-based manifold will be further analyzed.

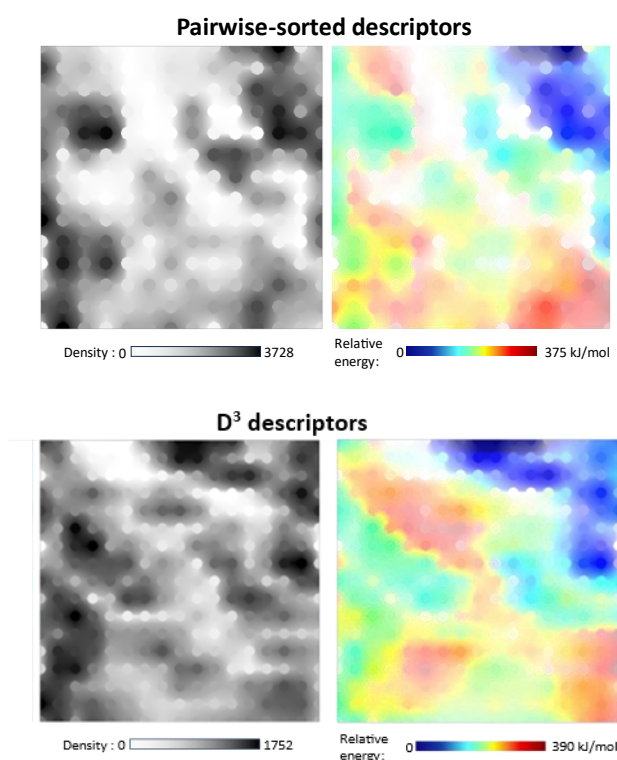


Figure 5: Relative energy (right) and density (left) landscapes for the DFT-based network, using the PSDB and D^3 descriptors.

Positions of the observed 2D motifs on the energy landscape are shown in Figure 6. Some of them overlap, which explains why the number of dots (85) in Figure 6 is smaller than the

number of 2D structural motifs (122) observed in the DFT simulations (Table 1). Examples of overlapping structural motifs are given in Table S1 in Supporting Information.

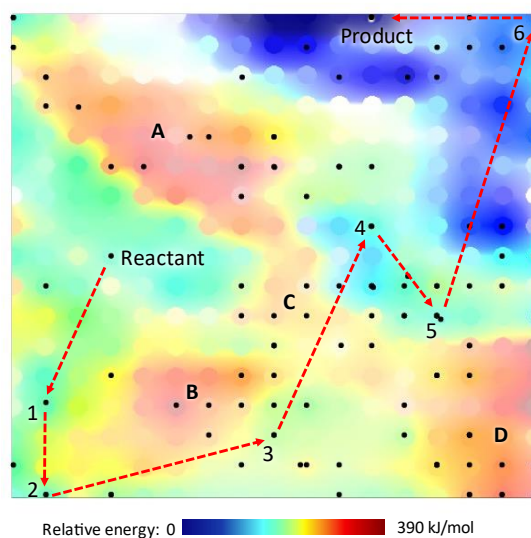
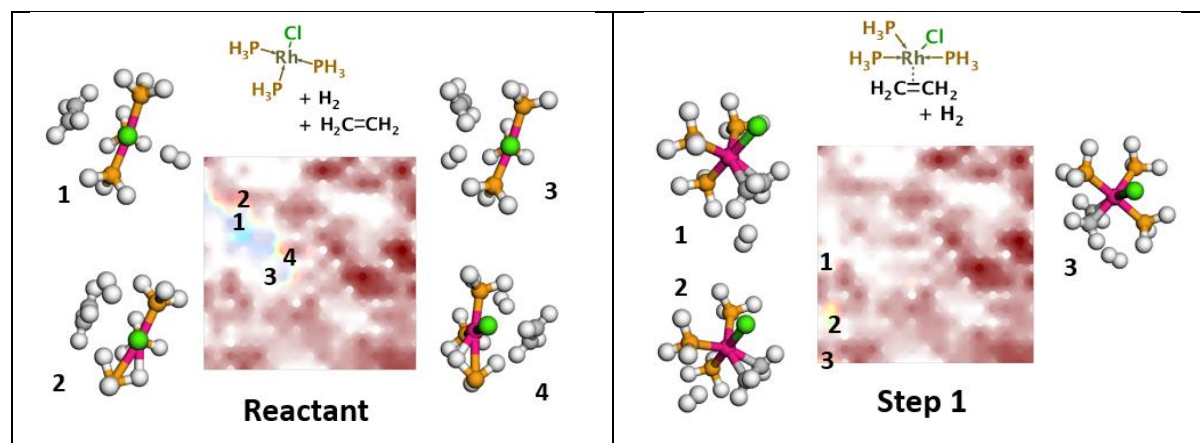


Figure 6. DFT energy landscape obtained with the D^3 descriptors. Each dot corresponds to a 3D structure representing a particular 2D motif (see eq. 3). The sequence of reaction paths (in red) connect the 2D motifs corresponding to reactant (**R**), product (**P**) and reaction steps **1-6**. 2D motifs and related 3D structures populating reaction steps areas and high energy zones **A-D** are shown in Figures 7 and 8, respectively.



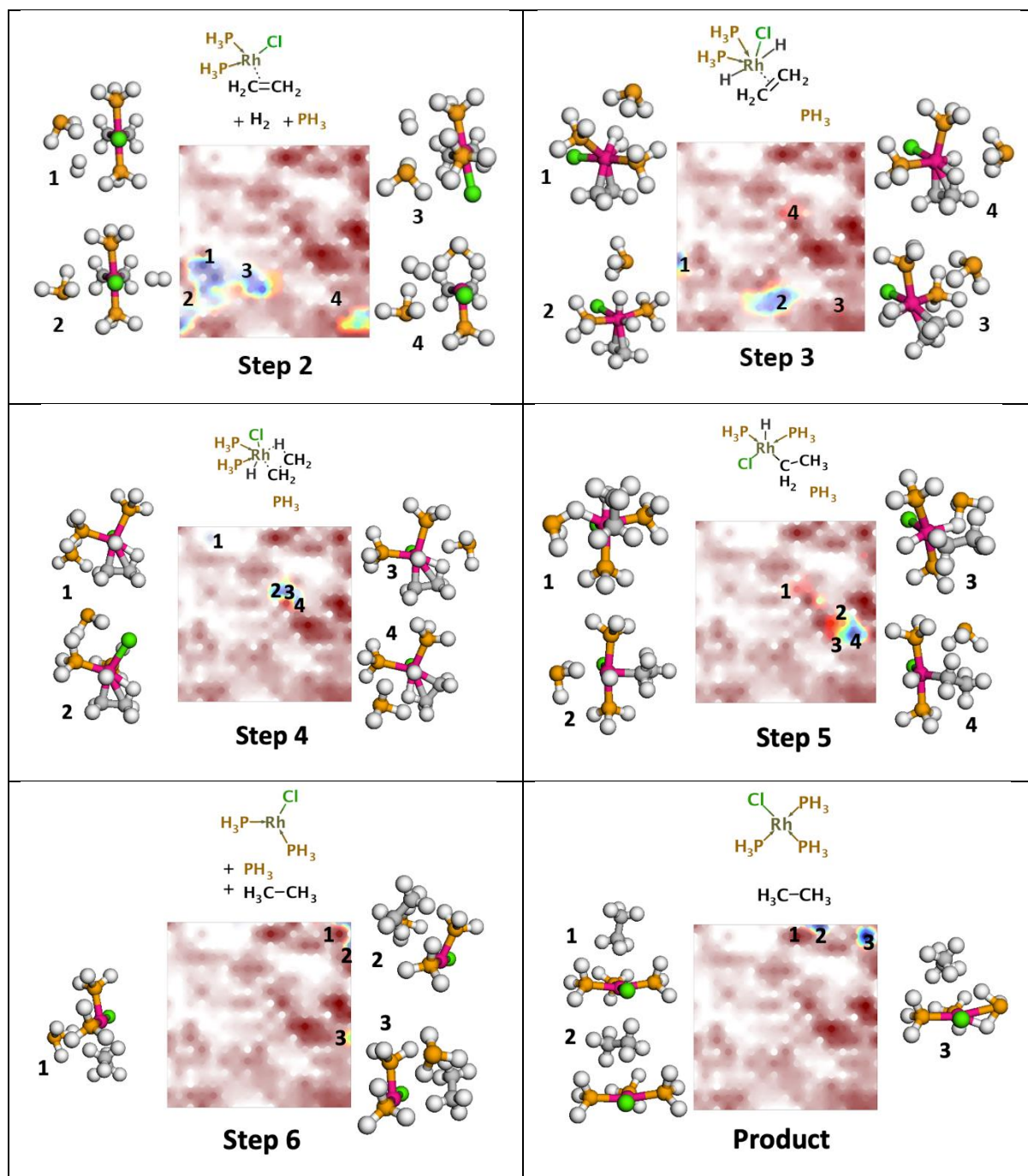


Figure 7. Landscapes of the 2D motifs and examples of representative 3D structures corresponding to different steps of reaction path (see Figure 6). On the maps, the color code shows highly (in blue) and low (dark red) populated areas by the given structures.

Figure 7 demonstrates a set of class landscapes describing relative population of EQ structures corresponding to different steps of the reaction path. In most of cases, these structures form a relatively tight cluster on the map. However, in the landscapes for steps 2, 3, 5, 6 and product several clusters were observed. This can be interpreted by formation of both low and high energy structures with different coordination patterns of PH_3 group or Cl and H atoms with respect to Rh or different position of uncoordinated H_2 or PH_3 species. The 3D structure

representative for a given 2D motif is always located in the low energy area (Figure 6). This is not surprising because the Boltzmann-like weighting of descriptors favors contribution of low-energy species.

The high-energy areas contain chemically unreasonable structures, for instance, those with a deeply unfavorable isolated Cl^- in vacuum (zone A) or H_2 and HCl formation from the catalyst decomposition (zone C), or an hypervalent P with trigonal bipyramidal shape (zone B), or even displaying a C-C dissociation (zone D), see Figure 8.

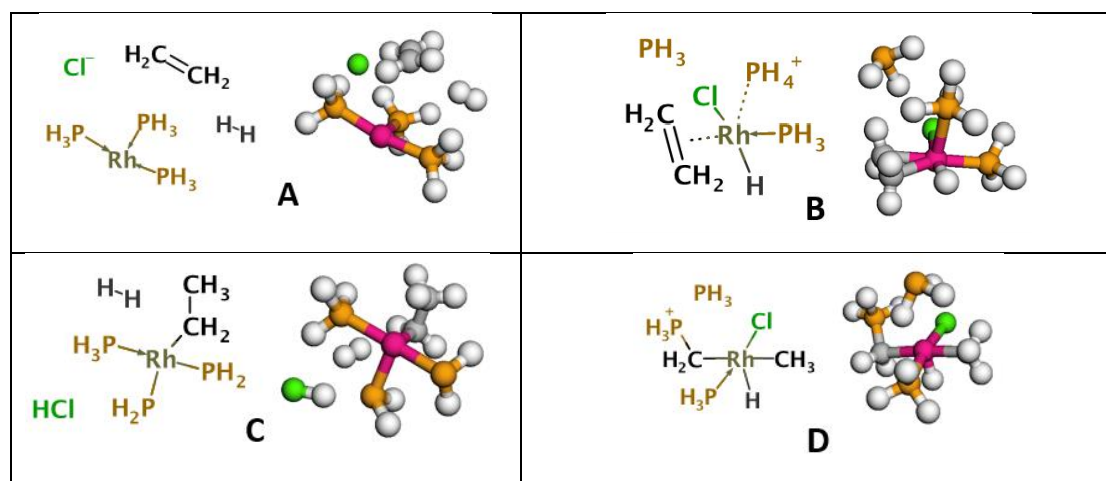
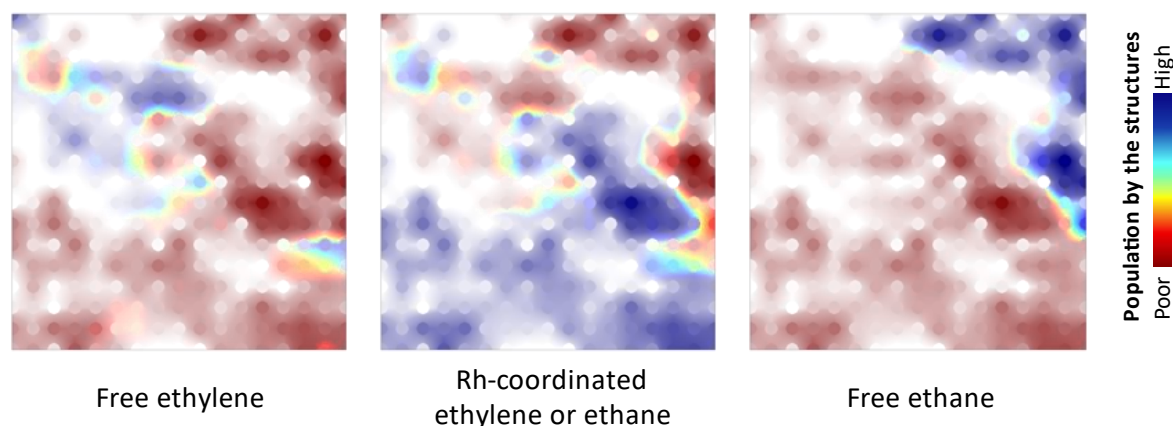


Figure 8. Examples of structural motifs and related 3D structures contained in high energy zones A-D (see Figure 6).

The class landscapes for alkanes/alkenes presence and their coordination states shown on Figure 9 exhibit remarkable discriminatory power, partitioning the GTM into three distinct region featuring uncoordinated ethylene or ethane (the upper left and upper right areas, respectively) and species with coordination of at least one carbon atom to the metal (bottom area). This suggest that GTM inherently produced a chemically meaningful mapping, that can be used for a chemical feature-based characterizations. As an example, here, by projecting and following each reaction step on the class landscapes of Figure 9, it appears clearly that the reaction starts with a free ethylene, then proceeds to its coordination and hydrogenation (while still being coordinated), followed finally by the release of a free ethane. With the appropriate class landscapes, a similar analysis could be performed based on any chemically relevant feature/property (e.g., the oxidation state of Rh or partial atomic charges on selected atoms).



Figures 9. Landscapes of the species including free ethylene (left), coordinated to Rh ethylene or ethane (center) and free ethane (right).

Monitoring the DFT network expansion.

In addition to “static” data distribution, we investigated the chronological expansion of the reaction path network driven by the AFIR algorithm. For this purpose, the manifold constructed for the entire DFT reaction path network was used to project the first $n\%$ ($n=1, 5, 20, 35, 50$ and 100) portions of path structures discovered in the DFT/AFIR run initiated from the reactants structure. Each portion corresponds to a “chronological milestone” of the reaction path network expansion. In such a way, 6 class landscapes showing data projected at the current chronological milestone compared to the previous ones (e.g., 5% compared to 1%) have been built, see Figure 10, left. On the “1%” landscape, the structures predominantly occupy the reactant area, whereas at 10%, the network explores the steps 1-5 zones. The “20%” landscape shows that the network not only encompasses the area related to the structures of step 6 but also reaches the product area. At 35% of the network expansion, the product area becomes dense. The 50% and 100% landscapes show that the search is shifting from exploration to refinement, mainly associated with the discovery of new reaction pathways in the products area.

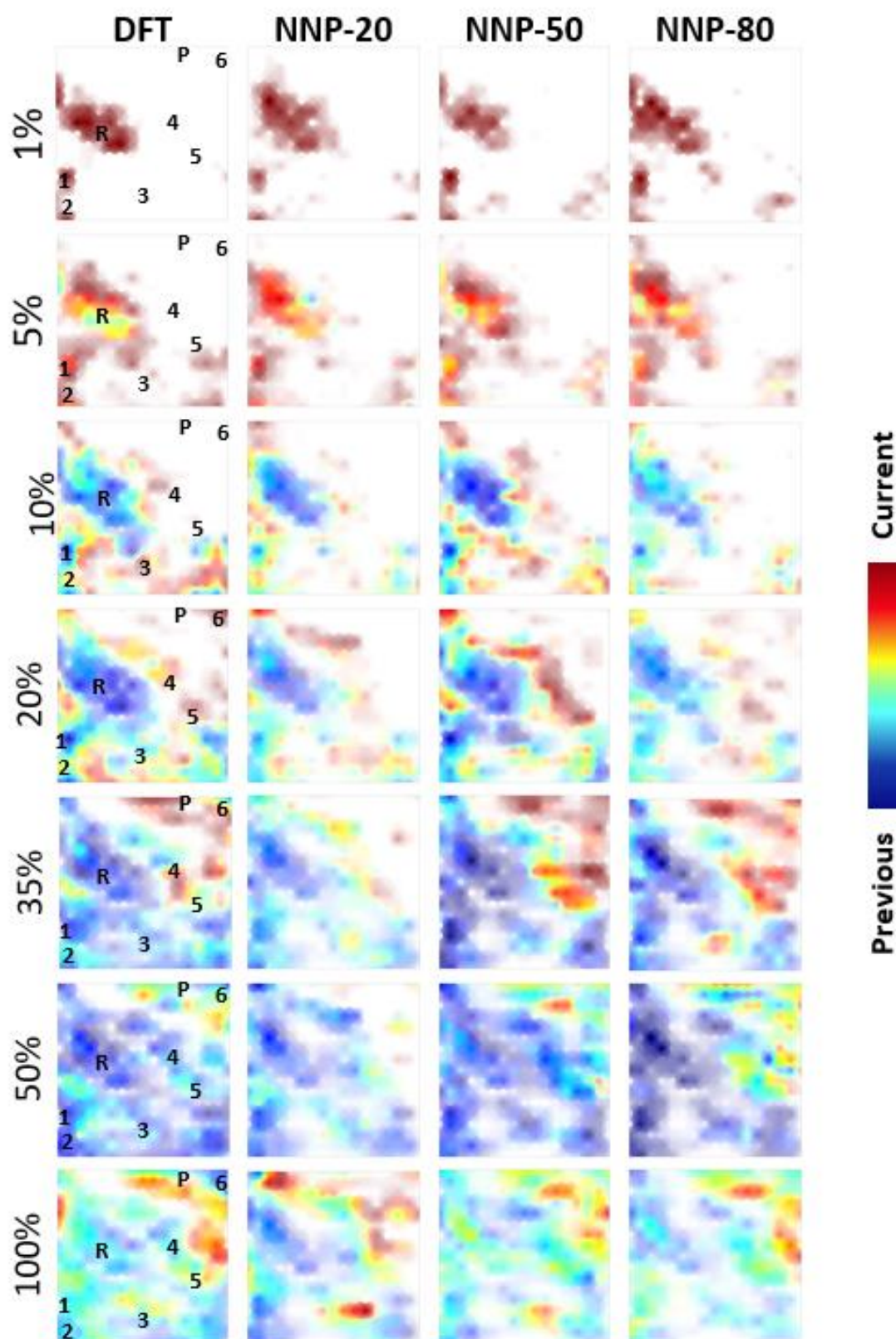


Figure 10: Classes landscapes of the networks explored by DFT and three NNP models at different AFIR exploration stages. The locations of the representative 3D structures corresponding to the 2D structural motifs of reactant (**R**) product (**P**) and reaction steps **1-6** are shown for the DFT network.

NNP-based reaction path network

Similar “chronological” analysis has been performed for three reaction path networks NNP-20, NNP-50 and NNP-80 obtained in NNP/AFIR runs using the model trained, respectively, on the first 20%, 50% and 80% of the DFT network (Figure 10). The “current/previous” class landscapes show that the NNP-20 network exploration significantly differs from the DFT, NNP-50 and NNP-80 ones. Thus, even in the first 50% of the network expansion, the NNP-20 model was not able to reach the zone of products, which can be explained by the fact that the training set contained just a few structures from the step 6 and product areas.

For more detailed analyses of the NNP networks, a series of regression landscapes were prepared, showing the distribution of predicted energies and the absolute error of energy predictions; as well as classification landscapes for overlapped DFT and NNP networks. They show that the NNP-20 trained on 20% DFT set generates many high energy structures mostly populating the areas poorly covered by its training dataset (see Figure 11, top left). This observation is consistent with the generation of unreasonable geometries outside of the training domain (Figure 12). Finally, Figure 11 (bottom left) shows a large discrepancy in the chemical spaces populated by the DFT/AFIR search and the NNP-20/AFIR search. When combining this landscape with the time evolution in Figure 10, we observe that some of the regions, which are very predominantly populated by NNP-20 geometries, correspond to areas that seems to “capture” the exploration at the expense of other regions of interest during the search. We believe these observations to characterize the presence of areas with deep (or even apparently unbounded) energy underestimations in the trained potential. Indeed, due to both the gradient-following nature of AFIR and the exploration focus around apparently kinetically accessible regions, the AFIR method is prone to preferentially sample these unphysical regions. Unsurprisingly, the DFT manifold is poorly adapted to describe NNP-20 data distribution, which is proved by loglikelihood distribution analysis shown in Figure 13.

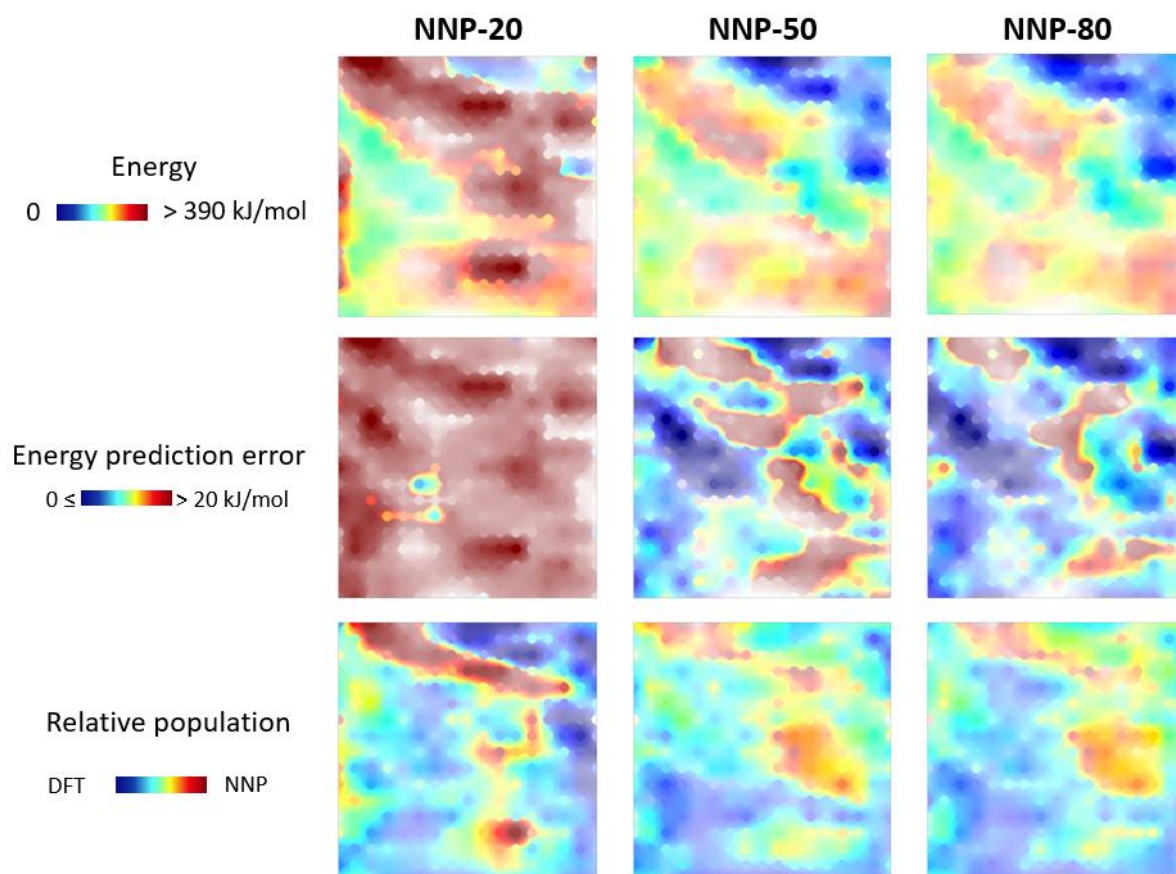


Figure 11. Analysis of three NNP runs trained on the first 20, 50 and 80% portions of the DFT network (*top*) Energy landscapes, (*middle*) landscape of mean absolute error of energy prediction, and (*bottom*) DFT vs NNP class landscapes.

Compared to the NNP-20 landscapes, both NNP-50 and NNP-80 landscapes show reasonable robustness during the search. Still, one observes some discrepancies between the NNP-50 energy landscape (Figure 11, top center) and the DFT energy landscape (Figure 7), especially apparent in high-energy regions, which is corroborated by the NNP-50 energy prediction error landscape (Figure 11, center). Notice that the NNP-80 and DFT energy landscapes (Figure 11, top right and Figure 7, respectively) are rather similar which is reflected by relatively small regions of high energy prediction errors observed in the NNP-80 landscape (Figure 11, middle right). Finally, one can see that the areas with domination of NNP generated structures on the landscapes for NNP-50 (Figure 10, bottom center) and NNP-80 (Figure 11, bottom right) runs roughly correspond to the zones with large prediction error shown in Figure 11, middle. The oversampled regions (i.e., with large relative population) characterizing high energy prediction error contain geometries for which NNP underestimates energy. These

structures are likely to appear more kinetically accessible, therefore misguidedly steering the reaction path search algorithm toward these regions.

The above results show that the NNP models have a limited extrapolation power. Only large enough training sets NNP-50 and NNP-80 containing sufficient number of main reactive species lead to the models with reasonable predictive performance.

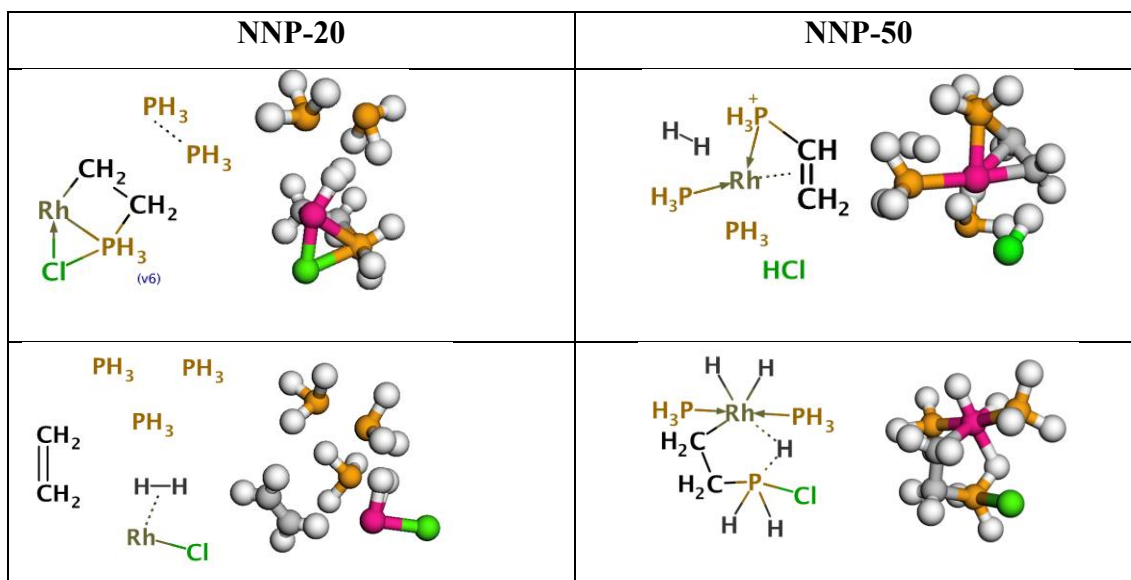


Figure 12. Examples of chemically unreasonable structures generated with the NNP-20 and NNP-50 models.

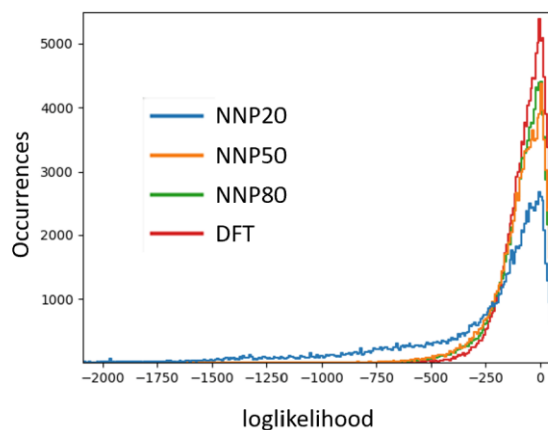


Figure 13. Loglikelihood (\mathcal{L}) distribution for the DFT, NNP-20, NNP-50 and NNP-80 data projected on the DFT manifold. One can see that the NNP-20 curve is shifted to the area of negative \mathcal{L} values which means that the related data are badly described by the DFT manifold.

Graph-based vs Chemography representation of reactions path networks

The graph-based representation is widely used in chemoinformatics in order to build chemical space networks (CSN)^[24] allowing to analyze structure-activity relations in heterogeneous compound data sets. In CSN, each node represents a compound, whereas each edge characterizes a pairwise relation between the compounds (e.g., common scaffold, molecular match pair, Tanimoto or Tversky similarity higher than a certain threshold). Size and color of nodes can be used to encode, respectively, the number of node connections and the compounds activity.

While CSNs provide both global and local views of activity landscapes they are—like other network representations— not applicable for visualization of very large datasets. With increasing numbers of data points, networks generally become difficult to navigate. Thus, to provide SPR views of increasingly large data sets, other methodologies should be considered. GTM is particularly suitable for this purpose, because, in parallel to considering individual objects, it allows to consider data distribution functions which significantly extends this approach to the analysis of ultra-large chemical libraries. On the other hand, GTM does not establish pairwise relationships between objects directly pointed in graph-based approaches.

With respect to the reaction path networks, conventional graph-based representation allows to visualize reaction pathways, but is limited to accommodate information about geometry, topology and electronic parameters of chemical structures. Increasing of the number of nodes (EQs) leads to serious problem of the network visualization, as it is demonstrated in Figure 2. Comparison of two different generated reaction path networks is hardly possible within the above approach.

GTM method – an advanced chemography approach – allows to overcome these limitations. It allows to group together on a 2D maps chemical structures possessing similar geometry and to establish structure-property relationships with the help of property landscapes. It identifies a common frame able to accommodate different reaction path networks, which make possible their comparison. On the other hand, the reaction pathways information is missed and, therefore, the graph-based and chemography approaches complement each other. Kayastha et al.^[24] suggested to combine these two approaches: GTM can be used for “satellite” view of the ensemble of data whereas graph-based representation is particularly important to describe objects relations (a part of network) in selected zone of the chemical space.

Conclusion

This study demonstrates that Generative Topographic Mapping approach is an efficient and useful tool for reaction path network visualization and analysis. GTM represents both individual molecular structures and their statistical distribution on a 2-dimensional map. Because of its probabilistic nature, GTM is able to handle (ultra)large chemical libraries, and in particular, reaction path networks including $>10^5$ structures. By combining different GTM-based property landscapes, one can identify regions of chemical interest, perform fast and intuitive comparison of different reaction path networks and monitor the reaction path exploration.

GTM is a dimensionality reduction method and, therefore, its efficiency strongly depends on the choice of descriptors encoding 3D molecular structures. Proposed in this work, the Distance Distribution descriptors (D^3) are alignment-free and invariant by rotation, translation, and permutation of atoms with the same atomic number. The descriptor vector size depends on the number of atom types in the considered molecular structure, but, unlike the previously reported PSBD descriptors, not on the total number of atoms. The D^3 descriptors, on one hand, allow to distinguish different molecular structures, and, on the other hand, to group on GTM structures possessing similar geometries.

Since GTM manifold delineates a frame of the explored reaction space, a retrospective analysis allows to monitor reaction path network expansion. This analysis shows how quickly the path search reaches the kinetically relevant reaction species and provides an information about data density in different areas of reaction space. The latter is important to draw conclusion concerning the suitability of generated DFT data at the given step for the training of NNP model.

This study demonstrates that GTM is a unique tool for comparing different reaction path networks. The map accommodating two networks (e.g, DFT and NNP) clearly shows both overlapping areas and zones occupied exclusively by one particular network. If the manifold is built on DFT data, the loglikelihood parameter helps to identify the NNP “novelties”, i.e., structures which significantly differ from the DFT generated ones.

Last but not least, compared to the conventional graph-based representation, GTM doesn't explicitly show the reaction pathways. That's why the combination of the two above approaches for the visualization and analysis of large reaction path networks looks very promising.

Software tools used

Generative Topographic Maps were built using the ISIDA/GTM program).^[29] The Marvin tool was used for depicting chemical 2D structures [Marvin version 23.2, ChemAxon (<https://www.chemaxon.com>)].^[25] The 3Dmol tool was used to display 3D chemical structures.^[26] Plots were generated using matplotlib and seaborn.^[27,28]

Data and software availability

The Python code for generation of D³ descriptors is available at <https://github.com/icredd-cheminfo/DistanceDistributionDescriptors>

Authors contributions

AV conceived the study and supervised the project. PG developed the tools for reaction networks analysis using ISIDA/GTM and performed the calculations. RS developed the Python code for descriptors calculations. PG and RS analyzed the results with the support of YH and SM. All authors reviewed and edited the manuscript.

Conflict of interest statement. The authors declare no conflict of interest.

Acknowledgement

We thank Dr Dragos Horvath, Dr Gilles Marcou and Dr Fanny Bonachera from the University of Strasbourg for the help with the implementation of ISIDA/GTM program and stimulating discussion.

References

- [1] A. Fernández-Ramos, J. A. Miller, S. J. Klippenstein, D. G. Truhlar, *Chem. Rev.* **2006**, *106*, 4518–4584.
- [2] S. Kozuch, S. Shaik, *Acc. Chem. Res.* **2011**, *44*, 101–110.
- [3] J. N. Harvey, F. Himo, F. Maseras, L. Perrin, *ACS Catal.* **2019**, *9*, 6803–6813.
- [4] S. Maeda, Y. Harabuchi, H. Hayashi, T. Mita, *Annu. Rev. Phys. Chem.* **2023**, *74*, 287–311.
- [5] S. Maeda, K. Ohno, K. Morokuma, *Phys. Chem. Chem. Phys.* **2013**, *15*, 3683–3701.
- [6] A. L. Dewyer, A. J. Argüelles, P. M. Zimmerman, *WIREs Comput Mol Sci* **2018**, *8*, e1354.
- [7] G. N. Simm, A. C. Vaucher, M. Reiher, *J. Phys. Chem. A* **2019**, *123*, 385–399.
- [8] S. Maeda, K. Morokuma, *J. Chem. Phys.* **2010**, *132*, 241102.
- [9] S. Maeda, T. Taketsugu, K. Morokuma, *J. Comput. Chem.* **2014**, *35*, 166–173.
- [10] S. Maeda, Y. Harabuchi, *WIREs Comput Mol Sci* **2021**, *11*, e1538.
- [11] Y. Harabuchi, S. Maeda, ChemRxiv **2022**, DOI 10.26434/chemrxiv-2022-tl4vj.
- [12] R. Staub, P. Gantzer, Y. Harabuchi, S. Maeda, A. Varnek, *Molecules* **2023**, *28*, 4477.
- [13] O. M. Becker, M. Karplus, *J. Chem. Phys.* **1997**, *106*, 1495–1517.
- [14] D. J. Wales, *The J. Chem. Phys.* **2015**, *142*, 130901.

- [15] J. A. Joseph, K. Röder, D. Chakraborty, R. G. Mantell, D. J. Wales, *Chem. Commun.* **2017**, *53*, 6974–6988.
- [16] Y. Sumiya, S. Maeda, *Chem. Lett.* **2019**, *48*, 47–50.
- [17] A. J. Birch, D. H. Williamson, in *Organic Reactions* (Ed.: John Wiley & Sons, Inc.), John Wiley & Sons, Inc., Hoboken, NJ, USA, **2011**, pp. 1–186.
- [18] C. M. Bishop, M. Svensén, C. K. I. Williams, *Neural Computation* **1998**, *10*, 215–234.
- [19] H. A. Gaspar, I. I. Baskin, G. Marcou, D. Horvath, A. Varnek, *Molecular Informatics* **2015**, *34*, 348–356.
- [20] T. Tsutsumi, Y. Ono, T. Taketsugu, *Chem. Commun.* **2021**, *57*, 11734–11750.
- [21] F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi, M. Ceriotti, *Chem. Rev.* **2021**, *121*, 9759–9815.
- [22] K. T. Schütt, H. E. Saucedo, P.-J. Kindermans, A. Tkatchenko, K.-R. Müller, *J. Chem. Phys.* **2018**, *148*, 241722.
- [23] D. Horvath, J. Brown, G. Marcou, A. Varnek, *Challenges* **2014**, *5*, 450–472.
- [24] S. Kayastha, R. Kunimoto, D. Horvath, A. Varnek, J. Bajorath, *J Comput Aided Mol Des* **2017**, *31*, 961–977.
- [25] Marvin version 23.2, ChemAxon (<https://www.chemaxon.com>)
- [26] N. Rego, D. Koes, *Bioinformatics* **2015**, *31*, 1322–1324.
- [27] J. D. Hunter, *Comput. Sci. Eng.* **2007**, *9*, 90–95.
- [28] M. Waskom, *JOSS* **2021**, *6*, 3021.
- [29] R. Pikalyova, Yu. Zabolotna, D. Horvath, G. Marcou, A. Varnek *J. Chem. Inf. Model.*, **2023**, *63*, 5571–5582

Supporting information

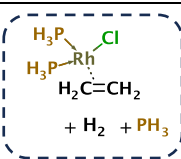
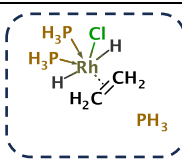
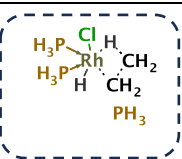
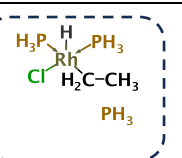
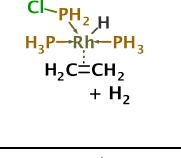
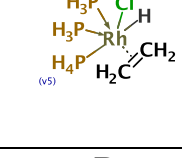
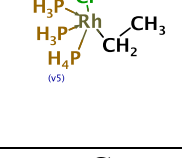
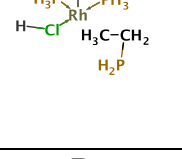
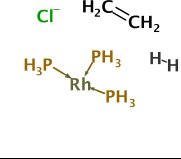
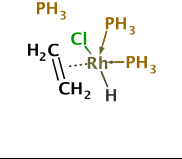
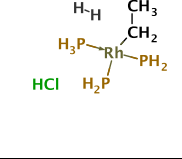
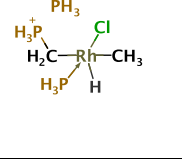
GTM-guided analysis of a Wilkinson's-like reaction network

Philippe Gantzer¹, Ruben Staub¹, Yu Harabuchi¹, Satoshi Maeda¹ and Alexandre Varnek*^{1,2}

¹ ICReDD

² University of Strasbourg

Table S1. 2D structural motifs of structures populating zones corresponding to reaction steps 2-5 and high energy zones A-D. The most kinetically relevant 2D motifs are given in dashed frames.

Zone	2	3	4	5
2D motifs	 $\text{H}_3\text{P}-\text{Rh}-\text{Cl}$ $\text{H}_3\text{P}-\text{Rh}-\text{Cl}$ $\text{H}_2\text{C}=\text{CH}_2$ $+ \text{H}_2 + \text{PH}_3$	 $\text{H}_3\text{P}-\text{Rh}-\text{Cl}$ $\text{H}_3\text{P}-\text{Rh}-\text{Cl}$ $\text{H}_2\text{C}=\text{CH}_2$ PH_3	 $\text{H}_3\text{P}-\text{Rh}-\text{Cl}$ $\text{H}_3\text{P}-\text{Rh}-\text{Cl}$ $\text{H}_2\text{C}=\text{CH}_2$ PH_3	 $\text{H}_3\text{P}-\text{Rh}-\text{Cl}$ $\text{H}_3\text{P}-\text{Rh}-\text{Cl}$ $\text{H}_2\text{C}-\text{CH}_3$ PH_3
	 $\text{Cl}-\text{PH}_2-\text{H}$ $\text{H}_3\text{P}-\text{Rh}-\text{PH}_3$ $\text{H}_2\text{C}=\text{CH}_2$ $+ \text{H}_2$	 $\text{H}_3\text{P}-\text{Rh}-\text{Cl}$ $\text{H}_3\text{P}-\text{Rh}-\text{Cl}$ H_4P $\text{H}_2\text{C}=\text{CH}_2$ $(v5)$	 $\text{H}_3\text{P}-\text{Rh}-\text{Cl}$ $\text{H}_3\text{P}-\text{Rh}-\text{Cl}$ H_4P CH_3 CH_2 $(v5)$	 $\text{H}_3\text{P}-\text{Rh}-\text{Cl}$ $\text{H}_3\text{P}-\text{Rh}-\text{Cl}$ $\text{H}_3\text{C}-\text{CH}_2$ H_2P
	A	B	C	D
	 Cl^- $\text{H}_2\text{C}=\text{CH}_2$ $\text{H}_3\text{P}-\text{Rh}-\text{PH}_3$ $\text{H}-\text{H}$	 PH_3 PH_3 $\text{H}_2\text{C}=\text{CH}_2$ Cl $\text{H}_3\text{P}-\text{Rh}-\text{PH}_3$ H	 $\text{H}-\text{H}$ CH_3 $\text{H}_3\text{P}-\text{Rh}-\text{PH}_2$ CH_2 HCl H_2P	 PH_3 Cl $\text{H}_3\text{P}-\text{Rh}-\text{PH}_3$ $\text{H}_2\text{C}-\text{CH}_3$ H_3P H