Statistics and Bias-Free Sampling of Reaction Mechanisms from Reaction Network Models

Dmitrij Rappoport*,†

† *Department of Chemistry, 1102 Natural Sciences 2, University of California, Irvine CA, 92697-2025*

E-mail: dmitrij@rappoport.org

Abstract

Selection bias is inevitable in manually curated computational reaction databases but can have a significant impact on generalizability of quantum chemical methods and machine learning models derived from these data sets. Here, we propose quasireaction subgraphs as a discrete, graph-based representation of reaction mechanisms that has a well-defined associated probability space and admits a similarity function using graph kernels. Quasireaction subgraphs are thus well suited for constructing representative or diverse data sets of reactions. Quasireaction subgraphs are defined as subgraphs of a network of formal bond breaks and bond formations (transition network) composed of all shortest paths between reactant and product nodes. However, due to their purely geometric construction, they do not guarantee that the corresponding reaction mechanisms are thermodynamically and kinetically feasible. As a result, a binary classification of feasible (reaction subgraphs) and infeasible (non-reactive subgraphs) must be applied after sampling. In this paper, we describe the construction and properties of quasireaction subgraphs and characterize the statistics of quasireaction subgraphs from CHO transition networks with up to six nonhydrogen atoms. We explore their clustering using Weisfeiler–Lehman graph kernels.

1 Introduction

Computational reaction databases[1–6] are a crucial source of information for benchmarking quantum chemical methods and training machine learning models. Large computational benchmark sets of barrier heights, for example, the Minnesota set of databases by Truhlar and co-workers,[1,7,8] the GMTKN55 database by Goerigk and Grimme,[2] the MGCDB84 database by Mardirossian and Head-Gordon,[3] and the BH9 database by DiLabio and co-workers,[5,6] are the results of years of careful collection and curation of experimental data and high-quality quantum chemical results. The collected data sets are often used as the "ground truth" in training machine learning models.[9,10]

The composition of these databases is in part limited by the availability of accurate experimental barrier heights, which tend to bias the data sets toward well-characterized reactions with simple mechanisms. The bulk of the current benchmark data for barrier heights of organic reactions is made up of few types of reactions including hydrogen transfer,[11,12] proton transfer,[13,14] nonhydrogen atom transfer, nucleophilic substitution, association, and unimolecular reactions,[12,15] pericyclic reactions,[16–21] S_N2 and E2 reactions,[22] radical rearrangements, pericyclic reactions, hydrogen and halogen transfer, B/Si reactions, nucleophilic substitution, and nucleophilic addition reactions.[5,6] In contrast, reactions resembling enzyme catalysis, are underrepresented in benchmark databases.[23]

Although computational databases play a key role in the development of quantum chemical methods, it is currently an open question how representative (bias-free) or diverse the underlying data sets are. The training of increasingly flexible machine learning models critically depends on data sets that are representative of broader chemical reactivity, which is sometimes referred to as "reaction space".[24] It is thus desirable to have methods for constructing bias-free samples of some provably comprehensive ground set of chemical reactions and for quantifying similarity of reactions, which can be used to assess the diversity of data sets,[25] see also Refs. 26–29 for applications in chemistry. Both groups of methods can be implemented in a principled way if we can construct a ground set of representations of chemical reactions that is equipped with a probability measure (probability space) and a similarity function. In order for the results obtained from this data set to be generalizable, the ground set should encompass chemical reactivity within well-defined limits.

Here, we introduce a discrete representation of chemical reaction mechanisms that possesses both a probability space and a similarity function definition. Thus, it is suitable for constructing bias-free data sets by sampling and for evaluating diversity of data sets using the similarity measure. This feature representation is based on subgraphs of a network of formal bond breaks and bond formations (transition network, TN).[30–32] TNs are discrete analogs of reactive potential energy surfaces (PESs), in which the continuous atomic dis-

placements are replaced by discrete structure changes. As a finite graph, each TN defines a discrete probability space of its subgraphs. The complete set of TN subgraphs can thus be obtained by enumeration. Additionally, a similarity function for subgraphs can be defined using graph kernels.[33,34]

However, the purely geometric construction of TN subgraphs cannot take into account the thermodynamic and kinetic feasibility of the corresponding reaction mechanisms. We note that if a reaction mechanism can be written as a molecular graph transformation, it can also be represented as a TN subgraph, which we denote as a reaction subgraph. On the other hand, a randomly selected TN subgraph may either be a reaction subgraph, or it may correspond to a hypothetical transformation that is thermodynamically or kinetically infeasible (non-reactive subgraph). We will refer to the complete set of TN subgraphs (both reactive and non-reactive) as quasireaction subgraphs to make the distinction clear.

The goal of this paper is to define quasireaction subgraphs as feature representations of reaction mechanisms and to characterize the statistics and clustering of quasireaction subgraphs from CHO TNs with up to six nonhydrogen atoms. Our focus in this paper is on graph theoretical methods. For the construction of bias-free data sets, an additional step is necessary, the binary classification of quasireaction subgraphs into reactive (positive class) and non-reactive (negative class). This classification can be performed by explicit transition state (TS) searches and will be addressed in detail in a future paper.

Several alternative approaches to constructing representative and diverse reaction databases should be mentioned. Most of these approaches are based on targeted subset selection from the curated databases. Data sets of difficult reactions were constructed by Patchkovskii and Ziegler.[35] Gould and Dale extracted a "poison subset" of the GMTKN55 set by selecting the data points with the largest errors.[36] An alternative strategy aims to select representative database subsets, for example, by minimizing the difference of the error measures between the subset and the full set,[37,38] by clustering,[39,40] or using feature selection methods.[41]

Very recently, approaches based on enumerated databases of organic compounds have

been published. A database of unimolecular rearrangements and their barrier heights was created by Green and co-workers using one-sided TS searches along predetermined reaction coordinates on the subset of the GDB-17 database with up to seven nonhydrogen atoms.[42,43] In a subsequent study, molecular structures near the TSs were sampled after refining TS structures using the nudged elastic band (NEB) method.[44] The resulting barrier heights were used to train a deep learning model for activation energies[45] and a message-passing neural network for predicting molecular energies in the vicinity of TSs.[46] However, these databases contain only a specific type of reactions, unimolecular rearrangements.

Computational databases based on graph enumeration have been used with successs in calculations of molecular energies and properties. Graph enumeration methods were used by Reymond and co-workers to create the GDB family of databases.[47–53] The GDB databases consist of neutral closed-shell organic molecules with up to 11, 13, and 17 nonhydrogen atoms (GDB-11 through GDB-17). The molecular structures were obtained by enumeration of non-isomorphic molecular graphs using the nauty/Traces graph isomorphism tools of McKay and Piperno,[54] followed by rule-based introduction of unsaturations and heteroatoms. With an eye toward drug-like molecules, the authors excluded compounds that were considered difficult to synthesize or too reactive, for example, anhydrides, hemiacetals, enols, and compounds containing heteroatom–heteratom bonds.

Several quantum chemical data sets were derived from subsets of GDB databases, for example, QM7 and QM9 data sets of von Lilienfeld and co-workers,[55–59] the Alchemy data set of Zhang and co-workers,[60] and the ANI data sets of non-equilibrium molecular structures by Roitberg, Isayev, Tretiak, and co-workers.[61–65] Interestingly, the QM9 data set showed a larger generalization error compared to the PC9 data set derived from the experimental PubChem database,[66–68] likely due to the absence of open-shell species and certain functional groups in GDB databases.[69,70] Combinatorial enumeration of molecular structures using graph-based substitutions is often used in high-throughput computational screening, for example, in the Harvard clean energy project of Aspuru-Guzik and co-workers.[71,72] An-

other group of approaches randomly samples structures for inclusion in the data set. The method of Wipf, Yang, Beratan, and co-workers constructed larger, drug-like molecules by a genetic-type algorithm on molecular graphs.[73] The "mindless" benchmarking approach of Korth and Grimme[74] generated eight-atomic molecules of random composition and optimized their structures using density functional theory (DFT). Sets of non-equilibrium molecular structures of selected molecules were sampled from molecular dynamics (MD) trajectories by the method of Müller, Tkatchenko, and co-workers.[75,76]

In comparison to the methods for constructing representative and diverse reaction databases described above, this work takes the opposite approach. The complete set of chemical reactions, for example, for CHO with up to six nonhydrogen atoms is not currently known. Therefore, instead of relying on a currently (experimentally or computationally) known subset of this reaction set and selecting its representative subsets, we construct a known superset (quasireaction subgraphs). This superset affords us the bias-free property by design but also necessitates binary classification of reactive and non-reactive subgraphs as a postprocessing step.

This paper is organized as follows. We describe the procedures for the construction of TNs, the sampling of quasireaction subgraphs, and their clustering using Weisfeiler–Lehman (WL) subtree and WL edge graph kernels[33,34] in Sec. 2. The statistics of the quasireaction subgraphs and the clustering results are presented in Sec. 3. The discussion is given in Sec. 4. We present our conclusions in Sec. 5.

2 Methods

2.1 Transition Networks

The reaction networks considered in this work are of the transition network (TN) type, that is, their network nodes contain collections of molecules subject to the fixed stoichiometry (atomic composition), and their network edges are stoichiometry-preserving transformations

from a predetermined rule set.[31,32] The stoichiometry-based TNs differ from the the chemical reaction networks (CRNs)[77–80] used in kinetic modeling, metabolism, or synthesis planning, which are referred to in the following as molecule-based CRNs for clarity. In molecule-based CRNs, network nodes correspond to individual molecules (reactants and products), while network edges are typically constructed between all reaction participants. TNs and molecule-based CRNs coincide only in the special case that all reactions have exactly one reactant and one product. The distinguishing feature of TNs is that they have an upper bound for the network size (number of nodes), determined by the number of ways the bonds may be distributed among the fixed set of atoms. In contrast, molecule-based CRNs do not have such natural upper limit since they can always continue producing larger and larger molecules, being unconstrained by stoichiometry. Moreover, due to the conserved stoichiometry, TNs can be understood as discretized versions of PESs, while molecule-based CRNs span many different PESs. In the remainder of this article, we only consider TNs. When referring to reaction networks, it is implied that they are of the TN type.

In this work, we consider TNs of stoichiometry $C_{\nu_C}H_{\nu_H}O_{\nu O}$ (CHO reaction networks) were constructed for $\nu_C + \nu_O = 2, \ldots, 6$ nonhydrogen atoms. The molecular graphs are represented by their SMILES strings,[81] ignoring stereochemical information. The stoichiometry-preserving transformations are chosen as polar bond breaks and bond formations that are consistent with the electronegativities of the elements carbon, hydrogen, and oxygen (normal polarity). These formal transformations describe the smallest discrete bonding changes that can be applied to molecular graphs and can be composed to represent reaction mechanisms. To incorporate the reactivity of multiple bonds, additional formal rules describing the "polarization" and "depolarization" of double and triple bonds are included in the rule set. Due to the composability of the formal transformations, any reaction mechanisms that consists of a transfer of one or more electron pairs ("arrow pushing"), can be represented within the TN. The complete rule set is shown in Table S1 of the Supplementary Information (SI). The reaction rules were encoded as SMARTS strings[82] and applied to the molecular structures

using the RDKit library.[83] The TN construction completed when no new nodes could be generated. As the rule set was chosen to be reversible (see Table S1 of the SI), nearly all network edges in the constructed network were reversible. The missing edges were due to inconsistent transformation between aromatic and Kekulé-type representations of molecular structures in RDKit. The TN construction was implemented in version 2 of the open-source colibri package.[84] All TNs were converted to undirected networks for further analysis. The networks were stored as compressed GraphML files.[85]

2.2 Quasireaction Subgraphs

To construct quasireaction subgraphs, all network nodes containing only neutral molecules (neutral nodes) were first identified. However, neutral nodes containing high-energy species were matched using a set of heuristic rules and excluded from consideration. The rules for matching high-energy species were: (i) more than 3 rings, (ii) triple and allene bonds in rings, (iii) double bonds at bridge atoms, and (iv) double bonds in fused 3-membered rings. See Table S2 of the SI for details. These heuristic rules help to exclude highly strained compounds from further analysis, among them several compounds that have been experimentally isolated, for example, benzvalene[86,87] and prismane.[88] The reaction mechanisms involving these compounds are likely to be significantly affected by strain, even if they are experimentally accessible. Nevertheless, it might be of interest to re-analyze them in future work.

A quasireaction subgraph is defined as the union of all simple paths between a pair of neutral nodes. In the following, we limit ourselves to quasireaction subgraphs composed of all shortest paths between pairs of neutral nodes. The paths that pass through neutral nodes other than the initial and final nodes are excluded. All internal nodes are thus non-neutral. All graphs are treated as unweighted with their edges labeled by the the applied reaction rules. The preference for shortest paths is equivalent to the principle of minimum chemical distance proposed by Ugi and co-workers in 1980.[89] However, as we described in previous

work, there are reactions, in which the shortest path does not correctly describe the reaction mechanism.[32] In these cases, it is preferable to extend the set of shortest paths between the reactant and product nodes to all simple paths of length $L' = L + s$, where L is the shortest path length and is the "slack" length $s \geq 0$. However, we will not consider this variant in the following. Shortest paths were computed for all pairs of neutral nodes using Dijkstra's algorithm. Quasireaction subgraphs were constructed only for pairs of neutral nodes with shortest path length $L \leq 8$. All graph algorithms were implemented using the NetworkX library.[90]

2.3 Clustering Using Graph Kernels

Quasireaction subgraphs are undirected graphs with discrete edge labels and thus can be classified into sharp isomorphism classes. However, this classification fails to take into account the similarities between graphs that differ by addition or removal of edges or by label substitutions. A more versatile classification approach uses clustering based on graph kernels.[91–93] Generally, a graph kernel $k(G, G')$ is a non-negative symmetric function of graphs G and G', which expresses a particular notion of their similarity. Pairs of similar graphs have higher values of $k(G, G')$ than dissimilar graphs. If the kernel function is normalized ($k(G, G) = 1$ for all G), then it can be straightforwardly converted to a distance function (metric) $d(G, G') = 1 - k(G, G')$, which can be used in standard clustering algorithms.

We used methods based on the Weisfeiler-Lehman (WL) graph isomorphism test[33] to identify discrete isomorphism classes of quasireaction subgraphs and to compute graph kernels for clustering. The WL algorithm performs an iterative relabeling of graph nodes based on the labels of their neighbors. For each node, the relabeling aggregates its label and the labels of the neighbor nodes into a multiset and compresses them into a new node label. Unique new labels are then added to the alphabet. This procedure is repeated for h iterations, after which each node label reflects the subtrees of height h rooted at the given node. In the WL isomorphism test, the multisets of the node labels of graphs G and G'

are compared after h relabeling iterations. The comparison can be simplified by hashing the multiset of node labels. However, the resulting isomorphism test is only approximate: Isomorphic labeled graphs are guaranteed to produce identical hashes but non-isomorphic graphs may have identical hashes with some (low) probability.

The WL algorithm is also the basis for the WL family of graph kernels.[33] Starting from a base graph kernel $k(G, G')$ between the graphs G and G', the WL kernel with h iterations is given by

$$k_{\mathrm{WL}}^{(h)}(G, G') = k(G_0, G'_0) + k(G_1, G'_1) + \ldots + k(G_h, G'_h)$$

where $G_0 = G$, $G'_0 = G'$ and the graphs G_i, G'_i for $i > 0$ are defined recursively by applying the WL relabeling procedure to G_{i-1} and G'_{i-1}, respectively. If the base graph kernel is the inner product between the vectors of node label counts, then WL algorithm produces the WL subtree graph kernel (WL-S).[33,34] Alternatively, the base graph kernel may be the inner product between the vectors of edge label counts (edge histogram), giving rise to the WL edge graph kernel (WL-E).[34] Since quasireaction subgraphs only have edge labels and no node labels, the WL-S graph kernel $k_{\text{WL-S}}(G, G')$ only compares the topological properties of the quasireaction subgraphs G and G', while the WL-E graph kernel $k_{\text{WL-E}}(G, G')$ additionally includes information about edge labels (reaction rules). The WL hashes for isomorphism tests were computed for $h = 3$ and hash length $w = 16$ using the NetworkX library.[90] The calculations of the WL-S and WL-E graph kernels used $h = 3$ and were performed with the GraKel library.[94]

To perform clustering, the normalized kernel matrix was first converted to a distance matrix and projected onto a two-dimensional feature space using the uniform manifold approximation and projection (UMAP) method.[95] The number of neighbors in the UMAP projection was $N_{\text{neighbors}} = 50$. Increasing the number of neighbors to 200 did not qualitatively change the results, while smaller values of $N_{\text{neighbors}}$ produced an relatively structureless distribution. The clusters were obtained by k-means clustering[96,97] with the optimal number of clusters k_{opt} determined from the plot of silhouette scores for $k = 3, \ldots, 49$ clusters.[98] The

implementation used the umap-learn library[99] for UMAP projection and the scikit-learn library[100] for k-means clustering and the evaluation of silhouette scores.

3 Results

3.1 Transition Network Statistics

The basic statistics of CHO TNs with up to six nonhydrogen atoms are shown in Fig. 1. A total of 159 TNs networks are analyzed. The TNs are grouped by the formal carbon oxidation state $\xi_C = (-\nu_H + 2\,\nu_O)/\nu_C$. The network size (number of nodes, N_{nodes}) ranges from a single node in CO to 49,637 nodes and 194,750 edges in the C_5H_6O network. The expected exponential increase in network size with the number of nonhydrogen atoms is observed in Fig. 1(a). Additionally, the network size varies by several orders of magnitude as a function of the carbon oxidation state, with the largest networks located at $\xi_C \approx -1$.

The counts of neutral nodes are shown in Fig. 1(b) and follow the same trend as network size. The networks containing the largest number of neutral nodes are C_5H_8O, C_5H_6O, and $C_4H_6O_2$ with carbon oxidation state of approximately -1. The combined set of neutral molecules from all TNs has 3246 molecules in total and 2222 molecules that do not contain strained motifs (i)–(iv), see Sec. 2.2. This set has comprehensive coverage of the experimentally known oxygen-containing molecules. A search of neutral CHO molecules with up to six nonhydrogen atoms in the CAS SciFinder database[101] yields 1600 results (ignoring stereo descriptors), of which 1375 do not contain strained motifs (i)–(iv). The search results include experimentally known compounds as well as those that are characterized only by computation. However, all of the 1375 neutral molecules that are returned by the SciFinder search and not excluded as high-energy are found in our combined set. The coverage of the GDB-7 data set[47] is similarly complete: all of the 379 molecules with up to 6 nonhydrogen atoms from the CO subset of the GDB-7 database are present in our combined set. The complete statistics of TNs generated in this work are given in Table S3 of the SI.

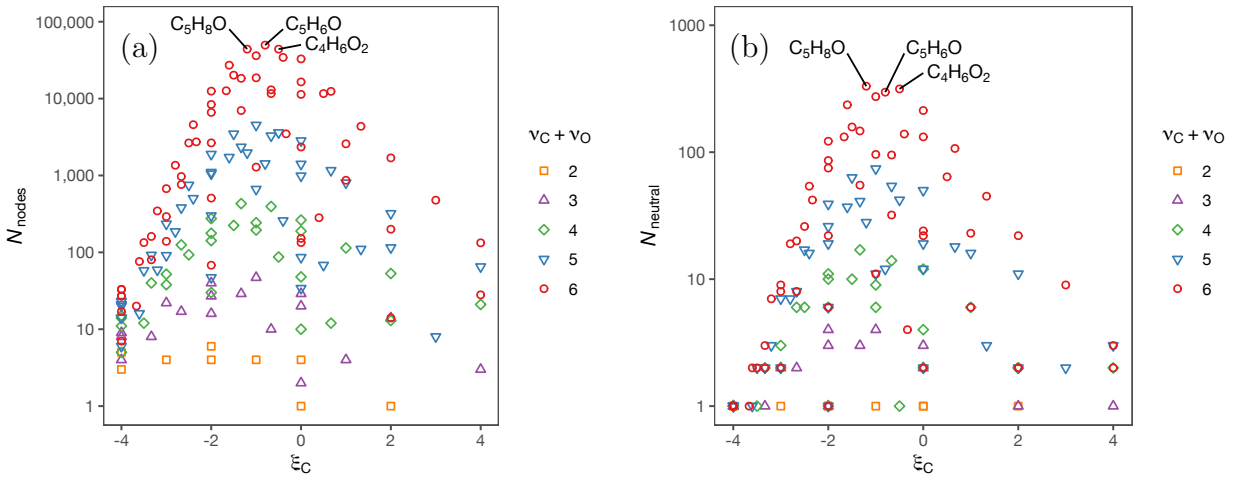


Figure 1: (a) Network size (number of nodes N_{nodes}) and (b) number of neutral nodes (N_{neutral}) in $C_{\nu_C}H_{\nu_H}O_{\nu O}$ reaction networks ($\nu_C + \nu_O = 2, \ldots, 6$) by carbon formal oxidation state ξ_C.

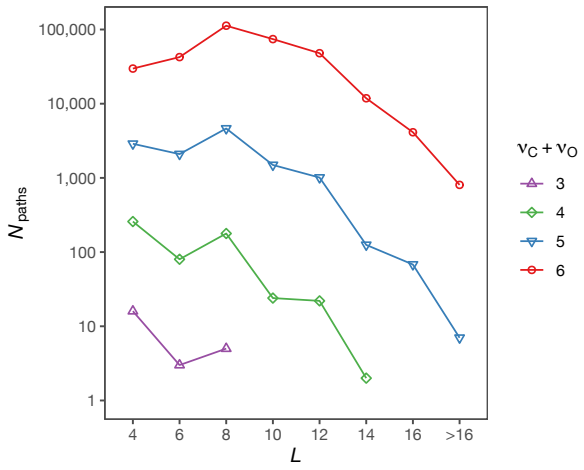


Figure 2: Counts of shortest paths N_{paths} between neutral nodes in $C_{\nu_C}H_{\nu_H}O_{\nu O}$ TNs ($\nu_C + \nu_O = 2, \ldots, 6$) by path length $L = 2, \ldots, 16, > 16$.

Each network edge in TNs describes exactly one bond break or bond formation. The shortest path length between a pair of nodes can thus be understood as the total number of bond changes in course of the reaction. Moreover, the transfer of one electron pair is equivalent to one polar bond breaking and one polar bond formation. The total number of transferred electron pairs is half of the shortest path length. The statistics of the shortest paths between pairs of neutral nodes are shown in Fig. 2. By construction, all paths between

neutral nodes have an even number of edges. The minimal shortest path distance is $L = 4$, while the maximal distance is bounded by the diameter of the network. One shortest path of length $L = 2$ was generated due to inaccuracies of the interconversion between the aromatic and the Kekulé structural representations.

The complexity of the reaction mechanism increases with the shortest paths length. At the same time, the number of simultaneous bond breaks and formations in most elementary reactions is limited. While some pericyclic reactions have been described with as many as 10 bonds rearranging in a concerted mechanism, such reactions are generally rare.[102–104] Based on this observation, we restrict our analysis to reaction mechanisms with shortest path length $L \leq 8$. Of the total 336,602 pairs of neutral nodes across all TNs, 194,896 fall into this category. In 118 pairs we find another neutral node at the midpoint of the shortest path. As discussed in Sec. 2.2, these paths are reducible to smaller structures. After excluding these pairs, we obtain a set of 194,778 pairs of neutral nodes, for which quasireaction subgraphs are constructed. The quasireaction subgraphs for these pairs will be characterized in the following. The complete data set of quasireaction subgraphs is available for public download from Zenodo.[105]

3.2 Quasireaction Subgraph Statistics

The quasireaction subgraph contains the union of all shortest paths between the initial and final nodes. Since it is an undirected graph, it simultaneously describes the mechanisms of the forward and reverse reactions. The alternative shortest paths describe the different decompositions of the overall reaction mechanism into discrete bond breaks and bond formations. We illustrate this by the reaction subgraph for water addition to ethene, $CH_2{=}CH_2 + H_2O \rightarrow CH_3CH_2OH$, which is shown in Fig. 3. It consists of 5 shortest paths of length $L = 4$. The combination of the shortest paths creates a subgraph with $N_{\text{nodes}} = 8$ and $N_{\text{edges}} = 10$.

Fig. 4 shows the statistics of the quasireaction subgraphs with shortest path length $L =$

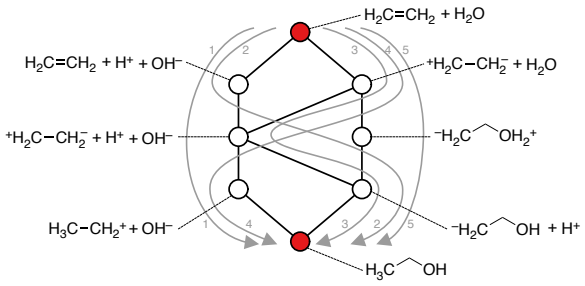


Figure 3: Shortest paths in the reaction mechanism of water addition to ethene. The reactant and product nodes are indicated by red solid circles, internal nodes are shown as empty circles. The shortest paths are indicated by gray arrows.

$4, 6, 8$ as a function of their number of nodes N_{nodes} and number of edges N_{edges}. The sizes of the empty circles indicate the subgraph counts for the given numbers of nodes and edges. Several node–edge size classes are significantly enriched among quasireaction subgraphs. 88 distinct classes were found for $L = 4$, of which $N_{\text{nodes}} = 7$, $N_{\text{edges}} = 8$ and $N_{\text{nodes}} = 8$, $N_{\text{edges}} = 10$ are by far the most frequent. Of the 932 classes for $L = 6$, the class with $N_{\text{nodes}} = 21$, $N_{\text{edges}} = 38$ is dominant. Even quasireaction subgraphs with $L = 8$ show enrichment of several node–edge size classes, in particular, $N_{\text{nodes}} = 52$, $N_{\text{edges}} = 105$. However, as many as 10,341 distinct classes are found for $L = 8$.

The quasireaction subgraphs belonging to the same node–edge size class are not necessarily isomorphic. We distinguish non-isomorphic subgraphs within the same class by an additional identifier P. The topologies of the most frequent subgraph patterns for $L = 4, 6, 8$ are shown in Fig. 5, Fig. 6, and Fig. 7, respectively. Their statistics are given in Table 1. We note that all of the most frequent node–edge size classes for $L = 4, 6$ are almost entirely composed of isomorphic subgraphs. The water addition and elimination reactions of Fig. 3 belong to the most frequent quasireaction subgraph pattern 8,10 A. The second most frequent pattern is 7,8 A and describes different orderings of two bond breaks and two bond formations, which would be customarily written as a linear reaction mechanism. An example of this pattern is the isomerization of 1,2,-butadiene to the more stable 1,3-butadiene isomer. The remaining three subgraph patterns 11,15 A, 12,16 A, and 13,18 A are obtained by fusing two basic 7,8 A motifs and adding extra nodes. The 11,15 A pattern describes, for example,

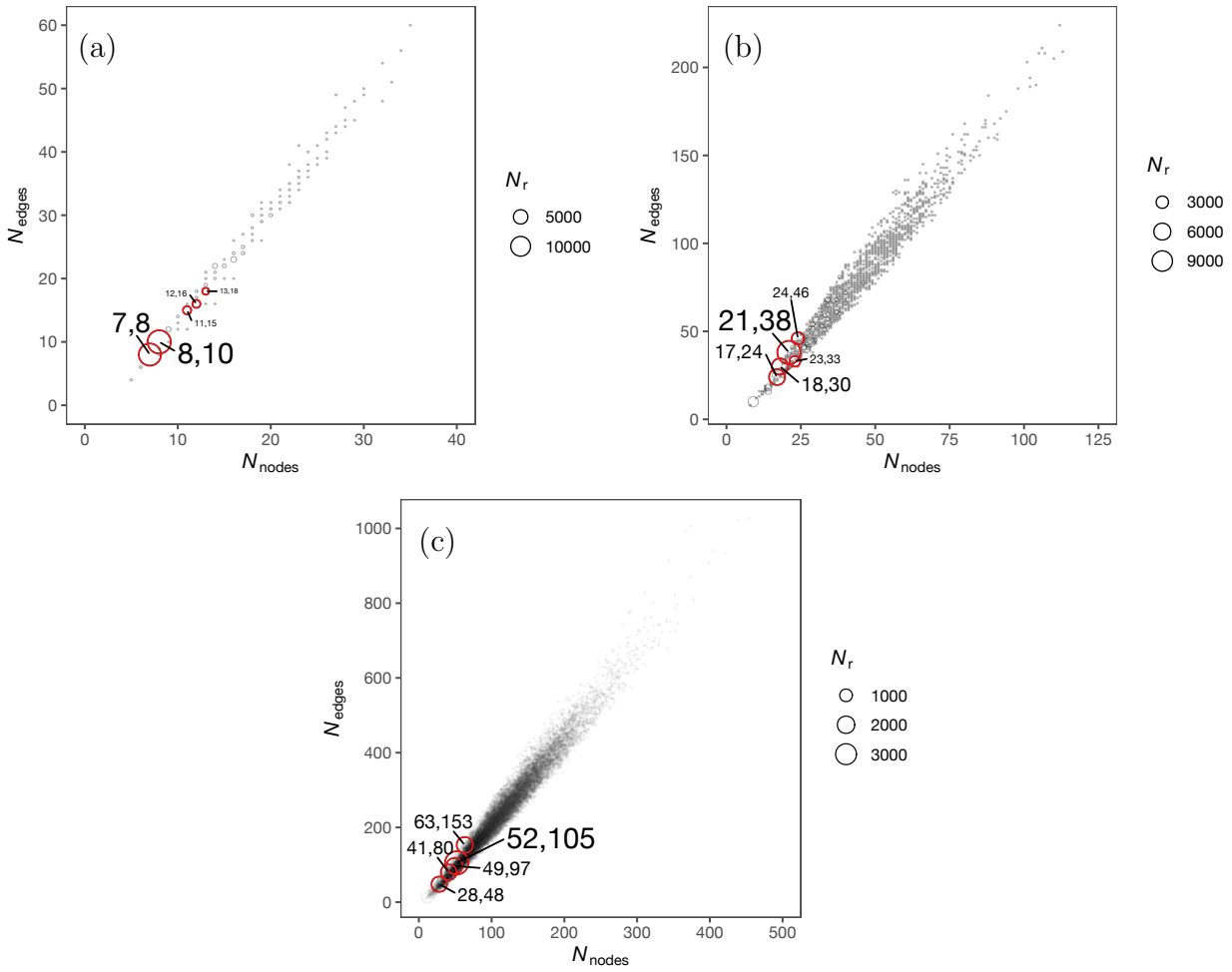


Figure 4: Counts of quasireaction subgraphs N_r by number of nodes N_{nodes} and edges N_{edges} with shortest path distance (a) $L = 4$, (b) $L = 6$, and (c) $L = 8$. The 5 most frequent N_{nodes},N_{edges} combinations for each value of L are shown in red.

the double bond migration in 1-butene → 2-butene. The 12,16 A pattern represents the hypothetical ring opening in 1-methylcyclopropene → 1-butyne. The 13,18 A pattern is found in the 1,3-hydroxyl migration in 3-buten-2-ol → 2-buten-1-ol. As the shortest path length increases, so does the variation in quasireaction subgraphs, while their symmetry tends to decrease. The largest node–edge size class $N_{nodes} = 52$, $N_{edges} = 105$ for $L = 8$ is split almost equally in two non-isomorphic subgraph patterns, denoted as 52,105 A and 52,105 B, see Fig. 7.

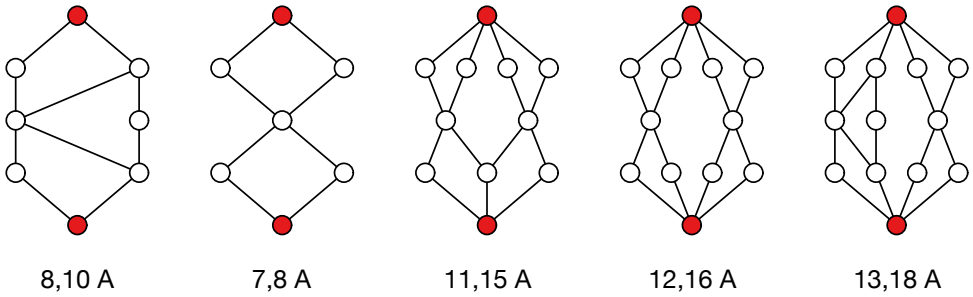


Figure 5: Most frequent quasireaction subgraph patterns with shortest path length $L = 4$. The patterns are labeled N_{nodes},N_{edges} P, where pattern identifier P distinguishes non-isomorphic patterns with the same numbers of nodes and edges. The reactant and product nodes are indicated by red solid circles, internal nodes are shown as empty circles.

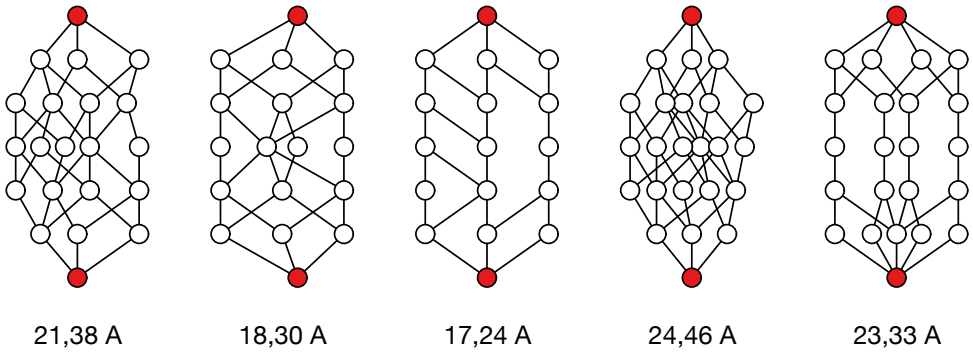


Figure 6: Most frequent quasireaction subgraph patterns with shortest path length $L = 6$.

3.3 Clustering of Quasireaction Subgraphs

Given the ground set of quasireaction subgraphs, the most straightforward sampling strategy is uniform sampling of a fixed fraction of the ground set. Alternatively, one could choose to sample examples by node–edge size class. However, because of the broad distribution of quasireaction subgraph patterns and the decreasing relative differences in their topologies, their classification discrete classes is too limiting. In the following we explore clustering approaches for the classification of the quasireaction subgraphs based on graph kernels. The similarity function defined by the graph kernel can be used to generate a maximally diverse subset.

In order to perform clustering, we uniformly sampled a subset of $N_{\text{sample}} = 10{,}000$

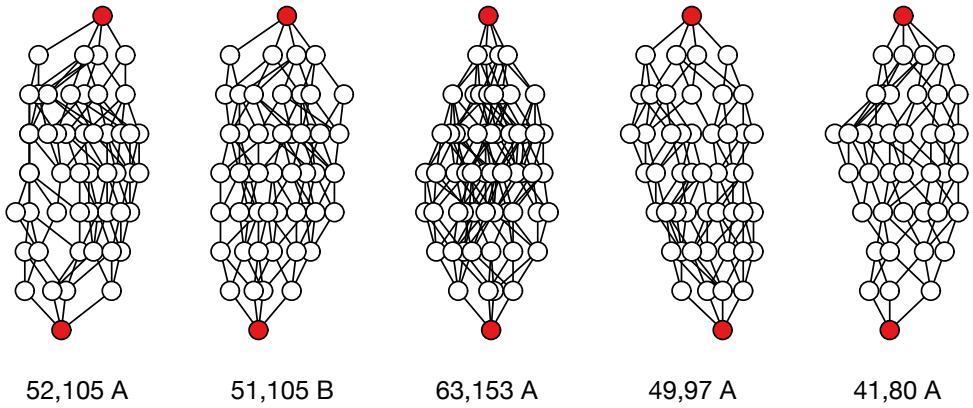


Figure 7: Most frequent quasireaction subgraph patterns with shortest path length $L = 8$.

Table 1: Counts of most frequent quasireaction subgraphs N_{r} by number of nodes N_{nodes} and edges N_{edges} with shortest path distance $L = 4, 6, 8$. The pattern identifier P distinguishes non-isomorphic patterns with the same numbers of nodes and edges.

$L = 4$				$L = 6$				$L = 8$			
N_{nodes}	N_{edges}	P	N_{r}	N_{nodes}	N_{edges}	P	N_{r}	N_{nodes}	N_{edges}	P	N_{r}
8	10	A	13822	21	38	A	11645	52	105	A	1896
7	8	A	12456	18	30	A	5440	52	105	B	1792
11	15	A	1312	17	24	A	5134	63	153	A	1623
12	16	A	1222	24	46	A	2867	49	97	A	1615
13	18	A	867	23	33	A	2091	41	80	A	1533
Total			32916				44655				117206

quasireaction subgraphs from the ground set for computing the WL-S and WL-E graph kernel matrices. The subset reproduces key topological characteristics of the ground set with good accuracy. The average node degree of the quasireaction subgraphs is $\bar{k} = 3.48 \pm 0.69$ on the subset and the average shortest length between the reactant and product nodes is $\bar{L} = 6.87 \pm 1.53$. Both characteristics match the averages over the ground set to two significant figures. As discussed in Sec. 2, the WL-S graph kernel is only sensitive to the topology of the quasireaction subgraphs, while the WL-E graph kernel additionally incorporates information about reaction rules as edge labels. The two-dimensional feature projections of the sampled quasireaction subgraphs using UMAP are displayed in Fig. 8. The five most

frequent subgraph patterns form well-separated clusters in the WL-S plot. The remaining quasireaction subgraphs forms a single giant cluster near the center that contains nearly half of the sample.

The plot of silhouette scores for $k = 3, \ldots, 49$ clusters in shown in Fig. S1 of the SI. The optimal number of clusters in the k-means algorithm is subject to some ambiguity. Here, k_{opt} was simply determined as the smallest value of k that corresponds to a local maximum of the silhouette score above 0.5. With the WL-S graph kernel, we obtain the optimal number of clusters as $k_{\text{opt}} = 7$. These findings indicate that sampled quasireaction subgraphs contain a set of preferred topological motifs for $L = 4, 6$ superimposed on a broad distribution of topological patterns. Including the edge labels in the WL-E graph kernel introduces much more structure in the distribution of quasireaction subgraph patterns. The optimal number of clusters with the WL-E kernel from the analysis of silhouette scores is $k_{\text{opt}} = 37$. The distribution of cluster sizes with the WL-S and WL-E graph kernels is shown in Fig. 9 and ranges from 58 to 1020. Due to the stochastic nature of the sample selection and UMAP feature projection, the number and sizes of clusters show some variation across multiple simulations. However, the qualitative findings are unaffected.

4 Discussion

4.1 Information Content of Reaction Subgraphs

The feature representation of reaction mechanisms as reaction subgraphs of discrete bond breaks and bond formations shares structural similarity with several representations developed in the past but offers some distinct benefits. The most closely related representation are the reaction matrices of Dugundji and Ugi.[30,106,107] The reaction matrix is defined as the difference of the bond–electron (BE) matrices of the products and the reactants. An example of the Dugundji–Ugi reaction matrix for the water addition to ethene is given in Fig. S2 of the SI. The entries of the reaction matrix reflect the changes in bond orders between pairs

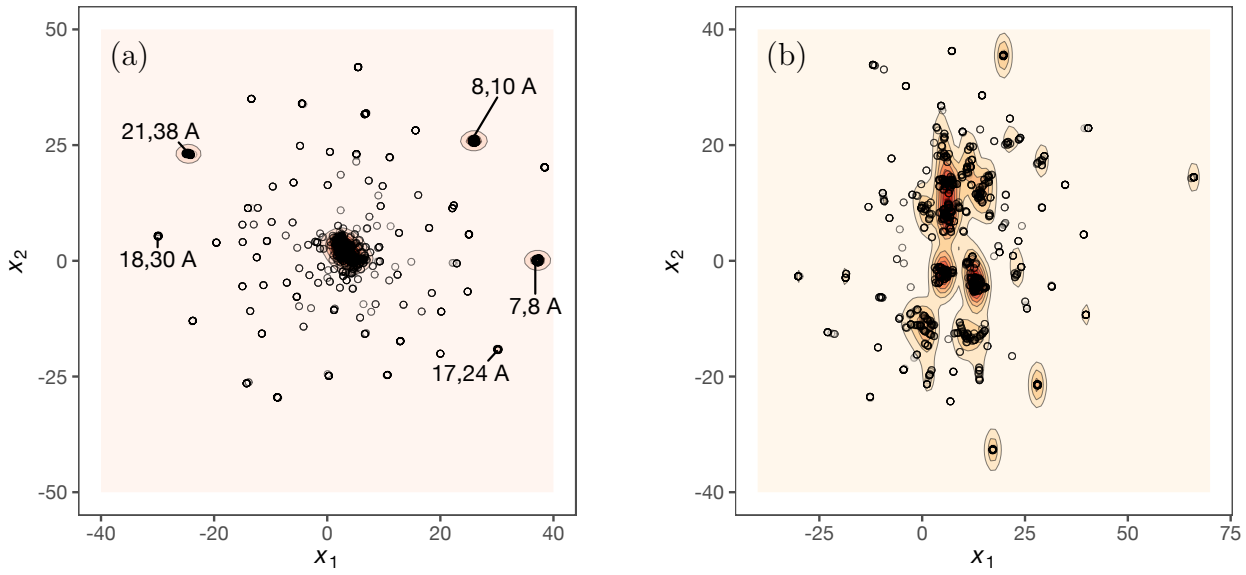


Figure 8: Density plot of two-dimensional UMAP feature embedding of $N_{\text{sample}} = 10{,}000$ quasireaction subgraphs using (a) WL subtree graph kernel (WL-S) and (b) WL edge graph kernel (WL-E). The five most frequent quasireaction subgraph patterns are labeled in the WL-S plot.

of atoms (and of numbers of lone pairs). Thus the reaction matrix can be represented as a sum of elementary reaction matrices for the individual bond breaks and bond formations. In comparison to reaction matrices, reaction subgraphs additionally encode information about the ordering of the bond break and bond formation steps and their relative topology. For example, both 7,8 A and 8,10 A subgraphs describe a reaction mechanism with two bond breaks and two bond formations (Fig. 5). While their reaction matrices are identical (up to a permutation), the two mechanisms differ in the number of reaction paths (4 for 7,8 A subgraphs but 5 for 8,10 A subgraphs).

Several reaction template notations[108–111] and the imaginary transition state representations of Fujita[112,113] similarly describe the total of bond breaks and bond formations but can additionally encode the topology of the molecular graph. In contrast, the reaction subgraphs represent the (discretized) topology of the reaction paths. The reaction paths can incorporate energy information, as we showed in previous work,[32] which allows to distinguish reactions that conform to the same reaction template but follow different mechanistic paths based on

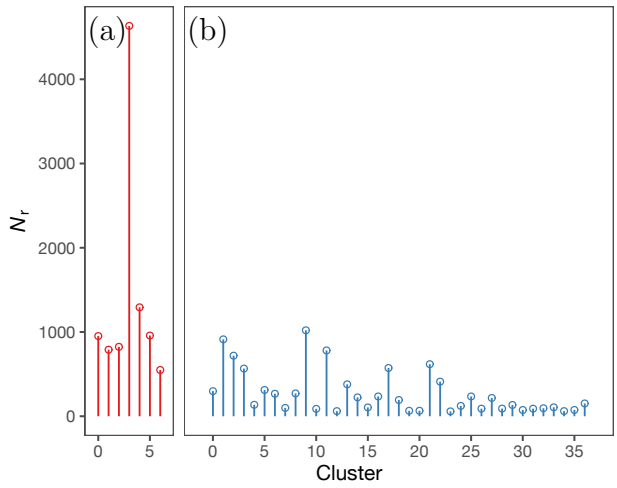


Figure 9: Counts of quasireaction subgraphs N_r per cluster in clustering $N_\mathrm{sample} = 10{,}000$ quasireaction subgraphs using (a) WL-S and (b) WL-E graph kernels.

their energetic landscapes.[114] For a simple example, consider again Fig. 3. Water addition to unsubstituted alkenes follows the electrophilic bimolecular addition ($\mathrm{Ad_E2}$) mechanism, in which the proton is first added to the double bond, followed by the formation of the C−O bond,[115,116] in line with the paths 1 and 4 in Fig. 3. On the other hand, the addition of water to acrolein has an isomorphic reaction subgraph (shown in Fig. S4 of the SI) but proceeds preferentially by the equivalents of the paths 2, 3, and 5.[117–119] These distinctions can be encoded empirically in reaction templates by fine-tuning the matched substructures. However, the approach offered by reaction subgraphs is more direct and better compatible with quantum chemical calculations.

4.2 Concerted Reactions

The decomposition into individual bond breaks and bond formations applies to both stepwise and concerted mechanisms.[32] Each transfer of an electron pair can be formally written as a combination of a polar bond break (bond electron pair becomes the lone pair of the electronegative bond partner) and a polar bond formation (lone pair becomes a bond electron pair). Note that the formal polarization and depolarization rule allow to extend this decomposition approach to reactions of double and triple bonds. For concerted reactions, in

which multiple electron pairs are transferred, we can perform this formal decomposition for each electron pair transfer. As an example, we show the reaction subgraph of the Diels–Alder reaction between butadiene and ethene in Fig. 10(a). For compactness, only one reaction path is shown explicitly. The path consists of three polarizations of double bonds and of three bond formations: two single and one double CC bond. The charge-separated structures are clearly high-energy species and do not occur as identifiable intermediates of the reaction mechanism. However, the true reaction path of these concerted reactions can be understood as a spatial average of the different discrete reaction paths. In this case, the bond order changes per step interpolate between those of the discrete reaction paths and should be capable of expressing different types of reaction mechanisms. The idea of the concerted reaction path as an interpolation between stepwise reaction paths is due to Jencks, who referred to the two-dimensional case as enforced concertedness,[114,120] see also earlier work by Critchlow.[121]

An additional refinement of the reaction subgraph representation is to attach non-negative weights $w_i, i = 1, \ldots, N_{\text{edges}}$ to the subgraph edges based on kinetic feasibility instead of the discrete rule type, as we did in Sec. 3.3. This approach can be applied to modeling both concerted and stepwise reaction mechanisms. In Ref. 32, we defined the heuristic kinetic feasibility (arc and karc) for reaction paths in TNs, which depended on the energies of the nodes along the reaction path. The karc criterion was successful in binary classification tasks of reaction paths as kinetically feasible or infeasible on a set of polar and pericyclic organic reactions. The appropriate weighting by kinetic feasibility allows to distinguish topologically isomorphic reaction subgraphs, for example, those in Fig. 3 and Fig. S4 of the SI. In combination with the reaction path interpolation discussed above, it can represent mechanistic continua that often encompass both stepwise and concerted mechanisms.[114,120,122] The weighting of the alternative paths, such as paths 1–5 in Fig. 3 can determine the preferred paths. As we showed out in Ref. 32, the preferred paths may not be shortest paths.

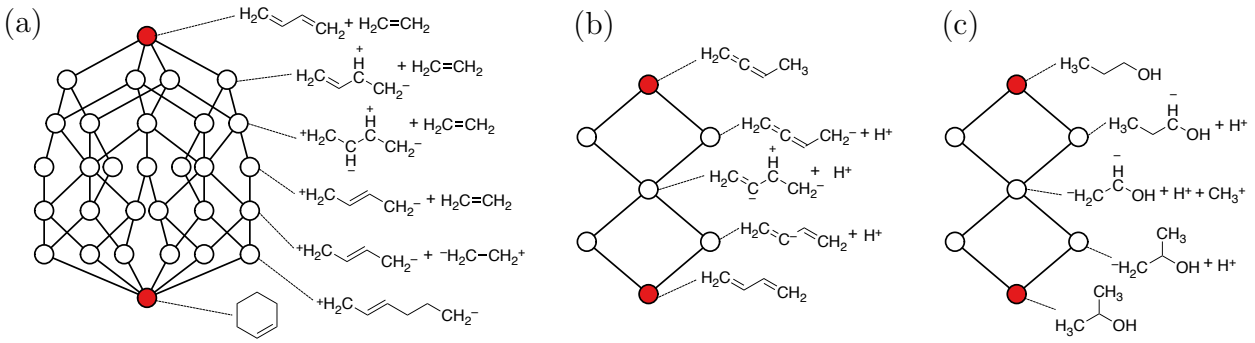


Figure 10: (a) Reaction subgraph of the Diels–Alder reaction between 1,3-butadiene and ethene ($L = 6$, $N_{\text{nodes}} = 30$, $N_{\text{edges}} = 55$); (b) Reaction subgraph of the 1,2-butadiene to 1,3-butadiene isomerization (bond breaks: C=C, C−H); (c) Non-reactive subgraph for a hypothetical isomerization of 1-propanol to 2-propanol via methyl shift (bond breaks: C−C, C−H).

4.3 Binary Classification of Reactive and Non-Reactive Subgraphs

In this work, a complete set of quasireaction subgraphs was constructed and characterized for CHO chemistry with up to six non-hydrogen atoms. This ground set would change if (choices made in this work are in parentheses) we expand the set of elements (CHO), increase the maximum number of non-hydrogen atoms ($\nu_{\text{C}} + \nu_{\text{O}} \leq 6$), expand the set of allowed bond breaks and bond formations (polar, normal electronegativity, see Table S1 of the SI), apply a different or no filtering rules for high-energy species (see Table S2 of the SI), or choose different maximal shortest-path length cutoff ($L \leq 8$) or slack length ($s = 0$). As these choices are not chemistry-specific (with the exception of the rules for identifying high-energy species), this set (or any other ground set generated by the other choices) is free from selection bias by design. The flip side of this (for the most part) chemistry-free generation algorithm is that we need to distinguish quasireaction subgraphs that correspond to thermodynamically and kinetically feasible reaction mechanisms (reactive subgraphs, positive class) and infeasible transformations (non-reactive subgraphs, negative class). The specific examples of the quasireaction subgraphs we had considered so far were from the positive class. Here, we examine two quasireaction subgraphs belonging to the same 7,8 A pattern, shown in Fig. 10(b) and (c). The isomerization of 1,2-butadiene to the

more stable 1,3-butadiene (Fig. 10(b)) involves the breaking (polarization) of the C=C bond and of the C−H bond. This reaction is known in the literature, however, in presence of catalysts.[123] We tentatively label this subgraph as reactive (member of the positive class). The topologically equivalent hypothetical isomerization of 1-propanol to 2-propanol via a methyl shift (Fig. 10(c)) would require a breaking of an unactivated C−C bond and of the C−H bond. This process does not appear kinetically feasible, and 1-propanol is likely to undergo dehydration instead. Therefore, we label this subgraph as non-reactive (belonging to the negative class).

In the above example, one could take advantage of empirical rules to conclude that breaking unactivated C−C bonds is energetically unfavorable and thus assign the negative label the quasireaction subgraph in Fig. 10(c). However, a more robust strategy would use quantum chemical calculations of thermodynamic and kinetic feasibility. The positive class could be defined algorithmically by the presence of a sufficiently low-energy TS along the reaction path connecting the structures of the reactants and of the products. Therefore, double-sided search methods such as variants of the nudged elastic band (NEB),[124–126] double-sided string methods,[127] and geodesic interpolation[128] methods are especially attractive. However, the topology of the quasireaction subgraph can also be utilized for TS searches. The energy profiles of the discretized reaction paths are consistent with Hammond's postulate,[129] which states that the TS, if it exists, is close in structure to the that of the high-energy intermediates. This suggests that the interpolation of the structures of the highest-energy internal nodes is a good starting structure for initializing TS searches. The instances, in which the TS search fails to converge or converges to a TS for a competing reaction should be labeled as members of the negative class. Similarly, the computed TSs with high barrier heights could be treated as negative instances. The availability of negative as well as positive instances may be of value for training robust machine learning models.

5 Conclusions

In this paper, we introduced quasireaction subgraphs, a discrete feature representation for reaction mechanisms that possesses a principled sampling approach and allows for a definition of similarity. These properties make the subgraph representation well suited for constructing data sets. This is because, for a given maximum stoichiometry and a rule set, a finite ground set of quasireaction subgraphs exists, from which a bias-free data set can be constructed, for example, by sampling or by maximum-diversity selection. The price of this construction, however, is that not all quasireaction subgraphs correspond to reactive transformations. Therefore, the sampling or diversity-oriented selection procedure must be combined with a binary classification to distinguish reaction subgraphs (positive class) from non-reactive subgraphs (negative class). In this paper, we focused on the definition, the generation, and the properties of quasireaction subgraphs. We will address the issue of the binary classification in a forthcoming paper.

A certain measure of selection bias may even be desired in data sets for some applications. For example, synthetically relevant reactions are collected and categorized in large synthetic databases. Apart from multiple commercial databases, the publicly available data set of reactions extracted from patent data is widely used for reaction predictions.[130] However, still larger public databases of chemical reactions are needed.[131] These databases include predominantly high-yield transformations and often have a bias toward drug-like molecules. This bias is entirely desirable for applications in synthesis but might be too narrow for exploring, for example, origins of life.[132] On the other hand, data sets designed to minimize selection bias, such as this work, are liable to contain more unknown or unfavorable reactions, in which synthetic chemists have no interest. They may also contain new, yet to be discovered reactions. However, for benchmarking quantum chemical methods or machine learning models of general applicability, selection bias is problematic as it typically improves the performance in one segment of the reaction space at the expense of others. In contrast, bias-free data sets help to achieve more uniform performance of the computational methods

and thus improve their predictive power.

Supporting Information Available

Reaction rules; high-energy species patterns; statistics of CHO transition networks; shortest path length statistics in CHO transition networks; silhouette scores of reaction subgraph clustering; comparison of Dugundji–Ugi reaction matrices and reaction subgraphs.

Acknowledgement

This work was supported in part by the National Science Foundation under Grant No. CHE-2227112.

References

(1) Verma, P.; Wang, Y.; Ghosh, S.; He, X.; Truhlar, D. G. Revised M11 Exchange-Correlation Functional for Electronic Excitation Energies and Ground-State Properties. *J. Phys. Chem. A* **2019**, *123*, 2966–2990.

(2) Goerigk, L.; Hansen, A.; Bauer, C.; Ehrlich, S.; Najibi, A.; Grimme, S. A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions. *Phys. Chem. Chem. Phys.* **2017**, *19*, 32184–32215.

(3) Mardirossian, N.; Head-Gordon, M. Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. *Mol. Phys.* **2017**, *115*, 1–58.

(4) Morgante, P.; Peverati, R. ACCDB: A collection of chemistry databases for broad computational purposes. *J. Comput. Chem.* **2019**, *40*, 839–848.

(5) Prasad, V. K.; Pei, Z.; Edelmann, S.; Otero-de-la Roza, A.; DiLabio, G. A. BH9, a New Comprehensive Benchmark Data Set for Barrier Heights and Reaction Energies: Assessment of Density Functional Approximations and Basis Set Incompleteness Potentials. *J. Chem. Theory Comput.* **2022**, *18*, 151–166.

(6) Prasad, V. K.; Pei, Z.; Edelmann, S.; Otero-de-la Roza, A.; DiLabio, G. A. Correction to "BH9, a New Comprehensive Benchmark Data Set for Barrier Heights and Reaction Energies: Assessment of Density Functional Approximations and Basis Set Incompleteness Potentials". *J. Chem. Theory Comput.* **2022**, *18*, 4041–4044.

(7) Peverati, R.; Truhlar, D. G. Quest for a universal density functional: the accuracy of density functionals across a broad spectrum of databases in chemistry and physics. *Philos. Trans. R. Soc., A* **2014**, *372*, 20120476.

(8) Wang, Y.; Verma, P.; Jin, X.; Truhlar, D. G.; He, X. Revised M06 density functional for main-group and transition-metal chemistry. *Proc. Nat. Acad. Sci. U.S.A.* **2018**, *115*, 10257–10262.

(9) Kirkpatrick, J.; McMorrow, B.; Turban, D. H. P.; Gaunt, A. L.; Spencer, J. S.; Matthews, A. G. D. G.; Obika, A.; Thiry, L.; Fortunato, M.; Pfau, D. et al. Pushing the frontiers of density functionals by solving the fractional electron problem. *Science* **2021**, *374*, 1385–1389.

(10) Liu, Y.; Zhang, C.; Liu, Z.; Truhlar, D. G.; Wang, Y.; He, X. Supervised learning of a chemistry functional with damped dispersion. *Nature Comput. Sci.* **2023**, *3*, 48–58.

(11) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. Multi-coefficient extrapolated density functional theory for thermochemistry and thermochemical kinetics. *Phys. Chem. Chem. Phys.* **2005**, *7*, 43–52.

(12) Zheng, J.; Zhao, Y.; Truhlar, D. G. The DBH24/08 Database and Its Use to Assess

Electronic Structure Model Chemistries for Chemical Reaction Barrier Heights. *J. Chem. Theory Comput.* **2009**, *5*, 808–821.

(13) Karton, A.; O'Reilly, R. J.; Radom, L. Assessment of Theoretical Procedures for Calculating Barrier Heights for a Diverse Set of Water-Catalyzed Proton-Transfer Reactions. *J. Phys. Chem. A* **2012**, *116*, 4211–4221.

(14) Mangiatordi, G. F.; Brémond, E.; Adamo, C. DFT and Proton Transfer Reactions: A Benchmark Study on Structure and Kinetics. *J. Chem. Theory Comput.* **2012**, *8*, 3082–3088.

(15) Zhao, Y.; González-García, N.; Truhlar, D. G. Benchmark Database of Barrier Heights for Heavy Atom Transfer, Nucleophilic Substitution, Association, and Unimolecular Reactions and Its Use to Test Theoretical Methods. *J. Phys. Chem. A* **2005**, *109*, 2012–2018.

(16) Guner, V.; Khuong, K. S.; Leach, A. G.; Lee, P. S.; Bartberger, M. D.; Houk, K. N. A Standard Set of Pericyclic Reactions of Hydrocarbons for the Benchmarking of Computational Methods: The Performance of ab Initio, Density Functional, CASSCF, CASPT2, and CBS-QB3 Methods for the Prediction of Activation Barriers, Reaction Energetics, and Transition State Geometries. *J. Phys. Chem. A* **2003**, *107*, 11445–11459.

(17) Ess, D. H.; Houk, K. N. Activation Energies of Pericyclic Reactions: Performance of DFT, MP2, and CBS-QB3 Methods for the Prediction of Activation Barriers and Reaction Energetics of 1,3-Dipolar Cycloadditions, and Revised Activation Enthalpies for a Standard Set of Hydrocarbon Pericyclic Reactions. *J. Phys. Chem. A* **2005**, *109*, 9542–9553.

(18) Dinadayalane, T. C.; Vijaya, R.; Smitha, A.; Sastry, G. N. Diels–Alder Reactivity of

Butadiene and Cyclic Five-Membered Dienes ((CH)4X, X = CH2, SiH2, O, NH, PH, and S) with Ethylene: A Benchmark Study. *J. Phys. Chem. A* **2002**, *106*, 1627–1633.

(19) Goerigk, L.; Grimme, S. A General Database for Main Group Thermochemistry, Kinetics, and Noncovalent Interactions—Assessment of Common and Reparameterized (meta-)GGA Density Functionals. *J. Chem. Theory Comput.* **2010**, *6*, 107–126.

(20) Karton, A.; Goerigk, L. Accurate reaction barrier heights of pericyclic reactions: Surprisingly large deviations for the CBS-QB3 composite method and their consequences in DFT benchmark studies. *J. Comput. Chem.* **2015**, *36*, 622–632.

(21) Yu, L.-J.; Sarrami, F.; O'Reilly, R. J.; Karton, A. Reaction barrier heights for cycloreversion of heterocyclic rings: An Achilles' heel for DFT and standard ab initio procedures. *Chem. Phys.* **2015**, *458*, 1–8.

(22) von Rudorff, G. F.; Heinen, S. N.; Bragato, M.; von Lilienfeld, O. A. Thousands of reactants and transition states for competing E2 and S_N2 reactions. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045026.

(23) Wappett, D. A.; Goerigk, L. A guide to benchmarking enzymatically catalysed reactions: the importance of accurate reference energies and the chemical environment. *Theor. Chem. Acc.* **2021**, *140*, 68.

(24) Stocker, S.; Csányi, G.; Reuter, K.; Margraf, J. T. Machine learning in chemical reaction space. *Nature Commun.* **2020**, *11*, 5505.

(25) Erkut, E. The discrete *p*-dispersion problem. *Eur. J. Oper. Res.* **1990**, *46*, 48–60.

(26) Gorse, A.-D. Diversity in Medicinal Chemistry Space. *Curr. Top. Med. Chem.* **2006**, *6*, 3–18.

(27) Huggins, D. J.; Venkitaraman, A. R.; Spring, D. R. Rational Methods for the Selection of Diverse Screening Compounds. *ACS Chem. Bio.* **2011**, *6*, 208–217.

(28) Willett, P. The Calculation of Molecular Structural Similarity: Principles and Practice. *Mol. Inf.* **2014**, *33*, 403–413.

(29) Koutsoukas, A.; Paricharak, S.; Galloway, W. R. J. D.; Spring, D. R.; IJzerman, A. P.; Glen, R. C.; Marcus, D.; Bender, A. How Diverse Are Diversity Assessment Methods? A Comparative Analysis and Benchmarking of Molecular Descriptor Space. *J. Chem. Inf. Model.* **2014**, *54*, 230–242.

(30) Dugundji, J.; Ugi, I. *Computers in Chemistry*; Springer: Berlin, Heidelberg, 1973; pp 19–64.

(31) Rappoport, D. Reaction Networks and the Metric Structure of Chemical Space(s). *J. Phys. Chem. A* **2019**, *123*, 2610–2620.

(32) Rappoport, D.; Aspuru-Guzik, A. Predicting Feasible Organic Reaction Pathways Using Heuristically Aided Quantum Chemistry. *J. Chem. Theory Comput.* **2019**, *15*, 4099–4112.

(33) Shervashidze, N.; Schweitzer, P.; van Leeuwen, E. J.; Mehlhorn, K.; Borgwardt, K. M. Weisfeiler–Lehman Graph Kernels. *J. Mach. Learn. Res.* **2011**, *12*, 2539–2561.

(34) Sugiyama, M.; Borgwardt, K. Halting in Random Walk Kernels. Advances in Neural Information Processing Systems 28. La Jolla CA, 2015; pp 1639–1647.

(35) Patchkovskii, S.; Ziegler, T. Improving "difficult" reaction barriers with self-interaction corrected density functional theory. *J. Chem. Phys.* **2002**, *116*, 7806–7813.

(36) Gould, T.; Dale, S. G. Poisoning density functional theory with benchmark sets of difficult systems. *Phys. Chem. Chem. Phys.* **2022**, *24*, 6398–6403.

(37) Zheng, J.; Zhao, Y.; Truhlar, D. G. Representative Benchmark Suites for Barrier Heights of Diverse Reaction Types and Assessment of Electronic Structure Methods for Thermochemical Kinetics. *J. Chem. Theory Comput.* **2007**, *3*, 569–582.

(38) Gould, T. 'Diet GMTKN55' offers accelerated benchmarking through a representative subset approach. *Phys. Chem. Chem. Phys.* **2018**, *20*, 27735–27739.

(39) Swann, E. T.; Fernandez, M.; Coote, M. L.; Barnard, A. S. Bias-Free Chemically Diverse Test Sets from Machine Learning. *ACS Comb. Sci.* **2017**, *19*, 544–554.

(40) Morgante, P.; Peverati, R. Statistically representative databases for density functional theory via data science. *Phys. Chem. Chem. Phys.* **2019**, *21*, 19092–19103.

(41) Chan, B. Formulation of Small Test Sets Using Large Test Sets for Efficient Assessment of Quantum Chemistry Methods. *J. Chem. Theory Comput.* **2018**, *14*, 4254–4262.

(42) Grambow, C. A.; Pattanaik, L.; Green, W. H. Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry. *Sci. Data* **2020**, *7*, 137.

(43) Spiekermann, K.; Pattanaik, L.; Green, W. H. High accuracy barrier heights, enthalpies, and rate coefficients for chemical reactions. *Sci. Data* **2022**, *9*, 417.

(44) Schreiner, M.; Bhowmik, A.; Vegge, T.; Busk, J.; Winther, O. Transition1x - a dataset for building generalizable reactive machine learning potentials. *Sci. Data* **2022**, *9*, 779.

(45) Grambow, C. A.; Pattanaik, L.; Green, W. H. Deep Learning of Activation Energies. *J. Phys. Chem. Lett.* **2020**, *11*, 2992–2997.

(46) Schreiner, M.; Bhowmik, A.; Vegge, T.; Jørgensen, P. B.; Winther, O. NeuralNEB—neural networks can find reaction paths fast. *Mach. Learn.: Sci. Technol.* **2022**, *3*, 045022.

(47) Fink, T.; Bruggesser, H.; Reymond, J.-L. Virtual exploration of the small-molecule chemical universe below 160 Daltons. *Angew. Chem. Int. Ed.* **2005**, *44*, 1504–1508.

(48) Fink, T.; Reymond, J.-L. Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and

analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *J. Chem. Inf. Model.* **2007**, *47*, 342–353.

(49) Blum, L. C.; Reymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732–8733.

(50) Blum, L. C.; Deursen, R. v.; Reymond, J.-L. Visualisation and subsets of the chemical universe database GDB-13 for virtual screening. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 637–647.

(51) Ruddigkeit, L.; Deursen, R. v.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875.

(52) Reymond, J.; Ruddigkeit, L.; Blum, L.; Deursen, R. v. The enumeration of chemical space. *WIREs Comput. Mol. Sci.* **2012**, *2*, 717–733.

(53) Reymond, J.-L. The Chemical Space Project. *Acc. Chem. Res.* **2015**, *48*, 722–730.

(54) McKay, B. D.; Piperno, A. Practical graph isomorphism, II. *J. Symb. Comput.* **2014**, *60*, 94–112.

(55) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.

(56) Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K.-R.; Lilienfeld, O. A. v. Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **2013**, *15*, 095003.

(57) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*, 140022.

(58) Kim, H.; Park, J. Y.; Choi, S. Energy refinement and analysis of structures in the QM9 database via a highly accurate quantum chemical method. *Sci. Data* **2019**, *6*, 109.

(59) Hoja, J.; Medrano Sandonas, L.; Ernst, B. G.; Vazquez-Mayagoitia, A.; DiStasio Jr., R. A.; Tkatchenko, A. QM7-X, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules. *Sci. Data* **2021**, *8*, 43.

(60) Chen, G.; Chen, P.; Hsieh, C.-Y.; Lee, C.-K.; Liao, B.; Liao, R.; Liu, W.; Qiu, J.; Sun, Q.; Tang, J. et al. Alchemy: A Quantum Chemistry Dataset for Benchmarking AI Models. 2019; arXiv:1906.09427 [cs.LG], `https://doi.org/10.48550/arXiv.1906.09427`.

(61) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1, A data set of 20 million calculated off-equilibrium conformations for organic molecules. *Sci. Data* **2017**, *4*, 170193.

(62) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **2017**, *8*, 3192–3203.

(63) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less is more: Sampling chemical space with active learning. *J. Chem. Phys.* **2018**, *148*, 241733.

(64) Smith, B. T., Justin S.and Nebgen; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nature Commun.* **2019**, *10*, 2903.

(65) Smith, J. S.; Zubatyuk, R.; Nebgen, B.; Lubbers, N.; Barros, K.; Roitberg, A. E.; Isayev, O.; Tretiak, S. The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. *Sci. Data* **2020**, *7*, 134.

(66) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **2009**, *37*, W623–W633.

(67) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B. et al. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* **2018**, *47*, D1102–D1109.

(68) Nakata, M.; Shimazaki, T. PubChemQC Project: A Large-Scale First-Principles Electronic Structure Database for Data-Driven Chemistry. *J. Chem. Inf. Model.* **2017**, *57*, 1300–1308.

(69) Glavatskikh, M.; Leguy, J.; Hunault, G.; Cauchy, T.; Da Mota, B. Dataset's chemical diversity limits the generalizability of machine learning predictions. *J. Cheminf.* **2019**, *11*, 69.

(70) Vazquez-Salazar, L. I.; Boittier, E. D.; Unke, O. T.; Meuwly, M. Impact of the Characteristics of Quantum Chemical Databases on Machine Learning Prediction of Tautomerization Energies. *J. Chem. Theory Comput.* **2021**, *17*, 4769–4785.

(71) Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway, A. M.; Aspuru-Guzik, A. The Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid. *J. Phys. Chem. Lett.* **2011**, *2*, 2241–2251.

(72) Lopez, S. A.; Pyzer-Knapp, E. O.; Simm, G. N.; Lutzow, T.; Li, K.; Seress, L. R.; Hachmann, J.; Aspuru-Guzik, A. The Harvard organic photovoltaic dataset. *Sci. Data* **2016**, *3*, 160086.

(73) Virshup, A. M.; Contreras-García, J.; Wipf, P.; Yang, W.; Beratan, D. N. Stochas-

tic Voyages into Uncharted Chemical Space Produce a Representative Library of All Possible Drug-Like Compounds. *J. Am. Chem. Soc.* **2013**, *135*, 7296–7303.

(74) Korth, M.; Grimme, S. "Mindless" DFT Benchmarking. *J. Chem. Theory Comput.* **2009**, *5*, 993–1003.

(75) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nature Commun.* **2017**, *8*, 13890.

(76) Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R. Machine learning of accurate energy-conserving molecular force fields. *Science Adv.* **2017**, *3*, e1603015.

(77) Hatzimanikatis, V.; Li, C.; Ionita, J. A.; Henry, C. S.; Jankowski, M. D.; Broadbelt, L. J. Exploring the diversity of complex metabolic networks. *Bioinformatics* **2005**, *21*, 1603–1609.

(78) Gao, C. W.; Allen, J. W.; Green, W. H.; West, R. H. Reaction Mechanism Generator: Automatic construction of chemical kinetic mechanisms. *Comput. Phys. Commun.* **2016**, *203*, 212–225.

(79) Unsleber, J. P.; Reiher, M. The Exploration of Chemical Reaction Networks. *Annu. Rev. Phys. Chem.* **2020**, *71*, 1–22.

(80) Wen, M.; Spotte-Smith, E. W. C.; Blau, S. M.; McDermott, M. J.; Krishnapriyan, A. S.; Persson, K. A. Chemical reaction networks and opportunities for machine learning. *Nature Comput. Sci.* **2023**, 1–13.

(81) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **1988**, *28*, 31–36.

(82) SMARTS – A Language for Describing Molecular Patterns. `https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html` (accessed 2023-02-28).

(83) Landrum, G., et al. RDKit: Open-Source Cheminformatics Software; version 2021.09.2. `https://zenodo.org/record/7541264` (accessed 2023-02-28).

(84) Rappoport, D. colibri2 is a distributed computational engine for chemical exploration. `https://bitbucket.org/rappoport/colibri2/` (accessed 2023-02-28).

(85) Brandes, U.; Eiglsperger, M.; Herman, I.; Himsolt, M.; Marshall, M. S. GraphML Progress Report Structural Layer Proposal. Graph Drawing. Berlin, Heidelberg, 2002; pp 501–512.

(86) Wilzbach, K. E.; Ritscher, J. S.; Kaplan, L. Benzvalene, the Tricyclic Valence Isomer of Benzene. *J. Am. Chem. Soc.* **1967**, *89*, 1031–1032.

(87) Katz, T. J.; Wang, E. J.; Acton, N. Benzvalene synthesis. *J. Am. Chem. Soc.* **1971**, *93*, 3782–3783.

(88) Katz, T. J.; Acton, N. Synthesis of prismane. *J. Am. Chem. Soc.* **1973**, *95*, 2738–2739.

(89) Jochum, C.; Gasteiger, J.; Ugi, I. The Principle of Minimum Chemical Distance (PMCD). *Angew. Chem. Int. Ed.* **1980**, *19*, 495–505.

(90) Hagberg, A. A.; Schult, D. A.; Swart, P. J. Exploring network structure, dynamics, and function using NetworkX. Proceedings of the 7th Python in Science Conference (SciPy2008). 2008; pp 11–15.

(91) Ghosh, S.; Das, N.; Gonçalves, T.; Quaresma, P.; Kundu, M. The journey of graph kernels through two decades. *Comput. Sci. Rev.* **2018**, *27*, 88–111.

(92) Kriege, N. M.; Johansson, F. D.; Morris, C. A survey on graph kernels. *Appl. Netw. Sci.* **2020**, *5*, 6.

(93) Nikolentzos, G.; Siglidis, G.; Vazirgiannis, M. Graph Kernels: A Survey. *JAIR* **2021**, *72*, 943–1027.

(94) Siglidis, G.; Nikolentzos, G.; Limnios, S.; Giatsidis, C.; Skianis, K.; Vazirgiannis, M. GraKeL: A Graph Kernel Library in Python. *J. Mach. Learn. Res.* **2020**, *21*, 1–5.

(95) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. **2018**, arXiv: 1802.03426 [stat.ML], `https://doi.org/10.48550/arXiv.1802.03426`.

(96) MacQueen, J. Classification and analysis of multivariate observations. 5th Berkeley Symp. Math. Statist. Probability. Berkeley and Los Angeles, 1967; pp 281–297.

(97) Lloyd, S. Least squares quantization in PCM. *IEEE Trans. Inf. Theor.* **1982**, *28*, 129–137.

(98) Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65.

(99) UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. `htttps://https://umap-learn.readthedocs.io/` (accessed 2023-02-28).

(100) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(101) American Chemical Society, CAS SciFinder. `https://scifinder-n.cas.org` (accessed 2023-02-28).

(102) Palazzo, T. A.; Mose, R.; Jørgensen, K. A. Cycloaddition Reactions: Why Is It So Challenging To Move from Six to Ten Electrons? *Angew. Chem. Int. Ed.* **2017**, *56*, 10033–10038.

(103) McLeod, D.; Thøgersen, M. K.; Jessen, N. I.; Jørgensen, K. A.; Jamieson, C. S.; Xue, X.-S.; Houk, K. N.; Liu, F.; Hoffmann, R. Expanding the Frontiers of Higher-Order Cycloadditions. *Acc. Chem. Res.* **2019**, *52*, 3488–3501.

(104) Jessen, N. I.; McLeod, D.; Jørgensen, K. A. Higher-order cycloadditions in the age of catalysis. *Chem* **2022**, *8*, 20–30.

(105) Rappoport, D. Discrete Feature Representations of CHO Reaction Mechanisms as Quasireaction Subgraphs. `https://doi.org/10.5281/zenodo.7905294` (accessed 2023-05-07).

(106) Dugundji, J.; Gillespie, P.; Marquarding, D.; Ugi, I.; Ramirez, F. In *Chemical Applications of Graph Theory*; Balaban, A. T., Ed.; Academic Press: New York, 1976; pp 108–174.

(107) Ugi, I.; Stein, N.; Knauer, M.; Gruber, B.; Bley, K.; Weidinger, R. New Elements in the Representation of the Logical Structure of Chemistry by Qualitative Mathematical Models and Corresponding Data Structures. *Top. Curr. Chem.* **1993**, *166*, 199–233.

(108) Chen, J. H.; Baldi, P. No Electron Left Behind: A Rule-Based Expert System To Predict Chemical Reactions and Reaction Mechanisms. *J. Chem. Inf. Model.* **2009**, *49*, 2034–2043.

(109) Kraut, H.; Eiblmaier, J.; Grethe, G.; Löw, P.; Matuszczyk, H.; Saller, H. Algorithm for Reaction Classification. *J. Chem. Inf. Model.* **2013**, *53*, 2884–2895.

(110) Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew. Chem. Int. Ed.* **2016**, *55*, 5904–5937.

(111) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **2017**, *3*, 434–443.

(112) Fujita, S. Description of organic reactions based on imaginary transition structures. 1. Introduction of new concepts. *J. Chem. Inf. Model.* **1986**, *26*, 205–212.

(113) Fujita, S. Graphic characterization and taxonomy of organic reactions. *J. Chem. Educ.* **1990**, *67*, 290–293.

(114) Jencks, W. P. Ingold Lecture. How does a reaction choose its mechanism? *Chem. Soc. Rev.* **1981**, *10*, 345–375.

(115) de la Mare, P. B. D., Bolton, R., Eds. *Electrophilic Additions to Unsaturated Systems*; Elsevier: Amsterdam, 1982.

(116) Schmid, G. H.; Garratt, D. G. In *Supplement A. The chemistry of double-bonded functional groups. Part 2*; Patai, S., Ed.; Wiley: London, 1977; Chapter 9, pp 725–912.

(117) Patai, S.; Rappoport, Z. In *The chemistry of alkenes*; Patai, S., Ed.; Interscience: London, 1964; Chapter 8, pp 469–584.

(118) Jin, J.; Hanefeld, U. The selective addition of water to C=C bonds; enzymes are the best chemists. *Chem. Commun.* **2011**, *47*, 2502–2510.

(119) Resch, V.; Hanefeld, U. The selective addition of water. *Catal. Sci. Technol.* **2015**, *5*, 1385–1399.

(120) Jencks, W. P. When is an intermediate not an intermediate? Enforced mechanisms of general acid-base, catalyzed, carbocation, carbanion, and ligand exchange reaction. *Acc. Chem. Res.* **1980**, *13*, 161–169.

(121) Critchlow, J. E. Prediction of transition state configuration in concerted reactions from the energy requirements of the separate processes. *J. Chem. Soc., Faraday Trans. 1* **1972**, *68*, 1774–1792.

(122) More O'Ferrall, R. A. Relationships between $E2$ and $E1cB$ mechanisms of β-elimination. *J. Chem. Soc. B* **1970**, 274–277.

(123) Slobodin, Y. M. *Zh. Ob. Khim.* **1935**, *5*, 48–52.

(124) Henkelman, G.; Jóhannesson, G.; Jónsson, H. In *Theoretical Methods in Condensed Phase Chemistry*; Schwartz, S. D., Ed.; Kluwer, 2002; Vol. 5; pp 269–302.

(125) Schlegel, H. B. Geometry optimization. *WIREs Comput. Mol. Sci.* **2011**, *1*, 790–809.

(126) Dewyer, A. L.; Argüelles, A. J.; Zimmerman, P. M. Methods for exploring reaction space in molecular systems. *WIREs Comput. Mol. Sci.* **2018**, *8*, e1354.

(127) Peters, B.; Heyden, A.; Bell, A. T.; Chakraborty, A. A growing string method for determining transition states: Comparison to the nudged elastic band and string methods. *J. Chem. Phys.* **2004**, *120*, 7877–7886.

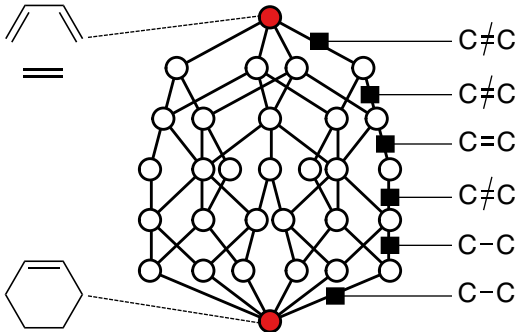
(128) Zhu, X.; Thompson, K. C.; Martínez, T. J. Geodesic interpolation for reaction pathways. *J. Chem. Phys.* **2019**, *150*, 164103.

(129) Hammond, G. S. A Correlation of Reaction Rates. *J. Am. Chem. Soc.* **1955**, *77*, 334–338.

(130) Lowe, D. Chemical reactions from US patents (1976–Sep2016). 2017; `https://doi.org/10.6084/m9.figshare.5104873.v1` (accessed 2023-02-28).

(131) Baldi, P. Call for a Public Open Database of All Chemical Reactions. *J. Chem. Inf. Model.* **2022**, *62*, 2011–2014.

(132) Wołos, A.; Roszak, R.; Żądło-Dobrowolska, A.; Beker, W.; Mikulak-Klucznik, B.; Spólnik, G.; Dygas, M.; Szymkuć, S.; Grzybowski, B. A. Synthetic connectivity, emergence, and self-regeneration in the network of prebiotic chemistry. *Science* **2020**, *369*, eaaw1955.



TOC Graphic