Chem Soc Rev

REVIEW ARTICLE

View Article Online
View Journal

Cite this: DOI: 10.1039/d6cs00079g

An overview of reaction outcome prediction with physics-based and data-driven methods

Joonyoung F. Joung, *[ab] Nicholas Casetti,[a] Priyanka Raghavan[a] and Connor W. Coley *[ac]

The prediction of reaction outcomes is a longstanding challenge in chemistry, with the ability to do so serving as a direct reflection of our understanding of chemical reactivity. Accurately predicting reaction products is crucial not only for synthetic planning but also for designing reaction pathways and experiments *in silico*. This review explores the diverse methodologies used to predict reaction outcomes, which can be broadly divided into two main categories. Some approaches predict reaction products and their likelihoods in a single step, while others break the task into two distinct parts: candidate enumeration and the subsequent prediction of product likelihoods. We examine both data-driven methods, such as graph-based and sequence-generation models, and physics-based methods, including potential energy surface exploration and reactive molecular dynamics. In addition, we discuss quantitative predictions of reaction selectivity, regioselectivity, stereoselectivity, and yield. This review summarizes trends and advances in reaction outcome prediction and briefly outlines future directions for the field.

Received 19th January 2026

DOI: 10.1039/d6cs00079g

rsc.li/chem-soc-rev

1 Introduction

One of the primary goals in chemistry is the precise manipulation or transformation of molecular structures by adding, removing, or rearranging specific components in a controlled and predictable manner, often under highly specific reaction conditions and constraints.[1] Achieving this goal requires a deep understanding of chemical reactivity, which chemists may begin developing during their undergraduate education. Through extensive study of reaction mechanisms, functional group behavior, and molecular interactions, students build a foundational knowledge of how molecules react under various conditions. Over time, this knowledge becomes more refined as one gains hands-on experience in the lab, learning to predict reaction outcomes and develop an intuitive understanding of reactivity patterns.

[a] *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. E-mail: ccoley@mit.edu*
[b] *Department of Chemistry, Kookmin University, 77 Jeongneung-ro, Seongbuk-gu, Seoul, 02707, Republic of Korea. E-mail: jjoung@kookmin.ac.kr*
[c] *Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

Joonyoung F. Joung

Joonyoung F. Joung is an Assistant Professor in the Department of Chemistry at Kookmin University, Republic of Korea. He received his PhD in Physical Chemistry from Korea University and conducted postdoctoral research at MIT. His research focuses on developing machine learning approaches for chemical problems that are difficult to address with conventional methods. In particular, he is interested in incorporating chemical knowledge and physical constraints into data-driven models to improve their reliability and interpretability in reaction and materials chemistry.

Nicholas Casetti

Nicholas Casetti is currently pursuing a PhD in Chemical Engineering at MIT in the Coley Lab. He received his BS in Chemical Engineering from Purdue University in 2022. His research primarily focuses on using machine-learned interatomic potentials for reaction mechanism prediction and elucidation.

However, chemical reactivity is inherently complex. While some transformations can be reliably predicted based on well-established heuristics and rules, accurate prediction of others remains elusive, requiring advanced reasoning and extensive experience.[2] It is not uncommon for experts to encounter reactions where outcomes deviate from expectations, reflecting the unpredictability of many chemical systems.[3–5] As such, computational models of outcome prediction are not only predictive tools but also reflections of the underlying logic of chemical reactivity that may elucidate yet-unknown trends.

The diversity of factors relevant to chemical reactions further complicates the task of predicting outcomes. Reactivity is influenced by electronic effects, steric hindrance, solvent environments, and diverse reaction conditions including photochemistry,[6–8] electrochemistry,[9–11] and high pressure,[12] as well as more ubiquitous factors such as temperature, concentration, and time, which are often not explicitly incorporated in current machine-learning models for reaction prediction. Under these more activating conditions, the intuition that many chemists rely on for conventional thermal reactions can fall short—as can many models. Unlike well-studied counterparts, these environments can involve excited states, electron transfer, or non-equilibrium dynamics, making it much harder to predict outcomes using established methods. The fact that chemical transformations become highly context-dependent motivates the development of new computational and theoretical frameworks to unravel their complex and diverse reactivity.

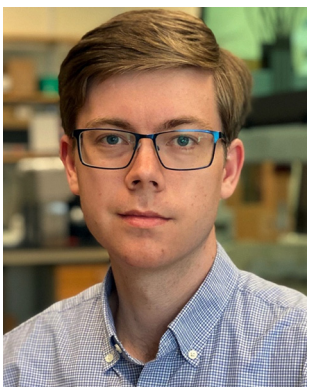
The term reaction outcome prediction can refer to a variety of modeling tasks, each pursuing different task formulations. Some approaches aim to predict the most likely product(s) and their associated likelihoods directly in a single step (Fig. 1a), while others simulate full mechanistic pathways using first-principles methods (Fig. 1b). In contrast, two-step approaches decouple the task of outcome prediction into first enumerating plausible candidate products and then scoring them according to their likelihoods (Fig. 1c). Other common formulations include predicting selectivity (Fig. 1d), or reaction yields under specified conditions (Fig. 1e); less common formulations may seek to predict reaction rates.

These different instantiations of the reaction outcome prediction task place more or less emphasis on the scope of potential transformation products relative to their relative likelihoods. For example, one might consider reaction yield prediction as a task where the set of relevant possible outcomes has been narrowed down to a single product and the challenge is exclusively on scoring in absolute terms; site selectivity prediction as a task where the set of possible outcomes is also relatively unambiguous; major product prediction as a task where the second stage of scoring need not be as quantitative; mechanism generation and simulation as a task where both defining the scope of possible transformations and quantitatively anticipating their rates can prove very challenging.

This review draws on that framing to clarify how different methods—whether rule-based, physics-driven, or data-driven-approach these tasks, and how each method implicitly defines what constitutes a "reaction outcome" in the first place. We pay close attention to the problem formulation, *i.e.*, how different tasks are posed. How are potential products defined? How are they scored? Is this approached as one simultaneous task or in two distinct parts? For most data-driven formulations of the problem, there are only weak distinctions between the settings of predicting "overall" transformations and predicting mechanistic transformations; the greatest differences are observed when considering the role of physics-based methods that use knowledge of the potential energy landscape to propose products, score products, or both. This distinction between one-step and two-step formulations also serves as a central organizational principle for the structure of this review. Rather than grouping methods solely by algorithmic paradigm (*e.g.*, physics-based or data-driven), we organize our discussion according to how each method conceptualizes the task of reaction outcome prediction. Specifically, we examine whether candidate outcomes are generated and scored simultaneously, or treated as distinct stages—an axis along which many of the most meaningful differences in methodology and assumptions emerge.

Priyanka Raghavan

Priyanka Raghavan received her BS in Chemical and Biomolecular Engineering from the University of California, Berkeley in 2020. She then received her PhD from MIT in 2025, working on data-driven methods for chemical reactivity prediction, under the supervision of Professor Connor Coley. She is now a Senior Scientist at AbbVie, building and applying machine learning models to accelerate early-stage drug discovery efforts for beyond-rule of 5 compounds.



Connor W. Coley

Connor W. Coley is an Associate Professor at MIT in the Department of Chemical Engineering and the Department of Electrical Engineering and Computer Science. His research group works at the interface of AI and chemistry to develop models that understand how molecules behave, interact, and react and use that knowledge to engineer new ones, with an emphasis on therapeutic discovery. He received his BS in Chemical Engineering from Caltech and his PhD from MIT, followed by postdoctoral training at the Broad Institute.

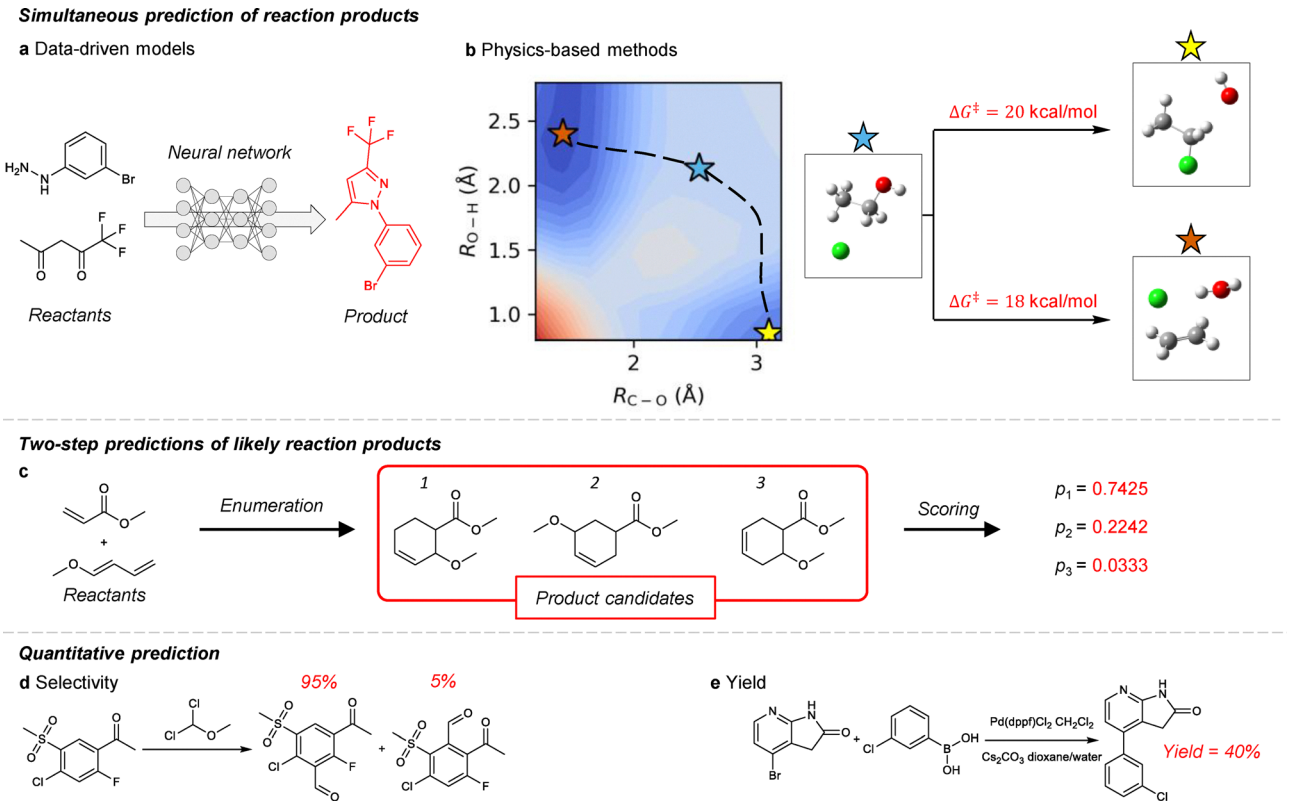


Fig. 1 Representative problem formulations in reaction outcome prediction, categorized by task type. (a) and (b) Simultaneous prediction of reaction products and their likelihoods. (a) A data-driven model takes reactants as input and directly predicts the most likely product(s) in a single step. The example shows an epoxide ring opening with ammonia, where both product identity and likelihood are predicted jointly. (b) A physics-based method explores the potential energy surface (PES) to identify feasible reaction pathways and products. Enumeration and scoring are handled implicitly *via* quantum chemical calculations of intermediates and transition states. (c) Two-step prediction of likely reaction products. The model first enumerates possible outcomes for a Diels–Alder reaction and then ranks them using a separate scoring step. (d) and (e) Quantitative prediction of other aspects of reaction outcomes. (d) In site selectivity prediction, a model assigns probabilities to alternative functionalization sites in an aromatic substrate undergoing C–H activation. Red labels indicate predicted site probabilities. (e) In reaction yield prediction, the model estimates the expected yield of a Suzuki–Miyaura coupling product under specified conditions. The predicted yield (40%) is shown above the product.

We first provide a brief background on the data and physical principles that underpin modern approaches to reaction outcome prediction. We then examine methods that perform joint prediction of outcomes and likelihoods, followed by approaches that decouple the task into separate stages of candidate enumeration and scoring. In the literature, the term "reaction" may refer either to an overall transformation from reactants to products or to an individual mechanistic step conventionally drawn as arrow-pushing diagrams. We then discuss models that aim to quantitatively predict other aspects of reaction outcomes, such as site selectivity and product yield. Finally, we highlight a select number of examples of reaction prediction tools and their application to advancing scientific understanding and discovery.

2 Background

2.1 Data behind data-driven methods

Data-driven methods leverage datasets of observed reaction outcomes (experimental or simulated) and cheminformatic or statistical learning algorithms to understand and formalize relationships between reaction inputs and outputs. Data-driven models are reliant on the quality and diversity of the datasets used for training, which directly inform their performance and generalizability.

2.1.1 Literature-derived experimental datasets. Large datasets that contain various types of chemical reactions are necessary to develop and achieve models with good predictive performance and broad domains of applicability.[13] These datasets minimally include information on reactants and products, sometimes the reaction conditions such as reagents and solvents, and occasionally intermediates or transition states. There are datasets that are publicly available such as the corpus of reactions extracted from the USPTO[14,15] or the Open Reaction Database,[16] readily-available commercial ones like Pistachio,[17] even larger "gold standard" datasets from Chemical Abstracts Service (CAS) or Reaxys[18] curated from the literature, and various proprietary sets originating from electronic lab notebooks of pharmaceutical companies. We summarize some of the most widely used and available experimental datasets in Table 1; many more have emerged in just the past couple of years, but have not yet been adopted as benchmark tasks.

The USPTO-derived reaction dataset is worth special mention due to its prominence as a fully open CC0-licensed dataset

Table 1 Overview of public and proprietary full reaction datasets commonly used as sources of experimental reaction data for model training

Name	Size	Description
USPTO[14,15]	424 621 reactions[a] 1 808 937 reaction[b]	A curated collection of chemical reactions extracted from United States patents, constructed by Lowe[14] through systematic text-mining, normalization, and atom-mapping procedures applied to the USPTO corpus.
USPTO-50k[19]	50 000 reactions	A subset pseudorandomly selected from the USPTO corpus[14,15] from reactions able to be classified or mapped by NameRxn[20] and Indigo.
USPTO-480k (USPTO-MIT)[21]	480 000 reactions	A dataset derived from the USPTO corpus.[14,15] This dataset was filtered to exclude stereochemical information.
USPTO-LEF[22]	349 898 reactions	This dataset consists of reactions from the USPTO-480k, filtered to focus on reactions that involve linear electron flow (LEF).
USPTO-STEREO[23]	1 002 970 reactions	A dataset derived from the USPTO corpus,[14,15] that retains stereochemistry information for reactions. It includes details about chiral centers and their transformations.
USPTO-Full[24]	1 100 105 reactions	A dataset derived from the USPTO corpus[14,15] and filtered by Dai *et al.*,[24] containing a wide variety of reaction types.
Pistachio[17]	Over 9 million reactions	A dataset of over 9 million reactions extracted from US & EPO patents, containing compound information (SMILES, trivial names), reaction details (reaction types, yields), and document info (publication date).
Ahneman *et al.*'s Buchwald–Hartwig amination[25]	4312 reactions	A high-throughput experimental dataset focused on the palladium-catalyzed Buchwald–Hartwig amination of aryl halides with 4-methylaniline, including yield, reaction conditions, and additives.
Perera *et al.*'s Suzuki cross-coupling[26]	5760 reactions	A high-throughput experimental dataset focused on the palladium-catalyzed Suzuki-Miyaura cross-coupling of aryl halides with boronic acids, including yield, reaction conditions, and additives.
HiTEA[27]	3083 Buchwald–Hartwig 2908 hetero hydrogenation 2567 homo hydrogenation 1575 Ullmann coupling	High-throughput experimental datasets collected over a decade from pharmaceutical companies' internal screening platforms, covering four reaction classes with detailed records of yields, conditions, and reagents.

[a] Reactions extracted from USPTO patents (2008–2011). [b] Reactions extracted from USPTO patents (1976–2016).

and the source of many data subsets used for model evaluation. It is built from the patent text-mining work of Daniel M. Lowe and contains data algorithmically extracted from U.S. patents dating from 1976 to September 2016.[14,15] Since the full dataset may contain reactions with missing components, questionable atom mapping, or other potential quality concerns, several versions of the USPTO dataset have been created based on different filtering methods.

The USPTO-15k subset consists of 15 000 curated reactions, originally used by Coley *et al.*[28] for reaction product prediction. The filtering process involved selecting reactions that could be mapped to a common set of reaction templates, with special care taken to eliminate data with unreliable atom mappings or incomplete reaction information. This dataset primarily focuses on reactions with well-defined and easily distinguishable reactants and products.

The USPTO-50k subset is a larger collection of 50 000 reactions able to be assigned a reaction class from the USPTO,[19] offering a sampling of reactions encountered in patent literature. The filtering process here focused on ensuring that only reactions with valid atom mappings and complete reaction schemes were included. This dataset was originally released in the context of reaction classification, but quickly became popular in retrosynthesis studies due to its curation and manageable size. The USPTO-480k dataset (sometimes referred to as USPTO-MIT) represents a larger and more specialized subset, containing 480 000 reactions that were created by Jin *et al.*[21] A key feature of this dataset is the requirement for reactive atoms to form a continuous reaction center, meaning that reactive atoms must form direct bonds between the reactants and as the product. From USPTO-480k, Bradshaw *et al.*[22] curated USPTO-LEF by further filtering out reactions that do not follow linear electron flow, resulting in 73% of the reactions. Neither dataset contains stereochemical information or reactions that alter aromatic bonds.

The USPTO-Full dataset is the most commonly used dataset that is intended to be a maximally-inclusive version of Lowe's USPTO dataset, with a cleaned version containing approximately 1 100 105 unique reactions after preprocessing to remove duplicates and reactions with incorrect atom mappings.[24] As the dataset with the largest number of reactions among the USPTO derivatives, it includes a relatively wide variety of reaction types covering standard chemical transformations from patent literature.

The USPTO-STEREO dataset is another important derivative, which specifically retains stereochemical information for reactions.[23] It contains 1 002 970 reactions, with single-product reactions constituting 92% of the dataset. The cleaning process involved removing 720 768 duplicates from Lowe's dataset by comparing reaction strings without atom mapping. Additionally, 780 reactions were removed because the SMILES strings could not be canonicalized with RDKit due to issues like invalid valence electron counts.

It is important to pay careful attention to what dataset is being described even if the name sounds familiar; for example, the USPTO-Full used by Tetko *et al.*[29] actually contains 4% fewer reactions than the USPTO-Full used by Dai *et al.*[24] due to additional ''cleaning'' by the authors. Even if this data cleaning

is fully justified, one must be hesitant to draw direct comparisons between performance metrics across datasets just because the dataset is referred to by the same name.

Reaction datasets minimally include the identities of reactants, agents, and products. These can be represented in two ways: either by combining the reactants and agents or by keeping them separate. In SMILES notation, the separated format is represented as reactants > agents > products, while the combined format is represented as reactants.agents ≫ products. The former is referred to as "separated", and the latter as "mixed" in some evaluations. Reactants are the compounds that contribute to the structure of the product, and when atom mapping is available, the atoms in the reactants that correspond to those in the product are identified. Atom mapping is a technique used to track the movement of atoms from reactants to products during a chemical reaction.[20,30–34] It assigns a unique identifier to each atom in the reactant molecules and ensures that these identifiers are preserved and mapped to the corresponding atoms in the product. Agents, by convention, are compounds such as solvents or catalysts that do not directly contribute to the product's structure. However, there are cases where small functional groups (*e.g.*, in the case of methylation, solvolysis) or even a single atom (*e.g.*, in the case of halogenations) may contribute to the product's structure, yet the compound is still classified by the dataset as an agent. In the "separated" format, the reaction prediction task is simplified by implicitly restricting possible outcomes to those that do not involve atoms from agents. "Mixed" is the more realistic and practical evaluation, but also more challenging from a statistical learning perspective, as the model must infer which of the chemical species are most likely to contribute to the product structure(s) from a larger pool.

2.1.2 High-throughput experimental datasets. Another valuable source of experimental data comes from high-throughput experiments (HTE), which are efficient for generating large combinatorial datasets.[25,35–38] HTE datasets typically focus on a single reaction type conducted under consistent reaction conditions. The yields are usually measured as assay yields, such as UV area percentages, percent conversions, or product-to-internal standard ratios.[25,39–42] While there are HTE-derived datasets available for specific reactions like Buchwald–Hartwig amination[25,37] and Suzuki cross-coupling reactions,[37] the inherently narrow scope of these datasets, due to their combinatorial reaction spaces, limits their generalizability beyond the specific reactions-and often, specific chemical species-they cover.

In contrast to literature-derived reaction datasets, which typically report only successful reactions that yield an identifiable product, high-throughput experimentation datasets often include measurements across a wide range of reaction conditions, including low-yielding or unproductive outcomes.[27,43,44] Because these experiments systematically explore combinatorial reaction spaces, they can capture both successful and unsuccessful conditions within the same dataset. Such data provide valuable information about reaction feasibility and can help mitigate survivor bias that arises when models are trained exclusively on successful transformations, thereby improving the robustness of machine learning models trained on these datasets.

2.1.3 Mechanistic reaction datasets. Datasets describing reaction mechanisms are less prevalent. Reaction mechanisms are often not reported in the literature because while the final product of a reaction can be characterized using conventional analytical techniques, reaction mechanisms are better described as hypotheses with varying degrees of experimental evidence to support them. The difficulty in documenting these mechanisms arises from the inherent ambiguity in defining a "ground truth mechanism", as multiple plausible mechanistic pathways may lead to the same product or be operative simultaneously.[45] This can occur due to factors like reaction conditions, which can influence the reaction pathway, or the presence of alternative reaction intermediates that lead to the same outcome. Furthermore, due to the transient nature of intermediates, direct observation is often not possible, and proposed mechanisms rely on indirect evidence, inference from overall reaction data, and referencing of "canonical" mechanisms. As a result, one major approach to constructing mechanistic datasets has been to dictate the plausible elementary steps or impute mechanisms from overall reactions, inferring the underlying steps and pathways based on available data, in parallel with computational datasets generated through automated exploration of potential energy surfaces. We summarize some of the most widely used and available mechanistic datasets in Table 2.

One method to generate mechanistic datasets is to do so through the application of expert-written mechanistic templates to different substrate combinations. Baldi has championed this approach of manually curating plausible elementary steps from organic chemistry textbooks and advanced organic chemistry books for over a decade, with recent releases of RMechDB and PMechDB.[46,47] RMechDB contains over 5300 elementary radical reaction steps, each carefully curated to represent plausible reaction pathways. On the other hand, PMechDB focuses on polar reactions, specifically those involving heterolytic bond cleavage and charged intermediates, and includes over 100 000 elementary reaction steps. While RMechDB primarily consists of manually curated data, PMechDB combines manually curated entries with computationally generated steps, expanding its coverage and diversity. Both databases use SMIRKS notation to represent the transformations.

A similar approach to generating mechanistic datasets is through imputation of overall reactions. This also involves constructing reaction templates that dictate transformations at the level of elementary steps, but rather than applying them solely in a forward enumerative sense, they are applied to full reactions (with defined reactants and products) to impute the intermediates.[48,49,51] These reaction templates are intended to cover widely agreed-upon and plausible reaction mechanisms, ensuring that they accurately reflect expert knowledge. Two mechanistic datasets have been made publicly available: Joung *et al.*[48] created a dataset by considering 175 reaction conditions across 86 reaction types. They applied templates in a designated order according to each reaction type to generate the mechanistic dataset. In contrast, Chen *et al.*[51] developed a mechanistic dataset using MechFinder, based on 63 general templates with high coverage, which were applied based on a lookup table to

Table 2 Overview of public and proprietary mechanistic datasets

Name	Size	Description
RMechDB[46]	5300 elementary steps	A public database of elementary radical reaction steps derived from textbooks and research literature, containing curated mechanistic pathways, transition states, and electron flow annotations for radical reactions.
PMechDB[47]	108 987 elementary steps	A publicly accessible database of elementary polar reaction steps, derived from curated literature and combinatorial methods, containing arrow-pushing mechanisms, and balanced reactions for various polar mechanisms.
Joung *et al.* 2024[48]	898 155 elementary steps	A large-scale dataset of elementary reactions derived from the USPTO-Full dataset using expert-curated reaction templates. This dataset includes detailed mechanistic steps, imputed intermediates, and side products.
FlowER dataset[49]	1 445 189 elementary steps	A fully balanced, large-scale mechanistic dataset derived from the USPTO-Full dataset. It includes 1220 expert-curated reaction templates across 252 reaction classes. The dataset ensures broad coverage of organic transformations and is designed to maintain mass and electron conservation.
Rad-6[50]	32 515 reactions	A dataset of reactions for open- and closed-shell organic molecules, derived from DFT calculations. It includes bond-breaking reactions and molecules with up to six non-hydrogen atoms.
mech-USPTO-31k[51]	31 364 reactions	A dataset derived from the USPTO corpus,[14,15] containing organic reactions annotated with 100 hand-encoded reaction types and 63 mechanistic classes. Each reaction is labeled with mechanistic steps in the form of arrow-pushing diagrams.
Reactron dataset[55,56]	*ca.* 2 850 000 elementary steps	A dataset derived from USPTO-480K,[21] containing reaction mechanism annotations generated by MechFinder.[51] Each reaction is labeled with mechanistic steps in the form of arrow-pushing diagrams. The dataset is currently partially available, with the full dataset comprising approximately 2.85M reaction mechanisms.
RGD-1[52]	176 992 reactions	Organic reactions of 413 519 molecules involving C, H, O, and N atoms, containing transition states, activation energies, and enthalpies of reaction.
Grambow *et al.*[53]	16 365 reactions	Organic reactions derived from an automated potential energy surface exploration method.
Transition1x[54]	10 073 reactions	A NEB/CINEB-based reaction-path dataset containing reactant, transition-state, and product geometries produced from automated PES exploration.

impute a mechanistic dataset for 31 364 overall reactions from the USPTO-50k dataset, excluding radical and organometallic reactions. This line of work has been further extended in the Reactron dataset,[51,55,56] which scales up mechanistic annotation to a larger set of reactions derived from USPTO-Full.[21] Contemporaneously, the FlowER dataset[49] was introduced, emphasizing fully balanced reactions that strictly adhere to mass, hydrogen, and electron conservation principles. Unlike the earlier dataset by Joung *et al.*,[48] FlowER explicitly considers acid-base reactions, significantly enriching its mechanistic diversity. This approach ensures comprehensive electron accounting, providing more chemically consistent mechanistic pathways suitable for advanced generative modeling. Similar adherence to conservation law is also enforced in datasets using MechFinder-based approaches.[51,56] Beyond these publicly available mechanistic datasets, several groups have developed large sets of mechanistic transforms that are used in rule-based reaction prediction and reaction discovery.[57–59] Because these collections are not released as standalone datasets, we discuss them separately in Section 4.1.2.

2.1.4 Simulated mechanistic datasets. While many mechanistic datasets rely on human curation or rule-based imputation from overall transformations, an alternative strategy involves generating reaction data entirely from first-principles (or semi-empirical) calculations. These computed datasets do not assume a particular reaction mechanism or transformation *a priori*, but instead derive reactivity and kinetics directly from quantum chemical calculations. This approach enables the systematic exploration of reaction pathways—including those that may be underrepresented or absent in experimental or literature-derived datasets—and provides access to quantities like activation energies, transition states, and enthalpies that are not available in large-scale experimental datasets. The following examples illustrate how such datasets are constructed by scanning potential energy surfaces, optimizing transition states, and validating reaction pathways using methods like density functional theory (DFT), offering a complementary view of chemical reactivity grounded in physical theory. Details of how physics-based methods explore chemical reactivity in the following section will provide additional clarity as to how these datasets were generated.

The Rad-6 database was created to explore the space of radical reactions, specifically focusing on organic molecules involving both closed- and open-shell systems.[50] It contains 10 712 molecules, primarily selected based on a graph-based enumeration of radical and non-radical systems.[60] These molecules, containing hydrogen, carbon, and oxygen atoms, were optimized using DFT calculations with the PBE0 functional and Tkatchenko-Scheffler dispersion corrections. The database is rich in radical fragments and unconventional motifs, such as poly-radicals, and includes molecules up to six non-hydrogen atoms. In addition to the molecular data, the Rad-6-RE dataset provides reaction energies calculated from the atomization energies of the molecules.

The Reaction Graph Depth 1 (RGD-1)[52] dataset was developed to comprehensively explore reaction spaces for organic reactions involving C, H, O, and N atoms. The dataset contains 176 992 reactions, with transition states, activation energies, and enthalpies of reaction. The reactions are derived from 413 519 molecules curated from PubChem, all having no more than 10 heavy atoms. The dataset was generated using a graphically-defined elementary reaction step (ERS) that applies a systematic enumeration process, allowing the identification of a wide range of reactions. Conformational sampling was used to generate diverse

reaction pathways, followed by transition state localization using Yet Another Reaction Program (YARP)[61] and verified using intrinsic reaction coordinate calculations.

Grambow *et al.*[53] created a large-scale dataset of elementary chemical reactions, focusing on the reactants, products, and transition states based on quantum chemical calculations. The dataset consists of 12 000 reactions computed using the ωB97X-D3/def2-TZVP level of theory and 16 365 reactions computed at the B97-D3/def2-mSVP level. These reactions include H, C, N, and O atoms, and the data was derived from an automated potential energy surface exploration method, using the single-ended growing string method[62,63] to optimize reaction paths and transition states. The results provide critical information such as atom-mapped SMILES, activation energies, and enthalpies of reaction.

Building on similar ideas of automated PES exploration, the Transition1x dataset[54] provides dense sampling of reaction pathways rather than focusing solely on discrete transition states. The dataset contains 10 073 reactions generated using nudged elastic band (NEB) and climbing-NEB (CINEB) calculations initiated from 11 961 reactant-product pairs. Each pathway includes reactant, product, and transition-state structures as well as intermediate images along the reaction coordinate, all computed at the ωB97X/6-31G(d) level of theory.

2.2 Physics behind physics-based methods

Physics-based methods use a first-principles approach to make predictions about reactivity. To do this, these methods employ computational approaches to probe the energetics of relevant chemical species (*i.e.* reactants, products, transition states). These calculated energy values are then linked to experimental results through a few theoretical underpinnings, one major one being transition state theory.

2.2.1 Transition state theory. In a first-principles framework, predicting chemical reactions requires identifying which products can form from a given set of reactants. A reactive process proceeds through transition states on the potential energy surface (PES). Physics-based approaches typically begin by locating reactants, products, and the saddle points that connect them, and transition state theory provides the formalism that links these stationary points to reaction kinetics. Transition state theory describes chemical reactivity in terms of a dividing surface that separates reactant and product configurations, with the transition state defined as the highest energy point along the minimum-energy path (MEP) that possesses exactly one imaginary vibrational frequency corresponding to motion along the reaction coordinate.[64]

Transition state theory relies on several key assumptions that connect the potential energy surface to reaction kinetics. It assumes that the reactant ensemble is in thermal equilibrium and forms a quasi-equilibrium with the transition state, allowing the relative population of the transition state to be expressed through Gibbs free energies. Nuclear motion is treated classically, and the reaction is assumed to proceed along a single well-defined minimum-energy path connecting reactants and products, without contributions from alternative pathways. A central requirement is the no-recrossing assumption, which states that trajectories crossing the dividing surface at the transition state do not return to the reactant region. These assumptions enable the derivation of the typical TST rate expression,

$$k = \frac{k_B T}{h} \exp\left(\frac{-\Delta G^{\ddagger}}{RT}\right), \quad (1)$$

where k is the rate constant, k_B is the Boltzmann constant, T is the temperature, h is Planck's constant, and $\Delta G^{\ddagger}$ is the change in Gibbs free energy between the reactant and transition state.

However, the simplifying assumptions also define the situations in which the theory may break down. For example, if there are large energetic releases from reactions that confound the dynamics,[65] trajectories are produced that do not remain on the minimum-energy path. The classical treatment of nuclear motion also neglects quantum tunneling, which allows particles to traverse barriers even when they lack sufficient classical energy.[66] In addition, the assumption that trajectories cross the dividing surface only once is often violated, and recrossing events can return flux to the reactant side and reduce the effective rate.[67] There are formulations of TST that seek to address some of the simplifications associated with the original theory. Variational transition state theory accounts for recrossing effects by redefining the dividing surface.[68] Multipath[69] and multistructure[70] variational TST account for recrossing and contributions from several reactive pathways and structures respectively. Although these formulations improve TST's modeling capability, conventional TST remains widely used for evaluating reaction rates from computed transition states.

Because transition states determine the activation free energy of an elementary step, locating the saddle point on the potential energy surface is essential for connecting molecular structure to reaction kinetics. Unlike reactant and product minima, transition states correspond to isolated first-order saddle points, and small deviations in geometry can drive the system toward either well. As a result, identifying a molecular configuration that lies sufficiently close to the saddle point is often challenging, and practical TS optimization requires an initial guess structure that captures the correct chemical transformation. These considerations motivate the development of algorithms that construct or refine such guesses, which are discussed in the following Section 4.2.2.

2.2.2 Conformer generation and search. Accurate three-dimensional molecular geometries are required for evaluating reaction energetics, since both transition-state structures and the thermodynamic quantities entering kinetic expressions depend on atomic arrangements. Likewise, the enthalpy and Gibbs free energy of reactants and products reflect the energetics of their accessible conformational states. Many molecules populate multiple conformational isomers arising from rotations of single bonds as shown in Fig. 2, and these conformers can differ substantially in energy, meaning that experimentally relevant thermodynamic quantities often correspond to a Boltzmann-weighted average over the conformational ensemble.[71] Consequently, physics-based methods must identify a representative set of low-energy conformers prior to any reliable assessment of

reaction energetics or transition-state searches. Although manual inspection of conformers may be feasible for small and rigid systems, the number of possible conformations increases combinatorially with the number of rotatable bonds (n_{rot}), often approximated as $3^{n_{rot}}$ when each torsion is sampled in three low-energy orientations. This rapid growth makes exhaustive manual enumeration impractical for most reaction systems, necessitating algorithmic conformer generation.

Given these challenges, conformer generation methods algorithmically construct three-dimensional geometries by sampling torsional degrees of freedom according to chemically reasonable constraints. Existing approaches fall broadly into systematic, stochastic, and machine-learned strategies, each differing in how torsion angles are selected and how the resulting geometries are filtered to produce a representative ensemble.

Systematic conformer generation methods enumerate conformational space by explicitly rotating each rotatable bond according to a set of allowed torsion angles derived from empirical structural data. These approaches rely on knowledge-based torsion libraries constructed from crystallographic or protein-bound ligand datasets, which encode the preferred low-energy torsional states observed in real molecules.[73–76] Conformers are assembled incrementally by building the molecular structure from a central fragment outward and applying all feasible torsion-angle assignments in either a depth-first or combined depth-first and breadth-first manner. Throughout this construction process, chemically invalid structures are removed through checks on bond lengths, bond angles, and steric clashes, often supported by VSEPR-derived ideal geometries and covalent radii. Many systematic generators incorporate specialized handling of ring systems or macrocycles, either through ring template libraries or through numerical optimization methods that rebuild macrocyclic torsions after enumeration.[75] Because systematic enumeration rapidly produces large numbers of candidate conformers, most algorithms apply RMSD-based clustering to identify a diverse and representative ensemble that remains within user-specified size limits.

Stochastic conformer generation methods explore conformational space by combining random coordinate perturbations with local energy minimization, rather than exhaustively enumerating torsional states. Distance-geometry approaches construct candidate geometries by sampling distance matrices that satisfy chemically reasonable upper and lower bounds, followed by force-field refinement.[77,78] Other stochastic strategies operate directly in internal-coordinate space by assigning random torsion angles and optimizing the resulting structures.[79–81] Energy-directed stochastic methods use molecular-dynamics trajectories perturbed along low-frequency vibrational modes to preferentially access low-strain conformations that may be difficult to locate through unbiased sampling.[82] These stochastic approaches provide a flexible framework for conformer generation in systems where combinatorial torsion enumeration becomes intractable.

Machine-learned conformer generation methods use generative models to predict three-dimensional molecular geometries directly from the molecular graph or from torsional degrees of freedom. Unlike systematic or stochastic approaches, which rely on predefined torsion libraries or random perturbations, machine-learned methods train on large datasets of experimentally derived or simulation-generated conformers to learn distributions over 3D structures. Graph neural networks can be used to predict local atomic frames and torsion angles conditioned on the molecular graph,[72,83,84] while diffusion-based approaches operate either in full Cartesian coordinate space or in torsion angle space to sample conformers through learned denoising trajectories.[85,86] Other models explicitly target the Boltzmann distribution by incorporating energy-based rewards into their training objectives.[87,88] These approaches have been shown to excel at generating low energy conformations however are known to struggle to recover ensembles that are distributed differently than their underlying reference structures (*e.g.*, bioactive conformations *vs.* low energy conformations).[89]

2.2.3 Energetic evaluation of molecular geometries. The previous sections focused on identifying or sampling the structures of reactants, products, and transition states, but determining where these configurations lie on the potential energy surface requires evaluating their electronic energies. These energies arise from solving approximate forms of the many-electron Schrödinger equation and determining various thermodynamic and kinetic quantities. Energetic evaluation involves computing the electronic energy, its gradients with respect to nuclear coordinates, and second derivatives that characterize equilibrium structures and transition states. These quantities enable the determination of reaction energies and barrier heights that enter models of reactivity such as transition state theory. Reliable energetic evaluation remains challenging because the exact solution to the many-electron Schrödinger equation is intractable, which necessitates a hierarchy of approximate methods that differ in the physical rigor of their underlying models, their computational cost, and their ability to describe diverse chemical environments. As a result, practical energetic evaluation spans a broad methodological landscape that includes *ab initio* wavefunction approaches, density functional theory (DFT), semiempirical quantum models, classical interatomic potentials, and neural network based machine-learned potentials. As summarized in Fig. 3, these approaches occupy different regions of the cost-accuracy spectrum, and the choice among them depends closely on the chemical complexity of the system being studied and on the level of fidelity required for modeling reaction pathways. Given this landscape, we outline the essential features of each class of methods related to the reaction modeling and more detailed other discussions available in the literature.[90–100]

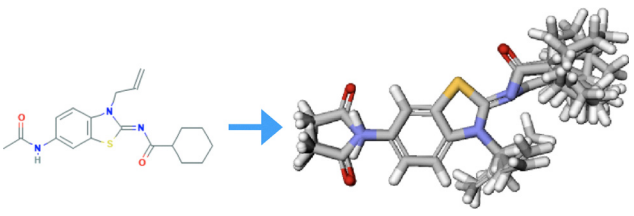


Fig. 2 Visualization of the many possible conformers available to a molecule. The number of accessible conformations combinatorially increases with the number of rotatable bonds. Reproduced with permission from ref. 72.

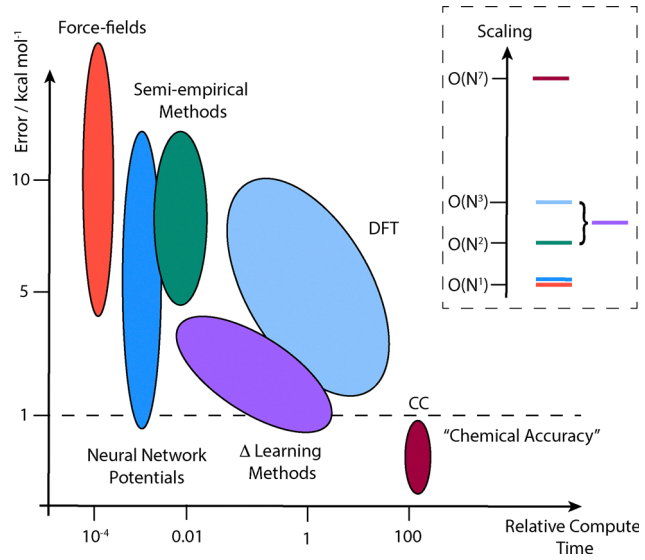


Fig. 3 Cost-accuracy tradeoff associated with methods used to evaluate the energy of a 3D geometry of a molecule or atomistic system. More accurate methods yield poorer scaling which leads to higher relative compute times. Reproduced with permission from ref. 101. Copyright 2023 John Wiley and Sons.

In practice, accurate energetic evaluation for reaction modeling involves several additional challenges beyond the choice of electronic structure method. Many reactions proceed on multiple potential energy surfaces, requiring the treatment of excited states, nonadiabatic effects, or spin-state crossings, particularly in photochemical systems and transition-metal catalysis.[102,103] Open-shell species, near-degenerate electronic configurations, and multireference character can further complicate the reliable description of reaction intermediates and transition states.[104] Environmental effects such as solvation and conformational flexibility introduce additional sources of uncertainty, as they can substantially alter relative energetics and barrier heights.[105] These factors make the construction of chemically accurate reaction energy profiles a nontrivial task, even when high-level electronic structure methods are employed.[106]

These methods estimate the wave function to determine the energy of a given molecular system. The Hartree–Fock method is one of the original methods to perform this calculation; it assumes the wave function can be approximated by a single Slater determinant.[107] Although Hartree–Fock has practical utility, it neglects electron correlation which limits its accuracy. There are several methods, post-Hartree–Fock methods, that build upon Hartree–Fock and address electron correlation. Full configuration interaction (CI) is an example of this that yields the exact solution to the non-relativistic Schrödinger equation, however is not practical for molecules with more than just a few atoms due to its exponential scaling.[108–110] Other post-Hartree–Fock methods employ estimations for electron correlation. For example, Møller–Plesset perturbation methods add electron correlation using Rayleigh–Schrödinger perturbation,[111] and coupled cluster methods use the exponential cluster operator to include electron correlation.[112,113] These methods trade achieving the exact solution for more favorable scaling ($O(N^5)$ and $O(N^7)$ respectively) allowing them to be used on larger molecules when compared to full CI.

The more common way to estimate the wave function is with density functional theory (DFT) calculations. The theory behind DFT are the Hohenberg–Kohn theorems which state that the ground state energy of a many electron system is uniquely defined by the electron density, and that an energy functional exists that is minimized at the ground state density.[114] By using the electron density, DFT can achieve better scaling than post-Hartree–Fock methods ($O(N^3)$) while still accounting for electron correlation. However, since the exact energy functional is not known, DFT remains an approximation and is generally less accurate than post-Hartree–Fock methods. Within DFT, there is a cost-accuracy trade-off where more accurate functionals and larger basis sets require more computational power but provide a more accurate description of the physics governing the system. For example, moving up Jacob's ladder to more accurate functionals involves an increasingly nonlocal description of exchange-correlation energy which provides a more exact description of electron density while necessitating an increased number of integral evaluations.[115] DFT has been used extensively in the characterization of reactive systems,[116,117] and is more often employed for providing *post hoc* insight into the results of experimentally performed reactions than for providing true predictive ability. It is worth noting too that "more accurate" methods are still context-dependent, and different chemical systems benefit from different design choices.[118]

Semi-empirical methods have emerged as a class of methods that reduce computational expense through the use of a less complete approximation for electronic integrals.[119–123] To mitigate the loss of fidelity, these methods use adjustable empirical parameters to correct energies based on experimental data or accurate theoretical data.[124] Semi-empirical methods can be several orders of magnitude faster than DFT[125] and have been shown to output reliable ground state and transition state geometries for certain reactions.[126–128] Although the accuracies of energies estimated by semi-empirical methods are often regarded as insufficient for reaction modeling,[126] the recently developed method g-xTB has demonstrated substantially more promising accuracy than its predecessors when estimating reaction barriers.[121]

Interatomic potentials are a set of methods that directly predict the energy of a molecular system from its geometry in order to reduce cost. Much of the computational expense associated with the above methods is the necessity of a self-consistent field method for calculating the wave function. Interatomic potentials circumvent this need. Force fields are the prototypical category of interatomic potentials parameterized by experimental data or quantum chemical calculations, but there are few force fields specifically parameterized for reactive geometries which allow them probe kinetics of reactions directly. Two such examples, ReaxFF[129,130] and the charge-optimized many-body potential (COMB),[131,132] were initially developed to handle hydrocarbons but were eventually expanded to handle more organic elements and metals. One

key difference between the two is ReaxFF is parameterized to reproduce reactive barrier heights whereas COMB is parameterized to reproduce elastic properties of materials.[130]

The traditional means of defining and parameterizing empirical force fields has been revisited in recent years through the lens of machine learning. Specifically, machine-learned interatomic potentials (or neural network potentials) have emerged as a new category of interatomic potentials parameterized by a neural network trained on quantum chemical calculations. This allows machine-learned interatomic potentials to, in principle, overcome the cost-accuracy trade-off by providing quantum chemistry accuracy at the cost of a single model inference pass;[133–135] their accuracy relies on having adequate training data. Certain machine-learned interatomic potentials are trained on both equilibrium and transition state-like geometries which can allow them to characterize geometries associated with reactions. Due to the millions of CPU hours it can take to generate training data, most reactive machine-learned interatomic potentials have been limited to narrow chemical spaces (*i.e.* small number of elements and reaction types) but still have been used to computationally explore chemical reactions.[136–141] One strategy to improve the performance of these potentials for a specific reaction is active learning. This entails sampling points in the chemical regime of interest where the model is uncertain with quantum chemical calculations. This minimizes the amount of training data necessary to achieve parity with a reference functional for specific reactions.[142–146] Furthermore, recent developments have resulted in increasingly general reactive machine-learned interatomic potentials. The dataset Open Molecules 2025 (OMol25) includes over 10 million DFT snapshots along reaction pathways with samples from organic and organometallic reactions; neutral, charged and radical reactions; and electrolyte reactions.[147] Models trained on this dataset should be expected to exhibit a broader domain of applicability. More targeted efforts have resulted in the development of reactive AIMNet models capable of modeling neutral reactions (AIMNet2-rxn),[148] Pd-catalyzed reactions (AIMNet2-Pd),[149] and radical reactions (AIMNet2-NSE).[150]

3 Simultaneous prediction of reaction products and their likelihoods

One formulation of the task of reaction outcome prediction is as a single-step mapping from reactants to products: predicting not only which product is formed, but also how likely that outcome is. This "end-to-end" formulation has become increasingly popular with the rise of deep learning, which excels at learning complex input-output relationships from large datasets. Rather than decoupling candidate generation and evaluation, these methods aim to directly infer the most likely outcome from the input.

This category also includes certain physics-based methods that perform simultaneous product prediction, for instance, by exploring potential energy surfaces and identifying both feasible products and their corresponding energetics. These two methodologies, data-driven and physics-based, will both be discussed in this section. Each method has its own way of defining and computing the "likelihood" of an outcome. For data-driven models, this often corresponds to predicted probabilities or ranked scores inferred from statistical patterns in historical data. For physics-based methods, likelihood is typically evaluated using computed activation energies or thermodynamic quantities such as Gibbs free energy, which reflect the kinetic or thermodynamic feasibility of a transformation. An illustrative comparison between these two paradigms is shown in Fig. 4. While these likelihoods are not directly comparable, treating them as instances of the same overarching task of assessing the favorability of possible outcomes helps to conceptually unify methods that are developed from different principles.

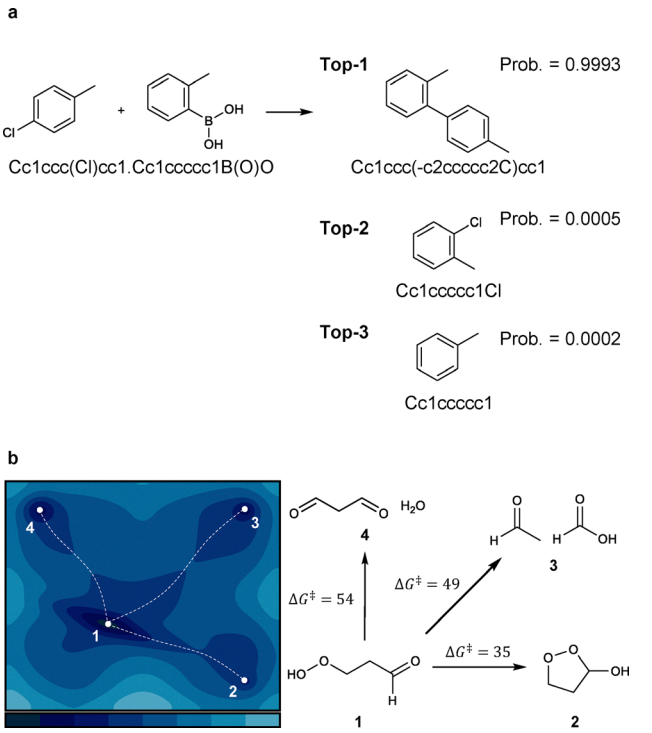


Fig. 4 Representative paradigms for the simultaneous prediction of reaction products and their likelihoods. (a) An example of an organic reaction with multiple reactants. End-to-end models aim to directly predict the most likely product and its associated likelihood in a single step, without separately enumerating candidates or scoring them *post hoc*. These models bypass mechanistic reasoning and instead learn a direct mapping from reactants to outcomes. Note that this often results in models learning to describe the data well (as illustrated by the "correctly" predicted top-1 product) but misunderstanding the importance of reagents and catalysts (as illustrated by the absence of an appropriate catalyst and base). (b) A toy potential energy surface for ketohydroperoxide paired with elementary steps for possible reactions. Physics-based models explore the local potential energy surface of the reactants to find feasible elementary steps. These elementary steps are evaluated by their activation barrier.

3.1 Data-driven models

Solving the end-to-end formulation in a data-driven manner requires a way of proposing or generating product molecules given a set of reactant molecules. Hence, use of statistical

learning approaches is quite popular. Conceptually, these models are learning $p(P|R)$ where R is a set of reactants, P is the product(s), and $p(P|R)$ defines a probability distribution over which to sample hypothetical products.[28,151]

Directly learning a normalized likelihood $p(P|R)$ is challenging due to the combinatorially large and discrete nature of the product space. Furthermore, product molecules are structured outputs such as graphs or sequences, making it difficult to define and estimate probability densities. As a result, most approaches will train a model to sample products from this distribution, $\hat{P} \sim p(P|R)$, rather than actually learning a normalized likelihood $p(P|R)$.[49,152,153]

After sampling one or more products, different models lend themselves to different ways of then estimating a likelihood that resembles or correlates with $p(P|R)$. A primary axis of organization is how one defines the model and representation of product molecules in order to perform this sampling.

These models differ in how they represent molecules—either as graphs[154,155] or as sequences[29,152]—and accordingly adopt different neural architectures and modeling assumptions. Because these representations encode different aspects of chemical structure and syntax, they lead to distinct modeling strategies, each with trade-offs in capturing topological detail, chemical validity, and predictive uncertainty.

The performance of various data-driven models for predicting reaction outcomes is summarized in Table 3. This table contains the vast majority of methods or models that predict reaction products from reactants in a data-driven manner, including both models that directly generate reaction products along with their associated likelihoods—discussed in Sections 3.1.1 and 3.1.2—and two-step models that decouple product enumeration and scoring, as described in Section 4. Most models represent molecules as either molecular graphs or SMILES sequences (or other line notations analogous to SMILES,[156,157]) though some adopt alternative encodings such as fingerprints. Model performance is most often reported as the top-k accuracy, defined as the proportion of test instances for which the correct product appears within the model's top k ranked predictions.

Despite nearly a decade of progress, top-1 accuracies on commonly used USPTO-derived benchmarks have largely plateaued around 90%. Rather than indicating a strict saturation of model performance, this trend likely reflects the fact that reaction outcome prediction is inherently under-specified in these datasets. In particular, the absence of detailed reaction conditions and procedures means that the mapping from reactants to products is not uniquely defined, introducing legitimate ambiguity in what constitutes the "major" product. In many cases, reported products may not correspond to a dominant ($>50\%$) outcome, but rather reflect experimentally convenient or selectively reported transformations. As a result, evaluation metrics based on exact product matching may penalize chemically reasonable predictions that differ from the recorded product, effectively imposing a ceiling on achievable top-1 accuracy. Additional factors, including reporting bias and data quality issues in patent-derived datasets, may further contribute to this performance limit.

3.1.1 Product prediction as graph generation. Graph-based models for reaction outcome prediction frame chemical reactions as transformations over molecular graphs, where atoms are represented as nodes and covalent bonds as edges. From an application perspective, these models are tasked with identifying which bonds are likely to change—break, form, or rearrange—given a set of reactant and agent molecules, and they can be instantiated as graph editing models that describe chemical reactions as a sequence of discrete modifications ("edits") to the molecular graph of the reactants.[159,161] By focusing on these graph-level changes, they aim to reconstruct the most likely product structures. The labels for the dataset used for training require that reactions be atom mapped in a consistent manner, so that cheminformatics tools can compare reactant and product structures to derive the set of edits that connect the two. These edit sequences provide interpretable stepwise changes that resemble mechanistic pathways, though they have no grounding in or alignment with actual mechanistic understanding.[22,163] These structural formulations define how models represent potential outcomes, but product prediction also requires deciding which of the generated outcomes is most plausible.

While many models produce and rank multiple candidate outcomes, the notion of 'scoring' varies significantly across frameworks. These range from learned likelihoods or autoregressive beam search scores[154,159] to empirical frequencies from repeated sampling.[49] These scores are generally only relative measures of plausibility and are not calibrated as absolute probabilities. Formally, after all, these models are trained to predict the major product of a reaction as it is recorded, not a quantitative branching ratio, yield, *etc.* Therefore, the term "likelihood" in this context refers not to a calibrated probability but to a more general notion of confidence or favorability assigned to predicted outcomes. Still, likelihood may and often does empirically correlate with accuracy.

In this section, we focus on models that produce product structures and associated likelihoods in a single integrated process. This includes models that predict a sequence of graph edits in an autoregressive manner so long as they do not separately evaluate or rank candidate outcomes after generation. In contrast, models that involve an explicit evaluation of predicted candidates (*e.g.*, *via* scoring, ranking, or binary feasibility assessment) are discussed later under two-step frameworks under Section 4.

3.1.1.1 Autoregressive graph editing. A notable early example of the one-step graph-editing paradigm is the Graph Transformation Policy Network (GTPN), which frames reaction prediction as a Markov Decision Process over molecular graphs.[154] Rather than relying on supervised learning to predict specific edits, GTPN uses reinforcement learning to learn a policy that sequentially generates "reaction triples"—consisting of a decision to continue or terminate editing, a pair of atoms, and a new bond type. These triples are predicted autoregressively and applied to update the molecular graph step by step. While the model includes a termination policy to decide when to stop, this mechanism functions as part of the generation process and

Table 3 Data-driven models for predicting reaction outcomes. Models are grouped by their prediction strategy[a]

Year	Model	Dataset	Setting	Representation	Top-1	Top-2	Top-3	Top-5	Top-10
Graph generation models (Section 3.1.1)									
2019	GTPN[154]	USPTO-15k	Separated	Graph	82.4		85.7	85.8	
2020	MPNN + ILP[158]	USPTO-480k	Separated	Graph	90.4	93.2	94.1	95.0	
2021	MEGAN[159]	USPTO-480k	Separated	Graph	89.3	92.7	94.4	95.6	96.7
2021	MolR[160]	USPTO-480k		Graph	88.2				
2021	NERF[161]	USPTO-480k		Graph	90.7	92.3	93.3	93.7	
2023	NERF* [162]	USPTO-480k		Graph	91.5	93.6	94.4	95.1	95.6
2023	ReactionSink[163]	USPTO-480k		Graph	91.3	93.3	94.0	94.5	94.9
2025	FlowER[49 b]	FlowER dataset	Mixed	Graph	88.4	96.3	97.8	98.4	98.5
2025	Reactron[56]	Reactron dataset	Mixed	Graph	96.4	97.6	97.8	—	—
Sequence generation models (Section 3.1.2)									
2018	Seq2Seq[23]	USPTO-480k	Separated	Sequence	83.2	87.7	89.2		
2019	Transformer[153]	USPTO-480k	Separated	Sequence	90.4	93.7	94.6	95.3	
2020	Augmented Transformer[29]	USPTO-480k	Separated	Sequence	92.0	95.4		97.0	
2022	Chemformer[164]	USPTO-480k	Separated	Sequence	92.5			94.9	95.1
2022	Chemformer-Large[164]	USPTO-480k	Separated	Sequence	92.8			94.9	95.0
2019	Transformer[153]	USPTO-480k	Mixed	Sequence	88.6	92.4	93.5	94.2	
2020	Augmented Transformer[29]	USPTO-480k	Mixed	Sequence	90.6	94.4		96.1	
2022	Graph2SMILES[155]	USPTO-480k	Mixed	Sequence	90.3		94.0	94.8	95.3
2022	Chemformer[164]	USPTO-480k	Mixed	Sequence	90.9			93.8	94.1
2022	Chemformer-Large[164]	USPTO-480k	Mixed	Sequence	91.3			93.7	94.0
2022	T5Chem[165]	USPTO-480k	Mixed	Sequence	88.9	92.9		95.2	
2022	R-SMILES[166]	USPTO-480k	Mixed	Sequence	91.0	95.0		96.8	97.0
2023	SeqAGraph[167]	USPTO-480k	Mixed	Graph	88.9		94.6	95.5	96.5
2023	BiG2S[168]	USPTO-480k	Mixed	Graph	89.0		94.6	95.6	96.6
2019	Transformer[153]	USPTO-STEREO	Separated	Sequence	78.1	84.0	85.8	87.1	
2022	Motif-Reaction[169]	USPTO-STEREO	Separated	Sequence	82.7		88.8	89.6	
2019	Transformer[153]	USPTO-STEREO	Mixed	Sequence	76.2	82.4	84.3	85.8	
2022	Graph2SMILES[155]	USPTO-STEREO	Mixed	Graph	78.1		84.6	85.8	86.8
2022	Motif-Reaction[169]	USPTO-STEREO	Mixed	Sequence	79.9		86.4	88.2	
2024	Transformer[48]	Subset of Pistachio	Mixed	Sequence	90.2	93.8	94.8	95.5	
2024	Graph2SMILES[48]	Subset of Pistachio	Mixed	Graph	98.3	98.6	98.7	98.7	
2024	Transformer[48 b]	Joung *et al.* 2024[48]	Mixed	Sequence	83.5	88.3	89.3	90.1	90.7
2024	Graph2SMILES[48 b]	Joung *et al.* 2024[48]	Mixed	Graph	88.7	91.0	91.3	91.5	91.6
Two-step models (Section 4)									
2022	LocalTransform[170]	USPTO-480k	Separated	Graph	92.3	95.6	96.5	97.2	
2017	WLDN[21]	USPTO-480k	Mixed	Graph	79.6		87.7	89.2	
2019	WLDN[171]	USPTO-480k	Mixed	Graph	85.6	90.5	92.8	93.4	
2021	MEGAN[159]	USPTO-480k	Mixed	Graph	86.3	90.3	92.4	94.0	95.4
2022	LocalTransform[170]	USPTO-480k	Mixed	Graph	90.8	94.8	95.7	96.3	
2019	GTPN[154]	USPTO-15k	Separated	Graph	83.4		85.7	86.7	
2019	GTPN[154]	USPTO-480k	Separated	Graph	83.2		86.0	86.5	
2018	Electro[22]	USPTO-LEF	Separated	Graph	87.0	92.6	94.5	95.9	
2017	Neural-Symbolic[172]	Reaxys		Fingerprint	78.0				98.0
2021	WLDN[173]	2225 Baeyer-Villiger oxidation		Graph	90.4	93.4	93.9		
2024	WLDN[48]	Subset of Pistachio	Mixed	Graph	95.0	97.1	97.5	97.6	
2024	WLDN[48 b]	Joung *et al.* 2024[48]	Mixed	Graph	79.4	86.5	87.4	88.0	88.3

[a] Datasets with the same name may differ in preprocessing or splits (*e.g.*, train/validation/test), making top-*k* accuracy values not directly comparable. [b] In mechanism prediction, multiple valid outcomes may exist, but models can assign only one per rank, imposing an upper bound on top-*k* accuracy.

does not involve evaluating or ranking alternative outcomes. The policy is optimized using an actor-critic objective, allowing GTPN to generate plausible product structures without predefined reaction rules or templates.

While GTPN learns edit sequences *via* reinforcement learning, an alternative strategy is to directly supervise autoregressive edit prediction. Autoregressive graph-editing is exemplified by the Molecule Edit Graph Attention Network (MEGAN) in Fig. 5a, which formulates reaction prediction as a sequence of graph edits applied to the input reactant graph.[159] Each edit corresponds to a discrete transformation, such as adding or removing atoms, modifying bonds, or adjusting atomic attributes like formal charge, and is selected from a learned vocabulary of chemically plausible actions. These edits are predicted autoregressively, one at a time, using masking mechanisms to ensure chemical validity at each step. Autoregressive graph-editing has also been extended beyond product prediction to model reaction mechanisms. For example, Reactron[56] operates at the level of electron flow by autoregressively predicting source-sink interactions that define arrow-pushing steps, thereby constructing reaction mechanisms as sequences of elementary transformations rather than directly generating the final product.

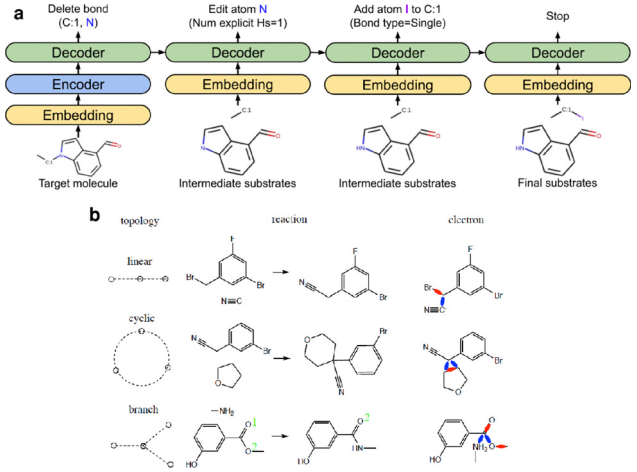


Fig. 5 Two representation models for graph-based models for reaction product prediction. (a) Prediction of sequence of graph edits (MEGAN). Reproduced with permission from ref. 159. Copyright 2021 American Chemical Society. (b) Non-autoregressive prediction (NERF). Reproduced with permission from ref. 161.

3.1.1.2 Non-autoregressive graph editing. The sequential generation of individual structural modifications to graphs through an autoregressive decoding process takes many steps, or iterations, to make a full prediction. At each step, the model conditions its next prediction on the edits made so far, producing a linearized sequence of graph transformations. This introduces some subtle modeling complexities. Sequences necessarily follow a certain generation order, are prone to error accumulation, and prevent efficient parallelization during inference. In contrast, non-autoregressive methods aim to predict the entire set of structural changes in a single step, without depending on previously generated edits. This one-shot prediction paradigm avoids the need to define an edit order, enables parallel decoding, and may be better suited to reactions involving complex or simultaneous transformations, which typically require a larger number of graph edits and thus longer sequences that are more prone to error propagation and harder to generate reliably in a stepwise manner.

A representative example of this one-shot prediction strategy is NERF (non-autoregressive electron redistribution framework), which frames reaction prediction as a global electron redistribution task over the molecular graph.[161] Instead of generating atom-level edits or bond changes sequentially, NERF samples latent vectors from a learned distribution and decodes them into predictions of how electrons are redistributed across the molecule. Specifically, the model estimates the movement probabilities of electrons between all atom pairs simultaneously, using attention-based mechanisms to infer bond formation and breaking events.

Building on the non-autoregressive formulation of NERF, ReactionSink addresses a key limitation in electron redistribution modeling: the failure to satisfy fundamental physical constraints simultaneously.[163] While NERF enforces a form of row-wise electron conservation through attention-based bond updates, it does not guarantee symmetry or full electron-counting consistency in its predicted bond changes. In particular, NERF applies row-wise normalization to attention weights, ensuring that each atom distributes a fixed amount of electron flow, but not that receiving atoms obey the same constraint. Moreover, its post-hoc symmetrization step can disrupt the initial conservation property, leading to inconsistencies between predicted donor and acceptor electron counts. ReactionSink introduces a refined decoder architecture that imposes both the electron-counting rule and the symmetry rule, ensuring doubly conservative predictions. This is achieved through the use of Sinkhorn normalization,[174] which transforms attention maps into doubly stochastic matrices. Compared to NERF, ReactionSink the improvement in top-1 accuracy on USPTO-480k is modest (from 90.7% to 91.3%). Note that this benchmarking dataset exhibits a plateau in top-1 accuracy (Table 3), making it increasingly difficult to distinguish model performance at the high-accuracy regime, as discussed further in Section 3.1.

A parallel line of work seeks to improve product diversity and uncertainty modeling within non-autoregressive frameworks. Guo *et al.*[162] augment NERF by eliminating the use of a latent prior and instead introducing two sources of controlled stochasticity: boosting, which trains an ensemble of specialized models to capture diverse reaction outcomes, and dropout, which introduces finer variations during inference. While their method does not explicitly enforce conservation laws like ReactionSink, it achieves improved product diversity on multi-selectivity reactions.

While these models operate within the non-autoregressive paradigm, they adopt an end-to-end formulation that trains on final product structures without capturing reaction intermediates. However, in many areas of chemistry, understanding the underlying reaction mechanism—rather than just the final product—is of interest. To capture this mechanistic detail, some machine learning approaches operate at the level of elementary steps.[48,49,56] Flow matching for Electron Redistribution (FlowER) implements this non-autoregressive paradigm by treating elementary step prediction as a flow-matching problem over the bond-electron matrix.[49] FlowER predicts mechanistic steps by learning a time-dependent vector field over interpolated bond-electron matrices, enabling the model to generate electron movements that update the bond–electron matrix in a physically consistent manner.

Graph-based models provide a powerful framework for chemical reaction prediction by explicitly representing molecular structures as graphs, where atoms are nodes and bonds are edges. Their primary strength lies in their ability to capture the local chemical environment and structural changes using message-passing algorithms, enabling interpretable predictions that align closely with chemical intuition, such as bond formation, bond breaking,[154,159,175] and electron redistribution.[49,161] Models like MEGAN[159] and ReactionSink[163] further enhance the reliability and accuracy of predictions by incorporating structural constraints and enforcing valence rules, ensuring that the predicted reactions are not only structurally valid but also consistent with fundamental chemical principles.

However, the sequential nature of graph-editing models (such as MEGAN[159] and GTPN[154]) can be computationally expensive, especially for complex reactions that involve numerous steps.

The process of iteratively predicting and updating graph structures poses significant scalability challenges, particularly when dealing with reactions requiring multiple transformations. While these patterns may resemble a reaction mechanism in some cases, they do not accurately represent true reaction mechanisms. To predict actual mechanisms, including electron flow (or arrow-pushing diagrams), the model would need to operate at the level of elementary steps, as demonstrated by models like FlowER,[49] and DeepMech,[175] as well as autoregressive approaches like Reactron,[56] which explicitly generate the sequence of mass- and charge-conserving mechanistic transformations rather than only the final product structure.

3.1.2 Product prediction as sequence generation. Molecular species in chemical reactions can be described using a string representation like SMILES.[176] Molecular graphs can be deterministically converted to SMILES strings and vice versa (though there is some subtlety to this related to stereochemistry, noncovalent bonding, *etc.*), meaning that the theoretical information content in each is arguably the same. However, SMILES strings and other line notations (*e.g.* DeepSMILES,[177] SELFIES,[178] and BigSMILES[179]) are amenable to borrowing techniques from natural language processing (NLP). The task of reaction prediction can be formulated as a sequence-to-sequence translation task: the goal is to map the concatenated strings of reactants and agents to the string of the major product(s), analogous to translating between languages that use the same alphabet.

The use of sequence-to-sequence models for chemical reaction prediction was first demonstrated, to our knowledge, by Nam and Kim,[152] who used the then state-of-the-art technique of a long short-term memory (LSTM)-based model (Fig. 6) operating on SMILES strings, trained on reaction data from the USPTO dataset prepared by Lowe.[14,15] Building on this idea and bringing significantly greater scale and practicality, Schwaller *et al.*[23] and demonstrated improved performance on larger-scale datasets, such as USPTO-480k, establishing sequence-to-sequence translation as a viable paradigm for practical reaction outcome prediction, favorably competing with the graph-based methods published shortly before (Table 3).

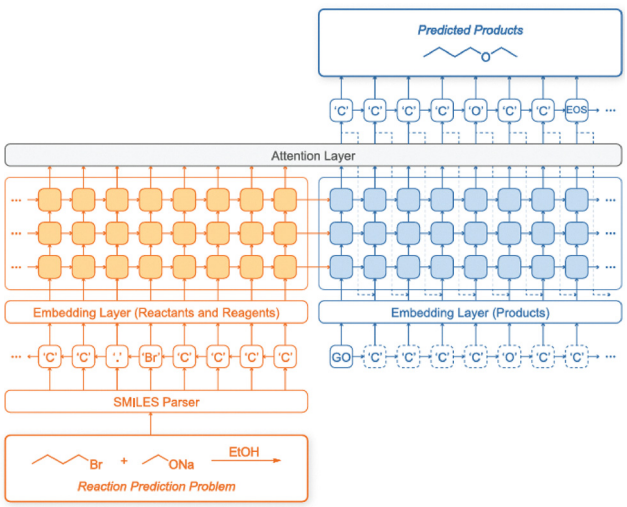


Fig. 6 Model architecture for seq2seq. Reproduced with permission from ref. 152.

Transformer-based models further advanced this approach by replacing recurrence with attention-based architectures,[180] allowing for improved modeling of long-range dependencies and better scalability to large datasets. When applied to chemical reaction prediction, these models improved the accuracy of product prediction over their predecessor LSTMs,[29,153,164] and are now widely used for SMILES-based modeling across reaction prediction tasks just as transformers serve as the backbone for the vast majority of autoregressive prediction tasks in contemporary deep learning.

The notion of "likelihood" in these sequence-generation models conveniently arises from natural language processing settings. To provide a ranked list of multiple candidate products, sequence generation models employ various beam search strategies. The predicted product is typically the SMILES string with the highest joint probability—computed as the product of per-token probabilities. While this provides a natural ranking, it does not guarantee chemical feasibility. In practice, models may assign high likelihoods to products that are syntactically well-formed but chemically implausible, especially when trained on data with limited constraints.[48,181,182] This disconnect between statistical likelihood and chemical likelihood remains a key limitation in SMILES-based architectures.

It is worth emphasizing that this class of models is conceptually distinct from graph-based approaches, which treat molecules as topological graphs and rely on atom- and bond-level features. Sequence-generation models operate entirely in token space, where SMILES strings are parsed into sequences of discrete units—tokens—such as atoms, bonds, or common substructures. These models must learn the SMILES syntax, chemical validity, reactivity, and conservation laws implicitly through exposure to large datasets. However, they do so quite effectively in practice.

In this section, we review representative sequence-generation models for reaction prediction, focusing on their ability to not only generate plausible product structures but also rank multiple outcomes by predicted likelihood. We discuss architectural innovations, data augmentation strategies, domain adaptation through transfer learning, and attempts to improve interpretability and physical fidelity. We also examine how recent models incorporate molecular structure features, including graph-based inputs and motif-aware representations, while retaining the efficiency and generality of sequence-based generation. A summary of their performance across datasets is provided in Table 3.

3.1.2.1 Evolution of sequence modeling approaches. As described earlier, the transformer architecture quickly became the standard for sequence translation tasks after its emergence in 2017, including to the task of product prediction with SMILES representations.[153] Two observations made early on with such SMILES-based models is (1) that they must learn the syntax of SMILES from scratch and (2) that the lack of uniqueness of SMILES strings lends itself to data augmentation, which may help address the first observation as well.

A single molecular structure can be represented by many valid SMILES strings, even if cheminformatics tools like RDKit contain canonicalization strategies to generate strings reproducibly. Handling this SMILES ambiguity has been a challenge in sequence-based prediction. Unlike graph-based approaches, which inherently preserve molecular connectivity regardless of atom ordering through permutation equivariance, SMILES-based models must recover structural information from a linear sequence of tokens. This makes them sensitive to the exact SMILES representation encountered during training and inference, with the risk of overfitting to specific token patterns rather than learning the underlying chemical structure.

One way to mitigate the sensitivity to specific SMILES representations is to expose the model to multiple valid encodings of the same molecule during training. By learning from varied token sequences corresponding to the same underlying structure, the model can better generalize to unseen SMILES formats at inference time. The Augmented Transformer[29] incorporated randomized SMILES augmentations throughout training and inference, applying them either to the input SMILES alone or to both input and output SMILES. This strategy exposed the model to diverse syntactic variations of the same underlying molecules, modestly improving its top-1 accuracy with full reaction augmentation.

In contrast to random augmentation, the R-SMILES approach[166] promotes consistent token ordering in the most chemically relevant regions of the SMILES string, specifically the reactive centers. By aligning reaction sites while allowing variability in non-reactive portions of the molecule, R-SMILES encourages substrings to be copied over from the reactants to products. Such randomized augmentation and structure-aware alignment strategies demonstrate how targeted exposure to multiple SMILES representations can improve the robustness of SMILES-based models. These techniques are now common components in recent sequence-based models.

Building upon this insight, Chemformer[164] introduced a pretrained Transformer architecture tailored for molecular tasks. Rather than training models from scratch, Chemformer adopted a two-stage process: unsupervised pretraining on large corpora of molecular SMILES (*e.g.*, from PubChem[183]) followed by task-specific fine-tuning on reaction data. This approach aimed to encode generalizable chemical priors such as SMILES grammar, connectivity patterns, and functional group identities into the model's weights before learning the specific task of reaction prediction. In forward prediction on the USPTO-480k dataset, this strategy leads to incremental differences in top-1 accuracy from 91.1% (random initialization) to 91.8% (with pretraining),[164] but did reduce the amount of task-specific data and training time needed to reach competitive performance.

While earlier models focused on a single prediction task, T5Chem[165] pursued a unified architecture that could support multiple tasks as text-to-text problems using task-specific prompts (*e.g.*, "Product:", "Reactants:", "Reagents:") for forward prediction, retrosynthesis, reaction classification, reagent suggestion, and yield estimation as shown in Fig. 7). This design enabled the model to share representations across tasks while learning task-specific behavior. In T5Chem,[165] the model achieved a top-1 accuracy of 88.9% for forward reaction prediction on USPTO-480k.

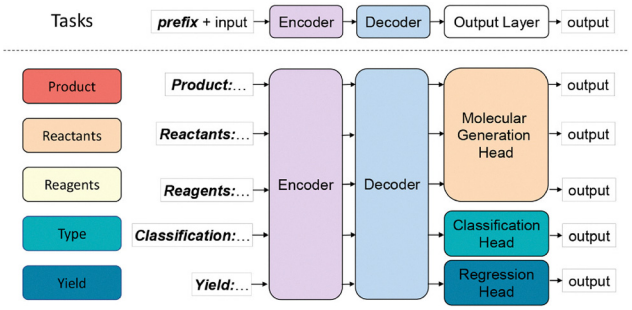


Fig. 7 Multitasking of T5Chem. Reproduced with permission from ref. 165. Copyright 2022 American Chemical Society.

Despite performing well on broad benchmark tasks and pushing accuracy incrementally higher, ranking alternatives remains a persistent challenge for sequence-based reaction prediction models. Model-assigned scores derived from the product of token-level probabilities do not necessarily correspond to chemical feasibility or mechanistic plausibility.[29,153,164] Several studies have observed that incorrect predictions often receive higher scores than chemically valid alternatives, reflecting a disconnect between statistical confidence and chemical correctness.[29,164] This gap indicates the need for better-calibrated scoring mechanisms to enable more reliable selection among plausible reaction products.

3.1.2.2 Adapting to data scarcity: transfer learning across chemical domains. As shown in Table 3, sequence generation models can achieve remarkably high top-1 accuracies (depending on the dataset and evaluation setting). However, a practitioner seeking to use such models for reaction outcome prediction may be disappointed by their real-world performance. As highlighted by Bradshaw *et al.*,[184] this gap is partly due to evaluation processes in which reactions are randomly divided between training and test sets. When split randomly, reactions linked to the same document or produced by the same author can end up in both the training and test sets. This overlap means the model may be evaluated on chemistry it has essentially already seen, with only minor differences, inflating reported performance. Deduplication of reactions only prevents the case where identical entries appear in train and test, yet near-identical entries are conceptually equivalent to data leakage. To investigate this effect, Bradshaw *et al.*[184] compared random reaction-level splits with more rigorous document- and author-level splits (Fig. 8), observing a substantial drop in top-*k* accuracy as the splitting criterion became more restrictive. In practice, when using a reaction predictor "in the real word," one is unlikely to encounter test reactions drawn from the same documents used for training, making document- or author-based splits more representative of real-world performance. These results show that models struggle to generalize to reactions dissimilar from those in the training set, and that a shift to a new reaction domain often exposes limitations driven by data scarcity.

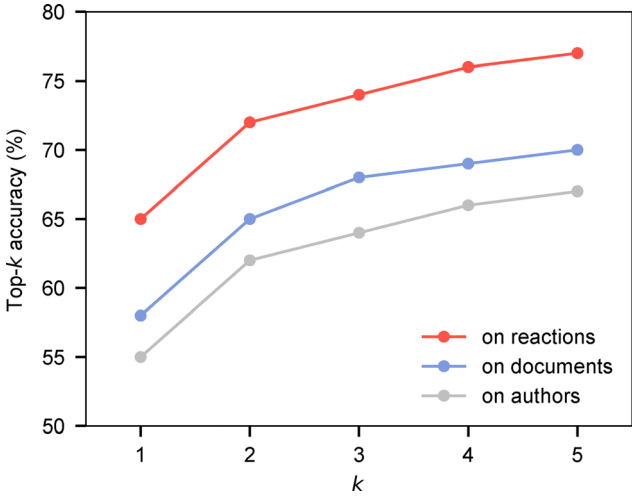


Fig. 8 Top-*k* accuracy of reaction prediction models under different data split strategies, replotted from data reported in Bradshaw *et al.*[184] "On reactions" indicates a random split over reactions, "on documents" splits by source document, and "on authors" holds out all reactions associated with specific authors.

One approach to mitigate these domain shift and data scarcity issues is transfer learning, where a model is first pretrained on a large corpus of general reactions and subsequently fine-tuned on a small set of domain-specific examples. This strategy can encode broad chemical knowledge during pretraining while adapting to the unique reactivity patterns of specialized domains during fine-tuning. Below, we summarize case studies illustrating how this paradigm enables accurate modeling in data-scarce reactions.

In carbohydrate chemistry, fine-tuning a Transformer pretrained on 370 000 general reactions with 25 000 domain-specific examples increased top-1 accuracy to 72.3% from no meaningful baseline.[185] For Heck reactions, the same pretraining raised accuracy from 66.3% to 94.9% using only 9959 examples.[186] In Baeyer–Villiger oxidations, accuracies improved from 64.7% to 83.8% (88.1% with SMILES augmentation) for one dataset[187] and from 58.4% to 81.8% (86.7% with augmentation) for another.[173] In an industrial setting, applying the Molecular Transformer to 147 392 reactions from Pfizer ELNs showed that training solely on internal data achieved 97.0% top-1 accuracy, compared to 82.9% top-3 when transferring from USPTO-Full.[188] These examples collectively show that pretraining on broad chemistry followed by domain-specific fine-tuning can recover or even exceed high accuracy in specialized chemical contexts with limited data. This ability to adapt pretrained sequence models to specialized chemical spaces-whether defined by functional group behavior, regioselectivity, or industrial context-suggests that their learned reactivity patterns are transferable and can be rapidly specialized in certain settings. A summary of these performance improvements is provided in Table 4.

3.1.2.3 Integrating structural information into sequence models. Despite the widespread adoption of SMILES as a representation language for organic molecules, its inherent ambiguity has long posed challenges for machine learning models.[160] As noted in the earlier section on SMILES augmentation and alignment, chemically identical molecules can have many different SMILES representations, depending on atom ordering. This syntactic variability limits the ability of models to directly associate structure with reactivity, often motivating the use of data augmentation or alignment schemes to increase robustness. While these strategies have shown practical value, they also increase training complexity and may introduce new sources of variability. An alternative strategy is to represent input molecules as molecular graphs, which naturally encode connectivity and are invariant to atom permutations. As discussed in Section 3.1.1, molecular graphs provide a direct and chemically intuitive representation of structure—arguably a more faithful encoding of how chemists themselves conceptualize molecules. Recent models have explored ways to incorporate such structural representations while retaining the generative strengths of autoregressive sequence decoders. These approaches aim to improve model fidelity to molecular structure, reduce reliance on augmentation, and more efficiently capture the inductive biases relevant to chemical reactivity.

Graph2SMILES[155] in Fig. 9 is a representative example of such a hybrid architecture. The model employs a graph encoder composed of a directed message passing neural network (D-MPNN) enriched with an attention mechanism and graph-aware positional embeddings. This encoder captures both local chemical environments and global molecular topology. A standard Transformer decoder generates product SMILES in an autoregressive manner. This architecture enables permutation-invariant input encoding without requiring SMILES augmentation. On the USPTO-480k dataset for forward reaction prediction, Graph2SMILES achieved a top-1 accuracy of 90.3% using canonical SMILES, closely matching the 90.6% accuracy of the Augmented Transformer[29] that relied on extensive SMILES augmentation. This shows that graph-based representations can offer comparable performance with fewer augmentation-related dependencies.

Table 4 Performance improvements *via* transfer learning (TL) across domain-specific reaction datasets. All models were pretrained on large general-purpose reaction datasets and fine-tuned on smaller, specialized subsets

Reaction class	Domain-specific data size	Pretraining dataset	Top-1 accuracy (scratch)	Top-1 accuracy (after TL)
Carbohydrate[185]	25 000	USPTO-480k	—[a]	72.3%
Heck[186]	9959	USPTO-480k	66.3%	94.9%
Baeyer–Villiger (Zhang)[187]	2254	USPTO-480k	64.7%	83.8% (88.1%[b])
Baeyer–Villiger (Wu)[173]	2254	USPTO-480k	58.4%	81.8% (86.7%[b])
Pfizer ELN[188]	147 392	USPTO-Full	82.9%[c]	97.0%

[a] The model trained from scratch did not yield meaningful predictions for carbohydrate reactions. [b] Accuracy after applying SMILES augmentation during fine-tuning. [c] Reported as top-3 accuracy when trained on USPTO and tested on Pfizer data.

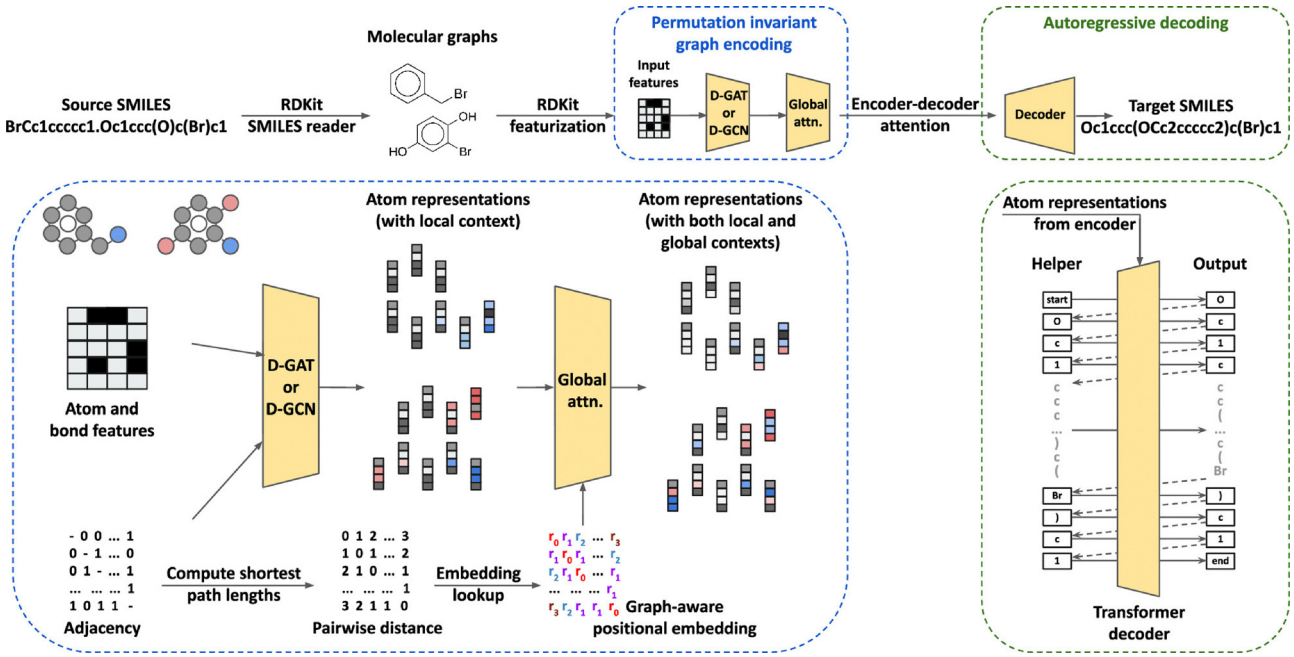


Fig. 9 Graph2SMILES model architecture. Reproduced with permission from ref. 155. Copyright 2022 American Chemical Society.

The value of this architectural shift has also been demonstrated in the context of mechanistic prediction. joungreproducing trained both standard Transformer and Graph2SMILES architectures on a large-scale dataset of over five million elementary reactions derived from expert-curated mechanistic templates. The performance gap between these two models on this dataset confirmed prior observations from ref. 155—namely, that graph-based encoders can improve performance even when outputs remain in SMILES format. Beyond performance, their results suggest a broader insight: it is not necessary to redesign the model architecture when moving from overall reaction prediction to mechanistic step prediction. Instead, this shift can be handled at the level of data preparation—by changing what the model is trained to predict, not how the model is built.

While Graph2SMILES focuses on replacing the input representation entirely with a graph, SeqAGraph[167] offers a complementary perspective: how to retain compatibility with SMILES-based data augmentation while still incorporating graph-based information. To achieve this, SeqAGraph includes an extra label for each graph input that specifies which atom was used as the starting point when generating the corresponding SMILES. This allows the model to distinguish between different randomized SMILES for the same molecule, while still treating the graph structure in a permutation-invariant way. As a result, standard sequence-based augmentation techniques can be applied without disrupting the benefits of graph-based encoding. Empirically, the model achieved a top-1 accuracy of 88.9% on USPTO-480k forward prediction, outperforming both Transformer and MPNN baselines and showing that simple additions to the input format can make graph-based models compatible with SMILES augmentation. Beyond SeqAGraph, BiG2S[168] extends the graph-to-sequence framework to both forward prediction and retrosynthesis within a shared encoder–decoder model, reporting a top-1 accuracy of 89.0% on USPTO-480k.

3.1.3 Limitations of data-driven models. While data-driven models have achieved high top-*k* accuracy in reaction prediction tasks, albeit mostly on random splits, there are certain limitations to both graph- and sequence-based methods.

Graph-based models are capable of encoding stereochemistry through explicit labels for features such as *cis*/*trans* isomerism, *E*/*Z* configurations, and *R*/*S* chirality. In practice, however, many graph-based reaction prediction models omit these annotations. As a result, stereochemical information is often not incorporated into the learned representation.

Sequence-based models built on SMILES (or related line notations) can handle the forms of stereochemistry that these notations explicitly encode, including *R*/*S* chirality (denoted using the "@" symbols) and *E*/*Z* configurations (represented by "\" and "/"). However, SMILES does not cover all stereochemical types, such as axial chirality (*e.g.*, BINOL derivatives), helical chirality, or certain forms of topological chirality. Consequently, sequence-based models can learn only the stereochemical distinctions that the SMILES representation itself defines and cannot capture other forms of stereochemistry outside this encoding. Although many of these limitations would, in principle, be resolved by using explicit three-dimensional coordinates, most data-driven architectures are designed around 2D graph or sequence inputs. Recent work has begun to explore machine learning architectures that explicitly incorporate three-dimensional information, including conformer-aware representations and equivariant neural networks that operate directly on molecular geometries.[189] Such approaches have shown promise in related problems where geometric and stereochemical effects are important, such as enantioselectivity prediction and bond dissociation energy estimation,[190] where

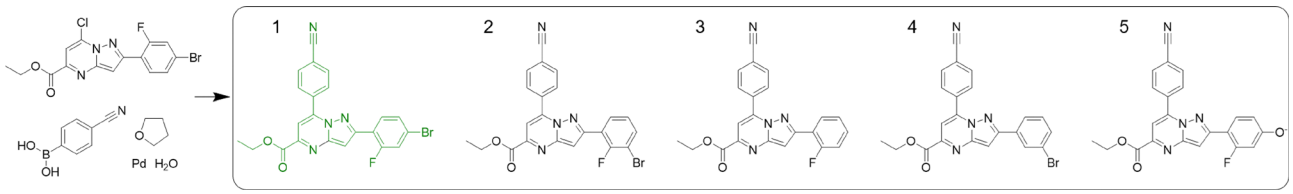


Fig. 10 Example failure mode of sequence generation models, where lower-ranked predictions are not chemically reasonable. Reproduced with permission from ref. 48.

conformer-level information can improve predictive accuracy. However, the value of incorporating three-dimensional structural information into large-scale reaction outcome prediction models remains uncertain, at least in terms of existing benchmarks, and is partially offset by the computational overhead imposed by conformer generation. As with analogous attempts to incorporate computed features into molecular representations,[191] any benefit of explicitly encoding 3D shape is likely to vanish as datasets grow larger.

A second limitation arises from the structure of the reaction datasets themselves. As discussed in Section 2.1, most large-scale experimental or literature-derived datasets provide only the reported major product for each reaction. Minor products, byproducts, and alternative regio- or stereochemical outcomes are rarely documented. As a result, models are not trained to reproduce the full distribution of experimentally formed species but rather to match the single outcome recorded in the dataset. This incomplete ground truth also means that lower-ranked predictions do not necessarily correspond to chemically plausible alternatives (*i.e.* minor products), since the underlying data do not contain information about the relative likelihoods of unreported products.

A further limitation is the widespread absence of reaction context. Key variables such as solvent, catalyst, stoichiometry, temperature, reaction time, concentration, or the order of reagent addition are missing or inconsistently recorded in most of the datasets summarized in Section 2.1. Because reaction outcomes can depend sensitively on these conditions, models trained only on reactant and product structures must infer reactivity patterns without access to the experimental factors that influence them. This lack of contextual information contributes to poor generalization under domain shifts, such as when evaluated on reactions drawn from different documents, authors, or chemical subdomains.

A further limitation of data-driven models is their tendency to hallucinate chemically implausible products.[48,181] Because data-driven models are trained to maximize statistical likelihood rather than to enforce chemical constraints, they may produce structures that are syntactically valid but violate basic principles of chemistry. Such hallucinations can be changes in the number of heavy atoms or hydrogens, violations of electron or valence conservation, incorrect stereochemical assignments, or unintended modifications to substituents peripheral to the true reaction center as shown in Fig. 10. Although several recent models incorporate explicit conservation constraints to prevent violations in atom or electron balance, these approaches do not eliminate hallucinations entirely.[49,56,175] The predicted structures may satisfy formal conservation laws yet still be chemically unreasonable, or they may be reasonable molecules that nonetheless could not arise from the given reactants under any feasible reaction mechanism.

3.2 Physics-based methods

Physics-based methods, in contrast to data-driven approaches, rely on fundamental principles of chemistry and physics to predict the most likely reaction outcomes. The core premise of these approaches is that by exploring the potential energy surface (PES) around the reactant (*e.g.*, with quantum chemistry or neural network surrogates), one can uncover the kinetically-accessible reaction pathways that correspond to the most likely elementary reactions (Fig. 11).[192,193] Physics-based approaches

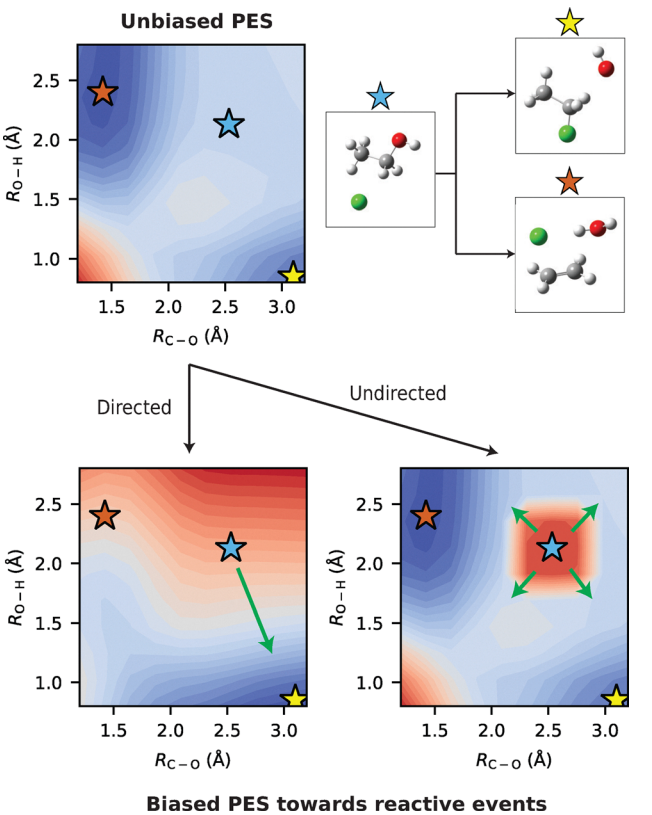


Fig. 11 Representative strategies for exploration of a potential energy surface (PES) for reaction outcomes. In the directed case, the PES is biased to drive the reactant towards a particular product. In the undirected case, the PES is biased to drive the reactant away from its initial conformational space thus driving it towards products that are close on the PES.

are valuable when physical models accurately recapitulate experimental observables as they can provide deep mechanistic insight into reactive systems. It is important to mention that although these methods (in both Sections 3.2.1 and 3.2.2) directly predict products, these methods tend to undergo post-processing to reevaluate kinetics and recalculate minimum energy paths *via* methods discussed in Section 4.2.2.

3.2.1 Directed potential energy surface exploration. Directed PES explorations take in an input reactant molecular geometry and a direction to explore, often defined as a reaction coordinate. Reaction coordinates are often defined by the vector corresponding to sets of atoms being pushed together or pulled apart. The output of directed PES explorations is one or more reaction pathways corresponding to a forcing function along the reaction coordinate. These reaction coordinates may be specified by the user or, perhaps more usefully, proposed algorithmically. Many directed PES exploration methods are built with algorithms to enumerate possible reaction coordinates which make them more applicable for single-ended outcome prediction. Their utility is limited for reactions that require consideration of multiple surfaces beyond the ground state.

An early example of this approach is the gradient extremal walking approach (GEWA) from Jørgensen *et al.*[194] In GEWA, the path connecting equilibrium structures is defined as the set of points where the gradient is minimized[195] thus following this path should yield the exact minimum energy path (MEP). Yielding the exact minimum energy path can be quite useful but is not necessary if the goal is solely to identify likely reaction products. An example of this is the anharmonic downward distortion following (ADDF) method from Ohno and Maeda.[196] Anharmonic downward distortions (ADDs) on a potential energy surface are deviations from the harmonic potential which are characteristic of reaction paths. ADDF searches for these ADDs by performing energy minimizations at various points along the harmonic potential surface to find downward bends in the PES. Selected ADDs are then followed and result in an approximate reactive path.

ADDF has demonstrated the ability to explore many different reacting systems[197] but was limited to unimolecular reactions before Maeda *et al.* extended the approach to bimolecular reactions with the artificial force induced reaction method (AFIR).[198–200] AFIR creates an artificial force between two sets of atoms and optimizes on a modified potential energy surface (which is the addition of the actual potential energy surface and the artificial force). If the force between the atoms is strong enough to overcome the barrier of the reaction, the optimization will result in an approximate reaction path. The force serves as a proxy for barrier height for the elementary step since high barrier steps will require a larger stimulus to induce the reactive event. By controlling the magnitude of the force, AFIR searches can be biased towards avoiding high barrier mechanistic steps that are kinetically less plausible. However, this means that multiple iterations of the same sets of atoms with different force constants may need to be run to extract all relevant pathways.

A different approach that conceptually achieves a similar goal without the need for iterative sampling is scanning along the reaction coordinate. Scanning involves forcing a simulation to follow a particular reaction coordinate through the whole corresponding elementary step. One method that implements this is Chemoton with its Newton trajectory scans.[201,202] In a Newton trajectory scan, atoms are forced together (as in AFIR) but the force is dynamically updated throughout the simulation to ensure that the atoms are bonded by the end of the simulation. Chemoton uses heuristic rules to automatically select the reactive sites (atoms/pairs of atoms) to push together. This represents part of a completely automated strategy to explore chemical reaction networks. A very similar approach developed by Zimmerman is the Single-Ended Growing String Method (SE-GSM).[62] This work evolved the Double-Ended Growing String Method (which will be discussed in a later section) from requiring a reactant and product conformer to only requiring a reactant conformer and a reaction coordinate. SE-GSM builds and optimizes a string along this reaction coordinate to generate the reaction path. Similar to Chemoton, ZStruct-2 is a mechanism exploration program that automatically provides the reaction coordinates to SE-GSM to drive the exploration of the reaction network.[203] The imposed activation (IACTA) approach from Lavigne *et al.*[204] expands upon the above methods by allowing reaction coordinates to be defined by bending and torsion angles as opposed to just interatomic distances. This approach is not self-driving, however, and thus requires user input of the reaction coordinate.

Perturbation of the vibrational modes of a molecule presents another method for exploring reactive events in a directed manner. This method, implemented in AutoMeKin,[205] is known as Transition State Search using Chemical Dynamics Simulations (TSSCDS).[206] To perform the search, a ''sizeable'' amount of energy is placed on the vibrational mode of interest during a molecular dynamics simulation. This stimulus biases reactive events which are then analyzed post-simulation for reaction paths. Users can input vibrational modes of interest into these simulations, but AutoMeKin also includes an option to automatically explore reactions with ChemKnow.[205]

3.2.2 Undirected potential energy surface exploration. Reaction outcomes can also be sampled on a PES by biasing systems towards reactive events in an undirected manner, *i.e.*, without predefining reaction coordinate(s) of interest. This can be done by performing reactive molecular dynamics, a subset of molecular dynamics specifically used to explore reactive transformations.

Despite being undirected, key to the implementation of this strategy is a means to bias the simulation in a manner that increases the rate at which reactions are observed. Local elevation metadynamics[207,208] adds an energetic penalty to molecular conformations that have already been visited during the simulation. This strategy is similar to the one used by bias potential driven dynamics,[209,210] where a Gaussian is added to the potential energy surface at the current geometry. In both local elevation and bias potential driven dynamics, molecules are therefore driven out of the local equilibrium wells that they would typically remain in if the simulation were unbiased. As a result, higher-energy configurations are sampled out of necessity

until configurations cross over reaction barriers into new local minima.

Rather than simply penalizing already-seen configurations, simulations can be run under other types of forcing conditions. One such condition is the inclusion of an electric field[211] as has been used to explore prebiotic Urey Miller chemistry.[211] Another condition is high temperature and high pressure as used in the original implementation of the *ab initio* nanoreactor.[212,213] Wang *et al.* performed molecular dynamics simulations at 2000 K with a simulated piston that compresses to encourage collisions between molecules. The nanoreactor has been used to explore Miller–Urey chemistry[213] and has also been applied to glycine synthesis,[213] nitromethane decomposition,[214] and phenyl radical oxidation chemistry.[215] Grimme combined bias potential driven dynamics and a wall potential to explore reactivity of Miller–Urey chemistry (as a common test case), ethyne oligomerization, and the oxidation of hydrocarbons, among others.[216]

The major advantage of such reactive simulations is that they do not require human input in deciding how to bias the search process. However, to successfully yield interesting reactive events, the reactants must exist in many replicates in the system and the simulation must be run for long enough to observe rare events, even with biasing. These simulations can be very computationally expensive when compared to directed approaches.

4 Two-step predictions of likely reaction products

Two-step predictions, in contrast to the direct prediction approaches, separate the process of identifying likely reaction products into distinct stages of defining (or sampling) the set of candidates and scoring them. This decoupled approach allows for a more flexible and potentially more accurate prediction by first generating a range of candidate products and then evaluating them based on different kinetic, thermodynamic, or other criteria. There are many approaches to each of these steps.

4.1 Defining candidate reaction products

The set of possible reaction products generated for a set of reactants must be comprehensive enough to include all relevant outcomes, but ideally not so broad that scoring them becomes intractable or infeasible. In the parlance of binary classification, we want to ensure that we have sufficient recall without completely sacrificing precision. Sampling candidate reaction products can therefore be approached with various levels of exhaustivity, from generic chemistry-agnostic enumeration to a focused, learned sampling.

4.1.1 Exhaustive or nearly-exhaustive enumeration. The most inclusive way to sample candidate reaction products is to consider "all" of the possible transformations allowed based on the reacting species. This approach ensures that no potential product is overlooked, providing a comprehensive set of candidates. This approach is typically applied in a mechanistic setting as opposed to a total reaction setting due to the comparatively smaller set of possible transformations. Exhaustive enumeration was most notably championed by Ugi in his formalization of the bond-electron matrix[217] and corresponding software IGOR (Interactive Generation of Organic Reactions).[218,219] IGOR generates reaction outcomes by modeling electron redistribution with reaction matrices that represent the flow of electrons between the reactants the products. These matrices can be enumerated in a brute force manner with transition tables to ensure that only allowable valence states are accessed. There are several other methods that utilize this formalism for the enumeration of products.[61,220,221] Other methods enumerate modifications to the bond connectivity matrix of the reactants.[222–224]

Perfectly exhaustive enumeration of all transformations is generally not tractable. In many elementary steps, only a small number of bonds are broken and formed which means enumerating graph edits with only a small number of bond rearrangements can allow graph-based search to be near exhaustive and tractable. This has led these methods to constrain enumeration to allow only up to 2 broken and 2 formed bonds ("b2f2"; Fig. 12). Many reaction types adhere to this constraint, yet some require consideration of more bond breaking/forming events like Diels-Alder, electrocyclizations, and Claisen rearrangements.[225] One area of weakness for graph-based enumeration is stereoisomerism. Since graph-based enumeration ignores stereochemical information, methods that employ this enumeration are not capable of explicitly exploring stereoselectivity without an additional subsequent enumeration of stereoisomers.

4.1.2 Rule-based enumeration. Human-guided or expert-curated rules are used to enumerate candidate products by applying explicit chemical logic rather than patterns derived from reaction data. These rules may be implemented as symbolic transformations, mechanistic operators, or structural heuristics, each defining how local bonding environments can change under specified conditions. In this section, we highlight two major classes of rule-based approaches: symbolic expert systems built from manually encoded logical rules, and mechanistic frameworks that model bond rearrangements using physicochemical reasoning.

4.1.2.1 Symbolic expert systems that use hand-crafted logical rules. Many of the earliest efforts to computationally enumerate chemical reactions were developed in the context of retrosynthetic analysis. These systems focused on deconstructing a target molecule into simpler precursors by applying manually

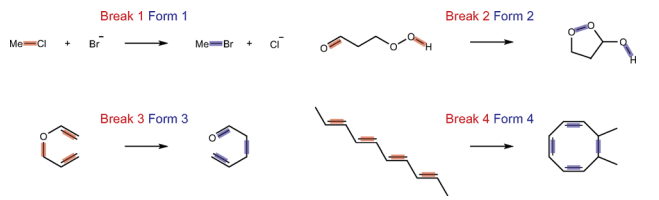


Fig. 12 Representative examples of reactions and their label in the "bnfn" formalism where "bn" indicates the number of bonds broken and "fn" indicates the number of bonds formed. Bonds labeled red are broken over the course of the reaction and bonds labeled blue are formed over the course of the reaction.

encoded rules that captured chemists' intuition about feasible bond disconnections. While this article's focus is on forward product prediction, these retrosynthetic systems laid the groundwork for rule-based approaches to candidate generation by formalizing chemical logic into a programmable decision framework.

This class of methods is often referred to as "symbolic" because the rules operate on discrete structural features—such as the presence or absence of certain substructures or bonding patterns—rather than relying on statistical associations or numerical optimization. Chemical transformations are defined in terms of human-readable logic (*e.g.*, "If a molecule contains a carbonyl adjacent to a leaving group, apply an elimination reaction"), and the system applies these rules conditionally based on the molecular structure at hand. Symbolic systems are inherently interpretable, and their behavior can be traced to explicitly defined rules.

Among the most influential examples, the LHASA system[226–230] pioneered the formalization of retrosynthetic planning as a set of modular transformations encoded in custom-designed languages. Using PATRAN to specify reactive patterns and CHMTRN to capture logical conditions, LHASA allowed chemists to express context such as steric hindrance, electronic effects, or unstable motifs in an executable form. At the level of individual transforms, these conditions primarily acted as binary applicability filters, determining whether a transformation could be invoked. SECS[231–234] shared this symbolic foundation but emphasized planning strategies and user interaction, combining transformation rules with heuristic search and a chemist-in-the-loop graphical interface. Both systems, though differing in implementation, demonstrated that codified chemical logic could meaningfully constrain the space of plausible transformations while preserving interpretability. Their impact extended into forward enumeration: the SAVI project[235,236] directly repurposed the LHASA infrastructure to generate synthetically accessible molecules at scale by applying predefined transformations to commercially available building blocks to enumerate billion-member libraries of synthetically-tractable products. More broadly, as in many earlier rule-based forward enumeration efforts predating SAVI, such systems are not necessarily followed by a secondary scoring step to re-rank the products they generate.

Building on earlier symbolic expert systems such as LHASA, AIPHOS (Artificial Intelligence for Planning and Handling Organic Synthesis) was an early example of applying structural pattern recognition to reaction prediction, using heuristic rules.[237] The system identified reactive centers by analyzing molecular substructures—such as functional groups, ring systems, and conjugated π-systems—and applied generalized patterns to propose plausible bond changes associated with substitution, addition, or elimination reactions. Unlike symbolic expert systems, AIPHOS did not require predefined, atom-mapped rules for each transformation; instead, it employed a pattern-based knowledge base to match local structural environments to known reaction types. Later developments further abstracted the concept of reactivity by encoding reaction sites as bit sequences that summarized the bonding and stereochemical features of atoms up to several bonds away.[238]

Rule-based enumeration underpins the Reaction Mechanism Generator (RMG).[239–241] Originally developed for combustion chemistry, RMG automates the construction of detailed kinetic models by systematically applying reaction families—predefined sets of transformation rules curated by chemists—to a given pool of reactants. Chemical species are represented as graphs, as described in Section 3.1.1, and elementary steps are generated by matching reactive substructures to family templates and applying the associated bond rearrangements. The initial set of reaction families, established by Green and co-workers, encoded mechanistic knowledge and kinetic data from the literature to capture key pyrolysis and oxidation chemistries. This foundation of roughly thirty reaction families has since been continuously expanded. These families are not derived automatically from datasets but are hand-crafted symbolic rules, with the scope of possible outcomes determined entirely by explicit chemical knowledge. By repeatedly applying these rules to expand the reaction network, RMG produces a comprehensive yet chemically constrained ensemble of plausible products. RMG has been broadly applied across diverse chemistries, demonstrating how symbolic expert systems can scale forward enumeration to thousands of reactions while maintaining transparency through explicitly encoded logic.[241] RMG was preceded by NetGen, a framework for automated network generation introduced by Broadbelt and co-workers.[242] In NetGen, molecules are similarly represented as graphs or matrices, and chemical transformations are encoded as reaction operators that define how substructures can change during a reaction. These operators, corresponding to reaction families such as oxidation, cleavage, or condensation, are predefined by chemists.

Building on this foundation, the Biochemical Network Integrated Computational Explorer (BNICE)[243–249] extends these formalisms to metabolic and biodegradation chemistry. Instead of general chemical operators, BNICE encodes generalized enzyme functions derived from the Enzyme Commission (EC) classification system, with each rule describing bond rearrangements catalyzed by a class of enzymes. Given an initial substrate, BNICE identifies functional groups and applies all relevant enzyme rules to generate new products, repeating this process iteratively to construct metabolic reaction networks. These rules are manually curated from biochemical databases such as KEGG[250] and University of Minnesota Biocatalysis/Biodegradation Database (UM-BBD).[251] In practice, BNICE has been used to reconstruct known metabolic routes, propose novel biodegradation pathways for xenobiotics, and analyze the combinatorial space of natural product biosynthesis.

Together, these symbolic expert systems demonstrate how hand-crafted logical rules provided early solutions for enumerating chemically reasonable reactions. Although largely superseded by data-driven methods, their modularity, transparency, and chemically grounded constraints established enduring design principles that continue to shape modern approaches to reaction prediction.

4.1.2.2 Mechanistic rules based on physicochemical reasoning. In contrast to symbolic systems that apply fixed transformation

rules based on structural patterns, mechanistic enumeration approaches aim to model chemical reactivity using principles from physical organic chemistry. These systems incorporate physicochemical reasoning—such as acid-base behavior, bond strengths, electronic structure, or thermodynamic constraints—to evaluate or prioritize candidate reactions. Rather than relying solely on predefined reaction templates, they simulate or infer plausible transformations based on the underlying properties of the reacting molecules. Mechanistic models attempt to explain not just what transformations might occur, but why they are chemically reasonable. Some systems use atom- and bond-level descriptors to compute local reactivity, such as pK_a values or bond dissociation energies (*e.g.*, CAMEO,[252–256] EROS,[257–259]) while others simulate bond-electron redistribution or apply quantum mechanical filters to reject implausible steps (*e.g.*, IGOR,[218,219,260] ROBIA,[261,262]). This class of approaches allows for broader generalization across chemical space while still grounding enumeration in interpretable chemical principles.

CAMEO (Computer-Assisted Mechanistic Evaluation of Organic Reactions) was among the first systems to generate reaction candidates through mechanistic reasoning rather than symbolic rule application.[252–256] It encodes mechanistic modules that mimic physical organic reasoning rather than fixed structural templates. After perceiving functional groups and structural features, the program assigns relative acidities (values) and nucleophilic or electrophilic sites, then applies predefined mechanistic steps such as proton transfer, substitution, elimination, or addition. Competing pathways, such as proton transfer *versus* organometallic addition or halogen–metal exchange, are resolved using heuristics based on relative base strength and leaving group ability, so that only chemically reasonable transformations are instantiated. The resulting products are then screened to exclude unstable or implausible structures, yielding a set of mechanistically-grounded reaction outcomes.

EROS (Elaboration of Reactions for Organic Synthesis) applies generalized reaction schemes known as RG*mn*, where *m* bonds are broken and *n* bonds are formed in a concerted rearrangement.[257–259] These schemes represent broad mechanistic motifs such as substitutions, eliminations, or pericyclic transformations, and are instantiated only when matching substructures are present. To control combinatorial growth, EROS employs physicochemical filters, including atomic charges, bond dissociation energies, or polarizability, estimated automatically by the PETRA (Parameter Estimation for the Treatment of Reactivity Applications) package. By combining formal reaction generators with descriptor-based constraints, EROS produces chemically plausible reaction outcomes that extend beyond predefined named transformations.

ROBIA (Reaction Outcomes by Informatics Analysis) similarly models transformations through stepwise mechanistic simulation rather than by substructure substitutions.[261,262] It identifies functional groups and reactive sites in the input molecules, then applies curated transformation scripts corresponding to mechanistic classes such as enolate formation, aldol condensation, or Diels–Alder cycloaddition. Each transformation generates intermediates and all possible stereoisomers, with unstable structures filtered *post hoc*.

SOPHIA (System for Organic Reaction Prediction by Heuristic Approach) extended the principles of AIPHOS by introducing a more systematic and generalizable framework for pattern-based reaction enumeration.[263,264] Rather than relying on predefined rules, SOPHIA constructed a reaction knowledge base by abstracting common transformation patterns from curated reaction examples. Each reaction center was encoded based on structural and bonding features—including atom types, bond orders, and neighboring environments up to the α, β, and γ positions—allowing the system to represent reactivity in a flexible, transferable format.

During enumeration, SOPHIA scanned the input molecule for substructures matching any of the abstracted patterns and applied the corresponding bond rearrangements to generate product candidates. These patterns were not exact subgraph matches but generalized structural motifs that captured common reactivity features. Unlike template-based methods (Section 4.1.3), SOPHIA did not require atom-mapped reactions or preservation of stereochemistry. Instead, it used structural plausibility and heuristic criteria to filter unreasonable candidates, enabling enumeration even in cases where no directly analogous transformation had been observed.

4.1.3 Template application using reaction data. As discussed in Section 4.1.2, early systems often relied on top-down rule construction, where a limited number of manually crafted disconnection rules were applied across many molecules. These approaches typically focused on changes at the reaction center, sometimes ignoring the surrounding chemical environment. Rule application often depended on expert-defined heuristics, such as fixed decision trees or prioritization rules, which were difficult to generalize beyond a narrow chemical domain. As a result, templates were either too general to be chemically meaningful or too specific to be broadly applicable.[239]

More recent approaches have shifted toward extracting transformation rules directly from reaction data.[265,266] These methods identify recurring structural changes from large collections of known reactions, allowing templates to be constructed algorithmically rather than encoded by hand. The underlying philosophy remains the same, which focuses on chemical transformations guided by rules or templates. Data-driven template-based enumeration provides a systematic way to explore chemically plausible outcomes while balancing reaction scope, structural specificity, and computational feasibility.[170,267]

Although this discussion focuses on forward prediction, template extraction is equally central to retrosynthetic analysis.[268] Methods ranging from systems with manually encoded rules to approaches that automatically extract retrosynthetic templates from reaction databases,[269] as well as data-driven approaches that identify minimal reaction cores,[265] SMARTS-based templates with careful treatment of stereochemistry,[266] and generalized transformation patterns[270,271] demonstrate that retrosynthetic template construction and application are mature techniques that share the same template-based strategies in the forward direction.

4.1.3.1 Forward reaction enumeration using template libraries. To construct a template library, a key challenge lies in how to

algorithmically define templates from reactions in a dataset. The reaction center corresponds to the substructures that undergo changes between reactants and products, and a template is obtained by describing the chemical transformation around this center. The extent to which the surrounding environment of the reaction center is included determines the balance between generality and specificity. Incorporating a larger neighborhood captures reactivity more accurately but limits applicability to cases where the exact substructure is present, whereas restricting the context yields templates that are more general but less chemically precise. This trade-off directly affects the scope of product candidates that can be enumerated.

The balance between generality and specificity in template definition can be illustrated by comparing different levels of representation Fig. 13. A highly specific template, such as those generated by RDChiral,[266] preserves surroundings of the reaction center, stereochemistry, and electronic states ensuring that the encoded transformation closely reflects the original chemical context. In contrast, local templates[270] focus only on the immediate connectivity changes at the reaction center, while generalized templates[170,267] abstract the transformation to broader symbolic patterns that can be applied across diverse chemical classes. As shown in Fig. 13, increasing specificity improves chemical fidelity but narrows the scope of applicability, whereas generalized templates expand coverage at the cost of reduced structural precision.

Once constructed, these templates can be systematically applied to reactants, where only those containing the required reaction center are activated. This procedure enumerates a diverse set of possible products, regardless of whether they are chemically plausible in practice.[272] Alternatively, template libraries can be restricted to a representative set of well-characterized reaction types, allowing them to generate expected products from given reactants in a more controlled manner.

4.1.3.2 Mechanistic dataset imputation. Templates can also be curated at the level of elementary steps rather than overall reactions; in particular, when applied to existing reaction datasets, such templates enable the imputation of mechanistic information from overall reactions. As a foundational step toward this goal, Chen and Baldi[273] developed more than 1500 elementary transformation rules, each annotated with electron flow specifications that explicitly describe how electrons are redistributed during a reaction. This work demonstrated how templates can represent mechanistic detail rather than only overall transformations.

Building on this idea, Joung *et al.*[48] showed that forward template enumeration can be applied in a constrained setting to impute mechanistic information from overall reactions. In this framework, a fixed sequence of mechanistically plausible transformation templates was applied to the reported reactants within a given reaction class, and only the path that reproduced the experimentally observed product was retained. This procedure enabled the construction of a mechanistic dataset in which each overall reaction was expanded into a plausible sequence of elementary steps. A subsequent study introduced the FlowER dataset,[49] which employed stricter constraints by explicitly enforcing mass, proton, and electron conservation during template-based mechanistic reconstruction. As noted in Section 2.1, overall reactions are widely documented, but mechanistic annotations are rare. These studies illustrate how forward template enumeration, when guided by class-level priors or conservation rules, can be used to derive mechanistic interpretations from existing reaction data.

Another recent example of forward template enumeration used for mechanistic reconstruction is the MechFinder framework introduced by Chen *et al.*[51] In this system, reaction templates are extracted from atom-mapped overall reactions and applied in the forward direction to generate intermediate structures along a proposed mechanistic path. A machine learning model is used to select the appropriate mechanistic template for each reaction, based on the local transformation pattern. Each mechanistic template specifies a predefined sequence of elementary steps, which is then deterministically applied to the reactants to yield a multi-step transformation matching the reported product.

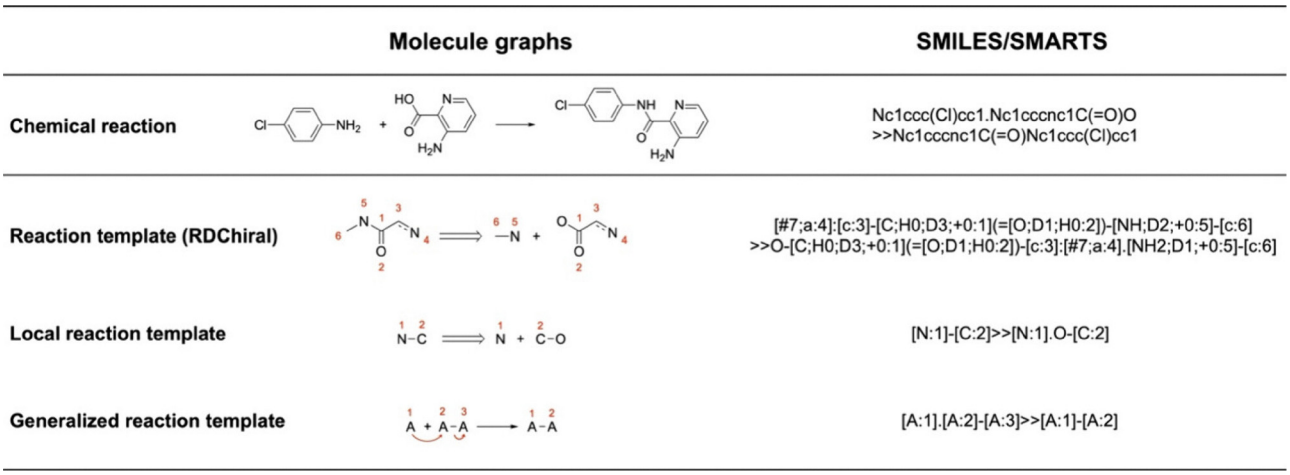


Fig. 13 Representative example illustrating how reaction templates can be extracted from a chemical reaction and expressed at increasingly abstract levels of representation. Reproduced with permission from ref. 267. Copyright 2024 American Chemical Society.

4.1.3.3 Template-guided reaction network exploration. One important application of a well-curated template library is the systematic exploration of reaction networks to uncover new chemical transformations. By repeatedly applying all available templates to a given set of reactants, one can generate a deep and wide network rooted in the initial reactants and populated with diverse intermediates and products. For example, Kluczynik *et al.*[59] curated a library of 489 mechanistic templates for carbocation rearrangements, informed by 715 solution-phase examples, and used these templates to systematically explore carbocation reaction networks. Starting from a carbocation input, templates for rearrangement, quenching, and related transformations were iteratively applied until either all carbocations were quenched or a predefined iteration limit was reached, generating a complex reaction network. This network was subsequently pruned using physical organic heuristics, and quantum chemical calculations were employed to estimate energies, enabling the calculation of rate constants and relative abundances. Through this workflow, the authors identified a novel tail-to-head terpene cyclization by systematic network exploration.

This idea can be further extended to the discovery of multicomponent reactions. Such reactions allow the construction of complex scaffolds from simple starting materials in a single step, and many of them have historically been uncovered largely by serendipity. By leveraging a mechanistic template library, it becomes possible to construct reaction networks in which three or more reactants participate simultaneously. When combined with the evaluation strategies discussed in the following Section 4.2, these networks can be systematically assessed to identify novel multicomponent reaction pathways.[58,274,275]

4.1.3.4 Template-based enumeration for site-selectivity. In the rule-based enumeration approaches discussed so far, the typical strategy involves applying all available transformation templates to exhaustively generate plausible reaction products. This is necessary when the underlying transformation itself is uncertain, requiring broad exploration across multiple reaction types and reactive sites. However, when the scope of the question is narrow and the transformation is known, such as in site-selectivity predictions—enumeration becomes simple. In these cases, the goal is not to identify which transformation might occur, but where it is likely to occur. Template application can therefore be restricted to a single transformation rule applied at multiple candidate sites.

Several studies have applied template-based enumeration to generate candidate outcomes for site-selectivity problems, where the transformation is fixed and the goal is to identify all chemically plausible application sites within a molecule. Tomberg *et al.*[276] used a template for electrophilic aromatic substitution and applied it exhaustively to all eligible positions on aromatic systems to enumerate possible substitution products. Guan *et al.*[277] followed a similar approach for nucleophilic aromatic substitution, systematically generating candidates by applying the S_NAr template across all reactive aryl halide sites. Hoque *et al.*[278] used template-based enumeration to identify all hydrogen atoms that could plausibly participate in radical abstraction reactions. These examples show how, when the scope of the question is narrow, enumeration can be reduced to the application of a single transformation rule across multiple candidate sites. Similar strategies have been adopted in a range of studies dealing with selectivity.

4.1.4 Reactivity rules learned from data. Machine learning-based enumeration provides a data-driven framework for generating plausible reaction outcomes that differs from traditional rule-based systems in how transformation patterns are specified and applied. Predefined templates or expert-curated transformation rules may suffer from limited coverage and run the risk of providing insufficient coverage, while exhaustive methods run the risk of being too inclusive. In contrast, data-driven approaches parameterize transformation rules from reaction datasets, learning statistical regularities in localized atomic and bonding changes and applying them to propose hypothetical products on new combinations of substrates. These methods typically operate by predicting localized atomic or bonding changes, from which full candidate products are constructed. By learning such transformation patterns directly from large-scale reaction data, ML-based enumeration can, in principle, expand the accessible chemical space beyond the reach of static rule sets.

Broadly, current ML-based enumeration methods fall into two categories, as illustrated in Fig. 14. The first approach focuses on identifying pairs of reactive atoms—often formulated as an electron-donating (source) and electron-accepting (sink)

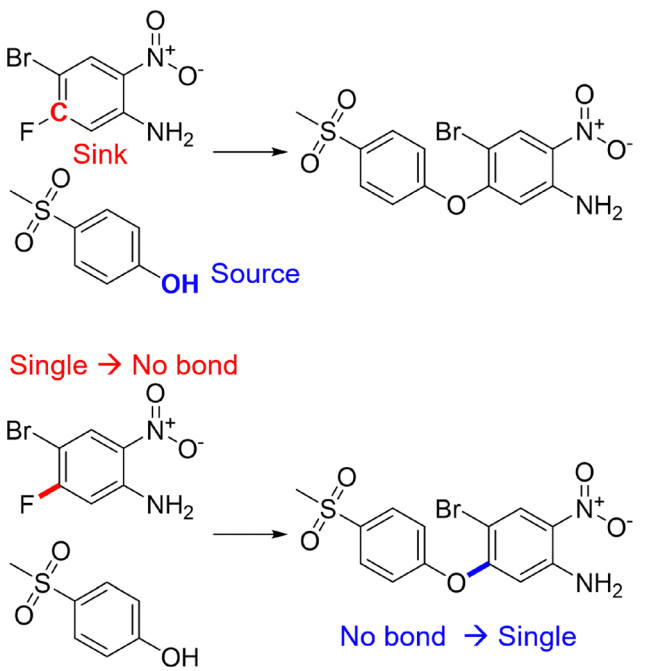


Fig. 14 Schematic illustration of two representative ML-based reaction enumeration formulations. In the top example, the model identifies a reactive source-sink atom pair, which implicitly defines the bond formation event. In the bottom example, the model directly predicts changes in bond order between atom pairs, demonstrated by a single-to-no-bond and no-bond-to-single transformation, without explicitly labeling source and sink roles.

sites in polar reaction mechanisms—based on their chemical environment and topological context.[22,56,279–284] While this formulation aligns naturally with polar reaction mechanisms, related extensions have adapted similar enumeration principles to radical or concerted reactions by modifying how reactive centers and bond changes are represented.[285,286] These reactive pairs are then enumerated to construct plausible mechanistic steps or reaction products. The second approach predicts, for each atom pair, the likelihood of a change in bond order after the reaction.[21,48,171] These bond-centric predictions are assembled into full product candidates under chemical constraints such as valence and connectivity. Despite differences in formulation, both strategies aim to sample a manageable set of candidate outcomes that are chemically plausible and consistent with learned reactivity patterns.

4.1.4.1 Prediction of candidate reacting atoms (source-sink pairs). Source-sink prediction methods aim to identify the specific pairs of atoms involved in electron flow during a reaction, namely, the electron-donating (source) and electron-accepting (sink) sites. These models typically operate in two stages: first, they assign reactivity scores to individual atoms (or orbitals), and then generate candidate transformations by enumerating high-scoring source–sink pairs. Depending on the formulation, the models may use hand-crafted chemical features, molecular orbital descriptors, graph-based neural architectures, or attention mechanisms to represent reactivity.

The Baldi group has developed a sequence of models that frame reaction prediction as a two-stage process centered on identifying reactive source and sink sites. The earliest of these, ReactionPredictor,[281,282] treats reactions as elementary mechanistic steps involving electron flow from a source to a sink atom or orbital. In the first stage, atom-level reactivity is predicted using classifiers trained on graph-topological and physicochemical descriptors, such as atom type, partial charge, and local bonding environment. The second stage performs exhaustive enumeration of all source-sink atom pairs to generate candidate mechanistic steps. The extended version of ReactionPredictor[282] supports not only polar but also radical and pericyclic mechanisms, and introduces a representation based on idealized molecular orbitals (MOs) such as lone pairs and π-bonds (donors), or empty orbitals and σ*-bonds (acceptors).

Later work by the group elevated orbitals from a conceptual enumeration unit to the primary learning objective, introducing orbital-level reactivity prediction to more explicitly reflect the electronic structure of molecules. In this model,[285] each atom is associated with a set of possible molecular orbitals (MOs)—such as lone pairs, π-bonds, or empty orbitals—and each MO is encoded using a tree-structured fingerprint that captures its local chemical environment, replacing hand-crafted atomic descriptors with a learned orbital-centered representation. A fully connected neural network is trained to assign reactivity scores to these orbitals using a balanced dataset of reactive and nonreactive examples. High-scoring source and sink orbitals are then paired to enumerate plausible elementary steps. While preserving the two-stage framework of ReactionPredictor, this approach replaces hand-crafted, heuristic-based atomic descriptors with learned representations that more directly reflect chemically meaningful orbital interactions.

A related line of work explored ML-based forward enumeration by leveraging learned reactivity models to expand reaction networks. Tavakoli *et al.*[287] applied a data-driven reactivity predictor to multi-step network expansion, where possible reaction pathways were enumerated from given reactants based on pairwise donor–acceptor combinations. In this framework, the model scores the likelihood of electron donation and acceptance at the atom level, and these scores are used to guide the systematic construction of reaction pathways.

The underlying reactivity model was introduced earlier by Tavakoli *et al.*,[284] who framed nucleophilicity and electrophilicity prediction as supervised learning tasks using methyl cation and anion affinities (MCA*, MAA*) as training targets. The model takes as input a combination of graph-topological and physicochemical atom-level descriptors and is implemented using graph neural network architectures, including Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs). These learned reactivity scores provide the quantitative basis for the subsequent enumeration and expansion of reaction networks.

More recently, the RMechRP model[286] introduced a contrastive learning framework for mechanism prediction in radical reactions. Rather than independently classifying individual atoms or orbitals, RMechRP directly learns to score pairs of orbitals by contrasting the true reactive pair against all other candidate pairs within the same reaction context. This formulation collapses reactive site identification and ranking into a single learning objective, avoiding an explicit two-stage enumeration-and-filtering pipeline. By scoring orbital pairs in their full molecular context, the model is able to prioritize chemically relevant interactions while maintaining consistency with radical reaction mechanisms.

Several models have adopted the idea of identifying reactive donor and acceptor sites as a basis for enumerating plausible mechanistic steps. These approaches differ in how they represent reactivity, the source of supervision, and the structure of the enumeration procedure. QC-RP[283] uses quantum chemical descriptors—such as condensed Fukui indices, atomic charges, and HOMO/LUMO energies—to identify likely reactive atoms, which are then paired to generate candidate steps. In contrast, ELECTRO[22] learns to generate entire sequences of source–sink transitions that trace a linear electron flow (LEF) from reactants to products. This path-based formulation mimics arrow-pushing diagrams but is trained on electron flow sequences heuristically extracted from atom-mapped LEF-type reactions, rather than on individual atom-level descriptors. Reactron[56] also constructs multi-step pathways iteratively, with broad coverage of reaction types and close adherence to well-curated ground truth mechanisms. One reactive step is predicted at a time, the molecular graph is updated, and enumeration continues until the product is reached. While these models vary in training data and representation—ranging from pairwise descriptors to path-based and graph-based updates—they share

the common goal of learning how electron redistribution events drive chemical transformations.

4.1.4.2 Prediction of candidate bond changes (graph edits). Bond change prediction approaches treat reaction enumeration as a graph editing problem, which is slightly decoupled from the notion of electron redistribution. These models predict how atomic connectivity changes by scoring possible bond-order modifications between atom pairs. A representative example is the Weisfeiler–Lehman Difference Network (WLDN),[21,171] a graph-convolutional neural network model that predicts the likelihood of bond order changes across all atom pairs, including the possibility of no bond change. Each bond is scored independently, and enumerated candidates are generated by selecting the top K bond changes (typically $K = 16$) and considering all combinations involving up to a fixed number of simultaneous edits, usually five. The total number of candidates is constrained by the binomial sum $\sum_{n=1}^{5} \binom{16}{n}$, significantly reducing computational cost while retaining broad coverage of plausible reaction outcomes. To ensure chemical validity, valence constraints and graph-theoretic rules are applied during enumeration. This formulation eliminates the need to assign explicit mechanistic roles to atoms, enabling scalable, template-free enumeration. However, because bond edits are predicted independently, product structures must be assembled *post hoc*, and key transformations may be missed if critical edits are misranked.

Joung *et al.*[48] adopt the same WLDN architecture but apply it to model individual elementary steps, expanding the output space to include changes in hydrogen count and formal charge. This additional complexity reflects the mechanistic nature of the data, which includes stepwise transformations and chemically valid intermediates. As a result, enumeration must consider combinations of bond, proton, and charge edits. The WLDN-based edit prediction thus becomes one component of a larger pipeline aimed at reproducing full mechanistic pathways rather than predicting only the final product.

4.2 Scoring candidate reaction products

Once a set of candidate products is generated or provided, the next step is to predict their relative likelihoods. For data-driven methods, this is most commonly done in a qualitative manner, for example, to at least identify the major product of the reaction. Physics-based methods are in principle more quantitative and more suitable for estimating values like reaction rates, but tend to be applied in qualitative way. In both cases, quantitative predictions of branching ratios, rates, yields, *etc.* tend to be reserved for settings with narrower domains of applicability where the number of potential products is limited (*e.g.*, in selectivity prediction).

4.2.1 Physics-based perspective on scoring. For elementary reactions, a physics-based approach involves calculating activation energies along the potential energy surface connecting reactants and products. Transition state theory, combined with the Arrhenius or Eyring equations, then provides rate constants that can be used to estimate product distributions. Because locating transition states is computationally demanding, thermodynamic quantities such as Gibbs free energies of reaction are often used as proxies for feasibility, identifying products likely to form spontaneously. While applying this thermodynamic perspective to multi-step mechanisms is less rigorous, it can still correlate with experimental outcomes. However, as reaction systems increase in complexity, identifying meaningful transition states and minimum energy paths becomes intractable, reducing the practicality of purely physics-based scoring and motivating more abstract or data-driven notions of product likelihood.

4.2.2 Kinetic evaluation. The most rigorous way to score candidate reaction products is through kinetic evaluation, where rate constants are derived from the Gibbs free energy of activation using transition state theory (Fig. 15). This approach provides a direct, physically grounded means of comparing product-forming pathways but is typically restricted to reactions comprising a single mechanistic step and practically limited to the types of reactions that can be accurately modeled with computational chemistry. The rate constants can be used to directly simulate the system to derive concentration profiles. These rate constants are calculated by evaluation of the transition state and subsequently the activation energy which can occur through systematic exploration of the potential energy surface or direct prediction by a data-driven model.

4.2.2.1 Kinetic simulation with rate constants. The primary method for tying physics-based computational insights directly to experimental results is by simulating the kinetics from the calculated parameters. Kinetic simulations require rate constants which can be estimated from the Gibbs free energy of activation using transition state theory (discussed in greater detail in Section 2.2.1). These simulations are performed by solving the differential equations that govern the balance of

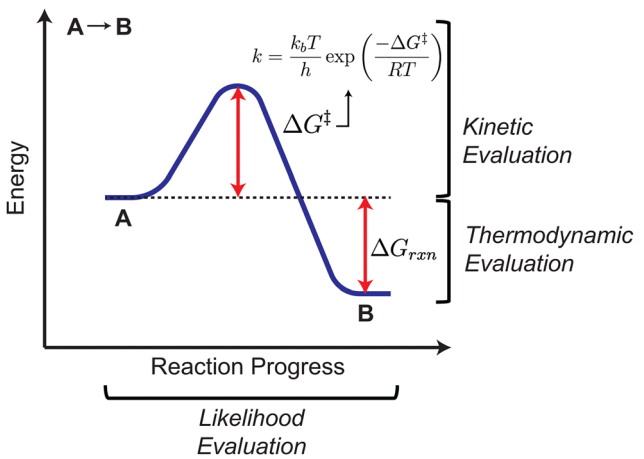


Fig. 15 There are a variety of ways to evaluate a chemical reaction. Kinetic evaluation is typically implemented by estimating the activation barrier of the reaction. Thermodynamic evaluation involves determining the difference in energy or enthalpy between the reactants and products. Likelihood evaluation uses heuristics and data to determine how "likely" a reaction is in an overall, abstract sense.

each species and yield concentration profiles for the species represented in the chemical reaction network (CRN). When using a fully explored CRN (*i.e.*, all kinetically relevant reactions are represented), direct predictions can be made about the experimental product yield. Even with a partial CRN, these concentration profiles can help prune kinetically irrelevant pathways and select species for further exploration. This strategy, referred to as flux-based screening, is common in physics-based exploration tools that emphasize deep exploration of CRNs like the reaction mechanism generator (RMG),[288,289] global reaction route mapping (GRRM),[290,291] Chemoton,[292,293] yet another reaction predictor (YARP),[294] and High-Performance Reaction Generation (HiPRGen).[295] Some of the ongoing challenges associated with this strategy are the combinatorial explosion of potential bimolecular reactions (which scales as the square of the number of species[289]), error propagation throughout the CRN due to imperfect estimation of rate constants, and the simulation cost of large CRNs, particularly if stiff. A couple of tools[293,294] have started to include uncertainty as a parameter to help mitigate the negative impacts of inaccurate rate constant on simulation results and prevent kinetically relevant pathways from being pruned. For CRNs that yield large, stiff systems of ordinary differential equations, methods like model lumping[296–298] and rate constant matrix contraction[290,299] have been developed to reduce the size of the system to reduce computational expense. Kinetic Monte Carlo[300,301] achieves a similar goal with stochastic sampling.[302]

4.2.2.2 Transition state searches. Estimating an activation energy involves optimizing the transition state of a given mechanistic step. To do this, it is necessary to identify the molecular conformation that lies on the saddle point of the potential energy surface spanning the reactant and product wells. In practice, this means establishing a ''guess'' conformation that is expected to be near the saddle point followed by a local optimization to the saddle point. These guess conformations are often manually enumerated by computational chemists; however, many algorithms exist to automatically establish this guess (see Table 5). These double-ended algorithms search for the minimum energy reaction path between an input reactant and product; the highest energy geometry along this reaction path is then used as an input to a subsequent local saddle point optimization. This is in contrast to the directed PES exploration methods discussed in Section 3.2.1 which were single-ended searches that moved an input reactant along a reaction coordinate for the purposes of revealing potential reaction outcomes. Workflows that explore mechanistic pathways will often use a double-ended algorithm followed by a saddle point optimization to find transition states and calculate activation barriers for scoring elementary steps.

One of the original algorithms to perform this task is the elastic band method.[306–309] The elastic band method places artificial springs between the geometries along the path and optimizes the entire path. The springs ensure that the geometries fall into their respective well allowing the geometries to span the whole reaction path. While this method can yield a valid reaction path, the highest energy geometry may not be near the transition state as the path may take a shortcut through a curved pathway due to the tension from the springs. This behavior resulted in the development of two of the most commonly used double-ended reaction path optimization methods: nudged elastic band (NEB)[311–313] and the string method.[319] NEB implements ''nudging'' to decouple the role of the geometry relaxations and springs in the elastic band thus ensuring that optimization to the global minimum will yield the MEP. A further improvement to NEB, climbing image nudged elastic band (CI-NEB), pushes a geometry on the reaction path to the highest energy point. This helps bring the highest energy geometry towards the transition state.

The string method represents a departure from the elastic band formulation. The string method defines and optimizes towards a string such that the energy gradient perpendicular to the path from a geometry relaxation is 0 along the reaction path (which is the definition of the MEP).[319] Intermittent reparameterization ensures geometries are equally spaced along the path. All of the previously discussed methods require the user

Table 5 Methods for identifying the minimum energy path between a reactant and a product

Year	Method	Description
1983	MaxFlux[303–305]	Finds pathway with maximum flux for diffusive dynamics assuming isotropic friction
1987	Elastic band[306–309]	Images along pathway are optimized with artificial springs between them
1994	Synchronous transit-guided quasi-Newton[310]	Linear/quadratic synchronous transit followed by an optimization with quasi-Newton or eigenvector following
1994	Nudged elastic band[311–313]	Elastic band with nudging mechanism to avoid corner cutting
1996	Variational reaction coordinate method[314–317]	Optimizes the line integral of the potential energy gradient norm
2000	Climbing image nudged elastic band[318]	Nudged elastic band where highest energy image is pushed to the top of the reaction path
2002	String method[319]	Minimize perpendicular energy gradient along a string between reactant and product
2004	Growing string method[62,320,321]	String method where the reactant and product propagate along the string toward each other
2004	Hamilton-Jacobi[322]	Hamilton-Jacobi equations are solved with the fast marching method to determine the MEP
2008	Spline Saddle[323,324]	Cubic splines representing the reaction path are optimized to the saddle point
2014	Image dependent pair potential interpolation[325]	Nudged elastic band with an atomic distance-based pair potential objective function, used to provide initial path guesses
2018	ReaDuct[326]	Spline saddle where the whole MEP is optimized
2019	Geodesic interpolation[327]	Optimize a geodesic in internal coordinates on the Morse potential, used to provide initial path guesses
2021	Energy weighted nudged elastic band[328]	Nudged elastic band where the springs are weighted by energy
2024	TS-Tools[329]	Constrained optimizations push the reactant geometry to the product geometry

to input a path between the reactant and product. If this input path (which often is a linear interpolation between the reactant and product geometries) is far away from the MEP, these algorithms may struggle to converge. The double ended growing string method (DE-GSM) addresses this by building the path between the reactant and product during the optimization.[62,320,321] The convergence to the MEP from both NEB and GSM have resulted in their widespread use to explore reaction paths.

After performing any of these reaction path optimizations, a transition state-like conformer can be extracted at the highest energy image along the path. Whether or not the applied method converges exactly to an MEP, it is often prudent to further refine this geometry to achieve a tighter convergence criteria. This can be done with a local saddle point optimization. Local saddle point optimizations follow the eigenvector associated with the largest absolute negative eigenvalue of the molecular hessian to the saddle point on the PES. This approach is shared amongst several popular methods like the Bernie algorithm,[330] rational function optimization (RFO)[331,332] and its successor restricted step partitioned rational function optimization (RSPRFO),[333] and the dimer method.[334,335] Due to these methods' reliance on the molecule's Hessian, convergence is quite slow which makes having an accurate guess for the transition state vital for convergence. This further motivates machine learning methods to generate transition state guesses discussed in the following section.

Once a saddle point is identified, it is important to identify whether that saddle point corresponds to the transition state of the mechanistic step of interest. Although this can be done qualitatively by identifying whether the negative frequency associated with the saddle point corresponds to the bond formation of interest, a more rigorous approach is to perform an intrinsic reaction coordination calculation.[336,337] This calculation establishes the reactant and product associated with the transition state by following the paths of steepest descent from the saddle point.

4.2.2.3 Transition state generation. Depending on the size of molecules and search method utilized, computing transition states can be quite expensive. To partially mitigate this cost, there have been efforts to directly predict transition state structures using machine learning. These methods learn to predict transition state structures from the simulated datasets discussed in Section 2.1.4. The fidelity of these predictions is not sufficient to use the structures as-generated, but they may serve as good initial guesses for subsequent saddle point optimization routines.

Early approaches formulated this task as a regression problem where the reactants and products are the inputs and the transition state geometry is the output. The reactants and products can be represented as 2D graphs as in Graph-2-Structure[338] but were more often represented as 3D geometries. Several architectures like kernel ridge regression (KRR) in Graph-2-Structure,[338] graph neural networks in TS-Gen,[339] tensor flow networks in TSNet,[340] and transformers in an unnamed method from Choi[341] were used as models for this task. An important attribute for each of these implementations is either equivariance or invariance to translation and rotation of the inputs (when the input is 3D). This is because translation and rotation of the input reactants/products geometries should either not change the output transition state coordinates (invariant) or change them in a reliable way (equivariant). This can be achieved with an equivariant architecture like the tensor flow networks in TSNet[340] or by predicting an invariant representation of the geometry like the distance matrix like in TSGen[339] and in Choi.[341]

More recent approaches for this task have used generative strategies to improve the accuracy of generated transitions states. A generative adversarial network (GAN) based approach was implemented in TS-GAN,[342] however significant performance enhancements were not seen until the introduction of diffusion. OA-ReactDiff[343] and TSDiff[344] train SE(3) equivariant graph neural networks to denoise from Gaussian noise to transition state structures. Both of these methods provide additional context to steer the diffusion towards the correct transition state. OA-ReactDiff uses a technique called "inpainting" where the reactant and product denoising trajectories are included in the model input whereas TSDiff includes a 2D graph representation of the reaction in its input. These methods provided state of the art results compared to previous methods, however diffusion models are more computationally expensive to inference than other generative approaches. To reduce computational expense, flow matching models for transition state generation were developed. Flow matching models learn the linear interpolation between a set of input data to a set of output data, the transition state structures for this task. The input data represents the reaction of interest; in React-OT[345,346] it is the reactant and product geometries, in GoFlow[347] it is the 2D condensed graph of reaction,[348] and in MolGEN[302] it is the 2D graphs of the reactant and product. These models represent the current state of the art for this task. For a more in-depth overview of transition state generation with machine learning, we point the readers to Beaglehole *et al.*[349]

4.2.2.4 Activation energy and rate constant prediction. The fastest way to evaluate the kinetics of an elementary step is by directly predicting activation energies and rate constants from the set of reactants and products. The subsequent discussion will explore machine learning methodologies that leverage datasets of quantum chemical calculations or experimental kinetics to predict these values thus bypassing costly transition state optimizations.

Similar to the previous section, many of these approaches use simulated datasets (Section 2.1.4) as training data. However, since these models directly predict experimental observables (activation energy/rate constant), they can also leverage experimental datasets. These datasets tend to be smaller than computational datasets but directly represent the ground truth value of interest. Most of the approaches discussed here predict the activation energy, but the activation energy and rate constant can be readily interconverted with the Eyring equation.

There has been much work to develop reaction input representations to learn reaction properties from data. A commonly employed 2D representation, the condensed graph of reaction

(CGR), superimposes the reactant and product 2D graphs to serve as input into a graph neural network.[348,350] A similar model, EquiReact, utilizes a graph neural network but instead takes 3D geometries of the reactant and product as input.[351,352] Benchmarking has shown that these input representations outperform baselines like reaction fingerprints[353] and that 3D methods outperform 2D methods when reaction mapping is not available.[354] A more in-depth review of the representations and architectures used for activation energy prediction can be found in De Landsheere *et al.*[355]

One strategy that has been explored for improving the performance of these models is providing them with extra information beyond the structure of the reactant and product. This information tends to be derived from the calculations that generate the activation energy training data like geometries and quantum mechanical descriptors of reactants, products, and transition states and has been shown to improve performance of these models, particularly in generalization to unseen compounds and in data-sparse regimes. One way to incorporate this information is directly provide it as input to the model. This has been employed with semi-empirical transition states,[356] semi-empirical descriptors,[357] DFT reaction energies,[358] and DFT orbital energies.[359] Although this can yield significant performance improvements, making predictions with these models is more computationally expensive since the additional information must be calculated for each prediction. To avoid this extra computational expense, models can be trained to jointly predict the extra information (descriptors, geometries, *etc.*) along with the activation energy in a multi-task learning or two-stage formulation. This approach is employed with QM descriptors in Stuyver and Coley[360] and with transition state geometries in Karwounopoulos *et al.*[361] and improves the generalization of the model in both cases.

There have been efforts to predict activation energies and rate constants from experimental datasets. A model with a CGR architecture has been successfully employed for the prediction of both E2[362] and S_N2[363] rate constants. However the size of these datasets (>1000 data points) is uncommon for most reaction classes so different approaches have been leveraged to make predictions in sparser data regimes. For example, as discussed in the previous paragraph, the strategy of providing additional information beyond the structure of the reactant and product in the model input has been used to improve predictions for nucleophilic aromatic substitution reactions. DFT features of the transition state[364] and DFT barrier heights[365] have served as inputs to models to improve accuracy. These models beat CGR baselines but degrade in performance when there is disagreement between DFT and experimental results. Another approach to training models on small experimental datasets is to train a model to predict parameters in a few-parameter empirical relationship. This can help prevent overfitting on minimal data as there are fewer parameters to learn and performance is grounded by the classical technique. Examples of this approach include learning the residuals of the Hammett model for S_N2 rate constants[366] and learning multivariate linear models based on QM and steric descriptors to predict inverse-electron demand Diels–Alder cycloaddition activation energies.[367]

4.2.3 Thermodynamic evaluation. Thermodynamic quantities such as Gibbs free energy and enthalpy provide insight into the spontaneity of a reaction and the relative stability of its products.[368] While a negative Gibbs free energy indicates thermodynamic favorability, it does not guarantee a rapid reaction or high yield. Nevertheless, free-energy calculations remain indispensable for evaluating reaction feasibility in large systems such as metabolic networks.[369–372]

Activation energy is often correlated with reaction enthalpy through linear free-energy relationships such as the Evans–Polanyi equation.[373] Although useful empirically, these relationships are not generally transferable and require prior knowledge of the reaction landscape. Reaction enthalpy itself can be obtained as the difference between the standard enthalpies of formation of products and reactants. One well-established approach for estimating the enthalpies of formation of arbitrary organic molecules are group additivity methods. Group additivity/group contribution methods provide a way to estimate enthalpy based on the functional groups present in a given molecule using experimental data. The first widely used group additivity model was the Benson group increment theory (BGIT).[374–376] This model accounts for atomic, bond, and group contributions to the heat of formation where groups account for the local environment surrounding an atom. The model's parameters are derived from experimentally calculated heats of formation. The Gronert model[377] extended BGIT by accounting for 1,2- and 1,3-interactions, improving accuracy for alkanes and trends in bond dissociation energies. More recently, the topology-automated component increment theory (TCIT)[378] has replaced empirical parameters with quantum chemically derived ones, enhancing transferability across diverse molecular classes.

Beyond these classical approaches, machine learning models and hybrid quantum–statistical frameworks have emerged for predicting both enthalpies and Gibbs free energies directly from molecular structure. There are many parallels between these modeling efforts and the activation energy prediction models discussed in Section 4.2.2.4. For example, simulated datasets (Section 2.1.4) are commonly used to train these models; however, since transition state energies aren't necessary, thermodynamic models can leverage more data like ground state quantum chemical calculations[379] or heats of formation.[380] Thermodynamic ML models also share model architectures[348,381–383] and augmentation strategies[384–386] with kinetic ML models. A common formulation for thermodynamic models is Δ-ML which means using an ML model to correct a less accurate calculation towards a more accurate calculation or experimental value.[384,385,387] This is accomplished by training a model to predict the difference between the less accurate method and a target value. This mirrors the input augmentation strategy discussed previously and allows researchers to produce higher accuracy thermodynamic predictions at a lower computational expense.[385,387–391] Together, these classical and data-driven strategies link thermodynamic evaluation to kinetic feasibility, offering complementary metrics for assessing reaction likelihood when explicit transition-state information is unavailable.

4.2.4 General "likelihood" evaluation. While kinetic (Section 4.2.2) and thermodynamic (Section 4.2.3) evaluations provide principled frameworks for assessing the feasibility and relative stability of reaction outcomes, obtaining precise activation energies or thermodynamic parameters for every possible candidate can often be computationally prohibitive or infeasible. Moreover, reaction outcome prediction that is not operating at the mechanistic level is not amenable to these approaches. As a practical alternative, general likelihood evaluation strategies have been developed to qualitatively assess whether candidate reaction outcomes are chemically reasonable and plausible based solely on the structural information of the given reactants and observed reaction products.

Broadly, methods for general likelihood evaluation can be grouped into three major categories: (1) binary classification of reaction feasibility, which rapidly evaluates the validity of reactant–product pairs in an absolute sense, one-at-a-time; (2) template-ranking approaches, which select and prioritize reaction templates to intrinsically guide outcome prediction; and (3) enumerated candidate ranking, which explicitly considers multiple potential outcomes and ranks their relative likelihoods. Each reflects a different modeling choice in how to assign or learn scores over candidate reaction outcomes, and they differ in terms of computational cost, interpretability, and the form of output they provide.

4.2.4.1 Binary classification of reaction feasibility. Feasibility classifiers determine whether a proposed reactant–product pair represents a chemically reasonable transformation. They are especially useful in large-scale synthesis planning, where rapid filtering prevents combinatorial explosion. Although they do not explicitly rank outcomes, their binary decisions can often be interpreted as assigning coarse-grained likelihood estimates.

A major challenge in training feasibility classifiers lies in obtaining negative examples—reactions that are chemically unreasonable or fail to occur—since failed or low-yield reactions are rarely reported in literature. As a result, most reaction datasets are heavily biased toward successful transformations. To construct balanced training data, studies such as the in-scope filter[392] and the fast filter[272] generated artificial negatives by applying extracted reaction templates to random sets of reactants (as discussed in Section 4.1.3. Each template, originally derived from a known successful reaction (A + B → C) was reapplied to produce hypothetical products (D, E, ...) not observed in the dataset. The known reaction yielding C was treated as a positive example, while all other hypothetical outcomes were labeled as negatives, enabling the classifier to distinguish feasible from infeasible transformations. To further mitigate data imbalance, Wollenhaupt *et al.*[393] later demonstrated that supplementing a large dataset with a small number of curated positive reactions, 14 500 rare cases from CAS collection, can substantially improve classifier performance on underrepresented reaction types without compromising general accuracy. An alternative perspective was introduced by Strieth-Kalthoff *et al.*,[394] who showed that feasibility classifiers can be effectively trained using only positive and unlabeled reaction data, without explicitly constructing negative examples. By framing reaction feasibility as a positive-unlabeled learning problem, their work demonstrated that reliable classifiers can be obtained despite strong reporting biases in the literature, challenging the assumption that large numbers of explicit negative reactions are required for training.

4.2.4.2 Multi-class classification of reaction templates. Template-ranking models estimate reaction likelihood by scoring transformation rules (Sections 4.1.2 and 4.1.3) prior to any product enumeration. Rather than evaluating complete reactant–product pairs or scoring enumerated candidates, these models assign probabilities directly over the space of applicable transformations, using only information from the input molecules. By shifting the decision point upstream, this strategy helps restrict the search space to plausible outcomes early in the prediction process, especially in workflows involving complex or multi-site reactions, where multiple products could arise from different reactive centers or competing mechanisms, enumerating and filtering all candidate products would be inefficient or ill-defined.

A representative implementation of this approach is the reaction type classifier introduced by Wei *et al.*,[395] in which a neural network is trained to predict the most likely reaction class from a set of 17 categories including a null class for cases where no reaction occurs. The input consists of concatenated fingerprints of reactants and reagents, and the model outputs a probability distribution over reaction types, each associated with a predefined SMARTS rule. Building on this idea, Segler and Waller[172] proposed a neural-symbolic framework that generalizes reaction-type classification to thousands of automatically extracted transformation rules. In this model, each reaction rule (*i.e.* template) is treated as a distinct class, and a neural network is trained to predict the probability distribution over these rules given the molecular fingerprints of the input reactants. Though these models are trained solely to recapitulate the template that matches what is recorded experimentally, these models implicitly learn which transformations are compatible with a given molecular context, prioritizing feasible rules while deprioritizing chemically inconsistent ones.

While template-ranking models effectively prioritize plausible transformations, predicting the most probable template does not uniquely determine the product structure. Many reactions, especially those involving multiple reactive sites or competing reaction centers, can yield several possible atom mappings that satisfy the same transformation rule, requiring subsequent modeling of selectivity to obtain unique product representations.

4.2.4.3 Candidate product ranking. The approaches discussed above evaluate whether a transformation is chemically feasible in absolute terms or prioritize among possible reaction templates before product generation. Candidate-ranking methods operate at a downstream stage, where multiple potential outcomes have already been enumerated and the goal is to determine which among them is most likely to occur. This shift from feasibility assessment to relative ranking reframes the

task from ''can this reaction happen?'' to ''which of these plausible outcomes most likely happened?''. Models in this class span a continuum from local, stepwise assessments of electron-transfer events to global scoring of complete reactant–product pairs.

At the most microscopic level, reaction likelihood can be evaluated in terms of a single electron-transfer, predicting the likelihood of interactions between electron-rich and electron-poor centers, often referred to as source-sink pairs. This framing naturally mirrors the conventions of arrow-pushing diagrams, where reaction mechanisms are represented as a sequence of localized electron movements. As a result, models that predict source–sink pairs step-by-step can be interpreted as approximating reaction mechanisms, with each pairwise decision corresponding to a plausible elementary event in a multistep pathway.[396]

Several models take the source–sink interaction as the fundamental unit of prediction, but differ in how they generate and evaluate candidate pairs. One strategy is to construct reaction pathways iteratively, predicting one pair at a time while deciding at each step whether the reaction should continue or terminate. This framework can be seen in ELECTRO,[22] which generates sequences of electron movements resembling arrow-pushing diagrams, although it does not explicitly score the plausibility of individual steps. Reactron,[56] in contrast, builds on this approach by associating each predicted electron move with a learned local compatibility score that guides sequential generation. Importantly, this score is used during the construction of the reaction pathway rather than to rank a set of pre-enumerated candidate outcomes. Rather than evaluating complete reactant-product pairs, Reactron updates the molecular structure after each predicted source-sink move, producing a sequence of intermediate-like structures that reflect the evolving electron flow. Although the model progresses through structures that include chemically meaningful intermediates—as these are present in the training mechanisms—it does not explicitly segment them as discrete steps, but instead models the reaction as a continuous chain of movements. Unlike graph-edit models such as MEGAN,[159] which generate and rank final reaction products using beam search over action sequences, Reactron does not produce scores for a set of multiple finished products, but instead predicts the evolution of a single reaction trajectory at a time.

A one-shot scoring strategy evaluates all candidate source–sink pairs in parallel, enabling direct comparison of alternative electron-transfer events based on local chemical context. Such models differ in how reactivity is represented, ranging from explicit quantum or physicochemical descriptors[283] to learned graph-based representations.[285–287]

Beyond electron-level modeling, candidate-ranking can also be defined at the level of complete reactant–product pairs. Here, reactivity is encoded as the structural difference between reactants and products. In the model of Coley *et al.*,[28] each candidate is represented by an edit vector that records bond formations, bond cleavages, and changes in hydrogen count; these vectors are processed by a neural network and converted to probabilities *via* softmax normalization. The Weisfeiler–Lehman Difference Network (WLDN)[21,171] generalizes this concept using a graph neural network that featurizes each atom by the difference between its reactant and product embeddings, improving expressiveness and accuracy by capturing both local edits and global context.

Difference-based and embedding-based strategies share the same underlying goal: assigning likelihoods to enumerated products by learning how structural changes correlate with known reactivity. Another approach is embedding-based methods learn a latent representation of the joint reactant–product pair that implicitly captures their transformation. These models use message-passing neural networks or attention-based encoders to process the full reaction context and assign a likelihood score based on the resulting features. For instance, LocalTransform[170] and its extension in Chen *et al.*[267] use GNNs to embed both the reactants and candidate products together and learn a compatibility function over the joint representation. Rather than focusing on structural deltas, these models learn to assess whether a given reactant–product combination is chemically coherent, modeling reactivity as a function of their overall configuration.

5 Quantitative prediction of other aspects of reaction outcomes

While the sections above focused on identifying what reactions occur and which products form, many important problems in reaction modeling concern how much they occur. In such cases, the space of possible outcomes, such as competing products or reactive sites, is already known. The task is to quantify their relative or absolute extents under specified conditions. Rather than discovering new transformations, these tasks model the distribution of outcomes within a known reaction type.

Within this category, two problems have received particular attention: selectivity and yield prediction. Both quantify the extent of reaction progress but differ in focus: selectivity describes the relative preference among multiple plausible pathways, while yield measures the overall efficiency of product formation. The following subsections review machine learning approaches to these two domains.

5.1 Selectivity prediction

Selectivity prediction operates under the premise that the plausible products or reactive sites are already known, and the key question becomes which of these sites will dominate under the reaction conditions. In other words, it is not about discovering new outcomes, but rather quantifying the extent to which each known pathway is favored. This makes selectivity prediction a fundamentally comparative problem—distinguishing between competing transition states or reaction channels—often modeled as a regression or ranking task.[397–399]

Selectivity encompasses multiple levels of preference, including chemoselectivity, regioselectivity, and stereoselectivity, each presenting distinct modeling challenges. While chemoselectivity overlaps with major product prediction, regio- and stereoselectivity require finer discrimination among closely related alternatives.

Because these differences often arise from subtle steric or electronic effects, many models incorporate physically interpretable descriptors, such as Sterimol parameters,[400] to complement learned molecular representations. The following subsections survey representative machine learning approaches to regioselectivity and stereoselectivity, illustrating how empirical observables and chemical priors are integrated into quantitative models of reaction outcomes.

5.1.1 Regioselectivity prediction. Regioselectivity prediction addresses the problem of identifying which specific site within a molecule is most likely to undergo transformation when multiple positions are chemically plausible.[401,402] This task is particularly critical in contexts such as C–H activation and substitution reactions, where even subtle differences in steric or electronic environments can influence the outcome. In contrast to general product prediction, which involves generating entire product structures, regioselectivity models typically operate at the level of predefined reaction sites, assigning each a quantitative score or rank that reflects its relative reactivity.

Regioselectivity prediction can be approached in two distinct ways, depending on how models define and evaluate reactivity at different sites within a molecule. Ranking-based approaches treat the task as a relative comparison among candidate positions. Given a predefined set of possible sites, these models assign a score, rank, or probability to each, selecting the most likely site of reaction based on the highest value. This formulation can be found in models trained on product annotations or experimental site distributions, as in cytochrome P450 metabolism,[403] nucleophilic aromatic substitution,[277] and aromatic C–H functionalization.[404,405] In many cases, these models output soft probability distributions over sites. For example, a model may assign 70% likelihood to one position and 20% to another.[404] This type of label can often be derived from existing reaction databases without the need for quantum mechanical calculations, making it suitable for large-scale or high-throughput settings.[406–409] It has been demonstrated that intentional dataset design with active learning can improve the performance of these models in data sparse regimes.[410]

In contrast, site-wise regression models are trained to predict an absolute reactivity-related value for each candidate site. These values are typically physical or thermodynamic quantities, such as bond dissociation energy (BDE), free energy barrier ($\Delta G^{\ddagger}$), or hydricity, which is the free energy required for heterolytic cleavage of a C–H bond to form a hydride ion (H^-).[277,411,412] Such models require a numerical label for each site, often obtained from quantum chemical calculations. Examples include predicting BDEs for sp^3 or sp^2 C–H bonds,[403,411,413–418] computing activation energies for radical substitutions,[411] or estimating hydricity values to assess hydrogen atom transfer reactivity.[419] These labels are more costly to generate than categorical site annotations, but they enable detailed mechanistic interpretation and generalization beyond training examples.

The modeling strategies used in ranking- and regression-based approaches differ both in architecture and in how transparently they reflect underlying chemical principles. Ranking models often rely on neural architectures such as message-passing networks or attention mechanisms that learn from molecular graphs or SMILES strings without explicitly defined input features.[404,420] While these models can achieve strong predictive performance, the source of their predictions is often difficult to interpret, as the learned representations do not correspond directly to chemically meaningful quantities.[397] In contrast, many site-wise regression models use descriptors grounded in physical organic chemistry, such as partial atomic charges,[421] Sterimol parameters,[405,421] Fukui indices,[422] or quantum-derived values like BDE[413–416,418] and orbital energies.[405,421] These descriptors have been employed in models that predict site selectivity in contexts such as C–H oxidation[421,422] and electrophilic aromatic substitution.[276,405,423] Because these input features are chemically interpretable and often mechanistically motivated, regression models are more amenable to rational analysis and *post hoc* interpretation than deep learning models trained end-to-end on structural data.[411]

Some recent studies combine these two perspectives into hybrid frameworks that aim to balance computational efficiency with mechanistic fidelity. These models use a ranking component to rapidly screen or prioritize candidate sites and apply regression or quantum calculations only to cases with uncertain or conflicting predictions. For example, Guan *et al.*[277] predicted site selectivity for nucleophilic aromatic substitution using a graph neural network, and triggered DFT calculations when the model's confidence was low. Similarly, Seumer *et al.*[412] employed a two-step strategy in which machine learning models first narrowed down reactive sites for Pd-catalyzed C–H activation, and quantum chemical methods were then used to evaluate the relative energies of plausible intermediates. A related philosophy also appears in descriptor-based studies that integrate empirical models with quantum refinement.[411,422,424] Such hybrid strategies offer a practical solution for workflows that demand both speed and reliability.

5.1.2 Stereoselectivity prediction. The task of predicting enantioselectivity is often framed as a regression problem, where the target is either the enantiomeric ratio (er) or the corresponding free energy difference ($\Delta\Delta G^{\ddagger}$) between competing enantio-determining transition states.[425] At room temperature (25 °C), a $\Delta\Delta G^{\ddagger}$ of approximately 1.36 kcal mol^{-1} corresponds to a 90 : 10 er, while a difference of around 6 kcal mol^{-1} leads to near-complete selectivity (99 : 1).[426] Because this metric is continuous, centered near zero for unselective reactions, and experimentally quantifiable with high precision, it provides a suitable target for model training in well-characterized reaction classes.[426] As in regioselectivity prediction, many models rely on quantum chemical descriptors and steric parameters.[397,427] However, the limited availability of large-scale datasets containing experimentally measured er or ee values has restricted the development of models based on learned molecular representations.[428]

Despite the increasing use of machine learning in modeling stereoselectivity, the availability of large-scale labeled data remains a fundamental limitation.[426,428,429] Experimentally determined er or ee values are labor-intensive to obtain and are rarely included in open-access reaction databases.[426] This scarcity is particularly problematic in asymmetric catalysis, where

the majority of reported reactions are biased toward high selectivity and favorable outcomes, leading to a skewed and incomplete representation of chemical space.[426] As a result, most datasets are restricted to narrow chemical domains, such as specific ligand scaffolds, transition metal complexes, or single catalytic motifs, limiting model generalizability across different reaction families.[430] Because of this locality and data sparsity, most current models adopt a descriptor-based framework tailored to specific reaction classes.[431] These models rely on physical organic descriptors or transition-state-derived features to achieve interpretability and predictive performance in constrained chemical spaces.[431,432] Attempts to build models using learned molecular representations remain limited due to data scarcity and concerns about extrapolation performance beyond the training domain.[428]

Most enantioselectivity prediction models adopt a regression framework and can be categorized based on how chemical information is represented and interpreted. A widely used strategy involves multivariate linear regression models built on steric and electronic descriptors, such as Charton values, Sterimol parameters, and classical linear free-energy descriptors.[427,433,434] These features are selected to reflect underlying physical organic principles and have been applied in early studies of asymmetric catalysis using multivariate regression approaches.[427,434] Because these models rely on human-selected, mechanistically interpretable inputs, they are well-suited for reaction classes with established transition state models and allow for direct rationalization of selectivity trends. In addition to these descriptor-based approaches, several studies have incorporated explicit three-dimensional steric environments derived from quantum chemical calculations into machine learning models for stereoselectivity prediction.[435]

Building on these descriptor-based approaches, several studies have explicitly integrated quantum chemical calculations with statistical or machine learning models to predict enantioselectivity. Early work by Huang *et al.*[436] used computed transition states to extract geometric and energetic features, which were then used in regression models. More recent studies have used a hybrid workflow that integrates computed steric maps or electronic descriptors with machine learning architectures to predict $\Delta\Delta G^{\ddagger}$ across multiple ligand–substrate combinations.[430,431,437–440] While some recent works have explored fully learned molecular representations, their applicability has remained limited due to the small size and chemical locality of existing datasets.[428]

Taken together, most high-performing models in this domain still rely on physical descriptors that can be mapped back to mechanistic reasoning. This allows the model to reveal how structural features, such as steric bulk or electronic effects positioned near the reactive site and in carbohydrate-based systems where structural conformation drives selectivity,[432,441] influence the formation of one enantiomer over the other. In contrast, deep models that learn molecular features end-to-end may offer greater representational flexibility but have seen more limited application in enantioselectivity prediction, largely due to the small size of available datasets.[397,428,442]

Most models for enantioselectivity prediction have been developed and evaluated within narrow chemical domains, typically focusing on a specific ligand scaffold, transition metal center, or reaction class. As a result, their predictive performance tends to degrade when applied to systems outside the original training distribution. This limited scope has motivated efforts to explore the transferability of structure–selectivity relationships across different catalytic contexts. Representative examples include the work of Reid and Sigman,[443] who demonstrated that a common descriptor space for bisphosphine ligands could be reused across multiple reaction types and metal centers. Related studies have explored broader condition spaces within a fixed reaction class,[25] or incorporated detailed mechanistic and quantum-chemical descriptors to model selectivity trends,[444,445] but generally remain constrained to singular catalytic systems. Their analysis showed that while model performance remained reaction-dependent, certain steric and electronic descriptors exhibited consistent correlations with selectivity across reaction classes.

5.2 Yield prediction

Yield is a practical metric used to assess the efficiency of a reaction in terms of the ratio of the amount of product obtained to the theoretical maximum, typically expressed as a percentage. High yields indicate reactions that proceed efficiently toward the desired product. As the most widely reported performance metric in experimental chemistry, yield has become a major target for data-driven modeling.

From a modeling perspective, yield prediction is typically framed as a regression problem, where descriptors of reactants, reagents, and conditions are concatenated and used as model input. Types of representations vary, with common choices being circular fingerprints or DFT-derived descriptors; graph-based encodings have been explored more recently as well. Efforts vary with respect to the types of datasets, models, and representations evaluated, and in terms of success, are a mixed bag of results, reflecting both the promise and the challenges of reliably predicting yield.

A common paradigm is to train a model on high-throughput experimental (HTE) data. Traditionally, these datasets represent the exhaustive exploration of a few substrate and reagent components for a single reaction type, and are attractive due to the use of automation for data generation, which allows for increased reaction thorughput, and reduces confounding factors by holding most procedural variables constant. Examples include two publicly-available Suzuki and Buchwald–Hartwig datasets, which are often used to benchmark yield prediction models.[25,446] Due to the combinatorial nature of this data, models are generally able to interpolate well. Even input features encoding no chemical information (*i.e.* one-hot encodings) often perform on par with chemically-rich features (*i.e.* structural/electronic information) on interpolation tasks.[447] However, the limited chemical diversity may hinder them from extrapolating to out-of-distribution components.[37,448–450] Nevertheless, as automation capabilities expand, an emergent trend has been the release of broader HTE datasets, covering multiple reaction types and a larger range of substrates,[43,451,452] or covering more industrially-relevant data,[408,453] potentially offering opportunities for more comprehensive model development.

Another common approach is to build a "global model" by training on large databases of published reactions spanning multiple reaction types and several diverse substrates (*i.e.* USPTO, CAS, Pistachio, Reaxys). While having such a model is undoubtedly useful in concept, it often sees little success in practice,[454] as models struggle to learn across multiple mechanisms and are hindered by implicit experimental differences, reactivity cliffs, and publication bias towards high-yielding reactions.[455] To at least minimize the effect of mechanistic changes, one may also consider building a model on literature data extracted for a single reaction type; however, it should be noted that model performance has generally been sub-optimal even in this setting as confounders remain.[456]

Yield is of course affected by numerous experimental factors such as concentrations, reaction scale, and the precise order and timing of addition, most of which goes unreported in commonly-used datasets.[457] Even when reaction procedures are shared with a special interest in reproducibility, outside of the context of data-driven modeling, it has been observed that a substantial fraction of reactions cannot be satisfactorily reproduced;[458,459] it is unsurprising then that models cannot predict yields under these circumstances. Modeling is further complicated by the skewed yield distributions in reported data, particularly in the literature where high-yielding examples dominate and negative data are scarce, limiting extrapolation to unfavorable substrates or conditions[460], whereas HTE datasets provide more balanced yield distributions that support robust prediction across the full yield range.[25,446] Reported yields may include both reactivity and purification, which is particularly true for datasets that combine information from multiple sources (*i.e.*, literature databases). Prediction of isolated yields is ill-posed as details of workup are not included in structured datasets.[461] Other performance metrics such as *enantio-* and regioselectivity are less subject to the aforementioned experimental differences. Thus, they have generally seen more success in modeling within specific reaction classes than yield, as discussed in the previous section.

It should be noted that these difficulties in modeling yield, which have often led to subpar model performance, do not necessarily mean that these models cannot be useful when applied to practical workflows. In discovery chemistry settings, for example, a binary classification yield prediction model (0 *vs.* $>0\%$ yield) is still very useful and can reduce the impact of dataset noise from variation in experimental conditions. Indeed, such classification models have been shown to be successful, both *in silico*[462] and in deployment for pharmaceutical discovery efforts.[463] Additionally, training on curated datasets that are more similar data to the desired task (*i.e.* medicinal relevance) has shown greater success in real-world extrapolation,[408,453] particularly when combined with active learning.[464]

6 Example applications of prototypical workflows

Several integrated workflows implement the concepts discussed above to perform end-to-end reaction prediction across diverse chemical domains. As many of these tools have already been introduced in earlier sections, this section briefly recapitulates their underlying design principles and highlights representative applications that demonstrate their utility.

6.1 Reaction mechanism generator

Reaction mechanism generator (RMG)[239–241] is an end-to-end workflow that proposes elementary steps to perform forward prediction. RMG utilizes a two-step prediction scheme where a rules-based enumeration (Section 4.1.2) is utilized to sample candidate products and a kinetic parameter estimation framework (Section 4.2.2) for scoring candidate products. The kinetic parameter estimation framework relies on a database of both experimental and computational kinetic and thermochemical parameters which allow RMG to calculate rate constants on the fly. These rate constants are then used in kinetic simulations to estimate species yields and identify dominant pathways.

RMG has been applied to numerous chemically complex systems. In modeling the steam cracking of *n*-hexane,[465] RMG generated a reaction network comprising 1178 elementary steps with calculated rate constants. The calculated rate constants were then used to perform kinetic modeling. The yields from the model matched well with experimental data from a pilot plant which allowed investigators to probe the reaction pathways that were kinetically relevant. Similar analyses have been performed for syngas production from bio-oil gasification,[466] combustion and pyrolysis of iso-butanol,[467] among many other chemical processes. RMG represents a sensible method choice in these applications since the elementary steps and their corresponding kinetics of these industrial processes are well understood. This means the template-based strategy can yield a robust chemical reaction network and the kinetic parameter estimation and kinetic simulations can yield meaningful concentration profiles.

6.2 Yet another reaction program

Yet another reaction program (YARP) is another end-to-end workflow that proposes elementary steps with a two-step scheme. Candidates are first generated through exhaustive graph-based enumeration (Section 4.1.1) and subsequently evaluated by computing activation barriers for each elementary step (Section 4.2.2). These barriers are calculated on the fly using semi-empirical GFN2-xTB and DFT methods to enable efficient evaluation across a wide chemical space. YARP also performs thermodynamic screening (Section 4.2.3) to reduce the number of high-level barrier calculations necessary for full network expansion.

A representative application is the exploration of the glucose pyrolysis reaction network.[468] YARP was used to explore over 31 000 elementary steps and was able to identify around 7000 kinetically relevant elementary steps (as defined by being below a barrier threshold of 45 kcal mol^{-1}). In this exploration, YARP was able to identify well-known reaction pathways in glucose pyrolysis as well as identifying new, low-barrier pathways to major experimental products. YARP has also been used for other complex reaction network explorations of prebiotic

ribonucleic acid formation from hydrogen cyanide[469] and Li-ion electrolyte degradation.[470] YARP constitutes a well-justified choice of method in these types of explorations due to the mechanistic uncertainty and emphasis on product retrieval associated with analyzing systems without well-established mechanisms. The exhaustive graph-based enumeration allows YARP to explore a broader set of potential mechanistic paths and the importance of retrieving products and pathways to these products (rather than accurate kinetic profiles like the previous example) justifies YARP's utilization of semi-empirical methods to accelerate kinetic evaluations.

6.3 Global reaction route mapping with AFIR

The global reaction route mapping (GRRM)[197] strategy for the artificial force induced reaction (AFIR)[198–200] method performs simultaneous prediction of products and likelihoods in an end-to-end fashion. AFIR conducts directed potential energy surface exploration (Section 3.2.1) by placing an artificial force between two sets of atoms to initiate reactive events. GRRM automatically selects sets of atoms to use for AFIR runs which ultimately yields the mechanistic network. A representative example of its use is the *in silico* reaction design of a multicomponent reaction with difluorocarbene.[471] In this study, GRRM used AFIR to build reaction networks for combinations of unsaturated compounds with difluorocarbene, followed by kinetic modeling based on computed barrier heights to identify the most favorable products. The cycloaddition product predicted *in silico* was subsequently confirmed experimentally. Additionally, GRRM with AFIR has been used in mechanistic explorations of the Passerini reaction,[472] Bignelli reaction,[473] $HCo(CO)_3$-catalyzed hydroformylation,[474] and more. Mechanistic exploration and imputation of synthetically relevant systems results in an emphasis on searching for unknown low barrier pathways which makes GRRM with AFIR particularly suitable for these applications. The undirected search performed by GRRM allows for the consideration of a wide range of potential mechanisms, and the artificial force used to explore the potential energy surface by AFIR helps to filter out kinetically irrelevant steps.

7 Conclusion

The prediction of chemical reaction outcomes remains one of the most fundamental challenges in chemistry, reflecting both our theoretical understanding of reactivity and our ability to translate that understanding into practical predictive tools. This review has surveyed the field's progress across both data-driven and physics-based paradigms, as well as hybrid systems such as RMG, YARP, and GRRM that integrate algorithmic enumeration with first-principles reaction modeling. Data-driven models have achieved remarkable performance in predicting reaction products and pathways by learning statistical distributions from large reaction corpora, at least according to existing benchmarks of in-distribution performance, while physics-based frameworks continue to provide interpretable mechanistic insights grounded in potential energy surfaces and kinetic theory. Together, these approaches illustrate the breadth of modern reactivity modeling.

Despite these advances, several challenges remain. The data-driven paradigm is often limited by the quality and representativeness of training data, which frequently omit reaction conditions, side products, or stereochemical details. As a result, learned models may generate statistically plausible yet chemically inconsistent outcomes, occasionally violating mass or electron conservation. Commenting on data availability being a limitation of data-driven methods is trite, but there is a deeper issue here. It is not merely that the field would benefit from more data, it is that the data that is currently collected into structured databases for the purposes of devising benchmarks are fundamentally incomplete. Reaction product prediction—in virtually every instantiation—is ill-posed. High-throughput experimentation within a single laboratory can mitigate confounding variables, but any dataset assembled from multiple literature sources is missing critical information about reaction conditions and procedures that we know to influence reaction outcomes. Some of that procedural richness exists in unstructured text in the original articles that describe a reaction, even if not in reaction databases we use, providing some hope that this information is already accessible to large language models. However, even when considering hypothetical models that learn "all" of the nuance of experimental procedures from published articles, one must acknowledge that challenges in reproducibility[458] raise important questions regarding what we can reasonably expect our models to learn.

On the other hand, physics-based methods are generally free of confounding variables and offer mechanistic rigor, but remain computationally demanding and scale poorly to complex systems. Even when substantial computational resources are devoted to them, the predictive accuracy of physics-based methods degrades for many systems of practical interest due to incomplete treatments of electron correlation, solvation, or anharmonicity.[475–477] With physics-based methods, the challenge is not that experimental datasets may omit important procedural details, but that these methods may not have a natural way to account for them; consider, for example, how a DFT model of reaction feasibility could account for the rate of addition. Moreover, these methods scale poorly with system size, making routine application to complex reactive networks infeasible. Recent advances in machine-learned interatomic potentials like smaller, more specialized models[478] and more efficient architectures[479] will help to address the problem of scalability. Addressing the problem of accuracy necessitates different types of contributions.

Looking ahead, progress in reaction outcome prediction will rely on continued advances within both data-driven and physics-based paradigms. Machine learning approaches are expected to benefit from more chemically faithful representations and curated datasets that capture negative outcomes (so-called "failed reactions"), richer reaction conditions, side products or impurities, and more. Such datasets are essential to more accurately reflect real experimental outcomes and to prevent models from overfitting to idealized or incomplete representations of chemical reactivity. Equally important is the transition from merely

predicting the identities of products to kinetics-informed models. Accurate predictions of quantitative reaction rates and product distributions would reflect a more complete understanding of reactivity and provide more utility than existing models. Predictive models that able to consider the role of reaction conditions and experimental protocols are a prerequisite for hypothetical future *in silico* condition optimization workflows.

For physics-based methods, improving the speed and accuracy of energetic oracles will be crucial for expanding the domain of applicability of these methods. Recent advances in machine-learned interatomic potentials like smaller, more specialized models[478] and more efficient architectures[479] will help to address the problem of scalability. The creation of the OMol25 dataset[147] highlights a broader trend of large-scale data generation to train foundation machine-learned interatomic potentials capable of providing the speed necessary to simulate increasingly complex systems. We expect that as the coverage of these datasets continues to improve, the resulting models will be capable of predictive simulations that provide actionable insights to experimental chemists, even if inheriting the limitations of its associated training data.[480,481]

Beyond purely technical advances, improving the usability and accessibility of computational tools will be essential to bridge the gap between methodological advances and routine adoption in experimental settings. Despite strong predictive performance in certain settings, reaction prediction models remain underutilized in practice, likely due to a combination of limited awareness, barriers to effective use, and a lack of confidence in model predictions. One often requires extensive experience to have (and often extensive patience to develop) an intuition for when a model prediction can be trusted, or when a seemingly-valid proposal actually reflects an idiosyncratic failure mode. Certainly, deep expertise is beneficial when understanding how to select an appropriate physics-based framework (*e.g.*, functional and basis set in DFT) for a given chemical system. In many cases, computational methods also remain difficult to install, configure, or integrate into existing workflows, particularly for researchers without formal training in programming. Improving the convenience of deploying models coming out of academic labs may require user-friendly interfaces and well-designed APIs or MCPs (*e.g.*, for tool-calling by large language models), as well as robust software engineering practices including maintenance, documentation, and validation. All of these usability concerns are rather easily addressed with modern coding agents, which may not yet have fully permeated into mainstream computational chemistry practices.

Another question that is fair to ask, critically, is for what purpose are these models actually useful? Among the many reaction outcome prediction tasks, there are several settings where existing predictive models have a clear role; *e.g.*, anticipating reaction success or failure in a parallel chemistry campaign to maximize the number of products for subsequent testing;[463] gaining confidence that a late-stage C–H functionalization will be selective before embarking on a long total synthesis or screening a precious intermediate;[410] constructing and parameterizing detailed kinetic models to embed in computational fluid dynamics simulations and estimate the effect of additives on ignition delay;[482] suggesting reasonable initial reaction conditions (*e.g.*, solvent, reagents, temperature) to accelerate experimental optimization workflows;[483] and screening synthetic routes for feasibility using a model trained on DFT calculations.[484] Yet there are many more hypothetical workflows that remain out of reach. Fully *in silico* reaction condition screening; a perfect oracle for reaction feasibility to guide retrosynthetic analysis; development of a "digital twin" reaction flask emulator for the chemistry lab with which to train laboratory agents; *de novo* catalyst design with reliable prediction of rates and turnovers; flowsheet modeling for pharmaceutical manufacturing that predicts heat and mass transfer-dependent product distributions; predicting reaction rates and time-dependent product distributions under realistic conditions, enabling quantitative control over selectivity and yield; autonomous exploration and discovery of complex reaction networks to identify previously unknown transformations. To address these more ambitious use cases, the next generation of reaction prediction models should aim to be able to generalize across disparate chemical domains, reason over reaction conditions, rank competing pathways, and quantitatively describe the dynamic evolution of reaction mixtures. Ultimately, these methods should enable the discovery of new, rather than the recapitulation of known, chemical reactivity.

Author contributions

J. F. J. drafted the discussion of data-driven methods; N. C. drafted the discussion of physics-based methods; P. R. drafted the discussion of the prediction of other aspects of reaction outcomes; C. W. C. drafted the discussion in the introduction and conclusion. All authors contributed to the revision of all sections.

Conflicts of interest

There are no conflicts to declare.

Data availability

No original code or data was developed in the preparation of this manuscript.

Acknowledgements

This work was supported by the NSF Center for Computer Assisted Synthesis (C-CAS) under Grant CHE-2202693 and the Machine Learning for Pharmaceutical Discovery and Synthesis consortium. J. F. J. additionally thanks the National Research Foundation of Korea (NRF) for grants funded by the Korean government (MIST) (no. RS-2025-16072756 and RS-2026-25484674). We thank Tim Pinkhassik and Zhengkai Tu for providing helpful comments on the manuscript.

Notes and references

1 E. J. Corey, *Angew. Chem., Int. Ed. Engl.*, 1991, **30**, 455–465.
2 Q. Zhu and C. Liu, *Pure Appl. Chem*, 2021, **93**, 1463–1472.
3 S. M. Roopan, A. Bharathi, J. Palaniraja, K. Anand and R. Gengan, *RSC Adv.*, 2015, **5**, 38640–38645.
4 Y. Jiang, M. Liu, M. Wang, Y. Lei, Q. Ding, H. Wu and X. Huang, *Org. Biomol. Chem.*, 2022, **20**, 7770–7775.
5 M. Cortes-Clerget, J. Yu, J. R. Kincaid, P. Walde, F. Gallou and B. H. Lipshutz, *Chem. Sci.*, 2021, **12**, 4237–4266.
6 M. Oelgemöller and N. Hoffmann, *Org. Biomol. Chem.*, 2016, **14**, 7392–7442.
7 R. I. Patel, S. Sharma and A. Sharma, *Org. Chem. Front.*, 2021, **8**, 3166–3200.
8 M. Latrache and N. Hoffmann, *Chem. Soc. Rev.*, 2021, **50**, 7418–7435.
9 L. F. Novaes, J. Liu, Y. Shen, L. Lu, J. M. Meinhardt and S. Lin, *Chem. Soc. Rev.*, 2021, **50**, 7941–8002.
10 B. Kaboudin, M. Behroozi and S. Sadighi, *RSC Adv.*, 2022, **12**, 30466–30479.
11 M. Regnier, C. Vega, D. I. Ioannou and T. Noël, *Chem. Soc. Rev.*, 2024, **53**, 10741–10760.
12 A. Y. Rulev and F. I. Zubkov, *Org. Biomol. Chem.*, 2022, **20**, 2320–2355.
13 P. Baldi, *J. Chem. Inf. Model.*, 2021, **62**, 2011–2014.
14 D. M. Lowe, PhD thesis, University of Cambridge, 2012.
15 D. Lowe, *Legacy reaction extraction data (1976–2013)*, 2020, https://figshare.com/articles/dataset/Legacy_reaction_extraction_data_1976-2013_/12084729.
16 S. M. Kearnes, M. R. Maser, M. Wleklinski, A. Kast, A. G. Doyle, S. D. Dreher, J. M. Hawkins, K. F. Jensen and C. W. Coley, *J. Am. Chem. Soc.*, 2021, **143**, 18820–18826.
17 NextMove Software, Pistachio, https://www.nextmovesoftware.com/pistachio.html, Accessed: 2020-11-19.
18 Reaxys, https://www.reaxys.com.
19 N. Schneider, N. Stiefl and G. A. Landrum, *J. Chem. Inf. Model.*, 2016, **56**, 2336–2346.
20 NameRxn, https://www.nextmovesoftware.com/namerxn.html.
21 W. Jin, C. Coley, R. Barzilay and T. Jaakkola, Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network, *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017.
22 J. Bradshaw, M. J. Kusner, B. Paige, M. H. Segler and J. M. Hernández-Lobato, *arXiv*, 2018, preprint, arXiv:1805.10970, DOI: 10.48550/arXiv.1805.10970.
23 P. Schwaller, T. Gaudin, D. Lanyi, C. Bekas and T. Laino, *Chem. Sci.*, 2018, **9**, 6091–6098.
24 H. Dai, C. Li, C. Coley, B. Dai and L. Song, Retrosynthesis Prediction with Conditional Graph Logic Network, *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, Canada, 2019.
25 D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher and A. G. Doyle, *Science*, 2018, **360**, 186–190.
26 D. Perera, J. W. Tucker, S. Brahmbhatt, C. J. Helal, A. Chong, W. Farrell, P. Richardson and N. W. Sach, *Science*, 2018, **359**, 429–434.
27 E. King-Smith, S. Berritt, L. Bernier, X. Hou, J. L. Klug-McLeod, J. Mustakis, N. W. Sach, J. W. Tucker, Q. Yang and R. M. Howard, *et al.*, *Nat. Chem.*, 2024, **16**, 633–643.
28 C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green and K. F. Jensen, *ACS Cent. Sci.*, 2017, **3**, 434–443.
29 I. V. Tetko, P. Karpov, R. Van Deursen and G. Godin, *Nat. Commun.*, 2020, **11**, 5575.
30 M. F. Lynch and P. Willett, *J. Chem. Inf. Comput. Sci.*, 1978, **18**, 154–159.
31 W. Jaworski, S. Szymkuć, B. Mikulak-Klucznik, K. Piecuch, T. Klucznik, M. Kaźmierowski, J. Rydzewski, A. Gambin and B. A. Grzybowski, *Nat. Commun.*, 2019, **10**, 1434.
32 P. Schwaller, B. Hoover, J.-L. Reymond, H. Strobelt and T. Laino, *Sci. Adv.*, 2021, **7**, eabe4166.
33 R. Nugmanov, N. Dyubankova, A. Gedich and J. K. Wegner, *J. Chem. Inf. Model.*, 2022, **62**, 3307–3315.
34 S. Chen, S. An, R. Babazade and Y. Jung, *Nat. Commun.*, 2024, **15**, 2250.
35 Z. Fu, X. Li, Z. Wang, Z. Li, X. Liu, X. Wu, J. Zhao, X. Ding, X. Wan and F. Zhong, *et al.*, *Org. Chem. Front.*, 2020, **7**, 2269–2277.
36 J. Götz, M. K. Jackl, C. Jindakun, A. N. Marziale, J. André, D. J. Gosling, C. Springer, M. Palmieri, M. Reck and A. Luneau, *et al.*, *Sci. Adv.*, 2023, **9**, eadj2314.
37 M. Saebi, B. Nan, J. E. Herr, J. Wahlers, Z. Guo, A. M. Zurański, T. Kogej, P.-O. Norrby, A. G. Doyle and N. V. Chawla, *et al.*, *Chem. Sci.*, 2023, **14**, 4997–5005.
38 M. Fitzner, G. Wuitschik, R. Koller, J.-M. Adam and T. Schindler, *ACS Omega*, 2023, **8**, 3017–3025.
39 A. Cook, R. Clément and S. G. Newman, *Nat. Protoc.*, 2021, **16**, 1152–1169.
40 M. Christensen, L. P. Yunker, F. Adedeji, F. Häse, L. M. Roch, T. Gensch, G. dos Passos Gomes, T. Zepel, M. S. Sigman and A. Aspuru-Guzik, *et al.*, *Commun. Chem.*, 2021, **4**, 112.
41 A. J. Rago, A. Vasilopoulos, A. W. Dombrowski and Y. Wang, *Org. Lett.*, 2022, **24**, 8487–8492.
42 N. J. Gesmundo, N. P. Tu, K. A. Sarris and Y. Wang, *ACS Med. Chem. Lett.*, 2023, **14**, 521–529.
43 H. Zhong, Y. Liu, H. Sun, Y. Liu, R. Zhang, B. Li, Y. Yang, Y. Huang, F. Yang, F. S. Mak, K. Foo, S. Lin, T. Yu, P. Wang and X. Wang, *Nat. Commun.*, 2025, **16**, 4522.
44 J. Ahlbrecht, M. D. Lutz, V. Jost, M. Farber, S. Brase and G. Wuitschik, *ACS Cent. Sci.*, 2026, **12**, 222–232.
45 J. Xie and W. L. Hase, *Science*, 2016, **352**, 32–33.
46 M. Tavakoli, Y. T. T. Chiu, P. Baldi, A. M. Carlton and D. Van Vranken, *J. Chem. Inf. Model.*, 2023, **63**, 1114–1123.
47 M. Tavakoli, R. J. Miller, M. C. Angel, M. A. Pfeiffer, E. S. Gutman, A. D. Mood, D. Van Vranken and P. Baldi, *J. Chem. Inf. Model.*, 2024, **64**, 1975–1983.
48 J. F. Joung, M. H. Fong, J. Roh, Z. Tu, J. Bradshaw and C. W. Coley, *Angew. Chem., Int. Ed.*, 2024, e202411296.
49 J. F. Joung, M. H. Fong, N. Casetti, J. P. Liles, N. S. Dassanayake and C. W. Coley, *Nature*, 2025, **645**, 115–123.
50 S. Stocker, G. Csanyi, K. Reuter and J. T. Margraf, *Nat. Commun.*, 2020, **11**, 5505.
51 S. Chen, R. Babazade, T. Kim, S. Han and Y. Jung, *Sci. Data*, 2024, **11**, 863.

52 Q. Zhao, S. M. Vaddadi, M. Woulfe, L. A. Ogunfowora, S. S. Garimella, O. Isayev and B. M. Savoie, *Sci. Data*, 2023, **10**, 145.
53 C. A. Grambow, L. Pattanaik and W. H. Green, *Sci. Data*, 2020, **7**, 137.
54 M. Schreiner, A. Bhowmik, T. Vegge, J. Busk and O. Winther, *Sci. Data*, 2022, **9**, 779.
55 S. Chen, 2.85M reaction mechanisms from the USPTO reaction dataset, 2025, https://figshare.com/articles/dataset/Reaction_mechanisms_for_training_Reactron/28398056.
56 S. Chen, K. S. Park, T. Kim, S. Han and Y. Jung, *arXiv*, 2025, preprint, arXiv:2503.10197, DOI: 10.48550/arXiv.2503.10197.
57 B. Mahjour, Y. Shen, W. Liu and T. Cernak, *Nature*, 2020, **580**, 71–75.
58 R. Roszak, L. Gadina, A. Wołos, A. Makkawi, B. Mikulak-Klucznik, Y. Bilgi, K. Molga, P. Gołebiowska, O. Popik and T. Klucznik, *et al.*, *Nat. Commun.*, 2024, **15**, 10285.
59 T. Klucznik, L.-D. Syntrivanis, S. Baś, B. Mikulak-Klucznik, M. Moskal, S. Szymkuć, J. Mlynarski, L. Gadina, W. Beker and M. D. Burke, *et al.*, *Nature*, 2024, **625**, 508–515.
60 J. T. Margraf and K. Reuter, *ACS Omega*, 2019, **4**, 3370–3379.
61 Q. Zhao and B. M. Savoie, *Nat. Comput. Sci.*, 2021, **1**, 479–490.
62 P. M. Zimmerman, *J. Chem. Phys.*, 2013, **138**, 184102.
63 P. M. Zimmerman, *J. Comput. Chem.*, 2015, **36**, 601–611.
64 H. Eyring, *J. Chem. Phys.*, 1935, **3**, 107–115.
65 D. I. Sverdlik and G. W. Koeppl, *Chem. Phys. Lett.*, 1978, **59**, 449–453.
66 D. J. Tantillo, *Advances in Physical Organic Chemistry*, Academic Press, 2021, vol. 55, pp. 1–16.
67 Y. Oyola and D. A. Singleton, *J. Am. Chem. Soc.*, 2009, **131**, 3130–3131.
68 D. G. Truhlar and B. C. Garrett, *Annu. Rev. Phys. Chem.*, 1984, **35**, 159–189.
69 T. Yu, J. Zheng and D. G. Truhlar, *J. Phys. Chem. A*, 2012, **116**, 297–308.
70 T. Yu, J. Zheng and D. G. Truhlar, *Chem. Sci.*, 2011, **2**, 2199–2213.
71 W. F. Van Gunsteren, R. M. Brunne, P. Gros, R. Van Schaik, C. A. Schiffer and A. E. Torda, *Methods in Enzymology*, Academic Press, 1994, vol. 239, pp. 619–654.
72 O.-E. Ganea, L. Pattanaik, C. W. Coley, R. Barzilay, K. F. Jensen, W. H. Green and T. S. Jaakkola, GeoMol: Torsional Geometric Generation of Molecular 3D Conformer Ensembles, *arXiv*, 2021, preprint, arXiv:2106.07802 [physics], DOI: 10.48550/arXiv.2106.07802, https://arxiv.org/abs/2106.07802.
73 A. E. Cleves and A. N. Jain, *J. Comput. Aided Mol. Des.*, 2017, **31**, 419–439.
74 D. Sindhikara, S. A. Spronk, T. Day, K. Borrelli, D. L. Cheney and S. L. Posy, *J. Chem. Inf. Model.*, 2017, **57**, 1881–1894.
75 N.-O. Friedrich, F. Flachsenberg, A. Meyder, K. Sommer, J. Kirchmair and M. Rarey, *J. Chem. Inf. Model.*, 2019, **59**, 731–742.
76 T. Seidel, C. Permann, O. Wieder, S. M. Kohlbacher and T. Langer, *J. Chem. Inf. Model.*, 2023, **63**, 5549–5570.
77 M. J. Vainio and M. S. Johnson, *J. Chem. Inf. Model.*, 2007, **47**, 2462–2474.
78 S. Riniker and G. A. Landrum, *J. Chem. Inf. Model.*, 2015, **55**, 2562–2574.
79 Macrocycle Conformations—Applications, https://docs.eyesopen.com/applications/omega/theory/macrocycle_theory.html.
80 K. S. Watts, P. Dalal, A. J. Tebben, D. L. Cheney and J. C. Shelley, *J. Chem. Inf. Model.*, 2014, **54**, 2680–2696.
81 S. Wang, J. Witek, G. A. Landrum and S. Riniker, *J. Chem. Inf. Model.*, 2020, **60**, 2044–2058.
82 P. Labute, *J. Chem. Inf. Model.*, 2010, **50**, 792–800.
83 C. Shi, S. Luo, M. Xu and J. Tang, Learning Gradient Fields for Molecular Conformation Generation, *arXiv*, 2021, preprint, arXiv:2105.03902 [cs], DOI: 10.48550/arXiv.2105.03902, https://arxiv.org/abs/2105.03902.
84 J. Zhu, Y. Xia, C. Liu, L. Wu, S. Xie, Y. Wang, T. Wang, T. Qin, W. Zhou, H. Li, H. Liu and T.-Y. Liu, Direct Molecular Conformation Generation, *arXiv*, 2022, preprint, arXiv:2202.01356 [cs], DOI: 10.48550/arXiv.2202.01356, https://arxiv.org/abs/2202.01356.
85 M. Xu, L. Yu, Y. Song, C. Shi, S. Ermon and J. Tang, GeoDiff: a Geometric Diffusion Model for Molecular Conformation Generation, *arXiv*, 2022, preprint, arXiv:2203.02923 [cs], DOI: 10.48550/arXiv.2203.02923, https://arxiv.org/abs/2203.02923.
86 B. Jing, G. Corso, J. Chang, R. Barzilay and T. Jaakkola, Torsional Diffusion for Molecular Conformer Generation, *arXiv*, 2023, preprint, arXiv:2206.01729 [physics], DOI: 10.48550/arXiv.2206.01729, https://arxiv.org/abs/2206.01729.
87 R. Jiang, T. Gogineni, J. Kammeraad, Y. He, A. Tewari and P. M. Zimmerman, *J. Comput. Chem.*, 2022, **43**, 1880–1886.
88 A. Volokhova, M. Koziarski, A. Hernández-García, C.-H. Liu, S. Miret, P. Lemos, L. Thiede, Z. Yan, A. Aspuru-Guzik and Y. Bengio, *Digit. Discovery*, 2024, **3**, 1038–1047.
89 Z. Wang, H. Zhong, J. Zhang, P. Pan, D. Wang, H. Liu, X. Yao, T. Hou and Y. Kang, *J. Chem. Inf. Model.*, 2023, **63**, 6525–6536.
90 W. J. Hehre, L. Radom, P. von and J. Pople, in *Ab initio molecular orbital theory*, ed. W. J. Hehre, Wiley, New York, 1986.
91 J. B. Foresman and A. Frisch, *Exploring chemistry with electronic structure methods*, Gaussian, Inc, Wallingford, CT USA, 3rd edn, 2015.
92 V. Butera, *Phys. Chem. Chem. Phys.*, 2024, **26**, 7950–7970.
93 P. Geerlings, F. De Proft and W. Langenaeker, *Chem. Rev.*, 2003, **103**, 1793–1874.
94 A. D. Becke, *J. Chem. Phys.*, 2014, **140**, 18A301.
95 M. Bursch, J.-M. Mewes, A. Hansen and S. Grimme, *Angew. Chem., Int. Ed.*, 2022, **61**, e202205735.
96 K. Burke, *J. Chem. Phys*, 2012, **136**, 150901.
97 H. Hayashi, S. Maeda and T. Mita, *Chem. Sci.*, 2023, **14**, 11601–11616.
98 K. Burke and L. O. Wagner, *Int. J. Quantum Chem.*, 2013, **113**, 96–101.
99 P. Morgante and R. Peverati, *Int. J. Quantum Chem.*, 2020, **120**, e26332.

100 N. Mardirossian and M. Head-Gordon, *Mol. Phys.*, 2017, **115**, 2315–2372.
101 N. Casetti, J. E. Alfonso-Ramos, C. W. Coley and T. Stuyver, *Chem. – Eur. J.*, 2023, **29**, e202301957.
102 T. R. Nelson, A. J. White, J. A. Bjorgaard, A. E. Sifain, Y. Zhang, B. Nebgen, S. Fernandez-Alberti, D. Mozyrsky, A. E. Roitberg and S. Tretiak, *Chem. Rev.*, 2020, **120**, 2215–2287.
103 H. Lischka, D. Nachtigallova, A. J. Aquino, P. G. Szalay, F. Plasser, F. B. Machado and M. Barbatti, *Chem. Rev.*, 2018, **118**, 7293–7361.
104 K. D. Vogiatzis, M. V. Polynski, J. K. Kirkland, J. Townsend, A. Hashemi, C. Liu and E. A. Pidko, *Chem. Rev.*, 2018, **119**, 2453–2523.
105 M. R. Dooley and S. Vyas, *Phys. Chem. Chem. Phys.*, 2025, **27**, 6867–6874.
106 S. M. Bachrach, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2014, **4**, 482–487.
107 J. C. Slater, *Phys. Rev.*, 1951, **81**, 385–390.
108 N. C. Handy, *Chem. Phys. Lett.*, 1980, **74**, 280–283.
109 P. E. M. Siegbahn, *Chem. Phys. Lett.*, 1984, **109**, 417–423.
110 P. J. Knowles and N. C. Handy, *Chem. Phys. Lett.*, 1984, **111**, 315–321.
111 K. Hirao, *Chem. Phys. Lett.*, 1992, **190**, 374–380.
112 F. Coester and H. Kümmel, *Nucl. Phys.*, 1960, **17**, 477–485.
113 J. Čížek, *J. Chem. Phys.*, 1966, **45**, 4256–4266.
114 P. Hohenberg and W. Kohn, *Phys. Rev.*, 1964, **136**, B864–B871.
115 J. P. Perdew and K. Schmidt, *AIP Conf. Proc.*, 2001, **577**, 1–20.
116 M. A. Chiacchio and L. Legnani, *Int. J. Mol. Sci.*, 2024, **25**, 1298.
117 W. M. C. Sameera and F. Maseras, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2012, **2**, 375–385.
118 C. Duan, A. Nandy, R. Meyer, N. Arunachalam and H. J. Kulik, *Nat. Comput. Sci.*, 2023, **3**, 38–47.
119 C. Bannwarth, S. Ehlert and S. Grimme, *J. Chem. Theory Comput.*, 2019, **15**, 1652–1671.
120 M. Friede, C. Hölzer, S. Ehlert and S. Grimme, dxtb—An efficient and fully differentiable framework for extended tight-binding, *J. Chem. Phys.*, 2024, **161**, 062501.
121 T. Froitzheim, M. Müller, A. Hansen and S. Grimme, g-xTB: A General-Purpose Extended Tight-Binding Electronic Structure Method For the Elements H to Lr (Z = 1–103), *ChemRxiv*, 2025, preprint, DOI: 10.26434/chemrxiv-2025-bjxvt, https://chemrxiv.org/engage/chemrxiv/article-details/685434533ba0887c335fc974.
122 J. J. P. Stewart, *J. Mol. Model.*, 2013, **19**, 1–32.
123 M. Gaus, Q. Cui and M. Elstner, *J. Chem. Theory Comput.*, 2011, **7**, 931–948.
124 W. Thiel, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2014, **4**, 145–157.
125 Q. Cui and M. Elstner, *Phys. Chem. Chem. Phys.*, 2014, **16**, 14368–14377.
126 M. Gruden, L. Andjeković, A. K. Jissy, S. Stepanović, M. Zlatar, Q. Cui and M. Elstner, *J. Comput. Chem.*, 2017, **38**, 2171–2185.
127 M. H. Rasmussen and J. H. Jensen, *PeerJ Phys. Chem*, 2020, **2**, e15.
128 T. Saito and Y. Takano, *Bull. Chem. Soc. Jpn*, 2018, **91**, 1377–1389.
129 A. C. T. van Duin, S. Dasgupta, F. Lorant and W. A. Goddard, *J. Phys. Chem. A*, 2001, **105**, 9396–9409.
130 T. P. Senftle, S. Hong, M. M. Islam, S. B. Kylasa, Y. Zheng, Y. K. Shin, C. Junkermeier, R. Engel-Herbert, M. J. Janik, H. M. Aktulga, T. Verstraelen, A. Grama and A. C. T. van Duin, *npj Comput. Mater.*, 2016, **2**, 1–14.
131 T. Liang, Y. K. Shin, Y.-T. Cheng, D. E. Yilmaz, K. G. Vishnu, O. Verners, C. Zou, S. R. Phillpot, S. B. Sinnott and A. C. T. V. Duin, *Ann. Rev. Mater. Res.*, 2013, **43**, 109–129.
132 Y. K. Shin, T.-R. Shan, T. Liang, M. J. Noordhoek, S. B. Sinnott, A. C. T. V. Duin and S. R. Phillpot, *MRS Bull.*, 2012, **37**, 504–512.
133 D. M. Anstine and O. Isayev, *J. Phys. Chem. A*, 2023, **127**, 2417–2431.
134 J. Behler, *Chem. Rev.*, 2021, **121**, 10037–10072.
135 Y.-W. Zhang, V. Sorkin, Z. H. Aitken, A. Politano, J. Behler, A. P. Thompson, T. W. Ko, S. P. Ong, O. Chalykh, D. Korogod, E. Podryabinkin, A. Shapeev, J. Li, Y. Mishin, Z. Pei, X. Liu, J. Kim, Y. Park, S. Hwang, S. Han, K. Sheriff, Y. Cao and R. Freitas, *Modell. Simul. Mater. Sci. Eng.*, 2025, **33**, 023301.
136 S. Käser, L. Itza Vazquez-Salazar, M. Meuwly and K. Töpfer, *Digital Discovery*, 2023, **2**, 28–58.
137 E. Kocer, T. W. Ko and J. Behler, Neural Network Potentials: A Concise Overview of Methods, *arXiv*, 2021, preprint, arXiv:2107.03727, DOI: 10.48550/arXiv.2107.03727.
138 J. Behler, *Angew. Chem., Int. Ed.*, 2017, **56**, 12828–12840.
139 Q. Hu, A. Johannesen, D. Graham and J. Goodpaster, Neural Network Potentials for Reactive Chemistry: CASPT2 Quality Potential Energy Surfaces for Bond Breaking, *ChemRxiv*, 2023, preprint, DOI: 10.26434/chemrxiv-2023-13cv, https://chemrxiv.org/engage/chemrxiv/article-details/6452ce3e27fccdb3ea725259.
140 S. Manzhos and T. J. Carrington, *Chem. Rev.*, 2021, **121**, 10187–10217.
141 P. Yoo, M. Sakano, S. Desai, M. M. Islam, P. Liao and A. Strachan, *npj Comput. Mater.*, 2021, **7**, 1–10.
142 G. S. Jung, J. Y. Choi and S. M. Lee, *Digit. Discovery*, 2024, **3**, 514–527.
143 T. A. Young, T. Johnston-Wood, V. L. Deringer and F. Duarte, *Chem. Sci.*, 2021, **12**, 10944–10955.
144 Q. Lin, L. Zhang, Y. Zhang and B. Jiang, *J. Chem. Theory Comput.*, 2021, **17**, 2691–2701.
145 B. Li, J. Xiao, Y. Gao, J. Z. Zhang and T. Zhu, *J. Chem. Inf. Model.*, 2025, **65**, 2297–2303.
146 R. Jinnouchi, K. Miwa, F. Karsai, G. Kresse and R. Asahi, *J. Phys. Chem. Lett.*, 2020, **11**, 6946–6955.
147 D. S. Levine, M. Shuaibi, E. W. C. Spotte-Smith, M. G. Taylor, M. R. Hasyim, K. Michel, I. Batatia, G. Csányi, M. Dzamba, P. Eastman, N. C. Frey, X. Fu, V. Gharakhanyan, A. S. Krishnapriyan, J. A. Rackers, S. Raja, A. Rizvi, A. S. Rosen, Z. Ulissi, S. Vargas, C. L. Zitnick, S. M. Blau and B. M. Wood,

The Open Molecules 2025 (OMol25) Dataset, Evaluations, and Models, *arXiv*, 2025, preprint, arXiv:2505.08762 [physics], DOI: 10.48550/arXiv.2505.08762, https://arxiv.org/abs/2505.08762.

148 D. M. Anstine, Q. Zhao, R. Zubatiuk, S. Zhang, V. Singla, F. Nikitin, B. M. Savoie and O. Isayev, AIMNet2-rxn: A Machine Learned Potential for Generalized Reaction Modeling on a Millions-of-Pathways Scale, *ChemRxiv*, 2025, preprint, DOI: 10.26434/chemrxiv-2025-hpdmg, https://chemrxiv.org/engage/chemrxiv/article-details/685505c9c1cb1ecda0f701de.

149 D. Anstine, R. Zubatyuk, L. Gallegos, R. Paton, O. Wiest, B. Nebgen, T. Jones, G. Gomes, S. Tretiak and O. Isayev, Transferable Machine Learning Interatomic Potential for Pd-Catalyzed Cross-Coupling Reactions, *ChemRxiv*, 2025, preprint, DOI: 10.26434/chemrxiv-2025-n36r6, https://chemrxiv.org/engage/chemrxiv/article-details/67d7b7f7fa469535b97c021a.

150 B. Kalita, R. Zubatyuk, D. M. Anstine, M. Bergeler, V. Settels, C. Stork, S. Spicher and O. Isayev, AIMNet2-NSE: A Transferable Reactive Neural Network Potential for Open-Shell Chemistry, *ChemRxiv*, 2025, preprint, DOI: 10.26434/chemrxiv-2025-kdg6n-v2, https://chemrxiv.org/engage/chemrxiv/article-details/688ae42f728bf9025e345e87.

151 M. H. Segler and M. P. Waller, *Chem. – Eur. J.*, 2017, **23**, 6118–6128.

152 J. Nam and J. Kim, *arXiv*, 2016, preprint, arXiv:1612.09529, DOI: 10.48550/arXiv.1612.09529.

153 P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas and A. A. Lee, *ACS Cent. Sci.*, 2019, **5**, 1572–1583.

154 K. Do, T. Tran and S. Venkatesh, Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 750–760.

155 Z. Tu and C. W. Coley, *J. Chem. Inf. Model.*, 2022, **62**, 3503–3513.

156 N. O'Boyle and A. Dalke, *ChemRxiv*, 2018, preprint, DOI: 10.26434/chemrxiv.7097960.v1.

157 M. Krenn, Q. Ai, S. Barthel, N. Carson, A. Frei, N. C. Frey, P. Friederich, T. Gaudin, A. A. Gayle and K. M. Jablonka, *et al.*, *Patterns*, 2022, **3**, 100588.

158 W. W. Qian, N. T. Russell, C. L. Simons, Y. Luo, M. D. Burke and J. Peng, *ChemRxiv*, 2020, preprint, DOI: 10.26434/chemrxiv.11659563.v1.

159 M. Sacha, M. Błaz, P. Byrski, P. Dabrowski-Tumanski, M. Chrominski, R. Loska, P. Włodarczyk-Pruszynski and S. Jastrzebski, *J. Chem. Inf. Model.*, 2021, **61**, 3273–3284.

160 H. Wang, W. Li, X. Jin, K. Cho, H. Ji, J. Han and M. D. Burke, *arXiv*, 2021, preprint, arXiv:2109.09888, DOI: 10.48550/arXiv.2109.09888.

161 H. Bi, H. Wang, C. Shi, C. Coley, J. Tang and H. Guo, International Conference on Machine Learning, 2021, pp. 904–913.

162 T. Guo, C. Ma, X. Chen, B. Nan, K. Guo, S. Pei, N. V. Chawla, O. Wiest and X. Zhang, *arXiv*, 2023, preprint, arXiv:2310.04674, DOI: 10.48550/arXiv.2310.04674.

163 Z. Meng, P. Zhao, Y. Yu and I. King, *arXiv*, 2023, preprint, arXiv:2306.06119, DOI: 10.48550/arXiv2306.06119.

164 R. Irwin, S. Dimitriadis, J. He and E. J. Bjerrum, *Mach. Learn.: Sci. Technol.*, 2022, **3**, 015022.

165 J. Lu and Y. Zhang, *J. Chem. Inf. Model.*, 2022, **62**, 1376–1387.

166 Z. Zhong, J. Song, Z. Feng, T. Liu, L. Jia, S. Yao, M. Wu, T. Hou and M. Song, *Chem. Sci.*, 2022, **13**, 9023–9034.

167 H. Hu, Y. Jiang, Y. Yang and J. X. Chen, Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, 2023, pp. 813–822.

168 H. Hu, Y. Jiang, Y. Yang and J. X. Chen, *Appl. Intell.*, 2023, **53**, 29620–29637.

169 M. Zhao, L. Fang, L. Tan, J.-G. Lou and Y. Lepage, *arXiv*, 2022, preprint, arXiv:2204.05919, DOI: 10.48550/arXiv.2204.05919.

170 S. Chen and Y. Jung, *Nat. Mach. Intell.*, 2022, **4**, 772–780.

171 C. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay and K. F. Jensen, *Chem. Sci.*, 2019, **10**, 370–377.

172 M. H. Segler and M. P. Waller, *Chem. – Eur. J.*, 2017, **23**, 5966–5971.

173 Y. Wu, C. Zhang, L. Wang and H. Duan, *Chem. Commun.*, 2021, **57**, 4114–4117.

174 R. Sinkhorn, *Canadian J. Math.*, 1966, **18**, 303–306.

175 M. Das, A. Hoque, M. Baranwal and R. B. Sunoj, *arXiv*, 2025, preprint, arXiv:2509.15872, DOI: 10.48550/arXiv.2509.15872.

176 D. Weininger, A. Weininger and J. L. Weininger, *J. Chem. Inf. Comput. Sci.*, 1989, **29**, 97–101.

177 N. O'Boyle and A. Dalke, DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures, *ChemRxiv*, 2018, preprint, https://chemrxiv.org/engage/chemrxiv/article-details/60c73ed6567dfe7e5fec388d.

178 M. Krenn, F. Häse, A. Nigam, P. Friederich and A. Aspuru-Guzik, *Mach. Learn.: Sci. Technol.*, 2020, **1**, 045024.

179 T.-S. Lin, C. W. Coley, H. Mochigase, H. K. Beech, W. Wang, Z. Wang, E. Woods, S. L. Craig, J. A. Johnson and J. A. Kalow, *et al.*, *ACS Cent. Sci.*, 2019, **5**, 1523–1531.

180 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, *CoRR*, 2017, abs/1706.03762.

181 X. Wang, C. Yao, Y. Zhang, J. Yu, H. Qiao, C. Zhang, Y. Wu, R. Bai and H. Duan, *J. Cheminform.*, 2022, **14**, 60.

182 P. Schwaller, A. C. Vaucher, R. Laplaza, C. Bunne, A. Krause, C. Corminboeuf and T. Laino, *Wiley Interdiscip. Rev., Comput. mol.*, 2022, **12**, e1604.

183 S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen and B. Yu, *et al.*, *Nucleic Acids Res.*, 2021, **49**, D1388–D1395.

184 J. Bradshaw, A. Zhang, B. Mahjour, D. E. Graff, M. H. Segler and C. W. Coley, *ACS Cent. Sci.*, 2025, **11**, 539–549.

185 G. Pesciullesi, P. Schwaller, T. Laino and J.-L. Reymond, *Nat. Commun.*, 2020, **11**, 4874.

186 L. Wang, C. Zhang, R. Bai, J. Li and H. Duan, *Chem. Commun.*, 2020, **56**, 9368–9371.

187 Y. Zhang, L. Wang, X. Wang, C. Zhang, J. Ge, J. Tang, A. Su and H. Duan, *Org. Chem. Front.*, 2021, **8**, 1415–1423.

188 A. A. Lee, Q. Yang, V. Sresht, P. Bolgar, X. Hou, J. L. Klug-McLeod and C. R. Butler, *Chem. Commun.*, 2019, **55**, 12152–12155.

189 Z. Wang, H. Yi, Z. You and Q. Jin, *ng. Appl. Artif. Intell.*, 2026, **163**, 112850.

190 Y. Zhu, J. Hwang, K. Adams, Z. Liu, B. Nan, B. Stenfors, Y. Du, J. Chauhan, O. Wiest and O. Isayev *et al.*, *arXiv*, 2023, preprint, arXiv:2310.00115, DOI: 10.48550/arXiv.2310.00115.
191 Y. Guan, C. W. Coley, H. Wu, D. Ranasinghe, E. Heid, T. J. Struble, L. Pattanaik, W. H. Green and K. F. Jensen, *Chem. Sci.*, 2021, **12**(6), 2198–2208.
192 S. Zev, M. Roth, J. N. SJ and D. T. Major, *ChemRxiv*, 2025, preprint, DOI: 10.26434/chemrxiv-2025-vmr1q-v2.
193 G. Li, H. Ling, C. Su, Z. Liu, G. Wang, M. Yang and S. Li, *ChemRxiv*, 2025, preprint, DOI: 10.26434/chemrxiv-2025-sm7f3-v2.
194 P. Jørgensen, H. J. A. Jensen and T. Helgaker, *Theor. Chim. Acta*, 1988, **73**, 55–65.
195 D. K. Hoffman, R. S. Nord and K. Ruedenberg, *Theor. Chim. Acta*, 1986, **69**, 265–279.
196 K. Ohno and S. Maeda, *Chem. Phys. Lett.*, 2004, **384**, 277–282.
197 S. Maeda, K. Ohno and K. Morokuma, *Phys. Chem. Chem. Phys.*, 2013, **15**, 3683–3701.
198 S. Maeda, Y. Harabuchi, M. Takagi, T. Taketsugu and K. Morokuma, *Chem. Rec.*, 2016, **16**, 2232–2248.
199 S. Maeda and K. Morokuma, *J. Chem. Phys.*, 2010, **132**, 241102.
200 S. Maeda, T. Taketsugu and K. Morokuma, *J. Comput. Chem.*, 2014, **35**, 166–173.
201 J. P. Unsleber, S. A. Grimmel and M. Reiher, *J. Chem. Theory Comput.*, 2022, **18**, 5393–5409.
202 M. Bensberg, S. Grimmel, L. Lang, G. N. Simm, J.-G. Sobez, M. Steiner, P. L. Türtscher, J. P. Unsleber, T. Weymuth and M. Reiher, *qcscine/chemoton: Release 3.1.0*, https://zenodo.org/records/10159640.
203 M. Jafari and P. M. Zimmerman, *Phys. Chem. Chem. Phys.*, 2018, **20**, 7721–7729.
204 C. Lavigne, G. Gomes, R. Pollice and A. Aspuru-Guzik, *Chem. Sci.*, 2022, **13**, 13857–13871.
205 E. Martínez-Núñez, G. L. Barnes, D. R. Glowacki, S. Kopec, D. Peláez, A. Rodríguez, R. Rodríguez-Fernández, R. J. Shannon, J. J. P. Stewart, P. G. Tahoces and S. A. Vazquez, *J. Comput. Chem.*, 2021, **42**, 2036–2048.
206 E. Martínez-Núñez, *Phys. Chem. Chem. Phys.*, 2015, **17**, 14912–14921.
207 T. Huber, A. E. Torda and W. F. van Gunsteren, *J. Comput. Aided Mol. Des.*, 1994, **8**, 695–708.
208 A. Laio and M. Parrinello, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 12562–12566.
209 M. Iannuzzi, A. Laio and M. Parrinello, *Phys. Rev. Lett.*, 2003, **90**, 238302.
210 C. Shang and Z.-P. Liu, *J. Chem. Theory Comput.*, 2013, **9**, 1838–1845.
211 A. M. Saitta and F. Saija, *Proc. Natl. Acad. Sci. U. S. A.*, 2014, **111**, 13768–13773.
212 L.-P. Wang, R. T. McGibbon, V. S. Pande and T. J. Martinez, *J. Chem. Theory Comput.*, 2016, **12**, 638–649.
213 L.-P. Wang, A. Titov, R. McGibbon, F. Liu, V. S. Pande and T. J. Martínez, *Nat. Chem.*, 2014, **6**, 1044–1048.
214 J. Ford, S. Seritan, X. Zhu, M. N. Sakano, M. M. Islam, A. Strachan and T. J. Martínez, *J. Phys. Chem. A*, 2021, **125**, 1447–1460.
215 A. M. Chang, J. Meisner, R. Xu and T. J. Martínez, *J. Phys. Chem. A*, 2023, **127**, 9580–9589.
216 S. Grimme, *J. Chem. Theory Comput.*, 2019, **15**, 2847–2862.
217 I. Ugi, J. Bauer, J. Brandt, J. Friedrich, J. Gasteiger, C. Jochum and W. Schubert, *Angew. Chem., Int. Ed. Engl.*, 1979, **18**, 111–123.
218 I. Ugi, J. Bauer, K. Bley, A. Dengler, A. Dietz, E. Fontain, B. Gruber, R. Herges, M. Knauer and K. Reitsam, *et al.*, *Angew. Chem., Int. Ed. Engl.*, 1993, **32**, 201–227.
219 J. Bauer, E. Fontain, D. Forstmeyer and I. Ugi, *Tetrahedron Comput. Methodol.*, 1988, **1**, 129–132.
220 Y. V. Suleimanov and W. H. Green, *J. Chem. Theory Comput.*, 2015, **11**, 4248–4259.
221 P. Ramos-Sánchez, J. N. Harvey and J. A. Gámez, *J. Comput. Chem.*, 2023, **44**, 27–42.
222 P. M. Zimmerman, *J. Comput. Chem.*, 2013, **34**, 1385–1392.
223 I. Ismail, H. B. V. A. Stuttaford-Fowler, C. Ochan Ashok, C. Robertson and S. Habershon, *J. Phys. Chem. A*, 2019, **123**, 3407–3417.
224 N. Casetti, D. Anstine, O. Isayev and C. W. Coley, *J. Chem. Theory Comput.*, 2025, **21**, 10362–10372.
225 Q. Zhao and B. M. Savoie, *Angew. Chem.*, 2022, **134**, e202210693.
226 E. J. Corey, *Pure Appl. Chem*, 1967, **14**, 19–38.
227 E. Corey, W. J. Howe and D. A. Pensak, *J. Am. Chem. Soc.*, 1974, **96**, 7724–7737.
228 D. A. Pensak and E. J. Corey, *Computer-assisted synthetic analysis. Methods for machine generation of synthetic intermediates involving multistep look-ahead*, ACS Publications, 1977.
229 E. Corey, A. P. Johnson and A. K. Long, *J. Org. Chem.*, 1980, **45**, 2051–2057.
230 E. J. Corey, A. K. Long and S. D. Rubenstein, *Science*, 1985, **228**, 408–418.
231 W. T. Wipke and T. M. Dyott, *J. Am. Chem. Soc.*, 1974, **96**, 4825–4834.
232 W. T. Wipke and P. Gund, *J. Am. Chem. Soc.*, 1976, **98**, 8107–8118.
233 W. Wipke, H. Braun, G. Smith, F. Choplin and W. Sieber, *SECS—simulation and evaluation of chemical synthesis: strategy and planning*, ACS Publications, 1977.
234 W. T. Wipke, G. I. Ouchi and S. Krishnan, *Artif. Intell.*, 1978, **11**, 173–193.
235 H. Patel, W.-D. Ihlenfeldt, P. N. Judson, Y. S. Moroz, Y. Pevzner, M. L. Peach, V. Delannée, N. I. Tarasova and M. C. Nicklaus, *Sci. Data*, 2020, **7**, 384.
236 H. Patel, W. Ihlenfeldt, P. Judson, Y. S. Moroz, Y. Pevzner, M. Peach, N. Tarasova and M. Nicklaus, *Sci. Data*, 2020, **7**, 384.
237 K. Funatsu and S.-I. Sasaki, *Tetrahedron Comput. Methodol.*, 1988, **1**, 27–37.
238 K. Satoh and K. Funatsu, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 316–325.
239 J. Song, PhD thesis, Massachusetts Institute of Technology, 2004.
240 C. W. Gao, J. W. Allen, W. H. Green and R. H. West, *Comput. Phys. Commun.*, 2016, **203**, 212–225.

241 M. Liu, A. Grinberg Dana, M. S. Johnson, M. J. Goldman, A. Jocher, A. M. Payne, C. A. Grambow, K. Han, N. W. Yee, E. J. Mazeau, K. Blondal, R. H. West, C. F. Goldsmith and W. H. Green, *J. Chem. Inf. Model.*, 2021, **61**, 2686–2696.
242 L. J. Broadbelt, S. M. Stark and M. T. Klein, *Ind. Eng. Chem. Res.*, 1994, **33**, 790–799.
243 J. González-Lergier, L. J. Broadbelt and V. Hatzimanikatis, *J. Am. Chem. Soc.*, 2005, **127**, 9930–9938.
244 S. D. Finley, L. J. Broadbelt and V. Hatzimanikatis, *Biotechnol. Bioeng.*, 2009, **104**, 1086–1097.
245 C. S. Henry, L. J. Broadbelt and V. Hatzimanikatis, *Biotechnol. Bioeng.*, 2010, **106**, 462–473.
246 S. D. Finley, L. J. Broadbelt and V. Hatzimanikatis, *BMC Syst. Biol.*, 2010, **4**, 7.
247 D. A. Pertusi, A. E. Stine, L. J. Broadbelt and K. E. Tyo, *Bioinformatics*, 2015, **31**, 1016–1024.
248 A. Stine, M. Zhang, S. Ro, S. Clendennen, M. C. Shelton, K. E. Tyo and L. J. Broadbelt, *Biotechnol. Progress*, 2016, **32**, 303–311.
249 X. Zhou, Z. J. Brentzel, G. A. Kraus, P. L. Keeling, J. A. Dumesic, B. H. Shanks and L. J. Broadbelt, *ACS Sustainable Chem. Eng.*, 2018, **7**, 2414–2428.
250 M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki and M. Hirakawa, *Nucleic Acids Res.*, 2006, **34**, D354–D357.
251 L. B. Ellis, D. Roe and L. P. Wackett, *Nucleic Acids Res.*, 2006, **34**, D517–D521.
252 T. D. Salatin and W. L. Jorgensen, *J. Org. Chem.*, 1980, **45**, 2043–2051.
253 B. L. Roos-Kozel and W. L. Jorgensen, *J. Chem. Inf. Comput. Sci.*, 1981, **21**, 101–111.
254 T. D. Salatin, D. McLaughlin and W. L. Jorgensen, *J. Org. Chem.*, 1981, **46**, 5284–5294.
255 C. E. Peishoff and W. L. Jorgensen, *J. Org. Chem.*, 1983, **48**, 1970–1979.
256 J. M. Fleischer, A. J. Gushurst and W. L. Jorgensen, *J. Org. Chem.*, 1995, **60**, 490–498.
257 J. Gasteiger, M. G. Hutchings, B. Christoph, L. Gann, C. Hiller, P. Löw, M. Marsili, H. Saller and K. Yuki, *Organic Synthesis, Reactions and Mechanisms*, 1987, pp. 19–73.
258 P. Röse and J. Gasteiger, *Anal. Chim. Acta*, 1990, **235**, 163–168.
259 R. Höllering, J. Gasteiger, L. Steinhauer, K.-P. Schulz and A. Herwig, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 482–494.
260 J. Bauer, *Tetrahedron Comput. Methodol.*, 1989, **2**, 269–280.
261 I. M. Socorro, K. Taylor and J. M. Goodman, *Org. Lett.*, 2005, **7**, 3541–3544.
262 I. M. Socorro and J. M. Goodman, *J. Chem. Inf. Model.*, 2006, **46**, 606–614.
263 H. Satoh and K. Funatsu, *J. Chem. Inf. Comput. Sci.*, 1995, **35**, 34–44.
264 H. Satoh and K. Funatsu, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 173–184.
265 I. A. Watson, J. Wang and C. A. Nicolaou, *J. Cheminform.*, 2019, **11**, 1–12.
266 C. W. Coley, W. H. Green and K. F. Jensen, *J. Chem. Inf. Model.*, 2019, **59**, 2529–2537.
267 S. Chen, J. Noh, J. Jang, S. Kim, G. H. Gu and Y. Jung, *Acc. Chem. Res.*, 2024, **57**, 1964–1972.
268 J. Meng, H. Yang, C. Li, H. Song, N. Xia and X. Jiang, *Angew. Chem., Int. Ed.*, 2025, **64**, e202515595.
269 J. Law, Z. Zsoldos, A. Simon, D. Reid, Y. Liu, S. Y. Khew, A. P. Johnson, S. Major, R. A. Wade and H. Y. Ando, *J. Chem. Inf. Model.*, 2009, **49**, 593–602.
270 S. Chen and Y. Jung, *JACS Au*, 2021, **1**, 1612–1620.
271 J. Roh, J. F. Joung, K. Yu, Z. Tu, G. L. Bartholomew, O. A. Santiago-Reyes, M. H. Fong, R. Sarpong, S. E. Reisman and C. W. Coley, *ACS Cent. Sci.*, 2026, **12**, 345–357.
272 C. W. Coley, D. A. Thomas III, J. A. Lummiss, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers and H. Gao, *et al.*, *Science*, 2019, **365**, eaax1566.
273 J. H. Chen and P. Baldi, *J. Chem. Inf. Model.*, 2009, **49**, 2034–2043.
274 S. Szymkuć, A. Wołos, R. Roszak and B. A. Grzybowski, *Nat. Commun.*, 2024, **15**, 10286.
275 M. Krzeszewski, O. Vakuliuk, M. Tasior, A. Wołos, R. Roszak, K. Molga, M. B. Teimouri, B. A. Grzybowski and D. T. Gryko, *J. Am. Chem. Soc.*, 2025, **147**, 15636–15644.
276 A. Tomberg, M. J. Johansson and P.-O. Norrby, *J. Org. Chem.*, 2018, **84**, 4695–4703.
277 Y. Guan, T. Lee, K. Wang, S. Yu and J. C. McWilliams, *J. Chem. Inf. Model.*, 2023, **63**, 3751–3760.
278 A. Hoque, M. Das, M. Baranwal and R. B. Sunoj, *arXiv*, 2024, preprint, arXiv:2407.10090, DOI: 10.48550/arXiv.2407.10090.
279 R. J. Miller, A. E. Dashuta, B. Rudisill, D. Van Vranken and P. Baldi, *J. Am. Chem. Soc.*, 2025, **147**, 41168–41176.
280 R. J. Miller, A. E. Dashuta, B. Rudisill, D. Van Vranken and P. Baldi, *arXiv*, 2025, preprint, arXiv:2504.15539, DOI: 10.48550/arXiv.2504.15539.
281 M. A. Kayala, C.-A. Azencott, J. H. Chen and P. Baldi, *J. Chem. Inf. Model.*, 2011, **51**, 2209–2222.
282 M. A. Kayala and P. Baldi, *J. Chem. Inf. Model.*, 2012, **52**, 2526–2540.
283 M. Fujinami, J. Seino and H. Nakai, *Bull. Chem. Soc. Jpn*, 2020, **93**, 685–693.
284 M. Tavakoli, A. Mood, D. Van Vranken and P. Baldi, *J. Chem. Inf. Model.*, 2022, **62**, 2121–2132.
285 D. Fooshee, A. Mood, E. Gutman, M. Tavakoli, G. Urban, F. Liu, N. Huynh, D. Van Vranken and P. Baldi, *Mol. Syst. Des. Eng.*, 2018, **3**, 442–452.
286 M. Tavakoli, P. Baldi, A. M. Carlton, Y. T. Chiu, A. Shmakov and D. Van Vranken, *Adv. Neural Inf. Process.*, 2023, **36**, 4080–4096.
287 M. Tavakoli, A. Shmakov, F. Ceccarelli and P. Baldi, *arXiv*, 2022, preprint, arXiv:2201.01196, DOI: 10.48550/arXiv.2201.01196.
288 R. G. Susnow, A. M. Dean, W. H. Green, P. Peczak and L. J. Broadbelt, *J. Phys. Chem. A*, 1997, **101**, 3731–3740.
289 K. Han, W. H. Green and R. H. West, *Comput. Chem. Eng.*, 2017, **100**, 1–8.
290 Y. Sumiya, T. Taketsugu and S. Maeda, *J. Comput. Chem.*, 2017, **38**, 101–109.
291 Y. Sumiya and S. Maeda, *Chem. Lett.*, 2020, **49**, 553–564.

292 M. Bensberg and M. Reiher, *Isr. J. Chem.*, 2023, **63**, e202200123.
293 M. Bensberg and M. Reiher, *J. Phys. Chem. A*, 2024, **128**, 4532–4547.
294 M. Woulfe and B. M. Savoie, *J. Chem. Theory Comput.*, 2025, **21**, 1276–1291.
295 D. Barter, E. W. C. Spotte-Smith, N. S. Redkar, A. Khanwale, S. Dwaraknath, K. A. Persson and S. M. Blau, *Digit. Discovery*, 2023, **2**, 123–137.
296 J. Wei and J. C. W. Kuo, *Ind. Eng. Chem. Fundam.*, 1969, **8**, 114–123.
297 J. C. Keck, *Prog. Energy Combust. Sci.*, 1990, **16**, 125–154.
298 T. Lu and C. K. Law, *Proc. Combust. Inst.*, 2005, **30**, 1333–1341.
299 Y. Sumiya, Y. Nagahata, T. Komatsuzaki, T. Taketsugu and S. Maeda, *J. Phys. Chem. A*, 2015, **119**, 11641–11649.
300 D. T. Gillespie, *Annu. Rev. Phys. Chem.*, 2007, **58**, 35–55.
301 A. Chatterjee and D. G. Vlachos, *J. Comput. – Aided Mater. Des.*, 2007, **14**, 253–308.
302 P. Tuo, J. Chen and J. Li, *Flow matching for reaction pathway generation*, *arXiv*, 2025, preprint, arXiv:2507.10530 [physics], DOI: 10.48550/arXiv.2507.10530, https://arxiv.org/abs/2507.10530.
303 S. Huo and J. E. Straub, *J. Chem. Phys*, 1997, **107**, 5000–5006.
304 M. Berkowitz, J. D. Morgan, J. A. McCammon and S. H. Northrup, *J. Chem. Phys*, 1983, **79**, 5563–5565.
305 R. Crehuet and M. J. Field, *J. Chem. Phys*, 2003, **118**, 9563–9571.
306 R. Elber and M. Karplus, *Chem. Phys. Lett.*, 1987, **139**, 375–380.
307 R. Czerminski and R. Elber, *J. Chem. Phys*, 1990, **92**, 5580–5601.
308 R. Czerminski and R. Elber, *Int. J. Quantum Chem.*, 1990, **38**, 167–185.
309 A. Ulitsky and R. Elber, *J. Chem. Phys*, 1990, **92**, 1510–1511.
310 C. Peng and H. Bernhard Schlegel, *Isr. J. Chem.*, 1993, **33**, 449–454.
311 G. Mills and H. Jónsson, *Phys. Rev. Lett.*, 1994, **72**, 1124–1127.
312 H. Jonsson, G. Mills and K. W. Jacobsen, *Classical and Quantum Dynamics in Condensed Phase Simulations*, World Scientific, 1998, pp. 385–404.
313 G. Henkelman and H. Jónsson, *J. Chem. Phys.*, 2000, **113**, 9978–9985.
314 R. Olender and R. Elber, *J. Chem. Phys*, 1996, **105**, 9299–9315.
315 A. B. Birkholz and H. B. Schlegel, *J. Chem. Phys*, 2015, **143**, 244101.
316 A. B. Birkholz and H. B. Schlegel, *J. Chem. Phys*, 2016, **144**, 184101.
317 E. Vanden-Eijnden and M. Heymann, *J. Chem. Phys*, 2008, **128**, 061103.
318 G. Henkelman, B. P. Uberuaga and H. Jónsson, *J. Chem. Phys.*, 2000, **113**, 9901–9904.
319 E. Weinan, W. Ren and E. Vanden-Eijnden, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2002, **66**, 052301.
320 B. Peters, A. Heyden, A. T. Bell and A. Chakraborty, *J. Chem. Phys*, 2004, **120**, 7877–7886.
321 E. Weinan, W. Ren and E. Vanden-Eijnden, *J. Chem. Phys.*, 2007, **126**, 164103.
322 B. K. Dey and P. W. Ayers, *Mol. Phys.*, 2006, **104**, 541–558.
323 R. Granot and R. Baer, *J. Chem. Phys*, 2008, **128**, 184111.
324 S. A. Ghasemi and S. Goedecker, *J. Chem. Phys*, 2011, **135**, 014108.
325 S. Smidstrup, A. Pedersen, K. Stokbro and H. Jónsson, *J. Chem. Phys.*, 2014, **140**, 214106.
326 A. C. Vaucher and M. Reiher, *J. Chem. Theory Comput.*, 2018, **14**, 3091–3099.
327 X. Zhu, K. C. Thompson and T. J. Martínez, *J. Chem. Phys.*, 2019, **150**, 164103.
328 V. Ásgeirsson, B. O. Birgisson, R. Bjornsson, U. Becker, F. Neese, C. Riplinger and H. Jónsson, *J. Chem. Theory Comput.*, 2021, **17**, 4929–4945.
329 T. Stuyver, *J. Comput. Chem.*, 2024, **45**, 2308–2317.
330 H. B. Schlegel, *J. Comput. Chem.*, 1982, **3**(2), 214–218.
331 J. Baker, *J. Comput. Chem.*, 1986, **7**, 385–395.
332 A. Banerjee, N. Adams, J. Simons and R. Shepard, *J. Phys. Chem.*, 1985, **89**, 52–57.
333 E. Besalú and J. M. Bofill, *Theor. Chem. Acc.*, 1998, **100**, 265–274.
334 G. Henkelman and H. Jónsson, *J. Chem. Phys.*, 1999, **111**, 7010–7022.
335 R. A. Olsen, G. J. Kroes, G. Henkelman, A. Arnaldsson and H. Jónsson, *J. Chem. Phys.*, 2004, **121**, 9776–9792.
336 K. Fukui, *Acc. Chem. Res.*, 1981, **14**, 363–368.
337 K. Fukui, *J. Phys. Chem.*, 1970, **74**, 4161–4163.
338 D. Lemm, G. F. von Rudorff and O. A. von Lilienfeld, *Nat. Commun.*, 2021, **12**, 4468.
339 L. Pattanaik, J. B. Ingraham, C. A. Grambow and W. H. Green, *Phys. Chem. Chem. Phys.*, 2020, **22**, 23618–23626.
340 R. Jackson, W. Zhang and J. Pearson, *Chem. Sci.*, 2021, **12**, 10022–10040.
341 S. Choi, *Nat. Commun.*, 2023, **14**, 1168.
342 M. Z. Makoś, N. Verma, E. C. Larson, M. Freindorf and E. Kraka, *J. Chem. Phys.*, 2021, **155**, 024116.
343 C. Duan, Y. Du, H. Jia and H. J. Kulik, *Nat. Comput. Sci.*, 2023, **3**, 1045–1055.
344 S. Kim, J. Woo and W. Y. Kim, *Nat. Commun.*, 2024, **15**, 341.
345 C. Duan, G.-H. Liu, Y. Du, T. Chen, Q. Zhao, H. Jia, C. P. Gomes, E. A. Theodorou and H. J. Kulik, *Nat. Mach. Intell.*, 2025, **7**, 615–626.
346 Q. Zhao, Y. Han, D. Zhang, J. Wang, P. Zhong, T. Cui, B. Yin, Y. Cao, H. Jia and C. Duan, *Adv. Sci.*, 2025, **12**, e06240.
347 L. Galustian, K. Mark, J. Karwounopoulos, M. P.-P. Kovar and E. Heid, *Digit. Discovery*, 2025, DOI: 10.1039/D5DD00283D.
348 E. Heid and W. H. Green, *J. Chem. Inf. Model.*, 2021, **62**, 2101–2110.
349 I. W. Beaglehole, M. J. Pemberton, E. H. E. Farrar and M. N. Grayson, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2025, **15**, e70025.
350 K. A. Spiekermann, X. Dong, A. Menon, W. H. Green, M. Pfeifle, F. Sandfort, O. Welz and M. Bergeler, *J. Phys. Chem. A*, 2024, **128**, 8384–8403.
351 P. van Gerwen, A. Fabrizio, M. D. Wodrich and C. Corminboeuf, *Mach. Learn.: Sci. Technol.*, 2022, **3**, 045005.

352 P. van Gerwen, K. R. Briling, C. Bunne, V. R. Somnath, R. Laplaza, A. Krause and C. Corminboeuf, *J. Chem. Inf. Model.*, 2024, **64**, 5771–5785.
353 P. v Gerwen, K. R. Briling, Y. C. Alonso, M. Franke and C. Corminboeuf, *Digit. Discovery*, 2024, **3**, 932–943.
354 P. van Gerwen, M. D. Wodrich, R. Laplaza and C. Corminboeuf, *Mach. Learn.: Sci. Technol.*, 2023, **4**, 048002.
355 J. De Landsheere, M. P.-P. Kovar, K. Mark, L. Ganser, L. Galustian, J. Karwounopoulos, C. Gerhaher and E. Heid, *ChemRxiv*, 2026, preprint, DOI: 10.26434/chemrxiv-2026-np10c.
356 E. H. E. Farrar and M. N. Grayson, *Chem. Sci.*, 2022, **13**, 7594–7603.
357 H.-C. Chang, M.-H. Tsai and Y.-P. Li, *J. Chem. Inf. Model.*, 2025, **65**, 1367–1377.
358 N. Lalith, A. R. Singh and J. A. Gauthier, *ChemPhysChem*, 2024, **25**, e202300933.
359 E. Marques, S. de Gendt, G. Pourtois and M. J. van Setten, *J. Chem. Inf. Model.*, 2023, **63**, 1454–1461.
360 T. Stuyver and C. W. Coley, *J. Chem. Phys.*, 2022, **156**, 2308–2317.
361 J. Karwounopoulos, J. D. Landsheere, L. Galustian, T. Jechtl and E. Heid, *Digit. Discovery*, 2025, **4**, 3208–3216.
362 T. I. Madzhidov, A. V. Bodrov, T. R. Gimadiev, R. I. Nugmanov, I. S. Antipin and A. A. Varnek, *J. Struct. Chem.*, 2015, **56**, 1227–1234.
363 T. Gimadiev, T. Madzhidov, I. Tetko, R. Nugmanov, I. Casciuc, O. Klimchuk, A. Bodrov, P. Polishchuk, I. Antipin and A. Varnek, *Mol. Inform.*, 2019, **38**, 1800104.
364 K. Jorner, T. Brinck, P.-O. Norrby and D. Buttar, *Chem. Sci.*, 2021, **12**, 1163–1175.
365 L. Tomme, I. Lengyel, F. H. Vermeire, C. V. Stevens and K. M. Van Geem, *Mach. Learn.: Sci. Technol.*, 2025, **6**, 035044.
366 M. Bragato, G. F. von Rudorff and O. A. von Lilienfeld, *Chem. Sci.*, 2020, **11**, 11859–11868.
367 J. M. Ravasco and J. A. Coelho, *J. Am. Chem. Soc.*, 2020, **142**, 4235–4241.
368 A. M. Tokita, T. Devergne, A. M. Saitta and J. Behler, *J. Chem. Phys.*, 2025, **162**, 17.
369 A. Kümmel, S. Panke and M. Heinemann, *Mol. Syst. Biol.*, 2006, **2**, 2006.0034.
370 C. S. Henry, L. J. Broadbelt and V. Hatzimanikatis, *Biophys. J.*, 2007, **92**, 1792–1805.
371 N. Zamboni, A. Kümmel and M. Heinemann, *BMC Bioinf.*, 2008, **9**, 199.
372 K. C. Soh and V. Hatzimanikatis, *Metabolic Flux Analysis*, Springer New York, New York, NY, 2014, vol. 1191, pp. 49–63.
373 M. G. Evans and M. Polanyi, *Trans. Faraday Soc.*, 1938, **34**, 11–24.
374 N. Cohen and S. W. Benson, *Chem. Rev.*, 1993, **93**, 2419–2438.
375 S. W. Benson and J. H. Buss, *J. Chem. Phys.*, 1958, **29**, 546–572.
376 S. W. Benson, F. R. Cruickshank, D. M. Golden, G. R. Haugen, H. E. O'Neal, A. S. Rodgers, R. Shaw and R. Walsh, *Chem. Rev.*, 1969, **69**, 279–324.
377 S. Gronert, *J. Org. Chem.*, 2006, **71**, 9560.
378 Q. Zhao and B. M. Savoie, *J. Chem. Inf. Model.*, 2020, **60**, 2199–2207.
379 M. D. Wodrich, C. Corminboeuf and S. E. Wheeler, *J. Phys. Chem. A*, 2012, **116**, 3436–3447.
380 P. Linstrom, *NIST Chemistry WebBook, NIST Standard Reference Database 69*, 1997, https://webbook.nist.gov/chemistry/.
381 B. Deng and T. Stuyver, *J. Chem. Inf. Model.*, 2025, **65**, 649–659.
382 L.-Y. Chen, T.-W. Hsu, T.-C. Hsiung and Y.-P. Li, *J. Phys. Chem. A*, 2022, **126**, 7548–7556.
383 F. N. O. Bruce, D. Zhang, X. Bai, S. Song, F. Wang, Q. Chu, D. Chen and Y. Li, *Fuel*, 2025, **384**, 133999.
384 J. Wu and X. Xu, *J. Chem. Phys*, 2007, **127**, 214105.
385 M. D. Wodrich and C. Corminboeuf, *J. Phys. Chem. A*, 2009, **113**, 3285–3290.
386 C. A. Grambow, Y.-P. Li and W. H. Green, *J. Phys. Chem. A*, 2019, **123**, 5826–5835.
387 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. von Lilienfeld, *J. Chem. Theory Comput.*, 2015, **11**, 2087–2096.
388 J. Sun, J. Wu, T. Song, L. Hu, K. Shan and G. Chen, *J. Phys. Chem. A*, 2014, **118**, 9120–9131.
389 X. Chen, P. Li, E. Hruska and F. Liu, *Phys. Chem. Chem. Phys.*, 2023, **25**, 13417–13428.
390 M. Yang, S. Wang, G. Song, L. Cheng and H. Ren, *J. Phys. Chem. A*, 2025, **129**, 5901–5910.
391 M. Jiang, Z. Wang, Y. Chen, W. Zhang, Z. Zhu, W. Yan, J. Wu and X. Xu, *J. Comput. Chem.*, 2025, **46**, e70081.
392 M. H. Segler, M. Preuss and M. P. Waller, *Nature*, 2018, **555**, 604–610.
393 M. Wollenhaupt, M. Villalba and O. Ravitz, *Predicting New Chemistry: The Impact of High-Quality Training Data on the Prediction of Reaction Outcomes*, 2021.
394 F. Strieth-Kalthoff, F. Sandfort, M. Kühnemund, F. R. Schäfer, H. Kuchen and F. Glorius, *Angew. Chem., Int. Ed.*, 2022, **61**, e202204647.
395 J. N. Wei, D. Duvenaud and A. Aspuru-Guzik, *ACS Cent. Sci.*, 2016, **2**, 725–732.
396 T. A. Neukomm, Z. Jončev and P. Schwaller, Teaching Language Models Mechanistic Explainability Through Arrow-Pushing, *arXiv*, 2025, preprint, arXiv:2512.05722 [cs], DOI: 10.48550/arXiv.2512.05722, https://arxiv.org/abs/2512.05722.
397 A. F. Zahrt, S. V. Athavale and S. E. Denmark, *Chem. Rev.*, 2019, **120**, 1620–1689.
398 L. M. Sigmund, M. Assante, M. J. Johansson, P.-O. Norrby, K. Jorner and M. Kabeshov, *Chem. Sci.*, 2025, 5383–5412.
399 J. P. Reid, I. O. Betinol and Y. Kuang, *Chem. Commun.*, 2023, **59**, 10711–10721.
400 A. Verloop, W. Hoogenstraaten and J. Tipker, *Drug Design*, 1976, **7**, 165–207.
401 Z. Zhang, J. Qiu, J. Zheng, Z. Yu, L. Su, Q. Lin, C. Zhang and K. Liao, *J. Chem. Inf. Model.*, 2025, **65**, 3420–3430.
402 S. Chen and R. Pollice, *Chem Catal.*, 2024, **4**, 101111.
403 K. Hasegawa, M. Koyama and K. Funatsu, *Mol. Inform.*, 2010, **29**, 243–249.
404 T. J. Struble, C. W. Coley and K. F. Jensen, *React. Chem. Eng.*, 2020, **5**, 896–902.

405 N. Ree, A. H. Göller and J. H. Jensen, *Digit. Discovery*, 2022, **1**, 108–114.
406 T. T. V. Tran, H. Tayara and K. T. Chong, *Pharmaceutics*, 2023, **15**, 1260.
407 E. King-Smith, F. A. Faber, U. Reilly, A. V. Sinitskiy, Q. Yang, B. Liu, D. Hyek and A. A. Lee, *Nat. Commun.*, 2024, **15**, 426.
408 D. F. Nippa, K. Atz, A. T. Müller, J. Wolfard, C. Isert, M. Binder, O. Scheidegger, D. B. Konrad, U. Grether and R. E. Martin, *et al.*, *Commun. Chem.*, 2023, **6**, 256.
409 J. Götz, E. Richards, I. A. Stepek, Y. Takahashi, Y.-L. Huang, L. Bertschi, B. Rubi and J. W. Bode, *Sci. Adv.*, 2025, **11**, eadw6047.
410 J. Schleinitz, A. Carretero-Cerdán, A. Gurajapu, Y. Harnik, G. Lee, A. Pandey, A. Milo and S. E. Reisman, *J. Am. Chem. Soc.*, 2025, **147**, 7476–7484.
411 X. Li, S.-Q. Zhang, L.-C. Xu and X. Hong, *Angew. Chem., Int. Ed.*, 2020, **59**, 13253–13259.
412 J. Seumer, N. Ree and J. H. Jensen, *J. Org. Chem.*, 2025, **21**, 1171–1182.
413 X. Qu, D. A. Latino and J. Aires-de Sousa, *J. Cheminform.*, 2013, **5**, 1–13.
414 H. Yu, Y. Wang, X. Wang, J. Zhang, S. Ye, Y. Huang, Y. Luo, E. Sharman, S. Chen and J. Jiang, *J. Phys. Chem. A*, 2020, **124**, 3844–3850.
415 M. Wen, S. M. Blau, E. W. C. Spotte-Smith, S. Dwaraknath and K. A. Persson, *Chem. Sci.*, 2021, **12**, 1858–1868.
416 W. Li, Y. Luan, Q. Zhang and J. Aires-de Sousa, *Mol. Inform.*, 2023, **42**, 2200193.
417 S. S. S. Vejaykummar, Y. Kim, S. Kim, P. C. S. John and R. S. Paton, *Digit. Discovery*, 2023, **2**, 1900–1910.
418 M. Kaneko, Y. Takano and T. Saito, *Chem. Lett.*, 2024, **53**, upae016.
419 R. M. Borup, N. Ree and J. H. Jensen, *ChemRxiv*, 2024, preprint, DOI: 10.26434/chemrxiv-2024-0nxcv-v.
420 B. Li, Y. Liu, H. Sun, R. Zhang, Y. Xie, K. Foo, F. S. Mak, R. Zhang, T. Yu and S. Lin, *et al.*, *Digit. Discovery*, 2024, **3**, 2019–2031.
421 A. J. Bischoff, B. M. Nelson, Z. L. Niemeyer, M. S. Sigman and M. Movassaghi, *J. Am. Chem. Soc.*, 2017, **139**, 15539–15547.
422 J. D. Griffin, D. B. Vogt, J. Du Bois and M. S. Sigman, *ACS Catal.*, 2021, **11**, 10479–10486.
423 M. Kruszyk, M. Jessing, J. L. Kristensen and M. Jørgensen, *J. Org. Chem.*, 2016, **81**, 5128–5134.
424 P. E. Gormisky and M. C. White, *J. Am. Chem. Soc.*, 2013, **135**, 14052–14055.
425 R. Asahara and T. Miyao, *ACS Omega*, 2022, **7**, 26952–26964.
426 M. Ruth, T. Gensch and P. R. Schreiner, *Angew. Chem., Int. Ed.*, 2024, **63**, e202410308.
427 K. C. Harper and M. S. Sigman, *Science*, 2011, **333**, 1875–1878.
428 S. Singh and R. B. Sunoj, *Acc. Chem. Res.*, 2023, **56**, 402–412.
429 W. L. Williams, L. Zeng, T. Gensch, M. S. Sigman, A. G. Doyle and E. V. Anslyn, *ACS Cent. Sci.*, 2021, **7**, 1622–1637.
430 S. Singh, M. Pareek, A. Changotra, S. Banerjee, B. Bhaskararao, P. Balamurugan and R. B. Sunoj, *Proc. Natl. Acad. Sci. U. S. A.*, 2020, **117**, 1339–1345.
431 M. Das, P. Sharma and R. B. Sunoj, *J. Chem. Phys.*, 2022, **156**, 114303.
432 S. Moon, S. Chatterjee, P. H. Seeberger and K. Gilmore, *Chem. Sci.*, 2021, **12**, 2931–2939.
433 J. D. Oslob, B. Åkermark, P. Helquist and P.-O. Norrby, *Organometallics*, 1997, **16**, 3015–3021.
434 E. N. Bess, A. J. Bischoff and M. S. Sigman, *Proc. Natl. Acad. Sci. U. S. A.*, 2014, **111**, 14698–14703.
435 A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow and S. E. Denmark, *Science*, 2019, **363**(6424), eaau5631.
436 H. Huang, H. Zong, G. Bian and L. Song, *J. Org. Chem.*, 2012, **77**, 10427–10434.
437 F. Sandfort, F. Strieth-Kalthoff, M. Kühnemund, C. Beecks and F. Glorius, *Chem*, 2020, **6**, 1379–1390.
438 J. J. Dotson, L. van Dijk, J. C. Timmerman, S. Grosslight, R. C. Walroth, F. Gosselin, K. Puntener, K. A. Mack and M. S. Sigman, *J. Am. Chem. Soc.*, 2022, **145**, 110–121.
439 L.-C. Xu, J. Frey, X. Hou, S.-Q. Zhang, Y.-Y. Li, J. C. Oliveira, S.-W. Li, L. Ackermann and X. Hong, *Nat. Synth.*, 2023, **2**, 321–330.
440 A. Hoque and R. B. Sunoj, *Digit. Discovery*, 2022, **1**, 926–940.
441 N. V. Faurschou, V. Friis, P. Raghavan, C. M. Pedersen and C. W. Coley, *J. Am. Chem. Soc.*, 2025, 36197–36209.
442 Y. Gong, D. Xue, G. Chuai, J. Yu and Q. Liu, *Chem. Sci.*, 2021, **12**, 14459–14472.
443 J. P. Reid and M. S. Sigman, *Nature*, 2019, **571**, 343–348.
444 S. M. Maley, D.-H. Kwon, N. Rollins, J. C. Stanley, O. L. Sydora, S. M. Bischof and D. H. Ess, *Chem. Sci.*, 2020, **11**, 9665–9674.
445 S. Gallarati, R. Fabregat, R. Laplaza, S. Bhattacharjee, M. D. Wodrich and C. Corminboeuf, *Chem. Sci.*, 2021, **12**, 6879–6889.
446 J. M. Granda, L. Donina, V. Dragone, D.-L. Long and L. Cronin, *Nature*, 2018, **559**, 377–381.
447 K. V. Chuang and M. J. Keiser, *Science*, 2018, **362**, eaat8603.
448 R. Shi, G. Yu, X. Huo and Y. Yang, *J. Cheminform.*, 2024, **16**, 22.
449 B. Li, S. Su, C. Zhu, J. Lin, X. Hu, L. Su, Z. Yu, K. Liao and H. Chen, *J. Cheminform.*, 2023, **15**, 72.
450 A. Sato, R. Asahara and T. Miyao, *ACS Omega*, 2024, **9**, 40907–40919.
451 E. King-Smith, S. Berritt, L. Bernier, X. Hou, J. Klug-McLeod, J. Mustakis, N. Sach, J. Tucker, Q. Yang, R. Howard and A. Lee, *ChemRxiv*, 2023, preprint, DOI: 10.26434/chemrxiv-2022-hjnmr-v2.
452 C. Zhang, Q. Lin, C. Yang, Y. Kong, Z. Yu and K. Liao, *Chem. Sci.*, 2025, **16**, 11809–11822.
453 D. F. Nippa, K. Atz, R. Hohler, A. T. Müller, A. Marx, C. Bartelmus, G. Wuitschik, I. Marzuoli, V. Jost, J. Wolfard, M. Binder, A. F. Stepan, D. B. Konrad, U. Grether, R. E. Martin and G. Schneider, *ChemRxiv*, 2022, preprint, DOI: 10.26434/chemrxiv-2022-gkxm6-v2.

454 P. Schwaller, A. C. Vaucher, T. Laino and J.-L. Reymond, *Mach. Learn.: Sci. Technol.*, 2021, **2**(1), 015016.
455 S. Newman-Stonebraker, S. Smith, J. Borowski, E. Peters, T. Gensch, H. Johnson, M. Sigman and A. Doyle, Linking Mechanistic Analysis of Catalytic Reactivity Cliffs to Ligand Classification, *ChemRxiv*, 2021, preprint, DOI: 10.26434/chemrxiv.14388557.v1.
456 J. Schleinitz, M. Langevin, Y. Smail, B. Wehnert, L. Grimaud and R. Vuilleumier, *J. Am. Chem. Soc.*, 2022, **144**(32), 14722–14730.
457 R. Mercado, S. M. Kearnes and C. W. Coley, *J. Chem. Inf. Model.*, 2023, **63**, 4253–4265.
458 R. G. Bergman and R. L. Danheiser, *Angew. Chem., Int. Ed.*, 2016, **55**, 12548–12549.
459 Detailing Experimental Procedures, 2025, https://cen.acs.org/articles/91/i21/Detailing-Experimental-Procedures.html.
460 M. P. Maloney, C. W. Coley, S. Genheden, N. Carson, P. Helquist, P.-O. Norrby and O. Wiest, *J. Org. Chem.*, 2023, **88**(9), 5239–5241.
461 P. Raghavan, B. C. Haas, M. E. Ruos, J. Schleinitz, A. G. Doyle, S. E. Reisman, M. S. Sigman and C. W. Coley, *ACS Cent. Sci.*, 2023, **9**, 2196–2204.
462 S. Ghosh, N. Jain and R. B. Sunoj, Efficient Machine Learning Approach for Yield Prediction in Chemical Reactions, *arXiv*, 2025, preprint, arXiv:2502.19976 [physics], DOI: 10.48550/arXiv.2502.19976, https://arxiv.org/abs/2502.19976.
463 P. Raghavan, A. J. Rago, P. Verma, M. M. Hassan, G. M. Goshu, A. W. Dombrowski, A. Pandey, C. W. Coley and Y. Wang, *J. Am. Chem. Soc.*, 2024, 15070–15084.
464 L. W. Souza, N. D. Ricke, B. C. Chaffin, M. E. Fortunato, S. Jiang, C. Soylu, T. C. Caya, S. H. Lau, K. A. Wieser, A. G. Doyle and K. L. Tan, *J. Am. Chem. Soc.*, 2025, **147**, 18747–18759.
465 K. M. Van Geem, M.-F. Reyniers, G. B. Marin, J. Song, W. H. Green and D. M. Matheu, *AIChE J.*, 2006, **52**, 718–730.
466 F. Seyedzadeh Khanshan and R. H. West, *Fuel*, 2016, **163**, 25–33.
467 S. S. Merchant, E. F. Zanoelo, R. L. Speth, M. R. Harper, K. M. Van Geem and W. H. Green, *Comb. Flame*, 2013, **160**, 1907–1929.
468 Q. Zhao and B. M. Savoie, *Proc. Natl. Acad. Sci. U. S. A.*, 2023, **120**, e2305884120.
469 Q. Zhao, S. S. Garimella and B. M. Savoie, *J. Am. Chem. Soc.*, 2023, **145**, 6135–6143.
470 H.-H. Hsu, T. Jin and B. M. Savoie, *J. Phys. Chem. Lett.*, 2025, **16**, 7685–7694.
471 H. Hayashi, H. Katsuyama, H. Takano, Y. Harabuchi, S. Maeda and T. Mita, *Nat. Synth.*, 2022, **1**, 804–814.
472 R. Staub, Y. Harabuchi, C. Seraphim, A. Varnek and S. Maeda, An accurate and efficient reaction path search with iteratively trained neural network potential: Answering the Passerini mechanism controversy, *ChemRxiv*, 2025, preprint, DOI: 10.26434/chemrxiv-2025-9h8dr.
473 M. Puripat, R. Ramozzi, M. Hatanaka, W. Parasuk, V. Parasuk and K. Morokuma, *J. Org. Chem.*, 2015, **80**, 6959–6967.
474 S. Maeda and K. Morokuma, *J. Chem. Theory Comput.*, 2012, **8**, 380–385.
475 W. H. Green, *AIChE J.*, 2020, **66**, e17059.
476 H. J. Kulik, *Isr. J. Chem.*, 2022, **62**, e202100016.
477 A. Grinberg Dana, K. M. Van Geem, C. Cavallotti and W. H. Green, *ACS Eng. Au*, 2026, **6**, 1–19.
478 I. Amin, S. Raja and A. Krishnapriyan, Towards Fast, Specialized Machine Learning Force Fields: Distilling Foundation Models via Energy Hessians, *arXiv*, 2025, preprint, arXiv:2501.09009 [physics], DOI: 10.48550/arXiv.2501.09009, https://arxiv.org/abs/2501.09009.
479 E. Qu, B. M. Wood, A. S. Krishnapriyan and Z. W. Ulissi, A recipe for scalable attention-based MLIPs: unlocking long-range accuracy with all-to-all node attention, *arXiv*, 2026, preprint, arXiv:2603.06567 [cs], DOI: 10.48550/arXiv.2603.06567, https://arxiv.org/abs/2603.06567.
480 E. C.-Y. Yuan, Y. Liu, J. Chen, P. Zhong, S. Raja, T. Kreiman, S. Vargas, W. Xu, M. Head-Gordon, C. Yang, S. M. Blau, B. Cheng, A. Krishnapriyan and T. Head-Gordon, *Nat. Rev. Chem.*, 2026, **10**, 212–230.
481 Y. Chiang, T. Kreiman, C. Zhang, M. C. Kuner, E. Weaver, I. Amin, H. Park, Y. Lim, J. Kim, D. Chrzan, A. Walsh, S. M. Blau, M. Asta and A. S. Krishnapriyan, MLIP Arena: Advancing Fairness and Transparency in Machine Learning Interatomic Potentials via an Open, Accessible Benchmark Platform, *arXiv*, 2025, preprint, arXiv:2509.20630 [physics], DOI: 10.48550/arXiv.2509.20630, https://arxiv.org/abs/2509.20630.
482 K. Zhang, N. Lokachari, E. Ninnemann, S. Khanniche, W. H. Green, H. J. Curran, S. S. Vasu and W. J. Pitz, *Proc. Combust. Inst.*, 2019, **37**, 657–665.
483 X. Sun, J. Liu, B. Mahjour, K. F. Jensen and C. W. Coley, *Chem. Sci.*, 2025, **16**, 18176–18189.
484 C. Li and R. A. Shenvi, *Nature*, 2025, **638**, 980–986.

