nature computational science

Perspective

https://doi.org/10.1038/s43588-022-00369-z

Chemical reaction networks and opportunities for machine learning

Received: 24 June 2022

Accepted: 8 November 2022

Published online: 16 January 2023

Check for updates

Mingjian Wen[1,2], Evan Walter Clark Spotte-Smith[3,4], Samuel M. Blau[2], Matthew J. McDermott[3,4], Aditi S. Krishnapriyan[5,6,7] & Kristin A. Persson[4,8]

Chemical reaction networks (CRNs), defined by sets of species and possible reactions between them, are widely used to interrogate chemical systems. To capture increasingly complex phenomena, CRNs can be leveraged alongside data-driven methods and machine learning (ML). In this Perspective, we assess the diverse strategies available for CRN construction and analysis in pursuit of a wide range of scientific goals, discuss ML techniques currently being applied to CRNs and outline future CRN-ML approaches, presenting scientific and technical challenges to overcome.

Computational research occupies a key role in studies of chemical reactivity. In domains such as gas phase thermochemistry[1,2], homogeneous[3] and heterogeneous[4,5] catalysis, electrochemistry[6,7] and atmospheric chemistry[8,9], short-lived intermediate species can be difficult or impossible to detect via experimental spectroscopy[10,11], making computational elucidation of reaction mechanisms critical to explain observed reaction outcomes and dynamics. Complex interactions in, for example, biochemical and cellular processes can often only be effectively disentangled using theoretical modeling[12,13]. Moreover, computational approaches are increasingly used to optimize industrial chemical processes[14] and enable novel materials syntheses[15]. Retrosynthesis planning tools in organic chemistry[16,17]—and, more recently, in materials chemistry[18,19]—can select from a combinatorial explosion of possible synthesis routes to maximize yield, minimize cost or minimize synthesis complexity, streamlining an otherwise extremely labor-intensive task.

Computational studies of reactivity are highly diverse, but a common approach is to interrogate chemical reaction networks (CRNs), sometimes called simply 'reaction networks' (Fig. 1). A CRN consists of a set of species, **S**, and a set of reactions, **R**, where each reaction is defined by its reactant and product species[20]. Analysis of CRNs usually requires the additional use of some properties, **P**, which further characterize the species and reactions in the network. For example, the reaction thermodynamics (such as reaction free energy, ΔG) and kinetics (for instance, free energy barrier, $\Delta G^{\ddagger}$, or rate coefficient, k) are often used to determine which reaction pathways are likely to proceed, as well as the stable or metastable species under conditions of interest. Reaction conditions—for instance, the need for a reaction to proceed in a particular solvent or in the presence of a catalyst—can also be thought of as reaction properties. Some properties may be especially necessary or useful in specific domains. For example, reaction yield, chemical safety, and precursor or process cost are relevant descriptors in synthetic applications. While the term 'reaction network' is overloaded in the literature, the general definition that we provide clarifies that even seemingly disparate examples, including reaction (hyper)graphs (Fig. 1a), energy diagrams (Fig. 1b), time dynamics (Fig. 1c) and ab initio surface explorations (Fig. 1d), are all fundamentally CRNs.

In recent years, there have been growing efforts to develop methods for the automatic exploration and characterization of CRNs using computational techniques[21]. At the same time, machine learning (ML) methods applied to chemical reactivity have exploded in popularity[22]. With CRNs being used to study ever-more complex chemical systems and increasing reliance on data-driven methods, integration between CRNs and ML (CRN-ML) is becoming both natural and critical. In this Perspective, we discuss developments in CRNs and ML, focusing on synergistic CRN-ML methods to explore reactive processes. We begin by surveying the literature, discussing various ways that CRNs can be constructed and analyzed. We then consider how ML can be used both as a data source or selection strategy for CRN inputs (meaning, **S**, **R** and **P**) and to aid CRN analysis yielding useful outputs (meaning, pathways, mechanisms and dynamics). Our discussion concludes by considering the rich opportunities for future development in CRN-ML,

[1]Chemical and Biomolecular Engineering, University of Houston, Houston, TX, USA. [2]Energy Technologies Area, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [3]Materials Science Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [4]Materials Science and Engineering, University of California, Berkeley, Berkeley, CA, USA. [5]Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [6]Chemical and Biomolecular Engineering, University of California, Berkeley, Berkeley, CA, USA. [7]Electrical Engineering and Computer Science, University of California, Berkeley, Berkeley, CA, USA. [8]Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. e-mail: kapersson@lbl.gov

https://doi.org/10.1038/s43588-022-00369-z

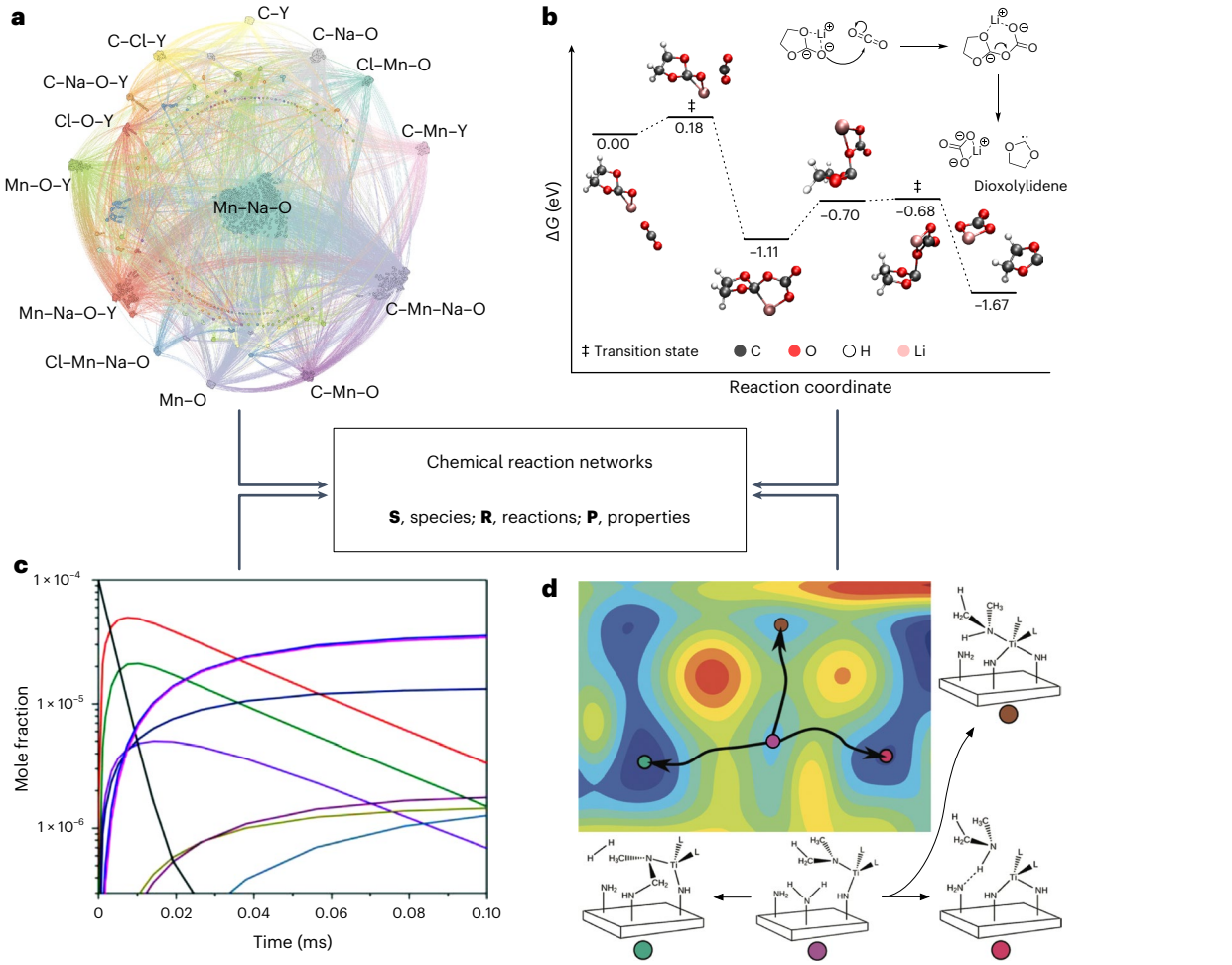

Fig. 1 | Diverse examples of CRNs. a, Illustration of a CRN for modeling the solid-state synthesis of $Y_2Mn_2O_7$ within the C–Cl–Mn–Na–O–Y chemical system. Nodes represent reactants and products, edges represent chemical reactions, and color indicates the chemical subsystem of the reaction. **b**, A portion of an electrochemical CRN depicted as an energy diagram in which the dioxolylidene carbene can be formed by the reaction of a doubly reduced lithium ethylene carbonate Li^+EC^{2-} with CO_2. **c**, Time dynamics of a combustion CRN, where lines depict simulated species concentration profiles, and realistic pathway competition and transient intermediate formation and consumption are observed. **d**, PES exploration during the construction of a catalytic CRN, where an initial reactant PES minimum (small purple circle, middle) is found to connect with three different product PES minima (small green, brown and red circles at the left, top and right of the PES, respectively) via three distinct paths that each traverse a different single reaction barrier. Panel **a** reproduced with permission from ref. 18 under a Creative Commons licence CC BY 4.0. Panels adapted with permission from: **b**, ref. 39 under a Creative Commons licence CC BY 4.0; **c**, ref. 149, Wiley; **d**, ref. 150, RSC.

and the challenges that the community must address before these opportunities can be realized.

Foundations of CRNs

At their core, CRNs are defined by a set of species **S**, a set of reactions **R** and, frequently, a set of properties **P**. However, this apparently simple structure obscures the many choices that must be made when constructing and/or analyzing CRNs. To construct a CRN (Fig. 2a), one must choose a strategy to construct the sets **S**, **R** and **P**, as well as a data source (mainly to populate **P**). The population of **S**, **R** and **P** can occur all at once or can be completed iteratively, adding batches of species and reactions over several generations. Once networks have been constructed, they can be analyzed (Fig. 2b) to obtain varied insights ranging from the key species involved in a chemical process, reaction pathways from initial reactants to species of interest and system time dynamics. Here we highlight these choices to provide the foundation motivating later discussion of ML applications in CRNs, providing specific examples of different CRN construction and analysis approaches and their associated challenges or limitations.

CRN construction

Species and reactions. In perhaps the simplest method of CRN construction, individual species and reactions are compiled in a bespoke or heuristic manner guided by chemical intuition and application-specific expertise[23,24]. This manual approach has the benefit that all species and reactions included in the network are presumed to be relevant to the process of interest. However, due to the human effort required and the reliance on intuition or comprehensive characterization, this strategy has limited predictive capacity and is constrained to small systems.

For studies of systems that cannot be easily probed experimentally or that involve many species and reactions, automated methods are essential. These methods broadly fall into two categories: those involving potential energy surface (PES) exploration and those which systematically enumerate species and reactions based on predefined rules. PES exploration techniques[25–27] use density functional theory (DFT) and related quantum chemical theories (for instance, wavefunction methods) to identify reactions proceeding from reactants of interest to various metastable intermediates and products. While PES exploration allows for the unbiased discovery of species and reactions

https://doi.org/10.1038/s43588-022-00369-z

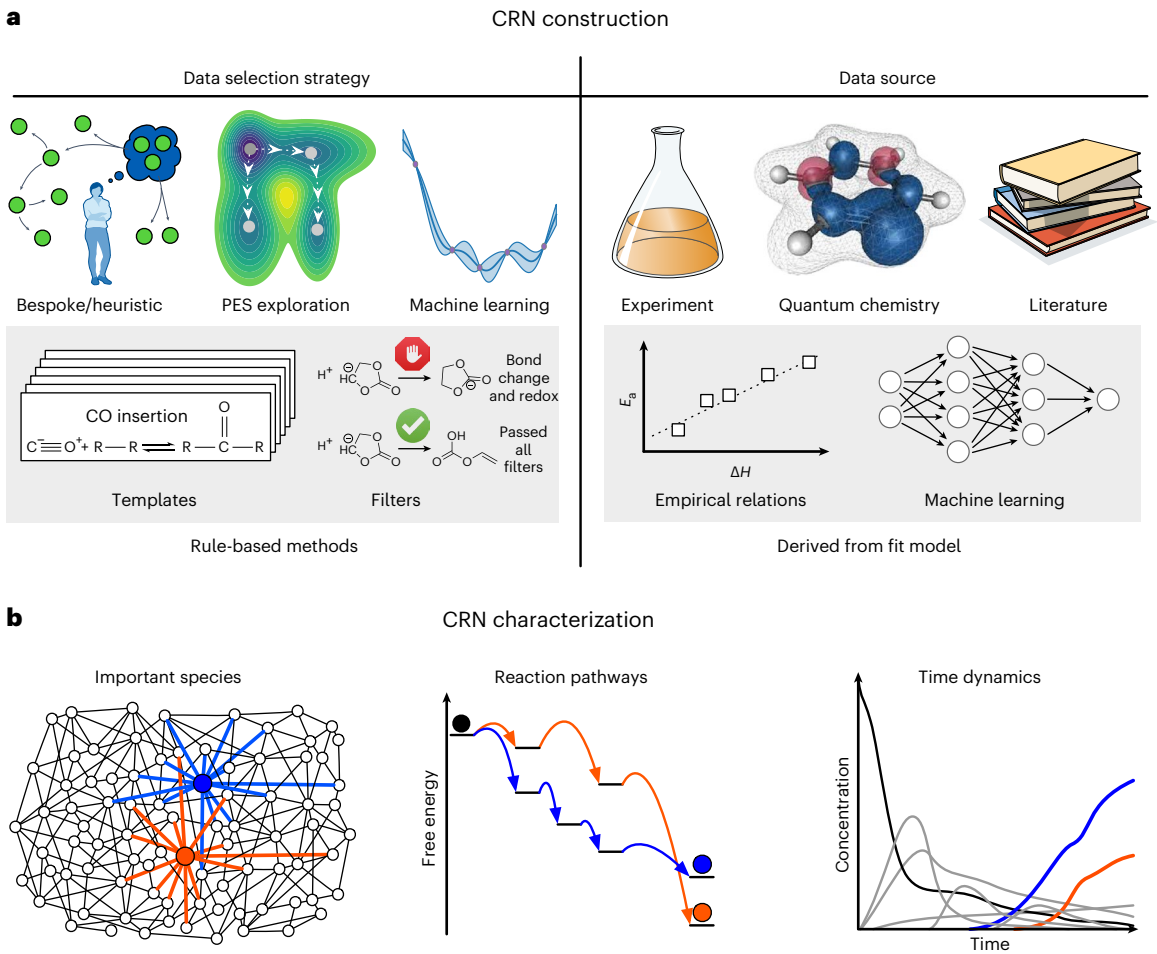

Fig. 2 | Construction and characterization of CRNs. a, CRN construction involves both a data selection strategy and a data source. Data selection strategies include bespoke or heuristic intuitive choice, quantum chemical PES exploration, ML or rule-based methods (meaning, templates and filters). Primary data sources include experiment, quantum chemistry and previous literature while secondary data sources derived from fit models include empirical relations (for example, group additivity for thermodynamics, linear scaling relations for kinetics) and ML models. **b**, Different methods of CRN characterization can yield useful insights into important species, reaction pathways and time dynamics. E_a, activation energy; ΔH, reaction enthalpy.

in complex systems, the high cost of ab initio quantum chemical calculations makes such approaches useful for only relatively small molecules (meaning, less than 10 heavy atoms) or reactions on very short timescales (meaning, about 10 ps). Zhao and Savoie[28] recently reduced the computational burden of PES exploration by combining DFT with cheaper semi-empirical methods in Yet Another Reaction Program (YARP). Despite its effectiveness for neutral organic molecules, even in the presence of catalytic surfaces[29], we note that YARP is limited by its semi-empirical engine, which is not reliable for charged and open-shell species[30].

When PES exploration is not feasible (or when elementary reaction steps are not needed), it is most common to use a set of rules to define **S** and **R**. In domains where reaction mechanisms are well characterized, such as organic chemistry[2,31,32], heterogeneous thermocatalysis[33,34], prebiotic chemistry[35,36] and biochemistry[37,38], reaction templates are often used. Such templates prescribe how molecules containing certain structural motifs can transform into other species. By successively applying these templates, one simultaneously defines new reactions and species via the templated products.

Because reaction templates are typically designed for specific, well-studied chemical systems (for example, aqueous chemistry), they cannot be applied universally. When exploring a novel or exotic type of chemistry, key reaction mechanisms may not even have been identified yet. An alternative approach is to use filters, prescribing what should be excluded (rather than included) from the network given some initial sets of species and reactions. For example, a filter could be applied such that no reaction involving more than a certain number of bonds forming or breaking should be included[3]. Recently, Barter et al. devised High-Performance Reaction Generation (HiPRGen)[39], which allows for user-defined species and reaction filters to construct CRNs following comprehensive reaction enumeration. A potential limitation of this filtering approach is that a set of species relevant to the chemical system of interest must be known at the time of network construction.

Like PES exploration, rule-based construction of **S** and **R** suffers from substantial drawbacks. As templates are, by definition, reactive motifs that have been previously observed, the use of templates biases CRNs towards well-studied chemistry and limits the ability of a CRN to discover novel reaction mechanisms. Templates are therefore inappropriate in domains where reaction mechanisms have not been thoroughly characterized. Even in domains where rule-based methods are appropriate and widely applied (for instance, organic synthesis), there is no guarantee that a reaction produced via a template or accepted through a set of chemical filters will actually occur. This is a substantial limitation, as even experts cannot easily predict which species and reactions will prove to be exceptions to the rules[40]. Accordingly, caution is always required when constructing and employing a rule-based CRN.

Regardless of how a network has been constructed, whether in one shot or iteratively, using templates, filters or PES exploration, network

incompleteness must be considered. In all but the most simple chemical systems, it is impractical to enumerate all possible chemical reactions; as a result, nearly every reported CRN is incomplete. However, the degree to which that incompleteness impacts the utility of the CRN depends on the chemical application and context. In retrosynthesis, CRN expansion is typically limited to the most promising reactions leading from a species of interest to some commercially available or easily synthesized precursors[41]. This incompleteness—ignoring irrelevant reactions that do not contribute to the desired synthesis—is known, acceptable and even advantageous to avoid scaling limitations. At the same time, there can also be unknown and undesirable forms of CRN incompleteness. If the set of templates used to generate a retrosynthetic CRN is flawed, missing reactions could prevent the identification of any viable synthetic paths to a target or cause a longer or more costly path to be found instead of the true optimal path. Considering the dynamics of a complex system, the absence of a reaction could yield only a small deviation in species concentrations or could fundamentally change the reactive competition, perhaps leading to the predicted formation of entirely different products. To resolve these problems, methods to identify and/or quantify network incompleteness are needed. Techniques to selectively and minimally expand a CRN, aiming to make a network more complete as the application demands without dramatically increasing network size, are also essential.

Reactive properties. Often, the process of obtaining necessary properties (for example, ΔG) occurs concurrently with the selection of **S** and **R**, meaning that the choice of technique used for the construction of **S**, **R** and **P** are frequently coupled. However, this is not a requirement, and properties can also be obtained either before or after species and reaction selection.

Reactive properties can be sourced from experiments, quantum chemistry, literature sources or fit models (specifically, empirical relations or ML). For a sufficiently small network composed of well-separated steps (meaning, not a reaction cascade), it may be possible to obtain experimental reaction energies or rates for all reactions. Recent advancements in high-throughput experimentation using a robotic platform[42,43] offer an exceptional avenue to expand the use of experimentally obtained reaction properties in CRNs. Even if this is not possible, the experimental literature can still be used to provide useful approximations. For example, Wołos et al.[36] surveyed the prebiotic chemistry literature to categorize different types of reaction based on their yields, ranging from trace (<3%) to high (≥80%).

When templates are used, it is often possible to apply fit models to approximate both thermodynamic and kinetic properties. Perhaps most famously, the Reaction Mechanism Generator (RMG)[2] leverages Benson group additivity[44] to estimate species thermodynamic properties (which are then used to calculate reaction thermodynamics) and combines databases of known rate coefficients with the Bell–Evans–Polanyi linear scaling relation[45,46] to predict reaction kinetics. Group additivity has also been exploited to predict reaction thermodynamics in metabolic networks[47]. This approach allows for the rapid prediction of rates for reactions following common organic mechanisms. When applied within sufficiently narrow families of molecules and reactions, group additivity and linear scaling relations can achieve admirable accuracy[48,49]. However, such trends frequently break down, even in relatively simple cases (for instance, single-atom chemisorption on transition-metal surfaces)[50]. Moreover, it is worth noting that the use of fit models relies on the availability of ample data on relevant species and reactions, meaning that such methods cannot be relied on for property prediction in sparsely explored domains.

Quantum chemistry is often applied to compute various properties of species and reactions, most importantly reaction (free) energies and energy barriers[51]. Importantly, DFT and related techniques can be employed even if PES exploration is not used to define **S** and **R**. For example, in their CRN to study solid-state materials synthesis pathways, McDermott et al.[18] used a combinatorial, filter-based approach for reaction enumeration from a known set of material compositions; they then determined reaction free energies by referencing DFT-calculated formation energies in the Materials Project database[52].

When relying on non-experimental sources for reaction properties, the role of the environment must be carefully taken into account. It is well established that the presence of interfaces can substantially affect reaction thermodynamics and kinetics[53,54]. In addition, many reactions are influenced by solvent effects[55–57] and the concentration of reactive and innocent species in solution. Notably, pH can have a tremendous impact on aqueous reactivity[58,59]. Treating these effects using quantum chemistry, namely by including explicit interfaces, explicit solvation shells and/or by calculating reaction properties under different conditions (for example, in the presence of hydronium or hydroxide ions, to simulate environments with different pH) can be computationally demanding. This therefore motivates efforts to develop low-cost methods to account for complex environmental effects.

CRN characterization

One of the most common applications of CRNs is to answer the question 'How might this species form?'. This amounts to searching for reaction pathways from initial reactants to the species of interest. A common pathfinding approach is to represent a CRN as a graph and use shortest-path algorithms. When the CRN consists only of reactions of the type A ⟶ B, a simple directed graph with nodes representing species and edges representing reactions (generally in the direction reactants ⟶ products) will suffice[3]. Because edges in conventional graphs cannot link more than two nodes, however, they cannot be used to treat more complex reactions with multiple reactants and/or products (for instance, A + B ⟶ C + D). For a more general treatment, CRNs must be represented either as a bipartite directed graph with separate species and reaction nodes[6,7,14] or as a directed hypergraph[60], in which edges can connect an arbitrary number of nodes.

To use a shortest-path algorithm on a CRN (hyper)graph representation, one must define how the cost of a reaction is calculated. This cost function could be based on chemical parameters such as the reaction thermodynamics[6,18], kinetics[61] or yield[36], or other properties such as the cost of reagents[41]. We note that choosing the cost function is one of the most challenging tasks in CRN pathfinding and there is no ideal universal cost function; instead, the choice of cost function depends on the available reaction descriptors and the application of interest. Common pathfinding algorithms are computationally expensive, scaling linearly or superlinearly with the size of the CRN[62,63]; they are, therefore, inappropriate for analysis of massive CRNs, such as those that emerge in organic retrosynthesis[14]. To overcome these scaling limitations, tailored search algorithms have been devised that can combine searches using multiple cost functions and employ a beam search with multiple priority queues to strategically limit the scope of the search and enhance robustness[41]. Stochastic methods, using either Monte Carlo tree search[64–66] or the Gillespie algorithm[39,67], offer another way to improve the over conventional shortest-path algorithms by efficiently sampling the reactive space with a focus on the most promising pathways.

When the species of interest in a network are not known a priori, it becomes important to identify key intermediates and products. A common approach is to interrogate the structure of the CRN. Like the Internet or social networks, CRNs often display a scale-free architecture[68]; for instance, the network of all organic reactions is scale free[69], as are some biochemical CRNs[70]. In a scale-free network, the fraction of species with k connections (equivalently, the fraction involved in k reactions, or in graph terms, the fraction of nodes with degree k) is described by a power law $P(k) \approx k^{-\gamma}$, where the exponent γ is a positive real number[68]. Such networks display 'hubs', key species with many connections that control reactive processes. By counting the degree of

each species in a CRN, Stocker et al.[71] identified hub species in natural gas combustion; a similar approach was taken by Wołos et al. in their analysis of prebiotic synthesis[36]. Thinking beyond individual important species, network structure has been used to identify a 'core' of organic chemistry[72], that is, a relatively small set of **S** and **R** that can be used to effectively reach the vast majority of other species in a small number of transformations. When system dynamics are available (see next paragraph), it is also possible to define the natural products of a CRN—species that are created in abundance and are formed much more than they are consumed—which can guide pathfinding and enable mechanistic discovery[39].

Reactive systems described by CRNs are dynamical, evolving over time as reactions occur. However, when reactive steps can be carried out in a controlled stepwise manner, reaction pathways fully prescribe how to transform reactants to end products, and many aspects of the dynamics can be safely abstracted away. In contrast, for applications involving reactive cascades where many reactions occur simultaneously and the individual steps of a chemical process cannot be easily separated (for example, via purification), such abstractions are insufficient. Rather, it is often essential to directly study temporal dynamics to capture reactive competition and understand the ultimate product(s) of a cascade process[73–75].

There are two main approaches to characterize the dynamics of a chemical system described by a CRN. In one, coupled rate equations are solved to determine the concentrations of species in **S** (ref. 76) or the probabilities of system states being occupied[77] as functions of time. As a simple example, for a CRN with $\mathbf{S} = \{A, B, C\}$ and **R** defined by

$$A \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} B \qquad B \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} C, \tag{1}$$

the changes in species concentrations can be expressed as the following set of differential equations

$$\begin{aligned} \frac{\mathrm{d}[A]}{\mathrm{d}t} &= -k_1[A] + k_{-1}[B] \\ \frac{\mathrm{d}[B]}{\mathrm{d}t} &= k_1[A] - k_{-1}[B] - k_2[B] + k_{-2}[C] \\ \frac{\mathrm{d}[C]}{\mathrm{d}t} &= k_2[B] - k_{-2}[C] \end{aligned} \tag{2}$$

where $[A]$, $[B]$ and $[C]$ are the species concentrations, and k_n and k_{-n} represent forward and reverse reaction rate coefficients, respectively. As an alternative to using coupled rate equations, one can employ stochastic methods such as kinetic Monte Carlo[78]. In these approaches, the system state is evolved one reaction at a time using random numbers weighted by reaction propensities (related to reaction rates)[67]. Both coupled differential equation methods and stochastic methods offer (in principle) exact solutions to the dynamics of a reactive system, although the algorithmic simplicity and the often comparably low cost of stochastic methods make them attractive for simulations involving vast numbers of species and reactions.

In many chemical applications, reaction rate coefficients often vary by many orders of magnitude. Rapid reactions cause species concentrations—and thus reaction rates—to change quickly relative to the timescales of interest (which will often be determined by more rare events or slow reactions), making it challenging to propagate CRN system dynamics either by solving coupled rate equations or using a stochastic approach (the dynamics are 'stiff'[79]). Simulations on stiff systems frequently require small time steps to capture the phenomena being modeled. As many applications require solving CRN dynamics repeatedly (for instance, decision-making and parameter estimation), numerical approaches can be computationally inefficient and can even become intractable.

Posing further problems for dynamical CRN studies, kinetic data are often limited in both availability and quality. As with other reaction properties, rate coefficients can be obtained by experiment or simulations (from calculated energy barriers), or could be derived using, for instance, scaling relations. In many domains, experimental rate coefficients are few (although high-throughput experiments could change this), meaning that computational methods are often relied on. Exceptional high-accuracy quantum chemical methods, such as coupled-cluster and multi-reference approaches, applied to study relatively simple reactions (for example, small-molecule reactions in the gas phase) can predict rate coefficients within 5–10% of experimental values[80]. In general, however, the quantitative accuracy of calculated rate coefficients leaves much to be desired[81].

Application of ML to CRNs

Modern ML methods have substantially expanded a chemist's toolbox, enabling data-driven modeling without relying heavily on expert analysis and chemical intuition. Combined with CRNs, ML methods have recently been applied to understand complex chemical systems. In this section, we discuss current applications and future opportunities of ML methods for CRN data selection, as a CRN data source and for CRN characterization, aiming to overcome the challenges laid out in the previous section.

ML for CRN data selection

Learning potential energy surfaces. A remarkable accomplishment in chemical ML is the development of surrogate ML interatomic potentials or force fields to approximate a PES. Different ML regression algorithms have been applied to develop interatomic potentials, including linear regression[82,83], kernel methods[84], neural networks (Fig. 3a)[85–87] and graph neural networks[88,89], among others. ML interatomic potentials can achieve a balance between accuracy and running speed, and thus have been widely applied to a number of systems, ranging from small molecules to crystals and biomolecules (see refs. 90,91 for recent topical reviews). Despite these successes, designing models to account for more diverse chemical space (for example, a system with dozens of atomic elements)[92,93] and to incorporate more complex interactions (including electrostatics[94,95], van der Waals forces[96] and magnetic states[97]) is less resolved and still a challenging task. These are areas of active, ongoing research.

In the context of CRNs, ML interatomic potentials can be used in place of a quantum chemical PES to identify species, **S**, and reactions, **R**, in an exploratory fashion. For example, Zeng et al.[98] performed reactive molecular dynamics simulations using an ML interatomic potential to study methane combustion, where reaction mechanisms were extracted via analysis of observed reactive events. The obtained methane combustion CRN from the extracted reactions was in excellent agreement with experimental observations. The capacity for interatomic potentials to extrapolate outside of their training data, however, remains questionable, particularly for complex systems[90], which calls into question their utility for predictive CRN construction via ML PES exploration for systems with poorly understood reactivity. For such systems, it may be more appropriate for ML interatomic potentials to be built on the fly and iteratively improved to accelerate the exploration of a quantum chemical PES. This approach has been demonstrated for molecular structure relaxation and transition state search in the GPMin (Gaussian process minimizer)[99] and ML-NEB (ML nudged elastic band)[100] methods, respectively.

Refining reaction properties. In the 'Reactive properties' section, we discussed how some reaction properties, particularly kinetic properties, are difficult to accurately predict, requiring the use of time-consuming experiments or costly computational methods. Systematically improving the quality of all reaction property data in a CRN (using experimental data instead of calculated data, or using a more advanced level of theory for calculations) might be prohibitively expensive.

https://doi.org/10.1038/s43588-022-00369-z

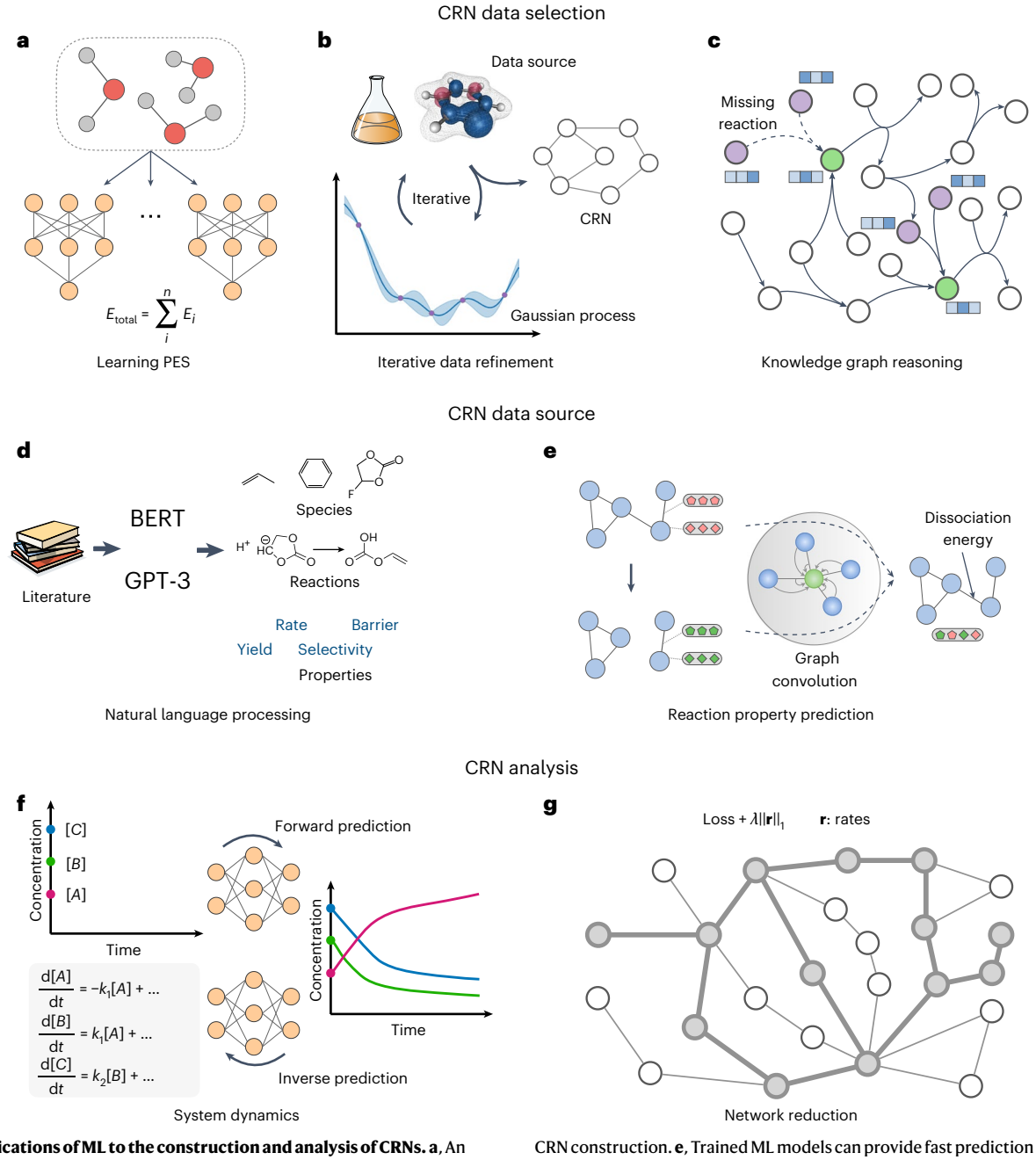

Fig. 3 | Applications of ML to the construction and analysis of CRNs. a, An ML interatomic potential model learns a PES, where the total potential energy E_{total} of a system of n atoms is obtained as the sum of individual atomic energies E_i. The PES can be explored to select species and reactions for a CRN. **b**, A CRN can be iteratively built by analyzing reactions and their properties with high uncertainty and then incorporating the corresponding refined data into the CRN. **c**, Knowledge graph reasoning can be used to identify missing reactions within a CRN. For example, given that the embeddings (blue squares) of the green molecules (products) are similar and that those of the purple molecules (reactants) are also similar, it is likely that there is a missing reaction (dashed lines) in the CRN. White circles and solid lines: other molecules and reactions. **d**, NLP can extract species, reactions and properties from the literature to aid in CRN construction. **e**, Trained ML models can provide fast prediction of reaction properties. For example, a graph neural network can combine the atom and bond features of the reactant and product molecules and then map the updated features to the reaction energy. **f**, Physics-informed neural networks can help to solve differential equations (for example, equation (2)) to evolve a reactive system state over time or can learn the CRN and the form of its dynamical equations from observed reactive trajectories. **g**, Sparse learning approaches using regularization can be employed to identify the skeleton of a CRN by eliminating unimportant species and reactions (for example, the light gray species and reactions) without affecting the model outcome. The regularization can be achieved by adding an L_1-norm term on reaction rates $\mathbf{r}$ to the loss function, with a Lagrangian multiplier λ to control the regularization strength.

An alternative would be to leverage ML to quantify the uncertainty in each datum via Gaussian processes[101] or Bayesian neural networks[102]. Data points with large uncertainty are considered less reliable and then selected for further analysis. For instance, this may involve calculating the properties of some species or reactions using DFT (Fig. 3b). In a study using heterogeneous catalysis CRN to investigate syngas formation on rhodium, Ulissi et al.[4] employed a Gaussian process to predict adsorption energies, which was combined with simple linear scaling relations (similar to the Bell–Evans–Polanyi relation discussed in the 'Reactive properties' section) to determine rate-limiting reactions to be more accurately calculated by DFT. Another approach, transfer learning, can also be used to refine reaction properties.

It first pretrains a model using low-quality but easily obtainable data (for example, from classical and/or semi-empirical quantum chemical calculations) and then fine-tunes the pretrained model on more limited high-quality data (experimental data or highly accurate DFT or wavefunction calculations). Transfer learning has already been widely applied to predict the structures of reaction transition states[103], reaction energy barriers[104] and rate coefficients[105,106]. However, to our knowledge, transfer learning has not been utilized to populate a CRN with reaction properties.

Addressing CRN incompleteness. After an initial CRN is constructed, it may be necessary to expand the network to address incompleteness (see also the 'Species and reactions' section). To maximize the efficiency of network expansion, it is desirable to train a model to suggest what data should be included in the CRN while acquiring such data using experimental measurements, quantum chemical calculations or ML. In some applications, particularly synthesis planning, CRNs gradually expand, with new species and reactions added to progress towards a well-defined goal such as a set of commercially available precursors. This could be achieved by iteratively expanding the network on the head species node[32], where one can determine which reactions to add by a cost function. In addition to the chemically informed cost functions discussed in the 'CRN characterization' section, ML models trained on large sets of literature reactions provide an alternative approach for cost function design[40,107]. Such models output a probability score for each compatible reaction and then select the most probable reaction for network expansion. However, ML models trained on only the chemical literature are not optimal because literature datasets are biased by the popularity of particular reactions[108]. This can be resolved by training on both literature data and expert-coded reaction rules, as demonstrated by the Grzybowski group[16,108]. As suggested by the recent work of Lan and An[109], deep reinforcement learning, which trains an 'agent' to make decisions based on a learned 'policy' (a function aiming to maximize a reward based on an objective function), could provide yet another means to select the most promising species and reactions to add to a network. Lan and An constructed their network describing ammonia synthesis on iron with no knowledge of reaction intermediates or mechanisms; however, if an existing CRN can be used to cheaply learn an initial policy, application to CRN expansion should be straightforward.

We anticipate that the ML CRN expansion can be further improved by using knowledge graphs. Knowledge graphs utilize graph-like data structures to store interlinked descriptions of entities (nodes) and their relations (edges)[110,111]. A common approach to using knowledge graphs is to generate embedding vectors of the species (nodes) and reactions (edges) while preserving their semantic meaning via scoring functions. The embeddings could then be used to identify missing links between the species nodes, in other words, identifying missing reactions that are not present in the network (Fig. 3c). This approach provides a systematic method to address the network incompleteness problem. Knowledge graph embeddings can also be used to assist other CRN tasks, such as ranking the species in a CRN to find key hubs (see the 'CRN characterization' section), or predicting reaction properties (see the 'Reaction property prediction' section). The major benefit of learning on a CRN knowledge graph is that this makes it possible for a learning algorithm to explicitly take advantage of the structure of the entire reaction space, which is missing if learning on only individual reactions. Learning on CRN graphs, however, requires new metrics to reflect the semantic meaning in reactions, which we believe should be task specific and carefully designed by domain experts.

ML as a CRN data source

Natural language processing. The scientific literature contains a wealth of prior experimental/theoretical data that may be utilized as a data source for CRN construction. While manual 'digitization' of CRN data from literature sources is possible and has been performed (for instance, from the origin of life literature by Wołos et al.[36]), this human-guided process is labor intensive and is not easily scalable to other chemistry domains, such as inorganic materials synthesis. This challenge has already been addressed by ML methods through the development of natural language processing (NLP) models for text extraction, such as BERT (bidirectional encoder representations from transformers)[112], GPT (generative pretrained transformer)[113] and models derived from these for application to scientific domains (for example, MatBERT[114] for materials science text) (Fig. 3d). Several literature-derived CRN datasets have already been created using these NLP approaches. For example, Kononova et al.[115] compiled an inorganic synthesis dataset consisting of over 4 million papers and over 188,000 paragraphs describing experimentally performed syntheses, extracting species, reactions and processing steps from each. Tshitoyan et al.[116] extracted species (together with other inorganic materials science vocabulary) from 3.3 million abstracts and generated word embeddings for them, which can be further leveraged for property prediction and reaction discovery. The Cronin group has designed not only an NLP-based tool (SynthReader)[117] for extracting synthesis procedures from the literature, but also a 'chemical programming language' (the XDL format)[118] for executing these steps on a robotic lab platform in a standardized fashion.

Despite the availability of these NLP-extracted datasets, there is little previous work exploring the construction and analysis of CRNs created with them as a primary data source. One of the challenges in using the literature as the main source of data for a CRN is that it introduces human bias by limiting the scope of chemical complexity considered to only that which has previously been observed, similar to the constraints of using prescriptive templates. It follows that a promising (and largely unexplored) opportunity for the use of these literature-derived datasets is the comparison between experimental and theoretical CRNs. By analyzing differences between theoretical CRNs and literature-derived CRNs, researchers may target areas where observed phenomena cannot yet be theoretically explained, or identify experimentally unexplored chemical spaces and new synthesis approaches that are predicted to be fruitful based on calculations. This type of comparison has become increasingly enabled in organic chemistry by the recent development of robotic lab platforms that can rapidly perform experiments in an automated fashion[118], including in a self-driving laboratory context[119]. These systems make it possible to not only validate and perform synthesis protocols extracted from the experimental literature[120] but also to accumulate sufficient reaction data to develop more comprehensive experimental CRNs, including those which contain 'negative' reaction data from failed experiments, which has been shown to greatly enhance the performance of ML models predicting reaction outcomes[121].

Reaction property prediction. Given a set of reactions, **R**, and a property of interest associated with each reaction, ML regression models can be to applied to approximate the relationship between each reaction and its associated property. Unlike ML interatomic potentials ('Learning potential energy surfaces' section), which need to satisfy symmetry requirements and can only take atomic number and coordinates as input[122], ML models for reaction property prediction (Fig. 3e) are more flexible. They can take a much wider set of features as input and have been applied to a variety of thermodynamic and kinetic reaction properties, such as reaction energies[71], bond dissociation energies[123,124], reaction energy barriers[104,125,126] and rate coefficients[127,128]. Once trained, a model can act as a data source and thus be employed to rapidly predict the properties of unseen reactions (kinetic properties are of particular interest), enabling the investigation of large-scale CRNs consisting of tens of thousands of reactions. For example, assuming a constant activation barrier of 0.3 eV for all reactions and using ML-predicted reaction energies, Stocker et al.[71] were able to perform

mean-field microkinetic simulation for a CRN containing 21,393 elementary reactions to study methane combustion. The flexibility in the choice of ML algorithms and training data for property prediction models makes them prolific in the literature. However, it can be very difficult for a user to select an existing model for CRN applications because model performance depends heavily on the ML algorithm, test data and, of course, the property to predict. Benchmarking these models on common data is in urgent need.

As discussed in the 'Reactive properties' section, environmental influences such as interfaces and solvent on reaction properties must be carefully considered when using them in CRNs. Environmental influences can be explicitly modelled by ML property prediction algorithms. For example, the FieldSchNet can model the interaction of molecules with arbitrary external fields, which enables it to describe implicit and explicit molecular environments, operating as a polarizable continuum model for solvation[129]. Models without an algorithmic consideration of environmental effects can still be used to predict properties of reactions in specific environments. They, however, would be best trained using a transfer learning approach (see 'Refining reaction properties' section), pretrained and fine-tuned on data without and with environmental effects, respectively, because the former is much easier to obtain.

ML for CRN analysis

System dynamics. Solving for the concentrations of different chemical species over time is known as the 'forward' solution of chemical dynamics (see also 'CRN characterization' section). One can also solve the 'inverse' problem where, given observed system dynamics, one seeks to recover the underlying reactions or rate equations (Fig. 3f). ML provides potential opportunities to solve both of these problems more efficiently. While ML could also be applied to stochastic methods such as kinetic Monte Carlo, here we focus on ML applied to methods based on coupled differential equations.

In the context of CRNs, preliminary work has looked at stabilizing neural network gradient calculations by scaling model outputs to mitigate the challenges associated with stiff dynamics (see 'CRN characterization' section)[130]. In addition, differential equations representing physical invariances can be added as 'soft' constraints to an ML objective function, thus penalizing the ML model to satisfy it[131]. Another approach to solving CRN dynamics is to use physics-informed ML models employing concepts such as quasi-steady-state kinetics to reduce the stiffness of the system and then train the ML model under the imposed soft constraint[132]. However, there are still many challenges associated with developing physics-informed models. For instance, optimization during the training process can be challenging and many ML models, such as neural networks, can struggle to converge[133,134]. They may also not preserve the correct inductive biases (for instance, continuity or conservation of energy), which may not be immediately apparent from standard ML training and testing pipelines without devising specific robustness testing strategies[135,136]. One potential direction forwards is to enforce physical constraints more precisely by incorporating differentiable numerical simulations into the training procedure via implicit differentiation[137,138]. On the efficiency side, instead of solving one specific parameterized differential equation at a time, it may be more fruitful to deduce the full family of equations by learning the mapping between parameterized differential equations and their solutions (and vice versa)[139,140]. The ML techniques applied to more accurately model highly nonlinear systems (such as chaotic systems[141]) may also provide insight for modeling CRNs displaying such behavior. Developing better ML techniques to overcome these challenges will be crucial to solving and understanding dynamic behavior in CRNs, particularly over long timescales.

Model reduction. Not every species and reaction in a CRN is equally important; as noted in the 'CRN characterization' section, CRNs often have a small number of highly connected—and therefore important—'hub' species and a larger number of peripheral species that participate in few reaction pathways. To improve the efficiency of CRN analysis, it is therefore useful to perform model reduction, eliminating species and reactions that have little or no effect on the outcome, thus yielding a simplified CRN while retaining the accuracy of the more extensive network. While model reduction is generally useful for accelerating CRN analysis, it has been most widely applied in the past to improve simulations of network time dynamics. Conventional model reduction methods include sensitivity analysis, timescale exploitation approaches and singular-value-decomposition-based approaches, among others[142].

ML techniques offer a data-driven approach for CRN model reduction. One such approach is to formulate model reduction as a sparse learning problem that optimizes the reaction rates and introduces regularization terms to enforce sparsity (Fig. 3g). For example, least-squares optimization with L1-norm regularization[143] and L2-norm regularization[144] can be used to identify reduced CRN systems. However, these methods are limited to data obtained from the equilibration phase, and are thus unable to recover the reaction dynamics. Katsoulakis and Vilanova[145] instead used variational inference, learning the probability distribution of different states in biochemical reaction networks. This approach allows for a simultaneous sensitivity analysis and optimization of a reduced network; moreover, variational inference allows one to perform stochastic sampling of a reactive space much more efficiently than Monte Carlo methods[146]. Rather than directly learning the dynamics of the complete or a reduced network, as in the 'System dynamics' section, Wang et al.[147] used a deep neural network to learn the error between the exact model and a guess reduced model. This predicted error allowed the authors to intelligently select which reduced model to select next for evaluation, accelerating the reduced model optimization by many orders of magnitude. A similar approach using Gaussian process ML to develop a surrogate model for stochastic CRNs was earlier conducted by Singh and Hellander[148].

We note that, for the most part, the ML methods discussed here have been successfully applied to only relatively small CRNs and toy models. It is therefore unclear what the accuracy and computational expense of ML model reduction may be for large complex chemical systems. The extension and quantification of the methods in these directions needs additional development.

Conclusions and outlook

Developing CRN-ML methodologies is a substantial research challenge demanding creativity and concerted effort. Thus far, applications of ML to CRNs have mainly focused on reducing the computational burden. Most models have been developed to either replace quantum chemical calculations or guide what reaction properties to collect by experiment or computation. These models are abundant largely because they can be relatively easily created by slight adjustment of existing ML methods that are well developed for other chemical applications. Closer integration of ML to address long-lasting challenges that are specific to CRNs (for instance, network expansion and model reduction) has already begun to emerge; however, such integration is not as straightforward and thus is still very sparse. We anticipate that designing new ML methods that take advantage of CRN characteristics (for example, the (hyper)graph structure and the sparsity of the system dynamics) should be a viable path forwards to address such challenges.

A second challenge is the dearth of high-quality data. Despite the advancement in high-throughput experiments and quantum chemical calculations in the past decade, it is still a formidable task to assemble a sufficiently large dataset for CRN-ML problems, especially for complex systems. Emerging ML techniques, such as NLP, active learning with iterative generation of data, geometric deep learning that allows for direct incorporation of chemical and physical constraints, and electronic structures calculations with learned density functionals, have great potential to alleviate the data scarcity problem.

Notably, we expect them to be leveraged in predicting the activation barriers and rates of individual reactions in a CRN, which is notoriously difficult but extremely important.

A third pressing challenge is actually technical in nature. A number of new computational tools and ML frameworks have been developed over the past several years with great potential to be applied to CRN problems. Despite that, at present, there are few standard tools available for the construction and analysis of CRNs. Open-source libraries and repositories are thankfully abundant, but most are research codes tailored to specific applications, limiting widespread utility. Moreover, there are no standard CRN problems that are well suited to benchmarking. We strongly encourage members of the CRN research community to collaborate on general-purpose software for CRNs and to develop open datasets and tasks to facilitate the testing of new CRN methodologies and the benchmarking of CRN-ML models. The development of such standards not only will aid existing research efforts but also may attract ML researchers and computer scientists to study CRNs.

If the existing challenges in combined CRN-ML studies of chemical reactivity can be overcome, we see substantial opportunities to expand the horizon of what is possible in computational studies of chemical reactivity. For systems that are already commonly studied using CRN approaches, ML offers potential avenues to allow for a greater degree of automation and more thorough exploration of chemical space, particularly for long-time processes that can only be reached deep in a chemical cascade. At the same time, ML could open the door for computational CRN studies in domains that cannot currently be tractably studied for reasons of scale (such as polymerization/depolymerization) and complexity (for instance, photoelectrocatalysis).

References

1. Manion, J. A., Sheen, D. A. & Awan, I. A. Evaluated kinetics of the reactions of H and CH_3 with *n*-alkanes: experiments with *n*-butane and a combustion model reaction network analysis. *J. Phys. Chem. A* **119**, 7637–7658 (2015).
2. Gao, C. W., Allen, J. W., Green, W. H. & West, R. H. Reaction mechanism generator: automatic construction of chemical kinetic mechanisms. *Comput. Phys. Commun.* **203**, 212–225 (2016).
3. Kim, Y., Kim, J. W., Kim, Z. & Kim, W. Y. Efficient prediction of reaction paths through molecular graph and reaction network analysis. *Chem. Sci.* **9**, 825–835 (2018).
4. Ulissi, Z. W., Medford, A. J., Bligaard, T. & Nørskov, J. K. To address surface reaction network complexity using scaling relations machine learning and DFT calculations. *Nat. Commun.* **8**, 14621 (2017).
5. Steiner, M. & Reiher, M. Autonomous reaction network exploration in homogeneous and heterogeneous catalysis. *Top. Catal.* **65**, 6–39 (2022).
6. Blau, S. M. et al. A chemically consistent graph architecture for massive reaction networks applied to solid-electrolyte interphase formation. *Chem. Sci.* **12**, 4931–4939 (2021).
7. Xie, X. et al. Data-driven prediction of formation mechanisms of lithium ethylene monocarbonate with an automated reaction network. *J. Am. Chem. Soc.* **143**, 13245–13258 (2021).
8. Centler, F. & Dittrich, P. Chemical organizations in atmospheric photochemistries—a new method to analyze chemical reaction networks. *Planet. Space Sci.* **55**, 413–428 (2007).
9. Heald, C. L. & Kroll, J. H. The fuel of atmospheric chemistry: toward a complete description of reactive organic carbon. *Sci. Adv.* **6**, eaay8967 (2020).
10. Zhang, J.-T., Wang, H.-Y., Zhang, X., Zhang, F. & Guo, Y.-L. Study of short-lived and early reaction intermediates in organocatalytic asymmetric amination reactions by ion-mobility mass spectrometry. *Catal. Sci. Technol.* **6**, 6637–6643 (2016).
11. Williams, P. J. H. et al. New approach to the detection of short-lived radical intermediates. *J. Am. Chem. Soc.* **144**, 15969–15976 (2022).
12. Tyson, J. J. & Novák, B. Functional motifs in biochemical reaction networks. *Annu. Rev. Phys. Chem.* **61**, 219–240 (2010).
13. Wong, A. S. Y. & Huck, W. T. S. Grip on complexity in chemical reaction networks. *Beilstein J. Org. Chem.* **13**, 1486–1497 (2017).
14. Kowalik, M. et al. Parallel optimization of synthetic pathways within the network of organic chemistry. *Angew. Chem. Int. Ed.* **51**, 7928–7932 (2012).
15. Todd, P. K. et al. Selectivity in yttrium manganese oxide synthesis via local chemical potentials in hyperdimensional phase space. *J. Am. Chem. Soc.* **143**, 15185–15194 (2021).
16. Mikulak-Klucznik, B. et al. Computational planning of the synthesis of complex natural products. *Nature* **588**, 83–88 (2020).
17. Wołos, A. et al. Computer-designed repurposing of chemical wastes into drugs. *Nature* **604**, 668–676 (2022).
18. McDermott, M. J., Dwaraknath, S. S. & Persson, K. A. A graph-based network for predicting chemical reaction pathways in solid-state materials synthesis. *Nat. Commun.* **12**, 3097 (2021).
19. Aykol, M., Montoya, J. H. & Hummelshøj, J. Rational solid-state synthesis routes for inorganic materials. *J. Am. Chem. Soc.* **143**, 9244–9259 (2021).
20. Feinberg, M. *Foundations of Chemical Reaction Network Theory* (Springer, 2019); https://doi.org/10.1007/978-3-030-03858-8
21. Unsleber, J. P. & Reiher, M. The exploration of chemical reaction networks. *Annu. Rev. Phys. Chem.* **71**, 121–142 (2020).
22. Meuwly, M. Machine learning for chemical reactions. *Chem. Rev.* **121**, 10218–10239 (2021).
23. Garza, A. J., Bell, A. T. & Head-Gordon, M. Mechanism of CO_2 reduction at copper surfaces: pathways to C2 products. *ACS Catal.* **8**, 1490–1499 (2018).
24. Lees, E. W., Bui, J. C., Song, D., Weber, A. Z. & Berlinguette, C. P. Continuum model to define the chemistry and mass transfer in a bicarbonate electrolyzer. *ACS Energy Lett.* **7**, 834–842 (2022).
25. Maeda, S., Harabuchi, Y., Takagi, M., Taketsugu, T. & Morokuma, K. Artificial force induced reaction (AFIR) method for exploring quantum chemical potential energy surfaces. *Chem. Rec.* **16**, 2232–2248 (2016).
26. Dewyer, A. L., Argüelles, A. J. & Zimmerman, P. M. Methods for exploring reaction space in molecular systems. *WIREs Comput. Mol. Sci.* **8**, 1354 (2018).
27. Simm, G. N., Vaucher, A. C. & Reiher, M. Exploration of reaction pathways and chemical transformation networks. *J. Phys. Chem. A* **123**, 385–399 (2019).
28. Zhao, Q. & Savoie, B. M. Simultaneously improving reaction coverage and computational cost in automated reaction prediction tasks. *Nat. Comput. Sci.* **1**, 479–490 (2021).
29. Zhao, Q., Xu, Y., Greeley, J. & Savoie, B. M. Deep reaction network exploration at a heterogeneous catalytic interface. *Nat. Commun.* **13**, 4860 (2022).
30. Blau, S., Spotte-Smith, E. W. C., Wood, B., Dwaraknath, S. & Persson, K. Accurate, automated density functional theory for complex molecules using on-the-fly error correction. Preprint at *ChemRxiv* https://doi.org/10.26434/chemrxiv.13076030.v1 (2020).
31. Gothard, C. M. et al. Rewiring chemistry: algorithmic discovery and experimental validation of one-pot reactions in the network of organic chemistry. *Angew. Chem. Int. Ed.* **51**, 7922–7927 (2012).
32. Szymkuć, S. et al. Computer-assisted synthetic planning: the end of the beginning. *Angew. Chem. Int. Ed.* **55**, 5904–5937 (2016).
33. Goldsmith, C. F. & West, R. H. Automatic generation of microkinetic mechanisms for heterogeneous catalysis. *J. Phys. Chem. C* **121**, 9970–9981 (2017).
34. Liu, M. et al. Reaction mechanism generator v3.0: advances in automatic mechanism generation. *J. Chem. Inf. Model.* **61**, 2686–2696 (2021).

35. Rappoport, D., Galvin, C. J., Zubarev, D. Y. & Aspuru-Guzik, A. Complex chemical reaction networks from heuristics-aided quantum chemistry. *J. Chem. Theory Comput.* **10**, 897–907 (2014).
36. Wolos, A. et al. Synthetic connectivity, emergence, and self-regeneration in the network of prebiotic chemistry. *Science* **369**, eaaw1955 (2020).
37. Liao, K. H. et al. Application of biologically based computer modeling to simple or complex mixtures. *Environ. Health Persp.* **110**, 957–963 (2002).
38. Wicker, J., Fenner, K., Ellis, L., Wackett, L. & Kramer, S. Predicting biodegradation products and pathways: a hybrid knowledge- and machine learning-based approach. *Bioinformatics* **26**, 814–821 (2010).
39. Barter, D. et al. Predictive stochastic analysis of massive filter-based electrochemical reaction networks. Preprint at *ChemRxiv* https://doi.org/10.26434/chemrxiv-2021-c2gp3-v2 (2022).
40. Segler, M. H., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).
41. Grzybowski, B.A., Badowski, T., Molga, K., Szymkuć, S.: Network search algorithms and scoring functions for advanced-level computerized synthesis planning. *WIREs Comput. Mol. Sci.* https://doi.org/10.1002/wcms.1630 (2022).
42. Granda, J. M., Donina, L., Dragone, V., Long, D.-L. & Cronin, L. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* **559**, 377–381 (2018).
43. Seifrid, M. et al. Autonomous chemical experiments: Challenges and perspectives on establishing a self-driving lab. *Acc. Chem. Res.* **55**, 2454–2466 (2022).
44. Benson, S. W. et al. Additivity rules for the estimation of thermochemical properties. *Chem. Rev.* **69**, 279–324 (1969).
45. Bell, R. P. & Hinshelwood, C. N. The theory of reactions involving proton transfers. *Proc. R. Soc. Lond. A* **154**, 414–429 (1936).
46. Evans, M. & Polanyi, M. Further considerations on the thermodynamics of chemical equilibria and reaction rates. *Trans. Faraday Soc.* **32**, 1333–1360 (1936).
47. Hatzimanikatis, V. et al. Exploring the diversity of complex metabolic networks. *Bioinformatics* **21**, 1603–1609 (2005).
48. Yu, J., Sumathi, R. & Green, W. H. Accurate and efficient method for predicting thermochemistry of polycyclic aromatic hydrocarbons—bond-centered group additivity. *J. Am. Chem. Soc.* **126**, 12685–12700 (2004).
49. Meng, Q. et al. A theoretical investigation on Bell–Evans–Polanyi correlations for hydrogen abstraction reactions of large biodiesel molecules by H and OH radicals. *Combust. Flame* **214**, 394–406 (2020).
50. Vijay, S., Kastlunger, G., Chan, K. & Nørskov, J. K. Limits to scaling relations between adsorption energies? *J. Chem. Phys.* **156**, 231102 (2022).
51. Mardirossian, N. & Head-Gordon, M. Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. *Mol. Phys.* **115**, 2315–2372 (2017).
52. Jain, A. et al. Commentary: The Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
53. Schlögl, R. Heterogeneous catalysis. *Angew. Chem. Int. Ed.* **54**, 3465–3520 (2015).
54. Wei, Z., Li, Y., Cooks, R.G., Yan, X.: Accelerated reaction kinetics in microdroplets: Overview and recent developments. *Annu. Rev. Phys. Chem.* **71**, 31–51 (2020).
55. Heitele, H. Dynamic solvent effects on electron-transfer reactions. *Angew. Chem. Int. Ed.* **32**, 359–377 (1993).
56. Cativiela, C., Garcia, J., Mayoral, J. & Salvatella, L. Modelling of solvent effects on the Diels–Alder reaction. *Chem. Soc. Rev.* **25**, 209–218 (1996).
57. Murzin, D. Y. Solvent effects in catalysis: implementation for modelling of kinetics. *Catal. Sci. Technol.* **6**, 5700–5713 (2016).
58. Eigen, M. Proton transfer, acid-base catalysis, and enzymatic hydrolysis. part I: elementary processes. *Angew. Chem. Int. Ed.* **3**, 1–19 (1964).
59. Cordes, E. & Bull, H. Mechanism and catalysis for hydrolysis of acetals, ketals, and ortho esters. *Chem. Rev.* **74**, 581–603 (1974).
60. Schwaller, P. et al. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci.* **11**, 3316–3325 (2020).
61. Robertson, C., Ismail, I. & Habershon, S. Traversing dense networks of elementary chemical reactions to predict minimum-energy reaction mechanisms. *ChemSystemsChem* **2**, 1900047 (2020).
62. Dijkstra, E. W. et al. A note on two problems in connexion with graphs. *Numer. Math.* **1**, 269–271 (1959).
63. Yen, J. Y. An algorithm for finding shortest routes from all source nodes to a given destination in general networks. *Quart. Appl. Math.* **27**, 526–530 (1970).
64. Browne, C. B. et al. A survey of Monte Carlo tree search methods. *IEEE Trans. Comput. Intell. AI Games* **4**, 1–43 (2012).
65. Wang, X. et al. Towards efficient discovery of green synthetic pathways with Monte Carlo tree search and reinforcement learning. *Chem. Sci.* **11**, 10959–10972 (2020).
66. Lee, K., Woo Kim, J. & Youn Kim, W. Efficient construction of a chemical reaction network guided by a monte carlo tree search. *ChemSystemsChem* **2**, 1900057 (2020).
67. Gillespie, D. T. Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.* **58**, 35–55 (2007).
68. Barabási, A.-L. Scale-free networks: A decade and beyond. *Science* **325**, 412–413 (2009).
69. Grzybowski, B. A., Bishop, K. J. M., Kowalczyk, B. & Wilmer, C. E. The 'wired' universe of organic chemistry. *Nat. Chem.* **1**, 31–36 (2009).
70. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A.-L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
71. Stocker, S., Csányi, G., Reuter, K. & Margraf, J. T. Machine learning in chemical reaction space. *Nat. Commun.* **11**, 5505 (2020).
72. Bishop, K. J. M., Klajn, R. & Grzybowski, B. A. The core and most useful molecules in organic chemistry. *Angew. Chem. Int. Ed.* **45**, 5348–5354 (2006).
73. Marshall, A. T. Using microkinetic models to understand electrocatalytic reactions. *Curr. Opin. Electrochem.* **7**, 75–80 (2018).
74. Vermeire, F. H. et al. Detailed kinetic modeling for the pyrolysis of a jet a surrogate. *Energy Fuels* **36**, 1304–1315 (2022).
75. Spotte-Smith, E. W. C. et al. Toward a mechanistic model of solid-electrolyte interphase formation and evolution in lithium-ion batteries. *ACS Energy Lett.* **7**, 1446–1453 (2022).
76. Zhang, H., Linford, J. C., Sandu, A. & Sander, R. Chemical mechanism solvers in air quality models. *Atmosphere* **2**, 510–532 (2011).
77. Miller, J. A. & Klippenstein, S. J. Master equation methods in gas phase chemical kinetics. *J. Phys. Chem. A* **110**, 10528–10544 (2006).
78. Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**, 2340–2361 (1977).
79. Byrne, G. D. & Hindmarsh, A. C. Stiff ODE solvers: a review of current and coming attractions. *J. Comput. Phys.* **70**, 1–62 (1987).
80. Klippenstein, S. J. From theoretical reaction dynamics to chemical modeling of combustion. *Proc. Combust. Inst.* **36**, 77–111 (2017).

81. Matera, S., Schneider, W. F., Heyden, A. & Savara, A. Progress in accurate chemical kinetic modeling, simulations, and parameter estimation for heterogeneous catalysis. *ACS Catal.* **9**, 6624–6647 (2019).
82. Thompson, A. P., Swiler, L. P., Trott, C. R., Foiles, S. M. & Tucker, G. J. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *J. Comput. Phys.* **285**, 316–330 (2015).
83. Drautz, R. Atomic cluster expansion for accurate and transferable interatomic potentials. *Phys. Rev. B* **99**, 014104 (2019).
84. Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403 (2010).
85. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
86. Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **8**, 3192–3203 (2017).
87. Zhang, L., Han, J., Wang, H., Car, R. & Weinan, E. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Phys. Rev. Lett.* **120**, 143001 (2018).
88. Schütt, K. et al. SchNet: a continuous-filter convolutional neural network for modeling quantum interactions. *Adv. Neural Inf. Process. Syst.* **30**, 992–1002 (2017).
89. Shui, Z. et al. Injecting domain knowledge from empirical interatomic potentials to neural networks for predicting material properties. Preprint at https://arxiv.org/abs/2210.08047 (2022).
90. Unke, O. T. et al. Machine learning force fields. *Chem. Rev.* **121**, 10142–10186 (2021).
91. Behler, J. Four generations of high-dimensional neural network potentials. *Chem. Rev.* **121**, 10037–10072 (2021).
92. Zubatyuk, R., Smith, J. S., Leszczynski, J. & Isayev, O. Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network. *Sci. Adv.* **5**, 6490 (2019).
93. Spicher, S. & Grimme, S. Robust atomistic modeling of materials, organometallic, and biochemical systems. *Angew. Chem. Int. Ed.* **59**, 15665–15673 (2020).
94. Xie, X., Persson, K. A. & Small, D. W. Incorporating electronic information into machine learning potential energy surfaces via approaching the ground-state electronic energy as a function of atom-based electronic populations. *J. Chem. Theory Comput.* **16**, 4256–4270 (2020).
95. Ko, T. W., Finkler, J. A., Goedecker, S. & Behler, J. A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer. *Nat. Commun.* **12**, 398 (2021).
96. Wen, M. & Tadmor, E. B. Hybrid neural network potential for multilayer graphene. *Phys. Rev. B* **100**, 195419 (2019).
97. Novikov, I., Grabowski, B., Körmann, F. & Shapeev, A. Magnetic moment tensor potentials for collinear spin-polarized materials reproduce different magnetic states of bcc Fe. *npj Comput. Mater.* **8**, 13 (2022).
98. Zeng, J., Cao, L., Xu, M., Zhu, T. & Zhang, J. Z. Complex reaction processes in combustion unraveled by neural network-based molecular dynamics simulation. *Nat. Commun.* **11**, 5713 (2020).
99. del Río, E. G., Mortensen, J. J. & Jacobsen, K. W. Local bayesian optimizer for atomic structures. *Phys. Rev. B* **100**, 104103 (2019).
100. Torres, J. A. G., Jennings, P. C., Hansen, M. H., Boes, J. R. & Bligaard, T. Low-scaling algorithm for nudged elastic band calculations using a surrogate machine learning model. *Phys. Rev. Lett.* **122**, 156001 (2019).
101. Williams, C. K. & Rasmussen, C. E. *Gaussian Processes for Machine Learning* (MIT Press, 2006).
102. Wen, M. & Tadmor, E. B. Uncertainty quantification in molecular simulations with dropout neural network potentials. *npj Comput. Mater.* **6**, 124 (2020).
103. Jackson, R., Zhang, W. & Pearson, J. TSNet: predicting transition state structures with tensor field networks and transfer learning. *Chem. Sci.* **12**, 10022–10040 (2021).
104. Spiekermann, K. A., Pattanaik, L. & Green, W. H. Fast predictions of reaction barrier heights: toward coupled-cluster accuracy. *J. Phys. Chem. A* **126**, 3976–3986 (2022).
105. Al Ibrahim, E. & Farooq, A. Transfer learning approach to multitarget temperature-dependent reaction rate prediction. *J. Phys. Chem. A* **126**, 4617–4629 (2022).
106. Wen, M., Blau, S. M., Xie, X., Dwaraknath, S. & Persson, K. Improving machine learning performance on small chemical reaction data with unsupervised contrastive pretraining. *Chem. Sci.* **13**, 1446–1458 (2022).
107. Coley, C. W., Green, W. H. & Jensen, K. F. Machine learning in computer-aided synthesis planning. *Acc. Chem. Res.* **51**, 1281–1289 (2018).
108. Badowski, T., Gajewska, E. P., Molga, K. & Grzybowski, B. A. Synergy between expert and machine-learning approaches allows for improved retrosynthetic planning. *Angew. Chem. Int. Ed.* **59**, 725–730 (2020).
109. Lan, T. & An, Q. Discovering catalytic reaction networks using deep reinforcement learning from first-principles. *J. Am. Chem. Soc.* **143**, 16804–16812 (2021).
110. Wang, Q., Mao, Z., Wang, B. & Guo, L. Knowledge graph embedding: a survey of approaches and applications. *IEEE Trans. Knowledge Data Eng.* **29**, 2724–2743 (2017).
111. Ji, S., Pan, S., Cambria, E., Marttinen, P. & Philip, S. Y. A survey on knowledge graphs: representation, acquisition, and applications. *IEEE Trans. Neural Netw. Learn. Syst.* **33**, 494–514 (2022).
112. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 6000–6010 (2017).
113. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
114. Trewartha, A. et al. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. *Patterns* **3**, 100488 (2022).
115. Kononova, O. et al. Text-mined dataset of inorganic materials synthesis recipes. *Sci. Data* **6**, 203 (2019).
116. Tshitoyan, V. et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
117. Mehr, S. H. M., Craven, M., Leonov, A. I., Keenan, G. & Cronin, L. A universal system for digitization and automatic execution of the chemical synthesis literature. *Science* **370**, 101–108 (2020).
118. Steiner, S. et al. Organic synthesis in a modular robotic system driven by a chemical programming language. *Science* **363**, aav2211 (2019).
119. Häse, F., Roch, L. M. & Aspuru-Guzik, A. Next-generation experimentation with self-driving laboratories. *Trends Chem.* **1**, 282–291 (2019).
120. Rohrbach, S. et al. Digitization and validation of a chemical synthesis literature database in the chempu. *Science* **377**, 172–180 (2022).
121. Raccuglia, P. et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).
122. Wen, M., Afshar, Y., Elliott, R. S. & Tadmor, E. B. KLIFF: a framework to develop physics-based and machine learning interatomic potentials. *Comput. Phys. Commun.* **272**, 108218 (2022).
123. St John, P. C., Guan, Y., Kim, Y., Kim, S. & Paton, R. S. Prediction of organic homolytic bond dissociation enthalpies at near chemical accuracy with sub-second computational cost. *Nat. Commun.* **11**, 2328 (2020).

124. Wen, M., Blau, S. M., Spotte-Smith, E. W. C., Dwaraknath, S. & Persson, K. A. BonDNet: a graph neural network for the prediction of bond dissociation energies for charged molecules. *Chem. Sci.* **12**, 1858–1868 (2020).
125. Grambow, C. A., Pattanaik, L. & Green, W. H. Deep learning of activation energies. *J. Phys. Chem. Lett.* **11**, 2992–2997 (2020).
126. Heinen, S., von Rudorff, G. F. & von Lilienfeld, O. A. Toward the design of chemical reactions: machine learning barriers of competing mechanisms in reactant space. *J. Chem. Phys.* **155**, 064105 (2021).
127. Houston, P. L., Nandi, A. & Bowman, J. M. A machine learning approach for prediction of rate constants. *J. Phys. Chem. Lett.* **10**, 5250–5258 (2019).
128. Jorner, K., Brinck, T., Norrby, P.-O. & Buttar, D. Machine learning meets mechanistic modelling for accurate prediction of experimental activation energies. *Chem. Sci.* **12**, 1163–1175 (2021).
129. Gastegger, M., Schütt, K. T. & Müller, K.-R. Machine learning of solvent effects on molecular spectra and reactions. *Chem. Sci.* **12**, 11473–11483 (2021).
130. Kim, S., Ji, W., Deng, S., Ma, Y. & Rackauckas, C. Stiff neural ordinary differential equations. *Chaos* **31**, 093122 (2021).
131. Raissi, M., Perdikaris, P. & Karniadakis, G. E. Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **378**, 686–707 (2019).
132. Ji, W., Qiu, W., Shi, Z., Pan, S. & Deng, S. Stiff-PINN: physics-informed neural network for stiff chemical kinetics. *J. Phys. Chem. A* **125**, 8098–8106 (2021).
133. Wang, S., Teng, Y. & Perdikaris, P. Understanding and mitigating gradient flow pathologies in physics-informed neural networks. *SIAM J. Sci. Comput.* **43**, 3055–3081 (2021).
134. Krishnapriyan, A., Gholami, A., Zhe, S., Kirby, R. & Mahoney, M. W. Characterizing possible failure modes in physics-informed neural networks. *Adv. Neural Inf. Process. Syst.* **34**, 26548–26560 (2021).
135. Krishnapriyan, A. S., Queiruga, A. F., Erichson, N. B. & Mahoney, M. W. Learning continuous models for continuous physics. Preprint at https://arxiv.org/abs/2202.08494 (2022).
136. Queiruga, A. F., Erichson, N. B., Taylor, D. & Mahoney, M.W. Continuous-in-depth neural networks. Preprint at https://arxiv.org/abs/2008.02389 (2020).
137. Amos, B., Jimenez, I., Sacks, J., Boots, B. & Kolter, J. Z. Differentiable MPC for end-to-end planning and control. *Adv. Neural Inf. Process. Syst.* **31**, 8299–8310 (2018).
138. Négiar, G., Mahoney, M. W. & Krishnapriyan, A. S. Learning differentiable solvers for systems with hard constraints. Preprint at https://arxiv.org/abs/2207.08675 (2022).
139. Lu, L., Jin, P., Pang, G., Zhang, Z. & Karniadakis, G. E. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nat. Mach. Intell.* **3**, 218–229 (2021).
140. Kovachki, N. et al. Neural operator: learning maps between function spaces. Preprint at https://arxiv.org/abs/2108.08481 (2022).
141. Gilpin, W. Chaos as an interpretable benchmark for forecasting and data-driven modelling. Preprint at https://arxiv.org/abs/2110.05266 (2021).
142. Snowden, T. J., van der Graaf, P. H. & Tindall, M. J. Methods of model reduction for large-scale biological systems: a survey of current methods and trends. *Bull. Math. Biol.* **79**, 1449–1486 (2017).
143. Yang, Q., Sing-Long, C. A. & Reed, E. J. Learning reduced kinetic monte carlo models of complex chemistry from molecular dynamics. *Chem. Sci.* **8**, 5781–5796 (2017).
144. Hoffmann, M., Fröhner, C. & Noé, F. Reactive SINDy: discovering governing reactions from concentration data. *J. Chem. Phys.* **150**, 025101 (2019).
145. Katsoulakis, M. A. & Vilanova, P. Data-driven, variational model reduction of high-dimensional reaction networks. *J. Comput. Phys.* **401**, 108997 (2020).
146. Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. Variational inference: a review for statisticians. *J. Am. Stat. Assoc.* **112**, 859–877 (2017).
147. Wang, Z. et al. A deep learning-based model reduction (DeePMR) method for simplifying chemical kinetics. Preprint at https://arxiv.org/abs/2201.02025 (2022).
148. Singh, P. & Hellander, A. Surrogate assisted model reduction for stochastic biochemical reaction networks. In *2017 Winter Simulation Conference* 1773–1783 (IEEE, 2017); https://doi.org/10.1109/WSC.2017.8247915
149. Chu, T.-C., Smith, M. C., Yang, J., Liu, M. & Green, W. H. Theoretical study on the HACA chemistry of naphthalenyl radicals and acetylene: the formation of $C_{12}H_8$, $C_{14}H_8$, and $C_{14}H_{10}$ species. *Int. J. Chem. Kinet.* **52**, 752–768 (2020).
150. Jafari, M. & M. Zimmerman, P. Uncovering reaction sequences on surfaces through graphical methods. *Phys. Chem. Chem. Phys.* **20**, 7721–7729 (2018).

Acknowledgements

This work is intellectually led by the Silicon Consortium Project directed by B. Cunningham under the Assistant Secretary for Energy Efficiency and Renewable Energy, Office of Vehicle Technologies of the US Department of Energy, contract number DE-AC02-05CH11231 (M.W. and E.W.C.S.-S.) with additional support from the start-up funds from the Presidential Frontier Faculty Program at the University of Houston (M.W.), the Joint Center for Energy Storage Research, an Energy Innovation Hub funded by the US Department of Energy, Office of Science, Basic Energy Sciences (E.W.C.S.-S.), the Laboratory Directed Research and Development Program of Lawrence Berkeley National Laboratory under US Department of Energy contract number DE-AC02-05CH11231 (S.M.B.), GENESIS: A Next Generation Synthesis Center, an Energy Frontier Research Center funded by the US Department of Energy, Office of Science, Basic Energy Sciences under award number DE-SC0019212 (M.J.M.), and the US Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Scientific Discovery through Advanced Computing (SciDAC) programme under contract number DE-AC02-05CH11231 (A.S.K.).

Author contributions

Conceptualization, investigation: M.W., E.W.C.S.-S., S.M.B., M.J.M. and A.S.K.; writing—original draft: M.W., E.W.C.S.-S., M.J.M. and A.S.K.; writing—review and editing: M.W., E.W.C.S.-S., S.M.B., M.J.M., A.S.K. and K.A.P.; visualization: M.W. and S.M.B.; project administration: M.W. and S.M.B.; funding acquisition: M.W., S.M.B., A.S.K. and K.A.P.; supervision: S.M.B. and K.A.P.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to Kristin A. Persson.

Peer review information *Nature Computational Science* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: Kaitlin McCardle, in collaboration with the *Nature Computational Science* team.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© Springer Nature America, Inc. 2023