Chemical Science

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: M. Ball, D. Horvath, T. Kogej, M. Kabeshov and A. Varnek, *Chem. Sci.*, 2025, DOI: 10.1039/D5SC03045E.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/chemical-science

View Article Online
DOI: 10.1039/D5SC03045E

Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Predicting Reaction Conditions: A Data-Driven Perspective†

Matt Ball[ab], Dragos Horvath[b], Thierry Kogej[a], Mikhail Kabeshov[a] and Alexandre Varnek[b*]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

The selection of optimal reaction conditions is a critical challenge in synthetic chemistry, influencing the efficiency, sustainability, and scalability of chemical processes. While machine learning (ML) has emerged as a promising tool for predicting reaction conditions in computer-aided synthesis planning (CASP), existing approaches face many significant challenges, including data quality, sparsity, choice of reaction representation and method evaluation. Recent studies have suggested that these models may fail to surpass literature-derived popularity baselines, underscoring these problems. In this work, we provide a critical review of state-of-the-art ML techniques, identifying innovations which have addressed the key challenges facing researchers when modelling conditions. To illustrate how relevant reaction representations can improve existing models, we perform a case study of heteroaromatic Suzuki-Miyaura reactions, derived from US patent data (USPTO). Using Condensed Graph of Reaction-based inputs, we demonstrate how this alternative representation can enhance the predictive power of a model beyond popularity baselines. Finally, we propose future directions for the field beyond improving data quality, suggesting potential options to mitigate data issues prevalent in existing literature data. This perspective aims to guide researchers in understanding and overcoming current limitations in computational reaction condition prediction.

1 Introduction

Selecting the 'optimal' reaction conditions for a given chemical transformation is a critical yet often time-consuming challenge faced by synthetic chemists daily. Despite its prevalence, tools capable of reliably and consistently predicting 'optimal' reaction conditions remain scarce and are limited by data quality[1], scalability[2] and generalisability[3]. The prediction of reaction conditions forms a critical component of Computer Aided Synthesis Planning (CASP), complementing tasks like forward synthesis prediction[4–6], retrosynthesis prediction[7–14], feasibility assessment[15] and reaction yield prediction[16–22]. With the ever-increasing amount of publicly-available data for chemical reaction modelling through projects such as the Open Reaction Database (ORD)[23,24], we might expect data-driven approaches to continue to improve. However notable challenges[1,25–29] face those building models to predict 'optimal' reaction conditions which must be considered.

We can start by defining what 'conditions' consist of. Conditions are the contents 'above the arrow' in a chemical reaction, defining the physicochemical environment under which a reaction occurs - see Fig. 1. This can consist of 'reagents': chemical species which take part in the reaction, but do not contribute a heavy atom to the product. Examples of 'reagents' include solvents, catalysts, ligands and bases in the case of a Suzuki coupling, but the scope of these 'reagents' will vary as a function of the reaction type being investigated. 'Conditions' are also comprised of physical parameters like temperature, pressure and time (and countless more), all of which influence the rate and feasibility of a reaction.

For modelling purposes, conditions can be encoded in the form of some vector, **c**. The definition of such a vector is a key challenge in reaction informatics: what is the best way to encode the ensemble of different species and parameters - reagents, temperature and pressure for example - in a single numeric vector? This vector requires a clear structure, containing elements associated with reagents and thermodynamic parameter values. At its most simple, this is a one-hot encoded vector, where the presence of a species is marked by the corresponding entry in the vector, and this is frequently used in condition prediction[30–32]. To make these labels more general, simple empirical categories can be used, like 'hydrophobic/polar/protic' for solvent, or '(Lewis) acid/base' for catalyst. Whilst the predictions of these targets may be less specific, they can help mitigate data sparsity which will be discussed in a later section. Moving towards the continuous space, descriptors might be calculated from the structure of the reagents[33–35]. Alternatively, agents may be characterised by their experimental properties, like dielectric constant or Kamlet-

[a] Molecular AI, Discovery Sciences RD, AstraZeneca, 431 83 Gothenburg, Sweden
[b] Laboratory of Cheminformatics, University of Strasbourg, 67081 Strasbourg, France
* varnek@unistra.fr

† Supplementary Information available: [details of any supplementary information available should be included here]. See DOI: 00.0000/00000000.

View Article Online
DOI: 10.1039/D5SC03045E

Taft values[36].

Therefore, for some reaction **r** under conditions **c**, reactivity modelling can be formulated as:

$$\hat{y} = f(\mathbf{r};\mathbf{c}) \tag{1}$$

The reaction outcome $\hat{y}$ can be - from the most accurate and less available, to the most empirical and more common - a reaction rate constant, a yield value or simply a binary classifier (feasible/infeasible). $\hat{y}$ can therefore be categorical (in the case of feasibility) or continuous. Any continuous prediction can be reformulated as a categorical one, by selecting a cutoff for 'acceptable' values for yield, rate etc. In general, feasibility models are the most popular, given that the presence of a reaction in a reaction database implies its feasibility, unless explicitly labelled as 'failed', which is unfortunately not customary[27,28]. Therefore, in the absence of negative data, feasibility models act as 'one-class classifiers' - for two-class classifiers, either experimental failures or assumedly infeasible 'decoy' examples must be provided[37].

This formulation is the generalisation of single-molecule quantitative structure-property relationship (QSPR) approaches to reactions. But there are additional challenges that must be considered in reaction informatics: the added complexity of reactions (compared to single molecules), resulting from the consideration of multiple reacting species and how they interact; in addition to the increased data pressures, like quality and sparsity, that the consideration of reaction conditions impose.

Like classical QSPR, Eq. 1 can be used to obtain 'optimal' condition predictions either directly or indirectly. By selecting different conditions, we can evaluate $f(\mathbf{r};\mathbf{c})$ to predict the reaction outcome of interest. Then, we can select the conditions which lead to the most desirable outcome. This is equivalent to selecting the set of conditions **c**, from the available set $\mathscr{C}$, that maximises the objective function $f(\mathbf{r};\mathbf{c})$ for a given reaction **r**. Formally this can be expressed as:

$$\mathbf{c_{opt}} = \underset{\mathbf{c}\in\mathscr{C}}{\operatorname{argmax}}\ f(\mathbf{r};\mathbf{c}) \tag{2}$$

where argmax denotes the value of **c** in $\mathscr{C}$ that yields the highest value of f. These 'reactant-specific' conditions may be best for a single pair of reactants, but may not be optimal for other, related reactants[38]. These conditions have applications in late-stage, scale-up chemistry, for example to optimise the production of a single molecule. Accordingly, vector **c** would, in this context, be rather detailed - predict the precise combinations of agents and thermodynamic parameters likely to work for the envisaged synthesis. Alternatively, 'general' conditions are also of interest, where a set of conditions may give strong outcomes for a range of related reactants, but are not highly specialised to a single reactant-pair[39,40]. Conditions of this type could be applied across a High-Throughput Experimentation (HTE) plate of a specific reaction type *R* using a variety of substrates in library synthesis or a 'robustness' screen. In this case, Eq. 2 changes to:

$$\mathbf{c_{opt}} = \underset{\mathbf{c}\in\mathscr{C}}{\operatorname{argmax}}\ \phi(f(\mathbf{r};\mathbf{c}),\mathscr{R}) \tag{3}$$

Where ϕ is some aggregation function, like the mean or a count of outcomes above a threshold[41]. These approaches are taken in pursuit of reaction optimisation via Bayesian optimisation[34,42–44] or Bandit optimisation[40,41], and often involves an iterative learning workflow to enhance the exploration of chemical space. The approach may work with both detailed and coarse condition vectors **c** - find specific setups which score consensually best of reaction type *R*, or predict the generic trend (e.g. "polar solvent, Lewis acid catalyst at room temperature").

This paper will mainly focus on the 'direct' prediction of conditions: what are the conditions required for the reaction to proceed to give the desired outcome (e.g. a maximal yield, a feasible reaction etc.).

$$\hat{\mathbf{c}} = g(\mathbf{r};\mathbf{y}) \tag{4}$$

Here, the condition vector should be as supported by the training data - the sparsity of which may force adoption of coarse-grained condition vectors. In Eq. 4, like any 'inverse QSPR' approach, condition prediction requires navigating many-to-many mappings between reactions and viable conditions[2]. This is the idea that a single reaction can occur under multiple different conditions, and inversely that a single set of conditions can be used for multiple reactions. Due to this, machine learning (ML)-based approaches to reaction condition prediction are diverse and highly dependent on the problem setup and dataset used. Raghavan *et al.* introduce the concept of 'global' models and 'local' models[45]. 'Global' models are trained on large amounts of literature data, often spanning a wide range of reaction types and aim to generalise across reaction space, like those in Refs. 30 and 46, 47. In contrast, 'local' models might focus on a single reaction type and a well-defined set of reactants and conditions[45]. Examples of 'local'-type models span from models for conditions of Michael additions[48] to C-N couplings[49,50]. This classification of these models as 'local' is arbitrary, as the applicability domain of these models varies in focus and comprehensiveness on a continuum. Ultimately, as the focus of a model shifts from a 'global' analysis of all reactions listed in a database to targeted modelling of specific reactions around selected reactants, the 'conditions' requiring consideration may also collapse to a subset of locally relevant options. Subsequently, the methods applied to predicting conditions require adaptation, paying attention to the constraints of the dataset of interest, and the scope of the conditions to be predicted.

This work examines the unique challenges of reaction condition prediction, particularly those related to data quality, model design (input and output) and evaluation. We then move on to review the state-of-the-art ML approaches, highlighting their progress and limitations. Finally, we present a case study of heteroaromatic Suzuki-Miyaura reactions from the USPTO dataset curated by Beker *et al.*[1]. In particular, our case study assesses the impact that reaction representation - how a reaction equation is encoded - has on the predictive power of condition prediction models. Here, we utilise Condensed Graph of Reaction (CGR) fragment representations[51] to explore if this reaction encoding can improve models' predictive power, beyond a strong popularity baseline. To conclude, we provide an outlook on the field,

View Article Online
DOI: 10.1039/D5SC03045E

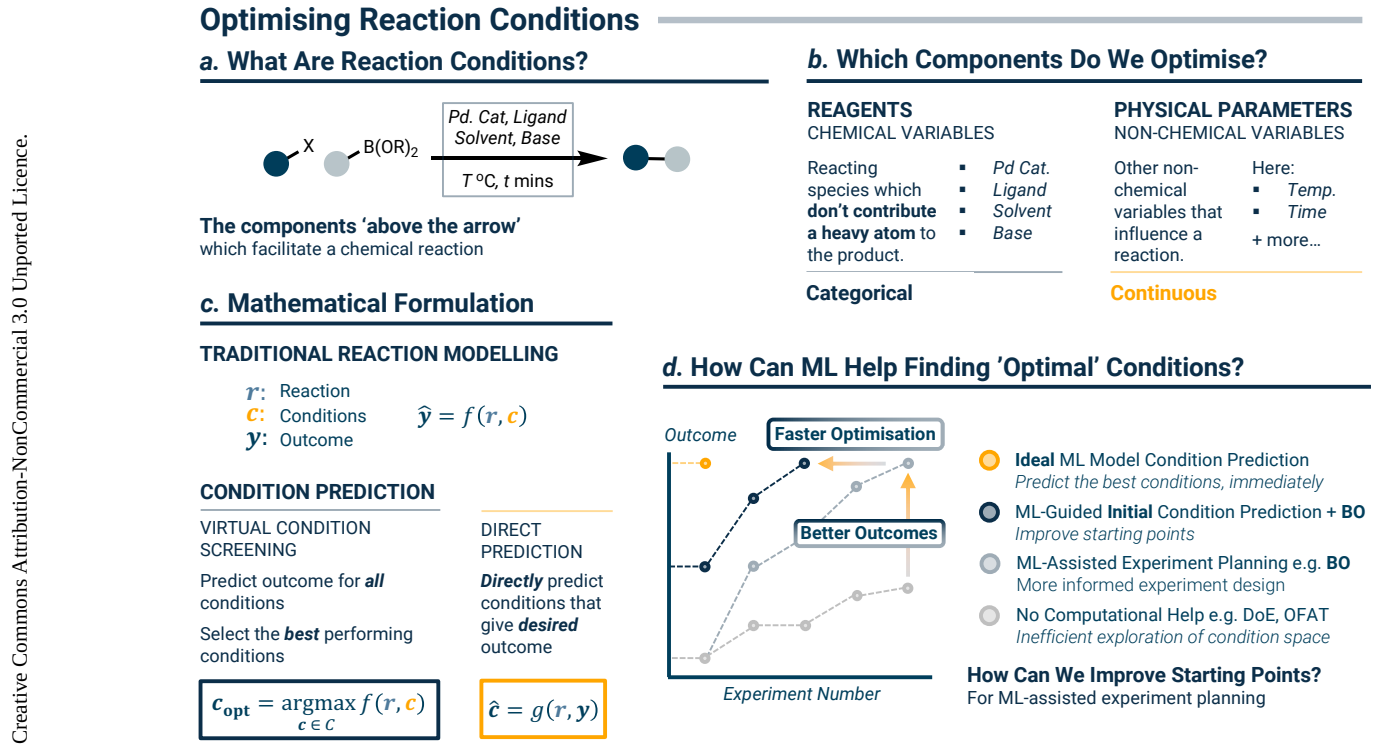

Fig. 1 Introducing reaction condition prediction. Because of the large possible scope of conditions, a decision must be made when creating models to limit the scope of 'conditions' considered.

identifying key directions for future research and development.

2 Data: Sources and Curation

From the formulation above, it can be seen that condition prediction is the 'inverse' problem of predicting reaction properties. Unfortunately, predicting some reaction properties (e.g reaction yield) is challenging[29], with many issues arising due to data. As discussed by Raghavan *et al.*, careful consideration must be paid to the selection of raw data source, curation protocols and underlying biases to produce datasets appropriate for reactivity modelling[45]. Due to the close relationship between yield and condition modelling, many of these problems are shared, as we discuss below. However, the many-to-many nature of condition prediction poses a unique obstacle for the construction and evaluation of condition prediction models. This section will briefly touch on challenges common across reaction/condition modelling, before moving onto problems specific to predicting conditions.

2.1 Data Sources

Since both reaction property prediction and condition prediction are modelling the same object: reactions, they naturally share the same data sources. There exist two main sources of reaction data: large-scale reaction databases (see Table 1) and smaller scale HTE/Electronic Lab Notebook (ELN) datasets (see Table 2).

As introduced by Raghavan *et al.* 'global', large-scale datasets typically cover a wide range of different reaction classes, with high substrate diversity but limited condition exploration for a given substrate[45]. A small collection of these can be found in Table 1. Models trained on these datasets are capable of suggesting conditions over a wide range of reaction types. However, Afonina *et al.* found that predictions of a 'global' model (Ref. 30) on a smaller, more focused dataset containing only hydrogenation reactions were not satisfactory, losing out to a simple popularity-based model. They hypothesised that the poorer performance is a result of the model not being biased towards a specific reaction type[2]. Even filtered versions of 'global' datasets to only include a single reaction type often lead to poor generalisability and applicability to industrial use cases, such as the screening of high-yielding conditions for new reactions[52]. In this case, it was suggested that these general datasets are too biased towards specific reagents for given reaction types to be useful for creating yield/condition prediction models that are useful for prospective applications.

Table 1 Examples of existing large-scale reaction datasets

Dataset	Open-Source?	# Reactions	Reference
USPTO (Curated from ORD)	Y	≈ 1.7 M	23,53,54
Pistachio	N	≈ 13.3 M	55
Reaxys	N	≈ 72 M	56,57
SciFinder	N	≈ 150 M	58

In contrast, the 'local', small-scale datasets cover a much smaller range of reaction classes, but with a smaller substrate diversity and higher condition exploration for each substrate. Models trained on this sort of data[21,38,59] can show more satisfactory results, and crucially better predictive propensity within their applicability domains[38] versus models trained on 'global' datasets.

View Article Online
DOI: 10.1039/D5SC03045E

The downside is that models trained on this data cannot be expected to generalise to other reaction types, due to the narrow scope of the training data. The other issue is data availability, as many smaller-scale datasets originate from proprietary ELNs within pharmaceutical companies.

Table 2 Examples of existing small-scale reaction datasets

Dataset	# Reactions	Reference
Suzuki HTE (2018)	5,760	60
Buchwald-Hartwig HTE (2018)	4,608	16
NiCOLit	2,003	18
Pd-catalysed C-H arylation	1,536	40
Amide Coupling	960	40

2.2 Data Curation

With a reaction dataset chosen, the next consideration is data quality. Raw reaction datasets are rarely ready for modelling. Errors in a chemical structure representation always complicate a modelling task and may cause technical problems[61]. However, despite the importance of reaction dataset curation, there have been few attempts to curate reaction datasets[62,63], particularly for reaction condition prediction[47,64]. The raw data sources mentioned in Table 1 have not been fully curated, and cannot be used immediately for modelling, requiring further curation before being used in a condition prediction tool.

Errors within chemical reaction data can arise in the form of missing reactants, reagents or products; mis-assigned reaction roles; incorrect SMILES representations and incorrect atom-mapping. There are a number of approaches for dealing with these issues, by either resolving the problems or removing the reaction from the dataset. As discussed by Gimadiev *et al.*, reactions should undergo 4 steps of curation before they can be used for reactivity modelling: chemical structures curation, transformation curation, reaction conditions curation and endpoints curation[63] (see Fig. 2).

The exact details of the chemical structures curation are usually a subset of steps suggested by Fourches *et al.*: detection of valence violations, ring aromatisation, normalisation of specific chemotypes, standardisation of tautomeric forms and the splitting of ions, among others[65]. 'Transformation curation' aims to resolve issues with unbalanced reactions, atom-to-atom mapping, reaction role assignment and duplicate detection. For unbalanced reactions, dealing with missing reagents, reactants and products can be done using ML tools by suggesting replacements for these missing species[46,66,67], and this improved data quality was shown to improve model performance in product prediction[46]. Alternatively, rule-based tools can be used to fill missing small molecules and balance reactions[68,69]. The same consideration needs to be paid to the representation of reaction conditions, where text-based entries for reaction conditions must be collected and mapped to the appropriate SMILES string[64].

When considering reaction condition prediction specifically, role-assignment of reagents is incredibly important but this is not trivial. Many existing 'global' approaches divide reagents into roles such as catalysts, solvents and agents (which encapsulate additives, acids, bases etc.)[30–32,47]. A single reagent can perform multiple different functions across different reaction types (or even within a single reaction), leading to challenges when assigning a reagent to a particular class. This is particularly pronounced when considering a wide range of reaction types, as is the case in 'global' models. For such models, it is often the case that a reagent role simply cannot be assigned beyond 'Agent', 'Solvent' or 'Catalyst'[30,47,70]. Therefore, there are a larger number of classes within this reagent type and subsequently, a more challenging classification problem. Another aspect of conditions curation is understanding which reagents take part in the reaction, and which 'reagents' are part of other procedural processes, for example workups or purification. More high-fidelity labelling of reaction roles could lead to higher quality datasets for condition modelling, as provided by modern databases such as ORD[23]. Furthermore, trusting the labelling of reaction roles from large datasets such as USPTO can lead to issues. Frequently reaction components are mislabelled[64], leading to ambiguity in what is a reactant versus a reagent. To rectify this, atom-mapped reaction equations can be used to determine what are reactants, by identifying which species contribute 'heavy atoms' to the product. Once reactions are in a standardised format and the roles of all components have been assigned, duplicate reactions need to be dropped. Duplicate entries are common, due to scientists adopting transformations reported elsewhere in the literature. Additional treatment of rare conditions may also be required, as Wigh *et al.* report that the removal of these entries can improve performance of condition prediction models[64]. It is crucial to adopt standardised curation protocols to not only benefit reactivity prediction tasks but enable fair comparisons of model performance.

3 Data: Sparsity and Bias

Even after curation, existing large-scale datasets face many issues that require consideration. Like the previous sections, there exist many common challenges to be addressed across reactivity modelling, the first of which is dataset bias.

3.1 Dataset Bias

Strieth-Kalthoff *et al.* explain the three key types of bias frequently found within chemical reaction datasets: experimental noise, selection bias and reporting bias[27] (see Fig. 3).

Briefly, experimental noise refers to noise caused by human or experimental error, for example, errors in experimental protocol, which caused the loss of product. This results in large variance in recorded yield for reactions performed under the same conditions as Voinarovska *et al.* show[29]. Depending on how yield information is used in the modelling process, the extent that this affects condition prediction models varies. For models which don't use yield information at all (and assume all non-zero-yield reactions are successful), these problems shouldn't affect model performance. Conversely, where yield information directly influences model training, high variance in yield could hypothetically lead to the incorrect 'optimal' conditions being identified.

Selection bias refers to the tendency of chemists to select established conditions (or the reagents that are simply available in

View Article Online
DOI: 10.1039/D5SC03045E

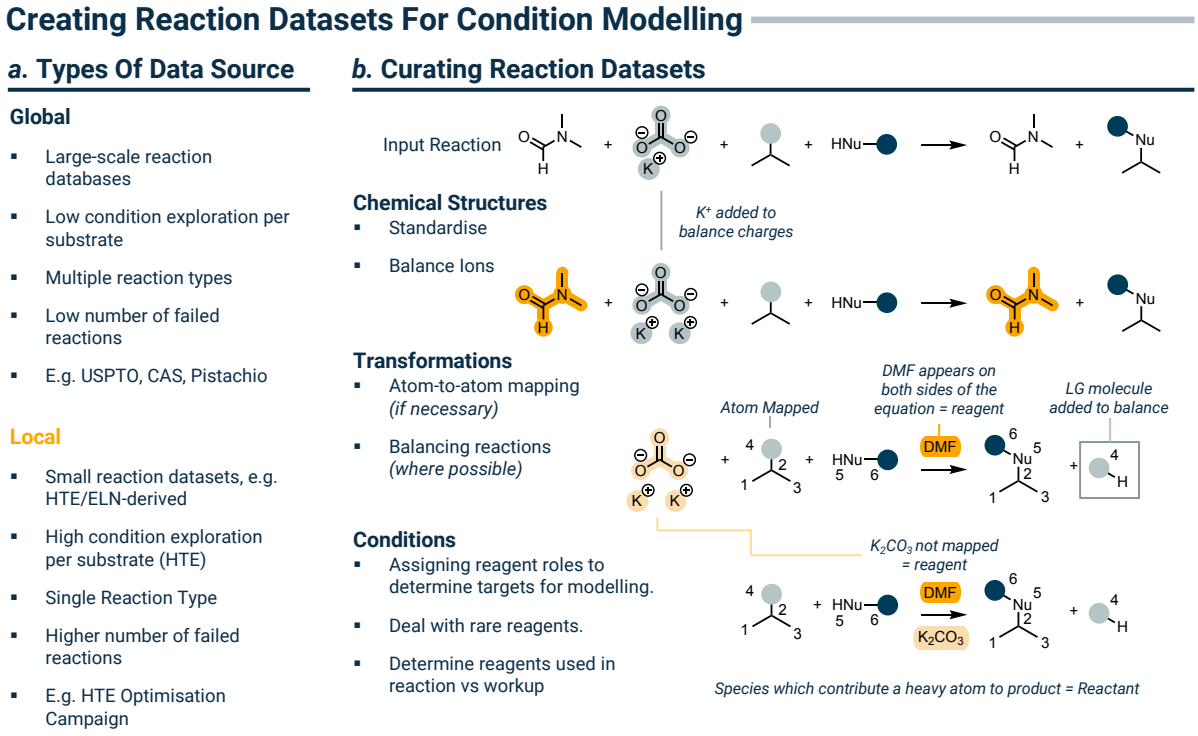

Fig. 2 Summarising key features of reaction data sources and the subsequent steps required to curate these sources.

the lab) when performing reactions, ultimately leading to large imbalances in the dataset where few conditions are explored and can lead to models that are trained on this data to learn little more than popularity trends[1].

The final type of bias discussed by Strieth-Kalthoff *et al.* concerns the reporting of results, and particularly the bias of high yielding 'successful' results. This issue is further exacerbated by the common practice of reporting only the optimal outcome from a series of identical experiments, often without accompanying error estimates, which further complicates modelling. As a result, there are large imbalances in the distributions of yields across a data source, which prevents models from learning which reactions don't work and ultimately reduces performance[27]. Maloney *et al.* called for an improvement in the reporting of experimental yields, and an increase in the amount of these 'low yielding' reactions being reported, thereby making these reactions more common in chemical reaction databases[28].

Despite these biases, there are approaches which can counteract this (although, they have their own issues which need to be considered). For example, it has been demonstrated that the introduction of synthetic 'negative' data (labelled, impossible reactions) in appropriate quantities can lead to improved performance in yield prediction[27] or retrosynthesis applications[37]. Alternative approaches include the sampling of 'hard negative' conditions. These 'hard negatives' (incorrect reagent or solvent predictions assigned a high probability by the model) were combined with true labels to generate diverse training examples to help the model distinguish between correct and incorrect conditions[32]. Schwaller *et al.* artificially expanded existing data via data augmentation using permuted and randomised reaction SMILES strings, resulting in an improvement in R^2 of up to 0.15 for a yield prediction model[71]. Various forms of data augmentation have also been applied in the prediction of retrosynthesis[7,72] or reaction products[73]. Other options to leverage existing chemical knowledge include transfer learning and this has shown promise in modelling reactivity[3,74], although in certain cases this 'transfer' of information can hinder the models' predictive capabilities via 'negative' transfer[75]. This emphasises that, although these strategies can aid the situation, care must be taken to ensure that the additional data is not causing a decrease in performance, and that the introduction of the new data is not bringing significant biases with it.

3.2 Data Sparsity

These biases both cause and propagate data sparsity - the scarcity of explored conditions per reaction relative to the vast combinatorial condition space (see Fig. 3). This sparsity hinders the development of reliable, robust models[27,29], especially in the case of 'global' datasets, where this problem is most acute with their broad scope of reaction types, and subsequently their reaction conditions. For these models it is difficult to learn the links between 'reactant reactivity' and 'reagent reactivity' (i.e. how harsh or mild the conditions are) due to the limited data available for each reactant pair. Beyond prediction, data sparsity also affects the way that condition prediction models can be evaluated, which we discuss below.

While data quality, bias, and sparsity are critical challenges in any reactivity modelling, they manifest themselves differently in

View Article Online
DOI: 10.1039/D5SC03045E

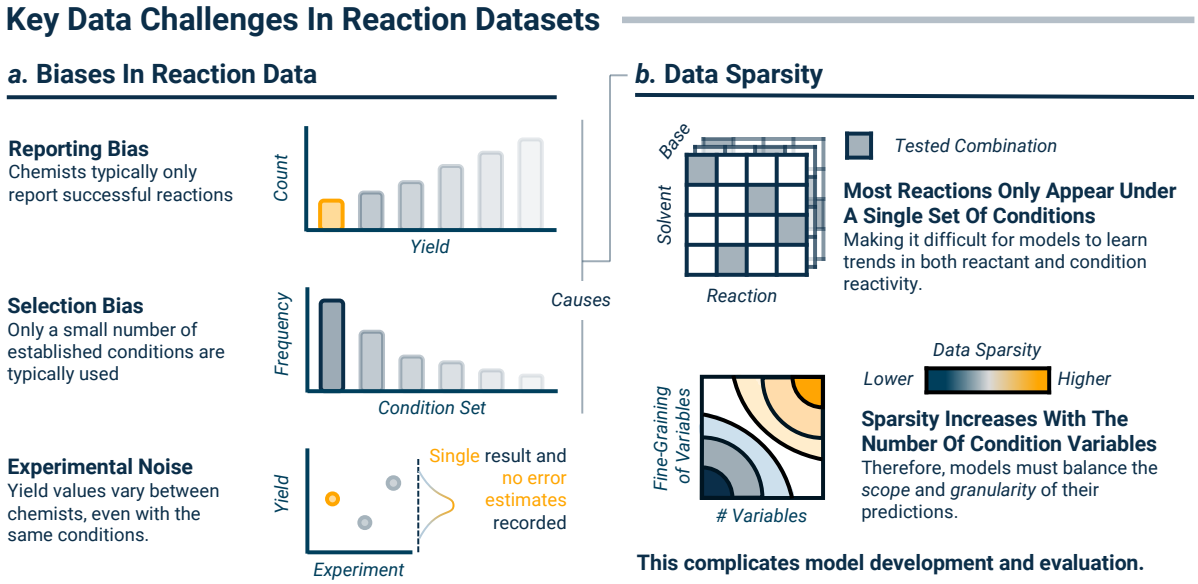

Fig. 3 Key data challenges faced when using reaction datasets. The biases in literature data sources directly result in the sparsity of reaction-condition pairs that are observed. These biases often interact, for example, a single transformation may result in multiple different outcomes for the same set of conditions (experimental noise), where only the 'best' outcome is then recorded in a database (reporting bias). This sparsity impacts the condition scope that models can reliably predict, since models may not 'see' an appropriate amount of condition space in training.

condition modelling due to the many-to-many relationship between reactions and potential conditions, which we will explore now.

4 The Many-to-Many Nature of Condition Prediction

Arguably the most important consideration in condition modelling is that a single reaction can succeed under multiple valid conditions. This many-to-many relationship, combined with the dataset biases and sparsity discussed above, make the modelling of reaction conditions particularly challenging. This challenge presents itself in two forms: model design (specifically the selection of input and output) and model evaluation.

4.1 Model Input: Representing Reaction Equations

For condition prediction models, we would expect the input to be a reaction equation, and the output to be a set, or list of, viable conditions. The choice of encoding of both input and output will have a profound effect on the performance of such a model. For 'global' models, reaction descriptors should balance computational overhead (and storage considerations), with relevance to the problem at hand. For example, reactions have been represented in binary fingerprints, like Morgan fingerprints of the reactants and products[1] (or their difference[30]). Alternatively, reactions can be encoded by SMARTS strings and modelled with Natural Language Processing (NLP) methods[46,47]. Finally, graph-based representations of the reactants and products[31,76] or the CGR[2,77] can be used. For smaller-scale ('local') models more computationally intensive descriptors can be calculated, due to the smaller dataset size. Examples of such representations for reactions here might include Density Functional Theory (DFT) properties of the reactive atoms[59,78], or custom 3D descriptors which can account for complex steric effects[38].

In our case study (see later), we use the CGR, aiming to strike a balance between an informative representation and computational cost. The CGR encodes a reaction as a single pseudo-molecule, arising from the superposition of the reactant and product graphs of molecules in a reaction[51] (see Fig. S2). This pseudo-molecule contains 'dynamic' bonds, representing the bonds that are broken or made during the reaction. However, this requires atom mapping, which is not trivial, and even state-of-the-art computational tools[79,80] cannot achieve perfect accuracy[81]. The requirement for atom mapping aside, CGRs have emerged as a powerful representation for chemical reactions and have shown strong performance when used as input for the prediction of reaction properties such as activation energies, rate constants and protecting group reactivity[82–84].

4.2 Model Output: Encoding Conditions

The direct prediction of reaction conditions using ML methods is tributary to the level of coarseness chosen (or enforced by data sparsity) for the condition vector to predict, and to the level of 'globality' of the approach - see Fig 4. Ideally, a model would be able to predict the exact combination, stoichiometry, and order of addition for all reagents in a reaction (and physical parameters like temperature, pressure, flow rate etc.). Due to the aforementioned challenges with existing literature data, increasing the number of variables to model exacerbates the data sparsity problem. Consequently, the choice of variables to predict forms a critical and difficult aspect of condition modelling. Variables must balance practical utility for synthetic chemists with the constraints imposed by data availability.

View Article Online
DOI: 10.1039/D5SC03045E

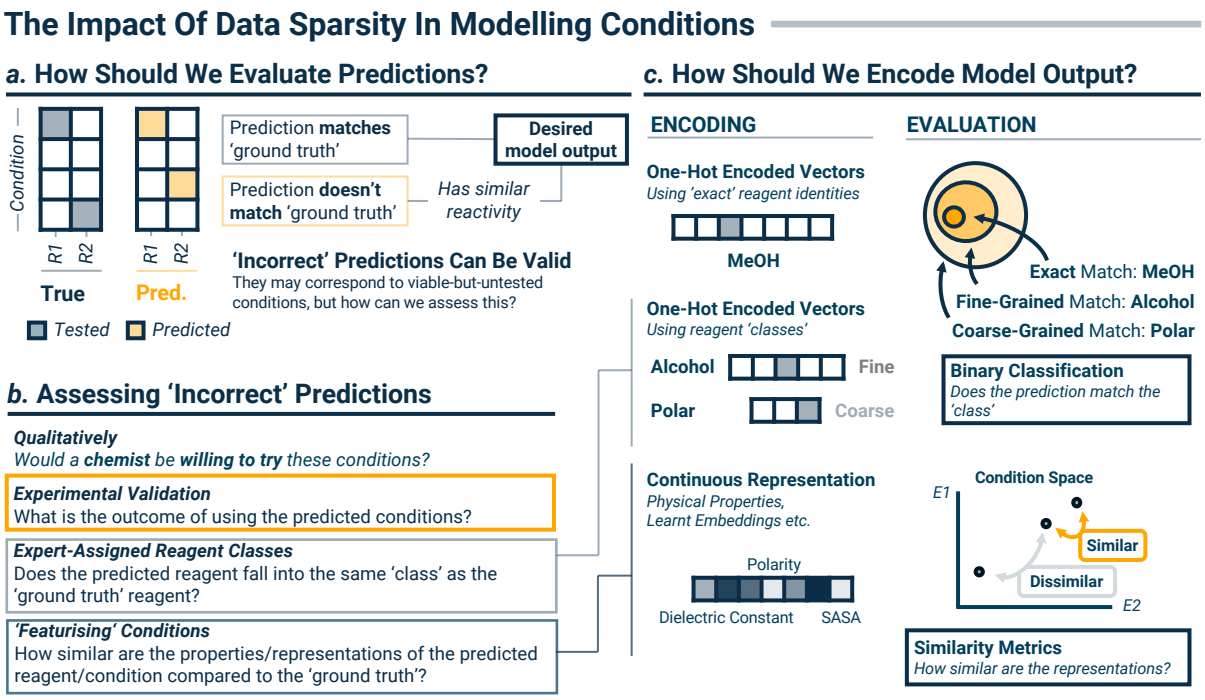

Fig. 4 The impact of data sparsity on the prediction of conditions. Because many reactions only have a single set of conditions associated with the record, analysis of these predictions are often limited to binary correct/incorrect evaluations, where more nuanced analysis of the incorrect predictions may be required. Consideration of conditions beyond simple one-hot encoding of identity could present a way of quantifying the similarity between an 'incorrect' prediction and the 'ground truth'. In any case, the best way of assessing 'incorrect' predictions is to test them experimentally.

Most often, a one-hot encoded **c** vector is targeted, where the presence of a given reagent is indicated by a binary label to be predicted by the approach. Continuous variables, like temperature or pressure, are often treated in the same way by 'binning' the variable into discrete categories[2] (or can be modelled as a regression task[30,47]).

When modelling with 'local' datasets, where data sparsity may be less pronounced, modelling variables at a higher fidelity may be possible. As a further benefit of this scenario, some condition factors may be sine-qua-non prerequisites for the given class of reactions, and therefore already known - hence no longer explicitly included in the output vector **c**. In contrast, when modelling with 'global' datasets, where the prerequisites for the conditions may vary across reaction types, this is not possible.

4.3 Model Evaluation

Unlike reaction property prediction, where the magnitude of a prediction's error can be gauged quantitatively using root mean squared error (RMSE) and related metrics, condition prediction necessitates careful consideration of classification-based metrics and the inherent ambiguity concerning 'optimal' conditions.

'Global' models are often evaluated using classification metrics like top-k accuracy[30,32,47,85]. However, the 'ground truth' in literature-derived datasets is inherently ambiguous: multiple valid conditions may exist for a reaction, but only a subset are documented. For example, a model predicting methanol instead of the 'ground truth' ethanol for a polar protic solvent is penalised equivalently to one predicting toluene, even though methanol is chemically plausible but untested. Conventional metrics fail to distinguish between chemically invalid predictions and valid-but-unexplored alternatives.

For 'local' models, multiple condition sets may be successfully applied to a single reaction. In such cases, ranking-based evaluation metrics such as mean reciprocal rank or the Kendall tau coefficient can be used to assess performance[78], with 'true' rankings based on the outcomes of each condition set. Similarly, when yield prediction is being used to 'screen' conditions, one could also use the Spearman correlation coefficient or average yield percentile ranking[35], which emphasise relative performance of conditions over absolute error. Of course, these approaches are less applicable for global models, where the ranking of all possible conditions is unfeasible, and a given reaction may only have a single condition label associated with it.

The 'gold-standard' for the evaluation of models would include the testing of predictions in the lab, alongside top-k accuracy, as done by Schilter *et al.*[86]. This is particularly important in the case of a model's prediction disagreeing with the 'ground truth'. However, access to experimental validation is not always possible (and is resource-intensive), but other *in silico* metrics could also be used. As an example, Wang *et al.* used the Solvent Similarity Index[87] to determine how similar the predictions of 'incorrect' solvents were to the ground truth[47]. Of course, no *in silico* metric of similarity can replace experimental validation, but it can help provide further information into the 'chemical reasoning' of a model.

Another alternative is to use condition clustering, where

View Article Online
DOI: 10.1039/D5SC03045E

reagents with similar chemical properties are categorised in the same cluster. The intuition behind this follows directly from above: in general, we might expect reagents with very similar chemical properties to react in the same way. This approach could be applied post-prediction[77], aiming to evaluate model performance whilst accounting for data sparsity and the many unlabelled positive examples in reaction condition datasets. On the other hand, such an approach could be applied in data preprocessing, reducing the number of classes that a model might need to predict, and subsequently improving performance[1,2]. We explore this concept further in our case study, see 7. This is comparable to the concept of 'binning' in yield prediction, where the underlying variance in yield data makes modelling of exact yields difficult[29]; but effective, useful tools can still be developed by considering yield as a discrete class including 'zero yield', 'low yield' or 'high yield' classes.

The final consideration to be made, like in any reactivity modelling, is testing that a model has learnt meaningful chemistry rather than exploiting underlying patterns in the data[88]. For example, using adversarial tests for unrelated representations of reactants (e.g. random or one-hot encoding) to illustrate the improvement that applying such a model can have on the problem of predicting appropriate conditions[78].

We have seen how condition prediction presents a unique challenge due to its inherently many-to-many nature. This complexity, combined with dataset sparsity and bias, impacts every stage of model development: from input representation and output encoding to evaluation. The choice of both input representation and output encoding is closely tied to the nature of the dataset and should be carefully considered, particularly for 'global' models. Furthermore, standard evaluation metrics in 'global' models often fall short, due to the ambiguity of 'ground truth' labels. Therefore, it is critical that the evaluation of such models should include experimental validation (in the ideal case), or at the very least careful analysis of 'incorrect' predictions, to gain a better insight into a model's performance.

5 State-of-the-Art Approaches to Predicting Conditions

These challenges set the stage for exploring state-of-the-art approaches that aim to overcome these limitations with more robust architectures and chemically aware strategies, as we will look at now. The following sections explore these developments, emphasising how models can still use imperfect datasets to deliver actionable predictions.

To begin, we refer back to the introduction, and the different definitions of 'optimal' conditions. The majority of existing approaches focus on selecting conditions for a given substrate pair which produce the highest yield of the desired product[38,43,59]. Though some methods focus on discovering and predicting 'general' reaction conditions[39,40,89]. We will predominantly focus on models of the former, analysing models based on their architecture. For models that aim to optimise the yield, we can classify these models in the same way we did with the data: 'global' models and 'local' models[45]. 'Global' models refer to models that are trained on large amounts of literature data contained within datasets such as US Patent & Trademark Office (USPTO)[53], Reaxys[56], Pistachio[90] or the Open Reaction Database[23], and can be applied to many different reaction types. On the contrary, 'local' models are trained on a single, specific reaction type (often) using HTE data.

Furthermore, it is important to distinguish possible problem setups employed to predict conditions (see Fig. 5). Most 'global' models aim to solve some form of 'classification' task: which reagent(s) from a selection of reagents are the most appropriate for the input reaction? With this in mind, we begin our analysis by analysing 'global' models.

5.1 'Global' Condition Prediction Models

5.1.1 Similarity Approaches

When synthetic chemists think about conditions for reactions with new substrates, conditions from similar reactions in the literature often function as a starting point. Some tools leverage the same idea by performing a similarity search for reactions and returning the conditions for the most similar reactions[84,91]. For example, the work of Refs. 84 and 91. The encoding of reactions differs in the two approaches, with Lin *et al.* using CGRs to predict conditions with respect to protecting group reactivity, with balanced accuracy of predictions varying between 85-95 %[84]. Conversely, Dobbelaere *et al.* demonstrate how a bond-electron matrix approach can be used to select initial conditions for a Heck reaction[91]. As another example, Walker *et al.* found that k-Nearest Neighbour (kNN) models were the best approach for predicting solvents for named organic reactions, beating support vector machine (SVM) and neural network methods. The kNN model achieved a top-1 accuracy of 69-80 % and an impressive top-3 accuracy of 91-99 % depending on the reaction type[92]. They also found that when the solvents were grouped using the similarity metrics employed by the kNN model, solvents formed distinct clusters corresponding to certain physical properties. These approaches provide inherent interpretability for predictions, where chemists can intuitively understand that conditions are being selected based on their suitability for similar reactions. A comparable approach was applied by Afonina *et al.* as a comparison model for hydrogenation reactions using a kNN model which showed good performance (improvement of 10 % over a popularity baseline for top-10 accuracy)[2], even outperforming other popular ML approaches (improvement of 3 % over the same baseline)[30].

However, the similarity approach does have issues, as similarity searching on large databases can become very slow, and requires special approaches like FAISS[93,94]. This can make similarity searching impractical, despite its interpretable nature. The other issue is that similar reactant structures can often exhibit very different reactivity, for example changing the substitution pattern of aromatic systems like indoles can cause vastly different reactions to occur (electrophiles reacting at C(3) versus reacting at nitrogen). Structural encodings must be able to capture this subtle change in reactivity, which might not be possible through simple fingerprints, and more complex DFT-based featurisation methods might be better suited to capturing these differences.

View Article Online
DOI: 10.1039/D5SC03045E

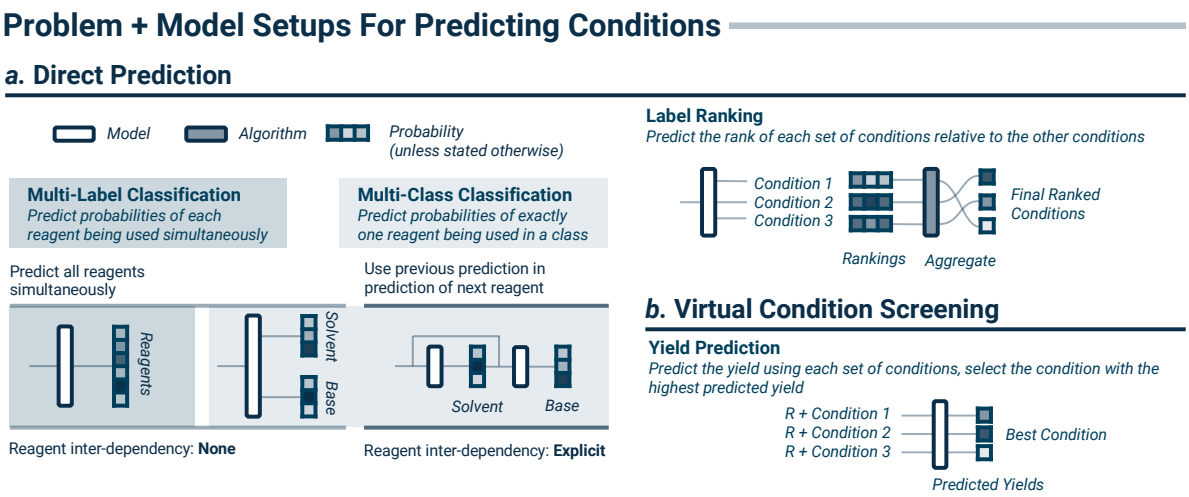

Fig. 5 Problem setups used when predicting conditions. For all 'direct prediction' models, the input is simply the reaction equation. Generally, 'global' models will adopt the task of predicting conditions as a multi-label classification task, or a series of multi-class classification tasks, aiming to predict the individual components of the conditions. For 'local' datasets, where the space of conditions can be enumerated, alternative approaches like label ranking and yield prediction can, and have, been used to predict highly performant sets of conditions.

Increasing the complexity of models can capture more of this reactivity information, which can lead to better performance which we will discuss now.

5.1.2 Feed Forward Neural Networks

Feed-forward neural networks have also been applied to this problem, aiming to capture more complex relationships between the reaction conditions and the reactants. The most notable example of this was Gao *et al.*, who used an RNN-like architecture, trained on over 10M reactions in the Reaxys database to predict reaction conditions consisting of two solvents, two reagents, a catalyst and temperature. The prediction was done sequentially, with the catalyst being predicted first, followed by two solvents, two reagents and finally the temperature. Here, reactions were encoded using simple Morgan fingerprints[95], illustrating how even 'simple' reaction representations can furnish strong results. This work also implements a 'close match' evaluation of predictions, using solvent similarity and feature-based Morgan fingerprints to determine matches. Using this metric they achieve an impressive 70 % top-10 accuracy across all reactions. The same reaction condition prediction architecture was also employed by Qian *et al.*, but alongside using product and reaction fingerprints, the input data was augmented with procedural information and demonstrated state-of-the-art performance, improving on SMILES-only methods by 17.2 % for top-1 accuracy[70]. This indicates that procedural information for similar reactions can be very useful in determining potential conditions for a new reaction of interest. Crucially, by allowing the model to use its previous reagent predictions, these approaches have the ability to capture the dependence *between* predicted reagents. Such a trait is important when thinking about reaction conditions, as it is important that reagents are compatible (e.g. all reagents are in the correct phase at the given reaction temperature). This idea is also exploited in the tree-based models used by Maser *et al.*, who highlight a drop-off in model performance when this information from the previous prediction is withheld[31].

Both Afonina *et al.* and Chen and Li also treat condition prediction as a classification problem, though with some similarities to 'label ranking' (see Fig. 5)[2,32]. Afonina *et al.* use a 'Likelihood ranking model' which enumerates all conditions including acid, base, temperature, pressure and catalyst, encoding reactions using ISIDA CGR fragment descriptors[51], before using a neural network to output the most likely conditions for that reaction. This approach showed strong performance, improving on the work of Gao *et al.* for hydrogenation reactions (73 % top-1 accuracy on a retrospective test set), although performance on the prospective test set showed that a popularity baseline was comparable in performance, achieving correct top-1 predictions 68 % of the time. This method requires the enumeration of all conditions, and for large datasets covering many reaction types, enumerating all combinations of conditions is computationally infeasible[2]. Chen and Li employed a neural network that shares many characteristics of the 'likelihood ranking model'. Using a two-stage condition generation and ranking approach, they leveraged a ranking model alongside a generation model to generate plausible conditions prior to ranking, avoiding the need for the enumeration of all possible conditions[32]. Again, this yielded good results, finding an exact match to the true condition with the top-1 suggestion 53 % of the time. Interestingly, in a short case study, the authors found that the model suggested conditions which were used in the publication but were not recorded in the reaction database. This reiterates the importance of not only recording all reactions performed in reaction databases, but also that care should be taken when evaluating models purely based on top-k accuracy.

The key to all feed forward neural network approaches is the choice of reaction descriptors. Whilst the fingerprints employed by Chen and Li and Gao *et al.* are computationally inexpensive to calculate, they may not capture the more complex electronic and steric effects that can explain reactivity patterns. With the devel-

View Article Online
DOI: 10.1039/D5SC03045E

opment of methods to estimate complex descriptors and features in computationally inexpensive ways[96,97], future models may be able to take advantage of this. Alternatively, researchers can look towards more complex architectures, such as graph neural networks and transformers to generate more information-rich encodings for the reactions in order to improve performance, which we will see in the next section.

5.1.3 Graph-Based Neural Networks

The structure of molecules naturally lends themselves to the graph representation, with nodes encoding atomic information and edges encoding bond information. Graph neural networks (GNNs) have been applied to many molecular tasks, including property prediction[98–101], synthesis planning[12,102] and generative molecular design[103–105]. Generally, these approaches use message passing neural networks (MPNNs)[106], or graph convolutional networks[107] to convert the molecular graph into a vector representation which can be used in downstream tasks. However, when thinking about reactions, the situation is more complicated. Reactions are composed of multiple disconnected graphs corresponding to the reactants, reagents and products. One method of dealing with this is using the CGR approach (explained in more detail in the case study)[51]. Once the reaction is encoded within the CGR pseudo-molecule, GNN methods can be applied to it, and the D-MPNN created by Heid and Green showed state-of-the-art performance on reaction property prediction tasks[83].

Applied to condition recommendation, the most notable examples of GNN application are Maser *et al.*, Kwon *et al.* and Wang *et al.*[31,77,85]. Maser *et al.* used 'attended relational' graph convolutional networks (AR-GCNs) to predict conditions for a collection of different coupling reactions, including Suzuki, Negishi and C-N couplings. The models showed good predictive performance over a popularity baseline (31-42 % improvement for top-1 predictions). In addition, this model has an accompanying analytical framework, providing interpretability analysis on the learned feature weights to understand the reasoning behind different predictions. However, the performance of this model was marginally worse (2 % for top-1 accuracy) compared to tree-based methods also used in the publication on the smaller Pauson-Khand dataset. The authors suggested that the smaller dataset size makes the GCN more prone to overfitting, which made tree-based modelling more suitable here[31].

Extending this approach, Kwon *et al.* used GNNs to encode both reactants and products, combining this with a variational autoencoder (VAE)[108] to predict conditions[85]. In comparison to both Gao *et al.* and Maser *et al.* this approach resulted in a higher accuracy when allowing multiple predictions from the VAE. However, this approach is more time-consuming versus the others, and no comparison was performed where the models from Refs. 30 and 31 could predict multiple conditions.

Finally, Wang *et al.* use a combination of templates and condition-clustering alongside a D-MPNN acting on CGRs. This work exemplifies one of the first uses of condition clustering to improve performance, by increasing the diversity of predictions, and acknowledging the many-to-many nature of condition prediction[77]. By incorporating this clustering, the top-1 accuracy of their method jumps from 45 % to 66 %, a significant increase. Zhang *et al.* take a slightly different approach; they encode their reactant and product as graphs before passing through their GNN pretrained on atom level and bond level tasks. These molecular level descriptors from the GNN are passed to a second NN along with a one-hot encoded reaction template, and this is used to predict the most likely solvents and catalysts for a reaction. However, in the prediction of the solvent and catalyst, the identity of the other reaction component is not considered[76]. Nonetheless, these models could predict the correct catalyst and solvent 59 % and 42 % of the time respectively. Ignoring the inter-dependence of the conditions is likely to lead to some drop in accuracy, because the identity of one reagent, along with the reaction will determine the identity of the other reagents. Modelling this dependence is a key part of reaction condition prediction.

GNNs clearly show promising performance in predicting appropriate conditions, indicating the representation that these models learn is comparable to (and sometimes better than), more simple fingerprint descriptors. Moving beyond graphs, reactions can also be described by their SMILES string, to which natural language processing (NLP) methods can be applied, and the final architecture we will look at are the transformer-based models.

5.1.4 Transformer Models

The transformer architecture[109] has shown application across many life sciences, covering areas like protein structure prediction[110,111], protein design[112] and in chemistry specifically, transformers have demonstrated utility in synthesis planning[4,9,13,17,113] among others. Extending this to condition prediction, Wang *et al.* created a condition prediction transformer, Parrot, which demonstrated very good performance, showing higher accuracies in a direct comparison to other condition prediction models (Refs. 30, 76 and 31) achieving a top-1 accuracy of 27 % for exact matches[47], on a subset of the USPTO and Reaxys datasets which the authors curated. This architecture uses a BERT-like[114] encoder to generate a reaction feature vector from a SMILES string, and a transformer decoder to predict the conditions sequentially, in the same order as the model proposed by Gao *et al.*. This was pretrained on 2 separate tasks, using masked language modelling (MLM) and masked reaction centre modelling (RCM) to guide the model to understand reaction centres. Furthermore, the transformers use of the attention mechanism allows for investigation of the attention weights to improve interpretability of predictions, another desirable feature of any model.

Another similar approach was taken by Andronov *et al.*, who repurposed the MolecularTransformer described by Schwaller *et al.* for reagent prediction[5,46]. The final example to leverage the transformer architecture is MM-RCR by Zhang *et al.*[115]. This uses a combination of the previous architectures, using a multimodal reaction input consisting of SMILES, Graphs and Text, on top of a large language model (LLM) to predict conditions. This achieves state-of-the-art performance on the same dataset curated by Wang *et al.*. Their ablation study demonstrates the benefits of a multi-modal representation, showing significant (up to 17 %) improvement over the same model using a single data modality.

View Article Online
DOI: 10.1039/D5SC03045E

To conclude this section, whilst 'global' condition prediction models are highly desirable (and many such models perform to a strong level), the level of detail that can be afforded without making the dataset too sparse means that finer grained details of a reaction such as timing, pH and others are often ignored, despite their importance to synthesis planning. Furthermore, the lack of consistent benchmarking datasets until the work of Wang *et al.* and Wigh *et al.* has meant there has not yet been a wide-scale comparison of the existing methods, including performance by reaction class or failure modes, which represents a potential area for future work. When tested on focused reaction datasets, these 'global' models can also struggle, as the exposure to many different types of reaction can add 'noise' to predictions, as found by Afonina *et al.* [2]. On the contrary, smaller-scale models can be tailored to specific reactions, allowing the aforementioned parameters to be predicted, and enabling the incorporation of domain-specific descriptors which enhance performance [38], as we will discuss now.

5.2 'Local' Condition Prediction Models

'Local' condition prediction models can be more specialised to the reaction of interest. By focusing on a single reaction type or even a single reactant pairing, reagents can be more easily classified into their respective roles, allowing for finer-grained modelling of condition components. With smaller datasets, the feasibility of computing more information-rich descriptors, such as those based on density function theory (DFT) [116], increases. Furthermore, as the condition space is smaller for single reaction types, it is now computationally feasible to enumerate conditions. This means that other approaches to condition prediction can be employed. For example, we can now rank all conditions against each other, like in label ranking, or alternatively, we can 'screen' conditions *in silico* by predicting the yield of reactant/reagent combinations and then testing the highest yielding predictions (see Fig. 5).

5.2.1 Label Ranking

An emerging approach which is particularly suited for small data is label ranking [78]. This method aims to directly compare conditions against one another, producing a ranked list of conditions for a given input equation. Shim *et al.* applied this to good effect on 'small-data' deoxyfluorination and C-heteroatom coupling reaction datasets. Notably, this approach performed well under both a full combinatorial regime where all reactant-condition pairs were available; and under the more realistic partially complete data regime, where some reactant-condition pairs were randomly masked out [78]. However, the authors also suggest that this approach can work on some larger condition datasets, using the C-N Ullman coupling dataset curated by Samha *et al.* [59] and they highlight the benefit of aggregating condition rankings as opposed to using a simple classifier.

In a related approach, Eshel *et al.* use classifiers to assign ranks in order to select conditions for aldehyde deuteration and C-H activation reactions. They incorporate expert knowledge about the reactivity of conditions relative to the substrates they are applied to inform the choice of ordinal ranking algorithms, therefore working in a similar manner to Shim *et al.* by ranking conditions against one another [117]. Both of these recent works suggest that ranking methods could be a strong approach for condition recommendation, particularly in the small-data regime.

5.2.2 Virtual Condition Screening

As referenced already, ML-based yield prediction is an entirely different topic on its own and is covered in depth, alongside their challenges, in other reviews [29,52,118]. The benefit of accurate yield prediction models is that they can be used to filter possible reaction conditions and by selecting the conditions that lead to the highest predicted yield to test we can find optimal conditions without wasting the time and resources required to screen all the conditions experimentally. Shields *et al.* use Bayesian Optimisation (BO) as a tool to predict and select conditions based on the observed yields in an iterative learning approach [43]. Rinehart *et al.* and Samha *et al.* both found success by modelling coupling reactions using a combination of ML models with custom descriptors [38,59], and Kwon *et al.* have used GNNs alongside BO to explore optimal reaction conditions [119].

All of the above approaches carry out their experiments in an 'Iterative Learning' [120] workflow, designing and creating datasets specifically for building models of reactivity.

The alternative approach is to use existing datasets. As previously discussed, a yield prediction model can be trained and applied to small, focused datasets, with conditions predicted to lead to the optimal outcomes being selected to test (see Figure 5b). As representative examples, Schwaller *et al.* created Yield-BERT, a transformer-based model to predict reaction yields, then trained this on a small fraction of a dataset and prospectively screened the rest of the dataset to identify promising conditions [17]. Atz *et al.* used a graph-transformer neural network in a similar manner to screen conditions for a Suzuki-type cross-coupling reaction [121]. Both examples exemplify how yield prediction can be incorporated into condition recommendation, provided conditions can be enumerated.

Of course, scaling these approaches to a 'global' level is challenging, requiring predictions for all possible combinations of conditions which would be computationally intensive. It is possible that these yield-prediction models could be used as a final computational screen of 'feasible' conditions suggested by a different model, analogous to Chen and Li [32]. However, for 'local' datasets, the yield-prediction route offers a viable method of evaluating and suggesting reaction conditions.

Whilst challenges like data sparsity and evaluation remain, we have seen how progress in reaction condition prediction can come from advances in model architecture. Although progress can also come from rethinking fundamental aspects of modelling, such as data representation. Having discussed the modelling earlier, we have seen how data representation can influence predictive performance. To illustrate this, we apply models to CGR-based reaction representations and demonstrate improved performance over traditional reaction representations.

6 The Importance of Reaction Representation: A Case Study

Beker *et al.* argued that machine learning-based models some-

View Article Online
DOI: 10.1039/D5SC03045E

Table 3 Examples of condition prediction models trained on large databases, instead of HTE data

Model	Type	Data Source	Source Code?	Reference
Reacon	GNN (+ FFNN)	USPTO	Y	77
2S-DNN	FFNN (×2)	Reaxys	Y	32
TextReact	FFNN	USPTO	Y	70
Molecular Transformer	Transformer	USPTO/Reaxys	Y	46
PARROT	Transformer	USPTO/Reaxys	Y	47
ReactionVAE	VAE	Reaxys	Y	85
CIMG	GNN	Dcaiku	Y	76
LRM	FFNN	Reaxys	N	30

times simply capture literature popularity and cannot provide significant improvement over predicting the most common conditions for a given reaction [1]. Their approach involved categorising solvents and bases into 6 or 13 and 7 expert-assigned classes, respectively, then training models to predict these classes. The best performing model, a feed-forward neural network based on Morgan fingerprints [95], could not significantly outperform simply picking the most popular classes for solvent/base.

To demonstrate the impact of reaction representation, we select a different method to encode reactions: CGR fragments. We wanted to see if this encoding could produce models of improved predictive power and crucially, outperform a challenging literature baseline [1]. Afonina *et al.* introduced a method combining a multitask neural network and likelihood ranking based on CGR fragments which can produce lists of viable conditions for hydrogenation reactions [2], and we adopt a similar strategy here.

The USPTO dataset was downloaded directly from Ref. 1, where the same curation as that publication was applied, splitting solvents into 6 'coarse' classes, 13 'fine' classes and the bases into 7 classes [1]. We choose not to predict the identity of the Pd source, ligand or temperature, to enable comparison with Ref. 1, who only predict solvent and base. Full details of the identities of the clusters can be found in the Supplementary Material. The reaction itself is split into reactants and products, leaving 2 reactants and a single product. Following this procedure, we perform atom mapping using Chython [79], and an additional duplicate check, removing all reactions with the same mapped reaction equation, 'coarse' solvent class, 'fine' solvent class and base class. This leaves us with fewer reactions (5,219) than the original publication (5,434).

We then split the dataset using 5x5 cross validation (CV), using stratified sampling of the 'fine' solvent class. Whilst this differs from Beker *et al.* who use random 5x5 CV, stratified sampling ensures that the model's evaluation is more accurate, given the unbalanced nature of both the base and solvent targets [1].

To generate the model input, ISIDA fragment descriptors [51] were generated for each reaction. We used the same procedure as set out in Ref. 2, generating atom and bond-centred fragments of length two to four atoms using ISIDA Fragmentor 2017, wrapped by CIMTools [122]. We used the same additional settings as that publication, namely Formal Charge encoding and all fragments formation, creating fragments with both 'dynamic' and 'regular' bonds. Fragments occurring fewer than five times were removed, and the resulting vectors were scaled to zero mean and unit standard deviation. Finally, incremental PCA was performed to get a final CGR fragment vector of length 1500 for each reaction. For a schematic, see Fig. S1.

We created four machine learning models based on vectors formed from the PCA projection of CGR fragment count vectors: a Random Forest (RF); a Gradient Boosting Machine (GBM); a similarity search (kNN) and a multitask neural network (MTNN), similar in architecture to the best model from the work of Beker *et al.* [1]. We used ChemProp [83,98,101], based on the D-MPNN architecture, as an additional test of CGRs as a reaction representation for condition prediction. For the RF, GBM, MTNN and D-MPNN models, the hyperparameters were tuned using Optuna [123], once per iteration of the 5x5 CV. These hyperparameters were used to test the models across the rest of the folds within that repetition.

As set out in Ref. 2, we transformed the independent predictions for solvents and bases into a ranked list of combinations of these reagents using a likelihood ranking approach. To do this we first enumerated all combinations of the solvents and bases. We then determine the probability of each combination by multiplying the probabilities for the solvent and base within each combination, and finally, ranking the combinations in order of probability. The only difference to Afonina *et al.* is that we do not take the mean of the negative log-likelihoods (and minimising), but rather maximise the probability directly. See Fig. S3 for more information.

6.1 Results

With the CGR fragments created, the models were then tested across the USPTO dataset. It can be seen from Fig. S4 that the CGR-based MTNN models outperformed the literature benchmark, and the best performing model from Ref. 1 (Morgan MTNN). Furthermore, the strong performance of the similarity-based search, comparable to the performance of the Morgan fingerprint-based model gives credence to the hypothesis that 'similar reactions react under similar conditions'. Although, an alternative interpretation could be that similar reactions may have been performed within the same laboratory, and therefore used similar conditions, highlighting the selection bias prevalent in reaction datasets.

Resulting statistical analysis, using the workflow suggested by Ash *et al.* [124] demonstrate that these results are statistically significant (see Fig. S9 and S10). The other CGR-based models were also tested, but for clarity of the plots, only the best model: the MTNN, was selected to be shown. Comparisons between the CGR-based methods can be found in the Supplementary Material. Therefore, to answer the question of the case study: 'Can machine learning methods improve significantly upon literature

View Article Online
DOI: 10.1039/D5SC03045E

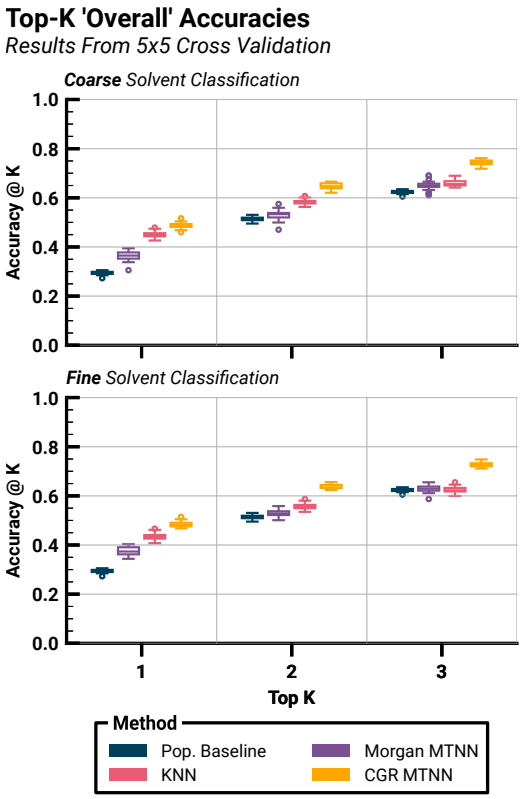

Fig. 6 Box plots of the distribution of top-k accuracies when predicting solvent and base, together. The CGR multitask network is highlighted in yellow, and outperforms the literature, similarity and existing machine learning baselines. We see the Morgan fingerprint-based model perform better than the literature popularity for top-1 accuracy, but becomes similar for k > 1. Plots for the individual solvent and base accuracies can be found in the supplementary material.

baselines on this dataset?', these results would suggest that an alternative representation, the Condensed Graph of Reaction, can outperform this baseline, on the independent predictions of solvents and bases.

However, synthetic chemists require combined predictions of all components in a chemical reaction, since solvents and bases may be incompatible, or not lead to a reaction, despite the individual components being sufficient in other cases. Therefore, we combined these independent predictions using the likelihood ranking approach, to give an indication of the performance of such a model when predicting combinations of reagents, the results of which can be found in Fig. 6. We can see that the gap between the CGR-based model and the popularity benchmark is now higher, and similarly with the Morgan fingerprint model. Although as we understand Beker *et al.* didn't include testing (or optimisation) for their Morgan fingerprint models on a combined reagent prediction task. Nonetheless, these results demonstrate that this CGR-based model can improve on the strong literature popularity benchmark. This is potentially because CGRs explicitly encode more information than Morgan fingerprints, where the transformation is not directly represented. Since the CGRs require atom mapping, the reaction centre is explicitly encoded, rather than this being implicitly encoded in other fingerprints based on the individual reactants.

Additionally, we wished to illustrate the benefit of expert-assigned reagent classification to enable fairer model evaluation. First, we generated 'exact' predictions for both the base and the solvent, Then the same clustering was applied post-prediction to highlight how clustering causes an increase in model accuracy, suggesting that when models are making 'incorrect' predictions, these predictions are still chemically relevant. The results of this can be seen in Fig. 7. It can also be seen that clustering in pre-processing can lead to improved performance, versus predicting the exact reagent and clustering post-prediction.

Our case study of Suzuki-Miyaura reactions demonstrates that existing machine learning methods can overcome popularity metrics, by using an appropriate representation. By using a CGR-based representation, we developed models that outperformed the existing state-of-the-art on the USPTO Suzuki dataset. Despite this, further improvement of the models is possible. Alternative classification metrics (see Fig. S7) show that despite the higher accuracies, these models still require improvements to truly 'learn' the underlying chemistry being modelled. This underscores the need for further improvements, either through the use of more complex architectures (though this doesn't always help, see Ref. 1) or through other strategies. For example, refining solvent and base clustering to address class imbalance and data sparsity (provided that clusters are both chemically meaningful and useful to end users). Furthermore, this modelling ignores the presence of many other variables in Suzuki reactions, like temperature, Pd-source and ligand. With an increase in the number of variables, the condition space expands: exacerbating the data sparsity problem and increasing the importance of methods to mitigate it.

Nonetheless, this example highlights the critical role of data representation in reaction condition modelling, aligning with our broader argument that thoughtful representation design is key to unlocking improvements in model performance. However, other challenges discussed in this perspective, such as data sparsity and selection bias, remain unresolved in this case study. Bridging these gaps will require continued exploration of strategies such as data augmentation, chemically informed clustering, or more advanced machine learning architectures.

Despite these limitations, the ability of models to outperform popularity benchmarks provides a step forward in bridging the gap between computational predictions and practical applications, even with existing literature data.

7 Conclusions

To conclude, this paper provides a brief introduction to the modelling and prediction of 'optimal' chemical reaction conditions, based on existing literature data.

The challenges with reaction data are well-documented[27–29], but we expect that with the increased awareness of synthetic chemists to the importance of holistic reporting of experiments that data quality in the future will continue to improve, resulting in improved models. Initiatives like ORD promote standardised recording of reaction data, which will act to counteract the existing biases. However, bridging the gap between existing datasets, and the 'ideal' datasets of the future will require continued inno-

View Article Online
DOI: 10.1039/D5SC03045E

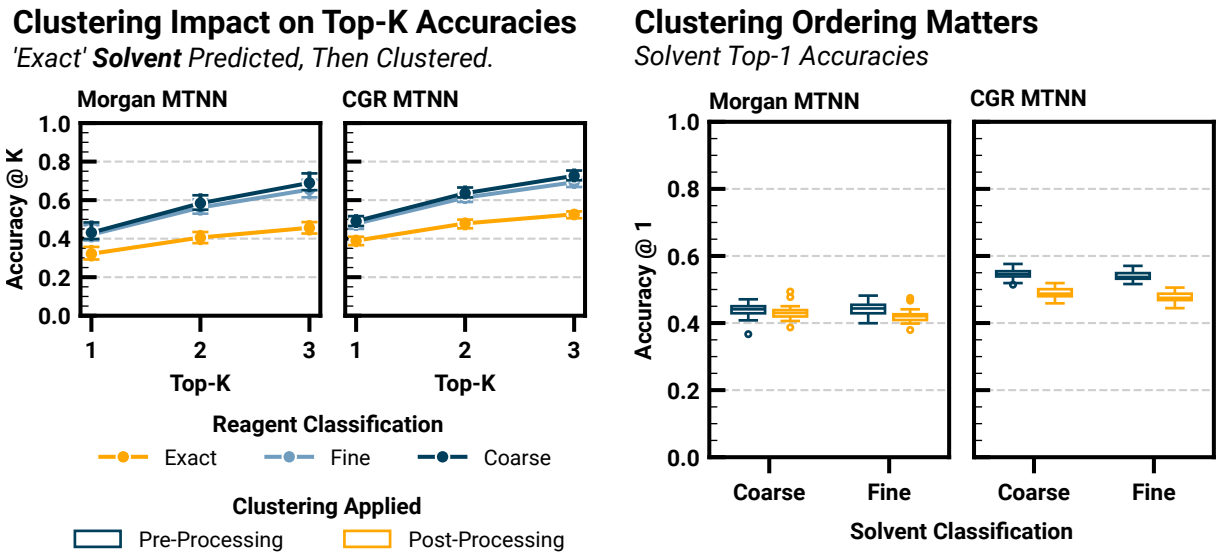

Fig. 7 Analysing the impact of clustering on solvent Top-K accuracies. Whilst we cannot say that clustering 'improves' model performance, because the clustering does not alter the underlying prediction, it can clearly be seen that the top-k accuracies increase across all methods. What this does highlight, is the need for care when evaluating model performance to understand how chemically relevant predictions are. Here we can see that despite low 'exact' accuracies, the model is still producing chemically relevant predictions in many cases. An equivalent plot for the bases can be found in Fig. S18. The right figure demonstrates that better results are obtained when training the models on the coarser solvent labels, rather than applying the clustering in post-processing.

vation, such as incorporating procedural data[70] data augmentation[71] and innovative sampling techniques[32] to maximise existing data and create generalisable, robust models.

In reaction condition prediction, the many-to-many relationship between a reaction equation and feasible conditions requires that models should predict multiple conditions for a single reaction equation, and the format of this output is dependent on the task at hand. Although the prediction of 'exact' reagents has its place in reaction optimisation, we believe that existing data requires condition predictions to adopt a more general condition encoding. As the scope of reactions considered increases - moving towards a 'global' model - and the data becomes sparser, we suggest that model outputs should generalise, for example through the categorisation of similar reagents in order to reduce the number of classes that a model is required to predict from. 'Local' models remain valuable in scenarios where data sparsity is less of a concern, such as carefully curated datasets with high condition coverage for each reaction equation. In this case, higher-fidelity condition predictions are possible, and the requirement for output 'generalisation' diminishes. With improving large-scale data quality, increasing fidelity of predictions from 'global' models may be possible in the future. In the meantime, the selection of an appropriate 'general' condition encoding remains an area for future work, and such a representation should incorporate chemical knowledge whilst compressing condition space to mitigate existing data concerns.

We provide an overview of existing models, through the lens of 'global' and 'local' models, following from the classifications of Raghavan *et al.* [45]. These different approaches have leveraged different representations, like strings (in the case of transformers), graphs and reaction fingerprints. Our case study highlights the critical role of reaction representation in reaction condition modelling, emphasising that thoughtful representation design is key to unlocking improvements in model performance. In particular, using reaction representations that explicitly encode the reaction transformation occurring, like the CGR, can improve upon the performance of other representations (like Morgan fingerprints).

By leveraging higher-quality data, the condition prediction models of the future will improve upon this current generation of models. However, during this transition period, we believe that developing novel encodings for both input and output of these models can enhance the practical applicability of these models to synthetic chemists.

Conflicts of interest

There are no conflicts to declare.

Data availability

All the source code and the dataset used to produce the reported results can be found at: https://github.com/Laboratoire-de-Chemoinformatique/CGR-Case-Study

References

1 W. Beker, R. Roszak, A. Wołos, N. H. Angello, V. Rathore, M. D. Burke and B. A. Grzybowski, *Journal of the American Chemical Society*, 2022, **144**, 4819–4827.
2 V. A. Afonina, D. A. Mazitov, A. Nurmukhametova, M. D. Shevelev, D. A. Khasanova, R. I. Nugmanov, V. A. Burilov, T. I. Madzhidov and A. Varnek, *International Journal of Molecular Sciences*, 2022, **23**, 248.
3 E. Shim, J. A. Kammeraad, Z. Xu, A. Tewari, T. Cernak and P. M. Zimmerman, *Chemical Science*, 2022, **13**, 6655–6668.

View Article Online
DOI: 10.1039/D5SC03045E

4 P. Schwaller, T. Gaudin, D. Lányi, C. Bekas and T. Laino, *Chemical Science*, 2018, **9**, 6091–6098.

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

6 H. Bi, H. Wang, C. Shi, C. Coley, J. Tang and H. Guo, Proceedings of the 38th International Conference on Machine Learning, 2021, pp. 904–913.

7 I. V. Tetko, P. Karpov, R. Van Deursen and G. Godin, *Nature Communications*, 2020, **11**, 5575.

8 S. Genheden, A. Thakkar, V. Chadimová, J.-L. Reymond, O. Engkvist and E. Bjerrum, *Journal of Cheminformatics*, 2020, **12**, 70.

9 P. Schwaller, R. Petraglia, V. Zullo, V. H. Nair, R. A. Haeuselmann, R. Pisoni, C. Bekas, A. Iuliano and T. Laino, *Chemical Science*, 2020, **11**, 3316–3325.

10 A. Thakkar, A. Vaucher, A. Byekwaso, P. Schwaller, A. Toniato and T. Laino, *Unbiasing Retrosynthesis Language Models with Disconnection Prompts*, 2022, https://chemrxiv.org/engage/chemrxiv/article-details/6328d0b8ba8a6d04fc551df7.

11 P. Torren-Peraire, A. K. Hassen, S. Genheden, J. Verhoeven, D.-A. Clevert, M. Preuss and I. Tetko, 2023.

12 T. Akhmetshin, D. Zankov, P. Gantzer, D. Babadeev, A. Pinigina, T. Madzhidov and A. Varnek, *SynPlanner: An End-to-End Tool for Synthesis Planning*, 2024, https://chemrxiv.org/engage/chemrxiv/article-details/66add90bc9c6a5c07ae65796.

13 A. M. Westerlund, L. Saigiridharan and S. Genheden, *Constrained Synthesis Planning with Disconnection-Aware Transformer and Multi-Objective Search*, 2024, https://chemrxiv.org/engage/chemrxiv/article-details/664ee4c291aefa6ce1c4fc8d.

14 A. M. Westerlund, S. Manohar Koki, S. Kancharla, A. Tibo, L. Saigiridharan, M. Kabeshov, R. Mercado and S. Genheden, *Journal of Chemical Information and Modeling*, 2024, **64**, 3021–3033.

15 D. Rappoport and A. Aspuru-Guzik, *Journal of Chemical Theory and Computation*, 2019, **15**, 4099–4112.

16 D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher and A. G. Doyle, *Science*, 2018, **360**, 186–190.

17 P. Schwaller, A. C. Vaucher, T. Laino and J.-L. Reymond, *Machine Learning: Science and Technology*, 2021, **2**, 015016.

18 J. Schleinitz, M. Langevin, Y. Smail, B. Wehnert, L. Grimaud and R. Vuilleumier, *Journal of the American Chemical Society*, 2022, **144**, 14722–14730.

19 A. Sato, T. Miyao and K. Funatsu, *Molecular Informatics*, 2022, **41**, 2100156.

20 X. Yin, C.-Y. Hsieh, X. Wang, Z. Wu, Q. Ye, H. Bao, Y. Deng, H. Chen, P. Luo, H. Liu, T. Hou and X. Yao, *Research*, 2024, **7**, 0292.

21 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, *Nature Chemistry*, 2024, **16**, 239–248.

22 P. Raghavan, A. J. Rago, P. Verma, M. M. Hassan, G. M. Goshu, A. W. Dombrowski, A. Pandey, C. W. Coley and Y. Wang, *Journal of the American Chemical Society*, 2024, **146**, 15070–15084.

23 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, *Journal of the American Chemical Society*, 2021, **143**, 18820–18826.

24 R. Mercado, S. M. Kearnes and C. W. Coley, *Journal of Chemical Information and Modeling*, 2023, **63**, 4253–4265.

25 G. Skoraczyński, P. Dittwald, B. Miasojedow, S. Szymkuć, E. P. Gajewska, B. A. Grzybowski and A. Gambin, *Scientific Reports*, 2017, **7**, 3582.

26 M. Fitzner, G. Wuitschik, R. J. Koller, J.-M. Adam, T. Schindler and J.-L. Reymond, *Chemical Science*, 2020, **11**, 13085–13093.

27 F. Strieth-Kalthoff, F. Sandfort, M. Kühnemund, F. R. Schäfer, H. Kuchen and F. Glorius, *Angewandte Chemie International Edition*, 2022, **61**, e202204647.

28 M. P. Maloney, C. W. Coley, S. Genheden, N. Carson, P. Helquist, P.-O. Norrby and O. Wiest, *The Journal of Organic Chemistry*, 2023, **88**, 5239–5241.

29 V. Voinarovska, M. Kabeshov, D. Dudenko, S. Genheden and I. V. Tetko, *Journal of Chemical Information and Modeling*, 2024, **64**, 42–56.

30 H. Gao, T. J. Struble, C. W. Coley, Y. Wang, W. H. Green and K. F. Jensen, *ACS Central Science*, 2018, **4**, 1465–1476.

31 M. R. Maser, A. Y. Cui, S. Ryou, T. J. DeLano, Y. Yue and S. E. Reisman, *Journal of Chemical Information and Modeling*, 2021, **61**, 156–166.

32 L.-Y. Chen and Y.-P. Li, *Journal of Cheminformatics*, 2024, **16**, 11.

33 B. C. Haas, A. E. Goetz, A. Bahamonde, J. C. McWilliams and M. S. Sigman, *Proceedings of the National Academy of Sciences*, 2022, **119**, e2118451119.

34 J. W. Sin, S. L. Chau, R. P. Burwood, K. Püntener, R. Bigler and P. Schwaller, *Highly Parallel Optimisation of Nickel-Catalysed Suzuki Reactions through Automation and Machine Intelligence*, 2024, https://chemrxiv.org/engage/chemrxiv/article-details/66f976b5cec5d6c142981e4e.

35 S. S. Gandhi, G. Z. Brown, S. Aikonen, J. S. Compton, P. Neves, J. I. Martinez Alvarado, I. I. Strambeanu, K. A. Leonard and A. G. Doyle, *ACS Catalysis*, 2025, **15**, 2292–2304.

36 T. Gimadiev, T. Madzhidov, I. Tetko, R. Nugmanov, I. Casciuc, O. Klimchuk, A. Bodrov, P. Polishchuk, I. Antipin and A. Varnek, *Molecular Informatics*, 2019, **38**, 1800104.

37 M. H. S. Segler, M. Preuss and M. P. Waller, *Nature*, 2018, **555**, 604–610.

38 N. I. Rinehart, R. K. Saunthwal, J. Wellauer, A. F. Zahrt, L. Schlemper, A. S. Shved, R. Bigler, S. Fantasia and S. E. Denmark, *Science*, 2023, **381**, 965–972.

39 N. H. Angello, V. Rathore, W. Beker, A. Wołos, E. R. Jira, R. Roszak, T. C. Wu, C. M. Schroeder, A. Aspuru-Guzik, B. A.

View Article Online
DOI: 10.1039/D5SC03045E

Grzybowski and M. D. Burke, *Science*, 2022, **378**, 399–405.
40 J. Y. Wang, J. M. Stevens, S. K. Kariofillis, M.-J. Tom, D. L. Golden, J. Li, J. E. Tabora, M. Parasram, B. J. Shields, D. N. Primer, B. Hao, D. Del Valle, S. DiSomma, A. Furman, G. G. Zipp, S. Melnikov, J. Paulson and A. G. Doyle, *Nature*, 2024, **626**, 1025–1033.
41 S. P. Schmid, E. M. Rajaonson, C. T. Ser, M. Haddadnia, S. X. Leong, A. Aspuru-Guzik, A. Kristiadi, K. Jorner and F. Strieth-Kalthoff, AI for Accelerated Materials Design - NeurIPS 2024, 2024.
42 F. Häse, L. M. Roch, C. Kreisbeck and A. Aspuru-Guzik, *ACS Central Science*, 2018, **4**, 1134–1145.
43 B. J. Shields, J. Stevens, J. Li, M. Parasram, F. Damani, J. I. M. Alvarado, J. M. Janey, R. P. Adams and A. G. Doyle, *Nature*, 2021, **590**, 89–96.
44 E. Braconi and E. Godineau, *ACS Sustainable Chemistry & Engineering*, 2023, **11**, 10545–10554.
45 P. Raghavan, B. C. Haas, M. E. Ruos, J. Schleinitz, A. G. Doyle, S. E. Reisman, M. S. Sigman and C. W. Coley, *ACS Central Science*, 2023, **9**, 2196–2204.
46 M. Andronov, V. Voinarovska, N. Andronova, M. Wand, D.-A. Clevert and J. Schmidhuber, *Chemical Science*, 2023, **14**, 3235–3246.
47 X. Wang, C.-Y. Hsieh, X. Yin, J. Wang, Y. Li, Y. Deng, D. Jiang, Z. Wu, H. Du, H. Chen, Y. Li, H. Liu, Y. Wang, P. Luo, T. Hou and X. Yao, *Research*, 2023, **6**, 0231.
48 G. Marcou, J. Aires de Sousa, D. A. R. S. Latino, A. de Luca, D. Horvath, V. Rietsch and A. Varnek, *Journal of Chemical Information and Modeling*, 2015, **55**, 239–250.
49 J. Li and M. D. Eastgate, *Reaction Chemistry & Engineering*, 2019, **4**, 1595–1607.
50 S. Genheden, A. Mårdh, G. Lahti, O. Engkvist, S. Olsson and T. Kogej, *Molecular Informatics*, 2022, **41**, 2100294.
51 A. Varnek, D. Fourches, F. Hoonakker and V. P. Solov'ev, *Journal of Computer-Aided Molecular Design*, 2005, **19**, 693–703.
52 M. Fitzner, G. Wuitschik, R. Koller, J.-M. Adam and T. Schindler, *ACS Omega*, 2023, **8**, 3017–3025.
53 *Chemical Reactions from US Patents (1976-Sep2016)*, 2017, `https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873/1`.
54 U. O. o. P. Affairs (OPA), *United States Patent and Trademark Office*, `https://www.uspto.gov/`.
55 W. A. Warr, W. Warr, B. Court and H. Chapel.
56 A. Lawson, J. Swienty-Busch, T. Geoui and D. Evans, *ACS Symposium Series*, 2014, vol. 1164, pp. 127–148.
57 *Reaxys Chemistry Database*, `https://www.reaxys.com`.
58 *Chemical Abstract Services*, `https://www.cas.org/`.
59 M. H. Samha, L. J. Karas, D. B. Vogt, E. C. Odogwu, J. Elward, J. M. Crawford, J. E. Steves and M. S. Sigman, *Science Advances*, 2024, **10**, eadn3478.
60 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.
61 D. Young, T. Martin, R. Venkatapathy and P. Harten, *QSAR & Combinatorial Science*, 2008, **27**, 1337–1345.
62 A. Toniato, P. Schwaller, A. Cardinale, J. Geluykens and T. Laino, *Unassisted Noise Reduction of Chemical Reaction Data Sets*, 2021, `http://arxiv.org/abs/2102.01399`.
63 T. R. Gimadiev, A. Lin, V. A. Afonina, D. Batyrshin, R. I. Nugmanov, T. Akhmetshin, P. Sidorov, N. Duybankova, J. Verhoeven, J. Wegner, H. Ceulemans, A. Gedich, T. I. Madzhidov and A. Varnek, *Molecular Informatics*, 2021, **40**, 2100119.
64 D. S. Wigh, J. Arrowsmith, A. Pomberger, K. C. Felton and A. A. Lapkin, *Journal of Chemical Information and Modeling*, 2024, **64**, 3790–3798.
65 D. Fourches, E. Muratov and A. Tropsha, *Journal of Chemical Information and Modeling*, 2016, **56**, 1243–1252.
66 F. Zipoli, Z. Ayadi, P. Schwaller, T. Laino and A. C. Vaucher, *Machine Learning: Science and Technology*, 2024, **5**, 025071.
67 C. Zhang, A. Arun and A. A. Lapkin, *ACS Omega*, 2024, **9**, 18385–18399.
68 R. Nugmanov, T. Madzhidov, I. Antipin and A. Varnek, 2018.
69 R. I. Nugmanov, R. N. Mukhametgaleev, T. Akhmetshin, T. R. Gimadiev, V. A. Afonina, T. I. Madzhidov and A. Varnek, *Journal of Chemical Information and Modeling*, 2019, **59**, 2516–2521.
70 Y. Qian, Z. Li, Z. Tu, C. W. Coley and R. Barzilay, *Predictive Chemistry Augmented with Text Retrieval*, 2023, `http://arxiv.org/abs/2312.04881`.
71 P. Schwaller, A. C. Vaucher, T. Laino and J.-L. Reymond, *Data Augmentation Strategies to Improve Reaction Yield Predictions and Estimate Uncertainty*, 2020, `https://chemrxiv.org/engage/chemrxiv/article-details/60c75258702a9b726c18c101`.
72 M. E. Fortunato, C. W. Coley, B. C. Barnes and K. F. Jensen, *Journal of Chemical Information and Modeling*, 2020, **60**, 3398–3407.
73 F. Jaume-Santero, A. Bornet, A. Valery, N. Naderi, D. Vicente Alvarez, D. Proios, A. Yazdani, C. Bournez, T. Fessard and D. Teodoro, *Journal of Chemical Information and Modeling*, 2023, **63**, 1914–1924.
74 E. King-Smith, F. A. Faber, U. Reilly, A. V. Sinitskiy, Q. Yang, B. Liu, D. Hyek and A. A. Lee, *Nature Communications*, 2024, **15**, 426.
75 E. Shim, A. Tewari, T. Cernak and P. M. Zimmerman, *Journal of Chemical Information and Modeling*, 2023, **63**, 3659–3668.
76 B. Zhang, X. Zhang, W. Du, Z. Song, G. Zhang, G. Zhang, Y. Wang, X. Chen, J. Jiang and Y. Luo, *Proceedings of the National Academy of Sciences*, 2022, **119**, e2212711119.
77 Z. Wang, K. Lin, J. Pei and L. Lai, *Chemical Science*, 2024.
78 E. Shim, A. Tewari, T. Cernak and P. M. Zimmerman, *Chemical Science*, 2025, **16**, 4109–4118.
79 R. Nugmanov, N. Dyubankova, A. Gedich and J. K. Wegner, *Journal of Chemical Information and Modeling*, 2022, **62**, 3307–3315.
80 *Extraction of Organic Chemistry Grammar from Unsupervised*

View Article Online
DOI: 10.1039/D5SC03045E

Learning of Chemical Reactions | Science Advances, https://www.science.org/doi/10.1126/sciadv.abe4166.

81 A. Lin, N. Dyubankova, T. I. Madzhidov, R. I. Nugmanov, J. Verhoeven, T. R. Gimadiev, V. A. Afonina, Z. Ibragimova, A. Rakhimbekova, P. Sidorov, A. Gedich, R. Suleymanov, R. Mukhametgaleev, J. Wegner, H. Ceulemans and A. Varnek, *Molecular Informatics*, 2022, **41**, 2100138.

82 T. I. Madzhidov, P. G. Polishchuk, R. I. Nugmanov, A. V. Bodrov, A. I. Lin, I. I. Baskin, A. A. Varnek and I. S. Antipin, *Russian Journal of Organic Chemistry*, 2014, **50**, 459–463.

83 E. Heid and W. H. Green, *Journal of Chemical Information and Modeling*, 2022, **62**, 2101–2110.

84 A. I. Lin, T. I. Madzhidov, O. Klimchuk, R. I. Nugmanov, I. S. Antipin and A. Varnek, *Journal of Chemical Information and Modeling*, 2016, **56**, 2140–2148.

85 Y. Kwon, S. Kim, Y.-S. Choi and S. Kang, *Journal of Chemical Information and Modeling*, 2022, **62**, 5952–5960.

86 O. T. Schilter, C. Baldassari, T. Laino and P. Schwaller, *Predicting Solvents with the Help of Artificial Intelligence*, 2023, https://chemrxiv.org/engage/chemrxiv/article-details/64d4ebad4a3f7d0c0df0eaa2.

87 *Solvent Similarity Index - Physical Chemistry Chemical Physics (RSC Publishing)*, https://pubs.rsc.org/en/content/articlelanding/2020/cp/d0cp01570a#.

88 K. V. Chuang and M. J. Keiser, *Science*, 2018, **362**, eaat8603.

89 I. W. Davies, *Nature*, 2019, **570**, 175–181.

90 *CINF 13: Pistachio - Search and Faceting of Large Reaction Databases*, 2017, https://www.slideshare.net/slideshow/cinf-13-pistachio-search-and-faceting-of-large-reaction-databases/78996582.

91 M. R. Dobbelaere, I. Lengyel, C. V. Stevens and K. M. Van Geem, *Journal of Cheminformatics*, 2024, **16**, 37.

92 E. Walker, J. Kammeraad, J. Goetz, M. T. Robo, A. Tewari and P. M. Zimmerman, *Journal of Chemical Information and Modeling*, 2019, **59**, 3645–3654.

93 J. Johnson, M. Douze and H. Jégou, *Billion-Scale Similarity Search with GPUs*, 2017, http://arxiv.org/abs/1702.08734.

94 *FAISS*, https://ai.meta.com/tools/faiss.

95 H. L. Morgan, *Journal of Chemical Documentation*, 1965, **5**, 107–113.

96 E. Caldeweyher, M. Elkin, G. Gheibi, M. Johansson, C. Sköld, P.-O. Norrby and J. F. Hartwig, *Journal of the American Chemical Society*, 2023, **145**, 17367–17376.

97 B. Haas, M. Hardy, S. Sowndarya S. V., K. Adams, C. Coley, R. Paton and M. Sigman, *Rapid Prediction of Conformationally-Dependent DFT-Level Descriptors Using Graph Neural Networks for Carboxylic Acids and Alkyl Amines*, 2024, https://chemrxiv.org/engage/chemrxiv/article-details/65d79de59138d23161bec6e6.

98 K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen and R. Barzilay, *Journal of Chemical Information and Modeling*, 2019, **59**, 3370–3388.

99 J. M. Stokes, K. Yang, K. Swanson, W. Jin, A. Cubillos-Ruiz, N. M. Donghia, C. R. MacNair, S. French, L. A. Carfrae, Z. Bloom-Ackermann, V. M. Tran, A. Chiappino-Pepe, A. H. Badran, I. W. Andrews, E. J. Chory, G. M. Church, E. D. Brown, T. S. Jaakkola, R. Barzilay and J. J. Collins, *Cell*, 2020, **180**, 688–702.e13.

100 D. Buterez, J. P. Janet, S. J. Kiddle, D. Oglic and P. Lió, *Nature Communications*, 2024, **15**, 1517.

101 E. Heid, K. P. Greenman, Y. Chung, S.-C. Li, D. E. Graff, F. H. Vermeire, H. Wu, W. H. Green and C. J. McGill, *Journal of Chemical Information and Modeling*, 2024, **64**, 9–17.

102 S. Luo, W. Gao, Z. Wu, J. Peng, C. W. Coley and J. Ma, *Projecting Molecules into Synthesizable Chemical Spaces*, 2024, https://arxiv.org/abs/2406.04628v1.

103 Z. Zhou, S. Kearnes, L. Li, R. N. Zare and P. Riley, *Scientific Reports*, 2019, **9**, 10752.

104 T. Akhmetshin, A. Lin, D. Mazitov, Y. Zabolotna, E. Ziaikin, T. Madzhidov and A. Varnek, *Journal of Chemical Information and Modeling*, 2022, **62**, 3524–3534.

105 X. Liu, K. Ye, H. W. T. van Vlijmen, A. P. IJzerman and G. J. P. van Westen, *Journal of Cheminformatics*, 2023, **15**, 24.

106 J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals and G. E. Dahl, *Neural Message Passing for Quantum Chemistry*, 2017, http://arxiv.org/abs/1704.01212.

107 T. N. Kipf and M. Welling, *Semi-Supervised Classification with Graph Convolutional Networks*, 2017, http://arxiv.org/abs/1609.02907.

108 D. P. Kingma and M. Welling, *Auto-Encoding Variational Bayes*, 2022, http://arxiv.org/abs/1312.6114.

109 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, *Attention Is All You Need*, 2023, http://arxiv.org/abs/1706.03762.

110 J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis, *Nature*, 2021, **596**, 583–589.

111 Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido and A. Rives, *Science*, 2023, **379**, 1123–1130.

112 E. Castro, A. Godavarthi, J. Rubinfien, K. Givechian, D. Bhaskar and S. Krishnaswamy, *Nature Machine Intelligence*, 2022, **4**, 840–851.

113 P. Schwaller, D. Probst, A. C. Vaucher, V. H. Nair, D. Kreutter, T. Laino and J.-L. Reymond, *Nature Machine Intelligence*, 2021, **3**, 144–152.

114 J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2019, http://arxiv.org/abs/1810.04805.

View Article Online
DOI: 10.1039/D5SC03045E

115 Y. Zhang, R. Yu, K. Zeng, D. Li, F. Zhu, X. Yang, Y. Jin and Y. Xu, *Text-Augmented Multimodal LLMs for Chemical Reaction Condition Recommendation*, 2024, https://arxiv.org/abs/2407.15141v1.

116 W. Kohn, A. D. Becke and R. G. Parr, *The Journal of Physical Chemistry*, 1996, **100**, 12974–12980.

117 I. L. Eshel, B. Shahar, S. Barranco, M. Pérez-Temprano and A. Milo, *Probability Guided Chemical Reaction Scopes*, 2025, https://chemrxiv.org/engage/chemrxiv/article-details/67a66bde6dde43c908fb7a13.

118 C. J. Taylor, A. Pomberger, K. C. Felton, R. Grainger, M. Barecka, T. W. Chamberlain, R. A. Bourne, C. N. Johnson and A. A. Lapkin, *Chemical Reviews*, 2023, **123**, 3089–3126.

119 Y. Kwon, D. Lee, J. W. Kim, Y.-S. Choi and S. Kim, *ACS Omega*, 2022, **7**, 44939–44950.

120 B. Settles, *Active Learning Literature Survey*, University of wisconsin-madison department of computer sciences technical Report, 2009.

121 K. Atz, D. F. Nippa, A. T. Müller, V. Jost, A. Anelli, M. Reutlinger, C. Kramer, R. E. Martin, U. Grether, G. Schneider and G. Wuitschik, *RSC Medicinal Chemistry*, 2024, **15**, 2310–2321.

122 A. Rakhimbekova, T. I. Madzhidov, R. I. Nugmanov, T. R. Gimadiev, I. I. Baskin and A. Varnek, *International Journal of Molecular Sciences*, 2020, **21**, 5542.

123 T. Akiba, S. Sano, T. Yanase, T. Ohta and M. Koyama, *Optuna: A Next-generation Hyperparameter Optimization Framework*, 2019, http://arxiv.org/abs/1907.10902.

124 J. R. Ash, C. Wognum, R. Rodríguez-Pérez, M. Aldeghi, A. C. Cheng, D.-A. Clevert, O. Engkvist, C. Fang, D. J. Price, J. M. Hughes-Oliver and W. P. Walters, *Practically Significant Method Comparison Protocols for Machine Learning in Small Molecule Drug Discovery.*, 2024, https://chemrxiv.org/engage/chemrxiv/article-details/672a91bd7be152b1d01a926b.

View Article Online
DOI: 10.1039/D5SC03045E

Data availability

All the source code and datasets used to produce the reported results can be found at: https://github.com/Laboratoirede-Chemoinformatique/CGR-Case-Study