Artificial Intelligence for Computer-Aided Synthesis In Flow: Analysis and Selection of Reaction Components

As Published: 10.3389/FCENG.2020.00005

Publisher: Frontiers Media SA

Persistent URL: https://hdl.handle.net/1721.1/135253

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

frontiers in Chemical Engineering

ORIGINAL RESEARCH
published: 04 August 2020
doi: 10.3389/fceng.2020.00005

Artificial Intelligence for Computer-Aided Synthesis *In Flow*: Analysis and Selection of Reaction Components

Pieter P. Plehiers[1,2], Connor W. Coley[1], Hanyu Gao[1], Florence H. Vermeire[1], Maarten R. Dobbelaere[2], Christian V. Stevens[3], Kevin M. Van Geem[2] and William H. Green[1]**

[1] Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA, United States, [2] Laboratory for Chemical Technology, Department of Materials, Textiles and Chemical Engineering, Ghent University, Ghent, Belgium, [3] SynBioC Research Group, Department of Green Chemistry and Technology, Faculty of Bioscience Engineering, Ghent University, Ghent, Belgium

OPEN ACCESS

Edited by:
René Schenkendorf,
Technische Universitat Braunschweig, Germany

Reviewed by:
Richard Anthony Bourne,
University of Leeds, United Kingdom
Alexei Lapkin,
University of Cambridge, United Kingdom

***Correspondence:**
Kevin M. Van Geem
kevin.vangeem@ugent.be
William H. Green
whgreen@mit.edu

Specialty section:
This article was submitted to Computational Methods in Chemical Engineering, a section of the journal Frontiers in Chemical Engineering

Received: 24 April 2020
Accepted: 26 June 2020
Published: 04 August 2020

Citation:
Plehiers PP, Coley CW, Gao H, Vermeire FH, Dobbelaere MR, Stevens CV, Van Geem KM and Green WH (2020) Artificial Intelligence for Computer-Aided Synthesis In Flow: Analysis and Selection of Reaction Components. Front. Chem. Eng. 2:5. doi: 10.3389/fceng.2020.00005

Computer-aided synthesis has received much attention in recent years. It is a challenging topic in itself, due to the high dimensionality of chemical and reaction space. It becomes even more challenging when the aim is to suggest syntheses that can be performed in continuous flow. Though continuous flow offers many potential benefits, not all reactions are suited to be operated continuously. In this work, three machine learning models have been developed to provide an assessment of whether a given reaction may benefit from continuous operation, what the likelihood of success in continuous flow is for a certain set of reaction components (i.e., reactants, reagents, solvents, catalysts, and products) and, if the likelihood of success is low, which alternative reaction components can be considered. The first model uses an abstract version of a reaction template, obtained via gaussian mixture modeling, to quantify its relative increase in publishing frequency in continuous flow, without relying on potentially ambiguously defined reaction templates. The second model is an artificial neural network that categorizes feasible and infeasible reaction components with a 75% success rate. A set of reaction components is considered to be feasible if there is an explicit reference to it being used in continuous synthesis in the database; all other reaction components are considered infeasible. While several cases that are "infeasible" by this definition, are classified as feasible by the neural network, further analysis shows that for many of these cases, it is at least plausible that they are in fact feasible – they simply have not been tested to (dis)prove this. The final model suggests alternative continuous flow components with a top-1 accuracy of 95%. Combined, they offer a black-box evaluation of whether a reaction and a set of reaction components can be considered promising for continuous syntheses.

Keywords: continuous synthesis, computer-aided synthesis planning, artificial neural networks, gaussian mixture model, reaction conditions

INTRODUCTION

The development of new active pharmaceutical ingredients (APIs) is a time-consuming and expensive process (DiMasi et al., 1991, 2003), with up to half of the total cost being spent in the pre-clinical phase (Adams and Van Brantner, 2006). Two important research topics in this phase are chemical discovery – identification of promising APIs – and chemical development – devising syntheses for the most promising ones. Methods to accelerate these tasks can be of crucial importance in reducing both the economic and time costs of drug development.

In the area of API identification, the advent of powerful machine learning techniques has provided many methods for molecule discovery (Bajorath, 2015; Schneider, 2017; Segler et al., 2018a; Zhang et al., 2018) and molecule property assessment (Burbidge et al., 2001; Ivanciuc, 2009; Ma et al., 2015; Maltarollo et al., 2015; Mayr et al., 2016; Ryu et al., 2018).

The search for syntheses for target molecules was formalized as retrosynthetic analysis in the 1960's (Corey, 1967; Corey and Wipke, 1969), and ever since, attempts have been made to automate it through computer-aided synthesis planning (CASP). While initially automated retrosynthetic tools and synthesis planners faced the skepticism of the chemical community (Gillies, 1996; Langley, 1998), new interest has been sparked (Cook et al., 2012; Warr, 2014; Szymkuć et al., 2016), particularly using machine learning algorithms (Bøgevig et al., 2015; Coley et al., 2018a; Segler et al., 2018b). In the task of CASP, one can identify several sub-challenges that have to be dealt with (Peplow, 2014; Szymkuć et al., 2016). A first is the retrosynthetic analysis of a molecule – iteratively identifying possible precursors. Initially, this was done via a rule-based approach, using a predefined set of chemical reaction templates, which were iteratively applied to the target molecules and its subsequent precursors (Law et al., 2009; Christ et al., 2012; Bøgevig et al., 2015). More recently, alternative approaches that do not rely on predefined chemical rules have been developed. Some suggest transformations based on the known "Network of Organic Chemistry" (Laidler and King, 1983; Fuller et al., 2012; Gothard et al., 2012; Kowalik et al., 2012; Cadeddu et al., 2014). Molecular similarity can also be used to suggest chemical transformations (Coley et al., 2017b). The currently most popular approach, however, is using various types of artificial neural networks to predict which transformations a given molecule can undergo (Segler et al., 2018b; Karpov et al., 2019; Lin et al., 2020; Zheng et al., 2020). Independent of the retrosynthetic approach, a crucial aspect of the analysis is mitigating the combinatorial explosion of possible reactions. Similarly to the retrosynthetic analysis step, distinction can be made between rule-based or heuristic methods (Huang et al., 2011) and more fuzzy methods, often based on artificial neural networks (Segler and Waller, 2017b; Coley et al., 2018b). Irrespective of the approach, these methods seek to eliminate reactions that cannot be performed in practice or that lead away from the target molecule. The effect of the reaction conditions can also be a factor in assessing whether or not to retain a certain reaction. This requires methods that annotate a reaction with a suitable set of reaction conditions, which has been the topic of several studies in the past few years (Marcou et al., 2015; Segler and Waller, 2017a; Gao et al., 2018; Nielsen et al., 2018; Li and Eastgate, 2019).

Both CASP and the recommendation of reaction conditions have progressed significantly – to the point of developing a robotic synthesis platform (Coley et al., 2019b). However, one characteristic that is attributed to such a "Robochemist" (Peplow, 2014) – the ability to automatically synthesize components in a continuous way – has not yet been explicitly addressed. Continuous flow can offer several benefits compared to traditional batch syntheses, including improved safety and control, higher process efficiency and more efficient process optimization (Plutschack et al., 2017). Additionally, continuous flow reactors are an important technology in moving to greener, more sustainable production through process intensification and, e.g., decreased solvent usage (Wiles and Watts, 2012). On the other hand, some batch syntheses can be equally sustainable as their continuous flow alternative, therefore the costs of switching existing batch processes to continuous ones cannot always be justified (Roberge et al., 2008; Calabrese and Pissavini, 2011; Teoh et al., 2016).

Running a process in a continuous fashion introduces many additional considerations for process design (Hartman and Jensen, 2009). Reactions in which some reagents are solids or those in which the equilibrium is driven in a certain direction via precipitation of the products are difficult to develop in flow as the solids can clog the channels (e.g., cross-couplings in organic solvents leading to precipitation of metal-halide salts). Improper combinations of solvents and reactants, products, or catalysts can also result in precipitation of certain compounds, again increasing the risk of clogging. Reactions requiring long residence times may require unpractically long reactors or inefficiently low flow rates. At very low flow rates, mixing becomes an important issue, implying that the use of a stirred spheroidal reactor is much more economical. As a result, it is very difficult to anticipate whether a synthesis can be performed continuously in an economic way. A detailed *in silico* evaluation would require the quantitative calculation of both solubilities of many different chemicals in organic solvents and reaction kinetics. Both of these are very difficult, especially at typical reaction temperatures and each form an entire research field on their own, for any single reaction. Nonetheless, a method has been developed to allow ab-initio exploration of organic chemistry (Wang et al., 2014). The alternative is performing time-consuming laboratory scale experiments. Given the above considerations and the vast number of reactions that must be evaluated, it is clear that it is very challenging to – with an acceptable precision – direct a CASP program toward syntheses with a high likelihood of being an economic option in continuous flow.

In this work we propose a data-driven method, based on a statistical analysis of published reactions, that can identify such continuous syntheses and can potentially guide retrosynthetic software toward such syntheses. Generally, the

Abbreviations: FCD, Flow Conditions Database; CASP, Computer-Assisted Synthesis Planning; FRD, Flow Reactions Database; EF, Enrichment Factor; PCA, Principal Component Analysis; API, Active Pharmaceutical Ingredient; GMM, Gaussian Mixture Model; ORD, Overall Reactions Database; QSAR, Quantitative Structure-Activity Relation; TPR, True Positive Rate; TNR, True Negative Rate; AUC, Area Under Curve; PRC, Precision Recall Curve.

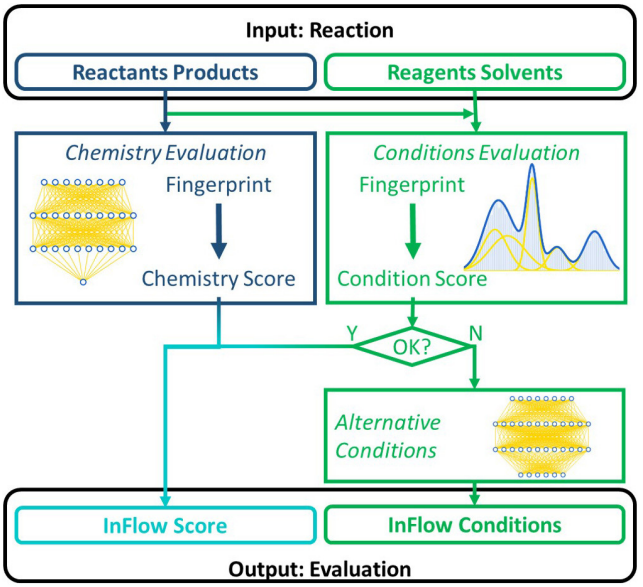


FIGURE 1 | Schematic overview of the algorithm to assess the suitability of a given reaction for continuous flow. A reaction case comprises reactants and (major) products and the (chemical) reaction conditions. The colors correspond to the different subtasks of the algorithm, which are discussed in different sections. Rounded fields are inputs/outputs, while the rectangular fields represent models.

aforementioned obstacles can be categorized into two main categories. A first general obstacle is the reaction chemistry. The second is related to the reaction conditions (reagents, solvents, catalysts, temperature, etc.). A schematic illustration of the combined algorithm is given in **Figure 1**. Reaction chemistry and conditions are evaluated separately. For those reactions of which the chemistry is evaluated positively, but the initially suggested conditions have been evaluated negatively, an alternative set of conditions is suggested. The final result for each reaction is a fast, numerical assessment of the likelihood that the reaction can be economically executed in a continuous flow reactor, combined with an alternative set of conditions for those reactions with a poor evaluation. This score can extend existing scoring methods already used to bias or filter the retrosynthetic tree search (Kowalik et al., 2012; Li and Eastgate, 2015; Coley et al., 2017a, 2018b).

DATA PREPARATION

The data-driven method described in this work is based on available literature data from the Reaxys database (Elsevier R&D Solutions, 2016). In what follows, an important distinction is be made between a *reaction* and an *instance* of a reaction. A reaction refers to a certain combination of reactants and products, *e.g.*, "C1=CC=CC1.COC(=O)C=C>>C1=CC(C2)C(CC(=O)OC)CC12" in SMILES notation (Weininger, 1988), and hence contains only information of the overall transformation that occurs during the reaction. An instance is a specific instantiation of that reaction—one specific published example—with additional details including the reagents, solvents, catalysts, temperature, etc. used. Therefore, an instance implicitly contains information about other chemical and physical properties of the reaction, such as the reaction rate. Instances in the Reaxys database may also contain an additional hand-curated annotation describing aspects of the implementation, including "flow," "flow reactor," and "micro reactor" in addition to more general statements like "inert atmosphere."

Our data processing pipeline (**Figure 2**) focuses on only reactions for which at least one instantiation is explicitly labeled as compatible with flow. Any reaction that has at least one instance that contains a keyword implying it has been performed in continuous flow, is added to the Flow Reactions Database (FRD). In the example in **Figure 2**, two of the three instances associated with the reaction contain a keyword indicating they have been used in continuous synthesis. Therefore, the example reaction is included in the FRD (without any reaction conditions). If none of the instances have such a keyword, the reaction is excluded from the set. Though the set of search keywords is made as inclusive as possible (*cfr.* **Supporting Information S-1.1**), it is not guaranteed that all continuous synthesis papers in Reaxys are extracted and considered. 7,042 such reactions are identified in the Reaxys database. This dataset is used to train the model that provides an assessment whether continuous synthesis is beneficial to the reaction in terms of the chemistry. For only the reactions in the FRD, all associated instances are processed into the Flow Conditions Database (FCD) by collecting their reported reaction conditions and assigning a label "flow" or "batch," depending on whether the instance has an explicit label indicating flow (or its conditions exactly match those of a flow-labeled instance of the same reaction). For the example, this would imply that three new instances are entered in the FCD: "4-chloroaniline + phenylboronic acic→4-chloro-N-phenylaniline with triethylamine in dichloromethane," "4-chloroaniline + phenylboronic acic→4-chloro-N-phenylaniline with pyridine and oxygen in dichloromethane" and "4-chloroaniline + phenylboronic acic→4-chloro-N-phenylaniline with potassium carbonate in acetonitrile." In total, 72,800 such instances are collected in the FCD, implying an average of 10.34 published sets of reaction conditions per reaction. For the example reaction, two instances carry the "flow" label. Overall, 11,275 of the 72,800 instances are assigned the "flow" label, indicating that on average each reaction in the FRD is associated with 1.6 flow-benefitting instances. The FCD is used to train a model to evaluate whether a reaction benefits from continuous synthesis in term of the applied reaction conditions and to train another model to predict flow-benefitting alternatives for batch conditions.

Before using the FRD or FCD for statistical learning, we examine the nature of these datasets. As one would expect, only a small fraction of known reaction types have been studied under continuous flow conditions. Following previously-reported heuristic extraction procedures (Law et al., 2009; Bøgevig et al., 2015; Coley et al., 2019a), 2,586 reaction templates are identified from the FRD, in contrast to 2.9 million reaction templates from the entire Reaxys database, of which 366,000 are represented by five or more different reaction examples. This indicates that the scientific community has been focusing on rather specific types of chemistry when continuous synthesis

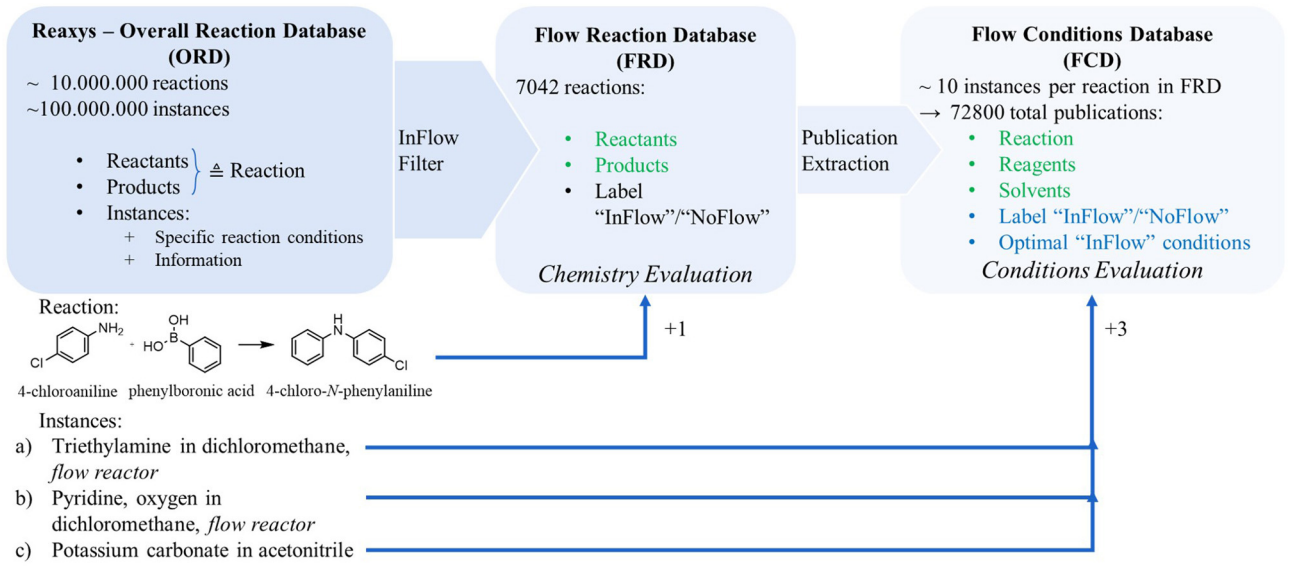


FIGURE 2 | Illustration of data extraction from the Reaxys reaction database. The InFlow filter selects reactions for which at least one publication mentions application of continuous flow. The publications contained in the reaction entries of the Flow Reaction Database (FRD) are expanded into a second database – the Flow Conditions Database (FCD). The bullets represent the data fields associated with each entry of the respective database. The green-colored bullets are used as input for the chemistry and reaction components evaluation, while the blue-colored bullets are the outputs the reaction components evaluation models are trained on.

is concerned (e.g., related to hazardous reagents or "forbidden chemistry"). **Figure 3** provides further support and shows that certain templates are much more popular in continuous synthesis when compared to their overall popularity in literature and *vice versa*. Of the 10 overall most popular templates, three are not found in the FRD at all, while only two are reported in the FRD with a significant frequency. In contrast to the other eight templates on the "most-popular" list, these two reactions commonly involve forcing conditions to achieve redox chemistry, such as hydrogen and harsh oxidizing agents. A second observation on the template popularity is made from **Figure 4A**, where the enrichment factor (EF) is defined as the ratio of the relative frequency of a template in the flow reaction database and the relative frequency of the same template in the overall database. **Figure 4A** shows that the majority of the templates enjoys a significantly higher popularity within the flow database than overall, with a median enrichment factor of 125. Hazardous reactions such as the ones mentioned above, but also azidations and nitrations, are examples that tend to have high enrichment factors. In those cases, increased process safety is one of the main benefits of continuous synthesis (Movsisyan et al., 2016).

These concepts of popularity and enrichment can be extended to the reaction conditions as well. **Figure 4B** indicates that there is no general enrichment for the solvents. **Figure 5** shows that the overall most popular solvents remain quite popular in continuous synthesis. There are, however, certain solvents that show a high preference when used in continuous synthesis. One example is cyclohexane, which is used as solvent in 2.7% of the reported continuous syntheses, ranking it as 10th most popular solvent. In the overall Reaxys database however, cyclohexane is used in only 0.3% of the syntheses, ranking it 28th. Other solvents with a high enrichment are N,N-dimethyl-acetamide, 2-methyl tetrahydrofuran, propanol, and nitrogen. More interesting is the case of the reagents. Reaxys distinguishes between a "catalyst" and a "reagent" in their tabulation, but in this work, no distinction is made between catalysts and reagents: the two fields are grouped into one that is henceforth referred to as "reagents." **Figure 4C** indicates that there are many reagents with an increased popularity in continuous syntheses. The relatively large number of reagents in the bottom left quarter of the plot is due to a large number of unique reagents being reported in Reaxys only once or twice. Many of these are very specific mixtures of metal(oxide)s – $Cr_{1.3}Fe_{0.7}O_3$ and $Cr_{1.1-}Fe_{0.9-}O_3$ would be classified as two different reagents – explaining why they are so rarely reported. In contrast, some very specific catalysts are very popular due to their commercial availability. Hydrogen and oxygen are ranked the 1st and 2nd reagents in continuous synthesis but only 4 and 78th overall. These two gasses are highly reactive and can pose important safety hazards, especially in processes operated at high pressures and temperatures (Gutmann et al., 2015). By operating such processes in continuous flow (micro) reactors, gas volumes can be reduced and equipment can be made safer for high-pressure operation (Gutmann et al., 2015; Kockmann et al., 2017). Organo-lithium reagents such as sec-butyl lithium and phenyl lithium are also found to have high EFs. Tert-butyl nitrite is almost 20 times as popular in continuous synthesis as overall. These reagents are again highly reactive and can cause safety hazards such as hotspots and runaway reactions in batch conditions. Due to the ability to achieve excellent heat transfer using continuous flow reactors with relatively small channel diameters, the exothermicity of reactions involving these

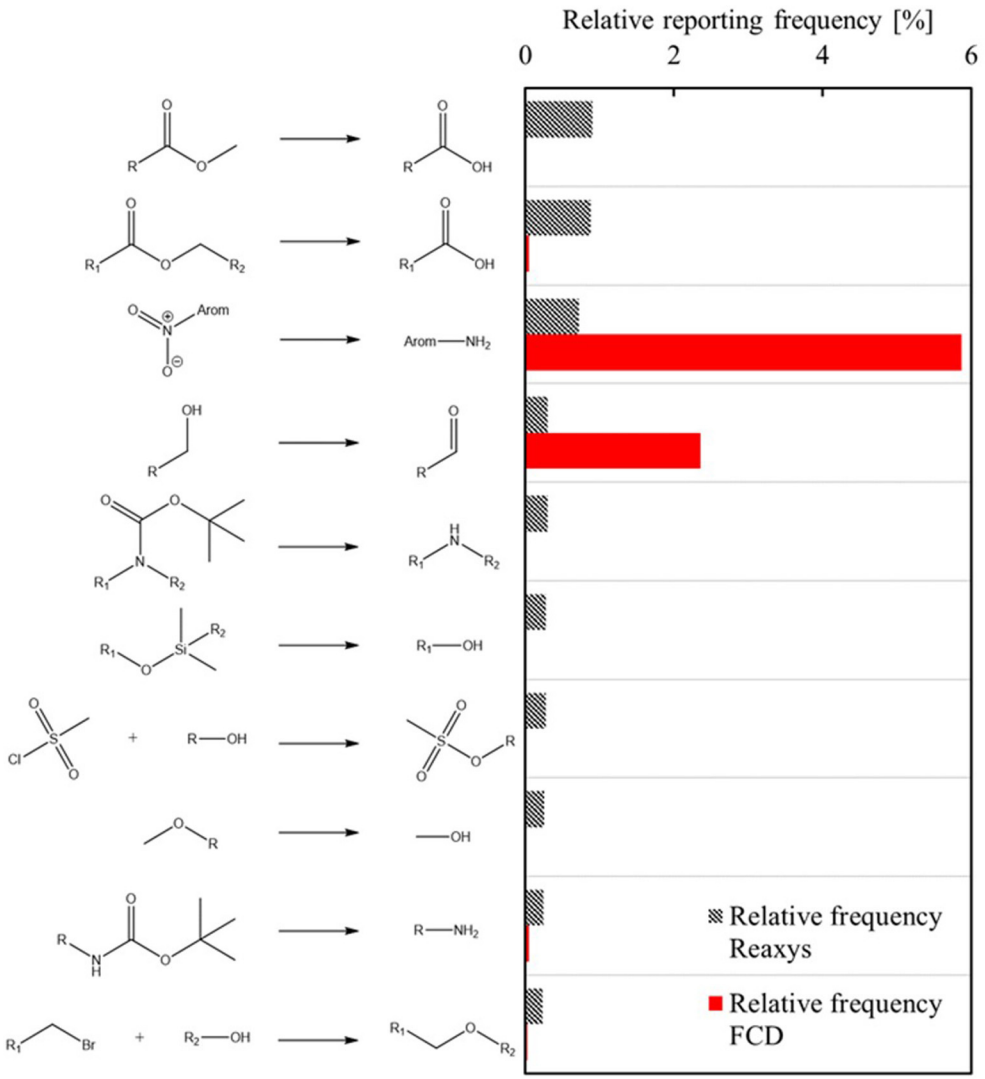


FIGURE 3 | Relative reporting frequency of the 10 overall most popular chemical templates and their respective relative frequencies in the flow reaction database. Numbers indicate popularity ranking within the flow reaction database. It is observed that protecting-group-reactions are well-represented overall, whereas they conclusively lack representation in the flow template set. Redox chemistry (reductions, oxidations) on the other hand, is well-represented in the flow template set.

reagents can be more easily controlled, again resulting in safer operation (Gutmann et al., 2015; Kockmann et al., 2017). A final stand-out reagent is rose bengal, which is 43 times more frequently reported in continuous synthesis than overall. Rose Bengal is a known good photocatalyst (Zhang et al., 2009, 2010). Other reagents that have been used as photocatalyst, such as titanium oxide (EF: 35) (Kitano et al., 2007) and zinc oxide (EF: 16) (Lee et al., 2016), also have high EFs. Photochemical reactions are a group of reactions that can greatly benefit from continuous operation as the higher surface-to-volume ratio and the small channel diameters of continuous flow reactors allow more light to penetrate into the reaction mixture (Gilmore and Seeberger, 2014). The high enrichment factors for typical photocatalysts indicates that photochemical syntheses are indeed disproportionately performed in continuous flow vs. batch.

The above analysis of the publication frequency of chemical reaction templates, solvents and reagents in continuous organic synthesis has shown that there are strong preferences toward specific types of chemistry and chemicals for continuous operation. In other words, both the FRD and FCD datasets are strongly biased. Several factors contribute to this bias. First of all, certain groups have built up a historical experience with certain reaction types, and stick to those when performing continuous syntheses, which can bias the performance of statistical models toward being more accurate for these reaction types than for others. Secondly, reactions that are easily carried out continuously are often preferred. Many issues encountered in continuous synthesis, such as the use of solids or multiphase flow can technically be resolved, though not always easily. Thirdly, given the somewhat larger scale of continuous flow

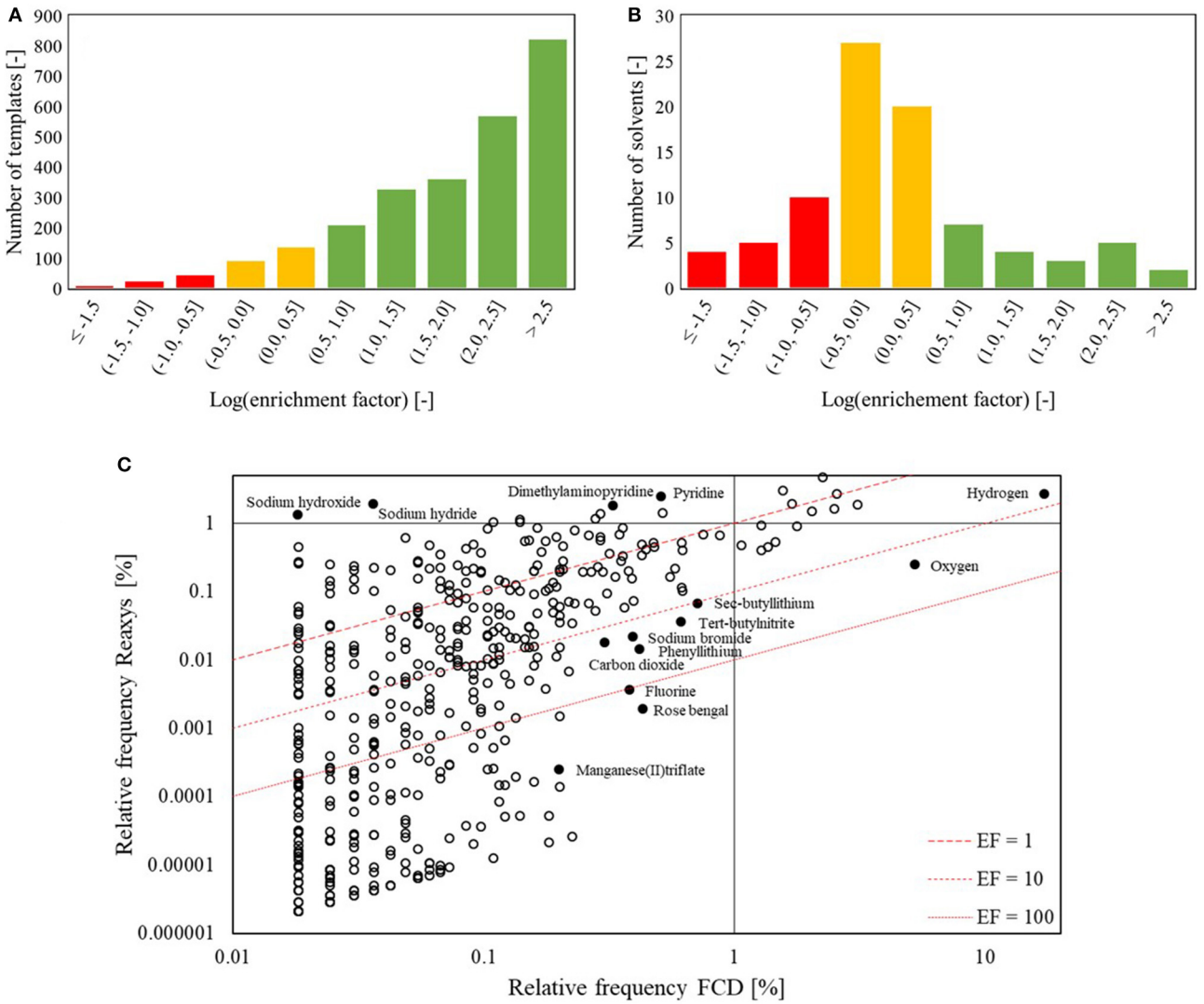


FIGURE 4 | Popularity enrichment of **(A)** the flow templates and **(B)** the reported solvents used in flow. This only includes the cases for which at least one example exists in flow, so cases with and EF of 0 are not included in the figure. *Enrichment factor* $= \frac{reported\ freqency\ FRD}{reported\ frequency\ Reaxys}$ Green: positive enrichment – Yellow: no or little enrichment – Red: negative enrichment. For the reagents **(C)**, the relative frequencies overall and in flow are plotted, along with lines of constant enrichment factors. Both the distributions of **(A)** reaction templates and **(C)** reagents/catalysts used in continuous flow show that many are over-represented relative to the broader reaction literature; **(B)** solvents used in continuous flow, in contrast, are not substantially different from those used overall.

units, cheaper chemicals are often preferred. These last two forms of bias are exactly what may be considered as *beneficial*, and are hence positive for the model performance. Analysis of the individual reagent enhancement factors has highlighted four such classes of reactions that benefit from continuous operation: (a) reactions involving gaseous reagents, (b) reactions that are highly exothermic and/or involve highly reactive reagents, (c) photochemical reactions, and (d) reactions that involve hazardous reagents. Reagents that have been reportedly used in these systems are found to have (very) high enrichment factors and their benefits from continuous flow are supported by literature. Given that for each of the highlighted classes, the benefits of continuous synthesis have been demonstrated (a variant of), the enrichment factor for reaction templates will be used in the next paragraph to assess which types of chemistry are likely to benefit from being carried out in continuous flow.

CHEMISTRY EVALUATION

Approach

As suggested in the introduction, the goal of the computer models developed here is not to definitively determine whether or not a given reaction is chemically suitable for continuous synthesis, but rather to recommend whether it might be beneficial to perform the given reaction in continuous flow, based on historical trends. Section Data Preparation indicates that certain types of chemistry

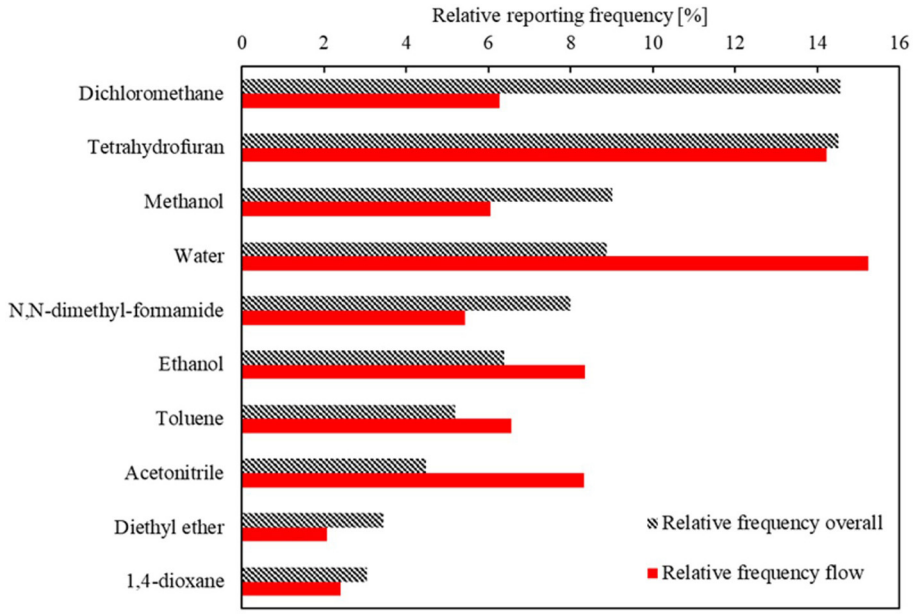


FIGURE 5 | Relative reporting frequency of the 10 overall most used solvents and their relative usage frequency in continuous synthesis.

are favored when the reactions are carried out in continuous flow (*cfr*. **Figure 3**).

Based on the analysis of the relative template popularity in literature, it can be inferred that templates with a high enrichment factor in the FRD are reactions that may benefit from continuous operation, while those with a low enrichment factor only benefit limitedly. The chemistry evaluation exploits this idea via an unsupervised clustering approach, based on Morgan fingerprints (Morgan, 1965). Each cluster is then assigned a score, based on the respective fractions of data in the FRD and data in the overall database that are assigned to that cluster. A Gaussian mixture model (GMM) or Gaussian clustering model based on Morgan fingerprints is preferred over directly using the templates and their corresponding enrichment factors, as reaction templates are potentially ambiguous (i.e., their definition is sensitive to the algorithmically-defined level of specificity or generality). The resulting clusters or categories can be seen as a generalized form of reaction templates. Additionally, a fingerprint can be constructed for any reaction, while – depending on the template database used – it is possible that no template is found for a new reaction that was not previously in the database.

The Morgan fingerprint of a reaction is first constructed by subtracting the reactant fingerprint from the product fingerprint (Schneider et al., 2015). This results in a fingerprint that contains information on what changes during the reaction and can therefore be interpreted as a fingerprint of the reaction template, without having to explicitly define the template. To improve the performance of the clustering approach, the fingerprints are first scaled to a standard deviation of 1 and centered to a zero mean. Next, the fingerprints are projected on the first 150 principal components (covering 70% of the total variance in the data). These 150-dimensional projections are used to construct a Gaussian clustering model with 500 categories – or as mentioned earlier, in this context, abstract reaction templates. The choice of the number of components is elucidated in the **Supporting Information**. The clustering model is built on a dataset that consists of a selection of 119,354 reactions drawn from all templates that have more than 50 reaction examples (a total of 29,948 templates). For each template, 4 examples are randomly chosen. This dataset is further referred to as ORD_{sel}. Only a limited selection of reactions is made, as using all available reactions in the ORD would result in an excessive computational cost for training the clustering model. During the optimization of the clustering model, no information about whether or not a reaction has been successfully performed in continuous flow is used. This information is only introduced during the final analysis step, in which the scores for each category of the GMM are determined. To do so, the selected reactions from the overall reaction database (ORD_{sel}) and the FRD reactions are independently categorized and the relative number of data points in each category is determined. $f_{i,FRD}$ and $f_{i,ORD_{sel}}$ are the fractions of the FRD and ORD_{sel} that are assigned to category *i*. The score for each category can then be calculated according to equation 1. A category score of zero indicates that there is no published evidence that reactions in that category benefit from continuous synthesis. The closer the score gets to one, the more evidence is available that such reactions can benefit from flow. The quantitative assessment of whether or not continuous synthesis is beneficial for a given reaction is then calculated via equation 2, where $p_i\,(reaction)$ is the probability that the reaction belongs to category *i*, as determined by the Gaussian clustering model. The meaning of the scores is similar to that of the category scores – 0 indicates that the reaction has no benefit

from flow based on published data, whereas values close to 1 indicate the opposite.

$$score_i = \frac{f_{i,FRD}}{f_{i,FRD} + f_{i,ORD_{sel}}} \quad (1)$$

$$score(reaction) = \sum_{i=0}^{\#clusters} p_i(reaction) \cdot score_i \quad (2)$$

Results

The gaussian clustering model distributes the template-based reactions fairly evenly across the 500 categories. As expected, the different categories are observed to contain reactions that are very similar in nature, but do not necessarily follow a single reaction template. This is illustrated by **Figure 6**, which shows

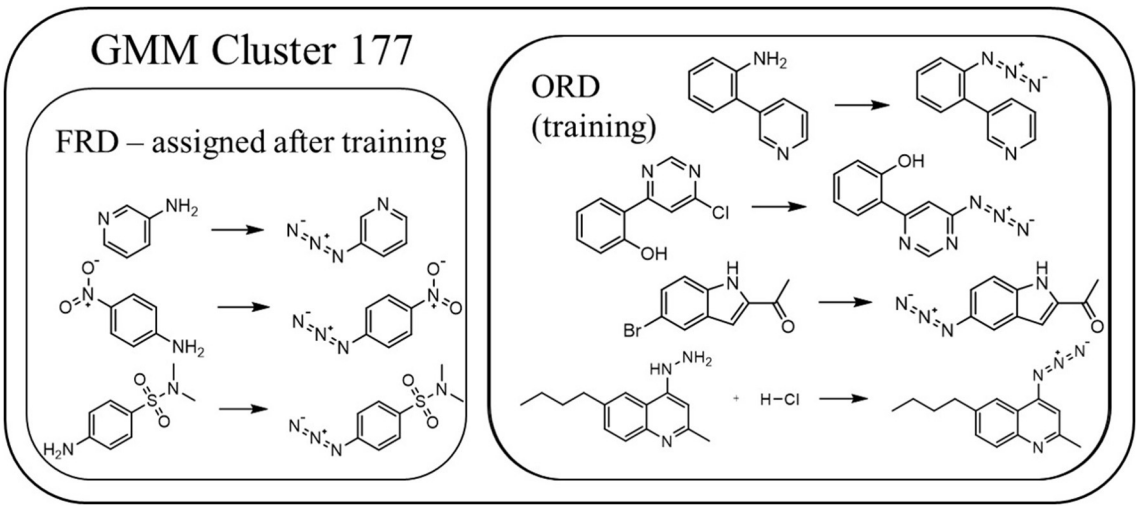


FIGURE 6 | Several examples of reactions categorized into category number 177. Only the reactions from the overall reactions database (ORD) were used to train the GMM. The reactions from the flow reactions database (FRD) are only assigned to the categories a posteriori.

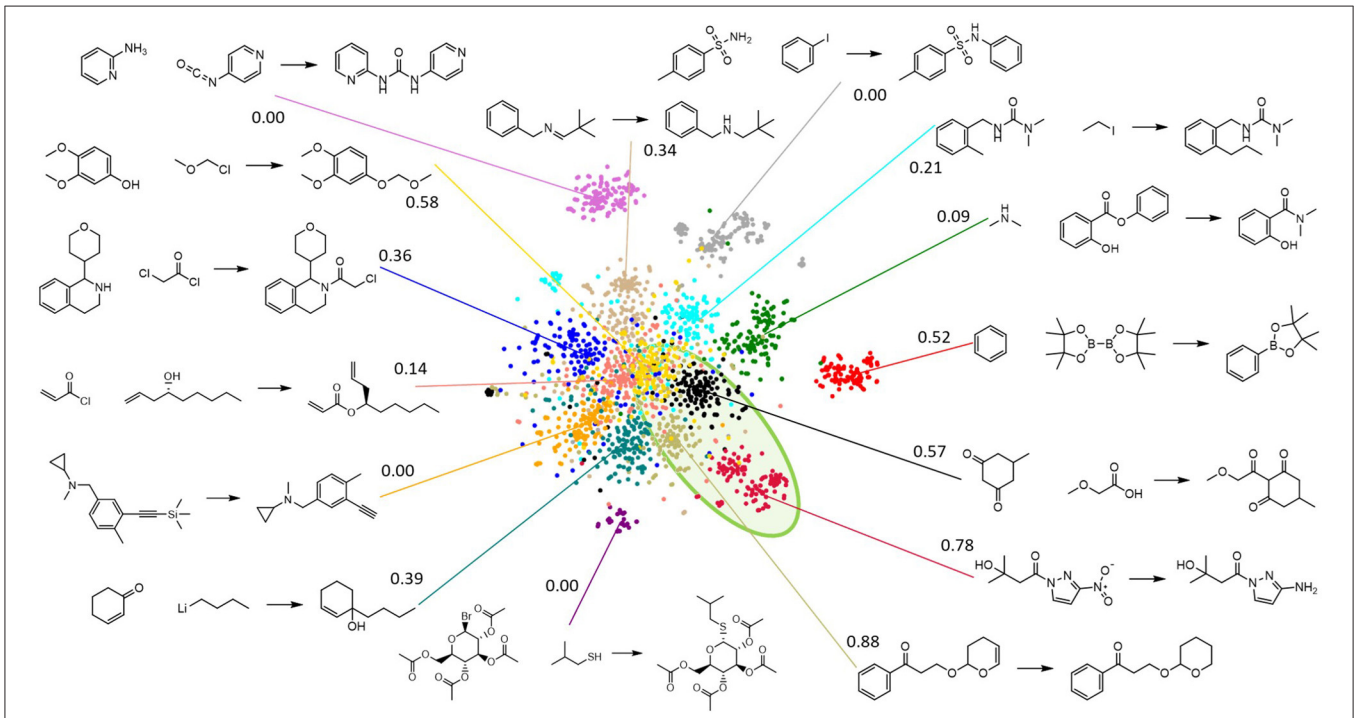


FIGURE 7 | Visualization of a fraction of the reaction space via t-SNE, where reactions are colored by the categories assigned via the Gaussian mixture model (GMM) for a subset of 15 categories (compared to the full model's 500). A representative example for each GMM category is shown, along with the score of each category. All categories (except the red one) with a score (cfr. equation 2) that is >0.5 are grouped closely together in the same highlighted green region of the reaction space.

several specific reactions that were assigned to the same category. In all these reactions, a functional group is converted to an azide, though the exact mechanism is clearly different. In fact, some reactions that follow the same template, but that have highly dissimilar reactants and products, are assigned to different categories. **Figure 7** visualizes the reaction space and the result of the GMM for a limited selection of reactions, drawn from 15 of the categories identified in the GMM. The high-dimensional (150 PCA components) reaction space is visualized in 2D via t-SNE (van der Maaten and Hinton, 2008) and each reaction is colored by the category it has been assigned to in the GMM. A number of the categories appear to form a larger cluster in the central region of the reaction space. While there are distinct differences between these clusters in types of reactions, reactants, and products, the overall reactions are observed to be quite similar and generally involve carboxylic substitutions or additions. Around this central agglomeration, there are several, more isolated reaction clusters. From the examples it is immediately clear that these categories describe very different types of chemistry compared to those in the central categories. Additionally, the reaction clusters with high scores are grouped closely together in a specific area of the reaction space.

As mentioned in the previous section, once the gaussian mixture model has been constructed based on the reactions from the ORD_{sel}, the reactions from the FRD are categorized as well. The reactions from the FRD are distributed across only 241 of the 500 categories, which is in agreement with the presumption of the used approach – continuous synthesis focusses on certain chemistries. Five reactions from the FRD are assigned to the example category 177. This corresponds to 0.07% of the FRD. Based on equation 1 and given that 114 reactions or 0.096% of the ORD_{sel} are assigned to category 177, it receives a score of 0.426. Any score >0 indicates that there is some benefit toward running reactions in that category in continuous flow, so the score of 0.426 indicates that the reactions in this category could be considered for flow, but that there are potentially stronger candidates as well. In section Data Preparation, azidation reactions were mentioned as reaction types with a high enrichment factor for which continuous synthesis provided improved process safety. **Figure 8** illustrates the distribution of the scores for the other categories. Very few categories group reactions with chemistry that is strongly over-represented in the FRD and therefore only a few types of chemistry are predicted to strongly benefit from continuous synthesis. In **Figure 7**, only a limited number of categories are seen to have a zero score – from **Figure 8**, however, it is clear that they are underrepresented in our selection with respect to their overall prevalence.

As mentioned before, the presented clustering and scoring method is essentially just another way of approaching the template enrichment discussed under Data Preparation. The major difference is that it does not depend on pre-defined reaction templates, but on a statistical analysis of the substructural changes that take place during a reaction as represented by a reaction fingerprint. Therefore, it can be used as a user-independent, fast screening tool to quantitatively assess whether the chemistry of a given reaction is similar to that of reactions which have been reported in continuous flow relatively more frequently than overall. This score can be used to assess whether a given reaction is likely to chemically benefit from continuous synthesis.

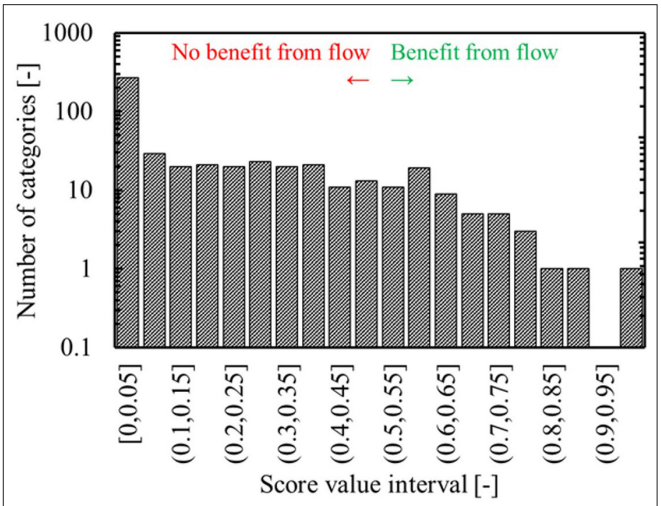


FIGURE 8 | Score distribution for the categories found by the gaussian mixture model – more than 50% of the categories have no reactions from the FRD assigned to them, resulting in a score of 0, indicating no published evidence of benefit from continuous synthesis.

REACTION COMPONENTS EVALUATION

Approach

To assess how beneficial continuous synthesis can be for a reaction, given a certain set of reaction components, an artificial neural network with an architecture as shown in **Figure 9A**, is trained on the FCD. Ideally, the model would evaluate the reaction *conditions* (reagents, solvents, residence times, temperature, etc.) instead of just the reaction *components* (reagents and solvents). However, the number of reactions in the Reaxys database for which the full set of reaction conditions is available, is extremely limited, and would not allow for the construction of a meaningful statistical model. A set of reaction components is assumed to benefit from continuous synthesis if the corresponding entry in the FCD has the "InFlow" label (as shown in **Figure 2**). These cases are further referred to as *benefitting cases*. If it carries the "NoFlow" label, no benefits are assumed. They are further described as *non-benefitting cases*. All chemicals in the input to the neural network – reactants, products, solvents and reagents – are represented by their 256-bit-long Morgan fingerprints (Morgan, 1965; Rogers and Hahn, 2010). Further details on the model architecture and input representation can be found in section 2.4.2.1 of the **Supporting Information**. The model is trained using 10-fold cross-validation and a 90-10 random training-test split of the reactions. During training, the model performance is monitored via the (average) validation area under the precision-recall curve (PRC) over the 10 cross-validation models. Early stopping is applied when this area attains a minimum, in order to prevent

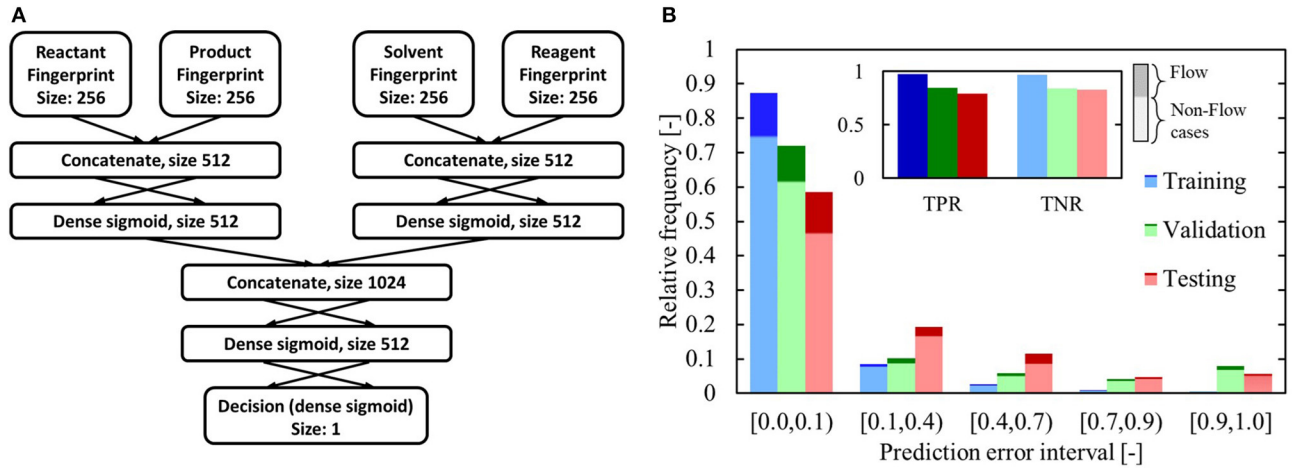


FIGURE 9 | (A) Classification network architecture. The input fingerprints are obtained via the Morgan algorithm. **(B)** Classification network performance on training, validation and test sets. The insert displays the true positive rate (TPR) and true negative rate (TNR) at a decision threshold of 0.5. The prediction error is the difference between the predicted score and the actual score (a benefitting case should be scored as 1, a non-benefitting one as 0). The darker shading indicates which fraction of the total cases in that category was labeled as benefitting, while the lighter shading indicates the fraction that was labeled as non-benefitting.

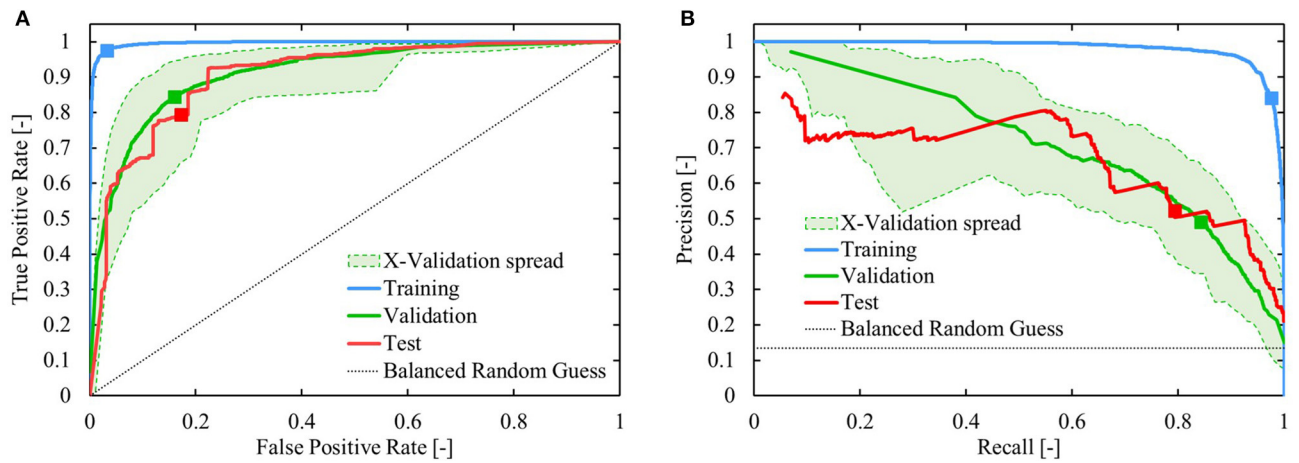


FIGURE 10 | Visualization of the performance of the reaction components evaluation model. **(A)** Receiver operating characteristic and **(B)** Precision-recall curve. The green area represents the envelope of the cross-validation performance. The "balanced randomized guess" in both **(A,B)** predicts 6.5 times more values in (0,0.5) than in (0.5,1). The squares indicate the position on the curves at a decision threshold of 0.5.

overtraining of the model. In order to account for the bias toward non-benefitting cases in the dataset – there are 6.457 non-benefitting cases for every benefitting case – all benefitting cases are given a weight of 6.457 during training.

Results

The performance of the trained network is presented in **Figure 9B**. Applying the classifier, using a decision threshold of 0.5, results in a correct prediction on 75% of the test set. Both the true positive rate (TPR) and true negative rate (TNR) are high for the test set, and comparable to the values for the validation dataset. The high TPR indicates that the model successfully identifies flow-benefitting cases, whereas the high TNR shows that the model is specific and can identify non-flow-benefitting cases. In **Figure 9B**, the prediction error interval (0.0, 0.1) collects all predictions that are >0.9 for the benefitting cases and <0.1 for the non-benefitting cases. The overall validation and testing performance of the model is quite similar, though there is a noticeable drop in performance compared to the training performance. Nonetheless, **Figure 10** shows that the model performs significantly better than a "balanced" random guess, where for every guess >0.5, 6.457 guesses are made below 0.5 to account for the sample balance in the dataset. In **Figure 10B**, the precision and corresponding recall at a threshold of 0.5 are singled out. For the test set, the precision is 0.52, whereas the recall is 0.8. This implies that a large number of non-benefitting cases is predicted to be benefitting. It should be

noted that the value of the threshold is subjective and tunable. For example, with a decision threshold of 0.377, the precision on the test set is still 0.5, though the recall has increased to 0.93.

One reason for the overall low precision is the imperfect assumption that reaction components not explicitly labeled as flow in Reaxys do not benefit from continuous flow synthesis. This is a very strict criterion, and it is quite conceivable that many of those cases *could* benefit from or be compatible with continuous flow. Next, an example is presented of a case where the model predicts a case, labeled as non-benefitting, to be benefitting. Though arguably, the prediction is understandable. After that, a number of additional cases – both benefitting and non-benefitting – are presented where the model makes the correct prediction.

The first example is the reaction of 4-(tetrahydropyran-2-yloxy)butan-1-ol to 2-(4-fluoro-butoxy)-tetrahydro-2H-pyran (**Figure 11A**). Of its three reported cases, two are labeled as flow (Baumann et al., 2008), while one is not (Akihiro et al., 2006). The flow cases were both performed using the same components, namely diethylamino-sulfur trifluoride as reagent and dichloromethane as solvent. The third case was performed with fluorosulfonyl fluoride, triethylamine tris(hydrogen fluoride) and triethylamine as reagents and acetonitrile as solvent. The two flow-benefitting cases are attributed a score of 0.996, while the case labeled as non-benefitting is attributed a score of 0.872. All three cases are therefore assessed as likely to benefit from continuous synthesis. In this case, solely based on the similarity and properties of the used solvents and reagents, given the reported success of the first two cases, it seems plausible that the second reaction can benefit as well. In all cases, the reagents contain ethylated amines and fluorinated sulfur compounds. Especially the latter present significant safety hazards, and as stated previously, reactions involving hazardous reagents or solvents are typical candidates to benefit from continuous syntheses. The literature provides some additional evidence to this plausibility. Fluorosulfonyl fluoride in combination with triethyl amine has been used as reagent in the Beckman rearrangement of ketoximines (Zhang et al., 2019). While the reported case does not state that it was carried out in flow, there are several examples of acid-catalyzed Beckman rearrangements that have been performed continuously (Curtin et al., 1993; Ko et al., 2000; Botella et al., 2007). Under the made assumption, the fact that other systems using similar reagents can benefit from continuous synthesis, is an indication that this reaction can also benefit. Further supportive evidence is found in the successful use in continuous flow of arylsulfonyl chlorides with acetonitrile as solvent (Malet-Sanz et al., 2010). Altogether, this seemingly erroneous assessment is understandable at least, and at best, the model identified a case that has been labeled as non-benefitting, but could, in reality, benefit from a continuous synthesis.

A second example is the synthesis of the fluorinated aziridine shown in **Figure 11B** (Baumann and Baxendale, 2016). This case is correctly predicted by the model, and provides an additional explanation for the erroneous prediction for the non-benefitting-labeled case in **Figure 11A**. Though the reactions themselves are quite different, the components used are very similar. Both cases use acetonitrile as solvent and triethylamine as one of the reagents. While **Figure 11B** uses methyl sulfonyl chloride where **Figure 11A** uses fluoro sulfonyl fluoride, both reagents are halogenated sulfonyl compounds.

Figure 11C (Josyula and Mitesh, 2014) and **Figure 11D** (Tsai et al., 2013; Han et al., 2015; Mallia et al., 2016) shows four more cases the model conclusively labeled correctly. The performance for the cases in **Figure 11E** varies again, with the first two being labeled correctly (Safari and Javadian, 2013; Monteiro et al., 2016). The third case is rather indecisive (Safari et al., 2013), whereas the last case is predicted incorrectly (Prevet et al., 2016). For these last two cases, the combination of polar and ionic reactants with polar solvents that are similar to those used in the benefitting case, is an explanation for the poor performance. Additionally, based on the solvents used in the benefitting case, there is no apparent reason why these two sets of components would benefit less from continuous synthesis.

As mentioned earlier, residence times and temperatures were not considered in the analysis as too few database entries are recorded with both. However, when investigating those cases for which reaction times are available, an interesting observation is made. For the reaction components that are labeled as benefitting, the incorrect predictions tend only to be made for cases that report short reaction times. Reactions that require long residence times are typically unpractical in continuous operation, which immediately limits the potential benefit those reactions can have from flow syntheses. The fact that the model predictions are in agreement with this observation is further evidence to the good performance of the model. This is illustrated in **Figure 12**, which shows a clearly decreasing trend of the maximal model prediction with increasing reported reaction time.

Overall, it can be concluded that given its assumptions, the model makes a reasonable assessment of which type of reaction components tend to benefit from continuous flow and which ones do not. For cases that appear to be miscategorized, the incorrect prediction can often be justified, based on similar cases present in the data and overall trends. Additionally, the assumption on which the model is built is quite weak and could label certain reaction components as non-benefitting from continuous synthesis, while they could benefit in reality. It is also important to keep in mind that this model has a descriptive nature, and does not explicitly answer the question of *why* a certain case does not benefit from continuous synthesis, which is important during the actual development of a new synthesis. However, currently, the main role of CASP is not to independently develop new syntheses, rather to support scientists by narrowing down the search space to a number of possibly routes, that are considered to be promising. The presented model can be used in a similar way, by flagging reactions that, based on historical literature data, are considered to benefit from continuously operated syntheses. After narrowing down the search synthesis search space to a number of plausible candidates, more detailed evaluations using detailed kinetics (Yuan et al., 2018; Koch et al., 2019; Konze et al., 2019; Unsleber and Reiher, 2020) could further evaluate the different possibilities to maximize the likelihood of success

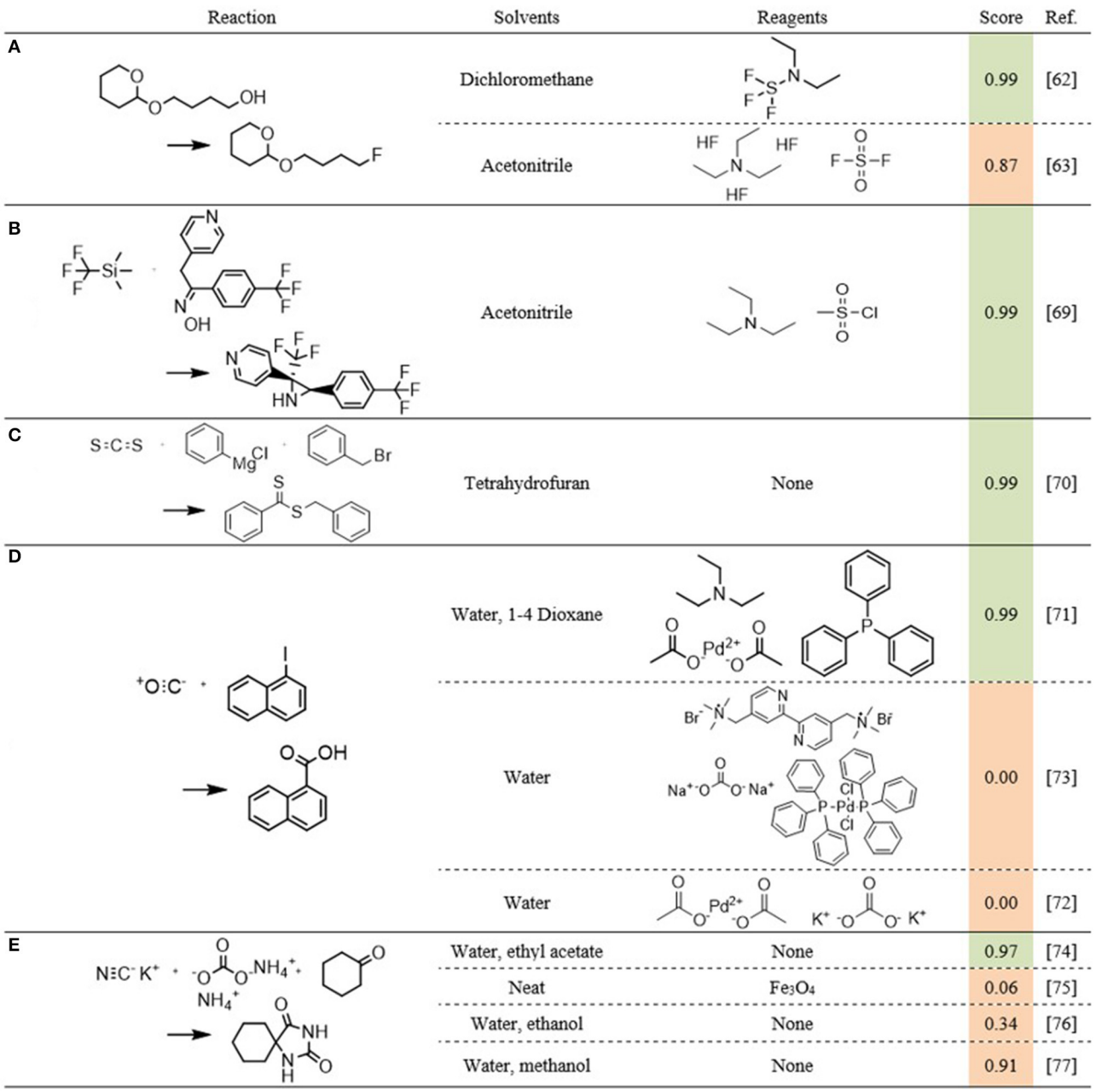


FIGURE 11 | Five reaction examples **(A-E)** for the reaction components scoring model. The score column is shaded by the case label: green is labeled as benefitting, orange is labeled as non-benefitting.

of the suggested syntheses. A possible improvement of the model to allow for a more explorative interpretation of its predictions, would be to include methods for the prediction of individual prediction confidence intervals. High epistemic uncertainties could be used as indicators for data scarcity on the queried reaction (Scalia et al., 2020), pointing it out as a potentially interesting case for an experimental study. The current dataset can then be augmented with the newly acquired information, improving the model performance. Similar active learning approaches have proven to be successful in other fields (Naik et al., 2013; Melnikov et al., 2018; Konze et al., 2019).

PROPOSITION OF ALTERNATIVE REACTION COMPONENTS

Approach

The method described in section Reaction Components Evaluation predicts how likely it is that a reaction benefits from continuous flow, given certain reaction components. For those

cases that this is considered highly improbable, one might want a CASP program to propose alternate components. A second neural network is trained to predict those components that are most likely to benefit from flow synthesis, for a given reaction. The assignment of the reaction components is done by iterating through the FRD, described previously. For each reaction, the components of the highest-yielding "InFlow"-labeled instance are treated as the optimal alternate components. The general architecture of the network drawn in **Figure 13**, is chosen along the same reasoning as the network described by Gao et al. (2018) for the prediction of reaction conditions in general. A one-hot-encoding method is used for the output, though in contrast to Gao's work, no minimal frequency is utilized. Because only 82 different solvents are reported in the flow conditions database and 980 reagents, it is not considered necessary to further reduce the variation in chemicals. As already mentioned previously, very few cases fully describe the reaction conditions, often leaving out one or more of e.g., reaction time, temperature and pressure. Including these parameters in the network output would significantly reduce the already limited amount of data the network could be trained on. Therefore, the prediction is again limited to the chemical components, *i.e.,* reagents and solvents, where reagents and catalysts are again grouped into a single class. Furthermore, in order to limit the complexity of the model, only two reagents and two solvents are allowed. While the selectable components are limited to those reported in the dataset, biased by the preferences of the scientists performing the experiments, they do represent the components that generally allow for the most straightforward use in continuous flow.

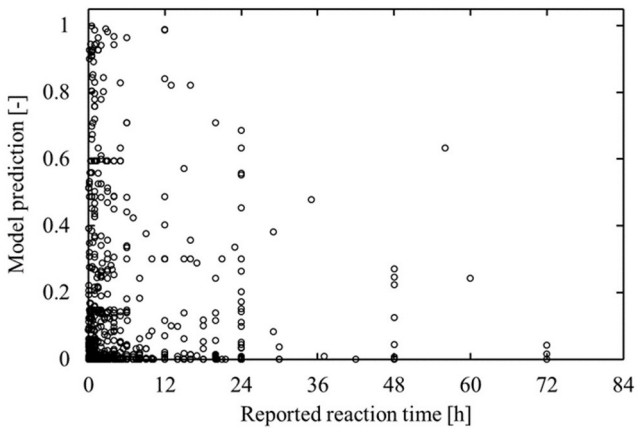


FIGURE 12 | Model prediction as function of the reported reaction time for non-flow-benefitting labeled cases in the test set. Although the model is not made aware of reaction times, it tends to assign reactions with high reported reaction times a lower flow-benefitting score.

Sigmoid activation functions are used in the hidden layers, while softmax activation is used in the output layers. A large number of reactions in the FRD are reported with no reagents or solvents at all, or at most one. As a result, the input strongly biases the model toward simply predicting "no reagents" and "no solvents" for any case, especially for the second reagent/solvent. To counteract this, three weights are used for the input. It is observed that 51.7% of the cases only report a single solvent or reagent. Therefore, such cases are assigned a weight factor of 0.517. Similarly, 22.6% of the cases report no solvents or reagents at all. Analogously, they are weighted by a factor 0.226. All other cases are attributed a weight of 1. The same 10-fold cross-validation approach as used for the components evaluation network is applied, with a 90-10 random training-test ratio. Early stopping is applied using the product of the accuracies on the four outputs as an overall accuracy metric.

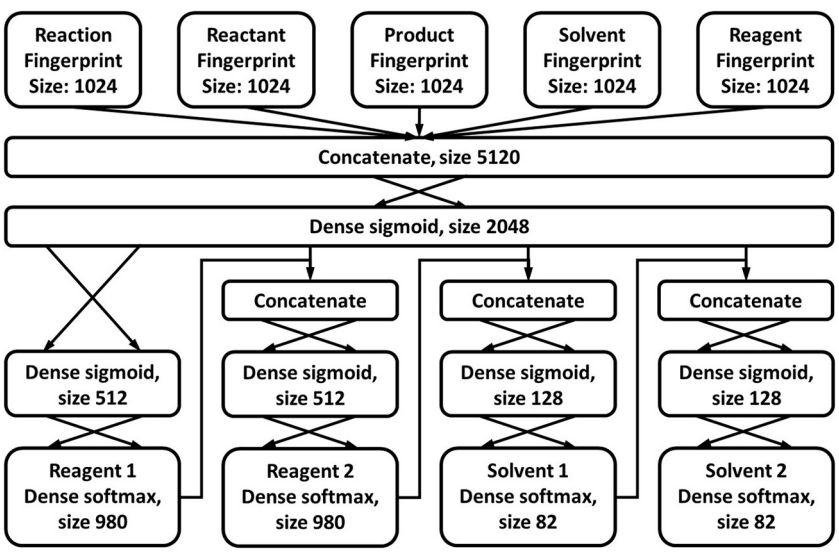


FIGURE 13 | Network architecture for suggesting alternative reaction components.

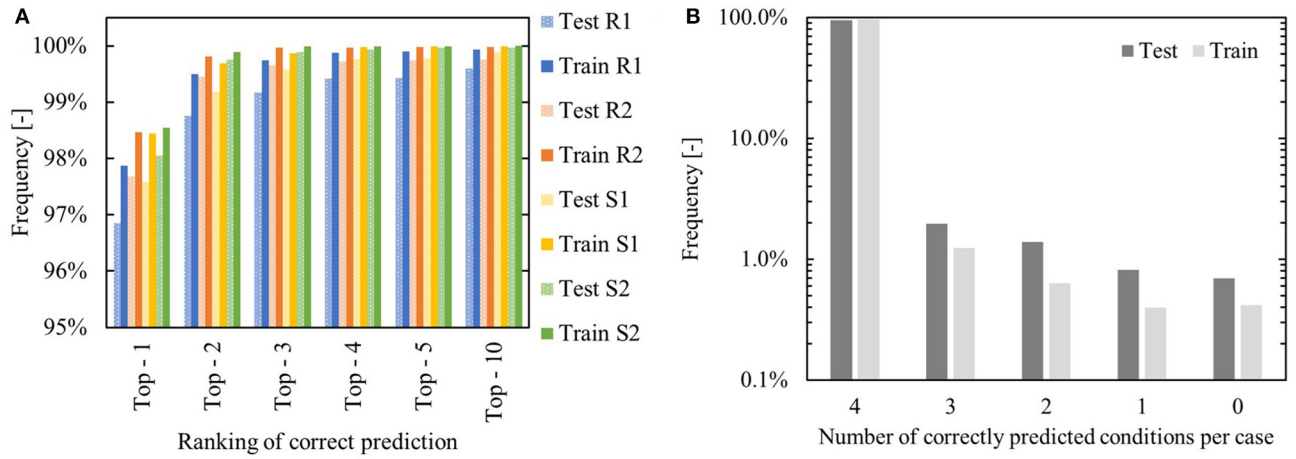


FIGURE 14 | Network performance for suggesting alternative reaction components. **(A)** "Top – x" performance – minimal rank required to encounter the true flow-compatible alternative. R,S refer to reagent and solvent. **(B)** Number of components that are predicted at rank 1 on a per-case basis.

Results

Figure 14 summarizes the performance of the network that suggests an alternative set of flow-benefitting reaction components. **Figure 14A** shows that in over 96% of test cases, the true alternative components are (individually) attributed the highest probability. In general, prediction of the solvents is found to be more accurate than prediction of the reagents, due to the fact that there are fewer options for the model to choose from. In almost 95% of the test reactions all four of the alternative components are predicted as the most likely combined alternative. The model performs only slightly better on the combined training and validation sets – in 97%, all four components are predicted as top-1.

In what follows, a number of specific cases of both successful and unsuccessful predictions will be discussed to illustrate the performance of the neural network.

The first one is the reaction of amplex red (10-acetyl-3,7-dihydroxyphenoxazine) to resorufin using horseradish peroxidase, hydrogen peroxide and a tris-HCl buffer as reagents and dimethyl sulfoxide as solvent [**Figure 15A** (Hellman et al., 2007)]. Given that this reaction has been performed in continuous flow, the model should preserve these components. As the model is limited to predicting two reagents and solvents, the last reported reagent (in this case hydrogen peroxide) is left out. The model predicts the following reagents: horseradish peroxidase (HRP)-APSN-APTES and hydrogen peroxide. As solvents it predicts an aqueous phosphate buffer and dimethyl sulfoxide (in that order). It is quite clear that the predicted reaction components are very similar to the ones that should have been predicted. The enzyme is a variation on the correct enzyme, reported by Seia et al. (2014). The acidic buffer is predicted as solvent, while it was listed as reagent in the reference case. The dimethyl sulfoxide is correctly predicted, but not as first solvent, resulting in it being classified as incorrect.

A second example is the oxidation of 4-methyl-1-indanol to 4-methyl-1-indanone [**Figure 15B** (Chorghade et al., 2013)]. This reaction has been performed in continuous flow using acetone as an (rather unconventional) oxidant and toluene as solvent. The model predicts four different components: sulfuric acid and hydrogen peroxide as reagents and 1,4-dioxane and water as solvents. As oxidizing components are required for the reaction, the proposed components chemically make sense – the combination of sulfuric acid and hydrogen peroxide is a well-known strong oxidizing agent, much stronger than acetone, reported by the "InFlow"-labeled case. As mentioned previously, slow reaction rates can negatively impact the possible benefit of continuous synthesis. Using a stronger oxidizing agent than reported for the "InFlow"-labeled case will hence only increase the potential of that reaction. Both 1-indanol[1] and 1-indanone[2] are soluble in water, so it is expected that the methylated derivatives will be as well, indicating that practical challenges such as component solubility are not expected to negatively impact the potential benefit of flow. The model makes its predictions with relatively high scores: 0.66, 0.8, 0.96, and 0.99, respectively. For all four chemicals however, the second ranked suggestion is the one that is listed as the correct suggestion.

The examples in **Figure 15C** (Brooke et al., 1961; McPake et al., 2012), **Figure 15D** (Chadwick et al., 2010; Pieri et al., 2014; Jong and Bradley, 2015) and **Figure 15E** (Jirkovsky, 1974; Pathak et al., 2014; OBrien and Cooper, 2016) show how the model correctly preserves the components for the flow-benefitting cases, while it adjusts the components for the non-benefitting cases.

Based on the overall performance and the cases shown above, it can be concluded that the alternative reaction components suggestions are close to the ones that were reported for the

[1]Chemical Book 1-Indanol. Available online at: https://www.chemicalbook.com/ChemicalProductProperty_EN_CB2685759.htm (accessed October 16, 2019).
[2]Chemical Book 1-Indanone. Available online at: https://www.chemicalbook.com/ProductChemicalPropertiesCB0384120_EN.htm (accessed October 16, 2019).

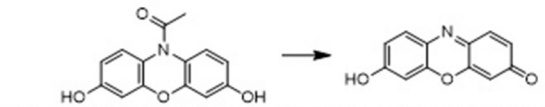
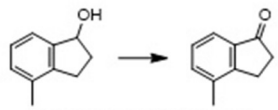
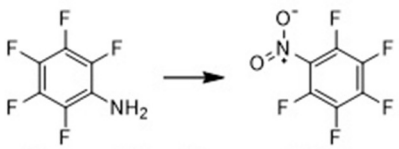
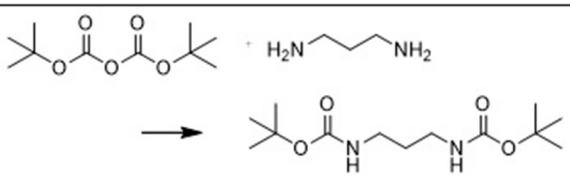
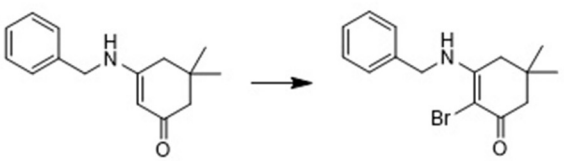
	Reaction		Input	Output	Ref.
A	 In dimethyl sulfoxide, with horseradish peroxidase, tris-HCl buffer and hydrogen peroxide	R1	horseradish peroxidase	horseradish peroxidase (HRP)-APSN-APTES	[78]
		R2	tris-HCl buffer	hydrogen peroxide	
		S1	dimethyl sulfoxide	aqueous phosphate buffer	
		S2	-	dimethyl sulfoxide	
B	 In toluene, with acetone	R1	Acetone	H_2SO_4	[80]
		R2	-	Hydrogen peroxide	
		S1	Toluene	1,4-Dioxane	
		S2	-	Water	
C	 In acetonitrile, with water and fluorine	R1	Water	Water	[83]
		R2	Fluorine	Fluorine	
		S1	Acetonitrile	Acetonitrile	
		S2	-	-	
		R1	Hydrogen peroxide	Water	[84]
		R2	Trifluoroacetic anhydride	Fluorine	
		S1	Dichloromethane	Acetonitrile	
		S2	-	-	
D	 In methanol	R1	-	-	[85]
		S1	chloroform	methanol	
		R1	-	-	[86]
		S1	1,4-dioxane	methanol	
		R1	-	-	[87]
		S1	methanol	methanol	
E	 In dichloromethane and water, with N-bromosuccinimide	R1	N-bromosuccinimide	N-bromosuccinimide	[88]
		R2	-	-	
		S1	dichloromethane	dichloromethane	
		S2	water	water	
		R1	bromomalononitrile	N-bromosuccinimide	[89]
		R2	-	-	
		S1	N,N-dimethylformamide	dichloromethane	
		S2	-	water	
		R1	N-bromosuccinimide	N-bromosuccinimide	[90]
		R2	-	-	
		S1	methanol	dichloromethane	
		S2	water	water	

FIGURE 15 | Five example reactions **(A–E)** from the test set for the active suggestion of flow-benefitting reaction components. The components listed in the *Reaction* column are those that were selected as the optimal flow components. The *Input* and *Output* columns list the chemicals that were, respectively, used as input to the model and obtained as suggestion from the model.

associated continuously operated reaction. Even for the cases with the poorest performance, the suggestions are still chemically acceptable, but experimental testing is required to ascertain this. Similarly to the previous models, this model does not explain why certain components are suggested. However, given that the output of the model is a ranked list of possible reaction components, it can still be of value for narrowing down the search space of solvents and reagents to a limited number, which have proven their usefulness in past continuous flow studies.

CONCLUSIONS

A data-driven, deep learning method for guiding retrosynthetic software toward syntheses that benefit from being executed in a continuous flow reactor has been presented. All data has

been sourced from Reaxys. This extracted flow chemistry data is naturally biased toward well-known reactions that can be performed relatively easily and cheaply in continuous flow. This bias is, however, aligned with what the method is learning to detect.

In the context of continuous organic synthesis, three questions have been identified. A first is whether or not a given reaction might chemically benefit from being performed in continuous synthesis. A second question is whether a given set of reaction components would allow the reaction to benefit from continuous operation. A final question is, if a proposed set of reaction components appears to offer no or little benefit, what are the reaction components that are likely to provide the greatest benefit. For each of these three questions, a descriptive, machine learning-based model has been developed. This implies that neither of the models give insights into why a reaction might or might not benefit from continuous synthesis. However, by providing a fast, quantitative assessment of its potential in continuous synthesis, large sets of reactions can be filtered down to a more manageable set. This set can subsequently be analyzed in more detail, e.g., using detailed computational methods to determine reaction rates and solubilities, in order to arrive to a definitive assessment.

To assess the potential benefit of continuous flow for a given reaction, a Gaussian mixture model has been developed, based on the difference between the Morgan fingerprints of the products and reactants. This unsupervised learning method identifies clusters of reaction types. Only a limited number of clusters contain reactions that have been performed in continuous flow. This is in line with the observation that only a very limited number of reaction templates have been reported in continuous synthesis. The ratio between the relative amount of flow and non-flow cases categorized into each cluster is used as measure for how likely it is that a reaction in that cluster will benefit from continuous synthesis.

For the two other questions, artificial neural network models have been developed. The first ANN classifies the combination of reaction and reaction components as benefitting or non-benefitting on a continuous scale. It successfully classifies 75% of all data correctly. However, the precision of the model is relatively low – only half of the cases predicted as benefitting were also explicitly labeled as benefitting from flow. As the method has been developed to direct computer-aided synthesis planners toward continuous syntheses where desirable, the fact that several assumedly non-benefitting reactions are still labeled as benefitting is not considered as problematic. Additionally, all reactions were labeled based on the assumption that if a reaction is not explicitly annotated as "performed in continuous flow" in Reaxys, it does not benefit from continuous synthesis. This is a very strict criterion, which likely led to several reactions being labeled incorrectly with respect to reality. In this regard, the low precision is an indication that the model actually learns which types of reactions and reaction components tend to benefit from flow. The second ANN predicts a combination of two reagents and two solvents from all reagents and solvents reported in successful continuous syntheses for those cases which received a low score by the first ANN. The second trained neural network achieves a combined top-1 accuracy of 95% on the test data. Even here, several of the incorrect predictions are chemically still acceptable and or components were predicted in a different order than listed in the dataset.

Combining the three models presented in this work will provide a preliminary assessment on whether or not it is useful to perform a reaction in continuous flow, and which alternative components could be evaluated. Based on that assessment, computer-aided synthesis planners can be directed toward suggesting more efficient, continuous syntheses wherever possible. In potential follow-up studies, the models could be extended with methods to predict individual uncertainties in order to suggest interesting reaction classes for which the current data do not allow conclusive evaluations. Expanding the dataset with new information on these reaction classes, will further improve the model performance. We also expect that with sufficient annotations, similar data-driven analyses might enable the biasing of CASP programs toward environmentally-friendly reactions or reactions that are conducive to scale-up.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: Data originates from Reaxys, which is a licensed database, curated by Elsevier. Our license does not allow disclosure of this data, we have provided the ids of the relevant data so researchers with their own license can access the data. Requests to access these datasets should be directed to William H. Green, whgreen@mit.edu or https://www.reaxys.com.

AUTHOR CONTRIBUTIONS

PP, CC, and HG conceived the idea and developed the theoretical approach. PP performed the simulations and optimization of the approach. PP, FV, and CC sourced the used datasets from Reaxys. PP, CC, and MD postprocessed the results and worked on the visual presentation of the data. CS provided counseling and discussions on the practical continuous synthesis related topics of the work. WG provided counseling during the conception of the project idea and supervised the project. PP wrote the manuscript, with counseling and assistance from KV. All authors discussed the results and thoroughly revised the manuscript.

FUNDING

PP acknowledges financial support from a doctoral fellowship from the Research Foundation – Flanders (FWO) and funding by the Commission for Educational Exchange between the United States of America, Belgium and Luxembourg under the Fulbright Program. CC received funding from the NSF Graduate Research Fellowship Program under Grant No. 1122374. This work was supported by the DARPA Make-It program under contract ARO W911NF-16-2-0023. The authors would like to acknowledge the financial support from the European Research Council under the European Union's Horizon 2020 research and innovation program (ERC grant agreement no. 818607).

ACKNOWLEDGMENTS

We thank Elsevier for access to the Reaxys API and their invaluable curation of the reaction data supporting this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fceng.2020.00005/full#supplementary-material

REFERENCES

Adams, C. P., and Van Brantner, V. (2006). Estimating the cost of new drug development: is it really $802 million? *Health Aff.* 25, 420–428. doi: 10.1377/hlthaff.25.2.420

Akihiro, I., Takashi, O., Manabu, Y., Hideyuki, T., Kenjin, I., Koji, U., et al. (2006). *Process for Production of Fluoro Derivative.* Japan: Central Glass Co, Ltd.

Bajorath, J. (2015). Computer-aided drug discovery. *F1000Res.* 4:F1000. doi: 10.12688/f1000research.6653.1

Baumann, M., and Baxendale, I. R. (2016). Continuous-flow synthesis of 2H-azirines and their diastereoselective transformation to aziridines. *Synlett* 27, 159–163. doi: 10.1055/s-0035-1560391

Baumann, M., Baxendale, I. R., and Ley, S. V. (2008). The use of diethylaminosulfur trifluoride (DAST) for fluorination in a continuous-flow microreactor. *Synlett* 2008, 2111–2114. doi: 10.1055/s-2008-1078026

Bøgevig, A., Federsel, H.-J., Huerta, F., Hutchings, M. G., Kraut, H., Langer, T., et al. (2015). Route design in the 21st century: the ICSYNTH software tool as an idea generator for synthesis prediction. *Organ. Process Res. Dev.* 19, 357–368. doi: 10.1021/op500373e

Botella, P., Corma, A., Iborra, S., Montón, R., Rodríguez, I., and Costa, V. (2007). Nanosized and delayered zeolitic materials for the liquid-phase Beckmann rearrangement of cyclododecanone oxime. *J. Catal.* 250, 161–170. doi: 10.1016/j.jcat.2007.05.020

Brooke, G. M., Burdon, J., and Tatlow, J. C. (1961). 172. Aromatic polyfluoro-compounds. Part VII. The reaction of pentafluoronitrobenzene with ammonia. *J. Chem. Soc.* 802–07. doi: 10.1039/jr9610000802

Burbidge, R., Trotter, M., Buxton, B., and Holden, S. (2001). Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem.* 26, 5–14. doi: 10.1016/S0097-8485(01)00094-8

Cadeddu, A., Wylie, E. K., Jurczak, J., Wampler-Doty, M., and Grzybowski, B. A. (2014). Organic chemistry as a language and the implications of chemical linguistics for structural and retrosynthetic analyses. *Angew. Chem.* 126, 8246–8250. doi: 10.1002/ange.201403708

Calabrese, G. S., and Pissavini, S. (2011). From batch to continuous flow processing in chemicals manufacturing. *AIChE J.* 57, 828–834. doi: 10.1002/aic.12598

Chadwick, J., Jones, M., Mercer, A. E., Stocks, P. A., Ward, S. A., Park, B. K., et al. (2010). Design, synthesis and antimalarial/anticancer evaluation of spermidine linked artemisinin conjugates designed to exploit polyamine transporters in Plasmodium falciparum and HL-60 cancer cell lines. *Bioorg. Med. Chem.* 18, 2586–2597. doi: 10.1016/j.bmc.2010.02.035

Chorghade, R., Battilocchio, C., Hawkins, J. M., and Ley, S. V. (2013). Sustainable flow oppenauer oxidation of secondary benzylic alcohols with a heterogeneous zirconia catalyst. *Organ. Lett.* 15, 5698–5701. doi: 10.1021/ol4027107

Christ, C. D., Zentgraf, M., and Kriegl, J. M. (2012). Mining electronic laboratory notebooks: analysis, retrosynthesis, and reaction based enumeration. *J. Chem. Inform. Model.* 52, 1745–1756. doi: 10.1021/ci300116p

Coley, C. W., Barzilay, R., Jaakkola, T. S., Green, W. H., and Jensen, K. F. (2017a). Prediction of organic reaction outcomes using machine learning. *ACS Cent. Sci.* 3, 434–443. doi: 10.1021/acscentsci.7b00064

Coley, C. W., Green, W. H., and Jensen, K. F. (2018a). Machine learning in computer-aided synthesis planning. *Acc. Chem. Res.* 51, 1281–1289. doi: 10.1021/acs.accounts.8b00087

Coley, C. W., Green, W. H., and Jensen, K. F. (2019a). RDChiral: an RDKit wrapper for handling stereochemistry in retrosynthetic template extraction and application. *J. Chem. Inform. Model.* 59, 2529–2537. doi: 10.1021/acs.jcim.9b00286

Coley, C. W., Rogers, L., Green, W. H., and Jensen, K. F. (2017b). Computer-assisted retrosynthesis based on molecular similarity. *ACS Cent. Sci.* 3, 1237–1245. doi: 10.1021/acscentsci.7b00355

Coley, C. W., Rogers, L., Green, W. H., and Jensen, K. F. (2018b). SCScore: synthetic complexity learned from a reaction corpus. *J. Chem. Inform. Model.* 58, 252–261. doi: 10.1021/acs.jcim.7b00622

Coley, C. W., Thomas, D. A., Lummiss, J. A. M., Jaworski, J. N., Breen, C. P., Schultz, V., et al. (2019b). A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* 365:*eaax*1566. doi: 10.1126/science.aax1566

Cook, A., Johnson, A. P., Law, J., Mirzazadeh, M., Ravitz, O., and Simon, A. (2012). Computer-aided synthesis design: 40 years on. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 2, 79–107. doi: 10.1002/wcms.61

Corey, E. J. (1967). General methods for the construction of complex molecules. *Pure Appl. Chem.* 14, 19–38. doi: 10.1351/pac196714010019

Corey, E. J., and Wipke, W. T. (1969). Computer-assisted design of complex organic syntheses. *Science* 166:178–192. doi: 10.1126/science.166.3902.178

Curtin, T., McMonagle, J. B., and Hodnett, B. K. (1993). "Beckmann rearrangement over solid acid catalysts," *in Studies in Surface Science and Catalysis, Vol. 75* eds L. Guczi, F. Solymosi, and P. Tétényi (Elsevier). doi: 10.1016/S0167-2991(08)64361-X

DiMasi, J. A., Hansen, R. W., and Grabowski, H. G. (2003). The price of innovation: new estimates of drug development costs. *J. Health Econ.* 22, 151–185. doi: 10.1016/S0167-6296(02)00126-1

DiMasi, J. A., Hansen, R. W., Grabowski, H. G., and Lasagna, L. (1991). Cost of innovation in the pharmaceutical industry. *J. Health Econ.* 10, 107–142. doi: 10.1016/0167-6296(91)90001-4

Elsevier R&D Solutions (2016). *Reaxys Fact Sheet*. Available online at: https://www.elsevier.com/__data/assets/pdf_file/0009/945837/RDS_FactSheet_Reaxys_Oct_2016-WEB.pdf

Fuller, P. E., Gothard, C. M., Gothard, N. A., Weckiewicz, A., and Grzybowski, B. A. (2012). Chemical network algorithms for the risk assessment and management of chemical threats. *Angew. Chem. Int. Ed. Engl.* 51, 7933–7937. doi: 10.1002/anie.201202210

Gao, H., Struble, T. J., Coley, C. W., Wang, Y., Green, W. H., and Jensen, K. F. (2018). Using machine learning to predict suitable conditions for organic reactions. *ACS Cent. Sci.* 4:1465–1476. doi: 10.1021/acscentsci.8b00357

Gillies, D. (1996). *Artificial Intelligence and Scientific Method.* Oxford University Press.

Gilmore, K., and Seeberger, P. H. (2014). Continuous flow photochemistry. *Chem. Rec.* 14, 410–418. doi: 10.1002/tcr.201402035

Gothard, C. M., Soh, S., Gothard, N. A., Kowalczyk, B., Wei, Y., Baytekin, B., et al. (2012). Rewiring chemistry: algorithmic discovery and experimental validation of one-pot reactions in the network of organic chemistry. *Angew. Chem. Int. Ed.* 51, 7922–7927. doi: 10.1002/anie.201202155

Gutmann, B., Cantillo, D., and Kappe, C. O. (2015). Continuous-flow technology—a tool for the safe manufacturing of active pharmaceutical ingredients. *Angew. Chem. Int. Ed.* 54, 6688–6728. doi: 10.1002/anie.201409318

Han, W., Jin, F., and Zhou, Q. (2015). Ligand-free palladium-catalyzed hydroxycarbonylation of aryl halides under ambient conditions: synthesis of aromatic carboxylic acids and aromatic esters. *Synthesis* 47, 1861–1868. doi: 10.1055/s-0034-1380497

Hartman, R. L., and Jensen, K. F. (2009). Microchemical systems for continuous-flow synthesis. *Lab Chip* 9, 2495–2507. doi: 10.1039/b906343a

Hellman, A. N., Rau, K. R., Yoon, H. H., Bae, S., Palmer, J. F., Phillips, K. S., et al. (2007). Laser-induced mixing in microfluidic channels. *Anal. Chem.* 79, 4484–4492. doi: 10.1021/ac070081i

Huang, Q., Li, L.-L., and Yang, S.-Y. (2011). RASA: a rapid retrosynthesis-based scoring method for the assessment of synthetic accessibility of drug-like molecules. *J. Chem. Inform. Model.* 51, 2768–2777. doi: 10.1021/ci100216g

Ivanciuc, O. (2009). Machine learning quantitative structure-activity relationships (QSAR) for peptides binding to the human amphiphysin-1 SH3 domain. *Curr. Proteom.* 6, 289–302. doi: 10.2174/157016409789973725

Jirkovsky, I. (1974). Studies on enaminoketones. *Can. J. Chem.* 52, 55–65. doi: 10.1139/v74-009

Jong, T., and Bradley, M. (2015). Flow-mediated synthesis of Boc, Fmoc, and Ddiv monoprotected diamines. *Organ. Lett.* 17, 422–425. doi: 10.1021/ol503343b

Josyula, K., and Mitesh, P. (2014). *Continuous Flow Synthesis of Dithioester Compounds*. Patent No. WO2014152453A3.

Karpov, P., Godin, G., and Tetko, I. V. (2019). "A transformer model for retrosynthesis," in *International Conference on Artificial Neural Networks*, eds I. V Tetko, V. Kůrková, P. Karpov, and F. Theis (Springer International Publishing), 817–830.

Kitano, M., Masaya, M., Michio, U., and Masakazu, A. (2007). Recent developments in titanium oxide-based photocatalysts. *Appl. Catal. A Gen.* 325, 1–14. doi: 10.1016/j.apcata.2007.03.013

Ko, Y., Kim, M. H., Kim, S. J., Seo, G., Kim, M.-Y., and Uh, Y. S. (2000). Vapor phase beckmann rearrangement of cyclohexanone oxime over a novel tantalum pillared-ilerite. *Chem. Commun.* 829–30. doi: 10.1039/b001466o

Koch, M., Duigou, T., and Faulon, J.-L. (2020). Reinforcement Learning for Bioretrosynthesis. *ACS Synthetic Biol.* 9, 157–168. doi: 10.1021/acssynbio.9b00447

Kockmann, N., Thenée, P., Fleischer-Trebes, C., Laudadio, G., and Noël, T. (2017). Safety assessment in development and operation of modular continuous-flow processes. *React. Chem. Eng.* 2, 258–280. doi: 10.1039/C7RE00021A

Konze, K. D., Bos, P. H., Dahlgren, M. K., Leswing, K., Tubert-Brohman, I., Bortolato, A., et al. (2019). Reaction-based enumeration, active learning, and free energy calculations to rapidly explore synthetically tractable chemical space and optimize potency of cyclin-dependent kinase 2 inhibitors. *J. Chem. Inform. Model.* 59, 3782–3793. doi: 10.1021/acs.jcim.9b00367

Kowalik, M., Gothard, C. M., Drews, A. M., Gothard, N. A., Weckiewicz, A., Fuller, P. E., et al. (2012). Parallel optimization of synthetic pathways within the network of organic chemistry. *Angew. Chem. Int. Ed.* 51, 7928–7932. doi: 10.1002/anie.201202209

Laidler, K. J., and King, M. C. (1983). Development of transition-state theory. *J. Phys. Chem.* 87, 2657–2664. doi: 10.1021/j100238a002

Langley, P. (1998). "The computer-aided discovery of scientific knowledge," in *Discovery Science: First International Conference*, eds S. Arikawa and H. Motoda (Berlin; Heidelberg: Springer), 25–39.

Law, J., Zsoldos, Z., Simon, A., Reid, D., Liu, Y., Khew, S. Y., et al. (2009). Route designer: a retrosynthetic analysis tool utilizing automated retrosynthetic rule generation. *J. Chem. Inform. Model.* 49, 593–602. doi: 10.1021/ci800228y

Lee, K. M., Lai, C. W., Ngai, K. S., and Juan, J. C. (2016). Recent developments of zinc oxide based photocatalyst in water treatment technology: a review. *Water Res.* 88, 428–448. doi: 10.1016/j.watres.2015.09.045

Li, J., and Eastgate, M. D. (2015). Current complexity: a tool for assessing the complexity of organic molecules. *Organ. Biomol. Chem.* 13, 7164–7176. doi: 10.1039/C5OB00709G

Li, J., and Eastgate, M. D. (2019). Making better decisions during synthetic route design: leveraging prediction to achieve greenness-by-design. *React. Chem. Eng.* 4, 1595–1607. doi: 10.1039/C9RE00019D

Lin, K., Xu, Y., Pei, J., and Lai, L. (2020). Automatic retrosynthetic route planning using template-free models. *Chem. Sci.* 11, 3355–3364. doi: 10.1039/C9SC03666K

Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E., and Svetnik, V. (2015). Deep neural nets as a method for quantitative structure–activity relationships. *J. Chem. Inform. Model.* 55, 263–274. doi: 10.1021/ci500747n

Malet-Sanz, L., Madrzak, J., Ley, S. V., and Baxendale, I. R. (2010). Preparation of arylsulfonyl chlorides by chlorosulfonylation of in situ generated diazonium salts using a continuous flow reactor. *Organ. Biomol. Chem.* 8, 5324–5332. doi: 10.1039/c0ob00450b

Mallia, C. J., Walter, G. C., and Baxendale, I. R. (2016). Flow carbonylation of sterically hindered ortho-substituted iodoarenes. *Beilstein J. Org. Chem.* 12, 1503–1511. doi: 10.3762/bjoc.12.147

Maltarollo, V. G., Gertrudes, J. C., Oliveira, P. R., and Honorio, K. M. (2015). Applying machine learning techniques for ADME-Tox prediction: a review. *Expert Opin. Drug Metab. Toxicol.* 11, 259–271. doi: 10.1517/17425255.2015.980814

Marcou, G., de Sousa, J. A., Latino, D. A. R. S., de Luca, A., Horvath, D., Rietsch, V., et al. (2015). Expert system for predicting reaction conditions: the michael reaction case. *J. Chem. Inform. Model.* 55, 239–250. doi: 10.1021/ci500698a

Mayr, A., Klambauer, G., Unterthiner, T., and Hochreiter, S. (2016). DeepTox: toxicity prediction using deep learning. *Front. Environ. Sci.* 3:80. doi: 10.3389/fenvs.2015.00080

McPake, C. B., Murray, C. B., and Sandford, G. (2012). Sequential continuous flow processes for the oxidation of amines and azides by using HOF·MeCN. *ChemSusChem* 5, 312–319. doi: 10.1002/cssc.201100423

Melnikov, A. A., Nautrup, H. P., Krenn, M., Dunjko, V., Tiersch, M., Zeilinger, A., et al. (2018). Active learning machine learns to create new quantum experiments. *Proc. Natl. Acad. Sci. U. S. A.* 115, 1221–1226. doi: 10.1073/pnas.1714936115

Monteiro, J. L., Pieber, B., Corrêa, A. G., and Kappe, C. O. (2016). Continuous synthesis of hydantoins: intensifying the bucherer–bergs reaction. *Synlett* 27, 83–87. doi: 10.1055/s-0035-1560317

Morgan, H. L. (1965). The generation of a unique machine description for chemical structures-A technique developed at chemical abstracts service. *J. Chem. Document.* 5, 107–113. doi: 10.1021/c160017a018

Movsisyan, M., Delbeke, E. I., Berton, J. K., Battilocchio, C., Ley, S. V., and Stevens, C. V. (2016). Taming hazardous chemistry by continuous flow technology. *Chem. Soc. Rev.* 45, 4892–4928. doi: 10.1039/C5CS00902B

Naik, A. W., Kangas, J. D., Langmead, C. J., and Murphy, R. F. (2013). Efficient modeling and active learning discovery of biological responses. *PLoS ONE* 8:*e*83996. doi: 10.1371/journal.pone.0083996

Nielsen, M. K., Ahneman, D. T., Riera, O., and Doyle, A. G. (2018). Deoxyfluorination with sulfonyl fluorides: navigating reaction space with machine learning. *J. Am. Chem. Soc.* 140, 5004–5008. doi: 10.1021/jacs.8b01523

OBrien, M., and Cooper, D. (2016). Continuous flow liquid–liquid separation using a computer-vision control system: the bromination of enaminones with N-bromosuccinimide. *Synlett* 27, 164–168. doi: 10.1055/s-0035-1560975

Pathak, S., Kundu, A., and Pramanik, A. (2014). Monobromomalononitrile: an efficient regioselective mono brominating agent towards active methylene compounds and enamines under mild conditions. *RSC Adv.* 4, 10180–10187. doi: 10.1039/C3RA46687F

Peplow, M. (2014). Organic synthesis: the robo-chemist. *Nature* 512:20. doi: 10.1038/512020a

Pieri, C., Borselli, D., Di Giorgio, C., De Méo, M., Bolla, J.-M., Vidal, N., et al. (2014). New ianthelliformisamine derivatives as antibiotic enhancers against resistant gram-negative bacteria. *J. Med. Chem.* 57, 4263–4272. doi: 10.1021/jm500194e

Plutschack, M. B., Pieber, B., Gilmore, K., and Seeberger, P. H. (2017). The hitchhikers guide to flow chemistry. *Chem. Rev.* 117, 11796–11893. doi: 10.1021/acs.chemrev.7b00183

Prevet, H., Flipo, M., Roussel, P., Deprez, B., and Willand, N. (2016). Microwave-assisted synthesis of functionalized spirohydantoins as 3-D privileged fragments for scouting the chemical space. *Tetrahedron Lett.* 57, 2888–2894. doi: 10.1016/j.tetlet.2016.05.065

Roberge, D. M., Zimmermann, B., Rainone, F., Gottsponer, M., Eyholzer, M., and Kockmann, N. (2008). Microreactor technology and continuous processes in the fine chemical and pharmaceutical industry: is the revolution underway? *Org. Process Res. Dev.* 12, 905–910. doi: 10.1021/op8001273

Rogers, D., and Hahn, M. (2010). Extended-connectivity fingerprints. *J. Chem. Inform. Model.* 50, 742–754. doi: 10.1021/ci100050t

Ryu, S., Lim, J., Hong, S. H., and Kim, W. Y. (2018). Deeply learning molecular structure-property relationships using attention- and gate-augmented graph convolutional network. *arXiv [Preprint]. arXiv:10988.*

Safari, J., Gandomi-Ravandi, S., and Javadian, L. (2013). Microwave-promoted facile and rapid synthesis procedure for the efficient synthesis of 5,5-disubstituted hydantoins. *Synthetic Commun.* 43, 3115–3120. doi: 10.1080/00397911.2012.730647

Safari, J., and Javadian, L. (2013). A one-pot synthesis of 5,5-disubstituted hydantoin derivatives using magnetic Fe3O4 nanoparticles as a

reusable heterogeneous catalyst. *Comp. Rendus Chim.* 16, 1165–1171. doi: 10.1016/j.crci.2013.06.005

Scalia, G., Grambow, C. A., Pernici, B., Li, Y.-P., and Green, W. H. (2020). Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. *J. Chem. Inf. Model.* 60, 2697–2717. doi: 10.1021/acs.jcim.9b00975

Schneider, G. (2017). Automating drug discovery. *Nat. Rev. Drug Discov.* 17, 97–113. doi: 10.1038/nrd.2017.232

Schneider, N., Lowe, D. M., Sayle, R. A., and Landrum, G. A. (2015). Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity. *J. Chem. Inform. Model.* 55, 39–53. doi: 10.1021/ci5006614

Segler, M. H. S., Kogej, T., Tyrchan, C., and Waller, M. P. (2018a). Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* 4, 120–131. doi: 10.1021/acscentsci.7b00512

Segler, M. H. S., Preuss, M., and Waller, M. P. (2018b). Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* 555, 604–610. doi: 10.1038/nature25978

Segler, M. H. S., and Waller, M. P. (2017a). Modelling chemical reasoning to predict and invent reactions. *Chem. A Eur. J.* 23, 6118–6128. doi: 10.1002/chem.201604556

Segler, M. H. S., and Waller, M. P. (2017b). Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chem. A Eur. J.* 23, 5966–5971. doi: 10.1002/chem.201605499

Seia, M. A., Stege, P. W., Pereira, S. V., De Vito, I. E., Raba, J., and Messina, G. A. (2014). Silica nanoparticle-based microfluidic immunosensor with laser-induced fluorescence detection for the quantification of immunoreactive trypsin. *Anal. Biochem.* 463, 31–37. doi: 10.1016/j.ab.2014.06.016

Szymkuć, S., Gajewska, E. P., Klucznik, T., Molga, K., Dittwald, P., Startek, M., et al. (2016). Computer-assisted synthetic planning: the end of the beginning. *Angew. Chem. Int. Ed.* 55, 5904–5937. doi: 10.1002/anie.201506101

Teoh, S. K., Rathi, C., and Sharratt, P. (2016). Practical assessment methodology for converting fine chemicals processes from batch to continuous. *Org. Process Res. Dev.* 20, 414–431. doi: 10.1021/acs.oprd.5b00001

Tsai, S.-W., Huang, S.-H., Lee, H.-S., and Tsai, F.-Y. (2013). A reusable palladium(II)/cationic 2,2′-bipyridyl catalytic system for hydroxycarbonylation of aryl iodides in water. *J. Chin. Chem. Soc.* 60, 769–772. doi: 10.1002/jccs.201200595

Unsleber, J. P., and Reiher, M. (2020). The exploration of chemical reaction networks. *Annu. Rev. Phys. Chem.* 71, 121–142. doi: 10.1146/annurev-physchem-071119-040123

van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605. Available online at: http://www.jmlr.org/papers/v9/vandermaaten08a.html

Wang, L.-P., Titov, A., McGibbon, R., Liu, F., Pande, V. S., and Martínez, T. J. (2014). Discovering chemistry with an ab initio nanoreactor. *Nat. Chem.* 6, 1044–1048. doi: 10.1038/nchem.2099

Warr, W. A. (2014). A short review of chemical reaction database systems, computer-aided synthesis design, reaction prediction and synthetic feasibility. *Mol. Inform.* 33, 469–476. doi: 10.1002/minf.201400052

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inform. Comput. Sci.* 28, 31–36. doi: 10.1021/ci00057a005

Wiles, C., and Watts, P. (2012). Continuous flow reactors: a perspective. *Green Chem.* 14, 38–54. doi: 10.1039/C1GC16022B

Yuan, S., Qin, J.-S., Li, J., Huang, L., Feng, L., Fang, Y., et al. (2018). Retrosynthesis of multi-component metal–organic frameworks. *Nat. Commun.* 9:808. doi: 10.1038/s41467-018-03102-5

Zhang, G., Zhao, Y., Xuan, L., and Ding, C. (2019). SO2F2-activated efficient beckmann rearrangement of ketoximes for accessing amides and lactams. *Eur. J. Org. Chem.* 2019, 4911–4915. doi: 10.1002/ejoc.201900844

Zhang, L., Mao, H., Liu, L., Du, J., and Gani, R. (2018). A machine learning based computer-aided molecular design/screening methodology for fragrance molecules. *Comput. Chem. Eng.* 115, 295–308. doi: 10.1016/j.compchemeng.2018.04.018

Zhang, P., Wang, M., Dong, J., Li, X., Wang, F., Wu, L., et al. (2010). Photocatalytic hydrogen production from water by noble-metal-free molecular catalyst systems containing rose bengal and the cobaloximes of BFx-bridged oxime ligands. *J. Phys. Chem. C* 114, 15868–15874. doi: 10.1021/jp106512a

Zhang, X., Jin, Z., Li, Y., Li, S., and Lu, G. (2009). Efficient photocatalytic hydrogen evolution from water without an electron mediator over Pt–rose bengal catalysts. *J. Phys. Chem. C* 113, 2630–2635. doi: 10.1021/jp8085717

Zheng, S., Rao, J., Zhang, Z., Xu, J., and Yang, Y. (2020). Predicting retrosynthetic reactions using self-corrected transformer neural networks. *J. Chem. Inf. Model.* 60, 47–55. doi: 10.1021/acs.jcim.9b00949

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.