Research Article
doi.org/10.1002/cmtd.202200062
Chemistry Europe
European Chemical Societies Publishing

www.chemistrymethods.org

Reaction Impurity Prediction using a Data Mining Approach**

Adarsh Arun,[a, c] Zhen Guo,[b, c] Simon Sung,[c] and Alexei A. Lapkin*[a, b, c]

Automated prediction of reaction impurities is useful in early-stage reaction development, synthesis planning and optimization. Existing reaction predictors are catered towards *main* product prediction, and are often black-box, making it difficult to troubleshoot erroneous outcomes. This work aims to present an automated, interpretable impurity prediction workflow based on data mining large chemical reaction databases. A 14-step workflow was implemented in Python and RDKit using Reaxys® data. Evaluation of potential chemical reactions between functional groups present in the same reaction environment in the user-supplied query species can be accurately performed by directly mining the Reaxys® database for similar or 'analogue' reactions involving these functional groups. Reaction templates can then be extracted from analogue reactions and applied to the relevant species in the original query to return impurities and transformations of interest. Three proof-of-concept case studies (paracetamol, agomelatine and lersivirine) were conducted, with the workflow correctly suggesting impurities within the top two outcomes. At all stages, suggested impurities can be traced back to the originating template and analogue reaction in the literature, allowing for closer inspection and user validation. Ultimately, this work could be useful as a benchmark for more sophisticated algorithms or models since it is interpretable, as opposed to purely black-box solutions.

Introduction

Early-stage knowledge of impurities and conditions under which they may form is crucial for the rapid design of robust, scalable, and sustainable synthesis pathways for target molecules. This is especially relevant for the pharmaceutical industry where impurity tolerance is low so as to maximize the safety and efficacy of the final drug product. Traditionally, for a given synthesis pathway and target molecule, chemical intuition and analytical techniques are employed in tandem to conduct rigorous impurity profiling.[1] However, these approaches can be time consuming, iterative, and expensive. Prediction of impurities *a priori* is an attractive option to alleviate some of these drawbacks and quickly identify or optimize synthesis routes. Notably, given the development of large chemical reaction databases such as Reaxys®[2][1] (>148 million substances and >54 million reactions), leveraging this available chemical data to make informed predictions is more viable now than ever before. Therefore, the primary aim of this work is to propose an automated impurity prediction workflow based on data mining Reaxys®.

This work is set in the broader context of the field of Computer-Aided Synthesis Planning (CASP), which has made strides in the automated discovery and design of chemical synthesis routes over the past several decades.[3–6] CASP methods have been classified as encompassing (but not limited to) three main tasks:[3] i) retrosynthetic analysis, which starts at a target molecule and works backwards towards simpler precursors to identify suitable synthetic routes; ii) forward reaction prediction, which starts with precursors and predicts potential products; and iii) reaction conditions prediction (catalyst, solvent, temperature etc. required for a reaction to occur). Impurity prediction can be regarded as a subtask of forward reaction prediction and involves predicting byproducts of a reaction given a set of precursors (reactants, reagents, main products etc.) and may also optionally include reaction conditions (temperature, catalyst, solvent etc.).

Existing data-driven methodologies for forward reaction prediction focus primarily on predicting **main products** based on reaction data in large chemical databases. Byproducts and impurities are often deprioritized because examples of negative and failed reactions are rare in literature.[3] Approaches usually fall under one of two categories:

[a] *A. Arun, Prof. A. A. Lapkin*
Department of Chemical Engineering and Biotechnology
University of Cambridge
Philippa Fawcett Drive
Cambridge CB3 0AS (UK)
E-mail: aal35@cam.ac.uk
[b] *Dr. Z. Guo, Prof. A. A. Lapkin*
Chemical Data Intelligence (CDI) Pte Ltd
Robinson Road, #02-00
068898 Singapore (Singapore)
[c] *A. Arun, Dr. Z. Guo, Dr. S. Sung, Prof. A. A. Lapkin*
Cambridge Centre for Advanced Research and Education in Singapore
1 CREATE Way
CREATE Tower #05-05
138602 Singapore (Singapore)

[**] *A previous version of this manuscript has been deposited on a preprint server (https://doi.org/10.26434/chemrxiv-2022-0btmt).*

Supporting information for this article is available on the WWW under https://doi.org/10.1002/cmtd.202200062

[1] Copyright © 2022 Elsevier Limited except certain content provided by third parties. Reaxys® is a trademark of Elsevier Limited.

I. **Rule/template-based**: These methods use manually encoded reaction/transformation rules[7–11] or automatically extracted reaction/transformation rules[12,13] from chemical databases often in the form of reaction templates to predict products. Recent methods have employed neural network classifiers and other machine learning (ML) algorithms to prioritize templates and transformations.[12,13]
II. **Rule/template-free**: These methods do not rely on rules or templates, instead leveraging advanced ML techniques such as graph neural networks[14] and transformer architectures based on NLP (Natural Language Processing)[15] to suggest products.

There are also other approaches, such as knowledge graphs,[16] stochastic block models based on large reaction networks,[17] and mechanistic ML models using a variety of descriptors for quantitative reaction prediction.[18–22] A recent review by Thakkar et al.[4] provides a comprehensive assessment of all these methods.

Recently, state-of-the-art data-driven methods have gravitated towards black-box models with limited interpretability or transparency.[15] As a result, it is harder to understand how or why certain products are predicted, introducing ambiguity in diagnosing erroneous outcomes. This is further exacerbated by the relatively little focus given to the underlying data used to train and test models, which can contain biases,[23–25] and lack diversity. At present, it can be challenging to pinpoint where the issues lie: the quality of the underlying data, or the type of model employed, or both. Although the field is rapidly progressing to address these drawbacks and challenges,[26,27] impurity prediction has not yet been explicitly addressed within this context hence the aim of this work. An automated impurity prediction workflow that is interpretable and transparent is possible with data mining. Additionally, the potential and pitfalls of chemical data in impurity prediction can be highlighted clearly and the workflow could serve as a benchmark when designing, refining, and troubleshooting ML approaches in the future.

Method Development

An overview

Impurity prediction fundamentally involves assessing how functional groups present in reactants, reagents, solvents and main products (henceforth referred to as *query species*) may react with each other under the specified conditions. It is proposed that this assessment can be performed by directly mining a database of *analogue* reactions, here defined as reactions involving functional group fragments derived from the query species. Our source of the original reaction data is Reaxys®, obtained with Elsevier's permission. As data in Reaxys® may be incomplete or contain erroneous information, it is necessary to prepare the data for this workflow. This involves removing duplicate reactions and ensuring that reactions are stoichiometrically balanced. Extracting reaction templates from the analogue reactions and applying them to relevant query species can suggest potential impurities and transformations of interest. Reaction conditions such as temperature and catalysts can be leveraged to further filter and only keep the realistic impurities. In keeping with interpretability, at all points, the suggested impurities can be traced back to the originating template and analogue reaction in the literature.

Figure 1 illustrates the proposed workflow for impurity prediction. The workflow consists of fourteen steps split across four main modules: I. Data mining, II. Data processing, III. Impurity prediction and IV. Impurity ranking.

The 14-step workflow is fully automated and was implemented in Python using RDKit,[28] a popular open-source cheminformatics library that facilitates reaction representation and manipulation. The workflow was run on a Linux (Ubuntu V 16.04.6 LTS) 32-core server using Ray,[29] an open-source Python library for distributed computing.

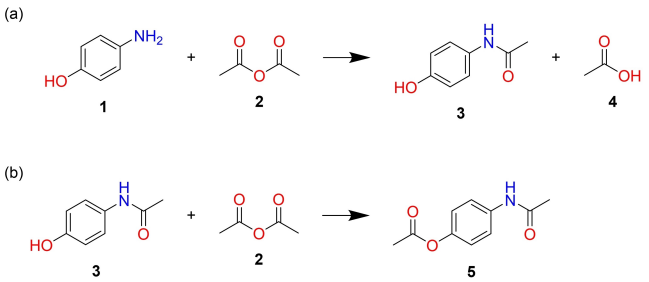
An illustrative case study, and step-by-step guide

To more clearly explain the workflow presented in Figure 1, an illustrative case study is outlined here, based on paracetamol (acetaminophen) synthesis, Scheme 1. This is a widely practiced reaction that involves acetylation of 4-aminophenol (**1**) with acetic anhydride (**2**) to produce paracetamol (**3**) and acetic acid (**4**). Based on impurity studies,[30] and shown in Scheme 1b, a possible impurity or by-product is 4-acetamidophenyl acetate (**5**), which is obtained when **3** is further acetylated by **2** in an overreaction. The proposed workflow should be able to predict this impurity and is explained in more detail in the following sections.

Data mining

The data mining portion of the workflow consists of steps 1 to 4 in Figure 1 and involves mining the Reaxys® database for analogue reactions. Figure 2 illustrates all these steps in the context of paracetamol synthesis.

The first step in the workflow is to input query species, here defined as user-supplied reactants, reagents, solvents, and main products that may interact with each other, in the form of a



Scheme 1. (a) Reaction scheme for paracetamol synthesis. 4-Aminophenol (**1**) is acetylated by acetic anhydride (**2**) to give paracetamol (**3**) and acetic acid (**4**). (b) Reaction scheme for impurity formation: 4-Acetamidophenyl acetate (**5**) is produced through an overreaction of paracetamol (**3**) with acetic anhydride (**2**). Note that reactions are not balanced.

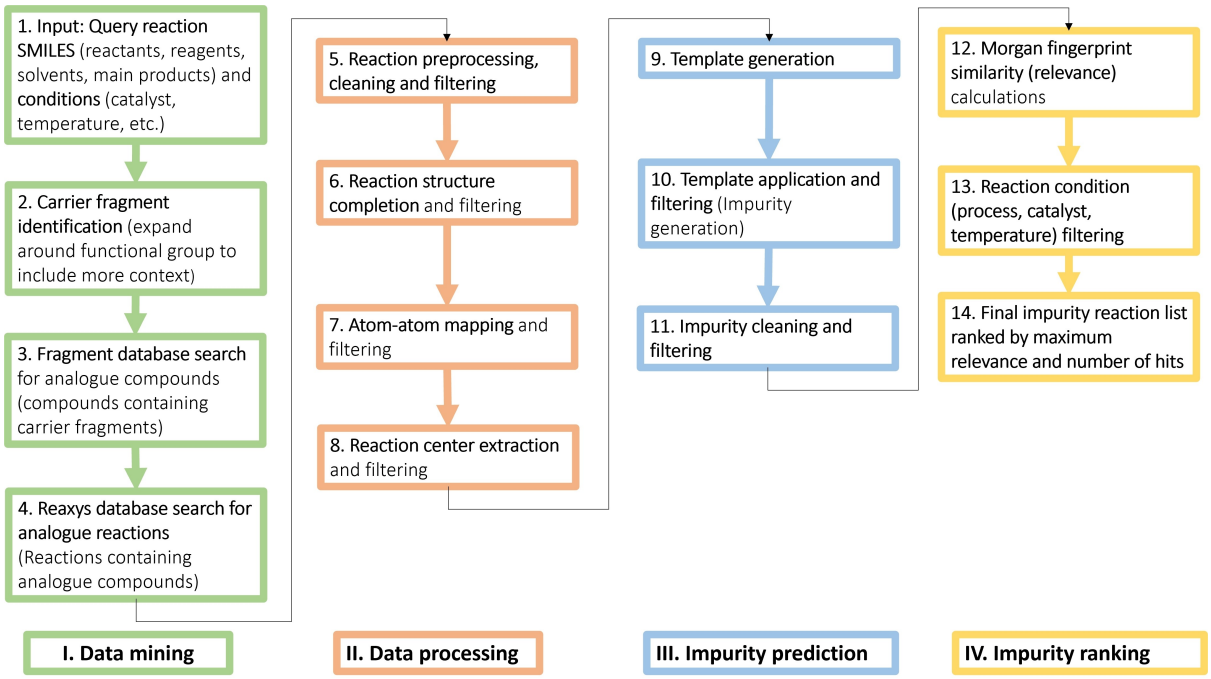


Figure 1. An illustration of the developed workflow for automated impurity prediction based on data mining. Fourteen steps are split across four modules: I. Data mining, II. Data processing, III. Impurity prediction and IV. Impurity ranking.

reaction SMILES (Simplified Molecular-Input Line-Entry System) string, which is the most widely adopted form of representation. Reactants, reagents, and solvents are not distinguished in this step as all query species are assumed to potentially react. Additionally, distinguishing these species would require *a priori* knowledge and would defeat the motives behind reaction prediction. However, reaction conditions can be optionally provided separately (e.g., catalyst, temperature, special processes) if already known. In this case, query species 1–4 are inputted.

Step 2 involves identifying functional group (FG) fragments in each user-supplied query species. Approaches in literature are divided into two general approaches: using manually curated substructure lists to identify FGs[31–33] or automatic identification methods.[34] Ertl's algorithm[34] is a most recent example of the latter, which automatically extracts FGs iterating through atoms in a molecule, and is easily implemented in cheminformatics libraries such as CDK[35] and RDKit.[36] The RDKit implemented algorithm was thus adapted in this approach to account for bonded hydrogens as well as the first degree environment (nearest neighbors) around the functional group. For instance, **1** contains hydroxyl (OH) and amine (NH_2) functional groups (highlighted in blue in Figure 2) which were expanded to include the aromatic carbons (highlighted in green in Figure 2), forming FG fragments C_{ar}–OH and C_{ar}–NH_2. These are henceforth referred to as carrier fragments. It is important to note that the user can choose to employ higher-degree environments to capture more molecular context, especially relevant groups further away that influence reactivity. However, keeping in mind that there is an inherent trade-off between template specificity and generalizability, the first-degree environment is used in this work. The final output of this step is a list of carrier fragments that belong to each query species, in this case: aromatic hydroxyl (belonging to **1** and **3**), aromatic amine (belonging to **1**), anhydride (belonging to **2**), acetamido (belonging to **3**) and carboxylic (belonging to **4**) fragments.

Step 3 involves finding analogue species that contain any of the identified carrier fragments. For this, a subset of the Reaxys® database (*circa* 17 million reactions) was extracted and the carrier fragment identification algorithm in step 2 was applied to every species contained in the Reaxys® data after canonicalization, leading to a carrier fragment database for faster querying (For more details, refer to Section S1 of the Supporting Information). Thus, for each carrier fragment, a list of analogue species that contains it can be retrieved. For instance, 3-aminophenol is an analogue species that contains both aromatic amine (at the *meta* position instead of *para*) and hydroxyl carrier fragments present in **1**, the query species. More than 1.9 million analogue species were obtained in this step.

In step 4, analogue reactions that only contain the aforementioned analogue or query species are extracted from the Reaxys® data. More specifically, all reactants and reagents in these reactions are required to be analogue or query species, ensuring that reactions between designated carrier fragments are possible. For instance, Figure 2a involves hydroquinone (analogue species, containing the aromatic amine carrier fragment) and acetic anhydride (query species) as reactants, and acetic acid (query species) as a reagent. The reader is referred to the Section S1 in the Supporting Information for more details on definitions of species roles in this work. The final result is a list of 243,600 analogue reactions, ready to be processed in the next portion of the workflow.

Research Article
doi.org/10.1002/cmtd.202200062
Chemistry Europe
European Chemical Societies Publishing

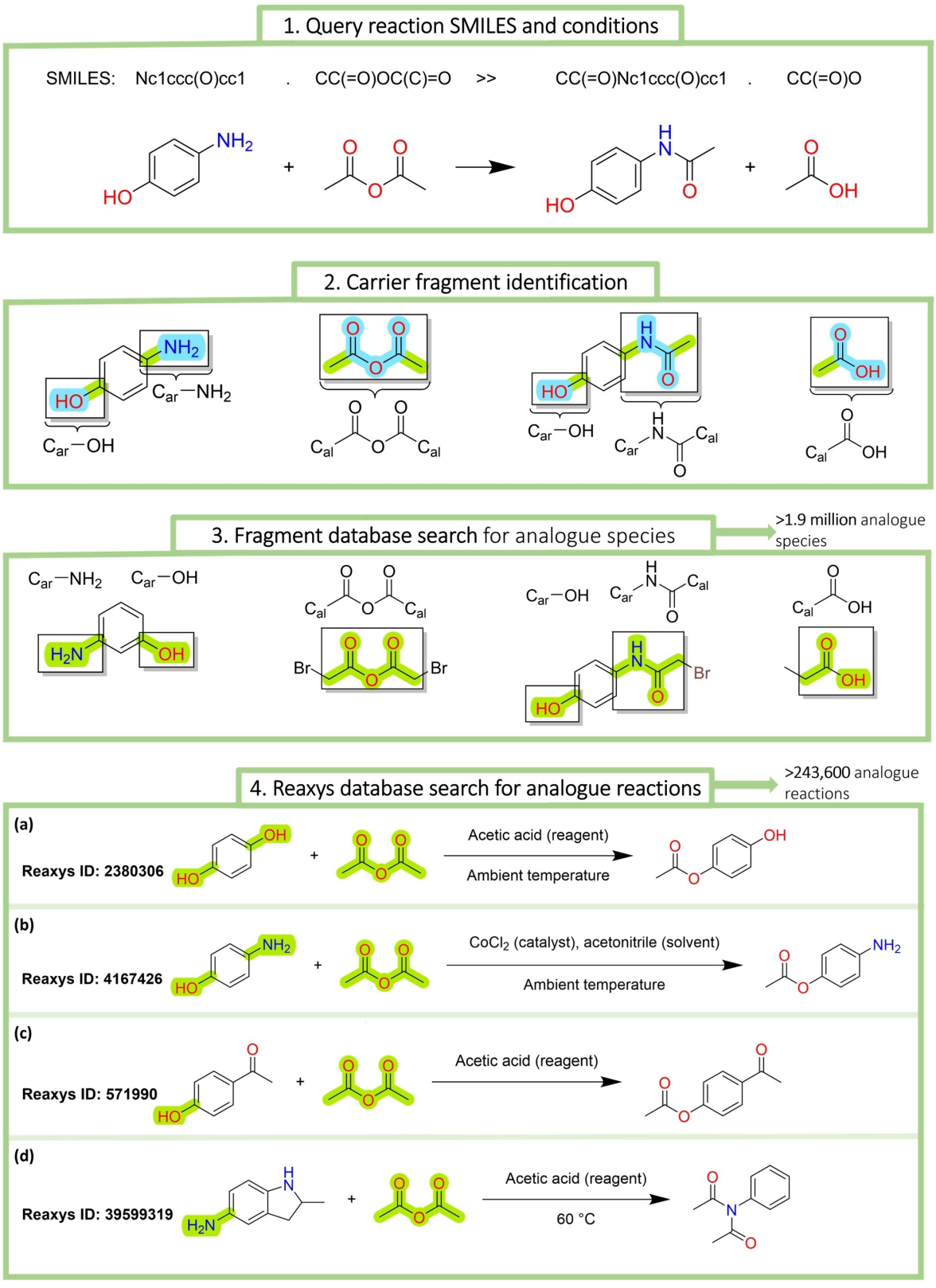


Figure 2. Steps 1 to 4 of the impurity prediction workflow with illustrative examples on the paracetamol synthesis. (a) to (c) represent valid analogue reactions, with Reaxys® IDs indicated on the left.

Data processing

The data processing portion of the workflow consists of steps 5 to 8 (refer to Figure 1), and involves processing the analogue reactions obtained in step 4 for template generation. Figure 3 illustrates all these steps in the context of paracetamol synthesis, giving relevant examples.

Step 5 is the pre-processing, cleaning and filtering the obtained analogue reactions. There are several potential issues with representing reactions extracted from Reaxys®, which have also been discussed in a recent review on transformation and structure data curation.[23] Firstly, as is the case in Figure 3b, invalid reactants, reagents and/or products may be present. These are species that have names but are either missing a SMILES representation and structure or contain an erroneous SMILES representation and structure when parsed by RDKit. In the case of the latter, valence errors are commonplace and in the case of the former, the species may be too complex or generic to identify an appropriate SMILES. These reactions are incomplete and are removed, together with reactions that are missing reactants and products. Additionally, analogue reactions may be missing all reaction conditions, which we define as encompassing reagents, solvents, catalysts, temperature, pressure, process (e.g., pyrolysis) and reaction time. In all these cases, the analogue reaction is of limited use, and is removed. This leaves *circa* 134,000 reactions, or 55% of the original analogue reaction set with valid reactants, reagents, products, and guaranteed to contain at least one of the previously mentioned reaction conditions as is the case in Figure 3a.

Step 6 resolves the issue of incomplete reaction structures that can be found in Reaxys®. As mentioned previously, some reaction records are imbalanced when directly taken from Reaxys®, due to missing species on the right-hand-side (RHS) and left-hand-side (LHS). Completely balanced reactions are needed to ensure that all possible products (including by-products) are present and that all reactant atoms can be traced to product atoms. Relatively few reports exist in the literature that aim to complete reaction structures, with notable recent approaches relying on transformers[37] and Condensed Graph of Reactions (CGR).[23]

Here, in keeping with interpretability, a balancing algorithm is implemented that considers the atom balance between the RHS and LHS. When there is an atom surplus on the RHS, reactants, reagents and solvents that address this surplus are added to the LHS. Likewise, when there is an atom surplus on the LHS, compatible small help species are added to the RHS. Help species are taken from a library of compounds ranging from water to small carboxylic acids and alcohols. For instance, Figure 3c contains an atom surplus on the RHS that matches acetic anhydride which is initially listed as a reagent in Reaxys®. Introducing acetic anhydride as a reactant creates an atom surplus on the LHS that can be balanced with a molecule of help species acetic acid on the RHS, thus properly balancing the reaction. In some cases, for instance in Figure 3d, the atom surplus on the RHS is caused by an untraceable group that cannot be resolved by reactants, reagents, or solvents on the LHS. These reactions are erroneous, missing important species and are removed. In cases where multiple candidate species can meet the atom surplus on either side of the reaction, an atom-to-atom mapper (detailed in step 7) is used to identify the most suitable species. A more detailed visualisation of the balancing algorithm is given in the Supporting Information Section S2.

Step 7 involves atom-to-atom mapping to establish a one-to-one correspondence between reactant and product atoms. Each product atom is mapped to a corresponding reactant atom by way of a common index number. This is important for later steps to extract reaction centres and generate templates. There are many atom-to-atom mapping tools that are available publicly such as NameRXN,[38] ChemAxon Standardizer,[39] Indigo,[40] Reaction Decoder Tool (RDT)[41] and IBM RXNMapper.[42] Specifically, RXNMapper is based on transformer models relying on NLP, which is template-free, and a recent benchmarking study[43] identified it as the most accurate tool (83.4% accuracy on a cleaned dataset published by Jaworski et al.[44]). Therefore, RXNMapper was utilized to map the remaining reactions. If some species are completely unmapped (due to multiple candidate species present after step 6), they are removed from the relevant reactions, which are then processed by steps 6 and 7 again. In some cases such as Figure 3f, due to complex species, the balanced reaction SMILES string length passed in exceeded the maximum sequence length that the mapper could cope with, resulting in errors. These reactions are removed, together with any imbalanced reactions that could not be balanced with help species in step 6. The final set of 64,499 reactions are fully balanced and mapped, ensuring every product atom can be traced back to a reactant atom.

Finally, in step 8, reaction centres of the remaining analogue reactions are obtained to ensure that they are within designated carrier fragments. The reaction centre constitutes all atoms that change in connectivity from reactants to products. For each mapping index number, corresponding atoms in reactant and product are compared with respect to eight criteria: atom SMARTS, atomic number, degree, formal charge, aromaticity, number of bonded hydrogens, number of radical electrons, and nearest neighbour identities (bond order and atomic number). If there are differences, that atom is part of the reaction centre. Additionally, the reaction centre must lie within carrier fragments of analogue species; this implies that relevant carrier fragments are reacting with each other and could lead to sensible impurities when applied to query species. For instance, in Figure 3g, atoms 8, 4 and 6 (highlighted in red) are identified as the reaction centre since they change with respect to the eight criteria. Atom 8 lies within the aromatic hydroxyl carrier fragment (highlighted in green), whilst atoms 4 and 6 lie within the anhydride carrier fragment (highlighted in green), so this is a reaction of interest. On the other hand, Figure 3h is invalid as the reaction centre, notably atoms 1 and 9 (aldehyde, circled in red) are outside the two valid aromatic hydroxyl fragments. The aldehyde involved is not a valid carrier fragment as it is not present in any query species. Other examples include reactions that don't contain a reaction centre or feature any changes in connectivity for example in the case of isomerism.

Research Article
doi.org/10.1002/cmtd.202200062

Chemistry Europe
European Chemical Societies Publishing

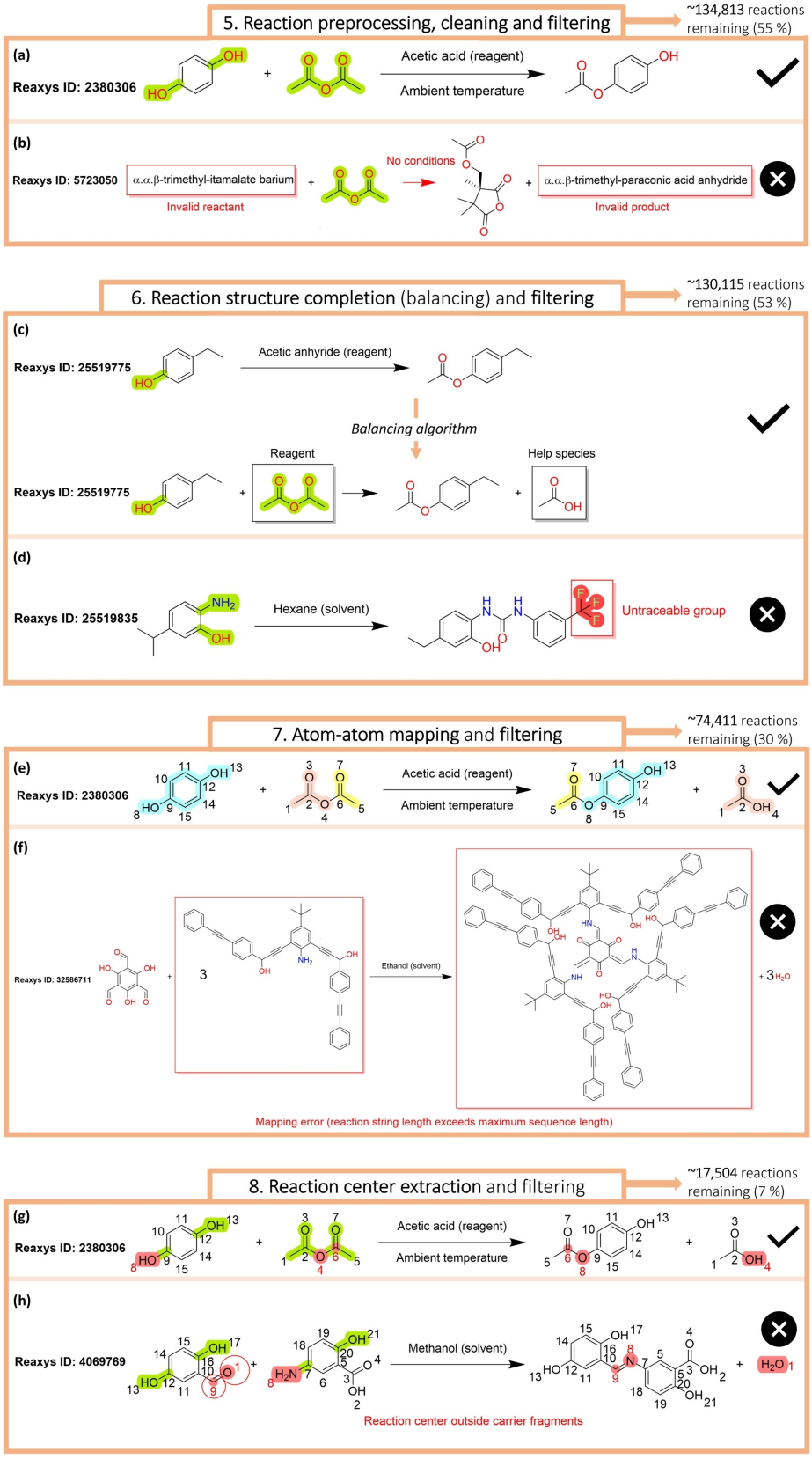


Figure 3. Steps 5 to 8 of the impurity prediction workflow (data processing) with illustrative examples on paracetamol synthesis. (a) to (g) are analogue reactions discussed in the text, with Reaxys® IDs indicated on the left. A tick or a cross is indicated on the right, representing a valid (kept) and invalid example (filtered out) respectively.

These types of reactions are removed, leaving a set of 17,504 analogue reactions that are guaranteed to feature interactions between carrier fragments.

Impurity prediction

The impurity prediction portion of the workflow consists of steps 9 to 11 in Figure 1, and involves generating and applying templates to suggest impurities. Figure 4 illustrates all these steps in the context of paracetamol synthesis, giving relevant examples.

Step 9 refers to template generation. Existing template-based workflows generate reaction templates from reaction centres, expanding around them to the nearest neighbours.[12,45] However, in this work, as carrier fragments already incorporate this knowledge, and involved carrier fragments can be determined based on the reaction centre, templates can be directly generated. For instance, in Figure 4a, atom 8, which is part of the reaction centre, belongs to the aromatic hydroxyl carrier fragment. Likewise, atoms 4 and 6 belong to the anhydride carrier fragment. Thus, the template can be generated with the aromatic hydroxyl and anhydride carrier fragments on the LHS, and corresponding product atom fragments on the RHS. Mapping indices are preserved, with the template reflecting acetylation of the aromatic hydroxyl carrier fragment. In cases where the reaction centre involves multiple carrier fragments in the same molecule, the Bellman-Ford algorithm implemented in RDKit is used to compute the shortest path between fragments, and a larger representative fragment is generated as part of the template. More details on template generation are provided in the Supporting Information Section S3.

Step 10 corresponds to template application and impurity prediction. As template species are carrier fragments, they can be aligned with respective query species, causing the desired transformation into impurities based on the atom mapping. In Figure 4b, the aromatic hydroxyl carrier fragment corresponds to either query species **3** (i) or **1** (ii), whilst the anhydride fragment corresponds to **2**. Applying the template to respective query species leads to impurity **5** and 4-aminophenyl acetate as well as acetic acid in both cases. Therefore, each analogue reaction can lead to one or more impurity reactions via a template depending on carrier fragments in the template and combinations of possible query species containing these carrier fragments.

However, templates are not always successfully applied. In Figure 4c, two aromatic hydroxyl carrier fragments are involved within the same analogue species whilst query species 1 only contains one. This causes the template application to fail as there are too many reacting fragments. In other cases, involved carrier fragments in the same analogue species may be too far apart, resulting in the final fragment unable to be aligned with the query species. After accounting for template failure, 9,980 reactions remain, 4.2% of the original analogue reaction set.

Finally, suggested impurities and impurity reactions are cleaned and filtered in step 11. This includes removing duplicate impurity reactions from the same analogue reaction, impurity reactions that do not have a transformation of interest (as is the case in Figure 4d, which suggests query species again), impurity reactions containing radicals, and self-reactions (without reagents and catalysts). This leaves 9,355 reactions or 4% of the original set.

Impurity ranking

The impurity ranking portion of the workflow consists of steps 12, 13 and 14 in Figure 1, and aims to rank suggested impurities considering reaction conditions, relevance scores and number of hits. Figure 5 illustrates only steps 12 and 13 in the context of the paracetamol synthesis, giving relevant examples. The reader is referred to Results and Discussion for more details on step 14.

Step 12 involves calculating Morgan fingerprint similarities between analogue and query species in suggested impurity reactions. This is based on the understanding that the most relevant analogue reactions will contain analogue species that not only consist of common reacting carrier fragments respective to query species but are also highly similar when comparing overall respective molecular structures. Fingerprints can capture a wider molecular context, at the expense of interpretability. In particular, ECFP (Extended Connectivity Fingerprints) or Morgan fingerprints[46] encode heavy atoms in circular layers up to a specified diameter, and are widely employed.[47] In this case, for each analogue reaction, 1024-bit ECFP4 fingerprints (radius of 2) were used to characterize molecules, and dice similarities[48] were computed between analogue reactants, reagents and respective query species in the suggested impurity reaction using bespoke RDKit functions. These similarities were averaged, with the final value reflecting the **relevance** of the analogue reaction with respect to each suggested impurity reaction, as shown in Figure 5a. The choice of using dice similarity was driven by the fact that it was easy to implement for circular fingerprints in RDKit and also by the fact that it is one of the best similarity measures alongside the Tanimoto index, cosine coefficient and Soergel distance, as verified by Bajusz et al.[48]

Step 13 is a crucial step and attempts to account for reaction conditions in filtering away unrealistic impurities. All impurity reactions (including main product) are assessed, with respect to the three metrics: i) **max relevance**, which indicates the highest relevance score across all analogue reactions that suggest the reaction, ii) **number of hits**, which indicates the number of analogue reactions that suggest the reaction, replicated across the temperature range indicated for each reaction record (if present) and iii) **temperature range**, which is identified by the 5^{th}–95^{th} percentile of the top 10% most relevant analogue reactions that suggested the reaction.

The user can specify a temperature range, but if this is absent, an assumption is made that impurity reactions should occur within the calculated temperature range for the main product. To this end, records that are missing temperature were

Research Article
doi.org/10.1002/cmtd.202200062
Chemistry Europe
European Chemical Societies Publishing

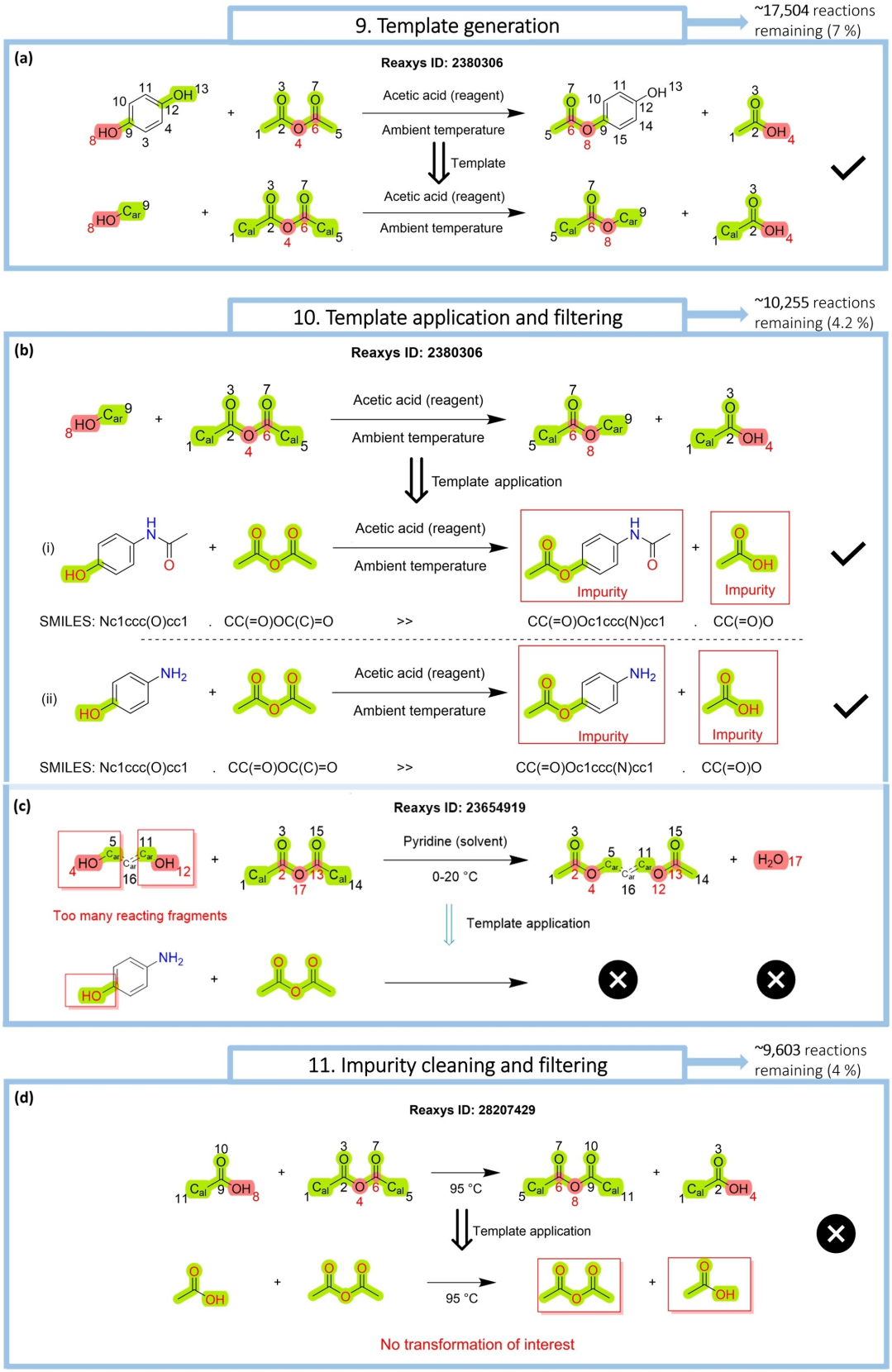


Figure 4. Steps 9 to 11 of the impurity prediction workflow with illustrative examples on paracetamol synthesis. (a) to (d) are analogue reactions discussed in the text, with Reaxys® IDs indicated on the top. A tick or a cross is indicated on the right, representing a valid (kept) and invalid example (filtered out) respectively.

Research Article
doi.org/10.1002/cmtd.202200062

Chemistry Europe
European Chemical Societies Publishing

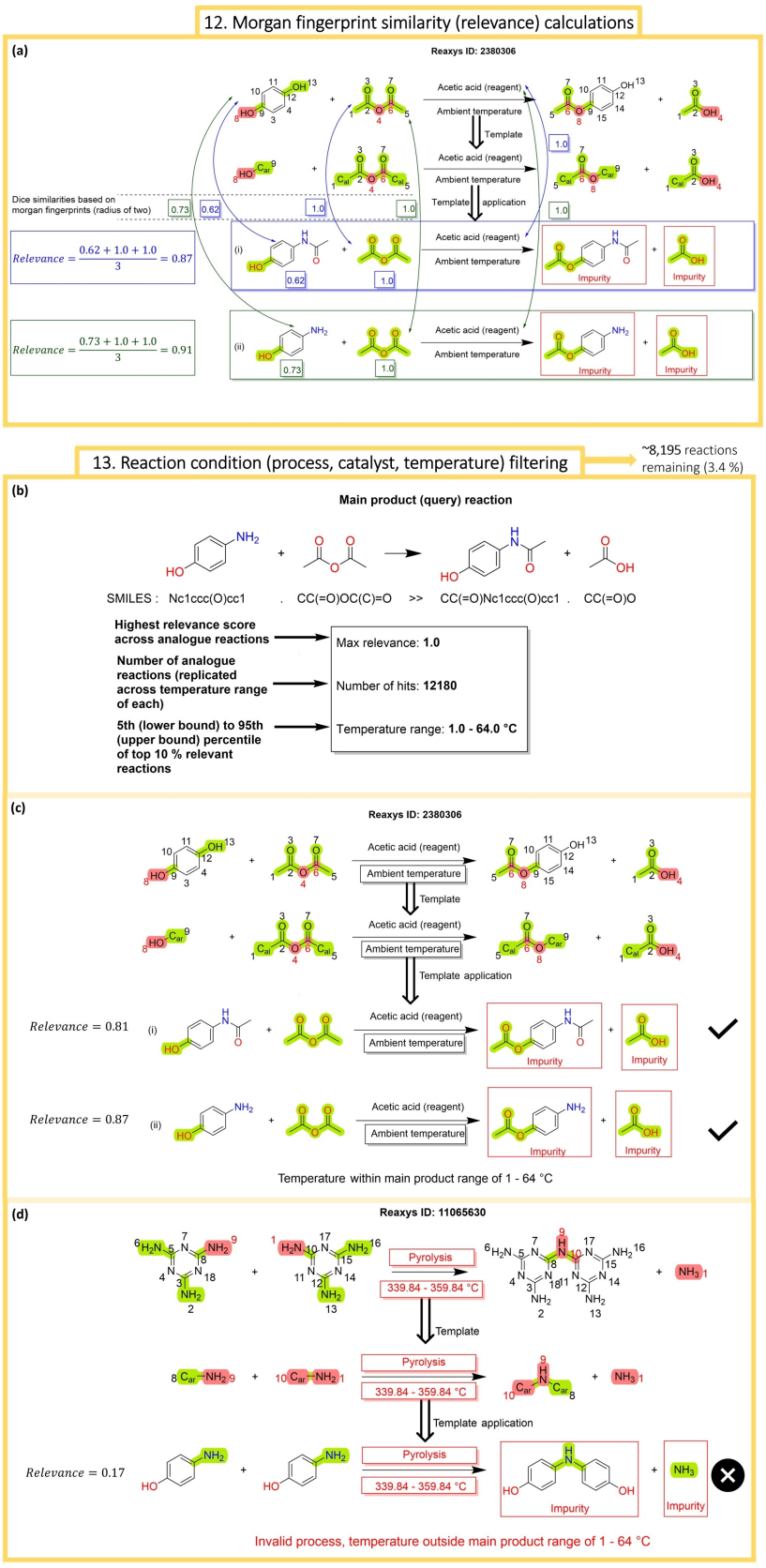


Figure 5. Steps 12 and 13 of the impurity prediction workflow with illustrative examples on paracetamol synthesis. (a) to (d) are analogue reactions discussed in the text, with Reaxys® IDs indicated on the top. A tick or a cross is indicated on the right, representing a valid (kept) and invalid example (filtered out) respectively.

not considered and were removed. It is also assumed that impurity reactions that require special processes (pyrolysis, gasification, enzymatic, bacterial, irradiation, sonication, electrochemical and others) and catalysts, if not specified by the user, are invalid. Solvents, pressure, and reaction time are also important conditions to consider but this is relegated to future work.

For example, in Figure 5b the main product temperature range for paracetamol synthesis was calculated as 1–64 °C and no catalysts or special processes were specified. In the case of Figure 5c, the impurity reaction is deemed valid it occurs within (overlaps with) the temperature range of the main product (ambient temperatures are assumed here to be 20 °C), and does not require special processes or any catalysts. On the other hand, the impurity reaction shown in Figure 5d is invalid because it requires pyrolysis and the identified temperature range 339–359 °C is far higher than the expected main product temperature range. Further examples of invalid reactions are shown in the Supporting Information Figure S5. The final step, which includes ranking the filtered impurity reaction list is elaborated in more detail in Results and Discussion and respective case studies.

Results and Discussion

The workflow highlighted in Method Development was applied to three case studies of increasing complexity: paracetamol synthesis (see the Paracetamol Case Study section), agomelatine synthesis (see the Agomelatine Case Study section) and lersivirine synthesis (see the Lersivirine Case Study section). With steps 1 to 13 elaborated previously, the final step of the workflow involves ranking the condition-filtered impurity reaction list by maximum relevance and number of hits. Impurities with only one hit were removed, as this was judged to be insufficient justification.

Paracetamol case study

Key results for the paracetamol case study are shown in Figure 6, which illustrates the query reaction inputted to the workflow, the main product reaction and the top two ranked impurity reactions. Relevant carrier fragments are highlighted in green throughout.

As shown in Figure 6b, the reaction yielding main product **3** is ranked highest in maximum relevance score (1.0) and also has the greatest number of hits (12180), as expected given the exact match to the query reaction. The main impurity highlighted in literature is **5** due to an overreaction of **3** with **2**. After accounting for the main product temperature range of 1–64 °C, the desired impurity reaction leading to **5** is ranked second in Figure 6c, with a maximum relevance score of 0.87 and 830 hits. The analogue reaction (2390306) with the maximum relevance score has been shown repeatedly throughout the workflow in Figures 2a, 3a, 3e, 3g, 4a, 4b, 5a, 5c.

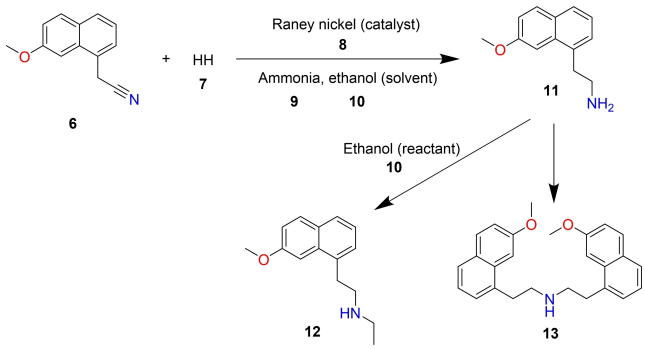
The impurity reaction ranked first features an acetylation of just the aromatic hydroxyl group, although this can be questioned due to the I) the likely reduced concentrations of starting material **1** after formation of main product **3** and II) the lower nucleophilicity of alcohols relative to amines, shown by the reduced hits (830 against 12180 for **3**). It is important to note that the workflow is designed such that all potential impurities are suggested without consideration of yields or concentrations of reacting species. Consideration of these factors would be challenging and require more data than is currently available in reaction databases. However, temperature ranges for both the main product and impurities can be ascertained, and therefore could indicate potential design spaces for optimization to minimise impurity formation.

Agomelatine case study

Agomelatine (N-[2-(7-methoxy-1-naphthyl)ethyl]acetamide) represents a new class of antidepressants.[49] There are several steps to its synthesis, see Scheme 2, where the first contains significant impurities as reported by Liu et al.[49] The synthesis involves hydrogenation of (7-methoxy-1-naphthyl)acetonitrile **(6)** forming an amine main product (intermediate) **11**. The workflow should be able to predict both the disubstituted impurity **13** and the ethylated impurity **12**.

As a point of reference, a comparison was drawn to existing state-of-the-art reaction predictors widely adopted by the community such as ASKCOS and IBM RXN. Results can be seen in Figure S6 in the Supporting Information; in both cases, the correct impurities were not suggested.

Results from the proposed workflow are shown in Figure 7. In addition to specifying the main product reaction outlined in Scheme 2 (query species **6**, **7**, **9**, **10** and **11**), catalyst **8** was also supplied to aid in reaction condition filtering. As shown in Figure 7b, the reaction yielding main product **11** via hydrogenation is ranked highest in maximum relevance score (1.0) and also has the greatest number of hits



Scheme 2. Reaction scheme for agomelatine synthesis. (7-methoxy-1-naphthyl)acetonitrile (**6**) reacts with hydrogen (**7**) with Raney nickel (**8**) as the catalyst in ammonia (**9**) and ethanol (**10**) to form the hydrogenated amine main product (intermediate) (**11**). The disubstituted compound (**13**) is formed as the main impurity, and the amine group of the intermediate can also be ethylated giving impurity (**12**). Note that reactions are not balanced.

Research Article
doi.org/10.1002/cmtd.202200062
Chemistry Europe
European Chemical Societies Publishing

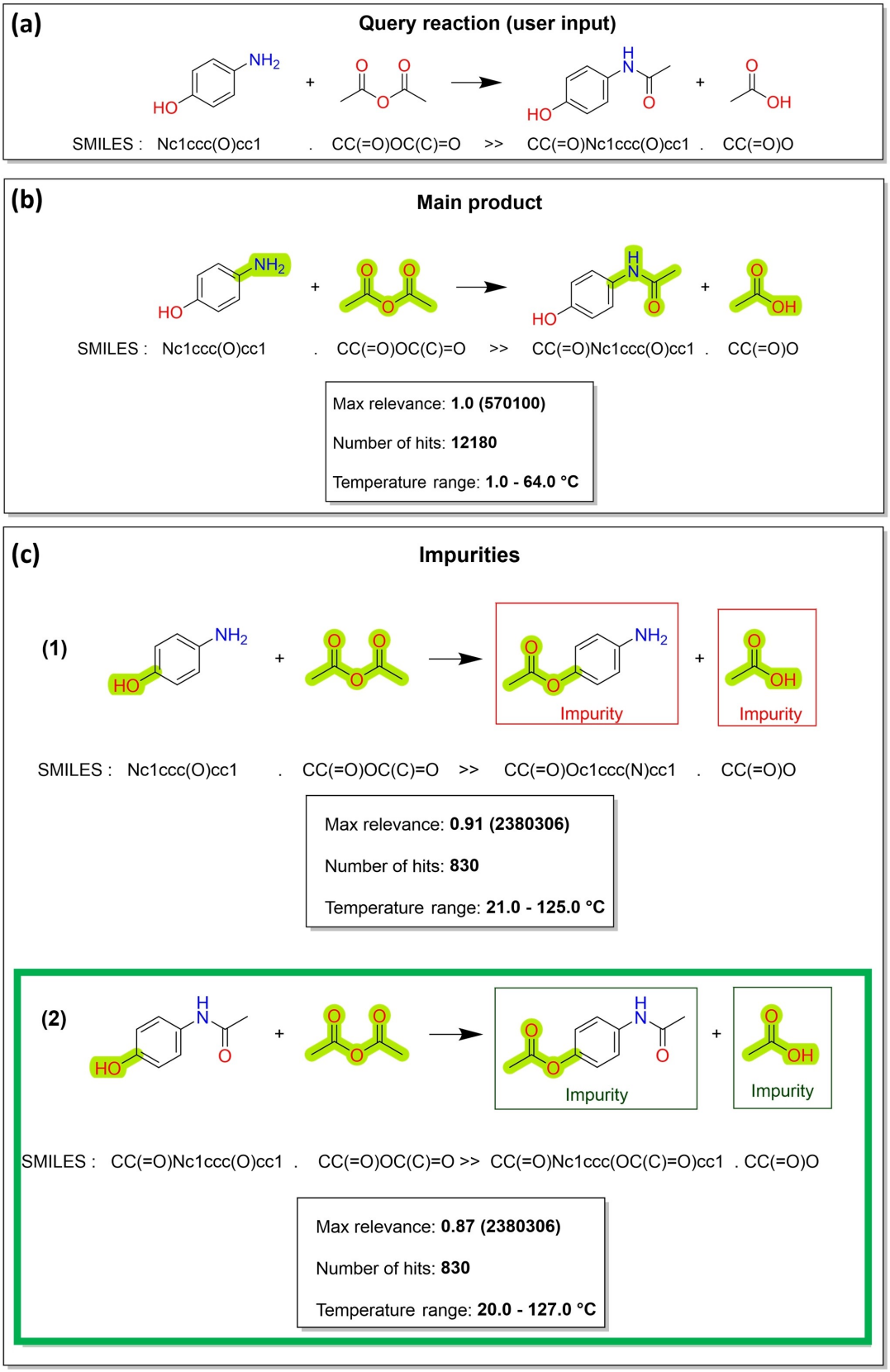


Figure 6. Summary of results for paracetamol case study: (a) Query reaction; (b) Main product reaction; (c) Top two ranked impurity reactions based on maximum relevance. Each reaction is captioned with a box containing the max relevance score with associated Reaxys® ID in brackets, number of hits and temperature range. The literature-identified impurity reaction is ranked second and is boxed in green. Relevant carrier fragments are highlighted in green.

Research Article
doi.org/10.1002/cmtd.202200062
Chemistry Europe
European Chemical Societies Publishing

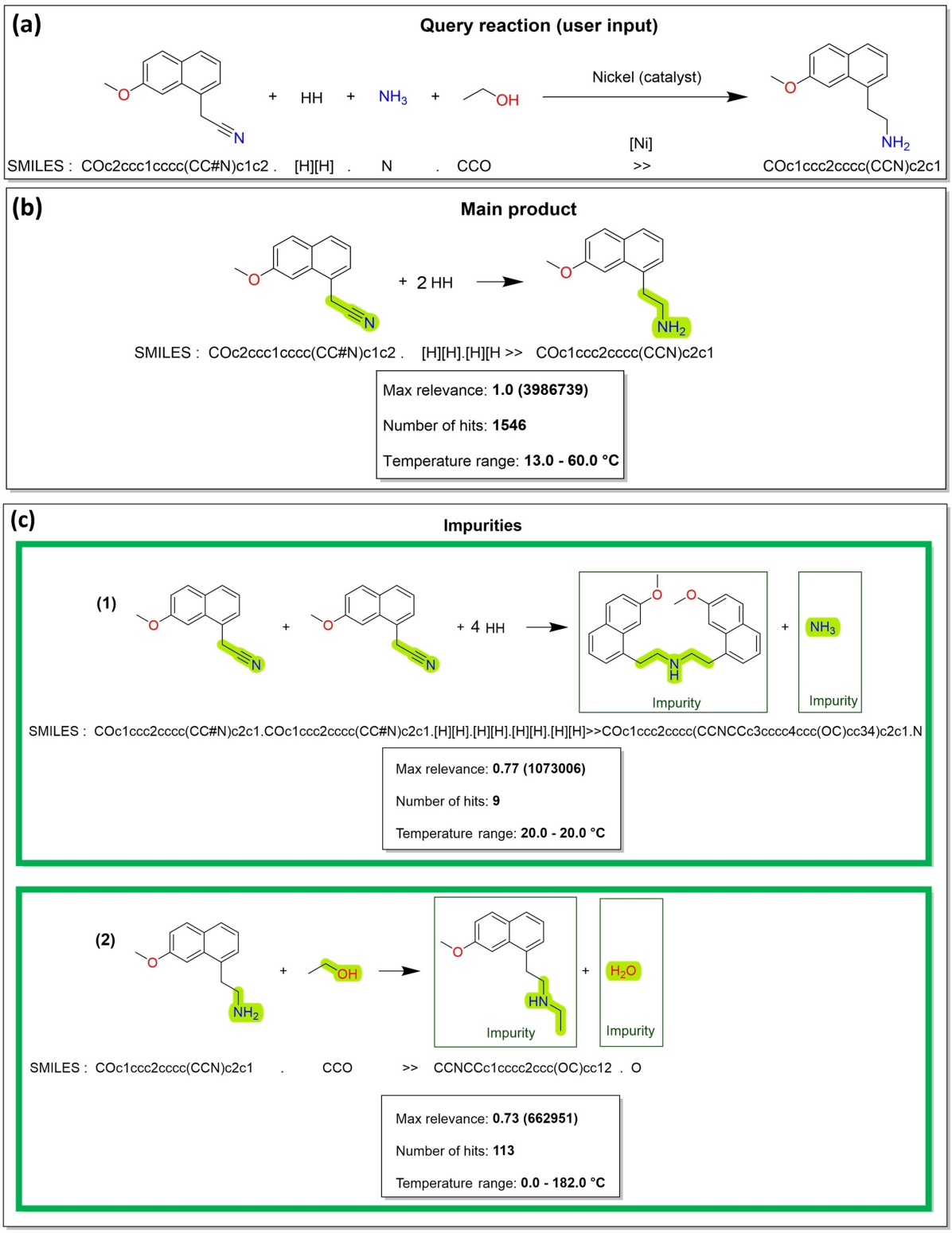


Figure 7. Summary of results for agomelatine case study: (a) Query reaction; (b) Main product reaction; (c) Top two ranked impurity reactions based on maximum relevance. Each reaction is captioned with a box containing the max relevance score with associated Reaxys® ID in brackets, number of hits and temperature range. The literature-identified impurity reactions are both suggested and boxed in green. Relevant carrier fragments are highlighted in green.

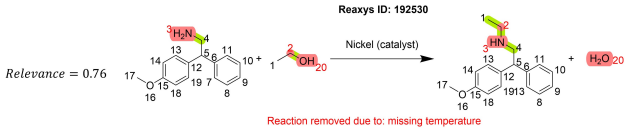
(1546), as expected. The most relevant analogue reaction (Reaxys® ID 3986739) is an exact match. After accounting for the main product temperature range of 13.0–60 °C, the desired impurity reaction leading to disubstituted impurity **13** (and ammonia as a smaller by-product) is ranked first in Figure 7c with a maximum relevance score of 0.77 and only 9

hits. The low number of hits suggests the potential difficulty in generating an adequate training set for conventional ML approaches to successfully predict this impurity. This also illustrates the importance of the maximum relevance score in identifying the most relevant impurities, in contrast to number of hits which need not always be reliable.

The interpretability of the workflow is demonstrated as the associated analogue reaction (1073006) with the highest relevance score is shown in Figure 8a together with the extracted template in Figure 8b and suggested impurity reaction leading to disubstituted impurity **13** in Figure 8c. Even though the analogue species differs slightly with respect to query species **6** (aromatic chlorine is present and ether is missing), the reaction also requires **10** and **8**, featuring conjugation and hydrogenation of two acetonitrile carrier fragments that lead to **13** as well as ammonia.

Similarly, ethylated impurity **12** is suggested in the impurity reaction **ranked second** in Figure 7c, with a maximum relevance score of 0.73 and 113 hits. Analogue reaction 662951, which has the corresponding highest relevance score is shown in Figure 9a, together with the extracted template in Figure 9b and the suggested impurity reaction in Figure 9c.

Another important feature of the workflow is the ability to trace relevant analogue reactions that were removed for a certain reason. For instance, Scheme 3 shows the analogue reaction (Reaxys® ID 192530) leading to a similar ethylated impurity with a higher relevance score of 0.76 compared to the analogue reaction after condition filtering shown in Figure 9 which has a relevance of 0.73. Despite this, the reaction was removed due to a missing temperature record. Thus, the interpretability in each step facilitates the diagnosis of blind spots in the workflow and highlighting of missing data.



Scheme 3. Reaction scheme of a highly relevant (0.76) ethylation reaction that forms impurity **12** and was filtered out due to missing temperatures. Relevant carrier fragments are highlighted in green, and atoms involved in the reaction centre are highlighted in red.

Lersivirine Case Study

The final case study involves synthesis of lersivirine (5-{[3,5-diethyl-1-(2-hydroxyethyl)-1-pyrazol-4-yl]oxy}isophthalonitrile), a drug for HIV treatment. The final stage of the synthesis has been outlined by Codina et al., see Scheme 4.[50] 5-(2-oxo-1-propanoylbutoxy)isophthalonitrile **(14)** reacts with 2-hydrazinoethanol **(15)** to yield lersivirine **(17)** and water **(18)**. The impurity reaction involves esterification of lersivirine by ethanoic acid **(19)**, forming impurity **(19)**. This impurity was also detected in an experimental set-up in the CARES lab in Singapore.

Figure S7 in the Supporting Information shows the impurity results using ASKCOS and IBM RXN; in both cases, the desired

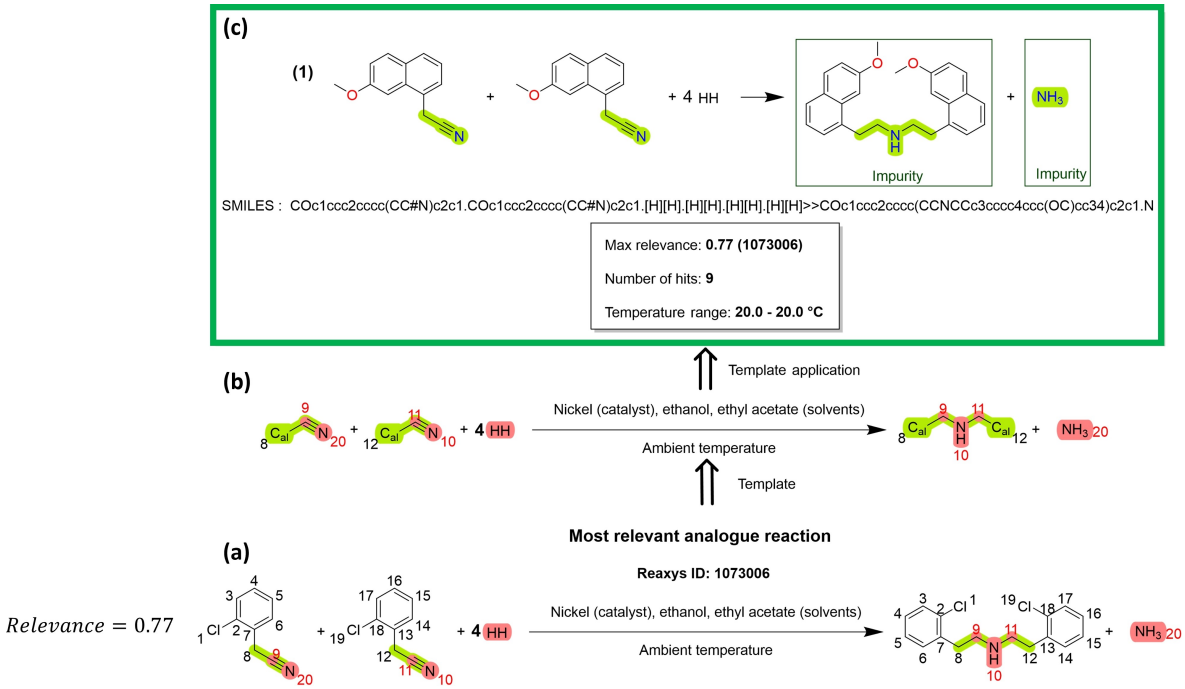


Figure 8. Most relevant analogue reaction leading to formation of disubstituted impurity **13**: (a) Balanced, mapped analogue reaction. Reaxys® ID is shown, along with reaction conditions, and relevance score on the left; (b) Extracted template from reaction; (c) Impurity reaction suggested by template application, captioned with a box containing the max relevance score with associated Reaxys® ID in brackets, number of hits and temperature range. Relevant carrier fragments are highlighted in green in (a), (b) and (c), and atoms involved in the reaction centre are highlighted in red in (a) and (b).

Research Article
doi.org/10.1002/cmtd.202200062
Chemistry Europe
European Chemical Societies Publishing

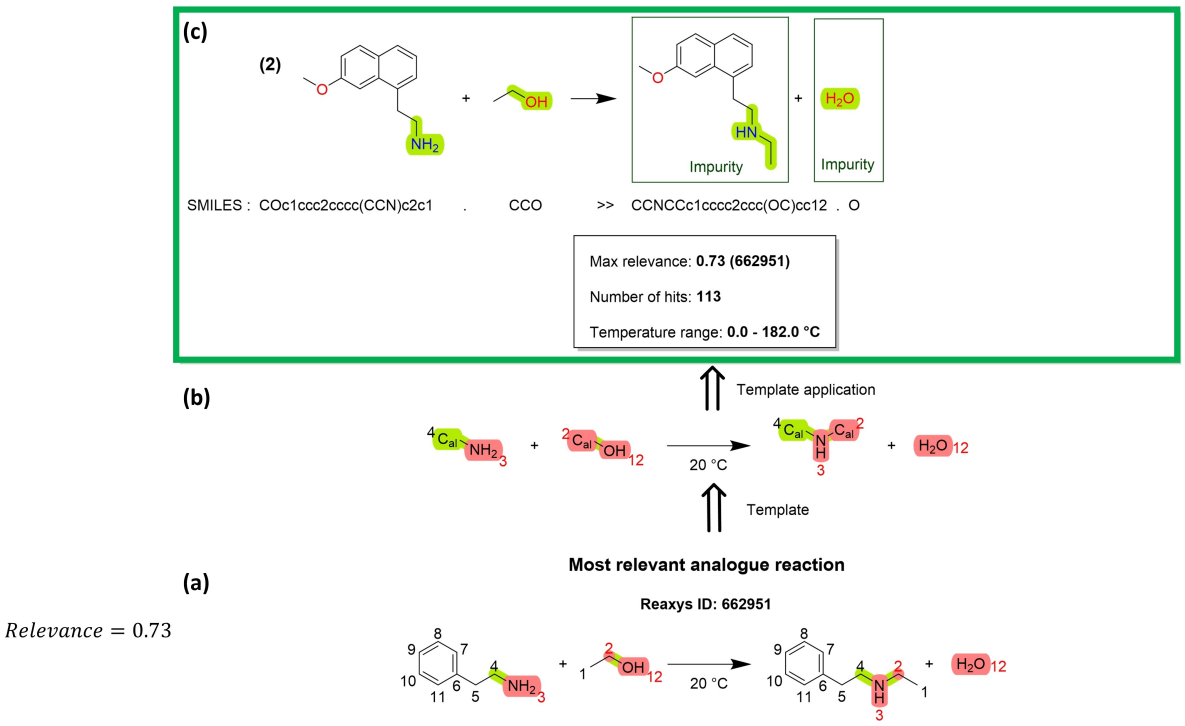
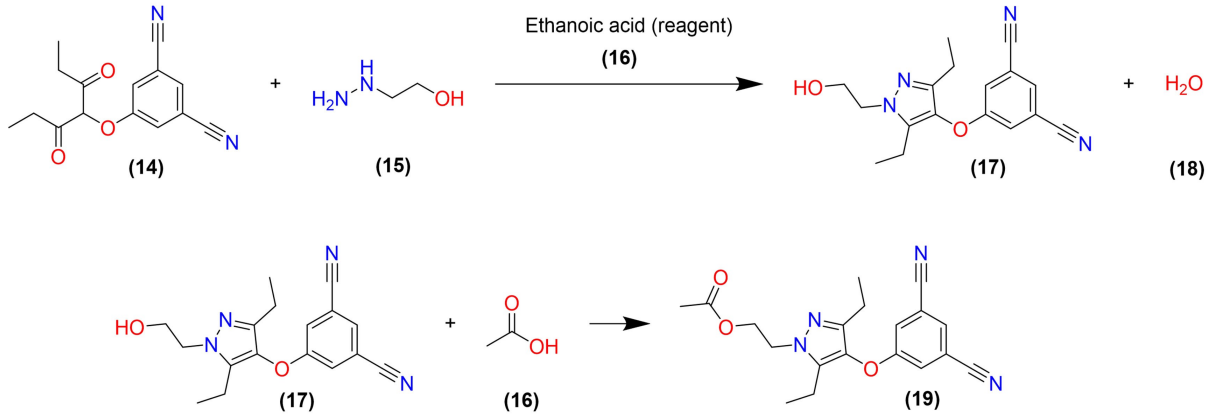


Figure 9. Most relevant analogue reaction leading to formation of ethylated impurity **12**: (a) Balanced, mapped analogue reaction; (b) Extracted template from reaction; (c) Impurity reaction suggested by template application, captioned with a box containing the max relevance score with associated Reaxys® ID in brackets, number of hits and temperature range. Relevant carrier fragments are highlighted in green in (a), (b) and (c), and atoms involved in the reaction centre are highlighted in red in (a) and (b).



Scheme 4. Reaction scheme for final stage of lersivirine synthesis. 5-(2-oxo-1-propanoylbutoxy)isophthalonitrile (**14**) reacts with 2-hydrazinoethanol (**15**), yielding lersivirine (**17**) and water (**18**). Impurity (**19**) is formed due to esterification of lersivirine by ethanoic acid (**16**). Note that reactions are not balanced.

impurity was not suggested. Figure 10 illustrates the results of the workflow when inputting the main product reaction in Scheme 4 (query species **14**, **15**, **16**, **17**, **18**). The reaction yielding main product **17** is ranked highest in maximum relevance score (0.91) with 139 hits as expected. After accounting for the main product temperature range of 20–50 °C, the desired impurity reaction (boxed in green) that leads to esterified impurity **19** (and **18** as a smaller by-product) is **ranked second** in Figure 10c with a maximum relevance score of 0.71 and 331 hits.

Analogue reaction 63199, which has the corresponding highest relevance score is shown in Figure 11a together with the extracted template in Figure 11b and the suggested impurity reaction in Figure 11c.

The impurity reaction ranked first in Figure 10c features esterification of **15**, which, like the paracetamol case study, could be questioned due to its reduced concentrations after formation of main product **17**.

Research Article
doi.org/10.1002/cmtd.202200062
Chemistry Europe
European Chemical Societies Publishing

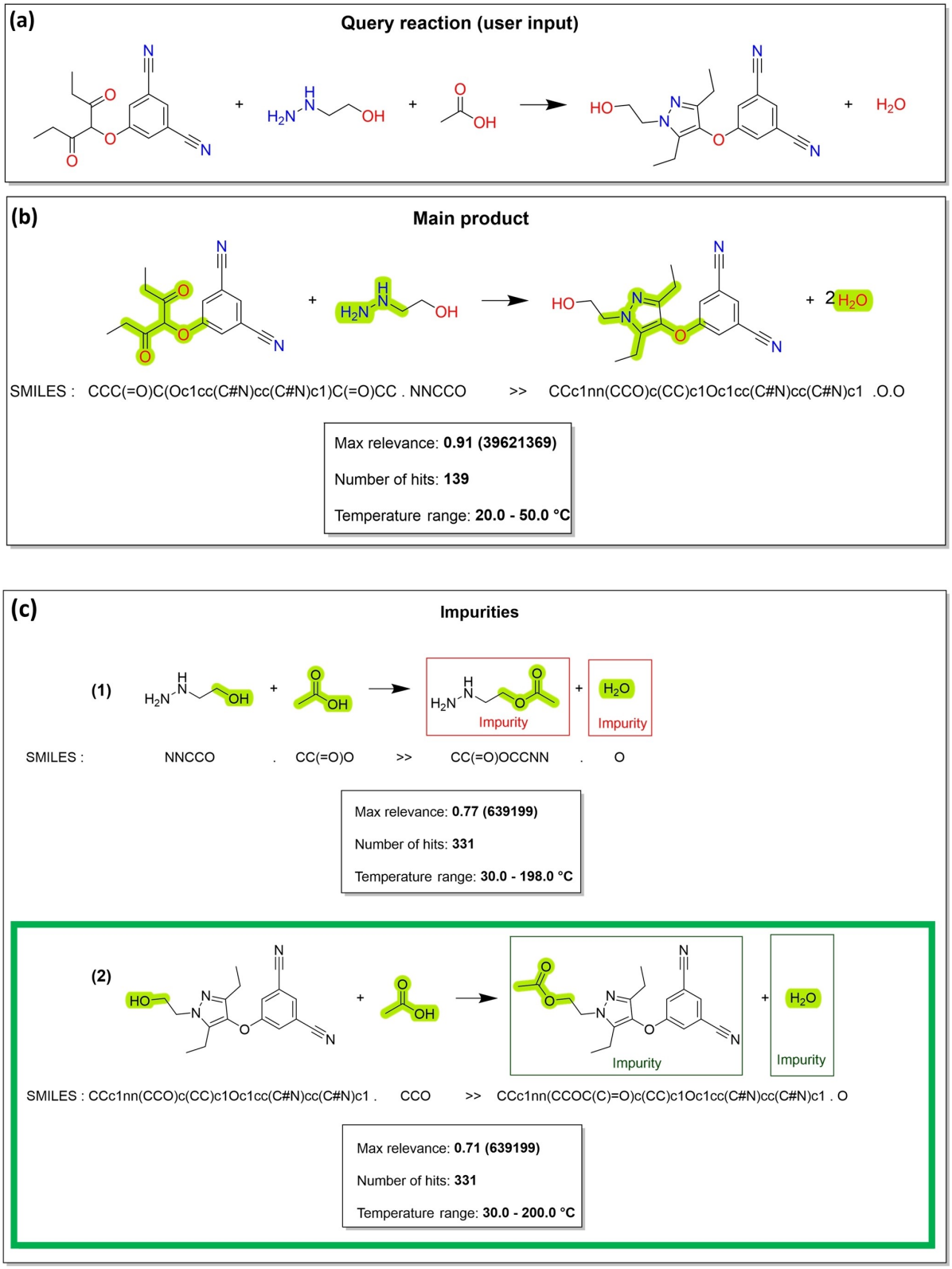


Figure 10. Summary of results for lersivirine case study: (a) Query reaction; (b) Main product reaction; (c) Top two ranked impurity reactions based on maximum relevance. Each reaction is captioned with a box containing the max relevance score with associated Reaxys® ID in brackets, number of hits and temperature range. The literature-identified impurity reaction is suggested and boxed in green. Relevant carrier fragments are highlighted in green.

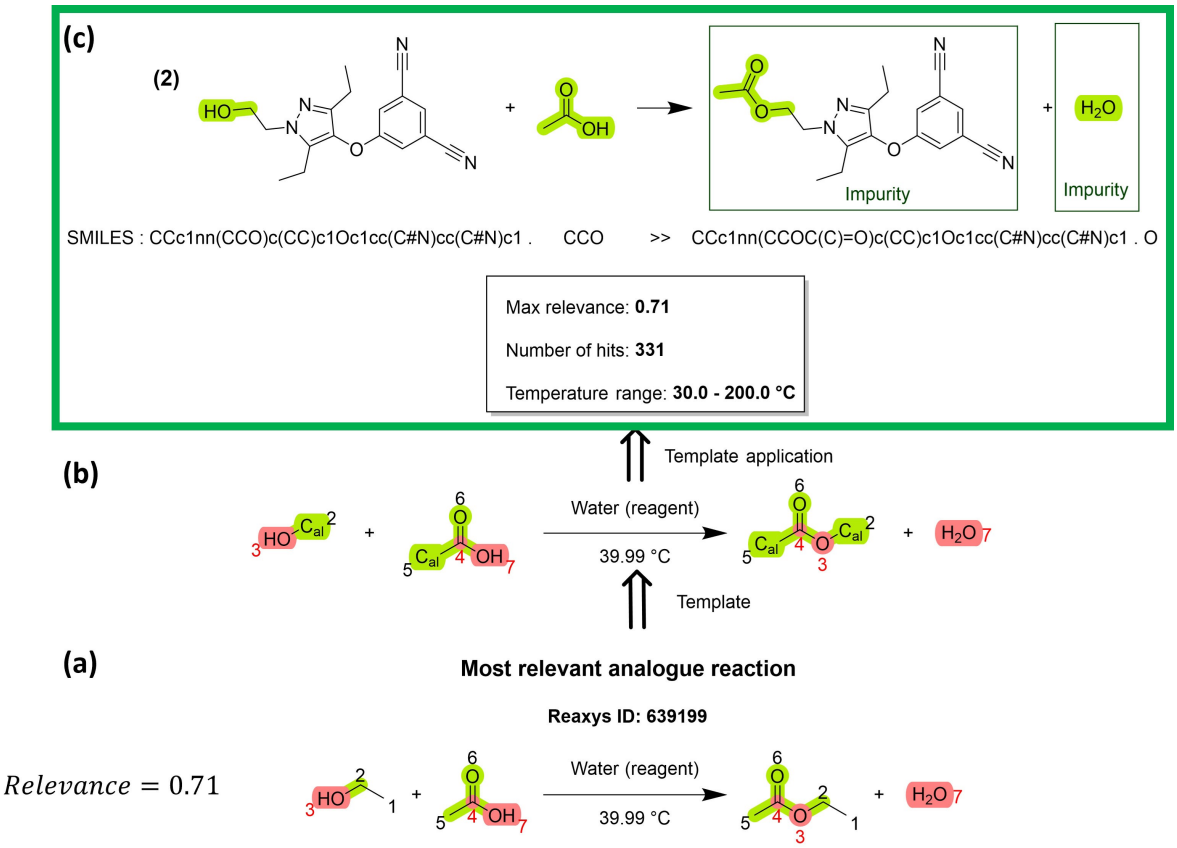


Figure 11. Most relevant analogue reaction leading to formation of esterified impurity **19**: (a) Balanced, mapped analogue reaction; (b) Extracted template from reaction; (c) Impurity reaction suggested by template application, captioned with a box containing the max relevance score with associated Reaxys® ID in brackets, number of hits and temperature range. Relevant carrier fragments are highlighted in green in (a), (b) and (c), and atoms involved in the reaction centre are highlighted in red in (a) and (b).

Conclusions

This work aims to present an automated impurity prediction workflow based on data mining large chemical reaction databases that is interpretable and transparent. Existing reaction predictors are catered towards main product reaction, and impurity prediction has rarely been tackled explicitly. Additionally, many predictors are black-box and not easy to interpret. To this end, an automated and modular 14-step workflow was developed using Python and RDKit, split into four modules: I. Data mining, II. Data processing, III. Impurity prediction and IV. Impurity ranking. Based on user-supplied query species, carrier fragments expanded from functional groups are extracted, and a list of analogue species for each fragment is retrieved. Analogue reactions containing these analogue species are retrieved by data mining Reaxys®, are cleaned, balanced and used to extract templates. Application of these templates to query species can suggest impurity reactions which are further filtered by reaction conditions and ranked by fingerprint similarity (relevance) as well as number of hits.

The workflow was successfully applied to three case studies: paracetamol synthesis, agomelatine synthesis, and lersivirine synthesis. In all cases, literature-identified impurities were suggested within the top two outcomes, and each analogue reaction could be traced to a Reaxys® reaction record. Additionally, highly relevant rejected reactions could be retrieved together with the reason for their removal.

Nonetheless, there are potential changes in future work that could be made to improve the functionality of the reported tool. While impurity awareness is important, it would be more useful to understand yields of impurities depending on concentrations of starting materials (e.g., distinguishing between major and minor impurities). Additionally, the workflow assumes that Reaxys® data is inherently correct which may not always be true, and it is also sensitive to missing condition data such as temperature. On average, only 50–60% of final analogue reaction lists contained a temperature record, which can lead to removal of important impurities (although this did not adversely impact results in the case studies). Consideration of the effects caused by other reaction components and conditions such as solvents, pressure and reaction times would also improve the quality of the suggested outcomes. Ultimately, there is a need to apply the workflow to a larger variety of case studies to assess how well it copes with reactions featuring more complex chemistries and multiple steps (e.g., ring forming, ring breaking, regioselective reactions, protective and depro-

tective group interactions). Assembling an exhaustive list of case studies where impurities are known from literature remains a challenge, requiring both manual and text mining approaches. However, this would allow for the use of the top-*k* accuracy metric (fraction of samples for which, given the reactants and main products, the recorded impurities are among the top-*k* predictions) to evaluate the proposed workflow in a statistically significant manner. Regardless, this work can serve as a benchmark and as an example for how well-curated prior reaction data can aid in the development of more sophisticated algorithms for reaction prediction with easily interpretable results.

Acknowledgements

This study was co-funded by Pharma Innovation Partnership in Singapore (PIPS) through project C4. This work was supported by National Research Foundation (NRF), Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) program: Cambridge Centre for Advanced Research and Education in Singapore Ltd (CARES, C4T project) (AAL & ZG). PhD scholarship of AA is co-funded by Chemical Data Intelligence (CDI) Pte Ltd and CARES Ltd within the framework of the project C4T.

We gratefully acknowledge collaboration with RELX Intellectual Properties SA and their technical support, which enabled us to mine Reaxys®. Copyright © 2020 Elsevier Limited except certain content provided by third parties. Reaxys® is a trademark of Elsevier Limited. Reaxys® data were made accessible to our research project via the Elsevier R&D Collaboration Network.

Reaxys® molecule and reaction data are accessible to users via Elsevier. Although the specific results from the case studies can only be obtained by using the Reaxys dataset, this workflow can, in principle, be applied to any custom dataset.

Conflict of Interest

Z.G. and A.A.L. are co-founders of Chemical Data Intelligence (CDI) Pte Ltd (cdi-sg.com), which was set-up to commercially exploit the chemical data networks.

Data Availability Statement

The data that support the findings of this study are available from Elsevier. Restrictions apply to the availability of these data, which were used under license for this study. Data are available at https://www.reaxys.com/#/login with the permission of Elsevier.

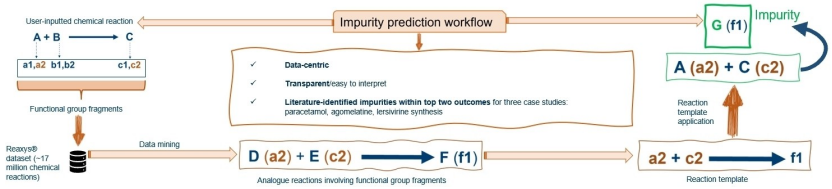
Keywords: Cheminformatics · Data mining · Impurity prediction · Reaction databases · Reaction prediction

[1] F. Qiu, D. L. Norwood, *J. Liq. Chromatogr. Relat. Technol.* **2007**, *30*, 877–935.
[2] Reaxys – An expert-curated chemistry database, https://www.elsevier.com/solutions/reaxys, Accessed September 30, 2021.
[3] S. Johansson, A. Thakkar, T. Kogej, E. Bjerrum, S. Genheden, T. Bastys, C. Kannas, A. Schliep, H. Chen, O. Engkvist, *Drug Discovery Today Technol.* **2020**, *32–33*, 65–72.
[4] A. Thakkar, S. Johansson, K. Jorner, D. Buttar, J. L. Reymond, O. Engkvist, *React. Chem. Eng.* **2021**, *6*, 27–51.
[5] O. Engkvist, P. O. Norrby, N. Selmi, Y. Hong Lam, Z. Peng, E. C. Sherer, W. Amberg, T. Erhard, L. A. Smyth, *Drug Discovery Today* **2018**, *23*, 1203–1218.
[6] T. J. Struble, J. C. Alvarez, S. P. Brown, M. Chytil, J. Cisar, R. L. DesJarlais, O. Engkvist, S. A. Frank, D. R. Greve, D. J. Griffin et al., *J. Med. Chem.* **2020**, *63*, 8667–8682.
[7] S. Szymkuć, E. P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2016**, *55*, 5904–5937; *Angew. Chem.* **2016**, *128*, 6004–6040.
[8] E. J. Corey, A. K. Long, S. D. Rubenstein, *Science (80-).* **1985**, *228*, 408–418.
[9] E. J. Corey, W. Todd Wipke, *Science (80-).* **1969**, *166*, 178–192.
[10] W. L. Jorgensen, E. R. Laird, A. J. Gushurst, J. M. Fleischer, S. A. Gothe, H. E. Helson, G. D. Paderes, S. Sinclair, *Pure Appl. Chem.* **1990**, *62*, 1921–1932.
[11] H. Satoh, K. Funatsu, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 173–184.
[12] C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green, K. F. Jensen, *ACS Cent. Sci.* **2017**, *3*, 434–443.
[13] M. H. S. Segler, M. P. Waller, *Chem. A Eur. J.* **2017**, *23*, 5966–5971.
[14] C. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay, K. F. Jensen, *Chem. Sci.* **2019**, *10*, 370–377.
[15] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas, A. A. Lee, *ACS Cent. Sci.* **2019**, *5*, 1572–1583.
[16] M. H. S. Segler, M. P. Waller, *Chem. A Eur. J.* **2017**, *23*, 6118–6128.
[17] P. M. Jacob, A. A. Lapkin, *ChemRxiv.* **2018**, preprint, DOI: 10.26434/chemrxiv.6954908.v1.
[18] D. Fooshee, A. Mood, E. Gutman, M. Tavakoli, G. Urban, F. Liu, N. Huynh, D. Van Vranken, P. Baldi, *Mol. Syst. Des. Eng.* **2018**, *3*, 442–452.
[19] M. A. Kayala, C. A. Azencott, J. H. Chen, P. Baldi, *J. Chem. Inf. Model.* **2011**, *51*, 2209–2222.
[20] P. Sadowski, D. Fooshee, N. Subrahmanya, P. Baldi, *J. Chem. Inf. Model.* **2016**, *56*, 2125–2128.
[21] M. Fujinami, J. Seino, H. Nakai, *Bull. Chem. Soc. Jpn.* **2020**, *93*, 685–693.
[22] D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, A. G. Doyle, *Science (80-).* **2018**, *360*, 186–190.
[23] 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, A. Varnek, *Mol. Inf.* **2021**, *40*, 2100119.
[24] T. Rodrigues, *Drug Discovery Today Technol.* **2019**, *32*, 3–8.
[25] X. Jia, A. Lynch, Y. Huang, M. Danielson, I. Lang'at, A. Milder, A. E. Ruby, H. Wang, S. A. Friedler, A. J. Norquist, J. Schrier, *Nature.* **2019**, *573*, 251–255.
[26] D. P. Kovács, W. McCorkindale, A. A. Lee, *Nat. Commun.* **2021**, *12*, 1695.
[27] A. Toniato, P. Schwaller, A. Cardinale, J. Geluykens, T. Laino, *Nat. Mach. Intell.* **2021**, *3*, 485–494.
[28] G. Landrum, RDKit: Open-source cheminformatics, http://www.rdkit.org, **2020**.
[29] P. Moritz, R. Nishihara, S. Wang, A. Tumanov, R. Liaw, E. Liang, M. Elibol, Z. Yang, W. Paul, M. I. Jordan, I. Stoica, In: *USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, **2018**, 561–577.
[30] R. N. Rao, A. Narasaraju, *Anal. Sci.* **2006**, *22*, 287–292.
[31] S. Soh, Y. Wei, B. Kowalczyk, C. M. Gothard, B. Baytekin, N. Gothard, B. A. Grzybowski, *Chem. Sci.* **2012**, *3*, 1497–1502.
[32] E. S. Salmina, N. Haider, I. V. Tetko, *Molecules.* **2016**, *21*, 1.
[33] T. Sterling, J. J. Irwin, *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337.
[34] P. Ertl, *J Cheminform.* **2017**, *9*, 36.
[35] S. Fritsch, S. Neumann, J. Schaub, C. Steinbeck, A. Zielesny, *J Cheminform.* **2019**, *11*, 37.
[36] R. Hall, G. Godin, RDKit, https://github.com/rdkit/rdkit/tree/master/Contrib/IFG, **2017**, Accessed November 15, 2021.
[37] A. C. Vaucher, P. Schwaller, T. Laino, **2020**, *ChemRxiv.* prepint DOI: 10.26434/chemrxiv.13273310.
[38] NextMove Software | NameRxn, https://www.nextmovesoftware.com/namerxn.html, Accessed November 17, 2021.

[39] JChem Engines | ChemAxon, https://chemaxon.com/products/jchem-engines, Accessed November 18, 2021.
[40] Indigo Toolkit, https://lifescience.opensource.epam.com/indigo/index.html, Accessed November 18, 2021.
[41] S. A. Rahman, G. Torrance, L. Baldacci, S. M. Cuesta, F. Fenninger, N. Gopal, S. Choudhary, J. W. May, G. L. Holliday, C. Steinbeck, J. M. Thornton, *Bioinformatics*. **2016**, *32*, 2065–2066.
[42] P. Schwaller, B. Hoover, J. L. Reymond, H. Strobelt, T. Laino, *Sci. Adv.* **2021**, *7*, eabe4166.
[43] A. Lin, N. Dyubankova, T. Madzhidov, R. Nugmanov, A. Rakhimbekova, Z. Ibragimova, T. Akhmetshin, T. R. Gimadiev, R. Suleymanov, J. Verhoeven, J. K. Wegner, H. Ceulemans, A. Varnek, *Mol. Inf.* **2021**, *41*, 2100138.
[44] W. Jaworski, S. Szymkuć, B. Mikulak-Klucznik, K. Piecuch, T. Klucznik, M. Kaźmierowski, J. Rydzewski, A. Gambin, B. A. Grzybowski, *Nat. Commun.* **2019**, *10*, 1434.
[45] C. W. Coley, W. H. Green, K. F. Jensen, *J. Chem. Inf. Model.* **2019**, *59*, 2529–2537.
[46] D. Rogers, M. Hahn, *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
[47] W. A. Warr, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 557–579.
[48] D. Bajusz, A. Rácz, K. Héberger, *J Cheminform.* **2015**, *7*, 20.
[49] Y. Liu, L. Chen, Y. Ji, *J. Pharm. Biomed. Anal.* **2013**, *81*, 193–201.
[50] A. Codina, A. Derrick, R. Joyce, F. Susanne, WO2013050873A1, **2012**, *Process for the preparation of lersivirine*.

Manuscript received: November 6, 2022

RESEARCH ARTICLE



An automated chemical reaction impurity prediction workflow is proposed, based on data mining chemical reaction databases (Reaxys®). The workflow aims to be data-centric, transparent, and interpretable. It was applied to three case studies: paracetamol, agomelatine, and lersivirine synthesis. In all cases, the literature-identified impurities were suggested within the top two outcomes.

*A. Arun, Dr. Z. Guo, Dr. S. Sung, Prof. A. A. Lapkin**

1 – 19

Reaction Impurity Prediction using a Data Mining Approach