Automated Integration of Structural, Biological and Metabolic Similarities to Sustain Read-Across

Domenico Gadaleta, Azadi Golbamaki Bakhtyari, Giovanna J. Lavado, Alessandra Roncaglioni and Emilio Benfenati

Laboratory of Environmental Chemistry and Toxicology, Department of Environmental Health Sciences, Istituto di Ricerche Farmacologiche Mario Negri – IRCCS, Milan, Italy

Abstract

Read-across (RAX) is a popular data-gap filling technique that uses category and analogue approaches to predict toxicological endpoints for a target. Despite its increasing relevance, RAX relies on human expert judgement and lacks a reproducible and automated protocol. It also only relies on structural similarity for identifying the analogues, while other aspects are often neglected. In this paper, we propose an automated procedure for the selection of analogues for data gap-filling. Analogues were identified with a decision algorithm that integrates three similarity metrics, each considering different toxicochemically relevant aspects (i.e., structural, biological and metabolic similarity). Structural filters based on the presence of maximum common substructures (MCS) and common functional groups were applied to narrow the chemical space for the analogues search. The procedure has been implemented as a workflow in KNIME and is freely available. The workflow provides informative tabular and graphical outputs to support toxicologists and risk assessors in drawing conclusion based on the RAX approach. The procedure has been validated for its predictive power on two datasets related to high-tier in vivo toxicological endpoints, i.e. human hepatotoxicity and drug-induced liver injury (DILI). The validation results gave good accuracy values (i.e., up to 0.79 for the binary hepatotoxicity classification and up to 0.67 for the three-class DILI classification) that were higher than those returned by RAX based on the sole use of structural similarity. Results confirmed the suitability of the procedure as a source of data to support regulatory decision-making.

1 Introduction

In the last 15 years, regulations in the field of chemical safety assessment have changed in that toxicological information for a large number of chemicals needs to be gathered prior to manufacture or import into the EU (EC, 2006). In vivo testing to meet these information requirements was clearly not feasible due to time, costs and the need to sacrifice an unacceptable number of animals. EU’s Registration, Evaluation, Authorisation and restriction of CHemicals (REACH) calls for the use of non-testing approaches to be used in the assessment of chemical substances where vertebrate animal testing should be seen as a last resort (ECHA, 2014).

Among non-testing methods, read-across (RAX) has proven to be an effective and widely used approach to provide toxicological information without the need to undertake animal testing ac cording to the 3Rs principle (Replace, Reduce, Refine) (Russell and Burch, 1959), with clear benefits in terms of money, time and animal savings. RAX is a data-gap filling technique to predict unknown toxicological endpoints for a chemical substance (target) by using the same endpoint information from one or more chemicals that are highly similar to the target (analogue(s)) (Patlewicz et al., 2013; Pradeep et al., 2017; OECD, 2014; ECHA, 2008).

The first step of RAX consists in identifying potential analogue(s) that may serve to fill toxicological data gaps of the target. This can be done using quantitative metrics to evaluate the similarity between the target and potential analogue(s). The following steps comprise gathering relevant data to determine analogue(s) suitability for the RAX. The presence of functional group(s) (e.g. aldehyde, epoxide, ester, specific metal ion) shared with the target, common constituents or chemical classes, similar carbon range numbers or the likelihood of common precursors and/or breakdown products are typical considerations when evaluating analogue(s) in the final RAX reasoning. Based on the retrieved data, one can assess the adequacy of the analogue(s) and use the most suitable ones to fill the data gaps for the target (Patlewicz et al., 2013, 2015, OECD, 2014; ECHA, 2008).

Two approaches for RAX exist. The “analogue approach” is based on a close number of structurally similar substances (usually a single analogue), while the “category approach” relies on a larger number of analogues, often included in the same chemical classes or showing some kind of trend in structures (e.g., an increasing number of carbon chain lengths).

Among non-testing methods, RAX is applied in various regulatory programs such as the OECD High Production Volume Programme (Bishop et al., 2012) and REACH. Indeed, REACH encouraged the possibility of using a category/analogue approach (ECHA, 2008) to address regulatory requirements for chemical substances, to the point that RAX has been used in up to 75% of...
analyzed registration dossiers for at least one endpoint (ECHA, 2014). This percentage is much higher than the ones related to other non-testing methods, e.g. quantitative structure activity relationships (QSARs). This is especially true for higher tier toxicological endpoints (such as reproductive and repeated dose toxicities) that refer to a wide range of adverse effects on different target organs and tissues. RAX proved to be more intuitive for regulators and more appropriate than QSAR/in vitro in appreciating the composite nature of complex in vivo endpoints (Patlewicz et al., 2013).

Despite its increasing use for regulatory purposes, RAX remains largely subjective and relies on human expert judgement for both analogue(s) selection and data interpretation (Patlewicz et al. 2017). This recently prompted European Chemical Agency (ECHA) in publishing their Read-Across Assessment Framework (RAAF) (ECHA, 2015) which complemented already existing regulatory technical guidance from ECHA (ECHA, 2008) and OECD (OECD, 2014). These documents exemplified all relevant aspects that should be evaluated to assure the acceptability of RAX proposals included in the REACH registrations. Despite this, the general lack of a well-defined and systematic decision workflow still hampers a consistent application of RAX (Patlewicz et al., 2017). Another limitation is that RAX traditionally relies on the chemical similarity principle. Indeed, REACH states that chemical structure should be the starting point for the definition of any category/analogous approach (ECHA, 2008). However, toxicity is a complex and composite processes, so that the accuracy of predictions based exclusively on structural similarity are often inadequate in handling complex mechanisms of toxicity (Ball et al., 2016). A disadvantage of the sole use of structural similarity is, for example, the existence of activity cliffs, i.e. group of compounds with high structural similarity but unexpectedly high activity (or property) differences (Cruz-Monteagudo et al., 2014). Ideally, therefore the RAX should be based on a weight of evidence (WoE) assessment of many different pieces of information, i.e. not only structures, but also metabolism, biology and physicochemical properties should be considered to substantiate the similarity between target and analogue(s) (Patlewicz et al., 2014, 2015).

These aspects highlight the need to define an automated RAX framework able to consider both structural and biological profiles of chemicals, still keeping the entire process transparent and easy to understand for regulators (Low et al., 2013). Several attempts are reported in the literature to exploit hazard (Luechtefeld et al., 2018) or High Throughput Screening (HTS) in vitro data to include biological similarity in RAX reasoning (Petrone et al., 2012; Russo et al., 2017; Shah et al., 2016; Grimm et al., 2016; Low et al., 2013). Large databases exist including thousands of HTS assay for a broad range of chemical substances. ToxCast by US EPA and PubChem are well-known examples. Biological phenotyping derived by the combination of HTS assays could serve to characterize the biological profile of target and analogue(s) in terms of a fingerprint that can be used to determine a quantitative biological similarity (Patlewicz et al., 2014). The increased availability of cheminformatics tools represents another possible improvement of traditional RAX. Automated tools exist that can be applied to analogue(s) identification (e.g., QSARToolbox), data retrieval or, if no experimental data are available, for the prediction of physicochemical, toxicokinetic and metabolic parameters for substances under study (Ball et al., 2016). Moreover, the same tools can be used to automatize the evaluation of similarity between target and analogue(s) and other steps of the RAX, facilitating the final expert judgement.

In this study, we present a novel automated workflow for analogue(s) selection for RAX based on a WoE approach that systematically compute and combine three similarity metrics between target and potential analogue(s). Given a target, the workflow automatically lists potential analogue(s) that are selected independently based on three similarity criteria (i.e. structural, biological, and metabolic similarities). A large collection of data retrieved from on-line databases or calculated with cheminformatics tools is provided in the final output of the workflow for aid experts in RAX reasoning. In the end, compound(s) included in multiple similarity lists (e.g., structural and biological lists) are suggested as most suitable analogue(s), and their activity is used to infer the activity of the target chemical. Finally, we present examples to evaluate the suitability of the described procedure to predict high-tier endpoints for the chemicals.

The entire workflow is implemented in KNIME (version 3.4) (Berthold et al., 2008) and made freely available to the scientific community and to regulators to aid expert reasoning and decision-making.

2 Methods

2.1 RAX workflow

An automated procedure for compute similarities between chemicals was implemented as a KNIME workflow (Berthold et al., 2008). The workflow accepts SMILES notation of target chemical as input and is connected to a pre-loaded dataset in which the analogue(s) search is performed (source dataset). Three separate lists of possible analogue(s) are retrieved from the source dataset based on three different and independent methods for compute similarity, i.e. 1) structural (StrS), 2) metabolic (MS) and 3) biological similarity (BS). Chemicals in the source dataset are ranked for each of the three similarity metrics, and then three different lists of top-ranked similar compounds are returned as output. The number of analogue(s) returned in each list can be customized by the user in order to meet specific endpoints requirements (e.g. increase the MS analogues if metabolism is known to be particularly relevant). By default, up to ten analogues are retrieved based on StrS, as a primary requisite for any RAX prediction (ECHA, 2015). Up to five analogues are retrieved based on MS and BS individually. Figure 1 shows the logical scheme of the RAX workflow, while a more detailed depiction of the KNIME implementation is included in Figure S1. The KNIME workflow is freely available for download. The user’s guide is available at the same GitHub link and in the supporting information.

The workflow also offers a rich tabular and graphical (pie-plot) outputs including relevant chemical and toxicological information for both the target and the analogue(s) (e.g., metabolites, number of biological assays with the same outcome, common

---

1 https://www.epa.gov/chemical-research/toxicity-forecasting
3 https://qsartoolbox.org/
4 doi:10.14573/altex.2002281s1
5 https://github.com/DGadaleta88/RAX_tool
6 https://doi.org/10.14573/altex.2002281s2
Fig. 1: Conceptual scheme of the RAX workflow

Similarities between the target and chemicals from the source dataset are computed, so that three independent lists of analogue(s) are obtained. The activity of chemicals included in one (or multiple) list(s) are used to infer the RAX prediction. A preliminary selection of source chemicals can be applied based on the presence of MCS and common functional group(s) (FGs) with the target.

Functional groups) which can be used by a toxicologist to support the RAX process and help regulators in decision making. The pie plots that graphically describe RAX results are created using the bokeh library in Python. Some examples are shown in Figure S2.

2.1.1 Structural similarity

StrS was based on MACCS fingerprints (166 bits) (Durant et al., 2002) that were calculated for both target and analogue(s) by the KNIME implementation of CDK toolkit. MACCS keys are commonly used in calculating structural similarity (Maggiora et al., 2014; Sánchez-Cruz and Medina-Franco, 2018). Bits codify for the presence of a given substructure in the molecule (e.g. a phenyl ring or a functional group). They are useful to disclose analogies in terms of those chemical features relevant for biological and toxicological activity. StrS was computed by means of Tanimoto coefficient (Willet et al., 1998).

In these regards, chemicals are identified as similar if they share a high number of biologically relevant moieties.

2.1.2 Metabolic similarity

“SyGMa metabolite” KNIME implementation was used to simulate metabolites of the target and the analogue(s) SyG Ma (Systematic Generation of possible Metabolites) is a freely available tool to simulate metabolism of chemicals. Metabolic rules implemented in SyGMa are derived combining expert knowledge and empirical analysis of proprietary data (i.e., MDL dataset) and cover the 70% of all known human metabolic reactions. Predictions made by SyGMa are associated to an empirical probability score that identifies more likely metabolic routes and reduce the number of false positives that are often generated by tools based only on expert rules (Ridder et al., 2008).

The tool allowed to calculate both specific metabolites and the type of metabolic pathways in humans that the parent undergone to generate the metabolite. In addition, SyGMa provides a free Python code and a KNIME implementation that allowed an easy integration in the workflow here presented.

Only one cycle of Phase I metabolism was considered for the present simulation. A MS score was calculated based on the number of common and exclusive metabolic pathways of the two compared chemicals.

\[
MS(A, B) = \frac{P(A, B)}{P(A, B) + P(A) + P(B)}
\]

\(P(A, B)\) are metabolic pathways in common between compounds A and B, while \(P(A)\) and \(P(B)\) are metabolic pathways that are present in one of the two compounds that are compared, but not in the other one.

The presence of structurally identical metabolites and/or parent compounds between the two compared chemicals was also reported in the final output of the workflow.

2.1.3 Biological similarity

BS calculation was based on HTS assays from PubChem. Assays were used to compile biological binary fingerprints for comparing target and analogue(s), with each bit of the fingerprint codifying the outcome of a specific assay. The REST version of the PUG (Power User Gateway) interface for accessing PubChem data was implemented in KNIME to automatically retrieve assays information for

7 http://bokeh.pydata.org/
8 https://cdk.github.io/
target and analogue(s)\(^9\). In particular, “Active” or “Probe” outcomes were flagged with 1 while “Inactive” with 0. “Inconclusive” and “Unspecified” outcomes were ignored. Duplicate assays resulting in different outcomes for the same chemical (e.g., “Active” and “Inactive”) were considered “Inconclusive”.

BS was calculated as proposed by Russo et al. (2017):

\[
BS(A, B) = \frac{\left| A_a \cap B_a \right| + \left| A_i \cap B_i \right| \cdot w}{\left| A_a \right| + \left| A_i \right|}
\]

\(A_a\) and \(A_i\) are active and inactive assays for the target \(A\), while \(B_a\) and \(B_i\) are active and inactive assays for the analogue \(B\). \(\cap\) indicates assays in common between the two compounds under investigation, while \(w\) is a weight assigned to common inactive assays that accounts the ratio of active to inactive bits in the target compound biological fingerprint:

\[
w = \frac{A_a}{A_i}
\]

The weight ranges from 0 to 1, giving to inactive data a fraction of the weight of active data. This variable was adopted by Russo et al. (2017) because the biosimilarity relies on active data more than inactive data, due to the extremely higher number of inactive assays reported in public HTS repositories comparing to the active ones. Given the unbalanced nature of HTS data, values lower than one are usually assigned to \(w\) and a lower weight was assigned to inactive assays compared to the active ones.

A confidence \(\text{Conf}_{BS(A,B)}\) index was assigned to BS to account for missing assays. The equation proposed by Russo et al. (2017) was modified in order to normalize the confidence in a 0-1 range, as follow:

\[
\text{Conf}_{BS(A,B)} = \frac{\left| A_a \cap B_a \right| + \left| A_i \cap B_i \right| \cdot w + \left| A_a \cap B_i \right| + \left| A_i \cap B_a \right|}{\left| A_a \right| + \left| A_i \right|}
\]

A lower weight was given to assays that are negatives for both compounds as explained above.

A final weighted BS (\(BS_{\text{weight}(A,B)}\)) is calculated as the product of \(B(A, B)\) and \(\text{Conf}_{BS(A,B)}\), that accounts both the degree of similarity of the two compared biological fingerprints, and the number of bits on which the comparison is based:

\[
BS_{\text{weight}(A,B)} = B(A, B) \cdot \text{Conf}_{BS(A,B)}
\]

Assays having no active compounds in the source dataset were not considered in the final fingerprints. In the same way, if either the target or the analogue had no positive bits in their fingerprints, the final similarity value was imposed to be equal to zero, to avoid large similarity values resulting from the exclusive comparison of negative assays.

2.1.4 Structural and functional group(s) filters
Pre-filters were implemented in the KNIME workflow to limit the search of potential analogue(s) to chemicals sharing relevant common structural features with the target. The user can decide to activate or deactivate each filter by modifying settings, that become particularly relevant when studying endpoints that are known to be related to a well common structural features with the target. The user can decide to activate or deactivate each filter by modifying settings.

Two independent filters were implemented:

1) **Maximum common substructures (MCS).** Chemicals in the source dataset were filtered based on the presence of a MCS with respect of the target. MCS was calculated using the RDKit MCS code\(^{10}\) implemented in KNIME. If the size of the MCS (i.e. the number of atoms in the common structural moiety) was greater than a given percentage of the size (i.e. number of atoms) of both the target and the analogue, the analogue was retained for the following searches. A default value of 50% was suggested that the user can manually customize in the settings.

2) **Functional groups.** Chemicals in the source dataset were filtered based on the presence of chemical functional groups in common with the target. The presence of 22 functional groups codified as SMARTS was verified for both the target and the possible analogue. Those functional groups codify for general chemical classes / categories (e.g. carboxylic acids or amine) and are relevant to describe the reactivity of chemicals. The collection of SMARTS codifying functional groups were retrieved and adapted from RDKit Functional Group Filter KNIME node and is available in Table S1\(^4\). If the percentage of common functional groups compared to the number of those present in the target was greater than a given percentage, the analogue was considered for the following searches. A threshold of 65% is the default value which can be customized.

2.1.5 Integration of similarities
The three independent lists of analogue(s) are integrated to identify a narrower range of chemicals for target data gap filling. If a given chemical is found in multiple similarity lists, it is considered a more suitable analogue compared to those found in a lower number of lists. For example, if an analogue is found in all of the three lists of similarity categories, a maximum suitability is assigned to that compound.

The prediction of activity for the target is made by averaging activities of analogue(s) included in one of the lists of similar. Chemicals simultaneously included in multiple lists are prioritized (e.g., using chemicals included in at least two out of three lists, or only those included simultaneously in all the three lists). The number of analogues used for prediction may vary based on the user’s decision and the degree of overlap of the similarity lists. It can be up to 20 (i.e., 10 from StrS, 5 from BS, 5 from MS), but the number is often reduced by the presence of chemicals in common between the similarity lists, or by the application of threshold for the selection of analogues (e.g., using only analogues that are included simultaneously in more similarity lists).

2.2 Source datasets
Two datasets (source datasets) were compiled from the literature including toxicological data related to high-tier in vivo endpoints and used for validating the proposed methodology:

1) The DILIrank dataset (Chen et al., 2011, 2016) is a collection of 1,036 FDA-approved drugs divided into four classes according to their potential for causing drug-induced liver injury (DILI). The DILI classification is the result of analysis of FDA-approved drug labeling documents and literature. Drugs are classified in three groups of DILI concern (Most-, Less-
and No-DILI concern), and one additional group (Ambiguous-DILI-concern) with undetermined causality. In the present work, the DILIRank dataset\(^1\) was downloaded. Compounds with Ambiguous-DILI-concern label were discarded.

2) Liu et al. (2015) dataset includes data for 667 compounds retrieved from ToxRef\(^2\). These data refer to three major groups of hepatic histopathologic effects, i.e. hypertrophy, injury, and proliferative lesions. In the present work, if at least one of the three histopathologic effects was positive, the chemical was classified as hepatotoxic, otherwise as non-hepatotoxic.

Chemical data included in each dataset were curated by means of a semi-automated in-house procedure described by Gadaleta et al. (2018). The procedure addresses the identification and removal of inorganic and organometallic compounds and mixtures, the neutralization of salts, the removal of duplicates (also checking for tautomeric forms). Finally, the resulting SMILES are converted to a standardized format. The procedure is implemented in KNIME and is freely available for download\(^3\). Entries with unspecified SMILES, and compounds with ambiguous classification were removed.

Information related to the stereoisomery was ignored because statistically not relevant. Indeed, out of the 15 couple of stereoisomers found in the DILIRank, only two cases (i.e., levofloxacin vs olfoxacin and amphetamine vs dextroamphetamine) showed differences in biological activity. No cases of stereoisomers with different activities were observed in the ToxRef dataset.

The final number of compounds included in each dataset and the distribution of activities is reported in Table 1. Details of the chemical space covered by the two datasets in terms of relevant physico-chemical properties (i.e., molecular weight –MW-, octanol-water partition coefficient –logP- and topological polar surface area –TPSA-) are also given in Table 1, while detailed information on single activity categories are given in Table S2\(^4\). The datasets and the information retrieved with the RAX workflow used to compute similarities are included in the supporting information\(^5\).

### Tab. 1: Source datasets implemented in the RAX workflow

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Categories</th>
<th>Total</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>logP</th>
<th>TPSA</th>
<th>MW</th>
</tr>
</thead>
<tbody>
<tr>
<td>DILIRank</td>
<td>No-DILI (C1)</td>
<td>691</td>
<td>252</td>
<td>260</td>
<td>179</td>
<td>-2.06 to 14.44 (2.8)</td>
<td>0.0 to 3115.35 (126.54)</td>
<td>60.02 to 7049.04 (431.39)</td>
</tr>
<tr>
<td></td>
<td>Less-DILI (C2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Most-DILI (C3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ToxRef</td>
<td>Non-Hepatotoxic (C1)</td>
<td>663</td>
<td>454</td>
<td>209</td>
<td>-</td>
<td>0.03 to 4.76 (2.37)</td>
<td>0.0 to 474.9 (58.9)</td>
<td>42.02 to 972.32 (272.57)</td>
</tr>
<tr>
<td></td>
<td>Hepatotoxic (C2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Results

It should be kept in mind that this workflow is not primary designed for batch calculation on large datasets, and one cannot expect to reach predicting accuracies at the same level of other methodologies (e.g. QSARs) specifically tailored for predicting large databases. One of the major strengths of this approach is that the information used to infer predictions is explicitly reported to the user and, unlike complex theoretical chemical descriptors used in QSAR modeling, it consists in sound chemical and toxicological data to facilitate use for regulators and scientists. In this regard, the authors propose to use the workflow for single chemical RAX predictions, so that experts can take case-by-case decisions on the suitability of the analogue(s) identified, based on the evaluation of data gathered.

With this in mind, the large-scale evaluation described is intended to provide an indication on the overall predictability of the approach. The RAX workflow is validated by predicting compounds included in the two source datasets (i.e., DILIRank and ToxRef). For each chemical, analogue(s) are identified among remaining compounds in the source dataset, then a prediction is returned as a majority vote of activities of selected analogue(s). Separate predictions are generated by considering analogue(s) simultaneously included in at least one, two or three similarity lists (i.e., structural, metabolic and biological). The effect of reducing the searchable list of analogue(s) by applying MCS and FGs pre-filters is also evaluated on accuracy. In order to determine a benchmark in prediction quality, chemicals in the two source datasets are also predicted based on the sole use of the single, closest analogue in terms of structural similarity. This is useful to evaluate if the combined use of multiple information and similarities represents an added value with respect to the traditional use of structural similarity. Full details on predictions are included in the supporting information\(^6\).

Figure 2 and Table 2 report balanced accuracies (BA) and percentage of predicted compounds (i.e., coverage) (%) for the two source datasets, for each combination of pre-filtering options and minimum number of similarity lists that should contain a chemical to keep it into account for prediction. Table 2 and Tables S3-S4\(^7\) include detailed statistics on the validation performed.

For the DILIRank source dataset, the “benchmark” BA obtained by using the closest structural analogue for predictions was equal to 0.632. For this dataset, the combined use of StrS, MS and BS slightly improved this result (i.e., BA = 0.642). BA is further improved when only analogue(s) matching at least two similarity lists are used (BA = 0.660), at the cost of a slight loss in number of predictions (i.e., 18% loss). The integration of the three similarity lists, on the other hand, does not give further improvements. This is likely related to the increased unbalancing of the dataset when considering only chemicals that are predicted under these conditions. Indeed, the “MostDILI” category showed the highest variation in percentage with respect of the initial distribution of activities (from 38% to 22% of the entire dataset). This category is associated to the highest reduction in classification performance, and also affects the performance of the whole dataset.

For the ToxRef database the simple integration of multiple similarities (BA = 0.639) does not improve the benchmark performance (i.e., BA = 0.698). On the other hand, BA is improved considerably when only analogue(s) included in two (BA = 0.719) or three (BA = 0.788) similarity lists are considered, even if in the last case the coverage is severely reduced to 0.141%.

---

3. https://github.com/DGadaleta88/data_curation_workflow
For each combination of similarity lists (i.e., number of lists including a single analogue) and pre-filtering method, the sensitivity (SEN), the specificity (SPE), the Balanced Accuracy (BA) and percentage (%) of predictions are reported. For the multi-category DILIRank database, SEN\text{avg} and SPE\text{avg} are the average of values computed separately for each class, while BA\text{avg} is the arithmetic mean SEN\text{avg} and SPE\text{avg}. The first row of the table refers to the benchmark performance related to the sole use of the closest structural neighbour to infer the prediction.

<table>
<thead>
<tr>
<th>Similarity lists</th>
<th>Pre-filtering</th>
<th>DILIRank</th>
<th>ToxRef</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SEN\text{avg}</td>
<td>SPE\text{avg}</td>
<td>BA\text{avg}</td>
</tr>
<tr>
<td>-</td>
<td>0.510</td>
<td>0.754</td>
<td>0.632</td>
</tr>
<tr>
<td>1</td>
<td>none</td>
<td>0.523</td>
<td>0.762</td>
</tr>
<tr>
<td>2</td>
<td>none</td>
<td>0.546</td>
<td>0.774</td>
</tr>
<tr>
<td>3</td>
<td>none</td>
<td>0.533</td>
<td>0.770</td>
</tr>
<tr>
<td>1 FG+MCS</td>
<td>FG+MCS</td>
<td>0.570</td>
<td>0.781</td>
</tr>
<tr>
<td>2 FG+MCS</td>
<td>FG+MCS</td>
<td>0.563</td>
<td>0.777</td>
</tr>
<tr>
<td>3 FG+MCS</td>
<td>FG+MCS</td>
<td>0.535</td>
<td>0.766</td>
</tr>
</tbody>
</table>

\textsuperscript{1}58 out of 691 SMILES in the dataset were unread from the RAX workflow and were not considered for statistical calculation.

\textsuperscript{2}18 out of 663 SMILES in the dataset were unread from the RAX workflow and were not considered for statistical calculation.

Low sensitivity values were observed in some cases. The reason is likely related to the degree of unbalance of the datasets (see Table 1). The issue of handling unbalanced datasets is also commonly observed for other in silico methodologies for predicting toxicity (i.e., QSAR), so that advanced strategies specifically designed to solve this problem have been proposed (Zakharov et al., 2014).

For the DILIRank dataset, improvements are observed by restricting the analogue search to those chemicals sharing common MCSs and FGs, with BA reaching a value of 0.670. On the other hand, the application of these filters on the ToxRef dataset does not improve the statistical performance. In this regard, the inspection of performance for single chemical categories revealed disappointing statistics for some of them (Tab. S5\textsuperscript{4}). In particular, aromatic amines (BA =0.58) and aromatic alcohols (BA =0.51) are relatively well-represented in the dataset, and the low sensitivities associated to these classes are likely to justify the drop in performance observed on the entire dataset. On the other hand, the methodology was found to perform better on aliphatic compounds such as carboxylic acids (BA =0.88), alcohols (BA =0.82) and halogens (BA =0.71). Overall, good performances were observed for the majority of the chemical categories in the DILIRank dataset, with the aliphatic carboxylic acids being the ones predicted with the higher accuracy (BA =0.80). Lower statistics were observed for aliphatic amines (BA =0.57) and aromatic halogens (BA =0.59). Detailed statistics for single chemical classes can be consulted from Table S5\textsuperscript{4}.

Generally speaking, results confirmed that the use of analogue(s) retrieving different type of similarities improve performance referred to the sole use of structural similarity. This is especially true when source compounds used for RAX belong to at least two orthogonal similarity lists of analogue(s). In this regard, they can be considered “close” to the target under multiple relevant aspects (e.g., chemical, toxicological, kinetic). In order to understand the contribution of different similarities to the final RAX prediction, the
activity of single analogues is evaluated for their coherence with the activity of the relative target. In Table 3, analogues are grouped based on the similarity list(s) in which they appear. For each combination, the number of analogues exactly matching the relative target activity are reported.

As expected, analogues simultaneously included in multiple similarity lists are more likely to match the activity of their target compound. Indeed, the combination of all the three similarity lists shows the highest percentages of source compounds having the same activity as of the target, i.e. 82% for ToxRef and 60% for DILIRank, reinforcing the initial hypothesis of this study.

This combination is followed by those integrating two out of three different similarities, in particular combinations including structural similarity (i.e., StrS and MS or StrS and BS) are always characterized by higher percentages of concordant analogues with respect of the one combining MS and BS.

As for source compounds included in a single similarity lists, those included in the StrS list are more often concordant with their target’s activity with respect to those included in the BS or MS lists.

### Tab. 3: Percentage of single analogues having the same activity as the target

The number of single selected source compounds (#scmpds), the number (#m_scmpds) and the percentage (%m_scmpds) of those matching their target’s activity are grouped based on the similarity lists in which they are included.

<table>
<thead>
<tr>
<th>Similarity</th>
<th>ToxRef</th>
<th>DILIRank</th>
</tr>
</thead>
<tbody>
<tr>
<td>STR, MET, BIO</td>
<td>#scmpds: 100</td>
<td>#scmpds: 82</td>
</tr>
<tr>
<td>STR, MET</td>
<td>1186</td>
<td>895</td>
</tr>
<tr>
<td>STR, BIO</td>
<td>145</td>
<td>108</td>
</tr>
<tr>
<td>MET, BIO</td>
<td>38</td>
<td>24</td>
</tr>
<tr>
<td>STR</td>
<td>5324</td>
<td>3506</td>
</tr>
<tr>
<td>MET</td>
<td>2795</td>
<td>1690</td>
</tr>
<tr>
<td>BIO</td>
<td>2250</td>
<td>1350</td>
</tr>
</tbody>
</table>

### 4 Discussion

#### 4.1 Integrated RAX strategy

This paper describes an automated approach to identify suitable analogue(s) for RAX. As in traditional RAX, the selection of analogue(s) is based on similarity with the target under study. However, while StrS is in many cases the only piece of information used to identify neighbors, this automated tool mathematically combined different approaches to evaluate similarity between the target and a series of putative analogue(s) from a source dataset to strength the evidence in supporting analogue(s) selection. Results in Tables 2-3 confirm the beneficial role of this strategy that leads to predictive improvements with respect of the sole use of the StrS.

The use of MCS and common FGs to narrow the list of candidate analogues does not seem to always carry evident improvements for the ToxRef dataset. In this specific case, hepatotoxicity is a complex endpoint and it is related to multiple mechanisms of action. Consequently, restricting the analogue(s) search to a single chemical category may be less effective with respect of other endpoints with a clear mechanistic link to specific substructure and/or functional groups (e.g., mutagenicity).

The methodology leads to good predictive performance on the two validation datasets, with BA values in the range of 0.632 – 0.670 for DILIRank and of 0.639 – 0.788 for ToxRef datasets. The approach suffers the use of unbalanced data for validation, leading sometimes to low sensitivity values. However, the reader should keep in mind that the methodology here presented does not aspire to reach the same level of predictive performance of other statistical approaches specifically tailored to large scale toxicity predictions. The main strength of this integrated RAX approach is the ability to provide results that are easy to understand, as well as a data-rich output that can be employed by the users to evaluate the final toxicity outcome based on their expertise.

Other examples are reported in the literature that combine different approaches to compute similarity for RAX. Notably, Wu et al. (2010) proposed and validated a robust evaluation framework to determine analogue(s) suitability for RAX that used several cheminformatics tools, e.g., for evaluating physicochemical properties of substances and simulate their metabolic pathway. Even though the decision framework proposed by Wu and coworkers takes in consideration all relevant aspects necessary for a RAX evaluation, no automated procedure is proposed and the intervention of various experts is required at various levels. Low et al. (2013) proposed an automated Chemical Biological Read-Across (CBRA) that used an integrated StrS and BS to predict the toxicity of chemicals. The approach was adapted by Shah et al. (2016) in their generalized read-across (GenRA) that compared the efficiency of structural and biological fingerprints (and a combination of both) to search analogue(s) for RAX. In both cases, only StrS and BS are considered, while other highly relevant pieces of information (e.g., physico-chemical properties, metabolism, reactivity and pharmacokinetics) are not addressed.

Results reported in Table 3 confirm that in our approach StrS still maintains a predominant role. It is accepted that the chemistry should be the starting point for the definition of similarity (ECHA, 2008; Ball et al., 2016), considering the strong correlation between the structure of compounds and their biological effect (Bender et al., 2004). For this reason, a higher number of neighbors based on StrS was selected (i.e., up to ten compared to up to five analogue(s) considered for other similarities). StrS identifies common behaviors between the target and the analogue(s) in terms of physico-chemical properties and, consequently, in terms of toxicokinetics. Indeed, the toxicity of a chemical depends on its absorption and excretion rates and the time that it effectively remains in the organism. Differences in these parameters could affect in vivo toxicity (due to differences in bioavailability) or in vitro toxicity (due to differences in solubility). On the contrary, high structural similarity values will result in analogies for most important physico-chemical properties (e.g., molecular weight, lipophilicity, solubility) and, as a consequence, in a similar toxicokinetic profile.
The role of MS becomes relevant when the metabolism of non-toxic chemicals leads to the production of harmful metabolites or alternatively, when metabolism can represent a detoxification process of a toxic substance. Metabolism of the analogue(s) and the target can have a significant impact on the overall RAX assessment. Indeed, the potential for two compared chemicals to diverge in their bioactivation pathway, may result in a different toxicological profile, and it may sensibly vary the conclusions drawn by using structural similarity alone (Patlewicz et al., 2013).

The use of BS has been explored extensively in the last years (Petrone et al., 2012; Russo et al., 2017; Shah et al., 2016; Grimm et al., 2016; Low et al.; 2013, Patlewicz et al., 2017). An approach that found large consensus and was also applied in this paper was to use outcomes (i.e., positive or negative) from a large number of HTS assays to build binary biological fingerprint of the target and the analogue(s). Fingerprints are well suited to compute similarity; indeed one can use the collective set of results from different assays to compare the target and the analogue(s) using classical mathematical methods, e.g. Tanimoto or Euclidean distances (Willett et al., 1998). Two chemicals characterized by a similar behavior on a large number of different biological assays are also likely to have a common toxicological profile.

4.2 Evaluation of uncertainty
The importance of an explicit strategy to characterize the uncertainty associated to RAX data-gap filling has been highlighted (OECD, 2007; Patlewicz et al., 2013; Blackburn et al., 2014). The number and the degree of suitability of analogue(s), the quality and quantity of the data considered, the nature and severity of the identified toxic effects, and the potency of the analogue(s) for those effects should be ideally evaluated in order to assess the effectiveness of the RAX and make transparent decisions.

The RAX tool offers several elements to quantify the uncertainty associated to RAX, e.g. the final number of analogue(s) used for the data gap filling, the criteria used to select them (e.g., use of functional groups and MCS filters) and the number of categories in which each analog falls. Ideally, a RAX sustained by a higher number of analogue(s) included in multiple lists is considered more reliable than a RAX based on few analogue(s) sharing few similarity lists. Another element that can be used to evaluate uncertainty associated to RAX is the consistence of activities across analogue(s) used. The greatest the percentage of analogue(s) having the same activity, the more the final prediction can be considered reliable. Currently, this RAX workflow was validated for qualitative prediction of toxicity (i.e. classification), because quantitative RAX was recognized as more challenging for the higher number of potential areas of uncertainty to address (Ball et al., 2016). As a drawback, increasing the number of similarities introduces further requirements to analogue(s) selection, with a consequent reduced coverage of the method.

4.3 RAX examples
Figures 3 and 4 provide four examples that show how the output of the RAX workflow should be interpreted and how the integrated similarities represent an advantage over the StrS alone.

A total of 18 analogue(s) were identified from the ToxRef to sustain the RAX of the 2-chlorophenol (CAS 95-57-8) (Figure 3A), a non-hepatotoxic compound. Phenol (CAS 108-95-2) is the only analogue that is simultaneously identified in three similarity lists. Despite not being ranked as a top analogue for StrS, this can be considered the most suitable analogue for sustaining RAX; indeed, the chemical has a very high BS with respect of the target (BS = 0.553, 299 negative assays and 2 positive assays shared with the target). In addition, it undergoes the same metabolic bio-transformations observed for the target, i.e. aromatic hydroxylation ortho- and para- to oxygen (MS = 1.00). In case of 2-chlorophenol, these biotransformations lead to the generation of 3-chlorocatechol and 2-chlorobenzene-1,4-diol, respectively. While phenol can be converted to catechol and hydroquinone, respectively. Hydroxylation may plausibly play a role in the detoxification of these two chemicals, because it increases the number of hydroxyl reactive centers that may undergo phase II reactions (i.e. conjugation) and be easily excreted from the body. Phenol has the same activity of the target, leading to a correct RAX prediction. On the other hand, the use of the activity of the top structural analogue (pentachlorophenol, CAS 87-86-5) leads to an incorrect prediction. Indeed, pentachlorophenol does not undergo hydroxylation that is observed for the 2-chlorophenol and that is a key process for detoxification. Moreover, despite sharing a good amount of common biological assay outcomes, pentachlorophenol also activates 41 additional assays that are not activated by the target (data not shown), drastically lowering the BS (i.e., 0.040).

The hydrazobenzene (CAS 122-66-7) (Figure 3B) is a hepatotoxic chemical from ToxRef. As in the previous example, the sole use of the top structural analogue (4-aminooazobenzene, CAS 60-09-3) leads to an underestimation of target’s toxicity. Indeed, 4-aminooazobenzene has a relatively low BS (i.e., 0.244) that leads to its exclusion from the top five biological analogues. It also differs from the target from a metabolic point of view because it undergoes additional biotransformations that are likely responsible of its detoxification. On the other hand, azobenzene (CAS 103-33-3) appears in all the three similarity lists, and it is more suitable as a RAX analogue despite it is less similar to the target from a structural point of view. Benzidine (CAS 92-87-5) and diphenylamine (CAS 122-39-4) are additional analogues of hydrazobenzene that are included in two out of three similarity lists. Benzidine has a lower MS score (only one shared pathway out of four total observed for the target and the analogue) while diphenylamine shows a different biological behavior (only one positive assay shared with the target). Overall when the analogue(s) that are in two or more lists are used, a prevalence of toxic chemicals (two out of three) is observed leading to a correct estimation of hydrazobenzene hepatotoxicity.

Amlodipine (CAS 88150-42-9) (Figure 4A) is classified as “Less-DILI” in the DILIrank dataset. Out of the 16 analogues identified in the RAX procedure, three of them are included in two or more similarity lists. Two of them (i.e., fenoldipine, CAS 72509-76-3, and nifedipine CAS 21829-25-4) have the same “Less-DILI” classification as the target, while the last one (i.e., clevidipine, CAS 167221-71-8) has a “No-DILI” classification. This is also the top structural analogue of amlodipine in the dataset. The unsuitability of the clevidipine as a source compound to sustain the RAX is related to the very low BS with respect to amlodipine (i.e., close to zero). Indeed, this chemical only shares five negative assay outcomes with the target, while the other two analogues are characterized by more than 200 common assay responses. Overall, fenoldipine is the most suitable analogue to sustain RAX, being included in all the three similarity lists, while nifedipine shows a relatively low StrS value (StrS = 0.514). Metabolism is also relevant; indeed, clevidipine is rapidly metabolized to its primary high-clearance metabolite (H152/81) after cleavage of the ester group at position 3 of the dihydropyridine core (Ericsson et al., 1999) (Figure 5). Due to absence of this rapid cleavage, amlodipine has a much longer half-life than nifedipine and a higher toxicity.
Fig. 3: RAX examples from the ToxRef source dataset
A) 2-chlorophenol (CAS 95-57-8), B) hydrazobenzene (CAS 122-66-7). HT, hepatotoxic; nHT, non-hepatotoxic.

The second example from DILIRank is dobutamine (CAS 34368-04-2) (Figure 4B), a “No-DILI” chemical. Two of the top structural analogues, i.e. propafenone (CAS 54063-53-5) (rank 1) and labetalol (36894-69-6) (rank 3), have discordant activity with respect of the target, and are indeed considered not suitable for sustaining RAX. The reason is that they show a very different metabolic profile with respect of the dobutamine that shares only four out of 19 pathways with propafenone, and five out of 13 with the labetalol. It is likely that some of the non-shared biotransformations are responsible of the target’s detoxification and/or analogues’ increased toxicity. In particular, propafenone and labetalol do not share the same catechol moiety that is present in dobutamine. This is the main site of the metabolic transformations (i.e., aromatic hydroxylation and conjugations) that are responsible of the compound’s clearance.

On the other hand, the three analogues isoprotenerol (CAS 7683-59-2), dopamine (CAS 51-61-6) and epinephrine (51-43-4) share the same catechol moiety as dobutamine. They share a higher number of detoxification pathways and show a higher MS values with respect of propafenone and labetalol. They are indeed characterized by the same activity as dobutamine. Isoprotenerol is also one
of the closest analogues in terms of biological profile (i.e., 194 negative assay responses and 12 positive ones in common with the target), despite it is ranked only seventh as structural analogue.

In addition to tabular outputs, Figure S2 shows the pie-plots that have been generated by the workflow to graphically describe RAX results for the four compounds considered. In these plots, all the putative analogue(s) are included as separate slices in clockwise order, starting from those included in multiple similarity lists (i.e., outlined in dark blue for three lists and light blue for two lists) with background colors that are descriptive of the experimental activity of each analogue(s).

A

<table>
<thead>
<tr>
<th>NAME</th>
<th>Structure</th>
<th>Rank</th>
<th>Activity</th>
<th>Structural Similarity</th>
<th>Metabolic Similarity</th>
<th>Common metabolic pathways</th>
<th>Biological Similarity</th>
<th>Common biological outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amlodipine</td>
<td><img src="image" alt="Structure" /></td>
<td>Target</td>
<td>Less-DILI</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Felodipine</td>
<td><img src="image" alt="Structure" /></td>
<td>STR2, MET1, BIO3</td>
<td>Less-DILI</td>
<td>0.754</td>
<td>0.818</td>
<td>9 out of 11</td>
<td>0.271</td>
<td>231 NEG 8 POS</td>
</tr>
<tr>
<td>Clevidipine</td>
<td><img src="image" alt="Structure" /></td>
<td>STR1, MET3</td>
<td>No-DILI</td>
<td>0.769</td>
<td>0.600</td>
<td>9 out of 15</td>
<td>c.c.a 0</td>
<td>5 NEG</td>
</tr>
<tr>
<td>Nifedipine</td>
<td><img src="image" alt="Structure" /></td>
<td>MET4, BIO1</td>
<td>Less-DILI</td>
<td>0.514</td>
<td>0.500</td>
<td>7 out of 14</td>
<td>0.302</td>
<td>217 NEG 14 POS</td>
</tr>
</tbody>
</table>

13 additional analogue(s)  
4 No-DILI, 8 Less-DILI, 1 Most-DILI

B

<table>
<thead>
<tr>
<th>NAME</th>
<th>Structure</th>
<th>Rank</th>
<th>Activity</th>
<th>Structural Similarity</th>
<th>Metabolic Similarity</th>
<th>Common metabolic pathways</th>
<th>Biological Similarity</th>
<th>Common biological outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dobutamine</td>
<td><img src="image" alt="Structure" /></td>
<td>Target</td>
<td>No-DILI</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Isoproterenol</td>
<td><img src="image" alt="Structure" /></td>
<td>STR7, MET5, BIO2</td>
<td>No-DILI</td>
<td>0.688</td>
<td>0.455</td>
<td>5 out of 11</td>
<td>0.331</td>
<td>194 NEG 12 POS</td>
</tr>
<tr>
<td>Dopamine</td>
<td><img src="image" alt="Structure" /></td>
<td>STR2, MET1</td>
<td>No-DILI</td>
<td>0.756</td>
<td>0.625</td>
<td>5 out of 8</td>
<td>0.243</td>
<td>177 NEG 6 POS</td>
</tr>
<tr>
<td>Epinephrine</td>
<td><img src="image" alt="Structure" /></td>
<td>STR7, MET3</td>
<td>No-DILI</td>
<td>0.688</td>
<td>0.500</td>
<td>5 out of 10</td>
<td>0.001</td>
<td>26 NEG 3 POS</td>
</tr>
<tr>
<td>Propafenone</td>
<td><img src="image" alt="Structure" /></td>
<td>STR1</td>
<td>Less-DILI</td>
<td>0.827</td>
<td>0.211</td>
<td>4 out of 19</td>
<td>0.258</td>
<td>166 NEG 16 POS</td>
</tr>
<tr>
<td>Labetalol</td>
<td><img src="image" alt="Structure" /></td>
<td>STR3</td>
<td>Most-DILI</td>
<td>0.745</td>
<td>0.385</td>
<td>5 out of 13</td>
<td>0.186</td>
<td>42 NEG 13 POS</td>
</tr>
</tbody>
</table>

11 additional analogue(s)  
6 No-DILI, 4 Less-DILI, 1 Most-DILI

Fig. 4: RAX example from the DILIRank source dataset  
A) amlodipine; B) dobutamine.
In this paper, we presented an automated tool that implements good practices for toxicological data-gap filling in RAX that have been described in the recent literature and technical guidelines. A great emphasis is put on the combined use of different types of similarities as a way to identify suitable analogue(s) with a high level of reliability. Another aspect that is worth of note is that the workflow provides a rich output in form of tables/graphs that give a strong basis to support RAX conclusions and to aid toxicologists and risk assessors in decision-making. The data (i.e., simulated metabolites, positive HTS assays, common substructures and/or functional groups) used to selected analogue(s) are made available in the workflow output and represent a relevant resource that the user can consider to assess the reliability of the conclusions drawn by the tool, e.g. by manually evaluating the consistence of toxicological and biological data gathered across the analogue(s).

On the other hand, a limitation of the tool is that it does not identifies a priori specific information relevant for different types of toxicities. Indeed, some specific properties can be key elements for making assessment for some endpoints (e.g., reactivity for mutagenicity, lipophilicity for bioaccumulation), while being less important for other endpoints. Each endpoint should be justified case-by-case, and RAX should be ideally endpoint specific (Patlewicz et al., 2015). In this regard, we invite the users to consider the workflow presented here mainly in combination with other source of evidence, as a visualization tool and a source of relevant data to aid the reasoning, more than an autonomous predictive tool. Further work is needed to develop variations of the workflow that are specific for a given endpoint. Adverse Outcome Pathways (AOPs) represent a promising resource that have been proposed to address endpoint specificities (Tollefsen et al., 2014; Patlewicz et al., 2015). Despite the use of AOPs is currently limited by the low number of validated AOPs, some efforts in applying computational tools to AOPs have been made (Gadaleta et al., 2018). In the future, the integration of computational predictions for molecular initiating events and key events in an AOP could be used to demonstrate that a set of chemicals have analogous biological behavior that is relevant for the toxicological endpoint of concern, providing new evidence for sustaining RAX results (Leist et al., 2017). This can reduce the use of new animal experiment, and improve the way to extract all relevant information from existing data in a more efficient and organized way where multiple features of heterogeneous nature are integrated.

References


doi:10.1016/j.comtox.2018.08.002


Conflict of interest
The authors declare that they have no conflicts of interest.

Acknowledgements
This work has received funding from the European Union’s Horizon 2020 research and innovation programme under Grant Agreement No. 681002 (EU-ToxRisk). We acknowledge the anonymous reviewers for their comments that significantly improved the manuscript.