EPA Public Access

Author manuscript

Comput Toxicol. Author manuscript; available in PMC 2019 October 30.

About author manuscripts

Submit a manuscript

Published in final edited form as:

Comput Toxicol. 2018 ; 8: 34–50. doi:10.1016/j.comtox.2018.07.001.

Extending the Generalised Read-Across approach (GenRA): A systematic analysis of the impact of physicochemical property information on read-across performance

George Helman^{a,b}, Imran Shah^b, Grace Patlewicz^{b,*}

^aOak Ridge Institute for Science and Education (ORISE), Oak Ridge, Tennessee, USA

^bNational Center for Computational Toxicology (NCCT), Office of Research and Development, US Environmental Protection Agency, 109 TW Alexander Dr, Research Triangle Park (RTP), NC 27711, USA

Abstract

Read-across is a useful data gap filling technique used within category and analogue approaches in regulatory hazard and risk assessment. Recently we developed an algorithmic, approach called Generalised Read-Across (GenRA) (Shah et al., 2016) which makes read-across predictions of toxicity effects using a similarity weighted average of source analogues characterised by their chemical and/or bioactivity descriptors. A default GenRA approach (termed baseline GenRA) relies on identifying 10 source analogues relative to a target substance that are structurally similar based on Morgan chemical fingerprints and computing an activity score to estimate presence or absence of *in vivo* toxicity. This current study investigated the impact that similarity in bioavailability plays in altering the local neighbourhood of source analogues as well as read-across performance relative to baseline GenRA using physicochemical property information as a surrogate for bioavailability. Two approaches were evaluated: 1) a filtering approach which restricted structurally related analogues based on their physicochemical properties; and 2) a search expansion approach which included additional analogues based on a combined structural and physicochemical similarity index. Filtering minimally improved performance, and was very dependent on the similarity threshold selected. The search expansion approach performed at least as well as the baseline GenRA, and showed up to a 9% improvement in read-across performance for at least 10 of the 50 organs considered. We summarise the overall impact that physicochemical information plays on GenRA performance, illustrate the improvement for a specific case study substance and describe how to select the most appropriate physicochemical similarity threshold to achieve optimal read-across performance depending on the toxicity effect and chemical of interest. The analyses show that physicochemical property information does result in a modest (up to 9% increase) improvement in structural based read-across predictions.

*Correspondence: Grace Patlewicz, Tel: +1 919 541 1540, patlewicz.grace@epa.gov.

Publisher's Disclaimer: The views expressed in this article are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Keywords

read-across; Generalised Read-Across (GenRA); similarity in bioavailability; physicochemical parameters; read-across performance

1. Introduction

1.1 Background context

Read-across is a widely used technique for filling data gaps for poorly studied substances within category and analogue approaches for regulatory purposes [1]. In principal, read-across predicts a property/endpoint for a substance of interest (target substance) from known information on the same property/endpoint from a 'similar' substance (source analogue) usually based on structural similarity [1,2–3]. Many web-based tools (reviewed in [4]) permit structure similarity searching which typically rely on a Tanimoto index [5] to limit the number of analogues identified. This requires some form of chemical fingerprinting technique to characterise the target substance as well as the inventory of potential source analogues to rapidly search and retrieve source analogues. A critical next step is to evaluate the validity and relevance of these analogues. The evaluation relies upon both general and endpoint specific considerations as well as the comprehensiveness of the underlying data that might exist for the candidate source analogues [3, 6–7]. The types of general considerations include other similarity contexts such as metabolic, reactivity, and bioactivity similarity as well as bioavailability similarity. For the latter, physicochemical properties such as the log of the octanol-water partition coefficient (logP), molecular weight, number of hydrogen bond donors or acceptors (as enshrined in the Lipinski rule of Five [8]) are convenient parameters to 'model' bioavailability. Arguably source analogues that are both structurally similar and possess a similar physicochemical profile should be more likely to exhibit greater similarity in toxicity than just by structure alone. Indeed, such data can be particularly helpful in highlighting differences in properties between a target substance and its source analogues that could impact *in vivo* or *in vitro* toxicity due to differences in bioavailability or solubility respectively [6, 9–10].

Whilst physicochemical similarity (as a model for bioavailability) has been highlighted as an important consideration in evaluating analogues for read-across in various technical guidance documents [1–2] as well as in the scientific literature [6, 9–13] and is often utilised in practice (for example, Pradeep et al. [22]), no systematic analysis has been performed to quantify how much this type of similarity improves read-across prediction. This is principally because read-across remains an expert driven approach that is endpoint and chemical specific which does not facilitate an objective assessment of overall performance. A systematic analysis is important for two main reasons: 1) understanding the utility of read-across for large numbers of substances; and, 2) evaluating the impact that different similarity contexts have on read-across performance.

Previously we have published on a systematic algorithmic approach to performing read-across called GenRA (Generalised Read-Across) [14] which builds upon earlier efforts by Low et al [15]. The GenRA approach predicts toxicity based on a similarity weighted

activity score from source analogues (nearest neighbours). An important feature of the approach was that it quantified the performance and the associated uncertainties in the predictions for large numbers of substances. The published GenRA approach [14] served to establish a baseline for assessing read-across performance relying on the identification of source analogues relative to a target substance that were similar based on chemical and/or bioactivity fingerprints and computing an activity score to estimate presence or absence of toxicity effects as determined within *in vivo* study types [14]. Under GenRA, the identification and evaluation of those analogues was not informed by either the toxicity effect of interest nor any other similarity context such as bioavailability (as modelled by physicochemical similarity), reactivity, or metabolism. In this study, we sought to systematically evaluate the impact that physicochemical similarity had on read-across performance using two different approaches. The first approach termed “filter” was intended to identify structurally similar source analogues and filter these on the basis of a physicochemical similarity threshold. The second approach was termed a “search expansion”. Here the intent was to “frontload” both structure and physicochemical property information into the analogue identification step (i.e., combining both parameters concurrently to identify source analogues).

Here we provide a brief overview of GenRA and the workflow followed to undertake the analysis. Results are summarised on an overall basis before highlighting the impact within chemical neighbourhoods (i.e., using the clusters defined in the original study in [14]) as well as exploring what differences are observed for one case study substance in terms of the specific analogues identified and the ensuing read-across predictions made.

2. Materials and Methods

2.1. Data

The dataset used in this analysis comprises the same training sets that had formed the basis of the GenRA analysis published [14]. Here we summarise each of the data sources used in brief.

2.1.1 Toxicity data—The toxicity data used were obtained from ToxRefDB, which contains *in vivo* health outcomes of hundreds of compounds from animal testing studies. All data are publicly available at ftp://newftp.epa.gov/COMPTOX/Animal_Tox_Data/. For this analysis, outcomes were taken for ~600 chemicals and the studies were aggregated to the level of study type and target site of the effects. Substances that produced a statistically significant effect for a particular study were categorised as ‘positive’ for that toxicity type and denoted with a ‘1’. A substance that did not produce any statistically significant treatment-related effects was categorised as ‘negative’ and denoted with a ‘0’. The study types included chronic toxicity (chr), developmental toxicity (dev), developmental neurotoxicity (dnt), neurotoxicity (neu), reproductive toxicity (rep), acute toxicity (acu), sub-acute toxicity (sac) and sub-chronic toxicity (sub). All other toxicity testing studies (i.e. where a specific guidance was not reported) were grouped into a category that was referred to as “other” (oth). The toxicity within those study types cover 126 different organs and effects.

2.1.2 Chemical structure data—Morgan fingerprints [16] were used as the default chemical fingerprint in baseline GenRA. These fingerprints were calculated using the freely available python library RDKit [17] and were represented as binary (bit) vectors where the elements themselves represented presence (1) or absence (0) of a certain structural feature.

2.1.3 Physicochemical data—Four physicochemical properties were selected as indicators of bioavailability. The four parameters were: hydrogen bond donors, hydrogen bond acceptors, molecular mass, and logP which feature in the Lipinski Rule of Five [8] as properties important for a xenobiotic's pharmacokinetics. These four properties were calculated using RDKit [17]. Of the 599 chemicals used in Shah et al. [14], RDKit failed to calculate properties for 18 of them, leaving a total of 581 substances for the analysis.

2.1.4 Bioactivity data—The *in vitro* assay data were generated from the high-throughput screening (HTS) of the 1,776 ToxCast phase I and phase II compounds and is publicly available (<https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data>). The data used in this study is available as supplementary information in reference [14]. The data collected was based on 821 HTS assays from 6 different technology platforms: ACEA Biosciences (ACEA); Apredica (APR); Attagene, Inc. (ATG); NovaScreen panel (NVS); Odyssey Thera (OT); and Tox21. In ToxCast, each assay datum was reported as the chemical concentration (micromolar) at half maximal efficacy (AC50). An overall activity call is also made based on whether there is a statistically significant concentration response. All active and inactive assay results were represented as binary values (active=1 and inactive=0, respectively). The bioactivity data was only used for the case study substance (see section 3.3 in the results).

2.1.4 Availability—All datasets are available from the EPA ftp site under NCCT publications (see ftp://newftp.epa.gov/COMPTOX/NCCT_Publication_Data/).

2.2. Summary of GenRA

A generalised category workflow has been described by Patlewicz et al. [4]. Figure 1 describes GenRA in the context of this workflow to frame the overall approach and provide context for the current analysis.

Analogue identification is the process of searching for similar analogues to the target substance, typically based on structural similarity. The data gap analysis for a target substance is intended to identify available toxicity and bioactivity data to inform the hazard profile. This drives the scope of the read-across required and the tactical approach of how best to identify analogues that will be fit for the intended purpose. In GenRA, the data gap analysis and the analogue identification steps are conducted at the same time. GenRA enables different analogue identification approaches to be undertaken, including structural similarity and bioactivity similarity. The two steps are conducted concurrently so that the available data for the source analogues can be explored alongside the data gaps for the target substance itself. In the default case, 'baseline GenRA', Morgan chemical fingerprints are used, with a Jaccard index [18–19] as a means of quantifying the pairwise similarity, to identify source analogues.

Analogue evaluation is intended to consider the relevance of the source analogues in terms of both the quantity and quality of their underlying data and their similarity with respect to structural, physicochemical, reactivity, toxicokinetic and toxicological characteristics. Analogue evaluation within GenRA is currently limited to considering the consistency, concordance, and potency of the source analogues relative to each other and to the target substance.

The next step is filling the data gap. Whilst several data gap filling techniques exist, (discussed in more detail in Patlewicz et al. [4]), GenRA fills data gaps by a similarity weighted read-across prediction. The prediction for a target substance is calculated as follows:

$$\frac{\sum_j^k S_j A_j}{\sum_j^k S_j} \quad (1)$$

where k is the number of source analogues in our neighborhood, j is an index of our source analogues ($j=1, \dots, k$), S_j is the Jaccard similarity [18–19] of the structural descriptors between the target and the source analogue, and A_j is its associated binary activity. The similarity weighted activity calculation provides a value that can range from 0 to 1. To interpret this value, a threshold needs to be assigned to determine whether a predicted value would be categorised as ‘positive’, ‘negative’ or ‘indeterminate’. One way of doing this is to apply an assignment rule on the prediction to assign an active or inactive effect for the prediction outcome (see equation 2).

$$A_j = \begin{cases} 1 & A_{pred} \geq 0.5 \\ 0 & A_{pred} < 0.5 \end{cases} \quad (2)$$

For example, if the above classifier set a threshold of 0.5 then a prediction (based on the activity score) of greater than or equal to 0.5 would denote a substance as “active” i.e., expected to exhibit toxicity and a prediction less than 0.5 would denote a substance as “inactive”. Henceforth, we refer to this as the naïve classifier.

Rather than use a default threshold, the published GenRA [14] utilises metrics to assess the performance and quantify the uncertainty of the read-across prediction. This is captured in the final step of the workflow as the uncertainty assessment. In GenRA [14], the prediction accuracy for each toxicity outcome is evaluated across all substances within a neighbourhood. Specifically, a receiver operating characteristic (ROC) curve is constructed using the predicted and true toxicity for the (k) nearest neighbours and at a specific similarity threshold (s). In the ROC analysis, the activity threshold is systematically varied from the minimum to maximum values to calculate the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) [20]. The performance is then measured by plotting the false positive rate against the true positive rate to generate the ROC curve. The area under the ROC curve (AUC) is then taken as a measure of performance of a given k and s . The significance of the AUC is then empirically estimated by constructing a

null distribution by permuting the true (known experimental) toxicity 100 times and calculating the fraction of times that the AUC is more extreme than what would be observed by chance alone; this is reported as the p-value.

2.3. Refined GenRA approach: Physicochemical similarity evaluation

The performance of baseline GenRA (structural similarity characterised by Morgan chemical fingerprints) [14] was compared to a refined GenRA approach where physicochemical similarity had been considered as part of the analogue identification and evaluation steps of the workflow in Figure 1. The “filtering” approach involved modifying the analogue evaluation step, whereas the “search expansion” involved modifying the analogue identification step.

2.3.1 Filtering approach—Ten source ($k = 10$) analogues were identified using the Jaccard similarity [18–19] between the Morgan chemical fingerprints [16] of a given target substance and the inventory of candidate source analogues. The physicochemical similarity between the source analogues and the target was then calculated using a generalised Jaccard similarity metric, which allows for continuous descriptors (equation 3).

$$S_{i,j,phys} = \frac{\sum_p (x_{pi} x_{pj})}{\sum_p (x_{pi})^2 + \sum_p (x_{pj})^2 - \sum_p (x_{pi} x_{pj})} \quad (3)$$

where $S_{i,j,phys}$ is the similarity between our target substance i and one of its source analogues j ($j=1,\dots,k$), x_{pi} is the value of physicochemical property p for our target substance i , and x_{pj} is the value of physicochemical property p for source analogue j . The summations are over p , the four physicochemical properties considered in this analysis. The four parameters were scaled such that a violation of the respective Lipinski rule [8] would correspond to a value of 1. For example, a chemical with 0 hydrogen bond donors would have a value of 0 for that property, 5 hydrogen bond donors would correspond to a value of 1, 3 hydrogen bond donors would correspond to a value of 0.6, etc.

A threshold value is then selected under which source analogues identified in the analogue identification step will be considered unsuitable and removed from subsequent analysis. The similarity metric (equation 4) is where θ is our chosen threshold value and $I()$ is the indicator function.

$$S_{i,j} = S_{i,j,struct} I(S_{i,j,phys} > \theta) \quad (4)$$

This would leave a subset of $k_\theta < k$ source analogues remaining in the neighbourhood. A grid search was conducted by allowing θ to vary over $[0,1]$ in increments of 0.05.

This approach of filtering mimics the practice of potentially removing source analogues if their physicochemical properties differ markedly from that of the target substance as their *in vivo* toxicity can differ (as discussed earlier).

2.3.2 Search expansion approach—The search expansion approach involved incorporating the physicochemical property information and chemical fingerprint information as characteristics in the similarity search. The similarity metric is represented by equation 5:

$$S_{i,j} = w_1 S_{i,j, struc} + w_2 S_{i,j, phys} \quad (5)$$

where $S_{i,j, struc}$ is the structural similarity between chemicals i and j , $S_{i,j, phys}$ is the physicochemical similarity (calculated by equation 3), and w_1 and w_2 are modifiable weights placed on the two contexts of similarity with $w_1 + w_2 = 1$. Thus, the 10 source analogues were identified based on both structural and physicochemical similarity. A grid search was performed by allowing w_1 to vary over $[0,1]$ in increments of 0.05 (since $w_2 = 1 - w_1$, w_2 also varies over the same interval).

2.4 Software

Data processing and analysis was conducted in the Python programming language (2.7.12, Python Software Foundation) using RDKit and the matplotlib package (v2.0.2) [21] for visualisation. Bioactivity heatmaps in section 3.3 were generated in the R programming language (R version 3.4.2) using the ggplot2 package.

3. Results and Discussion

3.1 Evaluating the refined GenRA approach which considers physicochemical similarity

The performance of the read-across predictions was evaluated using the area under the curve of the ROC plot (AUC) as discussed in section 2.2. Due to data paucity, predictions were aggregated to the organ level across all *in vivo* study types for the AUC calculation. This assumption is a reasonable aggregation given the intent was to use physicochemical information to model bioavailability, which is considered on a tissue or organ level in standard toxicokinetic modelling approaches. For the AUC calculations, at least 30 positive and 30 negative toxicity values were required to ensure a statistically robust comparison for the predictions derived. This resulted in 50 data-rich organs covering 575 chemicals.

3.1.1 Filtering—For each organ, the performance (as measured by the AUC and expressed as a percentage) did not typically change for low threshold (θ) values (typically θ ranging from 0 – 0.6 but a sharp decrease was noted at higher threshold values (0.8 and higher). This is illustrated in Figure 2 using liver and kidney as examples since these were associated with the largest number of results from which to assess the performance. A low threshold value equates to a subset of structural analogues where source analogues that have a low physicochemical similarity to the target have been removed, here the impact to the AUC performance is minimal. As the threshold increases, the neighbourhood of analogues is depleted by removal of analogues that do have a higher physicochemical similarity (i.e., by retaining only those analogues that have both high structurally and physiochemical similarity to the target). There is a risk that useful analogues could be removed from consideration if this threshold is too stringent. The only notable exception to this trend was the target organ vagina, which displayed a steady improvement as the threshold increased.

Table 1 contains the top 10 most improved organs that displayed an improvement in read-across performance of over 1% relative to the baseline GenRA (this was an arbitrary selection for illustrative purposes only). Additionally, a tabulation of the best performing threshold values for each of the target organs is provided in the supplementary information (Table A1).

3.1.2 Search expansion approach—The search expansion approach involved a similarity search using a weighted sum of Morgan chemical fingerprint distance and physicochemical distance (equation 3) to identify source analogues. The grid search was conducted over w_2 , where $w_2 + w_1 = 1$, i.e. when $w_2 = 0$, the search would be on the basis of structure alone and when $w_2 = 1$, the search would be on physicochemical characteristics alone.

The search expansion approach did not give rise to a marked improvement in performance over the baseline for the entire dataset though a much larger improvement (AUC increases of 5% or more) was noted for certain organs including large and small intestine, pancreas, ureter, and, urinary bladder

Figure 3 shows the AUC values plotted against the grid search over w_2 for the search expansion method for liver and kidney (selected to provide a direct comparison to the filtering approach). In the two plots, source analogues identified purely by physicochemical properties tended to give rise to poorer read-across performance compared with structural source analogues. This is not entirely unsurprising since one might expect that the structure would encode more relevant information for predicting toxicity than physicochemical information alone. However, it was also the case that often an interior point of the plot likely influenced by a combination of structure and physicochemical similarity gave rise to a neighborhood of source analogues that performed better than a purely structural neighborhood. Of the 50 organs analysed, 28 of them had such a point that was at least 1% better than the purely structural neighborhood. The physicochemical weights (w_2) giving rise to the best performance for all the target organs are shown in Table A2 of the supplementary information.

Table 2 reports the 10 organs where the greatest improvement in read-across performance was noted, ranging from 2% to 9.9%. Improvements of 1% or greater were noted for an additional 18 organs.

3.2. Evaluating the refined GenRA performance where only logP is used to account for physicochemical similarity

The filtering and search expansion approaches were repeated using only logP as the parameter to characterise physicochemical similarity to evaluate its significance in driving the differences in read-across performance over and above the baseline GenRA. LogP is often the key parameter used in modelling absorption or penetration hence there was an expectation that a similar performance could be observed compared with using all 4 parameters.

AUC plots comparing the logP analysis with all 4 physicochemical parameters can be found in the supplementary information. For the filtering approach, the same overall trend in performance across all threshold (θ) values was observed when using logP alone. For all organs, both curves displayed performance differences no greater than 4% at any point of the grid search. Using logP alone in the filtering approach gave rise to a comparable performance as using all 4 physicochemical properties. This is indicative that logP was the most influential parameter of the 4 properties selected to characterise physicochemical similarity.

In contrast, differences were observed in the search expansion approach when comparing logP with all 4 properties. For most organs, using all four properties performed better or comparable to only using logP when inspecting the AUC plots indicating that the other parameters were impactful to a limited extent in identifying analogues. Figure 4 shows the AUC plots for liver and kidney using all 4 properties and/or logP alone. Noticeable exceptions to this trend were found for target organs: Blood, Estrous Cycle, Prostate, Stomach, Ureter, and Age Landmark¹. For these organs, logP alone gave rise to better read-across performance than baseline GenRA. It is not clear from a biological perspective why these specific organs would follow a different trend per se – for blood and stomach, perhaps the passive diffusion that occurs for a substance first entering the body is more directly linked to logP than it is for the other properties considered. Figure 5 shows the AUC plots for Blood and Stomach to illustrate the differences in performance.

3.3. Illustrative case example using Butyl Benzyl Phthalate

The practical application of the impact that physicochemical property information had on read-across performance was illustrated using butyl benzyl phthalate as a case study substance. This substance was chosen since it is data-rich in terms of its available ToxRefDB *in vivo* toxicity data, and contains a well recognised ester structural feature. Butyl benzyl phthalate is a member of cluster 80 from the cluster analysis performed in the original manuscript [14].

It is worth noting how ToxRefDB data was aggregated together to put into context the insights derived from this case study which was intended as an illustration of the impact of the different GenRA predictions. Whilst the anti-androgenic mode of action of Butyl Benzyl phthalate has been studied [23], this type of information had not been explicitly reported in the version of ToxRefDB. The aggregations in ToxRefDB rely upon a systematic approach to determine positive vs negative toxicants which are not necessarily reflective of any regulatory use. In other words, positive and/or negative toxicants for a particular “-icity” were determined by grouping effects together and identifying the lowest LOAEL from that group for a given chemical or study. This is practical for research purposes, moreover it forms a part of a hybrid assessment approach to identify whether any specific substance might warrant further evaluation which could include an expert review to make a judgement on specific LOAELs and critical effects. This would explain any apparent discrepancy between the effects identified in ToxRefDB for this case study chemical and what might be

¹Age Landmark is clearly not a target organ but denotes a time period of development that is reported in developmental studies

identified by a detailed expert review. ToxRefDB is currently undergoing a substantial upgrade. The new version will include a customised effect grouping to calculate points of departure. A manuscript describing the differences and the way in which the data has been aggregated in the existing ToxRefDB is in preparation [24].

Figure 6 depicts the neighborhood for butyl benzyl phthalate based on the baseline GenRA approach. As would be anticipated from a structural search, baseline GenRA identified a set of 10 phthalates as source analogues. For the purposes of this case study, baseline GenRA was compared with the refined GenRA taking into account physicochemical similarity using the filtering and search expansion approaches for 25 chronic and developmental studies for which butyl benzyl phthalate is known to exhibit toxic effects.

Table 3 lists the studies where positive toxicity effects were observed as well as the baseline GenRA predictions based on the 10 source analogues depicted in Figure 6. Only 4 of the 25 toxicity effects were predicted correctly (16% accuracy), assuming a threshold of 0.5 (the naïve classifier from equation 2 to categorise positive vs. negative).

3.3.1 Filtering approach—A variety of parameter choices were evaluated for the filtering approach. The first was to evaluate the impact that setting a ‘modest’ filter to remove source analogues from the neighborhood that possessed a physicochemical similarity of less than 0.8. This filter removed three analogues from consideration (see Figure 7). The source analogues were Benzyl acetate [140-11-4], Di(2-ethylhexyl) phthalate [117-81-7] and Dimethyl phthalate [131-11-3] with physicochemical similarity indices 0.72, 0.75 and 0.77 respectively.

The results of this neighborhood in comparison to the baseline GenRA predictions are shown in table 4.

With respect to the naïve classifier applied in the baseline GenRA approach, the number of correct predictions improved markedly with 7 of the 25 toxicity effects now correctly predicted (28% accuracy), with 2 additional toxicity effects gaining an increased GenRA activity score. Seven of the toxicity effects decreased in terms of their activity scores compared to the baseline GenRA prediction.

A stricter filter was then applied to remove analogues with a physicochemical similarity of less than 0.9 from consideration. As can be seen in figure 8, this rejected 1 additional substance (Diethyl phthalate [84-66-2] (physicochemical similarity index = 0.85)) in comparison to the 0.8 filter but it did have a large impact on the predictions derived. (results in table 5).

In the case of Butyl benzyl phthalate, using physicochemical similarity filter of 0.8 improved the predictive accuracy from 16% to 28%. Increasing the value of the filter threshold to 0.9 produced a decrease in accuracy to 8%. The decrease in performance was primarily due to the exclusion of Diethyl phthalate [84-66-2] from the local neighbourhood, suggesting its importance in driving the accurate read-across prediction of the specific study types considered. This mirrors the general trend observed for the overall performance for specific target organs (discussed in section 3.1.1) where there was a sharp decrease in performance if

the filtering was too strict; due to highly relevant source analogues being erroneously removed from consideration.

3.3.2 Search expansion approach—A variety of parameter options were evaluated for the search expansion case study. In the first instance, 10 source analogues for butyl benzyl phthalate were identified where equal weights were placed on the structural similarity and physicochemical similarity contexts (using equation 5). Figure 9 shows the resulting neighborhood.

In addition to some of the chemicals from the baseline GenRA neighbourhood being re-ordered, 3 chemicals that were originally in the structural neighbourhood were superseded by new source analogues. Of note, was that these 3 new source analogues were not phthalates per se, which suggests that physicochemical properties may play an important role for this neighbourhood. The predictions for this neighbourhood relative to the baseline GenRA predictions are listed in table 6.

The predictions for this neighbourhood demonstrate a slight improvement over the baseline GenRA, correctly predicting 5 of the 25 effects (20% accuracy). Activity scores increased for an additional 7 effects. Eight toxicity effects decreased in their activity scores, but only developmental reproductive performance changes its prediction with respect to the naïve classifier.

Whilst these three non-phthalates clearly demonstrate an improvement in read-across performance, it raises a question of how relevant these specific analogues might be relative to the target butyl benzyl phthalate. To explore this aspect further, we compared the similarity in bioactivity profile of the analogues. The ToxCast hit call outcomes on a per assay technology basis (Attagene, Apredica, Novoscreen, BioSeek, Odyssey Thera and Tox21) of the neighbourhood were visualised as heatmaps to qualitatively compare the concordance in activity profiles across the analogues (Figure 10). The ToxCast hit calls were sourced from the supplementary information reported in [14]. The intention here was to inspect whether the bioactivity of the non-phthalates was concordant with the target and the remaining structurally similar analogues. Although Novoscreen assay profiles were also visualised in a heatmap, the output is not shown as the data availability was extremely sparse across the neighbourhood. All substances with exception of Isofenphos were associated with at least some ToxCast hit call data. Assays with more than 50% missing values (i.e., not tested) for the target and 9 source analogues were removed from consideration. The heatmap for the Attagene (AT) technology is shown in Figure 10(a). By visual inspection, greatest consistency in the assay profile between the target substance and the 2 non-phthalate analogues is most apparent for the assays targeting nuclear receptors. Even greater consistency was observed across the category for the Apredica (APR) and Odyssey Thera (OT) technologies (Figure 10(b) and Figure 10(c)). The available assay outcomes for the Bioseek (BSK) and Tox21 technologies revealed some inconsistency between the phthalate analogues themselves (Figure 10(c) and Figure 10(d)).

To quantify these apparent consistencies in bioactivity profile, pairwise Jaccard similarity on the basis of the bioactivity fingerprint was calculated between target and each source

analogue to investigate whether or not there is a difference on that basis between the phthalates and non-phthalates in the neighborhood. The Apredica technology was chosen for this calculation since all substances had been tested with that technology at the 24hr and 72 hr timepoints and therefore provided a common domain on which to calculate the fingerprint. The pairwise similarities for the entire GenRA relative to the target butyl benzyl phthalate were also calculated to provide context to interpret the index values in relative terms. Table 7 shows the chemicals in the neighbourhood together with their Apredica based bioactivity similarity indices and their physicochemical similarities. The non-phthalates appear to be reasonable analogues on a bioactivity basis (in fact, they have some of the highest similarities). Figure 11 provides a histogram of the pairwise similarity scores across the entire dataset to provide a context of the scores on a relative scale.

This assessment whilst preliminary for this specific case study, does lend support to the relevance of the non-phthalate analogues based on their bioactivity similarity. A generalised systematic analysis of the impact of bioactivity similarity will be the subject of a forthcoming manuscript.

Finally, a search using a distance metric that placed a very high weight on the physicochemical similarity relative to the structural similarity was conducted (i.e. the $w_2 = 0.9$ and $w_1 = 0.1$). The resulting neighborhood is shown in Figure 12, which is substantially different from the original neighbourhood based on structural similarity (Figure 6). This neighborhood contained 6 non-phthalates. The predictions relative to the baseline GenRA predictions are listed in table 8.

With respect to the naïve classifier, this neighbourhood predicted 9 of the 25 toxicity effects correctly (36% accuracy), with an improved activity score for 4 additional effects. Six toxicity effects had a lower predicted activity score in comparison to the baseline GenRA prediction, but only developmental reproductive performance changed its prediction with respect to the naïve classifier.

To explore the potential relevance of the additional non-phthalate analogues, similar heatmaps were constructed for the APR technology for the target and 9 analogues from Figure 12 (Figure 13).

Again, Apredica (24hr and 72hr time points) assays were used to calculate the bioactivity similarity for the neighbourhood reflected in Figure 12. The associated similarity scores are listed in Table 9. Cloquintocet-mexyl and 2,4-D-Butotyl both had similarity scores of 0 with respect to this array of assays, which raises questions about their appropriateness as analogues though it is worth noting that monobenzyl phthalate, which is in the baseline neighborhood, also has a similarity score of 0 on this basis.

3.4. Evaluating Performance by cluster

In the published GenRA analysis [14], all substances in the dataset were clustered into categories (neighbourhoods) using K-means clustering based on structural similarity. Here the performance using the search expansion approach (using all 4 properties) is summarised

with reference to these same clusters to enable a direct comparison with the original analysis.

Figure 14 shows the cluster-by-organ breakdown of the results by reflecting the optimal weight to place on physicochemical distance to give rise to the highest AUC value. Only those cluster and organ combinations that contained at least five true positive values and five true negative values and which had an improved AUC value of 0.1 or higher over the GenRA baseline are shown. This pared down our original dataset of 599 chemicals to 499. The cluster by organ breakdown in full showing the optimal weight and the associated AUC value are provided in the supplementary information.

In practice, this type of information is intended to facilitate analogue identification using the search expansion approach. The process could be applied as follows, first assign a new chemical into one of the existing chemical clusters (per the supplementary information reported in [14]), use the heatmap in Figure 14 (and the associated figure 6 data file in the supplementary information) to identify the optimal weight to use for the physicochemical similarity distance for the target organs of interest to generate a refined GenRA prediction for that associated effect. If the target organ is one of blood, stomach, ureter or age landmark, use the corresponding heatmap (Figure A1 in supplementary information) to determine the appropriate weight to apply for a logP only refined GenRA prediction.

3. Conclusions

Recently we developed an algorithmic, automated approach called Generalised Read-Across (GenRA) [14] which makes read-across predictions of toxicity effects using a similarity weighted average of source analogues characterised by their chemical and/or bioactivity descriptors. The GenRA approach served as a first step in establishing a baseline for read-across performance. A baseline GenRA approach relies on identifying source analogues relative to a target substance that are structurally similar based on Morgan chemical fingerprints and computing an activity score to estimate presence or absence of *in vivo* toxicity effects. The identification and evaluation of those analogues was not informed by the toxicity effect of interest nor other considerations such as similarity in metabolism, reactivity or bioavailability.

This study investigated the role that physicochemical property information as a surrogate for bioavailability plays in identifying and evaluating source analogues, as well as the read-across predictive performance. Physicochemical information was captured in 4 properties: logP, MW, number of hydrogen bond donors and number of hydrogen bond acceptors. Two approaches were evaluated: 1) a filtering approach which relied on evaluating structurally related analogues on the basis of their physicochemical properties; and, 2) a search expansion approach which relied upon identifying analogues on the basis of their structural and physicochemical characteristics at the same time. Performance was evaluated on the basis of the area under the curve (AUC) of a ROC curve. The filtering approach involved setting a physicochemical similarity threshold θ for identifying structural source analogues.

On the basis of this study, no marked improvement in performance was observed using the filtering approach, indeed there was a marked decrease in performance for high values of θ , indicating that using a very stringent threshold ($\theta > 0.9$) of physicochemical similarity could discard important source analogues from the prediction. On the other hand, a search expansion approach was found to generally improve read-across performance over and above the GenRA baseline. Of the 50 data-rich target organs included in the analysis, 10 of these showed improvements in performance from 2-9% over the baseline GenRA and a further 18 organs showed improvements greater than 1%. The analysis was repeated to compare whether physicochemical similarity as characterised by the 4 properties gave rise to better performance when using logP alone. In general, the same trends were observed in the filtering approach using either logP or all four properties showing that in general logP was the most influential physicochemical property driving the performance. For the search expansion approach, all 4 properties were found to give rise to better performance for the majority of organs aside from blood, stomach, ureter, prostate and age landmark where logP alone resulted in better performance. It is not clear why logP would be necessarily better in performance for these target organs – possibly a more direct link to absorption. Results for the search expansion approach were also summarised on the basis of chemical clusters that had been derived in the original analysis [14]. The intention was to help facilitate a customised analogue identification approach. A heatmap of cluster by organ was produced to provide a practical guide of what physicochemical weights to use depending on what specific target organ predictions were being made and what improvement in AUC could be expected. The intent was to propose default thresholds on what might be an appropriate balance between physicochemical and structural similarity for specific target organ GenRA predictions; i.e. the threshold for identifying analogues to predict liver effects for a specific type of chemical cluster would be different than the threshold for stomach effects for the same chemical cluster.

A specific case study substance, butyl benzyl phthalate was then used to illustrate the impact that changing specific similarity thresholds had on the read-across predictions for target organ toxicity effects observed in chronic and developmental studies. Search expansion introduced several very structurally dissimilar analogues yet resulted in a marked improvement in performance over the baseline GenRA for many of the target organs. The bioactivity profile of the non-phthalate analogues was explored visually on the basis of hit call profiles derived from ToxCast assays relative to the target substance to rationalise whether the biological similarity was consistent across the neighbourhood. There was a reasonable level of consistency for the Apredica technology where there was the greatest density of experimental data. Pairwise similarity scores on the basis of Apredica data were found to be higher for the non-phthalates relative to the dataset average. This provided preliminary evidence in support of the use of these non-phthalates as analogues to include in the neighbourhood to read-across to butyl benzyl phthalate itself.

Overall, these analyses demonstrate that there is a quantitative improvement in read-across performance when physicochemical property information is considered in concert with structural information.

This analysis is part of our ongoing research direction to systematically evaluate the impact that other contexts of similarity (metabolic, reactivity) have in driving read-across performance and in encoding these in an objective manner to facilitate reproducible and robust read-across predictions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

- [1]. OECD, Guidance on grouping of chemicals. OECD Series on Testing and Assessment No. 194. Organisation for Economic Co-operation and Development, Paris, France, 2014.
- [2]. ECHA, Guidance on information requirements and chemical safety assessment. Chapter R.6: QSARs and grouping of chemicals, 2008 http://echa.europa.eu/documents/10162/13632/information_requirements_r6_en.pdf (accessed 9 April 2018).
- [3]. ECETOC, Technical Report 116 Category approaches, read-across, (Q)SAR. <http://www.ecetoc.org/technical-reports>, 2012 (accessed 9 April 2018)
- [4]. Patlewicz G, Helman G, Pradeep P, Shah I, Navigating through the minefield of read-across tools. A review of in silico tools for grouping, Computational Toxicology 3 (2017) 1–18. [PubMed: 30221211]
- [5]. Tanimoto T, An Elementary Mathematical theory of Classification and Prediction. Internal IBM Technical Report. 1957.
- [6]. Patlewicz G, Ball N, Booth ED, Hulzebos E, Zvinavashe E, Hennes C, Use of category approaches, read-across and (Q)SAR: general considerations, Regul. Toxicol. Pharmacol 67 (2013) 1–12. 10.1016/j.yrtph.2013.06.002.
- [7]. Patlewicz GY, Fitzpatrick J, Current and future perspectives on the development, evaluation and application of in silico approaches for predicting toxicity, Chem. Res. Toxicol 29 (2016) 438–451. [PubMed: 26686752]
- [8]. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, Adv Drug Deliv Rev. 46 (2001) 3–26. [PubMed: 11259830]
- [9]. Wu S, Blackburn K, Amburgey J, Jaworska J, Federle T, A framework for using structural, reactivity, metabolic and physicochemical similarity to evaluate the suitability of analogs for SAR-based toxicological assessments, Regul. Toxicol. Pharmacol 56 (2010) 67–81. 10.1016/j.yrtph.2009.09.006. [PubMed: 19770017]
- [10]. Wang NC, Jay Zhao Q, Wesselkamper SC, Lambert JC, Petersen D, Hess-Wilson JK, Application of computational toxicological approaches in human health risk assessment. I. A tiered surrogate approach, Regul. Toxicol. Pharmacol 63 (2012) 10–19. 10.1016/j.yrtph.2012.02.006. [PubMed: 22369873]
- [11]. Patlewicz G, Ball N, Boogaard P, Becker RA, Hubesch B, Building scientific confidence in the development and evaluation of read-across, Regul. Toxicol. Pharmacol 72 (2015) 117–133. [PubMed: 25857293]
- [12]. Schultz TW, Amcoff P, Berggren E, Gautier F, Kalric M, Knight DJ, Mahony C, Schwarz M, White A, Cronin MTD, A strategy for structuring and reporting a read-across prediction of toxicity, Regul. Toxicol. Pharmacol 72 (2015) 586–601.
- [13]. Ball N, Cronin MT, Shen J, Blackburn K, Booth ED, Bouhifd M, Donley E, Egnash L, Hastings C, Juberg DR, Kleensang A, Kleinstreuer N, Kroese ED, Lee AC, Luechtefeld T, Maertens A, Marty S, Naciff JM, Palmer J, Pamies D, Penman M, Richarz AN, Russo DP, Stuard SB, Patlewicz G, van Ravenzwaay B, Wu S, Zhu H, Hartung T, Toward Good Read-Across Practice (GRAP) guidance, ALTEX 33 (2016) 149–166. 10.14573/altex.1601251. [PubMed: 26863606]
- [14]. Shah I, Liu J, Judson RS, Thomas RS, Patlewicz G, Systematically evaluating read-across prediction and performance using a local validity approach characterized by chemical structure

and bioactivity information, Regul. Toxicol. Pharmacol 79 (2016) 12–24. 10.1016/j.yrtph.2016.05.008. [PubMed: 27174420]

- [15]. Low Y, Sedykh A, Fourches D, Golbraikh A, Whelan M, Rusyn I, Tropsha A, Integrative chemical-biological read-across approach for chemical hazard classification, Chem. Res. Toxicol 26 (2013) 1199–1208. [PubMed: 23848138]
- [16]. Rogers D, Hahn M, Extended-Connectivity Fingerprints, J. Chem. Infor. Model 50 (2010) 742–754.
- [17]. Landrum G, RDKit. 2015, from www.rdkit.org.
- [18]. Jaccard P, Lois de distribution florale, Bulletin de la Société Vaudoise des Sciences Naturelles 38 (1902) 67–130.
- [19]. Jaccard P, The distribution of the flora in the alpine zone, New Phytologist 11 (1912) 37–50.
- [20]. Hanley JA, McNeil BJ, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1982) 29–36. [PubMed: 7063747]
- [21]. Droettboom M, matplotlib v2.0.2 [computer software]
- [22]. Pradeep P, Mansouri K, Patlewicz G, Judson RS, A systematic evaluation of analogs and automated read-across prediction of estrogenicity: A case study using hindered phenols, Computational Toxicology 4 (2017) 22–30. [PubMed: 30057968]
- [23]. Hotchkiss AK, Parks-Saldutti LG, Ostby JS, Lambright C, Furr J, Vandenberg JG, Gray LE Jr, A mixture of the “antiandrogens” linuron and butyl benzyl phthalate alters sexual differentiation of the male rat in a cumulative fashion, Biol Reprod 71 (2004) 1852–1861. [PubMed: 15286035]
- [24]. Watford S, Pharm LL, Shin R, Martin MT, Paul Friedman K, ToxRefDB version2.0: Improved utility for predictive and retrospective toxicology analyses. In preparation

Highlights

- GenRA approach is summarised in the context of the category workflow
- The impact of physicochemical information on read-across performance was assessed in 2 ways: filtering and search expansion
- Search expansion resulted in an up to 9% improvement in read-across performance for 10 of the 50 data rich target organs
- Results are summarised on a neighbourhood (chemical category) basis
- A case study substance is used to compare and contrast the read-across performance using the 2 approaches

Figure 1:
Category workflow in GenRA

Figure 2:
AUC plots for aggregated predictions for kidney and liver using the filtering method. Similar plots can be found for all 50 organs in the supplementary information.

Figure 3:
AUC plots for liver and kidney organs using the search expansion approach. Similar plots can be found for all 50 organs in the supplementary information.

Figure 4:
AUC plots for liver and kidney organs using the search expansion approach with all 4 properties vs logP alone

Figure 5:
AUC plots for blood and stomach using the search expansion approach with all 4 properties
vs logP alone

Figure 6:

Neighborhood identified using baseline GenRA. Butyl benzyl phthalate, the target substance lies in the centre. Numbers on the edges represent the pairwise structural similarity scores ('S') and physicochemical similarity score ('PC') between the target and a source analogue.

Figure 7: Neighborhood for the "modest" filter analysis. Butyl benzyl phthalate is in the centre. Numbers on edges represent the pairwise structural similarity scores ('S') and physicochemical similarity score ('PC') between butyl benzyl phthalate and a source analogue.

Figure 8:

Neighborhood for the stricter filter analysis. Butyl benzyl phthalate is in the center. Numbers on edges represent the pairwise structural similarity scores ('S') and physicochemical similarity score ('PC') between butyl benzyl phthalate and a source analogue.

Figure 9:

Neighbourhood resulting from search where structure and physicochemical characteristics were equally weighted. Butyl benzyl phthalate is in the centre. Numbers on edges represent the pairwise structural similarity scores ('S') and physicochemical similarity score ('PC') between butyl benzyl phthalate and a source analogue.

Figure 10:
ToxCast assay technology heatmaps on the basis of hit calls for butyl benzyl phthalate and its source analogues as depicted in figure 9

Figure 11.

Histogram of all pairwise bioactivity similar for the GenRA training set tested in the Apredica (24hr and 72 hr time points) relative to target butyl benzyl phthalate

Figure 12:

Neighbourhood for the high-weight for physicochemical similarity. Butyl benzyl phthalate is in the centre. Numbers on edges represent the pairwise structural similarity scores ('S') and physicochemical similarity score ('PC') between butyl benzyl phthalate and a source analogue.

Figure 13:
Heatmap for the APR technology for the target and source analogues depicted in Figure 12

Figure 14.

Cluster-by-organ heatmap for the search expansion approach using the GenRA clusters and all four physicochemical properties. The number in the box indicates the optimal weight to place on physicochemical similarity during the search expansion, and the colour density indicates the improvement in performance that was observed.

Table 1.

Organs where there was at least an 1% improvement in performance relative to baseline GenRA. The AUC increase is expressed as a percentage.

Organ	Improvement in performance between the Refined GenRA relative to the baseline GenRA
Gallbladder	2.5%
Vagina	2.3%
Age Landmark	2.3%
Sexual Development Landmark	2.0%
Seminal Vesicle	1.6%
Stomach	1.2%
Urinary Bladder	1.1%
Offspring Survival-Late	1.1%
Intestine Large	1.1%
Ureter	1.0%

Table 2:

Target organs where the largest improvement in read-across performance was found.

Organ	Improvement in performance between the Refined GenRA relative to the baseline GenRA
Ureter	9.9%
Gallbladder	6.2%
Estrous Cycle	5.3%
Seminal Vesicle	4.8%
Intestine Small	4.5%
Stomach	4.3%
Pancreas	3.6%
Vagina	3.2%
Sperm Measure	2.3%
Sexual Development Landmark	2.1%

Table 3:

Baseline GenRA similarity weighted activity scores for 25 studies for which butyl benzyl phthalate has positive effects. A GenRA active score of 0.5 or greater is categorised as positive (P).

Study Type	Organ	Baseline GenRA prediction	Overall predicted outcome w.r.t naïve classifier
Chronic	Body Weight	0.78	P
Chronic	Clinical Chemistry	0.27	N
Chronic	Food Consumption	0.00	N
Chronic	Hematology	0.00	N
Chronic	Kidney	0.27	N
Chronic	Liver	1.00	P
Chronic	Mortality	0.27	N
Chronic	Pancreas	0.27	N
Chronic	Prostate	0.00	N
Chronic	Skin	0.27	N
Chronic	Spleen	0.00	N
Chronic	Tissue NOS	0.00	N
Chronic	Urinary Bladder	0.00	N
Developmental	Body Weight	1.00	P
Developmental	Bone	0.27	N
Developmental	Clinical Signs	0.00	N
Developmental	Eye	0.17	N
Developmental	Heart	0.00	N
Developmental	Kidney	0.10	N
Developmental	Liver	0.48	N
Developmental	Offspring Survival-Early	0.33	N
Developmental	Ovary	0.00	N
Developmental	Reproductive Performance	0.50	P
Developmental	Ureter	0.00	N
Developmental	Water Consumption	0.21	N

Table 4:

Baseline GenRA and Refined GenRA predictions using filtering approach for physicochemical similarity (expressed as similarity weighted activity scores) for the same 25 studies for which butyl benzyl phthalate has positive effects.

Study Type	Organ	Baseline GenRA prediction	0.8 Filter refined GenRA prediction	Overall outcome prediction w.r.t naïve classifier
Chronic	Body Weight	0.78	0.55	P
Chronic	Clinical Chemistry	0.27	0.55	P
Chronic	Food Consumption	0	0	N
Chronic	Hematology	0	0	N
Chronic	Kidney	0.27	0.55	P
Chronic	Liver	1	1	P
Chronic	Mortality	0.27	0.55	P
Chronic	Pancreas	0.27	0	N
Chronic	Prostate	0	0	N
Chronic	Skin	0.27	0.55	P
Chronic	Spleen	0	0	N
Chronic	Tissue NOS	0	0	N
Chronic	Urinary Bladder	0	0	N
Developmental	Body Weight	1.00	1.00	P
Developmental	Bone	0.27	0.35	N
Developmental	Clinical Signs	0.00	0.00	N
Developmental	Eye	0.17	0.21	N
Developmental	Heart	0.00	0.00	N
Developmental	Kidney	0.10	0.00	N
Developmental	Liver	0.48	0.35	N
Developmental	Offspring Survival-Early	0.33	0.27	N
Developmental	Ovary	0.00	0.00	N
Developmental	Reproductive Performance	0.50	0.49	N
Developmental	Ureter	0.00	0.00	N
Developmental	Water Consumption	0.21	0.00	N

Table 5:

Baseline GenRA and Refined GenRA predictions (expressed as similarity weighted activity scores) for the same 25 studies for which butyl benzyl phthalate has positive effects.

Study Type	Organ	Baseline GenRA prediction	0.9 Filter Refined GenRA prediction	Overall outcome prediction w.r.t naïve classifier
Chronic	Body Weight	0.78	0	P
Chronic	Clinical Chemistry	0.27	0	N
Chronic	Food Consumption	0	0	N
Chronic	Hematology	0	0	N
Chronic	Kidney	0.27	0	N
Chronic	Liver	1	1	P
Chronic	Mortality	0.27	0	N
Chronic	Pancreas	0.27	0	N
Chronic	Prostate	0	0	N
Chronic	Skin	0.27	0	N
Chronic	Spleen	0	0	N
Chronic	Tissue NOS	0	0	N
Chronic	Urinary Bladder	0	0	N
Developmental	Body Weight	1.00	1.00	P
Developmental	Bone	0.27	0.35	N
Developmental	Clinical Signs	0.00	0.00	N
Developmental	Eye	0.17	0.21	N
Developmental	Heart	0.00	0.00	N
Developmental	Kidney	0.10	0.00	N
Developmental	Liver	0.48	0.35	N
Developmental	Offspring Survival-Early	0.33	0.27	N
Developmental	Ovary	0.00	0.00	N
Developmental	Reproductive Performance	0.50	0.49	N
Developmental	Ureter	0.00	0.00	N
Developmental	Water Consumption	0.21	0.00	N

Table 6:

Baseline GenRA and refined GenRA predictions (expressed as similarity weighted activity scores) for the same 25 studies for butyl benzyl phthalate.

Study Type	Organ	Baseline GenRA prediction	Equal weights refined GenRA prediction	Overall outcome predicted w.r.t naïve classifier
Chronic	Body Weight	0.78	0.79	P
Chronic	Clinical Chemistry	0.27	0.60	P
Chronic	Food Consumption	0	0.20	N
Chronic	Hematology	0	0.20	N
Chronic	Kidney	0.27	0.60	P
Chronic	Liver	1	0.80	P
Chronic	Mortality	0.27	0.40	N
Chronic	Pancreas	0.27	0	N
Chronic	Prostate	0	0	N
Chronic	Skin	0.27	0.21	N
Chronic	Spleen	0	0.20	N
Chronic	Tissue NOS	0	0	N
Chronic	Urinary Bladder	0	0	N
Developmental	Body Weight	1.00	1.00	P
Developmental	Bone	0.27	0.47	N
Developmental	Clinical Signs	0.00	0.22	N
Developmental	Eye	0.17	0.13	N
Developmental	Heart	0.00	0.00	N
Developmental	Kidney	0.10	0.11	N
Developmental	Liver	0.48	0.36	N
Developmental	Offspring Survival-Early	0.33	0.15	N
Developmental	Ovary	0.00	0.00	N
Developmental	Reproductive Performance	0.50	0.29	N
Developmental	Ureter	0.00	0.00	N
Developmental	Water Consumption	0.21	0.00	N

Table 7.

List of names and CasRNs for source analogues depicted in figure 10, as well as their bioactivity and physicochemical similarity scores.

Chemical	CasRN	Bioactivity Similarity (based on Apredica assays)	Physicochemical Similarity
Dibutyl phthalate	84-74-2	0.19	0.98
Monobenzyl phthalate	2528-16-7	0.0	0.95
Dipentyl phthalate	131-18-0	0.72	0.99
Dihexyl phthalate	84-75-3	0.63	0.91
Diisobutyl phthalate	84-69-5	0.62	0.95
Diallyl phthalate	131-17-9	0.11	0.95
Diethyl phthalate	84-66-2	0.06	0.85
Fenoxycarb	72490-01-8	0.65	0.98
Isofenphos	25311-71-1	Not tested	0.98
Cyhalofop-butyl	122008-85-9	0.70	0.98

Table 8:

Baseline GenRA and refined GenRA predictions (expressed as similarity weighted activity scores) for the same 25 studies for butyl benzyl phthalate.

Study Type	Organ	Baseline GenRA prediction	High weight search refined GenRA prediction	Overall outcome predicted w.r.t naïve classifier
Chronic	Body Weight	0.78	1	P
Chronic	Clinical Chemistry	0.27	0.80	P
Chronic	Food Consumption	0	0.60	P
Chronic	Hematology	0	0.80	P
Chronic	Kidney	0.27	0.80	P
Chronic	Liver	1	1	P
Chronic	Mortality	0.27	0.80	P
Chronic	Pancreas	0.27	0	N
Chronic	Prostate	0	0.20	N
Chronic	Skin	0.27	0.20	N
Chronic	Spleen	0	0.40	N
Chronic	Tissue NOS	0	0	N
Chronic	Urinary Bladder	0	0	N
Developmental	Body Weight	1.00	1.00	P
Developmental	Bone	0.27	0.33	N
Developmental	Clinical Signs	0.00	0.66	P
Developmental	Eye	0.17	0.00	N
Developmental	Heart	0.00	0.00	N
Developmental	Kidney	0.10	0.22	N
Developmental	Liver	0.48	0.22	N
Developmental	Offspring Survival-Early	0.33	0.12	N
Developmental	Ovary	0.00	0.00	N
Developmental	Reproductive Performance	0.50	0.33	N
Developmental	Ureter	0.00	0.00	N
Developmental	Water Consumption	0.21	0.22	N

Table 9.

List of names and CasRNs for source analogues depicted in figure 12, as well as their structural and physicochemical similarity scores.

Chemical Name	CasRN	Bioactivity Similarity (based on Apredica technology)	Physicochemical Similarity
Dibutyl phthalate	84-74-2	0.19	0.98
Dipentyl phthalate	131-18-0	0.72	0.99
Diisobutyl phthalate	84-69-5	0.62	0.99
Monobenzyl phthalate	2528-16-7	0.0	0.95
Pyriproxyfen	95737-68-1	0.74	.9995
Cloquintocet-mexyl	99607-70-2	0.0	0.996
2,4-D-Butotyl	1929-73-3	0.0	0.99
Chlorthal-dimethyl	1861-32-1	0.22	0.99
Fenpropathrin	39515-41-8	0.53	0.99
Cyhalofop-butyl	122008-85-9	0.70	0.98