Systematically evaluating read-across prediction and performance using a local validity approach characterized by chemical structure and bioactivity information

Imran Shah,^{1*} Jie Liu,^{2,3} Richard S Judson,¹ Russell S. Thomas¹, Grace Patlewicz¹

¹National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC, 27711, USA.

²Department of Information Science, University of Arkansas at Little Rock, AR, 72204, USA.

³Oak Ridge Institute for Science Education Fellow, National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC, 27711, USA.

*Corresponding author information:

Imran Shah

Tel: (919) 541 1391 Email: shah.imran@epa.gov

Abstract

Read-across is a popular data gap filling technique within category and analogue approaches for regulatory purposes. Acceptance of read-across remains an ongoing challenge with several efforts underway for identifying and addressing uncertainties. Here we demonstrate an algorithmic, automated approach to evaluate the utility of using *in vitro* bioactivity data (“bioactivity descriptors”, from EPA’s ToxCast program) in conjunction with chemical descriptor information to derive local validity domains (specific sets of nearest neighbors) to facilitate read-across for up to ten *in vivo* repeated dose toxicity study types. Over 3,239 different chemical structure descriptors were generated for a set of 1,778 chemicals and supplemented with the outcomes from 821 *in vitro* assays. The read-across prediction of toxicity for 600 chemicals with *in vivo* data was based on the similarity weighted endpoint outcomes of its nearest neighbors. The approach enabled a performance baseline for read-across predictions of specific study outcomes to be established. Bioactivity descriptors were often found to be more predictive of *in vivo* toxicity outcomes than chemical descriptors or a combination of both. This generalized read across (GenRA) forms a first step in systemizing read-across predictions and serves as a useful component of a screening level hazard assessment for new untested chemicals.

Word count (Abstract): 200

Keywords: read-across, nearest neighbors, local validity domains, (Q)SAR, KNN, bioactivity, ToxCast

1. INTRODUCTION

1.1 Problem formulation

Read-across is a widely used technique for filling data gaps within category and analogue approaches for regulatory purposes. In the European Union, under the Registration, Evaluation, Authorization and Restriction of Chemicals (REACH) regulation (EC, 2006), a ‘read-across’ or category approach has been used in up to 75% of analyzed dossiers for at least one endpoint (ECHA, 2014). A similar proportion of High Production Volume (HPV) chemicals submitted to the US EPA under the Toxic Substances Control Act (TSCA) have been initially evaluated using read-across approaches (Bishop et al., 2012). Despite the broad application of read-across approaches, regulatory acceptance remains a major hurdle primarily due to the lack of objectivity and clarity about how to practically address uncertainties. Currently, read-across is largely a subjective expert judgement-driven assessment in terms of both analogue selection and data interpretation (ECETOC, 2012; Cronin et al., 2013; Patlewicz et al., 2014a).

There is a continuum of non-testing methods¹ for relating chemical structural information (and *in vitro* bioactivity) to toxicity outcomes (Tollefsen et al., 2014). A specific approach is chosen dependent on the amount of available data, the endpoint of interest, and the regulatory context. When a large set of chemicals with known toxicity are available, then supervised machine learning methods can be used to automatically build and objectively evaluate the quantitative structure activity relationships (QSARs) derived. On the other hand, when there are only a handful of chemicals with toxicity data, then quantitative relationships between structural descriptors and activities are

¹ Non-testing approaches was a term coined during the development of the REACH Technical guidance to encompass the search and retrieval of existing information, the identification of structural alerts, to the development of (Q)SARs and the grouping of chemicals for read-across

1 more likely to be subject to “over-fitting,” often rendering those QSARs unreliable. In such data-poor
2 situations, read-across serves as a pragmatic option since experts can use prior knowledge to identify
3 appropriate (source) analogues and use them to infer activity of a new (target) chemical. Experts can
4 also substantiate the read-across prediction using mechanistic associations between structure and
5 activity, an explanatory quality that is often absent in QSARs, particularly global QSARs (Tollefsen et
6 al., 2014).

7 To gain regulatory acceptance, it is important to evaluate the uncertainty in read-across
8 predictions, and to improve scientific confidence by taking advantage of new *in vitro* bioactivity data
9 streams which have the potential to provide mechanistic information (Patlewicz et al., 2015a). We
10 believe this may be feasible by formulating read-across predictions in the context of a local validity
11 domain (Patlewicz et al., 2014a) that is amenable to objective evaluation using a QSAR-like framework
12 (OECD, 2004; 2007; ECHA, 2008).

13 In this study, we present a novel framework to formalize read-across based on a weight-of-
14 evidence approach that uses the activities of structurally related chemicals to infer the activity of a
15 new (target) chemical. We describe a unified approach to systematically evaluate confidence in read-
16 across predictions of toxicity using chemical structure, and a potential solution for reducing
17 uncertainty using *in vitro* bioactivity data. An approach was proposed by Low et al. (2013) to predict
18 the toxicity of a chemical by using a similarity-weighted activity of nearest neighbors (see Methods for
19 a more detailed description). Here we propose a generalization of the Low et al. (2013) approach,
20 called generalized read-across (GenRA), and use it to predict toxicity across structurally similar
21 neighborhoods in the ToxCast library containing 1,778 chemicals. Although GenRA is a naïve
22 approximation of expert judgement-driven read-across, we believe it is useful for establishing a

performance baseline upon which future improvements can be made. To reduce uncertainty of predictions of toxicity based on chemical structure alone, we evaluated the read-across performance by incorporating *in vitro* bioactivity descriptors based on 821 ToxCast assays. An important contribution of our work is the use of objective metrics to estimate confidence in read-across predictions of toxicity for untested target compounds using different number of source chemicals with known bioactivity.

1.2 Challenges with acceptance of current read-across approaches

Regulatory acceptance of read-across remains a particular challenge under REACH, especially for repeated dose toxicity endpoints as evaluated within guideline 90 day, reproductive or developmental toxicity studies (Ball et al., 2014; Patlewicz et al., 2014a). It has been postulated that formulating a framework to make explicit the critical information required for a read-across, and identifying the sources of uncertainties would be a helpful step forward, thus identifying where and what uncertainties may arise and practical ways to manage them (Patlewicz et al., 2014a). Various scientific experts have attempted to do this. Examples include the ECHA read-across assessment framework (RAAF), which is intended to assist evaluators in assessing read-across justifications within REACH dossiers (de Raat, 2014; ECHA, 2015). Researchers at P&G have proposed a framework to evaluate the uncertainties associated with a read-across framework particularly for Developmental and/or Reproductive Toxicity (DART) endpoints (Blackburn and Stuard, 2014, Blackburn et al., 2015). The EU Safety Evaluation Ultimately Replacing Animal Testing (SEURAT-1) Research Initiative developed templates to document uncertainty in read-across (Schultz et al., 2015). Finally, a read-across team established under the Long Range Initiative Programme of the European Chemistry Industry Council

(Cefic LRI) proposed scientific confidence considerations in the development and evaluation of read-across justifications (Patlewicz et al., 2015a).

Whilst all these efforts are helpful starting points, none have so far tackled practical ways of reducing uncertainties in a read-across prediction. The framework by Blackburn and Stuard (2014) put forward assessment factors to address uncertainties, these are appropriate for endpoints where a dose descriptor or point of departure (POD) such as a No Adverse Effect Level (NOAEL) can be established. An assessment factor could then be used to adjust the POD value being taken up in the subsequent risk assessment. However, use of the Blackburn and Stuard (2014) framework could result in a reference dose being derived (e.g., the derived no effect level (DNEL)), that is too low to be practically useful in the context of risk management. An assessment factor approach also can not address the uncertainty in a read-across for binary endpoints such as skin sensitization or mutagenicity. Patlewicz et al. (2015a) proposed two strategies in an effort to build scientific confidence in read-across for specific decision contexts, either making use of an adverse outcome pathway (AOP) (Ankley et al., 2010; Patlewicz et al., 2015b; Tollefsen et al., 2014) or through the use of bioactivity data such as that generated within ToxCast (Judson et al., 2010), assuming an AOP was not available or where the mechanism was insufficiently understood to develop one. Ultimately, the goal is to make use of appropriate mechanistic information to reduce the uncertainty in a read-across prediction (Patlewicz et al., 2015a)

Whilst the approaches proposed by Patlewicz et al. (2015a) may provide enhancements to current read-across, which is still very much rooted in chemical structural similarity, neither addresses the issue of read-across performance. Read-across remains an expert, judgement-driven assessment that is context dependent on the chemical and endpoint under consideration. There is presently an

1 absence of illustrative case studies to clarify how a regulatory agency will evaluate read-across
2 justifications, such that an objective framework to systematically evaluate read-across performance is
3 very much needed.

4 Indeed one can contrast the expert driven approach typically used in read-across with the
5 automated (Q)SAR approach, which also uses information on structurally-related chemicals, but does
6 so in a way that is objective, and amenable to making estimates of uncertainty in predictions (Cronin
7 et al., 2013; Patlewicz et al., 2015c). Both approaches are underpinned by the notion of toxicity being
8 a function of chemical structure, but they simply represent different regions along the continuum of
9 non-testing approaches. Thus, the principles described for (Q)SAR application such as those outlined
10 for REACH (ECHA, 2008) could be helpful. In particular, the framework for assessing a QSAR prediction
11 – whether the model addresses a defined endpoint, to what extent the QSAR is scientifically valid with
12 respect to the OECD Validation Principles (OECD, 2004; 2007) and whether the substance of interest
13 falls within the applicability domain (ECHA, 2008; Netzeva et al., 2005) could also be relevant for read-
14 across. For a QSAR, the applicability domain is usually extracted from the training set used to develop
15 the model itself, and there are a number of different approaches that can be employed to derive a
16 domain (Netzeva et al. 2005; Dimitrov et al. 2005). For SARs, the concept of an applicability domain is
17 still evolving (see Ellison et al. (2011) for a summary of several different approaches). More recently,
18 Patlewicz et al. (2014b) proposed the concept of a structural alert reliability within the context of the
19 hybrid expert system Tissue Metabolism Simulator for Skin Sensitisation (TIMES-SS) (Patlewicz et al.,
20 2007; Patlewicz et al., 2014b). The scientific confidence of relying on a structural alert was defined by
21 three considerations including: the extent to which the alert prediction was correct; the number of
22 chemicals underpinning the alert; and, whether a reaction chemistry mechanistic justification had
23 been proposed for the alert.

Another key consideration in the utility and confidence associated with the prediction from a (Q)SAR is the extent to which the activities caused by a chemical can be accurately predicted from the activities of “similar” chemicals. This prediction accuracy will depend on how similarity is defined resulting in an appropriate set of nearest neighbors or the so-named “local validity domain”. The notion of a local validity domain in the context of a read-across was discussed previously in Patlewicz et al. (2014a).

Using nearest neighbors to make predictions has been implemented in a variety of software tools. Within the OECD QSAR Toolbox (available at <http://www.qsartoolbox.org/>), a nearest neighbor approach is available to make an endpoint read-across prediction; however, the prediction is based on a single descriptor (either chemical or biological in nature). Other similarity software tools such as Toxmatch (Patlewicz et al., 2008) or ToxRead (Gini et al., 2014) enable nearest neighbor predictions to be made using different similarity indices. It is worth noting that nearest neighbor-based methods are common (Q)SAR approaches, for instance in the k-nearest neighbors (kNN) machine learning approach (Aha and Kibler, 1991).

Despite the progress made in the read-across field, the use of biological and chemical information in concert as a means of performing a read-across has not been investigated to any great extent and certainly not in the context of reducing uncertainty in traditional read-across predictions. The exception is the work of Low et al. (2013), who described an automated approach to assign the toxicity of a chemical by calculating the weighted aggregate of activities of k-nearest neighbors. The approach was termed Chemical-biological read-across (CBRA) because it integrated chemical structural descriptor information with bioactivity data to predict toxicity. This approach is described in brief in the methods section.

1 In this study, we present a novel framework to systematically derive a read-across prediction
2 for a new chemical based on the toxicities of neighboring chemicals as characterized by their chemical
3 descriptor and/or bioactivity information. In addition, we demonstrate a unified approach to
4 systematically evaluate the uncertainty in the read-across predictions of toxicity. This exercise forms
5 a first step in systemizing a read-across prediction approach that is agnostic to chemistry and
6 mechanistic insights, but could nonetheless serve as a useful tool as part of a screening-level
7 assessment for new untested chemicals.

8 **2. MATERIALS AND METHODS**

9 **2.1 Overview of the approach**

10 Our analysis used a large dataset of environmental and pharmaceutical chemicals from ToxCast
11 (Kavlock et al., 2012) and ToxRefDB (Martin et al., 2007). This dataset comprised 1,778 chemicals with
12 bioactivity measurements from ToxCast, including 821 high-throughput screening (HTS) assays, and
13 data from up to 10 different *in vivo* guideline testing study types. We first subcategorized the
14 chemicals into clusters using structural similarity. The initial clustering was intended to mirror the
15 workflow of identifying preliminary source analogues through structural similarity. Next, we evaluated
16 the predictivity of a modified CBRA approach, which we have termed GenRA (Generalized Read-
17 across), to systematically mine the neighborhoods of chemicals to predict *in vivo* study outcomes.
18 Using information about the known apical effects of chemicals, we evaluate confidence in GenRA
19 predictions using receiver operating characteristic (ROC) area under the curve (AUC). We visually
20 summarized the performance results across groups of structurally-related chemicals to intuitively
21 evaluate the read-across predictions made.

2.2 Data sources.

2.2.1 Bioactivity data.

The *in vitro* assay data were generated from the high-throughput screening of the 1,776 ToxCast phase I and phase II compounds and is publicly available (<http://www.epa.gov/ncct/toxcast/data.html>, version Dec2013). 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. For the machine learning analysis, all active and inactive assay results were converted to binary values (active=1 and inactive=0, respectively). The bioactivity data was stored in a matrix where rows corresponded to chemicals and columns to bioactivity assays (denoted as, X^{bio}). The bioactivity of each chemical (i) is given by: $X_i^{bio} = \{x_i^1, x_i^2, x_i^3, \dots, x_i^{bio}, \dots, x_i^{n_{bio}}\}$, where x_i^{bio} represents a bioactivity assay (bio) result. When more than one bioactivity values was available for a given chemical-assay pair, the results were combined by a logical OR operation (that is, $x^{bio} = x_i^{bio} \vee x_j^{bio}$). The total number of bioactivity assays are denoted as n^{bio} . All bioactivity data are provided as supplemental material (S1).

2.2.2 Toxicity data.

1 Toxicity data were obtained from ToxRefDB (Martin et al., 2007), which describes the *in vivo*
2 outcomes of hundreds of compounds from animal testing studies. All data are publicly available at
3 ftp://newftp.epa.gov/COMPTOX/Animal_Tox_Data/. ToxRefDB includes information about the effects
4 of chemicals across different species and target organs that were observed in US EPA, OECD or other
5 regulatory guideline or guideline like studies. For this analysis, *in vivo* outcomes for 600 chemicals
6 taken from ToxRefDB were aggregated at the level of a study type and target site of the effects (e.g.
7 the target organ affected). The study types included: chronic toxicity (chr), developmental toxicity
8 (dev), developmental neurotoxicity (dnt), multigenerational toxicity (mgr), neurotoxicity (neu),
9 reproductive toxicity (rep), acute toxicity (acu), sub-acute toxicity (sac), and sub-chronic toxicity (sub).
10 All other toxicity testing studies (i.e. where a specific guideline was not reported) were grouped into a
11 category that was referred to as “other” (oth). There were 129 unique targets in ToxRefDB (found in
12 the column “effect_target” of the ToxRefDB data file). If a chemical produced a significant effect in a
13 particular study type and target at any dose, then it was categorized as positive for that toxicity type.
14 For example, if a chemical produced hepatotoxicity or nephrotoxicity in a sub-chronic study then it
15 was considered a “sub-chronic liver toxicant” (denoted as, sub_liver), or “sub-chronic kidney toxicant”
16 (denoted as, sub_kidneys). For example, a subchronic study in rats was reported in ToxRefDB for
17 Perfluorooctanesulfonic acid, potassium salt PFOS-K (CAS# 2795-39-3), where liver hypertrophy was
18 reported in both sexes dosed at 0.25 and 1 mg/kg/day. These were statistically-significant treatment-
19 related effects, and therefore, resulted in an assignment of a “positive” call for subchronic liver
20 effects. If a substance did not produce any statistically significant treatment-related effects in a study,
21 then a negative call was assigned. We recognize that summarizing different effects across different
22 types of clinical outcomes in different species oversimplifies complex pathology. However, in the
23 interests of investigating whether any read-across inferences could be made and quantified for a large

number of substances, this type of aggregation was a necessary yet pragmatic approach in order to perform the subsequent computational analysis. In summary, the toxic effects of a chemical (denoted as, X^{tox}) were represented as a multi-dimensional vector, $X_i^{tox} = \{x_i^1, x_i^2, x_i^3, \dots, x_i^{tox}, \dots, x_i^{n_{tox}}\}$, in which each element states the existence of at least one statistically significant treatment-related effect in a particular *in vivo* study type. There were 574 unique types of toxic effects across the ten different study types and 129 targets. All toxicity data are provided as supplemental material (S2).

2.2.3 Chemical structure data.

The chemical descriptors generated relied upon structural fingerprints that have been previously published in the literature. Fingerprints were represented as binary (bit) vectors where the elements themselves represent the presence or absence of a certain feature. Three different methods for calculating structural fingerprints were used including: Morgan fingerprints (Rogers and Hahn 2010); and, 166 publicly available Molecular Access System (MACCS) keys and topological torsion descriptors (Nilakantan, Bauman et al. 1987). The chemical structure data was stored in a matrix in which rows corresponded to chemicals and columns to descriptors (denoted as, X^{chm}). The total number of chemical descriptors were denoted as n^{chm} . The chemical structure descriptors were calculated using the freely available python RDKit cheminformatics library (Landrum, 2015). All chemical structure data are provided as supplemental material (S3).

2.3 Clustering chemicals by structural similarity into local neighborhoods.

All chemicals were grouped by similarity across all structural descriptors using K-means clustering. To identify the optimal number of partitions (K) in X^{chm} , a subspace sampling approach was derived to evaluate cluster stability for $K=10$ to $K=120$. For each value of K , a random subset (defined by the

1 fraction, f_k) of chemicals (denoted as, X') was selected and then used to predict the clustering for all
 2 chemicals (denoted as, C'). This process was repeated n times yielding a set of predicted clusters,
 3 $C'_k = \{C'_{k1}, C'_{k2}, \dots, C'_{ki}, \dots, C'_{kn}\}$. The pairwise stability, denoted as s_{kij} , between clusters, C_{ki} and C_{kj} ,
 4 was calculated using the adjusted random score (Hubert and Arabie, 1985). The cluster stability score
 5 (s_k) for each value of K was calculated as the mean value across all pairwise scores s_{kij} (where
 6 $s_k = \frac{2}{n(n-1)} \sum_{i,j} s_{kij}$). We calculated s_k for $f_k = 0.7$ and identified the value of K (denoted as K^*)
 7 that produced the maximum cluster stability score (s_k^*). The optimal partitions for all chemicals
 8 (denoted as C^*) was then calculated by K-means analysis of the complete data using $K = K^*$.
 9 The chemical clustering C^* partitioned the chemicals into K^* clusters given by,
 10 $C^* = \{C_1, C_2, \dots, C_l, \dots, C_{K^*}\}$. Each cluster, C_l , contained a subset of chemicals, n_l , a matrix of chemical
 11 descriptors (denoted as, X_i^{chm} , where $X_i^{chm} \subset X^{chm}$), and a matrix of bioactivity assays (denoted as,
 12 X_i^{bio} , where $X_i^{bio} \subset X^{bio}$). The number of the subset of chemical and bioactivity descriptors present
 13 in C_l are given by n_l^{chm} and n_l^{bio} , respectively.

14 **2.4 Summary of CBRA approach**

15 The CBRA approach has been used to assign the toxicity of a compound (denoted as A_{pred}) using
 16 different descriptors to measure similarity. In the first method, A_{pred} was calculated purely by chemical
 17 descriptors as the Tanimoto similarity (S_i)-weighted aggregate of activities A_i of k -nearest neighbors
 18 using the following equation (Equation 1):

$$A_{pred,RA} = \frac{\sum_i^k S_i A_i}{\sum_i^k S_i} \quad [1]$$

The CBRA activity of a substance is then calculated using a combination of chemical and biological neighbors using the following equation (Equation 2):

$$A_{pred,CBRA} = \frac{\sum_i^{k_{bio}} S_i A_i + \sum_j^{k_{chm}} S_j A_j}{\sum_i^{k_{bio}} S_i + \sum_j^{k_{chm}} S_j} \quad [2]$$

Chemical activity is calculated using Equation 2 by pooling the descriptors to form a hybrid space (CBRA-hybrid).

2.5 Predicting toxicity using nearest neighbors – Generalized Read-Across (GenRA)

The CBRA approach was generalized to predict the toxicity outcome of the different study types of all chemicals using a weighted sum of toxicity information from nearest neighbors. For each chemical (denoted as c_i), the toxicity outcome was calculated using the activities of k -nearest neighbors (denoted by $NN_i = \{c_{i1}, c_{i2}, \dots, c_{ij}, \dots, c_{ik}\}$) based on the Jaccard similarity (also known as Tanimoto distance). The Jaccard similarity s_{ij} between two chemicals c_i and c_j was calculated using their vector representations (x_i and x_j , respectively) as $s_{ij} = \frac{\sum_l (x_{il} \wedge x_{jl})}{\sum_l (x_{il} \vee x_{jl})}$. For each pair of chemicals, the chemical similarity (s_{ij}^{chm}), biological similarity (s_{ij}^{bio}) and biological+chemical similarity (s_{ij}^{bc}) were calculated using the chemical (chm), biological (bio) and hybrid chemical and biological (bc) descriptors, respectively. The predicted read-across toxicity of a chemical (denoted as y^{tox}) was calculated based on k -nearest neighbors using Equation 3.

$$y_i^{tox} = \frac{\sum_j^k s_{ij}^\alpha x_j^{tox}}{\sum_j^k s_{ij}^\alpha} \quad [3]$$

Where x_j^{tox} , in this case, is the *in vivo* toxicity of chemical j . *In vivo* toxicity outcomes of chemicals were predicted using chemical, bioactivity and hybrid descriptors (that is $\alpha = \{chm, bio, bc\}$) for $k = 2$ to 10, where k is the number of nearest neighbors. In the case of hybrid descriptors (that is $\alpha = bc$), GenRA is identical to CBRA (Low et al., 2013).

2.6 Prediction accuracy of GenRA

The prediction accuracy of each toxicity outcome across all chemicals in a local neighborhood was evaluated in two ways. Receiver operating characteristic (ROC) analysis of the predicted (Y^{tox}) and true toxicities (X^{tox}) was conducted for k nearest neighbors (where the value of k ranged from 1 to the maximum number of chemicals in the neighborhood), and with a similarity threshold, s (where the value of s ranged from the minimum to maximum values of s across all unique pairwise comparisons in the neighborhood). The area under the curve (AUC) was then taken as a measure of performance for a given k and s value. In addition, the overall performance across all values of k and s in a local neighborhood was summarized as the volume under the surface (VUS) of the ROC scores.

In ROC analysis, the activity threshold (θ_o^{tox}) is systematically varied from the minimum ($\min(\theta^{tox})$) to the maximum value $\max(\theta^{tox})$ to calculate the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). The performance was measured by plotting the false positive

rate ($FPR = \frac{FP}{FP + TN}$) against the true positive rate ($TPR = \frac{TP}{TP + FN}$), which is known as the ROC

curve. The area under the ROC (called the AUC) was used as a measure of performance. The

significance was empirically estimated by constructing a null distribution by permuting X^{tox} 100 times and calculating the fraction of times the AUC was more extreme than what would be observed by chance (reported as the p-value). The predicted activity threshold, θ_o^{tox} , that produced maximum balanced accuracy (*BA*), sensitivity (*TPR*) and specificity (*1-FPR*) was also reported. The ROC scores calculated in this manner have a minimum value of 0 (which means no predictive ability) and a maximum value of 1 (which means perfect prediction).

The AUC was used to objectively measure the predictive performance for GenRA in a local neighborhood for a single value of k and s . The AUC can vary considerably as k and s change, which makes it difficult to compare the predictive performance of the 3 different descriptor approaches across a neighborhood or indeed the performance for the same neighborhood across different toxicity effects. The ROC scores for GenRA across the range of k and s in a neighborhood can be represented as a three dimensional surface. Calculating the volume under the ROC surface (VUS) provides a single metric to compare the performance across the entire neighborhood. Consider the ROC scores for a neighborhood with 11 chemicals and Jaccard similarity values ranging from 0.2 to 0.8. The ROC surface is defined by values of k ranging from 1 to 11 and values of s ranging from 0.2 to 0.8. In the ideal case, the ROC score for each value of k and s would be 1 and therefore the absolute VUS would be 100 ($VUS = \sum_{k=1}^{k=11} \sum_{s=0.2}^{s=0.8} ROC_{ks}$). In reality, the absolute VUS for a specific toxicity outcome will be less than the ideal value. If the absolute VUS was 60 for sake of argument, the relative VUS would be a fraction of the ideal VUS, i.e. 60/100 or 0.6. The utility of this relative VUS is that it can be compared across different toxicities and across the entire neighborhood. The relative VUS for each toxicity outcome for each neighborhood was calculated as a summary measure of performance for GenRA.

2.7 Software.

Data processing and analysis was conducted in the Python programming language (version 2.7) using RDKit (version 2014-09-02) (Landrum 2015), and the matplotlib package (version 0.99.1.2) (Hunter, 2007) for visualization. The source code is available from the authors upon request.

3 RESULTS

3.1 Chemical structure clusters.

Using cluster stability analysis, 100 clusters were found to be optimal in partitioning the chemicals. These 100 clusters are depicted as leaves in the circular dendrogram shown in Figure 1. The number of chemicals in each cluster is represented by the area of the circular leaf nodes. A subset of the clusters readily lent themselves to structural chemical class labels, including: 13 perfluorinated acids (PFAAs) (cluster 0), 22 pyrrolidones (cluster 3), 6 cyclodienes (cluster 4), 29 nitrobenzene-containing chemicals (cluster 5), and 12 benzoylurea chemicals (cluster 16). These labels were used for identification and illustrative purposes only. In many cases, the clusters could not be easily labelled by a single chemical class description. Examples include clusters 19 and 35. Cluster 19 comprised 117 chemicals whereas Cluster 35 contained 369 chemicals. In both, a broad range of different chemical classes were covered spanning from aromatic nitroso compounds to aliphatic diketones to glycols to carbamates.

This highlights a potential limitation of the structural features used in the automated clustering approach which formed groupings that an expert cannot readily identify with. In this analysis, a subset of the 61 clusters that contained at least two or more chemicals with associated *in vivo* animal testing data were selected for further evaluation. Fewer than 2 chemicals with associated *in vivo* animal

testing would not permit a read-across prediction to be derived. Information about the membership of chemicals in all chemical structure clusters and the assignment of clusters to known chemical groups is provided as supplemental material (S4).

3.2 Visualizing and predicting toxicity using nearest neighbors.

For each chemical in every cluster, the similarity was calculated, the nearest neighbors identified and the *in vivo* toxicity summary outcomes predicted using chemical, bioactivity and hybrid descriptors by using GenRA (applying Equation 3). For example, cluster 0 contained 13 Perfluoroalkyl Acids (PFAAs) that produced some effects in three study types including: developmental, multigenerational and subchronic studies. In all, the 13 PFAAs produced 50 different types of toxicity effects. Figure 2 shows an example of the chemical descriptor-based nearest-neighbors for perfluorohexanoic acid (PFHxA). In this case, eight nearest-neighbors of PFHxA were identified and visualized by organizing them radially in an anti-clockwise manner, starting with most similar chemical and ending with the least similar chemical. The connections between PFHxA and the nearest neighbors reflect the magnitude of their structural similarity. The closest nearest-neighbor of PFHxA based on the chemical structural similarity was perfluoroheptanoic acid (PFHpA), with a Jaccard similarity of 0.94. In Figure 2, the subscript c is used to denote similarity with respect to chemical descriptors. The other neighbors of PFHxA in descending similarity were: perfluorodecanoic acid (PFDA), perfluorononanoic acid (PFNA), perfluorooctanoic acid (PFOA), perfluoroundecanoic acid, ammonium salt of perfluorooctanoic acid, perfluorooctane sulfonic acid (PFOS) and potassium perfluorohexanesulfonate (PFHS-K) with Jaccard

1 similarities (to two decimal places) of 0.94, 0.91, 0.91, 0.91, 0.91, 0.63, 0.32 and 0.29, respectively. It
2 is worth noting that based on the 3,239 chemical structure descriptors and Jaccard metric used in this
3 analysis, PFHxA actually had the same level of similarity (0.91) with PFDA, PFNA and PFOA.

4 The visualization in Figure 2 also shows nine illustrative examples (out of the 50 effects for the 13
5 chemicals in cluster 0) that were observed *in vivo* as obtained from ToxRefDB including:
6 developmental effects on food consumption (dev_food_consumption), sexual developmental
7 landmark (dev_sexual_developmental_landmark); multigenerational effects on brain (mgr_brain),
8 and motor activity (mgr_motor_activity); subchronic effects on lungs (sub_lungs), sperm morphology
9 (sub_sperm_morphology), testes (sub_testes), and thymus (sub_thymus). The results across the 9
10 toxicity outcomes for each chemical were summarized visually in Figure 2 as a fingerprint in which
11 green denotes negative, red denotes positive and grey denotes not tested (e.g., PFHpA did not have
12 any guideline testing data in the current version of ToxRefDB). In this example, PFHxA was found to be
13 experimentally negative for subchronic effects on lungs, sperm morphology, and testes, and there
14 was no data on the developmental or multigenerational effects (that is, there was no information
15 about developmental or multigenerational studies in the version of ToxRefDB that was used in this
16 analysis).

17 After summarizing the effects of chemicals in terms of *in vivo* outcomes, the diverse effects for the
18 chemicals in local neighborhoods were mathematically analyzed using GenRA. Equation 3 (which
19 defines the similarity weighted toxicity estimation function) was used to predict the toxicity outcome
20 of PFHxA for each of the 9 *in vivo* outcomes using the nearest neighbors. The similarity weighted
21 activity scores for the 9 *in vivo* outcomes including: dev_food_consumption,
22 dev_sexual_developmental_landmark, dev_sperm_morphology, mgr_brain, mgr_motor_activity,

sub_lung, sub_sperm_morphology, sub_testes and sub_thymus, were found to be 0, 0, 0, 0, 1, 0, 0, 0. Thus PFHxA is predicted to have potential for motor activity effects in multigenerational studies but no effects for the other developmental, multigenerational or subchronic effects. The GenRA score calculated from Equation 3 is usually a numeric value between 0 and 1 that requires a threshold value (θ_o^{tox}) to assign a positive (1) or negative (0) prediction. This threshold is not defined arbitrarily but is determined by analyzing the GenRA scores within the neighborhood (see methods). In the case of PFHxA, a threshold of 0.5 was appropriate to determine the GenRA predictions, which are shown above the observed *in vivo* results for PFHxA in Figure 2.

3.3 Evaluating GenRA predictions for all chemicals in a neighborhood,

The GenRA score was calculated to estimate the *in vivo* effects for all chemicals in cluster 0. Translating the similarity-weighted read-across score derived from Equation 3 into a prediction of toxicity requires the determination of a threshold as noted above. The similarity-weighted GenRA scores for each *in vivo* outcome were used to calculate the range of scores across all chemicals in this neighborhood. The threshold score was then varied from the minimum value to the maximum value. As the threshold score varied, so did the predicted outcome for each chemical. For example, for a threshold of 0.1 (i.e. $\theta_o^{sub_testicle} = 0.1$), the subchronic testicular effects of PFHxA were predicted to be negative (because the GenRA score of 0 was less than the threshold of 0.1). Since PFHxA also did not produce any subchronic testicular effects, this prediction was considered a true negative (TN) at a score threshold of 0.1. Using a similar approach, the variation in TN, TP, FN and FP for all values of $\theta_o^{sub_testicle}$ for all chemicals in the neighborhood was analyzed (see Methods). The variation in

sensitivity and specificity for all similarity-weighted toxicity score thresholds was systematically analyzed by comparison with the true toxicity outcomes by ROC analysis, and the AUC score reported. Randomization was used to estimate the empirical significance of the AUC score and reported as the p-value. This procedure allowed for the predictive accuracy of the read-across for a given neighborhood using a single fixed number of neighbors (8 in this case) to be estimated.

3.4 Evaluating the read-across performance across an entire neighborhood.

The predictive accuracy depends on the size of the local neighborhood and this relationship was systematically evaluated for each cluster, *in vivo* outcome and descriptor type. The size of the neighborhood for each chemical was varied by adjusting two parameters, namely the number of nearest neighbors (k) and the minimum similarity threshold (s). Figure 3 shows the relationship between k , s and predictive accuracy (AUC) for multigenerational motor activity effects in cluster 0 as a contour graph. The row of graphs reflects the performance of the three sets of descriptors (chemical (chm), biological (bio) and hybrid (bc)). In this specific case, the performance was highly dependent on the type of descriptors used. Bioactivity descriptors were found overall to be poor predictors (evidenced by the blue color in the contour plot) whereas chemical structure descriptors could predict this *in vivo* outcome (as shown by the red color in the contour plot). This is also summarized in the VUS metric. There is a complex relationship between similarity thresholds (s) and nearest neighbors (k) that could accurately predict this outcome. Calculating the VUS provides as measure of the overall predictive accuracy across all values of k and s . The predictive accuracy of GenRA for multigenerational motor activity effects in cluster 0 for chemical, bioactivity and hybrid descriptors was determined to be 0.75, 0.05 and 0.11, respectively (to two decimal places) i.e. the predictive accuracy was indeed highest for chemical descriptors for this effect.

3.5 Summary of the toxicity predictions.

A summary of the predictive accuracy of GenRA for chemical, biological and hybrid chemical and biological descriptors for cluster 0 for multigenerational effects is shown in Table 1. Each row in Table 1 summarizes the performance results for each cluster-study-target combination using the 3 sets of descriptors. Performance results are summarized in the columns by each descriptor type. For instance, the first row shows the GenRA accuracy for predicting multigenerational effects on the adrenal gland (mgr_adrenal_gland) for 2 positive and 2 negative PFAAs. Using chemical structure descriptors, the predictive accuracy is as follows: VUS of 0.41, best ROC AUC score of 1 with p-value less than 0.1. The performance for bioactivity descriptors is poorer with metrics as follows: VUS of 0.1, best ROC AUC score of 0.55 with p-value greater than 0.1. For the hybrid bioactivity and chemical descriptors, the performance had intermediate values given by: VUS of 0.09, best ROC AUC score of 0.62 with p-value less than 0.1. For all the multigenerational effects in cluster 0, chemical descriptors outperform the bioactivity descriptors. The complete performance results for all clusters and toxicity outcomes are attached in supplemental material (S5).

A summary of GenRA predictions at the level of studies and descriptor types is given in Table 2 for cluster 0. Rows in Table 2 show the numbers of study-target organ toxicities that can be accurately (defined as ROC AUC>0.8) predicted by the three sets of descriptors. For example, for multigenerational studies (Table 2, row 4), 6 outcomes can be accurately predicted by GenRA using chemical descriptors, 4 outcomes can be accurately predicted using bioactivity descriptors, and only 1 outcome can be predicted using hybrid descriptors. Further, 5 out of the 6 outcomes can only be accurately predicted by chemical descriptors while all 4 out of the 4 outcomes can be predicted only by bioactivity descriptors. The identity of these outcomes can be found using supplemental material

(S5). Overall, for chemicals in cluster 0, chemical-structure based read-across predicted *in vivo* outcomes were more accurate than through using bioactivity descriptors.

Many examples of clusters where bioactivity based descriptors performed much better than chemical structure descriptors can be found using the first table (labeled, Summary) in supplemental material (S5). For instance, using the S5 Summary table, bioactivity descriptors were better at predicting developmental *in vivo* effects on mortality and the spleen than chemical descriptors for the chemicals in cluster 5 (nitrobenzenes).

While it is not feasible to discuss the results of applying GenRA to each cluster in turn, we have selected some illustrative examples. The Summary table in supplemental material S5 is a useful starting point for identifying clusters, and the types of toxicity effects observed. For instance, cluster 80 contained a set of 41 chemicals (18 tested in guideline studies and 23 untested chemicals) with an average of 113 chemical structure descriptors, 72 bioactivity descriptors and 12 apical outcomes. To identify the specific chemicals in this cluster (and citations for the *in vivo* studies) see the sheet labeled "Cluster-80" in supplemental material S4. The Summary table in supplemental material S5 shows the GenRA performance results for cluster 80 (in rows 221 to 225) in the same format as Table 2. For cluster 80, we found that chemical structure descriptors accurately predicted 37 chronic, 17 developmental, 25 multigenerational, 9 subacute and 29 subchronic effects. Bioactivity descriptors predicted 19 chronic, 14 developmental, 16 multigenerational, 2 subacute and 16 subchronic effects. Hybrid descriptors predicted 32 chronic, 17 developmental, 16 multigenerational, 9 subacute and 25 subchronic effects. Only chemical structure descriptors could accurately predict 6/37 chronic, 11/35 multigenerational, and 4/29 subchronic effects, bioactivity descriptors alone only predicted 1/19 chronic, 1/14 developmental, 2/2 subacute and 3/16 subchronic effects, and hybrid descriptors were

required to predict 1 chronic and 1 developmental effect. The last three columns show the number of cases (where each case is an *in vivo* outcome per cluster) in which chemical, bioactivity and hybrid descriptors, respectively, provide the best predictions.

The performance for specific effects are given in supplemental material S5 under in the sheet labeled, "Cluster-80," which has the same format as Table 1. A subset of these results are shown in Table 3. The columns show the performance results for chemical (chm), bioactivity (bio) and hybrid (bc) descriptors. For chronic effects on the kidneys, hybrid descriptors out-performed chemical and bioactivity descriptors based on AUC (which was 0.78, 0.78 and 1.00 for chm, bio and bc, respectively). This means that the optimal value for the GenRA AUC for hybrid descriptors for certain number of nearest neighbors (k) and similarity (s) is greater than the best AUC for chemical and bioactivity descriptors. It is important to note that the VUS for chm, bio and bc descriptors was 0.55, 0.27, and 0.32, respectively. The difference between AUC and VUS can be interpreted in terms of the relationship between predictive performance and the range of k and s. The VUS measures the performance across entire neighborhood (i.e. all k and s) whereas, the optimal AUC measures the performance in a specific region of the neighborhood (i.e. a narrow range of k and s). This has implications for using GenRA to predict *in vivo* effects of untested chemicals.

3.6 Comparing the performance for different descriptors.

For each of the 61 clusters (each with two or more chemicals), the accuracy of GenRA for predicting many *in vivo* outcomes in different types of guideline testing studies using three types of descriptors. The comparison of the different descriptors for predicting *in vivo* outcomes across all clusters are

1 shown in Table 3. For the purpose of this comparison, only high performance scores, where ROC
2 AUC>0.8 and p<0.01, were considered. The rows in Table 3 denote toxicity outcomes and the columns
3 denote descriptors (or combinations of descriptors). Each element shows the number of cases for
4 which the toxicity (row) could be accurately predicted by using the descriptor (column) of a certain
5 type. From left to right, the second column (labeled, Cases) shows the number of cluster-study-effect
6 combinations (called cases henceforth), which represent the unique number of study level effects
7 across all clusters; the third column (labeled, chm) shows results for chemical structure descriptors,
8 which accurately predicted chronic toxicity for 97 cases; the fourth column shows the results for
9 bioactivity descriptors (labeled, bio), which accurately predicted chronic toxicity for 133 cases; and
10 the fifth column shows the results for hybrid descriptors (labeled, bc), which accurately predicted
11 chronic toxicity for 67 cases. Columns 6, 7 and 8 (labeled, "chm>bio|bc", "bio>chm|bc" and
12 "bc>chm|bio") show the cases for which chemical, bioactivity and hybrid descriptors, respectively,
13 had the best accuracy. The ninth column (labeled, "Any") shows the number of cases in which any of
14 the descriptors produced accurate predictions. For chronic outcomes, for example, there were 203
15 cases in which bio, chm or bc descriptors produced accurate results. From top to bottom, the toxicity
16 outcomes are organized by different guideline testing study types including: chronic toxicity (chr),
17 developmental toxicity (dev), developmental neurotoxicity (dnt), multigenerational toxicity (mgr),
18 reproductive toxicity (rep), subacute toxicity (sac), subchronic toxicity (sub), and all types of toxicity
19 studies (All). For instance, there were 1115 cases of chronic toxicity, out of which 9% (97/1115) could
20 be predicted by chemical structure descriptors, 12% (133/1115) by bioactivity descriptors, and 6%
21 (67/1115) by hybrid descriptors. Similarly, out of the 84 developmental neurotoxicity cases, 19% could
22 be predicted by chemical descriptors and 31% by bioactivity descriptors. The last row (labeled, All)

shows a comparison of performance across all descriptors. Overall, bioactivity descriptors outperformed other descriptors for inferring toxicity using GenRA for 16% (570/3663) of the cases.

3.7 Illustrative case example of GenRA toxicity predictions using Benoxacor

The practical application of GenRA to predict the *in vivo* effects of benoxacor using chemical descriptors and 10 nearest neighbors (i.e. $k=10$) is shown in Figure 4(a). For the sake of brevity, the visualization is limited to chronic effects on brain (chr_brain); multigenerational effects on body weight (mgr_body_weight), bones (mgr_bone), large intestines (mgr_intestine_large), urinary bladder (mgr_urinary_bladder); and subchronic effects on the small intestine (sub_intestine_small), prostate gland (sub_prostate), spleen (sub_spleen), urinalysis (sub_urinalysis), water consumption (sub_water_consumption). GenRA predictions of toxicity (positive and negative predictions are shown as orange and blue, respectively) were consistent with the experimentally observed effects (positive, negative and unknown toxic effects in studies are shown as red, green and grey, respectively). In the case of benoxacor, GenRA correctly predicted the multigenerational effects on body weight and absence of effects on bone and large intestine. In addition, GenRA also predicted the potential absence of chronic effects on the brain and other subchronic effects.

The relationship between the AUC, k and s for different effects on across the local neighborhood of Benoxacor (cluster 80) are shown in Figure 4(b). In each graph, GenRA predictive performance, measured by AUC, number of neighbors (k) and Jaccard similarity (s), are shown as the color, x-, and y- axis, respectively. From top to bottom, each row of graphs shows the performance of GenRA for predicting brain (mgr_brain), body weight (mgr_body_weight), bones (mgr_bone), and

1 multigenerational effects on large intestines (mgr_intestine_large) and the urinary bladder
2 (mgr_urinary_bladder) across the local neighborhood of chemicals in this cluster. From left to right
3 the graphs show the performance for chemical (chm), bioactivity (bio) and hybrid (bc) descriptors.

4 The left most graph in the first row of Figure 4(b) shows the variation in GenRA performance
5 for predicting chronic brain effects using chemical structure descriptors. The extent of the local
6 neighborhood is defined by a Jaccard similarity (s) range from 0.0 to 0.6 and number of nearest
7 neighbors (k) from 2 to 12. This extent corresponds to the top left (k=2 and s=0.6) and bottom right
8 (k=12 and s=0.0) coordinates of the graph. As described earlier in methods, the lowest value of k=2
9 because GenRA predictions required at least one positive and one negative example (the minimum
10 value of s was rounded to 1 decimal place). After calculating the AUC for each value of k and s, we
11 interpolated the AUC to estimate a smoothed AUC surface for the neighborhood. The regions of k and
12 s that produced an AUC = 1 are shown in red and lower values of AUC are shown on a decreasing
13 scale (shown in the legend). For chronic brain effects using chemical descriptors the overall
14 performance of GenRA decreased as k increased or as s decreased. However, GenRA produced AUC >
15 0.8 across a nonlinear region of k and s values. Because there were regions in the neighborhood
16 (s>0.1 and k>10) for which the AUC was quite low, the overall VUS was 0.77.

17 The second graph on the first row of Figure 4(b) shows the variation in GenRA performance for
18 predicting chronic brain effects using bioactivity descriptors. The extent of the local neighborhood is
19 defined by a Jaccard similarity (s) range from 0.0 to 0.4 and number of nearest neighbors (k) from 2 to
20 12. The similarity range between chemical and bioactivity descriptors was different due to the varying
21 number of descriptors between the two spaces (among other factors). The maximum GenRA AUC was
22 0.94, which means that the bioactivity descriptors could predict chronic brain effects accurately. As

the visualization of AUC across the neighborhood shows, this only holds for a limited range of chemicals for which $s > 0.38$. This is expressed succinctly by the VUS, which had a value of just 0.5, which was 27% lower than the VUS for chemical structure descriptors.

4 DISCUSSION

The current approach to read-across is a subjective expert driven and qualitative assessment. The feasibility of deriving an algorithmic approach to read-across was investigated here using a large dataset of chemicals with available bioactivity data and *in vivo* data. We developed a novel read-across approach GenRA, which uses *in vitro* bioactivity data and chemical descriptor information to predict *in vivo* toxicity outcomes within local domains (nearest neighborhoods of chemicals). Objective measures of performance were established to evaluate the GenRA predictions – both in terms of the optimal performance for a given number of neighbors and similarity threshold and in terms of the overall local neighborhood.

Overall, the performance across all of the clusters suggests that in the absence of prior knowledge of the chemistries under study, there is merit in using bioactivity descriptor information to make read-across predictions for a number of the toxicity effects evaluated. This is evidenced in the performance of the bioactivity descriptors in Table 3. It should be noted that the ability of generating bioactivity descriptors requires experimental testing whereas chemical descriptors can be readily computed on the basis of chemical structure alone. However on a cluster by cluster basis, it is evident that read-across prediction accuracy is context dependent both on the chemistries under evaluation and the toxicity effect of interest. As shown in Table 1, for cluster 0, use of chemical descriptors outperformed bioactivity descriptors for several of the toxicity endpoints of interest. The

performances of the read-across predictions varied depending on the toxicity effect under consideration and the chemical cluster being assessed. This suggests that the selection of the starting analogues to be included in a category and the type of information used to characterize the similarity is very context dependent. This is in agreement with how analogues are currently identified and evaluated (Patlewicz et al., 2013, Wu et al., 2010; Patlewicz et al., 2015a). Typically analogue identification relies upon structural similarity, and the resulting analogues are evaluated for their physicochemical, reactivity and metabolic considerations.

It is also noteworthy to consider other explanations for the varied performance. One reason why bioactivity might be a promising predictor could be because most of the *in vitro* activity is associated with high concentration effects, which aligns with the high dose effects observed in *in vivo*. Chemical structure was found to be a good predictor of toxicity in many cases. This is not entirely surprising as structure-based descriptors were used to cluster the chemicals in the first place. Given the relatively few numbers of tested chemicals in each cluster (compared to the potential set of environmental chemicals), we cannot rule out the possibility that our results are influenced by spurious associations due to a small data set.

The selection of descriptors was uninformed by any knowledge of the endpoint or chemicals being studied, and hence contained potentially irrelevant or redundant information, which does introduce noise. This makes teasing out the features characterizing the similarity, and therefore the read-across prediction, itself a challenge. Furthermore, the dataset under study is biased to an extent since it was chosen based on the availability of bioactivity and *in vivo* study information rather than being systematically designed. However, because in this case, the classes of toxicity endpoints were so broad, it would not be logically possible to select specific classes of bioactivity. (The same can be

1 said about chemical descriptors – it is not clear that any expert could select the “right” descriptors to
2 measure similarity when looking at such a broad set of toxicity outcomes.) Any metabolic potential
3 was not taken into account, hence the neighbors selected and evaluated assumed that the structural
4 similarity component was the only rationale for forming the grouping. In addition, the HTS assays
5 underpinning the bioactivity data, for the most part, lack any significant metabolic capacity. In reality,
6 the metabolic rationale is a frequent hypothesis used to group substances together and metabolic
7 pathways are considered when evaluating the suitability and relevance of source analogues within a
8 specific read-across prediction. These shortcomings will be a focus of further study.

9 As an immediate next step, investigating whether there is any improvement in performance by
10 virtue of incorporating any knowledge about the chemical landscape in terms of driving the formation
11 of initial chemical groupings and their associated neighborhoods by exploiting existing SARs which
12 summarize knowledge for specific endpoints will be explored. Additionally, codifying some of other
13 standard expert judgement considerations in a more objective manner will be explored to inform the
14 derivation of the starting neighbors and to facilitate the assessment of read-across performance.

15 In terms of addressing the issue of uncertainty in traditional read-across as is currently
16 performed under regulations such as REACH, this approach provides a line of evidence in much the
17 same way as a QSAR or SAR prediction does. Rationalizing the biological context of the assays and
18 their relationship to the traditional study endpoints would be needed to provide the mechanistic
19 understanding required to address uncertainty in a qualitative manner in the current read-across
20 assessment process. In terms of the current practical utility of the approach so far, a baseline
21 prediction of likely hazards of concern in the absence of other data is provided which is of value in a
22 screening level hazard assessment. Of particular value are the performance metrics AUC and VUS for

evaluating confidence in predicting the toxicity of untested chemicals using GenRA. When the AUC is high and the VUS is high then one can be confident in GenRA predictions of toxicity from a broader set of neighboring analogues (defined by k and s). When AUC is high and VUS is low then the analogues must be selected more carefully over a narrower range of k and s. Finally, low values of AUC and VUS in a neighborhood indicate low confidence in any read-across predictions.

5 CONCLUSIONS

Read-across continues to generate interest as a data gap filling technique, but its acceptance for regulatory purposes remains challenged by effective means of addressing residual uncertainty and by the fact that evaluation of read-across cases is not objective (Patlewicz et al, 2014a). Identifying sources of uncertainty and practical ways of addressing them continues to be a major focus of study by many researchers (Schultz et al., 2015, Patlewicz et al., 2015a). Qualitative frameworks for characterizing uncertainties have been developed by researchers in private sector companies as well as by ECHA. Examples include those by Blackburn and Stuard, (2014) and by Patlewicz et al. (2015a). In the latter case, two practical routes for addressing uncertainties were proposed – either through using an AOP approach if the mechanism was known and an AOP that could be exploited was available, or through the use of bioactivity data as generated through programs such as the EPA ToxCast.

This study provides a first step in a tiered approach of investigating the utility of bioactivity data in concert with chemical structural information to make objective read-across predictions with quantitative measures of characterizing the uncertainties of those predictions. In this study, the selection of this type of information was not informed by any knowledge of the chemistry of the

1 substances under study or by the regulatory systemic toxicity endpoints of concern. An added
2 component was to explore the extent to which the performance of a read-across prediction could be
3 evaluated in a systematic manner. The resulting GenRA approach put forward is an adaptation of the
4 CBRA published by Low et al. (2013) which calculates a similarity weighted average of nearest
5 neighbors that have been characterized by chemical structural and bioactivity descriptors and a
6 Jaccard similarity metric. The overall read-across performance for all the neighborhoods evaluated
7 suggested that bioactivity descriptors were effective for a number of systemic toxicity effects studied.
8 However, when individual neighborhoods were probed across all the toxicity effects, the performance
9 varied. Nonetheless, the approach shows promise in moving towards codifying expert knowledge in
10 an objective quantitative framework and being able to demonstrate the performance of the read-
11 across systematically.

12 **Disclaimer:** *The views expressed in this article are those of the authors and do not necessarily reflect*
13 *the views or policies of the U.S. Environmental Protection Agency and the US Food and Drug*
14 *Administration. Mention of trade names or commercial products does not constitute endorsement or*
15 *recommendation for use.*

16

17 **Supporting Information Available:**

18 The following supporting information is provided:

19 S1 Bioactivity data for the chemicals as taken from the ToxCast.

20 S2 *in vivo* summary data as taken from ToxRefDB.

21 S3 All chemical descriptor information for chemicals.

1 S4 Information about the membership of chemicals in all chemical structure clusters and the
2 assignment of those chemicals into the respective clusters

3 S5 Predictive accuracy of chemical, biological and hybrid read-across for toxicity for all clusters
4
5
6
7

8 **Acknowledgement**

9 This work was supported in part by an appointment to the Research Participation Program at the
10 Office of Research and Development, U.S. Environmental Protection Agency, administered by the Oak
11 Ridge Institute for Science and Education through an interagency agreement between the U.S
12 Department of Energy and EPA.
13
14

References

- Aha, D., Kibler, D. 1991. Instance-based learning algorithms. *Machine Learning* 6, 37-66.
- Ankley, G. T., Bennett, R. S., Erickson, R. J., Hoff, D. J., Hornung, M. W., Johnson, R. D., Mount, D. R., Nichols, J. W., Russom, C. L., Schmieder, P. K., Serrano, J. A., Tietge, J. E., Villeneuve, D. L., 2010. Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. *Environ. Toxicol. Chem.* 29, 730-41.10.1002/etc.34
- Ball, N., Bartels, M.J., Budinsky, R.A., Klapacz, J., Hayes, S.M., Kirman, C.R, Patlewicz, G.Y., 2014. The challenge of using Read-Across within the EU REACH regulatory framework; how much uncertainty is too much? Dipropylene Glycol Methyl Ether Acetate, an exemplary case study. *Regul. Toxicol. Pharmacol.* 68, 212-221.
- Bishop, P.L., Manuppello, J.R., Willett, C.E., Sandler, J.E., 2012. Animal use and lessons learned in the U.S. High Production Volume Chemicals Challenge Program. *Environ. Health Perspect.* 120, 1631-1639.
- Blackburn, K., Stuard, S.B., 2014. A framework to facilitate consistent characterization of read across uncertainty. *Regul. Toxicol. Pharmacol.* 68, 353-362.
- Blackburn, K., Daston, G., Fisher, J., Lester, C., Naciff, J. M., Rufer E. S., Stuard, S. B., Woeller, K. 2015. A strategy for safety assessment of chemicals with data gaps for developmental and/or reproductive toxicity. *Regul Toxicol Pharmacol.* 72, 202-215. doi: 10.1016/j.yrtph.2015.04.006.
- Cronin, M.T.D, Madden, J., Enoch, S.J., Roberts, D.W. 2013. Chemical Toxicity Prediction: Category Formation and Read-across. Royal Society of Chemistry.

1 de Raat, K., 2014. Assessment of read-across: An ECHA perspective. Presented at the World Congress
2 for Animal Alternatives, 24-28th August 2014, Prague.

3 Dimitrov, S., Dimitrova, G., Pavlov, T., Dimitrova, N., Patlewicz, G., Niemala, J., Mekenyan, O. A. 2005.
4 Stepwise approach for defining the applicability domain of SAR and QSAR models. *J. Chem. Inf.*
5 *Comput. Sci.* 45, 839-849.

6 EC – European Commission. 2006. Regulation (EC) No 1907/2006 of the European Parliament and of
7 the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and
8 Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive
9 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No
10 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC,
11 93/105/EC and 2000/21/EC. *Off J Eur Union L396/1* of 30.12.2006. Commission of the European
12 Communities.

13 ECHA. 2008. Guidance on information requirements and chemical safety assessment. Chapter R.6:
14 QSARs and grouping of chemicals. Available at
15 http://echa.europa.eu/documents/10162/13632/information_requirements_r6_en.pdf

16 ECHA. 2014. The Use of Alternatives to Testing on Animals for the REACH Regulation. Second report
17 under Article 117(3) of the REACH Regulation. ECHA-14-A-07-EN. ISBN: 978-92-9244-593-5. DOI:
18 10.2823/22471

19 ECHA. 2015. Read-across Assessment Framework (RAAF). ECHA-15-R-07-EN

1 Ellison, C. M, Sherhod, R., Cronin, M. T., Enoch, S. J., Madden, J. C., Judson, P. N. 2011. Assessment of
2 methods to define the applicability domain of structural alert models. *J. Chem. Inf. Model.* 51(5), 975-
3 985. doi: 10.1021/ci1000967.

4 Ganter, B., Snyder, R. D., Halbert, D. N., Lee, M. D. 2006. Toxicogenomics in drug discovery and
5 development: mechanistic analysis of compound/class-dependent effects using the DrugMatrix
6 database. *Pharmacogenomics* 7(7), 1025-1044.

7 Gini, G., Franchi, A. M., Manganaro, A., Golbamaki, A., Benfenati, E. 2014. ToxRead: A tool to assist in
8 read across and its use to assess mutagenicity of chemicals. *SAR QSAR Env. Res.* 25, 999-1011.

9 Hubert, L., Arabie, P. 1985. Comparing Partitions. *J. Classification* 2(2-3), 193-218.

10 Hunter, J. D. 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* 9(3),
11 90-95.

12 Igarashi, Y., Nakatsu, N., Yamashita, T., Ono, A., Ohno, Y., Urushidani, T., Yamada, H. 2015. Open TG-
13 GATES: a large-scale toxicogenomics database. *Nucleic Acids Res.* 43(Database issue), D921-927.

14 Kavlock, R., Chandler, K., Houck, K., Hunter, S., Judson, R., Kleinstreuer, N., Knudsen, T., Martin, M.,
15 Padilla, S., Reif, D., Richard, A., Rotroff, D. Sipes, N., Dix, D. 2012. Update on EPA's ToxCast program:
16 providing high throughput decision support tools for chemical risk management. *Chem. Res. Toxicol.*
17 25, 1287-1302.

18 Landrum, G. 2015. RDKit. 2015, from www.rdkit.org.

19 Lock, E. F., Abdo, N., Huang, R., Xia, M., Kosyk, O., O'Shea, S. H., Zhou, Y. H., Sedykh, A., Tropsha, A.,
20 Austin, C. P., Tice, R. R., Wright, F. A., Rusyn, I. 2012. Quantitative high-throughput screening for
21 chemical toxicity in a population-based in vitro model. *Toxicol. Sci.* 126(2), 578-588.

1 Low, Y., Sedykh, A., Fourches, D., Golbraikh, A., Whelan, M., Rusyn, I., Tropsha, A. 2013. Integrative
2 chemical-biological read-across approach for chemical hazard classification. *Chem. Res. Toxicol.* **26**(8),
3 1199-1208.

4 Martin, M. T., Judson, R., Richard, A., Houck, K. A., Dix, D. J. 2007. ToxRefDB: Linking Regulatory
5 Toxicological Information on Environmental Chemicals with High-Throughput Screening and Genomic
6 Data. International Forum on Computational Toxicology.

7 Netzeva, T. I., Worth, A.P., Aldenberg, T., Benigni, R., Cronin, M. T. D, Gramatica, P., Jaworska, J. S.,
8 Kahn, S, Klopman, G., Marchant, C. A., Myatt, G., Nikolova-Jeliazkova, N., Patlewicz, G. Y., Perkins, R.,
9 Roberts, D. W., Schultz, T. W., Stanton, D. T., van de Sandt, J. J. M., Tong, W., Veith, G., Yang, C. 2005.
10 Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure-Activity
11 Relationships. The Report and Recommendations of ECVAM Workshop 52. *Altern. Lab. Anim.* 33(2),
12 155-173.

13 Nilakantan, R., Bauman, N., Dixon, J. S., Venkataraghavan, R. 1987. Topological Torsion - a New
14 Molecular Descriptor for Sar Applications - Comparison with Other Descriptors. *J. Chem. Infor.*
15 *Comput. Sci.* 27(2), 82-85.

16 Rogers, D. Hahn, M. 2010. Extended-Connectivity Fingerprints. *J. Chem. Infor. Model.* 50, 742-754.

17 OECD. 2004. OECD Principles for the Validation for Regulatory Purposes of (Quantitative) Structure-
18 Activity Relationship Models. <http://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf>

19 OECD. 2007. Guidance document on the Validation of (Quantitative) Structure-Activity Relationship
20 Models. OECD Series on Testing and Assessment No. 69. ENV/JM/MONO(2007)2

1 Patlewicz, G., Dimitrov, S., Low, L. K., Kern, P. S., Dimitrova, G. D., Comber, M. I., Aptula, A. O., Phillips,
2 R. D., Niemelä, J., Madsen, C., Wedebye, E. B., Roberts, D. W., Bailey, P. T., Mekenyan, O. G. 2007.
3 TIMES-SS - A promising tool for the assessment of skin sensitization hazard. A characterization with
4 respect to the OECD validation principles for (Q)SARs and an external evaluation for predictivity.
5 *Regul. Toxicol. Pharmacol.* 48(2), 225-239.

6 Patlewicz, G., Jeliaskova, N., Gallegos Saliner, A., Worth, A. P. 2008. Toxmatch – a new software tool to
7 aid in the development and evaluation of chemically similar groups. *SAR QSAR Environ. Res.* 19(3-4),
8 397-412.

9 Patlewicz, G., Ball, N., Booth, E. D., Hulzebos, E., Zvinavashee, E., Hennes, C. 2013. Use of Category
10 approaches, Read-across and (Q)SAR: General considerations. *Regul. Toxicol. Pharmacol.* 67, 1-12.

11 Patlewicz, G., Ball, N., Becker, R.A., Booth, E.D., Cronin, M.T.D., Kroese, D., Steup, D., Van Ravenzwaay,
12 B., Hartung, T., 2014a Food for thought..... Read-across approaches - misconceptions, promises and
13 challenges ahead. *ALTEX*, 31, 387-396.

14 Patlewicz, G., Kuseva, C., Mehmed, A., Popova, Y., Dimitrova, G., Ellis, G., Hunziker, R., Kern, P., Low,
15 L., Ringeissen, S., Roberts, D.W., Mekenyan, O., 2014b. TIMES-SS – recent refinements as a result of
16 an Industrial skin sensitisation consortium. *SAR QSAR Environ. Res.* 25, 367-391.
17 doi:10.1080/1062936X.2014.900520.

18 Patlewicz, G., Ball, N., Boogaard, P., Becker, R. A., Hubesch, B. 2015a. Building scientific confidence in
19 the development and evaluation of read-across. *Regul. Toxicol. Pharmacol.* 72, 117-133.

1 Patlewicz, G., Simon, T., Rowlands, J.C., Budinsky, R., Becker, R.A. 2015b. Proposing a scientific
2 confidence framework to support application of Adverse Outcome Pathways for regulatory purposes.
3 *Regul. Toxicol. Pharmacol.* 72, 17-25.

4 Patlewicz, G., Worth, A.P., Ball, N. 2015c. Validation of computational methods - in Validation of
5 Alternative Methods eds. Whelan, M., and Eskes, C. Springer Press UK.

6 Schultz, T. W., Amcoff, P., Berggren, E., Gautier, F., Kalric, M., Knight, D. J., Mahony, C., Schwarz, M.,
7 White, A., Cronin, M. T. D. 2015. A strategy for structuring and reporting a read-across prediction of
8 toxicity. *Regul. Toxicol. Pharmacol.* 72, 586-601.

9 Tollefsen, K.E., Scholz, S., Cronin, M.T., Edwards, S.W., de Knecht, J., Crofton, K., Garcia-Reyero, N.,
10 Hartung, T., Worth, A., Patlewicz, G., 2014. Applying Adverse Outcome Pathways (AOPs) to support
11 Integrated Approaches to Testing and Assessment (IATA). *Regul. Toxicol. Pharmacol.* 70, 629-640.

12 Wu, S., Blackburn, K., Amburgey, J., Jaworska, J., Federle, T., 2010. A framework for using structural,
13 reactivity, metabolic and physicochemical similarity to evaluate the suitability of analogs for SAR-
14 based toxicological assessments. *Regul. Toxicol. Pharmacol.* 56, 67-81.

15

16

17

1 Figures and Tables

1 GeneRA predictions of toxicity are shown above the experimental data where positive and negative
2 predictions are shown as orange and blue, respectively.

3

4 Figure 3. Relationship between performance (AUC), number of neighbors (k) and Jaccard similarity (s)
5 (shown as the color, x-, and y- axis, respectively) for multigenerational motor activity effects for
6 chemicals in cluster 0. From left to right the graphs show the AUC performance for chemical (chm),
7 bioactivity (bio) and hybrid (bc) descriptors.

8

9

10

11

1 Figure 4. Predicting toxicity endpoints using GenRA. (a) The nearest neighbors of benoxacor as a radial
2 tree organized in an anti-clockwise manner, starting with most similar chemical and ending with the
3 least similar chemical. The Jaccard similarity between chemicals is labeled on the edges. The toxicity
4 data for each chemical are summarized in terms of effects observed in guideline studies including:
5 chronic effects on brain (chr_brain); multigenerational effects on body weight (mgr_body_weight),
6 bones (mgr_bone), large intestines (mgr_intestine_large), urinary bladder (mgr_urinary_bladder); and
7 subchronic effects on the small intestine (sub_intestine_small), prostate gland (sub_prostate), spleen
8 (sub_spleen), urinalysis (sub_urinalysis), water consumption (sub_water_consumption). Positive,
9 negative and unknown toxic effects in studies are shown as red, green and grey, respectively. For
10 benoxacor the GenRA predictions of toxicity are shown above the experimental data where positive
11 and negative predictions are shown as orange and blue, respectively. (b) Relationship between
12 GenRA predictive performance as receiver operating characteristic (ROC) area under the curve (AUC),
13 number of neighbors (k) and Jaccard similarity (s), which are shown as the color, x-, and y- axis,
14 respectively, for chemicals in cluster 80. AUC performance results from 0 to 1 are represented to a
15 color scale from blue to red (see legend). The GenRA activity score thresholds that produced the best
16 balanced accuracy for each value of k and s are also shown as contours above the AUC performance
17 results (see Methods). From left to right the graphs show the AUC performance for chemical (chm),
18 bioactivity (bio) and hybrid (bc) descriptors. From top to bottom the graph show the performance
19 results for multigenerational effects on brain (mgr_brain), body weight (mgr_body_weight), bones
20 (mgr_bone), and multigenerational effects on large intestines (mgr_intestine_large) and the urinary
21 bladder (mgr_urinary_bladder).

Study	Study & Target	n_chm	chm	bio	bc
mgr	mgr_adrenal_gland	4	0.41 1.00 *(2+,2-,13)	0.10 0.50 (2+,2-,13)	0.09 0.62 *(2+,2-,13)
	mgr_clinical_signs	4	0.32 1.00 *(3+,1-,13)	0.05 0.33 (3+,1-,13)	0.05 0.33 (3+,1-,13)
	mgr_food_consumption	4	0.41 1.00 *(2+,2-,13)	0.10 0.50 (2+,2-,13)	0.09 0.62 *(2+,2-,13)
	mgr_motor_activity	4	0.74 1.00 *(1+,3-,13)	0.05 0.50 *(1+,3-,13)	0.11 1.00 *(1+,3-,13)
	mgr_offspring_survival-late	4	0.34 0.33 (3+,1-,13)	0.23 1.00 *(3+,1-,13)	0.07 0.50 *(3+,1-,13)
	mgr_pituitary_gland	4	0.32 1.00 *(1+,3-,13)	0.05 0.33 (1+,3-,13)	0.05 0.33 *(1+,3-,13)
	mgr_seminal_vesicle	4	0.34 0.33 (1+,3-,13)	0.23 1.00 *(1+,3-,13)	0.07 0.50 *(1+,3-,13)
	mgr_sperm_measure	4	0.34 0.33 *(1+,3-,13)	0.23 1.00 *(1+,3-,13)	0.07 0.50 *(1+,3-,13)
	mgr_spleen	4	0.34 0.33 (1+,3-,13)	0.23 1.00 *(1+,3-,13)	0.07 0.50 *(1+,3-,13)
	mgr_testes	4	0.32 1.00 *(1+,3-,13)	0.05 0.33 (1+,3-,13)	0.05 0.33 (1+,3-,13)

Table 1. Summary of GenRA predictive accuracy for toxicity outcomes (row) across descriptors (columns) for cluster 0. Each row shows the performance results for a study and target site. From left to right the first three columns show: study type (Study), study & target site (Study & Target), number of chemicals with *in vivo* data. the descriptor type used for GeneRA (d). The remaining columns show the GenRA predictions for chemical (chm), bioactivity (bio) and hybrid (bc) descriptors. GenRA accuracy results are reported using: the volume under the surface (VUS), the receiver operating characteristic (ROC) area under the curve (AUC), and p-value (p<0.1 is shown as *, p<0.05 as ** and p≥0.1 is left blank). The number of chemicals are given inside the parenthesis including: positive, and negative examples of the *in vivo* effect and the total number of chemicals in the cluster. For instance, row 1 summarizes the GenRA performance results for predicting multigenerational effects on the adrenal gland (mgr_adrenal_gland) for 2 positive and 2 negative perfluorinated chemicals (PFAAs) using chemical structure descriptors as follows: VUS of 0.41, best ROC AUC score 1 with p-value less than 0.1. Similarly the performance for bioactivity descriptors as follows: VUS of 0.1, best ROC AUC score 0.55 with p-value greater than 0.1.

1

CL	Study	chm	bio	bc	Only chm	Only bio	Only bc	chm> bio bc	bio>ch m bc	bc>ch m bio
0	chr	0	0	0	0	0	0	0	0	0
	dev	1	1	1	0	0	0	0	0	0
	mgr	6	4	1	5	4	0	5	4	0
	sub	15	4	1	13	3	0	14	3	0

2 Table 2. Summary of the number of studies (row) and effect sites that can be predicted with an ROC AUC >0.8. Each row shows the
3 performance results for a study type. From left to right the first two columns show the cluster (CL) and study type (Study). Columns 3 to 5
4 give the number of study and effect targets with an AUC > 0.8 for: chemical descriptors (chm), bioactivity descriptors (bio) and hybrid
5 descriptors (bc). Columns 6 to 8 show the number of study and effect targets for which only chemical, only bioactivity or only hybrid
6 descriptors produced an AUC > 0.8. Finally, columns 9 to 11 show the cases in which chemical descriptors outperformed all others
7 (chm>bio|bc), bioactivity descriptors outperform all others and hybrid descriptors outperform all other descriptors.

8

9

10

Study	Cases	chm	bio	bc	chm>bio bc	bio>chm bc	bc>chm bio	Any
chr	1115	97 / 9%	133/12%	67 / 6%	75 / 7%	107 / 10%	21 / 2%	203/18%
dev	712	63 / 9%	138/19%	50 / 7%	43 / 6%	117 / 16%	7 / 1%	167/23%
dnt	84	16 / 19%	26 / 31%	15 / 18%	5 / 6%	20 / 24%	2 / 2%	27 / 32%
mgr	668	78 /12%	126/19%	56 / 8%	46 / 7%	106 / 16%	4 / 1%	156/23%
rep	34	0 / 0%	0 / 0%	0 / 0%	0 / 0%	0 / 0%	0 / 0%	0 / 0%
sac	128	2 / 2%	7 / 5%	6 / 5%	1 / 1%	6 / 5%	2 / 2%	9 / 7%
sub	922	107/ 12%	140/15%	78 / 8%	64 / 7%	122 / 13%	21 / 2%	212/23%
All	3663	363/10%	570/16%	272 / 7%	234 / 6%	478 / 13%	57 / 2%	774/21%

1

2 Table 3. Comparison of different descriptor sets for their prediction of in vivo outcomes in different types of guideline testing studies. The
3 table shows the number of unique cluster-study-effect combinations for which the type of descriptors (columns) could predict toxicity
4 (rows) accurately (i.e. ROC AUC >0.80). The first column shows the type of toxicity study including: acute toxicity (acu), chronic toxicity
5 (chr), developmental toxicity (dev), developmental neurotoxicity (dnt), multigenerational toxicity (mgr), neurotoxicity (neu), other toxicity
6 (oth), reproductive toxicity (rep), subacute toxicity (sac) and subchronic toxicity (sub). The second column shows the number of unique
7 clusters and target sites for the effects. The remaining columns have the same meaning as those in Table 2.

Study	Effect	N	n	bc	bio	chm
chr	chr_brain	41	12	0.56 1.00 *(4+,8-,41)	0.50 0.94 *(4+,8-,41)	0.77 1.00 *(4+,8-,41)
mgr	mgr_body_weight	41	13	0.66 1.00 *(12+,1-,41)	0.69 1.00 *(12+,1-,41)	0.80 1.00 *(12+,1-,41)
	mgr_bone	41	13	0.46 1.00 *(2+,11-,41)	0.28 0.66 *(2+,11-,41)	0.78 1.00 *(2+,11-,41)
	mgr_intestine_large	41	13	0.66 1.00 *(1+,12-,41)	0.69 1.00 *(1+,12-,41)	0.80 1.00 *(1+,12-,41)
	mgr_urinary_bladder	41	13	0.65 1.00 *(1+,12-,41)	0.48 1.00 *(1+,12-,41)	0.79 1.00 *(1+,12-,41)
sub	sub_intestine_small	41	13	0.46 1.00 *(1+,12-,41)	0.18 0.50 *(1+,12-,41)	0.77 1.00 *(1+,12-,41)
	sub_prostate	41	13	0.67 1.00 *(1+,12-,41)	0.54 1.00 *(1+,12-,41)	0.76 1.00 *(1+,12-,41)
	sub_spleen	41	13	0.67 1.00 *(1+,12-,41)	0.54 1.00 *(1+,12-,41)	0.76 1.00 *(1+,12-,41)
	sub_urinalysis	41	13	0.72 1.00 *(4+,9-,41)	0.41 1.00 *(4+,9-,41)	0.84 1.00 *(4+,9-,41)
	sub_water_consumption	41	13	0.67 1.00 *(1+,12-,41)	0.54 1.00 *(1+,12-,41)	0.76 1.00 *(1+,12-,41)

Table 4. Summary of GenRA predictive accuracy for toxicity outcomes (row) across descriptors (columns) for cluster 80. Each row shows the performance results for a study and target site. From left to right the first three columns show: study type (Study), study & target site (Study & Target), number of chemicals with *in vivo* data. the descriptor type used for GenRA (d). The remaining columns show the GenRA predictions for chemical (chm), bioactivity (bio) and hybrid (bc) descriptors. GenRA accuracy results are reported using: the volume under the surface (VUS), the receiver operating characteristic (ROC) area under the curve (AUC), and p-value (p<0.1 is shown as *, p<0.05 as ** and p≥0.1 is left blank). The number of chemicals are given inside the parenthesis including: positive, and negative examples of the *in vivo* effect and the total number of chemicals in the cluster.