

A systematic evaluation of analogs and automated read-across prediction of estrogenicity: A case study using hindered phenols[☆]

Prachi Pradeep^{a,b,*}, Kamel Mansouri^{a,b}, Grace Patlewicz^b, Richard Judson^b

^aOak Ridge Institute for Science and Education, Oak Ridge, Tennessee

^bNational Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina

[☆]Short title: "A systematic evaluation of analogs and automated read-across prediction of estrogenicity"

*Corresponding author: Prachi Pradeep, Email address: pradeep.prachi@epa.gov

Abstract

Read-across is an important data gap filling technique used within category and analog approaches for regulatory hazard identification and risk assessment. Although much technical guidance is available that describes how to develop category/analog approaches, practical principles to evaluate and substantiate analog validity (suitability) are still lacking. This case study uses hindered phenols as an example chemical class to determine: (1) the capability of three structure fingerprint/descriptor methods (PubChem, ToxPrints and MoSS MCSS) to identify analogs for read-across to predict Estrogen Receptor (ER) binding activity and, (2) the utility of data confidence measures, physicochemical properties, and chemical R-group properties as filters to improve ER binding predictions. The training dataset comprised 462 hindered phenols and 257 non- hindered phenols. For each chemical

of interest (target), source analogs were identified from two datasets (hindered and non-hindered phenols) that had been characterized by a fingerprint/descriptor method and by two cut-offs: (1) minimum similarity distance (range: 0.1 - 0.9) and, (2) N closest analogs (range: 1 - 10). Analogs were then filtered using: (1) physicochemical properties of the phenol (termed global filtering) and, (2) physicochemical properties of the R-groups neighboring the active hydroxyl group (termed local filtering). A read-across prediction was made for each target chemical on the basis of a majority vote of the N closest analogs. The results demonstrate that: (1) concordance in ER activity increases with structural similarity, regardless of the structure fingerprint/descriptor method, (2) increased data confidence significantly improves read-across predictions, and (3) filtering analogs using global and local properties can help identify more suitable analogs. This case study illustrates that the quality of the underlying experimental data and use of endpoint relevant chemical descriptors to evaluate source analogs are critical to achieving robust read-across predictions.

Keywords: Read-across, analog identification, analog evaluation, quantitative uncertainty analysis, estrogen receptor (ER) binding

1 Introduction

Read-across is a well-established data-gap filling technique used within category and analog approaches for regulatory hazard identification and risk assessment [1]. In the read-across approach, endpoint information for one or more chemicals (source analogs) are used to predict the same endpoint for another chemical (target), which is considered “similar” (usually on the basis of structural similarity)[1-3]. There are a number of steps in the development of a category or analog approach. Slight variations of the exact number and name of these steps depends on the technical guidance and publication used [1, 4-6]. However, the two critical steps in the process are analog identification and analog evaluation [7, 8]. Analog identification is the process of searching for source analogs similar to the target chemical. Source analogs are usually identified based on structural similarity, using fingerprints that encode chemical information based on the presence or absence of certain structural features [5, 9]. A similarity index such as the Jaccard distance (Tanimoto index) [10] is then used as a threshold to limit the number of source analogs retrieved. Many web-based tools that permit structure searching include an algorithm to search for structurally similar chemicals in this manner. Common web based tools include ChemID plus [11], and ChemSpider [12]. This type of structural search tends to be general in scope, in the sense that no assumptions are made to limit the analog search on the basis of properties or parameters that might be pertinent to a specific endpoint. On the other hand, a search for source analogs informed by parameters relevant to the endpoint of interest would rely on descriptors which could affect chemical bioavailability and reactivity, such as physicochemical properties (*e.g.*, LogP, molecular volume), electronic properties (*e.g.*, energy of the lowest unoccupied orbital (eLUMO), energy of the highest occupied orbital (eHOMO) [5]. The next step, analog evaluation, gathers associated property and effect information for the source analogs to determine their relevance and suitability for the endpoint of interest. Many structure fingerprint and descriptor methods are available (free or commercial), each of which capture different aspects of chemical structure that are potentially relevant to different endpoints.

Despite available guidance [3, 4, 7] for analog/category approaches to read-across, guiding principles to evaluate analog validity for specific endpoints remains lacking [4, 8]. The similarity rationale underpinning source analog selection, as well as the quantity and quality of experimental data associated with the selected analogs, are important sources of uncertainty in read-across [4, 13]. Consequently, even though read-across is conceptually accepted by both regulatory agencies and industry, difficulties remain in the consistent application of read-across approaches in practice, which in turn has limited their acceptance for regulatory decisions [3, 14]. Efforts have been made to standardize and characterize a framework for documenting read-across justifications to help increase consistency and promote regulatory acceptance of read-across predictions [4, 13, 15]. Although several read-across studies have been published recently [16-18], successful examples are still lacking [14].

To establish improved and reproducible read-across predictions, this case study undertook a systematic analysis of analogs for read across predictions of *in vitro* ER binding. Hindered phenols were selected as an example chemical class. Hindered phenols are defined as phenols that contain one or more bulky functional groups ortho to the phenolic hydroxyl group, e.g., 3-chloro-4-hydroxybenzoic acid. Phenols, in general, are known to mimic the activity of estrogen and possess estrogenic activity resulting in the possibility of endocrine disruption [19, 20]. Endocrine disruption can lead to a wide range of health disorders in humans, including reproductive and developmental toxicity [21, 22]. Phenols can interact with the estrogen receptor (ER) due to the presence of the phenol hydroxyl group, which aids in binding with ER. Hindered phenols are expected to be less potent ER binders than non-hindered phenols because their bulky functional groups block the hydroxyl group-protein interaction [23].

An automated approach was developed for this case study to: (1) identify and evaluate the suitability (validity) of source analogs; and (2) evaluate and assess uncertainty due to confidence in data and analog suitability in read-across predictions. Specifically, the case study presents an analysis of the ability of three structure fingerprint/descriptor methods to identify source analogs to read-across ER

binding, and the use of data confidence measures, physicochemical properties, and chemical substituent functional (R) group physicochemical properties to evaluate the validity of the source analogs identified.

2 Methods

2.1 Dataset

The dataset of phenols used in this study was extracted from the prediction dataset constructed as part of CERAPP, the Collaborative Estrogen Receptor Activity Prediction Project [24]. This CERAPP prediction dataset (herein referred to as the source dataset) contained literature-derived curated data from a number of overlapping sources including Tox21 [25-29], U.S. FDA Estrogenic Activity Database (EADB) [30], METI (Ministry of Economy, Trade and Industry, Japan) database [31], and ChEMBL [32] for over 32,000 chemical structures. Each chemical in the CERAPP source dataset had been assigned a literature data source count such that there was <20% disagreement among different sources. For instance, if there were 4 independent publications of ER activity for a chemical, all 4 sources had to agree (i.e., ER binder or non-binder) for the chemical to be included in the source dataset. On the other hand, if there were 5 published reports for a particular chemical, the chemical would still be included in the source dataset if one reference disagreed on ER binding activity with the majority consensus. The majority ER activity outcome from all the sources was taken as the final outcome, 1 or 0 representing ER binder and non-binder, respectively. The literature data source count from CERAPP was used as a surrogate for data confidence in this study. The expectation was that the more consistent the literature reports were of ER activity (binder or non-binder), the more likely the activity could be relied upon to be reproduced in a subsequent experiment. A custom KNIME workflow (version 2.11.3) [33] was developed to extract phenols from the larger CERAPP chemical library and to categorize them as being hindered or not, based on the presence or absence of bulky groups at the ortho position. The final dataset used in this study

comprised 719 phenols with 462 hindered phenols (207 ER binders) and 257 non-hindered phenols (155 ER binders).

2.2 Chemical descriptors

One study aim was to analyze the ability of different structure descriptor approaches to identify source analogs for hindered phenols, and evaluate the adequacy of the analogs for ER read-across predictions. This enabled a baseline performance assessment to be made for the preliminary analogs identified. While a myriad of fingerprints/descriptors can be computed for chemical properties (structure, physicochemical, electronic), there are no published or systematic guidelines for evaluating the suitability of one descriptor type versus another for a specific endpoint.

Three common structure-based fingerprint/descriptors sets (PubChem [34], ToxPrints [35], and MoSS MCSS [36]) were used. PubChem fingerprints are 881 bits long where each bit represents the presence or absence of a specific substructure. The substructure categories spanned by a PubChem fingerprint include hierarchical element counts, rings in a canonical extended smallest set of smallest rings, ring set, simple atom pairs, simple atom nearest neighbors, detailed atom neighborhoods, simple SMARTS patterns, and complex SMARTS patterns. The PubChem fingerprints were generated in KNIME analytics platform (version 2.11.3) [33]. ToxPrint chemotypes (or ToxPrints) comprise 729 uniquely defined chemical features (<https://toxprint.org>) coded in XML-based Chemical Subgraphs and Reactions Markup Language (CSRML) [35]. The ToxPrints features were specifically designed to provide a broad coverage of inventories consisting of environmental and industrial chemicals including pesticides, cosmetics ingredients, food additives and drugs. The fingerprints represent a wide range of substructures comprising atoms, bonds, chains, groups and ring elements. The ToxPrints were generated within the publically available Chemotyper application (version 1.0.r12976, <https://chemotyper.org>). MoSS is a substructure miner algorithm implemented in KNIME analytics platform (version 2.11.3) [33] that calculates the size of the maximum common substructure (MCSS) between two chemicals. The MCSS of

two chemicals is the largest possible substructure that is present in both chemical structures; more similar the chemicals have larger MCSS sizes. The Jaccard distance (Tanimoto index) was used to calculate pairwise similarity indices for the all phenols in the datasets as characterized by the PubChem and ToxPrints descriptor sets. The similarity index ranges from 0-1, where 0 indicates least similar (dissimilar) and 1 indicates most similar (mostly chemical similarity by itself). These pairwise similarities were summarized in a similarity matrix. The similarity matrix is calculated using a component called a distance matrix in KNIME. For the third descriptor set, the MCSS itself was taken as the similarity index.

2.3 Read-across analysis workflow

Figure 1 summarizes the four steps of the read-across analysis workflow that was followed in this study. First, a set of structurally related analogs were identified using each of the three descriptor sets to determine the baseline performance of ER read-across predictions for the set of hindered phenols. Second, the target-analog pairs from each descriptor method were then used to evaluate two sources of uncertainty in read-across: (1) data confidence (as quantified by the number of consistent literature sources), and (2) analog validity. Third, the analogs were filtered on the basis of physicochemical properties expected to be relevant for ER binding to assess the improvement (if any) in performance of the read-across predictions compared with the baseline performance. Finally, a majority-vote based read-across prediction was made for each target based on the experimental binding data of the identified analogs were used. Each step is described in detail in the following subsections. The software code for the read-across analysis was developed in Python 2.7 [37] and is available as part of the supplementary information.

2.3.1 Selection of analogs

As described earlier, bulky groups ortho to the hydroxyl group hinder the ability of a phenol to bind to the ER [21]. Accordingly, the phenols in this dataset had been categorized into two classes: hindered and non-

hindered phenols. Each target hindered phenol had two datasets from which to identify source analogs: (1) hindered phenols, and (2) non-hindered phenols. Source analogs from each of those sets were identified using the Jaccard distance [37] for PubChem and ToxPrints fingerprints, whereas MCSS size was used as a similarity metric for the MoSS descriptor set. To ensure a pragmatic and practical number of source analogs were retrieved in each case, the number of analogs from each descriptor set was limited in 2 different ways by: (1) the Jaccard distance/MCSS size which was varied from 0.1 - 0.9 (distance ranges over the interval from 0 to 1), and (2) the number of closest analogs (from 1 - 10) and a similarity index ≥ 0.7 . This resulted in four sets of source analogs that were then assessed for their ability to predict ER binding for each target hindered phenol. Performance was characterized by accuracy (fraction of the total set of targets predicted correctly) and balanced accuracy (average of sensitivity and specificity). Balanced accuracy is useful in situations such as this where there are significantly unequal numbers of chemicals in the positive and negative classes. This enabled the determination of a baseline performance, which characterizes the performance of the read-across model when using a high quality dataset and selecting analogs using standard descriptor methods.

2.3.2 Evaluation of sources of uncertainty

Two sources of uncertainty were investigated in this study:

1. Data confidence: The impact of experimental variability was evaluated by systematically filtering the dataset such that the overall consensus ER binding outcome for each chemical was derived from k number of literature sources, where k ranged from 1 to 10. For each value of k , read-across predictions were made for each hindered phenol using N (1 - 10) analogs. Since the underlying data confidence is well known to impact any read-across prediction [4, 15], it was hypothesized that structural similarity based read-across predictions should improve as the confidence in the experimental data improved. The results of this analysis (discussed in Section 3.3) were used to

select an optimum value of k to serve as a surrogate measure of confidence in the experimental data. It should be noted that this optimum value of k was used to filter the phenol dataset into a “high quality” set for the subsequent steps of the read-across analysis workflow.

2. Analog validity: Analog validity was initially evaluated in terms of the concordance in experimental ER binding between each target-analog pair, where concordance is the ratio of number of analogs with the same experimental outcome as the target and the total number of target-analog pairs within each similarity threshold. Concordance analysis was performed for each of the four target-analog sets using: (1) each descriptor approach, and (2) analogs resulting from combinations of descriptor approaches *i.e.*, either binary combinations or all three descriptor approaches. The results of concordance analysis for analog validity (discussed in Section 3.1), were used to assess whether source analogs needed to be categorized separately as hindered or non-hindered phenols, as well as the ability of the 3 different structure descriptor methods to identify potential analogs for ER binding.

2.3.3 Filtering analogs to improve validity

Source analogs were also evaluated using a selection of physicochemical properties suggested by previous QSAR studies to be relevant for ER binding [38]. This analysis was performed using two separate approaches: (1) globally, using several whole molecule properties (*e.g.*, molecular volume, eHOMO, eLUMO), and (2) locally, using the neighborhood of the phenol hydroxyl group defined by neighboring R-groups. Size and chemical nature (*e.g.*, charge) of different R-groups were hypothesized to affect the ability of the active hydroxyl group to interact with the binding pocket due to steric hindrance or changes in the polarity of the phenol hydroxyl group. Several 2D and 3D properties (see below) were evaluated globally and locally. Properties able to discriminate ER binders from non-binders were used for filtering. Properties for both global and local filtering were computed using MOE software (version 2015.10) [39]. The final filtering criteria are described in detail below:

1. Global filtering: Source analogs were screened based on physicochemical property (LogP), a steric property (molecular volume) and electronic properties (hydrogen bond donors and acceptors). A Student's t-test was performed to test if these properties were significantly different for the phenols that bind to the ER compared to the non-binders. The results of the t-test are discussed in Section 3.3. Each of these properties for the selected source analog was compared to the original target chemical, and if it was not within an acceptable threshold, the analog was discarded until *N* closest analogs were selected for each target chemical within the 0.7 similarity threshold. The threshold derivation for each of the properties are discussed in Section 3.3. If none of the analogs satisfied the acceptance thresholds, then no read-across prediction could be made and the target was considered out of domain for read-across predictions.
2. Local filtering: Each phenol was decomposed into a basic phenol scaffold and substitutions (R-groups) at different positions relative to the hydroxyl group using a KNIME workflow. This decomposition led to R-groups obtained from 12 distinct substituent positions (auto-numbered by the algorithm), which included conjoined and bridging rings. The most commonly occupied substituent positions were R2, R3, R4 and R6 (Figure 2). A number of steric, physicochemical and electronic properties of the R-groups were systematically evaluated to identify those properties that potentially impacted ER binding at the phenolic hydroxyl group. The properties examined were LogP, acidity (pH = 7), basicity (pH = 7), molecular flexibility, topological polar surface area (TPSA), Lipinski donor count, Lipinski acceptor count, number of H-bond donor and acceptor atoms, and Van der Waals volume. A Student's t-test was performed to select those properties that were significantly different for the R-groups in ER binding phenols compared with non-binders. The selected properties of the same positioned R-groups, if any, between a target and an analog were compared, and if it was not within an acceptable threshold (as discussed in Section 3.3), the analog was discarded until *N* closest analogs were selected for each target chemical

within a similarity index of 0.7. If none of the analogs satisfied the acceptance thresholds, then the target was considered out of domain for read-across predictions. The results of the t-test, selected R-groups, properties, and selection of the acceptable range is discussed in Section 3.3.

Read-across ER binding predictions were derived for each target hindered phenol (after global and local analog filtering) to assess whether there was an improvement in the performance of the predictions relative to the baseline performance assessment in step 1 of the read-across analysis workflow. The improvement, if any, in performance was statistically validated by using the bootstrapping technique, which is a technique used to derive confidence intervals for any quantifiable quantity (accuracy and balanced accuracy in this case) [40]. In bootstrapping n random samples of data are generated by sampling with replacement from the original dataset. The model results from the various samples are then aggregated to derive estimates of errors and confidence intervals for the performance measures of a model.

3 Results and Discussion

3.1 Uncertainty analysis

1. Data confidence analysis: Structural similarity based read-across predictions improved with the confidence in experimental data, but there was a threshold for optimal performance. Figure 3 shows the trend of maximum read-across prediction accuracy and associated balanced accuracy using the PubChem fingerprints as a function of the number of literature sources k (1-10). The maximum accuracy was also dependent on the number of source analogs and the number of target hindered phenols for which predictions could be made. The same trend was observed using the other structure descriptor methods sets (results not shown), but PubChem yielded the best read-across prediction performance. As evidenced in Figure 3, the number of data sources did

improve prediction accuracy up to a threshold of 4 literature sources, but after that the performance rate increase drops off, as do the number of source analogs drops significantly and the number of phenols in the dataset that can be predicted declines drastically. For example, balanced accuracy is 69.6% when $k \geq 1$ and 8 source analogs are used. In contrast, using ≥ 4 literature sources and just 1 source analog from PubChem increases balanced accuracy to 85.3%. These results demonstrate the importance of robust experimental data underpinning the source analogs. Based on these findings, the remainder of the read-across analysis was performed using a reduced dataset of phenols that had ≥ 4 literature data sources. This reduced the original dataset of 719 phenols (of which 462 were hindered phenols) to 481 phenols of which 296 were hindered phenols (61.5% of the original set).

2. Analog validity analysis: Analog validity was evaluated in terms of the concordance in experimental ER binding between each target-analog pair. Concordance is the fraction of analogs that had the same call (binder vs. non-binder) as the target chemical. With a few exceptions, concordance in ER binding rises with increasing similarity (similarity cut-off range: 0.1 - 0.9) for each descriptor method as shown in Figure 4. There is a marked increase in concordance ($> 80\%$) at cut-offs above 0.7 (plots (a)- (c)) along with a decrease in coverage (number of chemicals for which an analog could be selected at the given similarity threshold) as shown in Table 1. The structure based fingerprints/descriptor methods PubChem and ToxPrints (or a combination of the two) performed better than the MoSS MCSS in the identification of analogs for ER binding prediction based on a balance between performance and coverage. Generally, combinations of descriptor methods did not result in significantly improved performance relative to using individual descriptor sets.

There was not a significant difference in read-across predictivity between the use of hindered phenols, non-hindered phenols or both hindered and non-hindered phenols as analogs.

The concordance as a function of similarity index plot for the both the set of phenols (plot (c)) was very similar to the other plots for the separate sets of phenols at a similarity threshold of 0.7. Figure 4(plot (d)) shows a histogram of concordance for each of the three descriptor methods using either hindered analogs, non-hindered analogs or both at a similarity threshold of 0.9. The number on top of each bar indicates the coverage as listed in the tables in Table 1. As shown, using both hindered and non-hindered analogs as analogs resulted in significantly greater coverage as compared to using either one of them without a big loss in performance. Given the lack of difference in the concordance across the three sets and the effect on coverage, the source analogs for the remainder of the analysis were not evaluated separately as hindered or non-hindered phenols, but solely on the basis of the similarity index and number of analogs.

3.2 Filtering analogs to improve validity

1. Global filtering: The distribution of chemical properties discussed in Section 2.3.3 (and reflected in Figure 1) were compared between ER binders and non-binders. The legend in each boxplot of Figure 5 lists the p-values from the Student's t-test to determine if the filtering properties were significantly different between ER binders and non-binders. The property space coverage for each of the selected properties (molecular volume, LogP and number of H bond donors and acceptors) across ER binders and non-binders are shown in Figure 5 as box plots (a)-(c). A subjective decision was made for each property value based on its range to determine a threshold (listed in Figure 1) for filtering source analogs. Source analogs were filtered out if their molecular volume was beyond $\pm 100\%$ of the molecular volume of the target chemical, if LogP was beyond ± 1 unit, and if the total number of hydrogen bond donors and acceptors were beyond ± 6 units. Source analogs that met these thresholds and had a similarity index ≥ 0.7 were used to make read-across predictions.

2. Local filtering: As discussed in Section 2.3.3, source analogs were also filtered locally to determine any improvement in prediction accuracy. Table 2 summarizes the t-test performed to determine which properties drove ER binding at the phenol hydroxyl group. LogP, total number of hydrogen bond donor and acceptors, and basicity were found to be most significant of the properties considered. A subjective decision was made in order to set the threshold for each property. A source analog was filtered out if the number of donor-acceptor atoms of the R2, R3 and R4 groups were beyond ± 2 units, if the LogP of the R2 and R3 groups was beyond ± 3 units, if the LogP of R6 group was beyond ± 2 units, and if the basicity of R2 and R3 groups were beyond ± 2 units as compared to the respective R-groups of the target. Source analogs that met these thresholds and had a similarity index ≥ 0.7 were used to make read-across predictions.

3.3 Read-across performance

The read-across analysis workflow in Figure 1 outlined three steps undertaken to systematically evaluate the utility of structure-based fingerprints/descriptors to identify source analogs for a set of hindered phenols and make read-across predictions of ER binding. Further it investigated the impact of data confidence measures, and global and local filtering (on the basis of physicochemical properties relevant to ER binding) on read-across predictions. Using PubChem fingerprints to characterize source analogs resulted in better read-across performance to predict ER binding for hindered phenols as compared to ToxPrints or MoSS MCSS fingerprints. Using the number of literature sources as a surrogate measure to account for data confidence lead to a significant improvement in read-across performance. Filtering source analogs on the basis of properties relevant to ER binding globally and locally also improved read-across performance, though to a much lesser extent. Figure 6 summarizes the performance (accuracy and balanced accuracy) of read-across as a function of the number of source analogs used, using PubChem fingerprints. The performance without accounting for data confidence and analog filtering, annotated in the plots as “no filtering”, is shown as blue bars. The best performance is found when only 1 source analog

was used for each target for read-across prediction (accuracy = 70% and balanced accuracy = 69%). The performance after accounting for data confidence considerations, annotated in the plots as “data confidence filtering” and reflected by green bars, results in a marked improvement in performance. The accuracy using 1 source analog for each target increases to 90%. We term this baseline performance of the read-across method, because we have largely filtered out the effect of data confidence. The results demonstrate that using only 1 (closest) analog with good quality data, performs as well as any other approach ((Q)SARs), which provides support for using the standard “analog” approach in read-across. We next examine whether analog filtering (global or local) yields performance improvements over baseline. The read-across accuracy increases to 93% using global filtering and to 91% using local filtering, when using 1 source analog for each target.

Next, the confidence intervals for accuracy and balanced accuracy were evaluated using the bootstrapping method described in Section 2.3 for all the three levels of data filtering: (1) data confidence, (2) global filtering, and (3) local filtering of analogs. For each level, the dataset was sampled with replacement to generate $n = 1000$ datasets of phenols. For each bootstrap sample, the analog selection and read-across analysis, as described in the methods, was repeated and the prediction metrics were calculated. Figure 7 shows the confidence intervals corresponding to each level of filtering. The non-overlapping confidence intervals demonstrate that the filtering methods did result in an improvement in the performance of the read-across predictions relative to the baseline performance. Overall, the results demonstrate that increased data confidence and filtering analogs using global and local properties can significantly improve read-across predictions.

4 Conclusions

Read-across is a valuable data gap filling technique for chemicals that lack empirical data. However, implementation of read-across methods is challenging because of difficulties in the selection and

evaluation of source analogs, and issues with the quality of data of the selected analogs. Read-across is not a “one size fits all” approach and, as found in this case study, using structural fingerprints alone does not guarantee valid analogs and good read-across performance. Addressing the key sources of uncertainty in read-across such as the quality of underlying experimental data and characterization of endpoint relevant properties led to improved read-across predictions.

This case study presents a read-across analysis workflow to systematically assess the baseline performance of reading across ER binding for a set of hindered phenols, and evaluating the impact of data confidence and endpoint relevant chemical descriptors on read-across performance. To summarize, the results of this case study illustrate that: (1) for each fingerprint/descriptor method, the concordance in biological activity increases with increasing structural similarity, (2) substructure-based fingerprint methods (PubChem and ToxPrints) tend to perform better than maximum common substructure based method (MoSS) for selection of analogs for ER binding prediction when a balance between performance and coverage is considered, (3) data validated from multiple sources increases the accuracy of read-across predictions, and (4) filtering analogs based on properties relevant to ER binding slightly improves the validity of analogs and subsequently prediction accuracy.

Acknowledgment: This work was supported in part by an appointment to the ORISE participant research program supported by an interagency agreement between the US EPA and DOE. The authors are thankful to Tara Barton-Maclaren, Matthew Gagne, and Nicholas Trefliak from Health Canada for the discussions and their suggestions in the development of this work.

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.

Supplemental files:

1. Input data files
2. Code files

5 References

1. OECD 2014. *Guidance on grouping of chemicals*. OECD Series on Testing and Assessment No. 194. Organisation for Economic Co-operation and Development, Paris, France. 2016.
2. Enoch, S.J., *Chemical Category Formation and Read-Across for the Prediction of Toxicity*, in *Recent advances in QSAR studies: Methods and applications*. 2010, Springer Netherlands. p. 209-219.
3. Patlewicz, G., et al., *Use of category approaches, read-across and (Q)SAR: General considerations*. Regulatory Toxicology and Pharmacology, 2013. **67**(1): p. 1-12.
4. Patlewicz, G., et al., *Building scientific confidence in the development and evaluation of read-across*. Regulatory Toxicology and Pharmacology, 2015. **72**(1): p. 117-133.
5. ECETOC, 2012. *Technical Report TR 116: category approaches, read-across, (Q)SAR*.
6. ECHA, 2008. *Guidance on information requirements and chemical safety assessment. Chapter R.6: QSARs and grouping of chemicals*.
7. Wu, S., et al., *A framework for using structural, reactivity, metabolic and physicochemical similarity to evaluate the suitability of analogs for SAR-based toxicological assessments*. Regulatory Toxicology and Pharmacology, 2010. **56**(1): p. 67-81.
8. Patlewicz, G., *Read-across approaches - misconceptions, promises and challenges ahead*. ALTEX, 2014. **31**(4): p. 387-396.
9. Jaworska, J. and N. Nikolova-Jeliazkova, *How can structural similarity analysis help in category formation? SAR and QSAR in Environmental Research*, 2007. **18**(3-4): p. 195-207.
10. Bajusz, D., A. Rácz, and K. Héberger, *Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?* Journal of Cheminformatics, 2015. **7**(1).
11. ChemID Plus. <https://chem.nlm.nih.gov/chemidplus>.

12. ChemSpider. <http://www.chemspider.com/>.
13. Schultz, T.W., et al., *A strategy for structuring and reporting a read-across prediction of toxicity*. Regulatory Toxicology and Pharmacology, 2015. **72**(3): p. 586-601.
14. Ball, N., *Toward Good Read-Across Practice (GRAP) guidance*. ALTEX, 2016.
15. Blackburn, K. and S.B. Stuard, *A framework to facilitate consistent characterization of read across uncertainty*. Regulatory Toxicology and Pharmacology, 2014. **68**(3): p. 353-362.
16. Ball, N., et al., *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*. Regulatory Toxicology and Pharmacology, 2014. **68**(2): p. 212-221.
17. Schultz, T.W., et al., *Read-across of 90-day rat oral repeated-dose toxicity: A case study for selected 2-alkyl-1-alkanols*. Computational Toxicology, 2017.
18. Schultz, T.W., et al., *Read-across of 90-day rat oral repeated-dose toxicity: A case study for selected n -alkanols*. Computational Toxicology, 2017.
19. Hu, J.-Y. and T. Aizawa, *Quantitative structure-activity relationships for estrogen receptor binding affinity of phenolic chemicals*. Water Research, 2003. **37**(6): p. 1213-1222.
20. Routledge, E.J. and J.P. Sumpter, *Structural Features of Alkylphenolic Chemicals Associated with Estrogenic Activity*. Journal of Biological Chemistry, 1997. **272**(6): p. 3280-3288.
21. Giulivo, M., et al., *Human exposure to endocrine disrupting compounds: Their role in reproductive systems, metabolic syndrome and breast cancer. A review*. Environmental Research, 2016. **151**: p. 251-264.
22. Safe, S.H., *Endocrine disruptors and human health--is there a problem? An update*. Environ Health Perspect, 2000. **108**(6): p. 487-93.
23. Miller, D., et al., *Estrogenic Activity of Phenolic Additives Determined by an In Vitro Yeast Bioassay*. Environmental Health Perspectives, 2001. **109**(2): p. 133.

24. Mansouri, K., et al., *CERAPP: Collaborative Estrogen Receptor Activity Prediction Project*. Environmental Health Perspectives, 2016. **124**(7).
25. Attene-Ramos, M.S., et al., *The Tox21 robotic platform for the assessment of environmental chemicals – from vision to reality*. Drug Discovery Today, 2013. **18**(15-16): p. 716-723.
26. Collins, F.S., G.M. Gray, and J.R. Bucher, *TOXICOLOGY: Transforming Environmental Health Protection*. Science, 2008. **319**(5865): p. 906-907.
27. Huang, R., et al., *Profiling of the Tox21 10K compound library for agonists and antagonists of the estrogen receptor alpha signaling pathway*. Scientific Reports, 2014. **4**.
28. Shukla, S.J., et al., *The future of toxicity testing: a focus on in vitro methods using a quantitative high-throughput screening platform*. Drug Discovery Today, 2010. **15**(23-24): p. 997-1007.
29. Tice, R.R., et al., *Improving the Human Hazard Characterization of Chemicals: A Tox21 Update*. Environmental Health Perspectives, 2013. **121**(7): p. 756-765.
30. Shen, J., et al., *EADB: An Estrogenic Activity Database for Assessing Potential Endocrine Activity*. Toxicological Sciences, 2013. **135**(2): p. 277-291.
31. *METI (Ministry of Economy Trade and Industry, Japan). 2002. Current Status of Testing Methods Development for Endocrine Disrupters. 6th Meeting of the Task Force on Endocrine Disrupters Testing and Assessment (EDTA). 24–2 June 2002. Tokyo, Japan.*
32. Gaulton, A., et al., *ChEMBL: a large-scale bioactivity database for drug discovery*. Nucleic Acids Research, 2012. **40**(D1): p. D1100-D1107.
33. Berthold, M.R., et al., *KNIME - the Konstanz information miner: version 2.0 and beyond*. ACM SIGKDD Explorations Newsletter, 2009. **11**(1): p. 26.
34. PubChem. <https://pubchem.ncbi.nlm.nih.gov/help.html>.
35. Yang, C., et al., *New Publicly Available Chemical Query Language, CSRML, To Support Chemotype Representations for Application to Data Mining and Modeling*. Vol. 55. 2015. 510--528.

36. Borgelt, C., T. Meinl, and M. Berthold. *MoSS: a program for molecular substructure mining*. 2005. ACM Press.
37. Python Software Foundation. *Python Language Reference, version 2.7*. Available at <http://www.python.org>.
38. *Review of QSAR Models and Software Tools for predicting Developmental and Reproductive Toxicity*. JRC Scientific and Technical Reviews.
39. *Molecular Operating Environment (MOE), 2013.08*; Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2016. 2016.
40. Low, L.Y., *The Jackknife, the Bootstrap and Other Resampling Plans* - Efron, B. Journal of the American Statistical Association, 1983. **78**(384): p. 987-987.

6 Figures

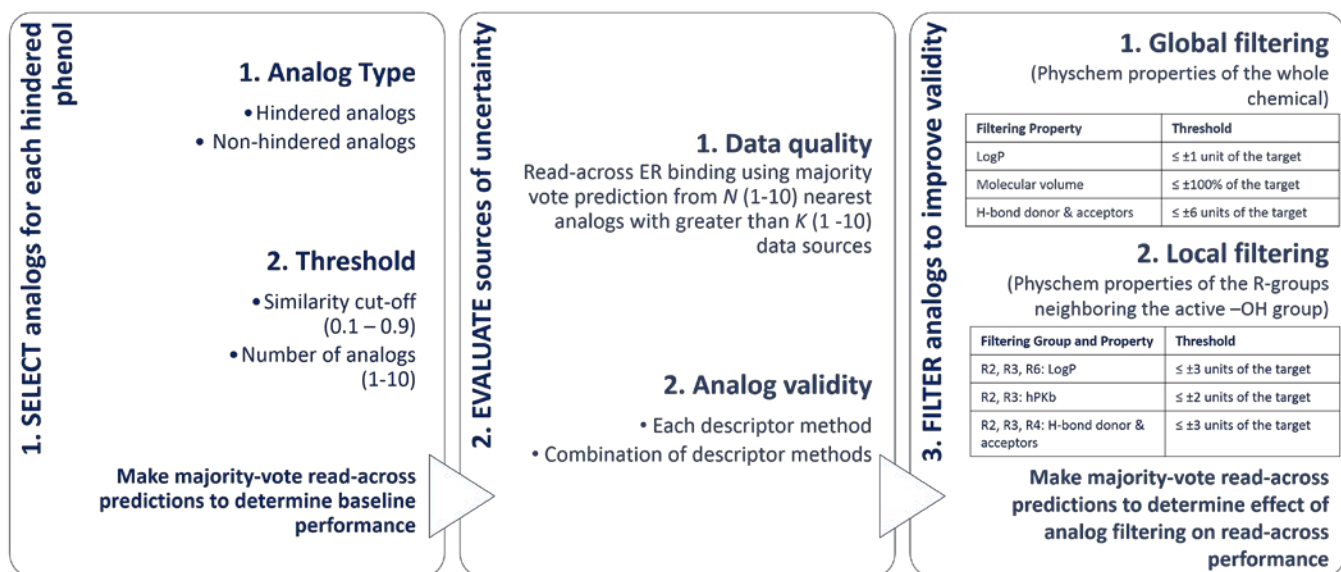


Figure 1: Overall read-across analysis workflow

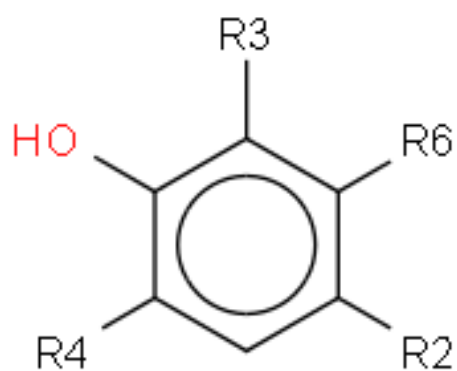


Figure 2: Each phenol was decomposed to extract the basic phenol scaffold to identify different substitution positions and R-groups using a KNIME workflow. The decomposition resulted in 12 distinct substituent positions, which included conjoined and bridging rings. R2, R3, R4, and R6 were the most commonly occupied substituent positions.

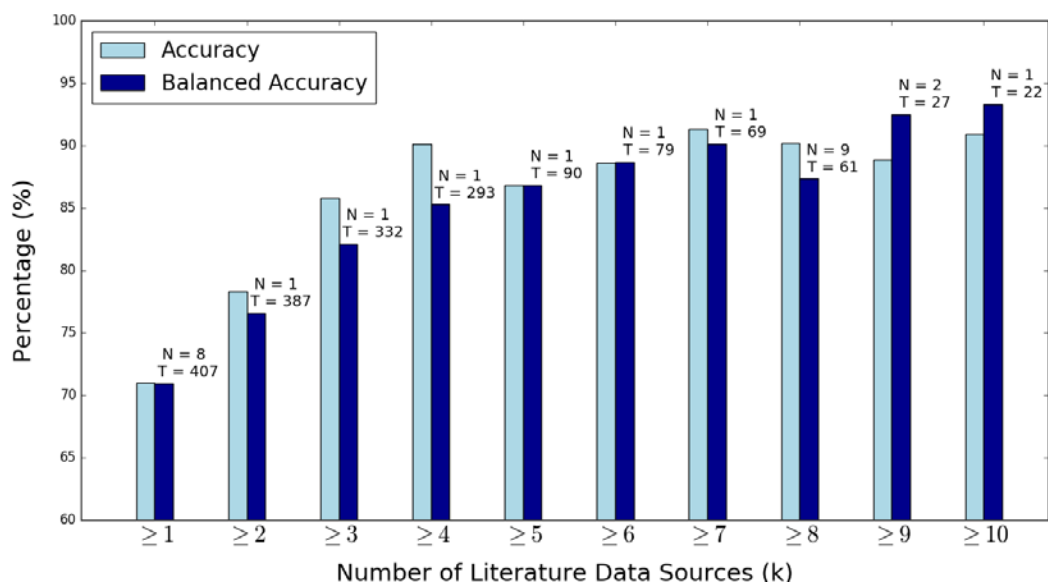


Figure 3: Literature data source analysis to observe the effect of data confidence on read-across predictions. N: number of analogs resulting in the best prediction, T: number of hindered phenols predicted from the restricted dataset. The x-axis corresponds to the threshold in number of data sources (measure of data confidence) and the y-axis corresponds to the maximum accuracy/balanced accuracy of prediction for the dataset. The text on top of each bar plot indicates the number of analogs resulting in the best prediction (N) and the number of hindered phenols that had at least N analogs (i.e. were predicted) from the restricted dataset (T).

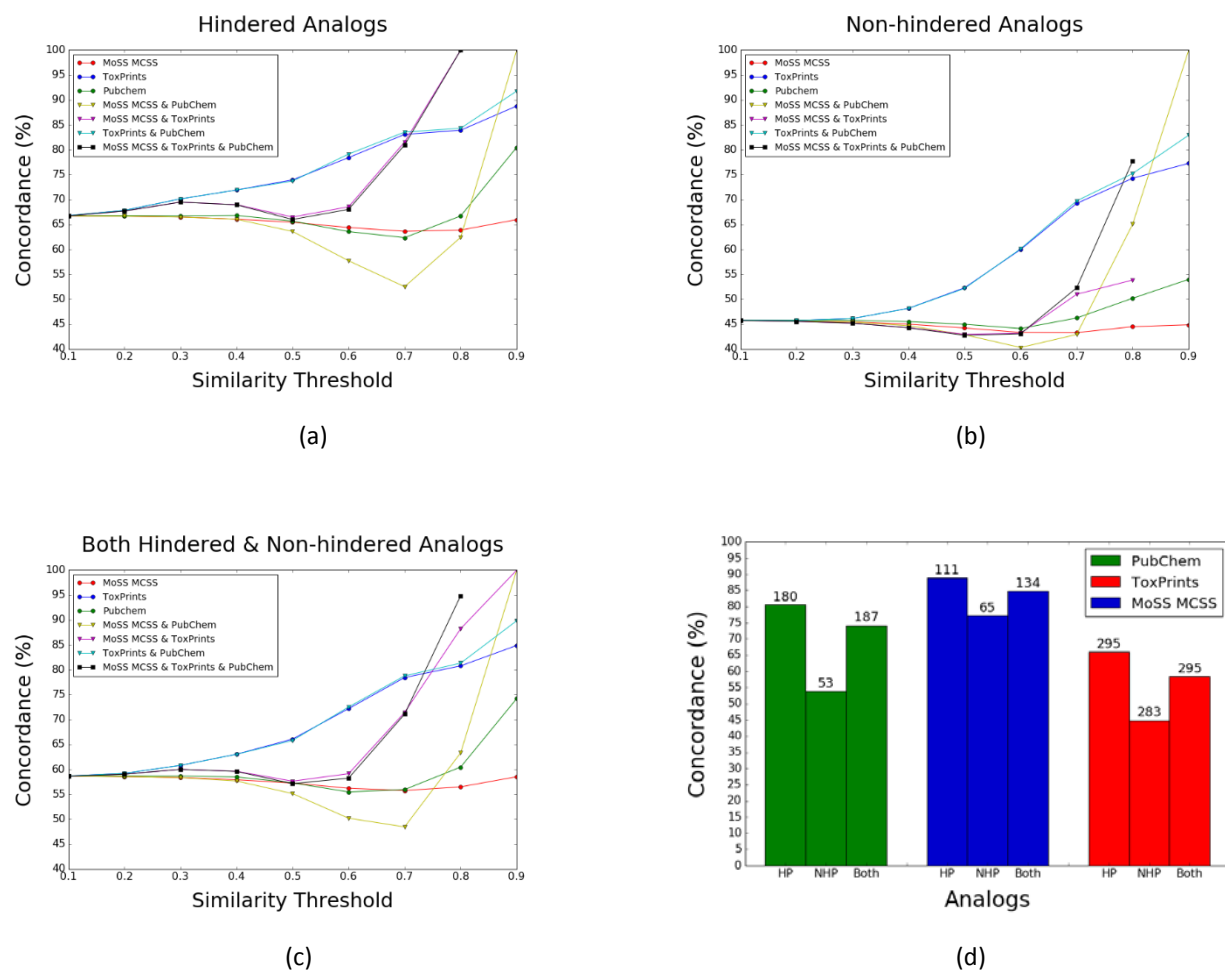
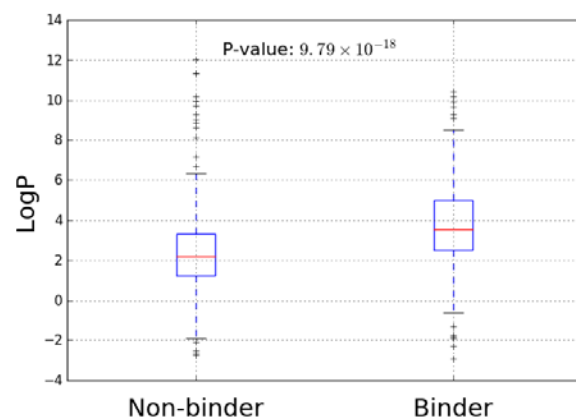
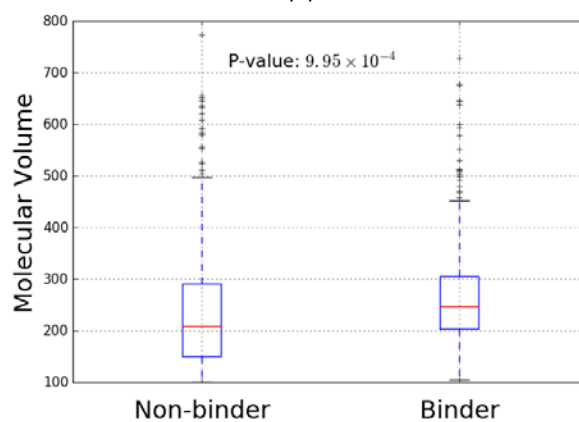


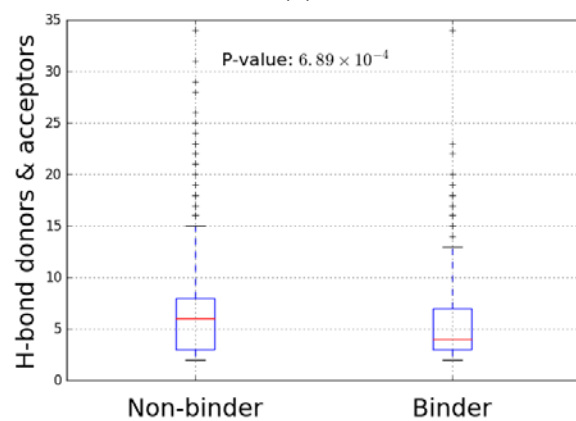
Figure 4: Concordance analysis using phenols with ≥ 4 literature data sources. (a) Using hindered phenols as analogs for each target hindered phenol, (b) Using non-hindered phenols as analogs for each target hindered phenol, (c) Using all phenols as analogs for each target hindered phenol. As shown, the concordance increases as similarity decreases regardless of the fingerprint method and the type of phenols (hindered, non-hindered or both) used as analogs. (d) Histogram of concordance for each of the three descriptor methods using either hindered analogs, non-hindered analogs or both at a similarity threshold of 0.9. The number on top of each bar indicates the coverage as listed in Table 1. As shown, the improvement in concordance using hindered versus non-hindered analogs is not significant as compared to the improvement in coverage when using both hindered and non-hindered analogs.



(a)

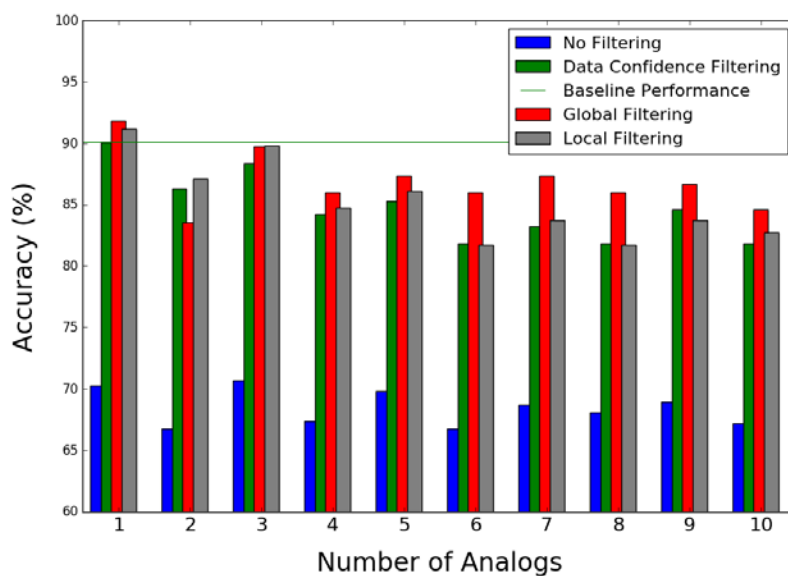


(b)

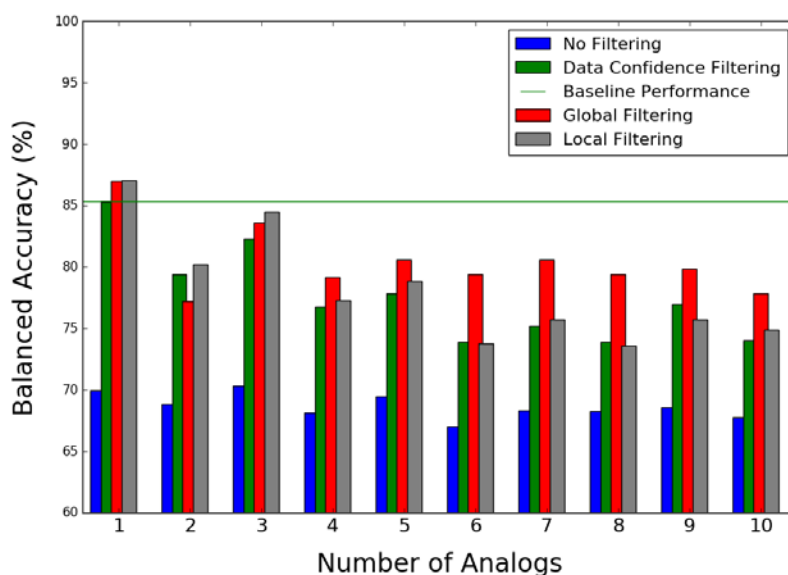


(c)

Figure 5: Box plot distributions of physicochemical properties of hindered phenols with respect to ER binders and non-binders. (a) LogP, (b) Molecular volume, (c) Total number of hydrogen bond donors and acceptors. The box for each plot goes from the first quartile to the third quartile. The red line within the box indicates the median value of the property. The median values for all properties are significantly different between ER binders and the non-binders. The value in the inset textbox is the p-value from the Student's t-test which was used to determine if the values of the global filtering properties are significantly different across ER binders and non-binders.



(a)



(b)

Figure 6: Effect of global and local filtering on analog quality and read-across predictions using PubChem fingerprints. (a) Accuracy, and (b) balanced accuracy. The green bar shows the prediction metrics when the dataset is filtered for data confidence. The best predictive performance, when filtering by data confidence alone, is seen when just one analog is selected for each target. This performance level has been established as the baseline performance level to help evaluate the ability of analog filtering to improve the adequacy of analogs and read-across performance.

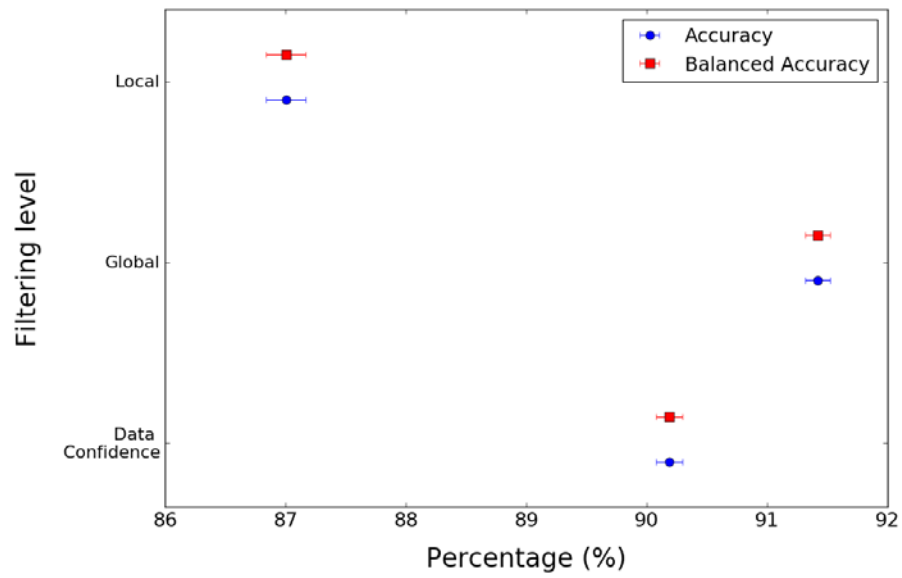


Figure 7: Confidence intervals for accuracy and balanced accuracy estimated using 1000 bootstrap samples. The non-overlapping confidence intervals demonstrate that the filtering methods improved the performance of the predictions relative to the baseline performance. The results are represented as mean (symbols) \pm standard deviation (error bars).

7 Tables

Threshold	Coverage		
	PubChem	ToxPrints	MoSS MCSS
0.1	296	296	296
0.2	296	296	296
0.3	296	296	296
0.4	296	296	296
0.5	296	290	296
0.6	296	276	296
0.7	292	245	296
0.8	266	190	296
0.9	180	111	295

(a)

Threshold	Coverage		
	PubChem	ToxPrints	MoSS MCSS
0.1	296	296	296
0.2	296	296	296
0.3	296	296	296
0.4	296	296	296
0.5	296	270	296
0.6	293	228	296
0.7	255	172	296
0.8	176	116	296
0.9	53	65	283

(b)

Threshold	Coverage		
	PubChem	ToxPrints	MoSS MCSS
0.1	296	296	296
0.2	296	296	296
0.3	296	296	296
0.4	296	292	296
0.5	296	294	296
0.6	296	281	296
0.7	293	255	296
0.8	270	202	296
0.9	187	134	295

(c)

Table 1: Coverage or the number of hindered phenols for which analogs could be selected for concordance analysis above the similarity threshold corresponding to Figure 4 using PubChem,

ToxPrints and MoSS MCSS descriptor methods. (a) Using hindered phenols as analogs, (b) using non-hindered phenols as analogs, and (c) using both hindered and non-hindered phenols as analogs. As shown, coverage decreases as the similarity threshold increases regardless of the type of phenols used as analogs.

R-group	Property	p-value
R2	Basicity	5.44×10^{-9}
R2	LogP	9.01×10^{-19}
R2	Total H-bond donors and	3.62×10^{-6}
R3	Basicity	5.44×10^{-9}
R3	LogP	9.02×10^{-19}
R3	Total H-bond donors and	3.62×10^{-6}
R4	Total H-bond donors and	3.46×10^{-2}
R6	LogP	1.79×10^{-2}

Table 2: R-group properties and p-values from the Student's t-test to determine if the values of the R-group properties are significantly different in phenols that are ER binders or non-binders.