# A Systematic Evaluation of Analogs and Automated Read-across Prediction of Estrogenicity


**Prachi Pradeep[1,2], Kamel Mansouri[1,2], Grace Patlewicz[2] and Richard Judson[2]**

[1]Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee

[2]National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina

Prachi Pradeep | pradeep.prachi@epa.gov | ORCID iD: 0000-0002-9219-4249 | 919-541-5150


www.epa.gov

## INTRODUCTION

Read-across is a data gap filling technique widely used within category and analog approaches to predict a biological property for a data-poor (target) chemical using known information from similar (source analog) chemical(s). Potential source analogs are typically identified based on structural similarity. Although much guidance has been published for read-across, practical principles for the identification and evaluation of the scientific validity of source analogs remains lacking. This case study explores how well 3 structure descriptor sets (Pubchem, Chemotyper and MoSS) are able to identify analogs for read-across and predict Estrogen Receptor (ER) binding activity for a specific class of chemicals: hindered phenols. ndered phenols are phenols with one or more bulky functional groups ortho to the hydroxyl group. E.g. 3-Chloro-4-hydroxybenzoic acid: (  )

For each target chemical, two sets of analogs (hindered and non-hindered) were selected using each descriptor set with two cut-offs: (1) Minimum similarity distance (range 0.1 - 0.9), and (2) Closest *N* analogs (range 1 - 10). The target-analog data was then used to evaluate two key sources of uncertainty in read-across: (1). Data quality - read-across predictions were evaluated for each target hindered phenol using *N* analogs and restricting the data set to include phenols with a threshold on literature data sources as a marker for experimental data quality, and (2). Analog validity - each target-analog pair was evaluated for its concordance with measured ER binding from literature using phenols with greater than or equal to four data sources. The analogs were then subsequently filtered to improve their validity using: (1) physchem properties of the phenol (global), and (2) physchem properties of the R-groups neighboring the active hydroxyl group (local). Subsequently, a majority vote prediction was made for each target phenol by reading-across from the closest *N* analogs.

The data set comprised 462 hindered phenols and 257 non-hindered phenols. The results demonstrate that: (1) The concordance in ER activity rises with increasing similarity, (2) data quality significantly reduces uncertainty in the quality of analogs and read-across predictions, and (3). filtering analogs using global and local properties results in better read-across predictivity. This case study demonstrates how biologically-relevant chemical descriptors can be used to identify valid analogs for read-across.

## OBJECTIVE

**To investigate the utility of various structure descriptor methods for identification of analogs for read-across ER predictions and to assess the improvement in uncertainty of predictions by utilizing data quality measures, physchem properties, and R-group properties for filtering of relevant analogs to ascertain better prediction of ER activity for hindered phenols.**

1. Structural source analogs were identified using 3 different chemical structure descriptor approaches (Pubchem, Chemotyper and MoSS MCSS) and Tanimoto index as a measure of similarity.
2. Concordance analysis and a read-across ER binding prediction was done for each target hindered phenol.

**Analog Selection Method**

| Descriptor Approach | Basis |
|---|---|
| Pubchem (P) | 881 bits fingerprints |
| MoSS MCSS (M) | Size of most common substructure |
| Chemotyper (C) | Chemical substructures fingerprint with pre-defined chemotypes |

**Underlying basis for each of the three chemical descriptor approaches**

**DATASET**

Curated data set from different over lapping sources including: Tox21, FDAEDKB, METI database, ChEMBL and other sources from CERAPP project.

**Target:** 462 hindered phenols
**Inventory of Source Analogs:** 719

**Target phenols (>=4 data sources):** 296
**Inventory of Source Analogs (>=4 data sources):** 481

## METHODS

**WORKFLOW**

**SELECT** analogs for each hindered phenol

**Analogs**
- Hindered analogs
- Non-hindered analogs

**Threshold**
- Similarity cut-off (0.1 – 0.9)
- Number of analogs (1-10)

**EVALUATE sources of uncertainty**

**1. Data quality**
(Impact of number of literature data sources as a measure of reliability of experimental ER outcomes)
- **Read-across** Estrogenicity using Majority Vote Prediction from 10 nearest analogs with greater than *N* data sources (1 -10)

**2. Analog validity**
(Concordance in experimental ER outcomes between each target-analog pair)
- Each descriptor approach
- Combination of descriptor approaches

**FILTER analogs to improve validity**

**1. Global filtering**
(Physchem properties of the whole chemical)

| Filtering Property | Threshold |
|---|---|
| LogP | <= (+/-) 1 unit of the target |
| Molecular volume | <= (+/-) 100% of the target |
| H-bond donor and acceptors | <= (+/-) 6 units of the target |

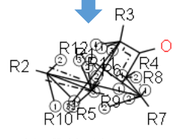
**2. Local filtering**
(Physchem properties of the R-groups neighboring the active –OH group)

| Filtering Group and Property | Threshold |
|---|---|
| R2, R3, R6: LogP | <= (+/-) 3 units of the target |
| R2, R3: hPKb | <= (+/-) 2 units of the target |
| R2, R3, R4: H-bond donor and acceptors | <= (+/-) 3 units of the target |

**R-group Decomposition**

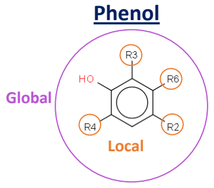
**Use the phenol scaffold to decompose each phenol into R-position substituents**

- Phenols
- KNIME Workflow




Phenol Scaffold with R-group positions

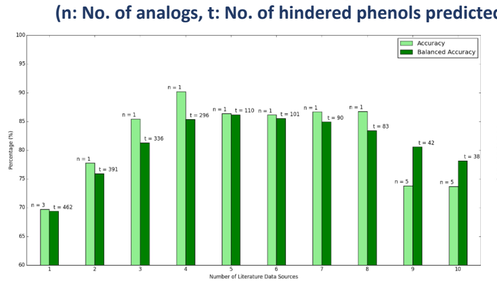
**Phenol**



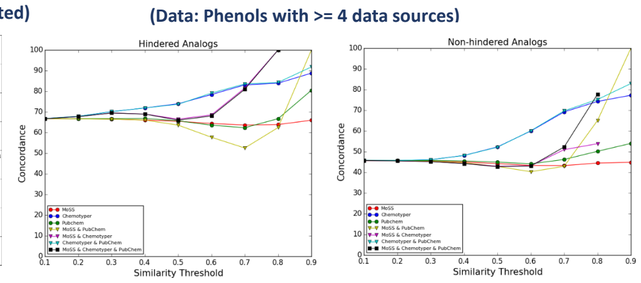

Most frequent R-group substitution positions

## UNCERTAINITY ANALYSIS

**1. Data Quality**
**(n: No. of analogs, t: No. of hindered phenols predicted)**

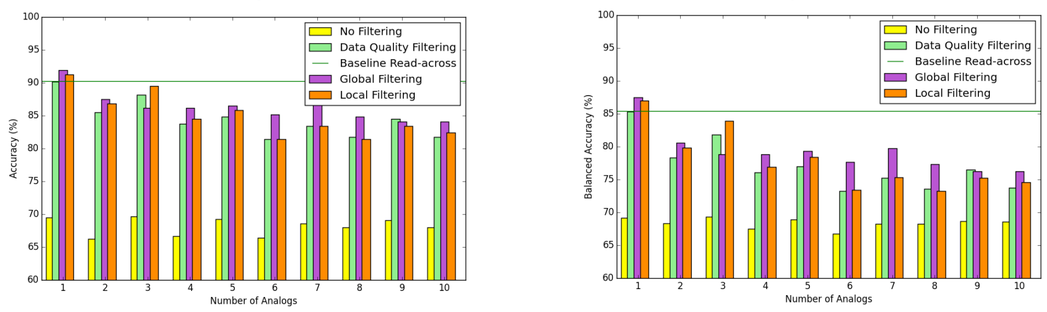



**2. Concordance**
**(Data: Phenols with >= 4 data sources)**

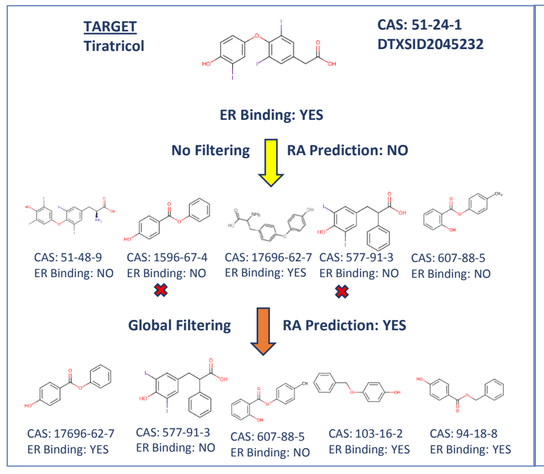


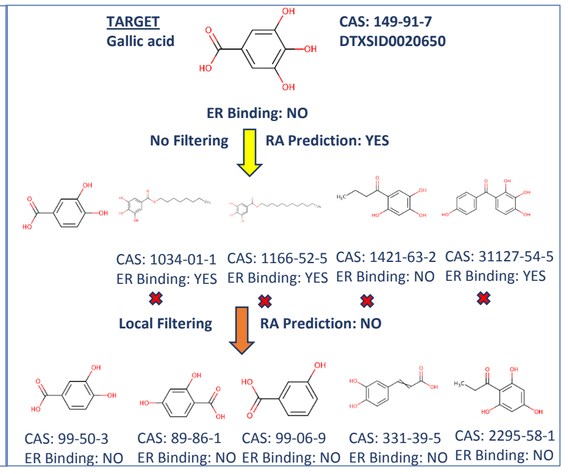

## READ-ACROSS RESULTS

**(Method:** PubChem, **Data:** Phenols with >= 4 data sources)




## ANALOG FILTERING ILLUSTRATIVE EXAMPLES




## CONCLUSION

- Concordance analysis using each target-analog pair (using a similarity cut-off (0.1 - 0.9)) indicates that the concordance in ER activity rises with increasing similarity.
- Data quality analysis illustrates the importance of using good data (validated from multiple sources) and its impact in reducing uncertainty in the quality of read-across predictions. Setting limits on data source thresholds drastically improves prediction accuracy.

Read-across predictions reveal that:

- Filtering of analogs based on conceptually simple steric and electronic properties improves the validity of analogs and subsequently prediction accuracy. (E.g. After data quality consideration using just 1 analog from PubChem (BA increase from 69.2 % to 85.3% ), BA increases to 87.5% when filtered by global properties and to 87.0% when filtered by local properties.)

**Using only one (nearest) analog with good quality data, performs as well as any other combination (balanced or total accuracy). This provides support for using the standard "analog" approach in read-across.**

Future Directions:

- Read-across is a conceptually simple and scientifically sound technique. However, identification of relevant and valid analogs for read-across prediction for any endpoint is not trivial.
- We see a complex interaction between the R-groups and their properties, and physchem properties of the chemical and ER binding. Future research will focus on employing machine learning techniques to identify properties that are most relevant to these interactions.


U.S. Environmental Protection Agency
Office of Research and Development