Building a compendium of expert driven read-across cases to facilitate an analysis of the contribution that different similarity contexts play in read-across

G Patlewicz[1*], Amanda Ross[2], HC Bledsoe[2], Janielle Vidal[2], Louis Groff[1], Brett Hagan[1,3], Imran Shah[1]
[1] CCTE, U.S. EPA, RTP, NC, USA, [2]ICF Corporation, Reston, VA, USA,[3]ORAU, Oak Ridge, TN, USA

Grace Patlewicz | Patlewicz.grace@epa.gov | ORCID: 0000-0003-3863-9689 | 919-541-1540

www.epa.gov

BACKGROUND

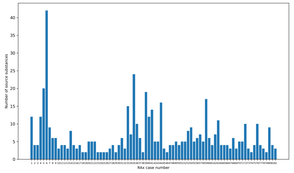
Read-across is a data-gap filling technique utilised to predict the toxicity of a target chemical using data from similar analogues. Read-across is predominantly performed as part of an expert-driven assessment which can impeded broad acceptance. Data-driven approaches such as Generalised Read-Across (GenRA) offer scope to generate reproducible read-across predictions where uncertainties and performance are quantified. A key issue is how to reconcile an expert-driven approach with a data driven approach both in terms of how analogues are identified and evaluated as well as how the read-across prediction is derived. An important component of analogue identification and selection is in understanding the contribution that different similarity contexts play, i.e. does structural similarity play a larger role in analogue selection compared with metabolism similarity. This study aimed to explore some of these considerations through building a compendium of expert-driven read-across assessments that had been published for repeated-dose toxicity endpoints.

METHODS

- Read-across cases of repeated dose toxicity were compiled from the published literature, EPA Provision Peer Review Toxicity Values (PPRTV) assessments as well as OECD IATA case studies.
- A structured excel sheet was created to capture specific information including the target substances being assessed, the candidate source analogues, the toxicity data being read across as well as the rationale use to identify and evaluate the analogues (so-named analogue evidence streams).
- A SOP was developed to ensure consistency in extractions. Extractions were performed by one individual which were then checked for completeness and consistency by a second individual.
- Target and source analogue identities were subsequently mapped to DSSTox content using the EPA CompTox Chemicals Dashboard to augment the information captured to include structures using SMILES.
- The freely available web application ClassyFire was used to categorise all discrete organic structures into classes using its chemistry ontology.
- Similarity contexts evaluated included structure and predicted metabolism.
- Metric learning approaches were attempted to predict to target-analogue associations from chemical structure and predicted metabolism information, the latter generated using OASIS TIMES.

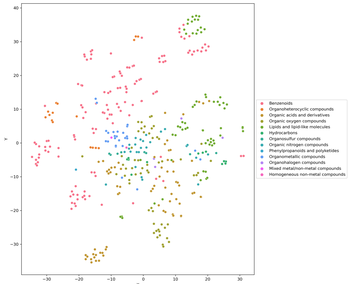
DATASET SUMMARY

- 82 Read-across examples cases were compiled from the three main sources.
- There were 22 unique decision contexts when aggregated by NAMs, technical guidance or regulatory purposes.
- Of the 82 examples, 68 captured regulatory purposes, the remainder were relatively evenly split between efforts to improve guidance or evaluate the utility of NAMs to substantiate read-across justifications.
- The balance of decision contexts is not unsurprising given the origin of the case studies, ~25 (30%) of the cases were taken from the US EPA PPRTV effort, 38% were OECD SIDs examples, 15% were OECD IATA case studies, 10% from journal articles with the remainder comprising a couple of examples each from ECETOC or Health Canada.
- Of the approaches used, there was a bias towards category approaches with 55% of cases utilising a category approach and the 36% being analogue approaches.
- All the EPA PPRTV cases relied on an analogue approach whereas in general over 90% of all other examples used a category approach.



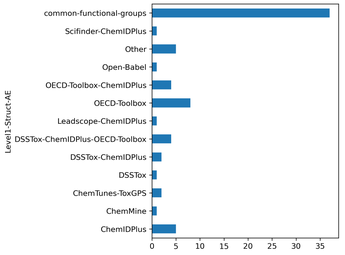
Median number of members within a case was 5 whereas the maximum was 42.

CLASSYFIRE



- The 82 read-across example cases were characterised by 497 individual substances of which 468 could be mapped to a discrete organic structure by way of SMILES
- To gain a perspective of the chemical diversity across these 468 substances, the chemistry ontology ClassyFire was used to assign structures into their respective chemical class.
- Some ~63% of the substances were members of either the "Benzene and substituted derivatives" class (25%), the "Fatty Acyls" class (14%), the "Organooxygen compounds" class (12%) or the "Carboxylic acids and derivatives" class (12%).
- A t-SNE plot based on Morgan chemical fingerprints shows the chemical landscape and is colour coded using a subset of the most populist chemical subclasses.

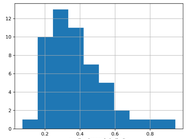
ANALOGUE EVIDENCE STREAMS



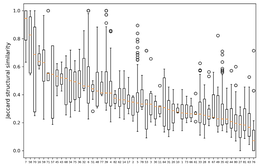
- Across the 82 examples, there were 77 different evidence streams characterising the basis for identifying and evaluating the source analogues in each case. The first component of the evidence stream characterising the primary means of identifying candidate source analogues was a structural one in 72 cases, and metabolism in 4 cases.
- There were 13 different approaches by which analogues were identified. The barplot highlights the main tools and approaches.
- The OECD Toolbox, DSSTox (within the EPA CompTox Chemicals Dashboard), the NIH's structure searching tool within ChemIDPlus or some combination of these tools were most common in terms of identifying structural analogues.
- However by far the most common means of identifying analogues was to look for common scaffolds based on functional groups.

STRUCTURAL SIMILARITY EVALUATION

- The pairwise Jaccard similarities were computed within each read-across example to explore how structurally similar the target and source analogues were amongst themselves.
- The median of the distribution of median values for each case study was determined to be 0.34. (see Fig a)
- There was a large degree of variation in pairwise similarities within each case study (See Fig b)



(a)

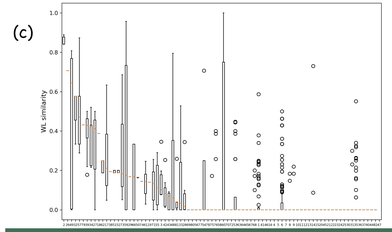


(b)

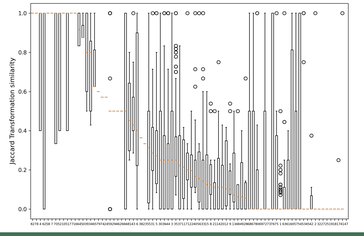
METABOLIC SIMILARITY EVALUATION

- The pairwise similarities were computed within each read-across example to explore how metabolic similar the target and source analogues were amongst themselves with respect to their metabolic graph (c) and transformation profile (d)
- There was a large degree of variation in pairwise similarities within each case study and the similarities were low overall.

(c)



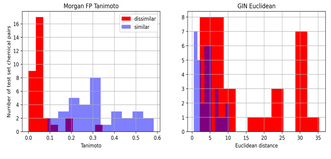
(d)



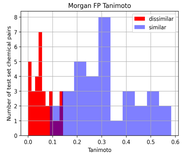
METRIC LEARNING EVALUATIONS

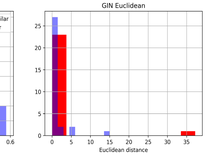
A deep learning metric learning approach was attempted as follows:

- A pairwise matrix was constructed for all substances with SMILES. If a pair of substances were both members of the same read-across case, it was denoted by label 0, otherwise by label 1. A random sampling was performed to downsample the dissimilar pairs.
- A Graph Isomorphism Network (GIN) was structured as a Siamese network such that each pair of substances could be fed into the network, contrastive loss was used as the loss function since it learns embeddings in which two similar substances have a low Euclidean distance and two dissimilar points have a large Euclidean distance. Two networks were investigated: 1) in one network, target-analogue smiles were used as inputs which were converted into pytorch-geometric graphs whereas the second network 2) used predicted metabolic graphs as inputs where the nodes were represented as bit vectors of Morgan fingerprints (FPs) and the edges were represented by the reaction pathway which was one encoded as a feature vector. The intent was to explore embeddings representing chemical structure information and predicted metabolism information.
- Performance was poor in both cases – using a threshold of 5 for network 1 – the accuracy was only 26%. The low structural similarities observed for the read-across pairs is likely to be contributing to this poor performance. The embeddings were unable to discriminate between the similar and dissimilar read-across pairs (Figure e). In network 2, the accuracy was higher at 45% but the discrimination remained poor using the GIN metabolism embeddings (Figure f).



(e)





(f)

CONCLUSIONS

- A set of read-across cases published in the literature and elsewhere were compiled.
- Preliminary work undertaken has evaluated the similarity of the substances within the cases from the perspective of structure and metabolism. Similarities appeared to be low and extremely variable across and within each case example.
- GCN models were attempted in an effort to derive embeddings that could differentiate between similar and dissimilar read-across substances. Further work will consider other metric learning approaches.

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