Toxicological Data Gap Filling of Ingredients; Comparison of Machine Learning Based Imputation and Traditional QSAR Methods

Thomas M WHITEHEAD[1], Joel STRICKLAND[1], Gareth J CONDUIT[1], Alexandre BORREL[2], **Daniel MUCS[3]**, Irene BASKERVILLE-ABRAHAM[3]

[1]Intellegens, Cambridge, United Kingdom [2]Inotiv, Research Triangle Park, NC, United States [3]Scientific and Regulatory Affairs, JT International SA, Geneva, Switzerland

Abstract Number/Poster Board number: 4366/P151

BACKGROUND

At almost every phase of the toxicological risk-assessment process, data gap filling is a recurring challenge. While there are many *in silico* methods available for the different toxicological endpoints, developing, maintaining, and producing predictions using these – such as batteries of Quantitative Structure Activity Relationship (QSAR) models (**Figure 1.**) – in a reproducible manner can be quite challenging and time consuming even for computational experts. Imputation (**Figure 2.**) offers a viable alternative to use multiple different traditional QSAR models simultaneously, and recent results suggest even superior performance compared to using individual structure-based models.

Figure 1: General representation of the QSAR process using AI generated images [1]

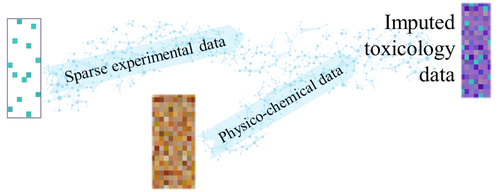


Figure 2: General concept of the imputation approach using sparse experimental, complete phys-chem data to acquire imputed toxicology data [2]

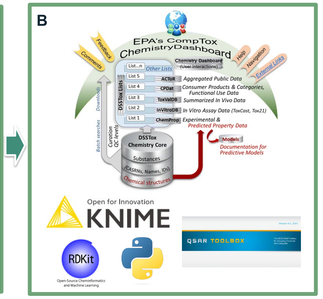
METHODS

In our study, a Machine Learning (ML) based imputation method was tested on an open-source data set [3] comprising of approximately 2500 ingredients (**Figure 3. A**) with limited *in vitro* and *in vivo* data obtained from the OECD QSAR Toolbox [4]. This was then compared to well-established single endpoint ML-based QSAR approaches. We have also explored the impact of augmenting the initial human health focused data with additional ecotoxicological data, to see if it improves model performance (**Figure 3. C**). Data processing was done using KNIME [5], RDKit [6] and Python (Scikit-Learn [7], **Figure 3. B**).

A

Table 1

(1) FL No.	(2) Chemical Name	(3) CAS No.	(4) JECFA No.	(5) CoE No.
01.001	Limonene	138-86-3		491
01.002	1-Isopropyl-4-methylbenzene	99-87-6	1325	620
01.003	Pin-2(10)-ene	127-91-3	1330	2114
01.004	Pin-2(3)-ene	80-56-8	1329	2113
01.005	Terpinolene	586-62-9	1331	2115
01.006	alpha-Phellandrene	99-83-2	1328	2117
01.007	beta-Caryophyllene	87-44-5	1324	2118



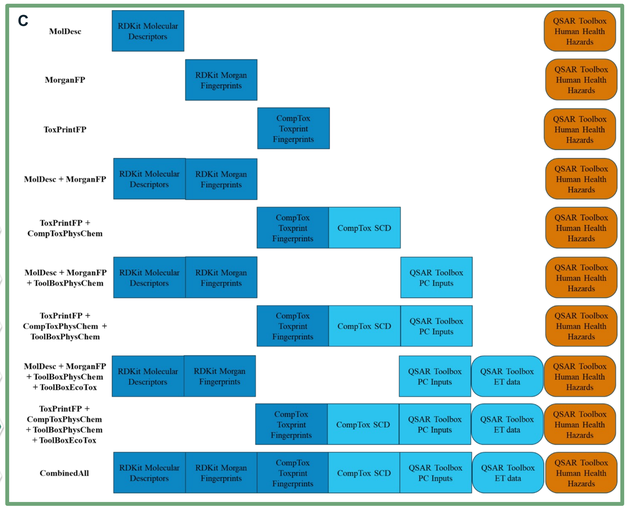


Figure 3: Overall workflow for creating the datasets: A) EU list of flavorings, B) Data sources and processing tools, C) Final datasets created from different combinations of the available feature and endpoint data

REFERENCES

[1] Images created with Microsoft Bing Copilot (DALL-E-3)
[2] Whitehead et al. "Quantifying the Benefits of Imputation over QSAR Methods in Toxicology Data Modeling" J. Chem. Inf. Model. December 13, 2023
[3] Table 1 of Commission Implementing Regulation (EU) No 872/2012
[4] OECD QSAR Toolbox version 4.5 https://qsartoolbox.org/
[5] Berthold et al. "KNIME - the Konstanz information miner", ACM SIGKDD Explorations Newsletter, 11 (1): 26 (2009)
[6] RDKit: Open-source cheminformatics. https://www.rdkit.org
[7] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

RESULTS

A comparative analysis was done between the traditional QSAR models and imputation models (**Figure 4.**), focusing on the same dataset to evaluate the additional benefits imputation models offer by learning from inter-endpoint relationships. Based on the hypothesis that more pertinent data can enhance the performance of machine learning models – in that case, through the inclusion of data on other 'Human Health Hazard' endpoints - we explored augmented datasets that contain additional chemical data (**Figure 5.**). As a last step we added more experimental data to the input data frame (**Figure 6.**). This data came in the form of ecotoxicological measurements from the QSAR Toolbox, and it was hypothesized that its inclusion might improve model performance, if biological information about ecotoxicity correlates with Human Health Hazards biological information. **Figure 7.** shows the variables used as input by the imputation model.

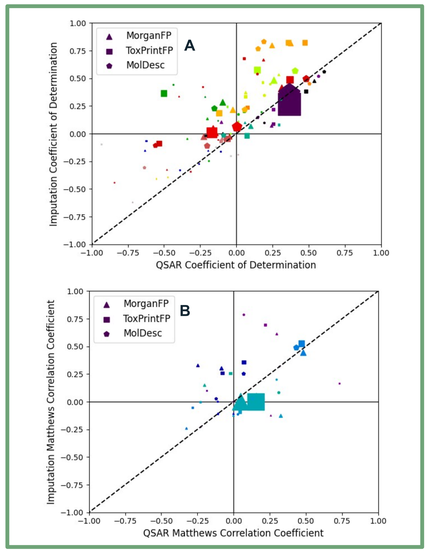


Figure 4: Overall performance for regression (A) and classification (B) endpoints. Color represent the endpoint, and size represents data volume

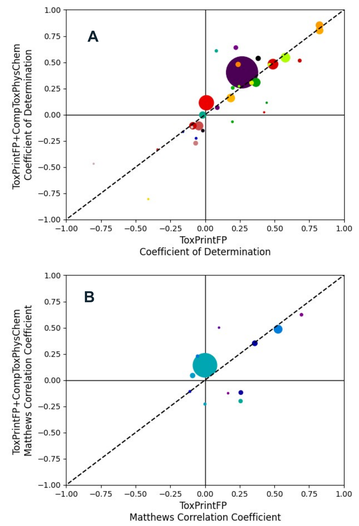


Figure 5: The impact of providing additional chemistry data on regression (A) and classification (B) endpoints. Color represent the endpoint, and size represents data volume

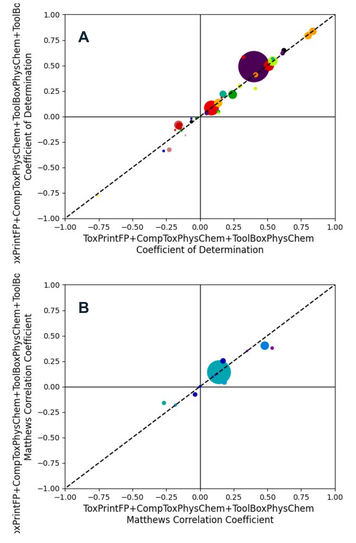


Figure 6: The impact of providing additional experimental data on regression (A) and classification (B) endpoints. Color represent the endpoint, and size represents data volume

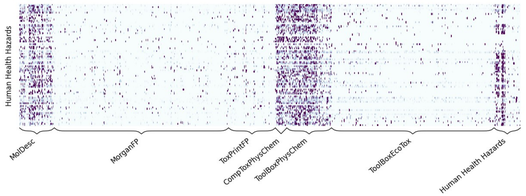


Figure 7: The variables used as input by the model (columns in the figure) to predict each of the 89 target endpoints (rows in the figure). Darker cells in the figure show stronger use of the input variable

CONCLUSION

On top of the improved performance, the imputation approach also showed its resilience to the inclusion of extraneous chemical or experimental data meaning it has a reduced need for laborious manual pre-processing tasks such as feature selection. Ultimately, this method can make data preparation for ML analysis more efficient and easier to manage compared to having multiple single endpoint QSAR models.