# Toxicological Data Gap Filling Using Machine Learning Approaches; Quantifying the Benefits of Imputation over QSAR Methods in Toxicology Data Modelling of Ingredients

*Thomas M Whitehead*†, Joel Strickland†, Gareth J Conduit†, Alexandre Borrel§, Daniel Mucs‡, Irene Baskerville-Abraham‡*

†Intellegens Ltd, Eagle Labs, Chesterton Road, Cambridge, CB4 3AZ, UK

§Inotiv, RTP, NC, USA

‡Scientific and Regulatory Affairs, JT International SA, 8, rue Kazem Radjavi, 1202, Geneva, Switzerland

ABSTRACT

Imputation machine learning (ML) surpasses traditional approaches in modeling toxicity data. The method was tested on an open-source data set comprising approximately 2500 ingredients with limited *in vitro* and *in vivo* data obtained from the OECD QSAR Toolbox. By leveraging the relationships between different toxicological endpoints, imputation extracts more valuable information from each data point compared to well-established single endpoint methods like ML-based Quantitative Structure Activity Relationship (QSAR) approaches. The inclusion of extraneous chemical or experimental data in the imputation models does not impact their performance. This finding is significant because such additional data usually introduces considerable noise and limits performance of single endpoint QSAR modeling. Consequently, imputation models eliminate the need for laborious manual pre-processing tasks such as feature selection, reducing the effort required to prepare data for ML analysis. This successful test conducted on open-source data validates the efficacy of imputation approaches in toxicity data analysis. This work opens the way for applying similar methods to other types of sparse toxicological data matrices and so we discuss the development of regulatory authority guidelines to accept imputation models.

# Introduction

In the past decade industry, academia, and regulatory scientists have invested significant effort to promote next generation risk assessment (NGRA). Their goals are to develop new approach methodologies (NAMs) for animal testing that have more human relevance for chemical toxicity testing and that reduce the numbers of animals used. By 2035 the US EPA plans to eliminate all mammalian study requests and funding (Gwinn et al., 2020). Similar efforts are ongoing in the European Union where guidance from the European Chemical Agency (ECHA) is encouraging the use of alternative methods such as Quantitative Structure Activity Relationship (QSAR) modeling and *in vitro* testing instead of animal testing (Westmoreland et al., 2022). NAMs open many perspectives in risk-based metrics as well as chemical prioritization and screening in a fast and cost-efficient way. However, with the development of NAMs, regulatory authorities such as the Organization for Economic Co-operation and Development (OECD) have established guidelines for the validation of QSAR models for regulatory purposes that are more constrained compared to general QSARs used primarily for R&D purposes (OECD, 2014).

Combining *in vitro* with *in silico* modeling to impute data is one of the NAMs methods currently in development. *in vitro* assays often lack comprehensive chemical testing and quality control leading to a rather sparse database from which to perform a complete and holistic toxicological risk assessment. QSAR models are now being leveraged to impute the data (DiMaggio et al., 2010; Kensert et al., 2018; Kovarich et al., 2019).

As an alternative to single-endpoint traditional QSAR models, various imputation approaches have been developed, ranging from empirical read-across methods or simple statistical models to complex machine or deep learning based approaches, or even encapsulating frameworks such as

conformal prediction or transfer learning (Kensert et al., 2018; Kovarich et al., 2019; Norinder et al., 2014; Pradeep et al., 2019; Simm et al., 2015; Simões et al., 2018; Sun et al., 2022)

Previous efforts have demonstrated that imputation allows for the combination of molecular descriptors with sparse bioactivity responses to enrich the QSAR training set and improve overall prediction for a specific endpoint (Whitehead et al., 2019)(Walter et al., 2022). However, none of these imputation models have been discussed in terms of their ability to fit within the current guidelines for potential use for regulatory purposes.

In this study we develop a data imputation model for ingredients that are recognized by the EU as part of the human health toxicological risk assessment for exposure routes such as inhalation or dermal uptake. Data were gathered from all the available information from publicly available sources using the OECD QSAR Toolbox. The sparsity of the available data already makes it challenging to perform the initial hazard assessment step, which provides a great case study for imputation. A particular emphasis will be placed on considering routes to regulatory acceptance.

# Material

We used a set of ingredients covered by EU Regulation 872/2012 (EU Commision, 2012) (https://eur-lex.europa.eu/eli/reg_impl/2012/872/oj). Chemicals were extracted using their CAS Registry Number (CAS RNs). To obtain the chemical structures, we conducted a batch search on the US EPA chemical dashboard (Williams et al., 2017) extracting the structures represented by their QSAR ready Simplified Molecular Input Line Entry System (QSARr SMILES). Chemicals without QSARr SMILES as well as duplicate based on their CAS RNs were omitted resulting in a set of 2,363 chemicals.

Experimental data to impute were extracted from the OECD QSAR Toolbox version 4.5 (*QSAR Toolbox*, n.d.) (referred to hereafter as 'QSAR Toolbox' data). The 89 Human Health Hazard endpoints taken as targets were all the endpoints under the 'Human Health Hazard' category in the OECD QSAR Toolbox data with 2 or more experimental datapoints present among the compounds considered. Of these endpoints, 69 were continuous and 20 were binary (positive/negative). Where data for binary endpoints was marked as 'equivocal' it was removed from the analysis: this removed 49 total data points from seven targets. Summary statistics are only reported on those endpoints with 10 or more data points present to minimize noise in the results.

The modelling data was further subdivided into modeling datasets focused on different modeling strategies and data.

First, three distinct sets of data focused on purely describing chemical structures: (i) Molecular Descriptors (MolDesc) using 121 calculated chemical properties using RDKit Python library

(version 2022.09.1), including molecular weight, SlogP, substructure counts, etc; Morgan Fingerprint (MorganFP) using structural fingerprints calculated with RDKit; and (iii) Tox Print fingerprint (ToxPrintFP) using structural fingerprints specifically deemed relevant for toxicity (Yang et al., 2015) computed using the EPA Chemical Dashboard.

Second, two sets focused on describing physico chemical properties: (iv) sparse calculated data from the US EPA CompTox Chemicals Dashboard (CompToxPhysChem) and (v) physical chemistry properties from the QSAR Toolbox data (ToolBoxPhysChem).

Finally, the last set (vi) consisted of ecotoxicological data (EcoTox) extracted from the QSAR Toolbox (ToolBoxEcoTox).

These datasets were combined in a hierarchical manner to examine the effect of different data types (computation vs experimental; molecular descriptor vs fingerprint vs physico-chemical information; etc.). These sets were combined into a dataset called “CombinedAll”. Combination of these modeling sets are presented in Figure 1 and the data volume of each combination is presented in Table 1

*Figure 1: Graphical representation of the datasets used in modelling and their relationships. Square boxes are computationally calculated data;, rounded boxes are experimental data. Dark blue represents complete input data; light blue represents sparse input data; orange represents (sparse) target data.*

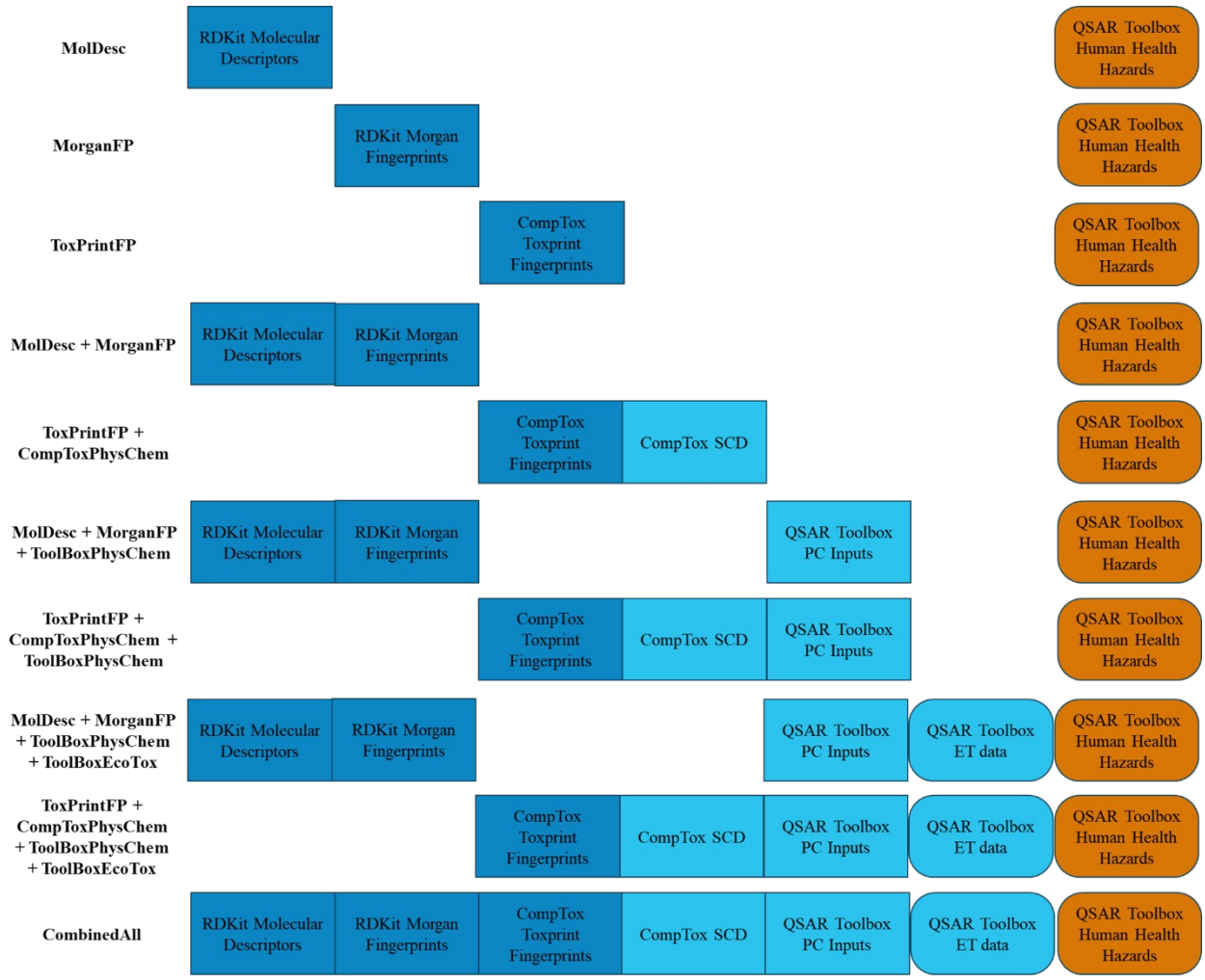



*Table 1: data volumes used in each of the modelling datasets.*

| Modelling dataset | Number of compounds | Number of complete input variables | Number of sparse input variables | Number of non-target experimental endpoints | Number of target experimental endpoints |
|---|---|---|---|---|---|
| MolDesc | 2363 | 113 | 0 | 0 | 89 |
| MorganFP | 2363 | 1020 | 0 | 0 | 89 |
| ToxPrintFP | 2363 | 276 | 0 | 0 | 89 |

| | | | | | |
|---|---|---|---|---|---|
| MolDesc+MorganFP | 2363 | 1133 | 0 | 0 | 89 |
| ToxPrintFP+CompToxPhysChem | 2363 | 278 | 32 | 0 | 89 |
| MolDesc+MorganFP+ToolBoxPhysChem | 2363 | 1133 | 112 | 0 | 89 |
| ToxPrintFP+CompToxPhysChem+ToolBoxPhysChem | 2363 | 278 | 144 | 0 | 89 |
| MolDesc+MorganFP+ToolBoxPhysChem+ToolBoxEcoTox | 2363 | 1133 | 112 | 657 | 89 |
| ToxPrintFP+CompToxPhysChem+ToolBoxPhysChem+ToolBoxEcoTox | 2363 | 278 | 144 | 657 | 89 |
| CombinedAll | 2363 | 1411 | 144 | 657 | 89 |

# Methods

## Machine learning methods

We used random forest (RFs) for the QSAR models, using the implementations in scikit-learn (Pedregosa et al., 2011). Random Forest (RF) is a powerful and versatile machine learning algorithm that works on the principle of ensemble learning. RF is often used in QSAR studies because it can handle large, complex datasets with high dimensionality and multi-collinearity, typical in QSAR studies. RF is often considered the “gold standard” in QSAR modelling (Kwon et al., 2019) due to its robustness, performance, interpretability, efficiency, and non-linearity. Despite the exceptional performance of RFs in QSAR modelling, they possess certain limitations. For instance, they do not natively handle missing values and are primarily designed for single target predictions. We do note here, that RFs have limitations when it comes to extrapolating new data for regression problems (continuous variables), however for the sake of the imputation comparison study, we deemed this method sufficient for these endpoints as well.

AlchemiteTM is a machine learning-based tool used for making predictions from sparse and noisy data, making it highly effective in situations where there are numerous missing values (Whitehead et al., 2019). Previous studies (Mahmoud et al., 2021) have shown an improved performance over traditional QSAR methods. Unlike other approaches, AlchemiteTM simultaneously learns correlations between all input and output variables, enabling interpolation and extrapolation in multi-dimensional space (Irwin et al., 2020). Incorporating additional data enhances the robustness of models, as a larger dataset allows AlchemiteTM to learn and generalise better, minimising overfitting. Additionally, by considering all output variables jointly, AlchemiteTM improves prediction accuracy by accounting for interdependencies and

influences among the outputs (Tse et al., 2021). Notably, AlchemiteTM accurately quantifies data and model uncertainty, providing a measure of prediction confidence. By identifying the most uncertain datapoints, AlchemiteTM focuses on these challenging instances during training, leading to improved modelling performance. This comprehensive approach, along with the ability to propagate uncertainty throughout the modelling process, enables AlchemiteTM to capture dependencies and correlations that may go unnoticed when considering only a single output or a subset of data, as is the case with a conventional RF approach.

## Analysis methods

All evaluation was carried out using 10-fold cross-validation: the dataset was randomly split by compound into 10 disjoint segments; 9 segments were combined into a training set with the last predicted against; then this was repeated so that each segment was used as the prediction set exactly once. In this way all the data in each set was predicted against without being used as input in the same model. Accuracy measurements were calculated by concatenating all 10 prediction sets and calculating accuracy metrics across all present data points per endpoint. Compared to the more common method of calculating accuracy metrics on each prediction set individually and then averaging, the concatenation method increases the statistical power of the metric by providing more data points to each calculation but does not provide access to an estimate of the variance in the metric; it is however equally blind with respect to the training data on each fold. For endpoints with only a small number of datapoints present (all except three endpoints had fewer than 200 datapoints to test against) the concatenation approach is more suitable.

We use the Coefficient of determination ($R^2$) as the performance metric for continuous endpoints: this metric is defined as

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \overline{y})^2}$$

where $f_i$ are the model predictions for the $i$th datapoint, with corresponding experimental measurements $y_i$, and $\overline{y}$ is the mean experimental measurement.

We use the Matthews correlation coefficient (MCC) as the performance metric for binary endpoints: this metric is defined as

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where TP is the number of true positive observations (i.e. the experimental value is positive and the model predicts positive), TN is the number of true negative observations, FP is the number of false positive observations, and FN is the number of false negative observations. MCC is a symmetric cost function that rewards success at capturing both binary classes equally.

For carrying out statistical comparisons of performance across endpoints between models we use Wilcoxon signed-rank tests implemented in SciPy(Virtanen et al., 2020). We use Wilcoxon signed-rank tests rather than t-tests because both $R^2$ and MCC have finite ranges, and (particularly for the binary endpoints) the number of endpoints is small enough to impact the reliability of the Gaussian approximation in a t-test.

# Results

## Imputation compared to QSAR

We first compare traditional QSAR models to imputation models on the same dataset to compare the extra value that learning from inter-endpoint relationships can provide. As QSAR models are traditionally trained on chemical structure related data, meaning molecular descriptors or fingerprints as feature vectors, and require complete input information, we trained QSAR models on the ‘MorganFP’, ‘ToxPrintFP’, and ‘MolDesc’ datasets.

In Figure 22 we compare the performance of these RF based QSAR models against models trained using the imputation method, separately for continuous binary variables. In each case a point is plotted for each target variable for each of the three modelling datasets, with points of the same color corresponding to the same target endpoint within each graph, and the size of the points corresponding to the number of datapoints that were present in the target endpoint. We only include those endpoints that have 10 or more data points in order to avoid cluttering the plots.

We observe that in general the imputation models outperform the corresponding QSAR models. Using the one-tailed Wilcoxon test the continuous variable imputation performance on endpoints with 10 or more data points significantly outperforms the corresponding QSAR models for the ‘MorganFP’, ‘ToxPrintFP’, and ‘MolDesc’ sets with p-values 3e-6, 5e-5, and 1e-3 respectively. This indicates that the use of inter-endpoint relationships by the imputation method have significantly improved the predictive accuracy of the models, validating the

expectation that leveraging relevant additional data in machine learning models can improve their performance. None of the imputation models show evidence of different performance on average from the corresponding QSAR models for binary endpoints.

We can also use Figure 2 to investigate how the input data sets compare on the QSAR and imputation models. For the continuous endpoints there is no evidence of a statistically significant difference between the imputation models on the different datasets, and for the categorical endpoints only the ‘ToxPrintFP’ shows a significant improvement over the ‘MorganFP’ dataset using the imputation models (p-value 6e-3). This provides evidence that in general an expert should be able to select whichever input information from these sets is most chemically informative to extract maximum value from the modelling process without impacting the imputation performance. Generally, the ‘MolDesc’ set contains the most chemically interpretable information, although the ‘ToxPrintFP’ fingerprints are expected to capture properties more relevant to toxicological endpoints.

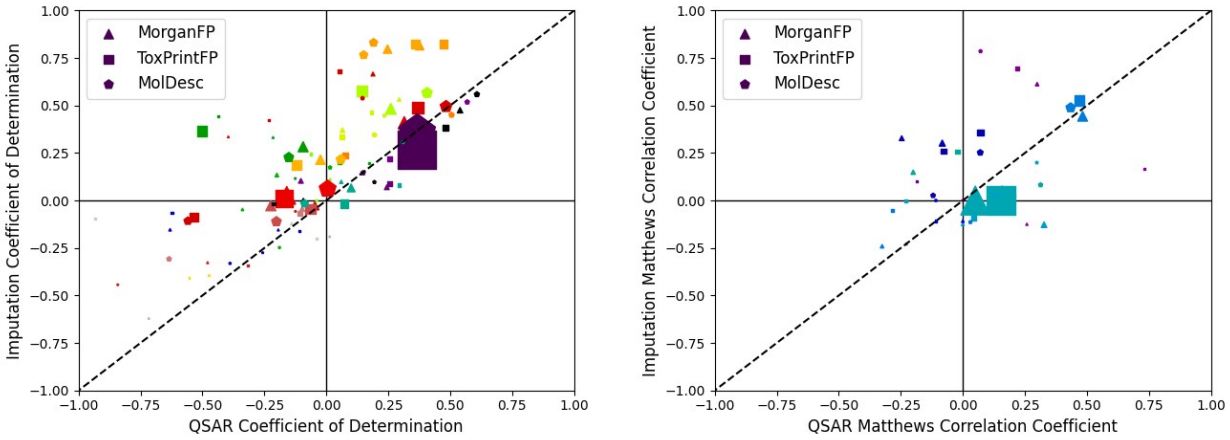
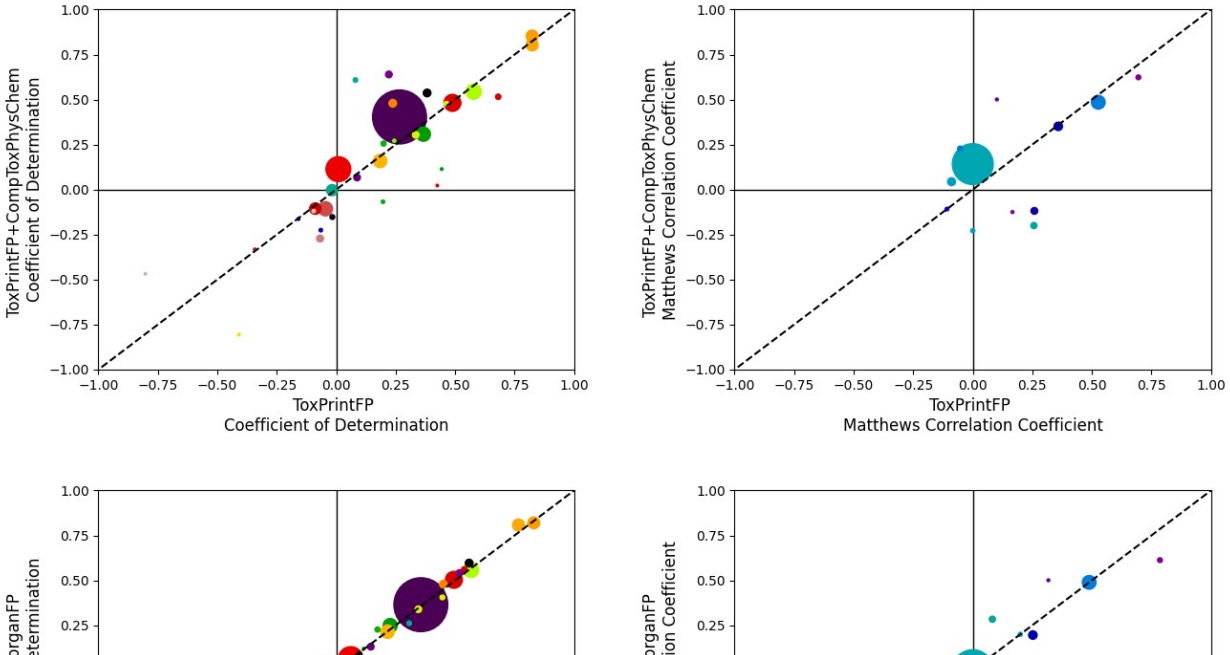



*Figure 2 performance of imputation and QSAR models using the same datasets for the continuous (left) and binary (right) endpoints separately.*

## Providing additional chemical data

Given the observation in the previous section that the provision of more relevant data to the machine learning models (there in the form of data on other ‘Human Health Hazard’ endpoints) improves the model performance, in this section we examine further data sets (summarised in Table 1) that include more input data information. All models are “imputation” models in the vein of the previous section, directly including inter-endpoint relationships.

We first compare the addition of further chemically-relevant information to the ‘ToxPrintFP’ dataset to generate the ‘ToxPrintFP+CompToxPhysChem’ set of all the information extracted from the US EPA CompTox Chemicals Dashboard and, separately, the combination of the ‘MolDesc’ and ‘MorganFP’ sets into the ‘MolDesc+MorganFP’ set of all the information obtained from RDKit. The results of these combinations are shown in Figure 3: 3 below: no combinations show a significant improvement over the ‘ToxPrintFP’/’MolDesc’/’MorganFP’ datasets alone, indicating that in general all of the chemical information provided by the additional data in the ‘ToxPrintFP+CompToxPhysChem’ and ‘MolDesc+MorganFP’ datasets is already captured in the simpler input information already available to the models.



To investigate whether providing a different source of data to the models would improve the performance further, we next examine the 'ToxPrintFP + CompToxPhysChem + ToolBoxPhysChem' and 'MolDesc + MorganFP + ToolBoxPhysChem' datasets, which supplement the previous datasets with physical chemistry data from QSAR Toolbox. The results are shown in Figure 4: in no case does the addition of the physical chemistry information significantly improve the model performance compared to the versions without the QSAR Toolbox data. As expected from the results in Figure 3, neither the 'ToxPrintFP + CompToxPhysChem + ToolBoxPhysChem' nor 'MolDesc + MorganFP + ToolBoxPhysChem' models show any evidence of significant improvements over the simplest 'MorganFP', 'ToxPrintFP', or 'MolDesc' imputation models either, indicating all the relevant chemical information is provided in each of those simplest data sets.

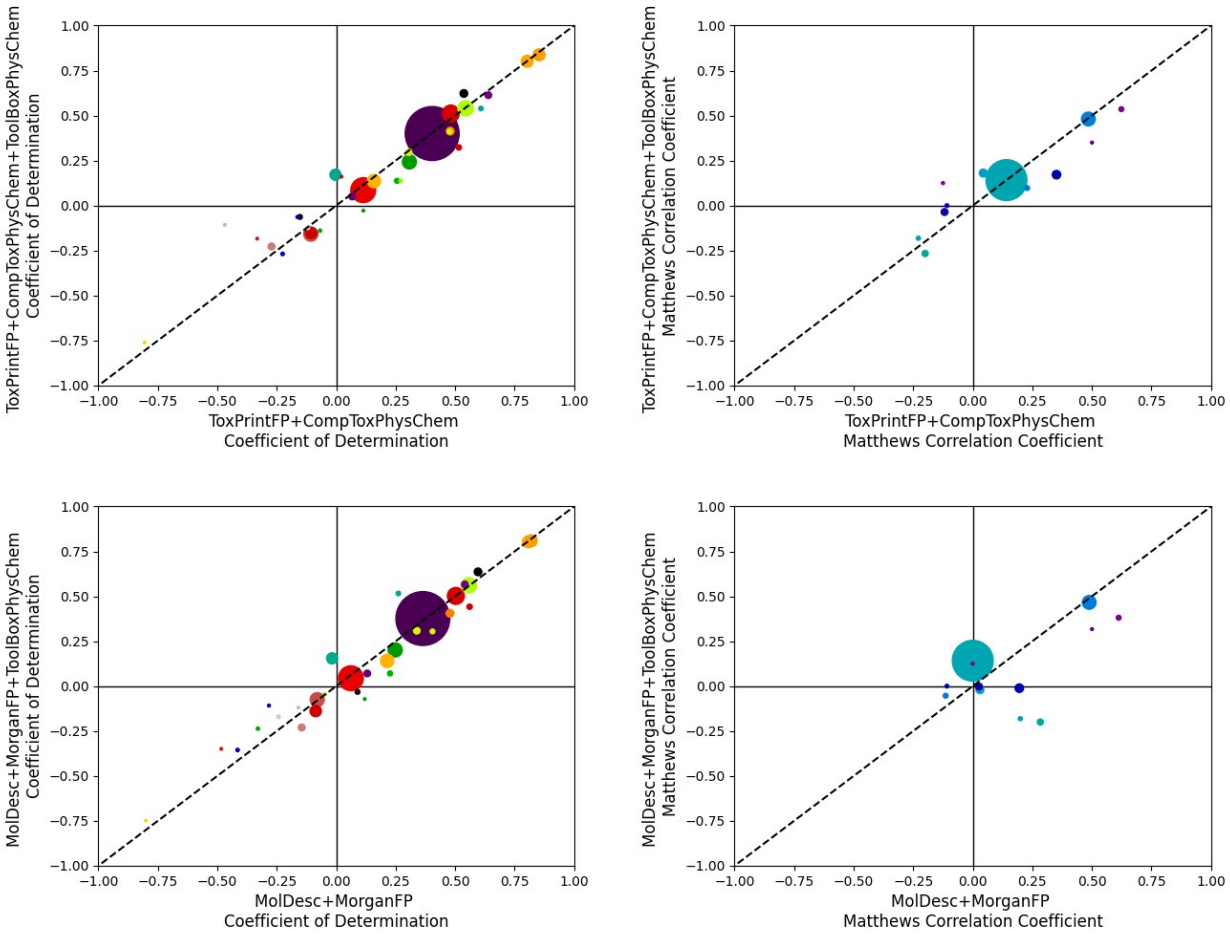
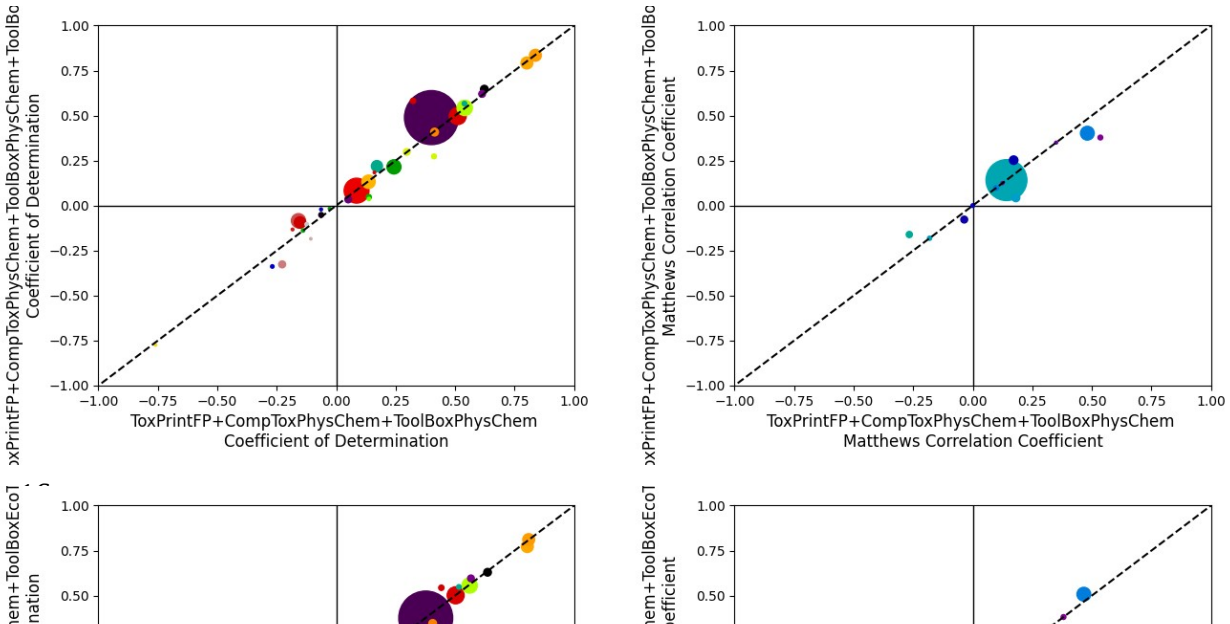


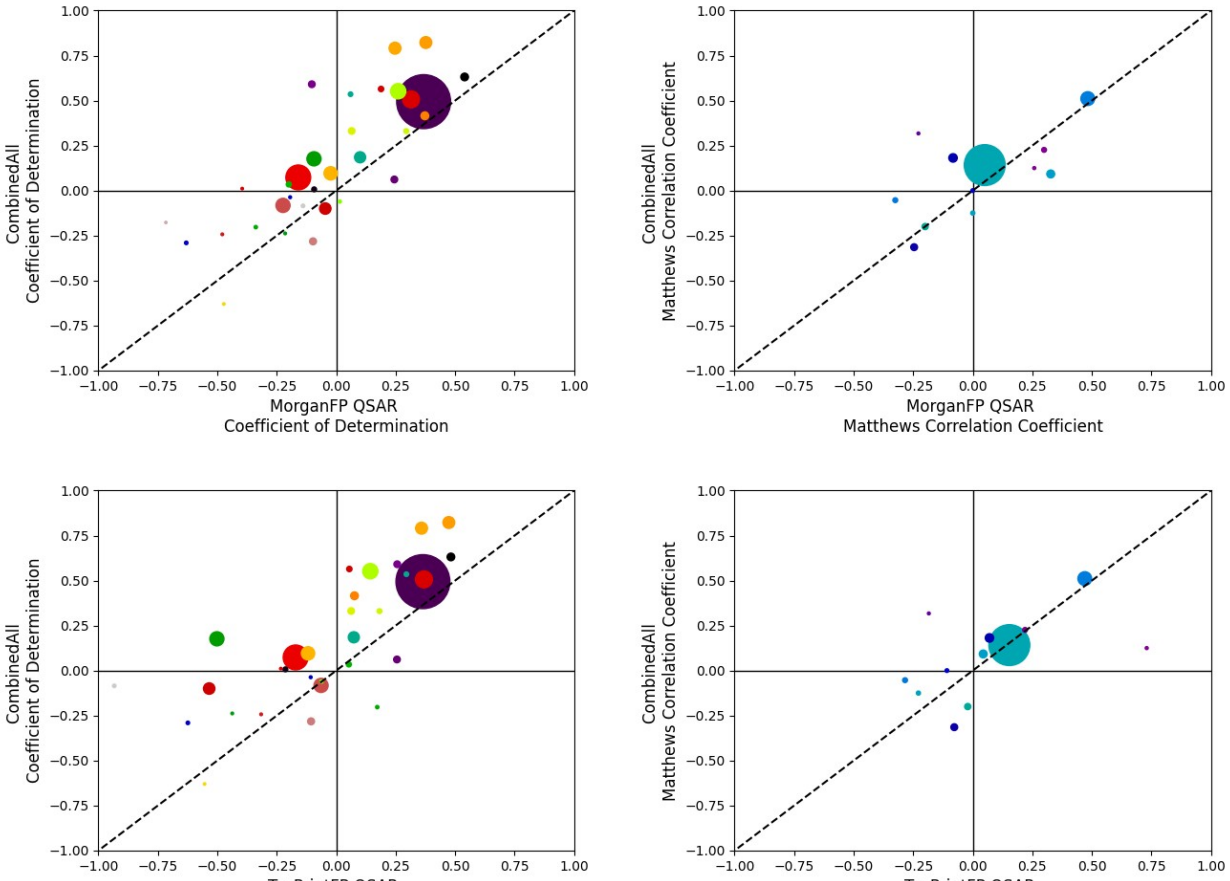

*Figure 4: performance of the ToxPrintFP+CompToxPhysChem+ToolboxPhysChem models compared to the ToxPrintFP+CompToxPhysChem models and performance of the MolDesc+MorganFP+ToolboxPhysChem compared to MolDesc+MorganFP respectively for the continuous (left) and binary (right) endpoints separately. The size of the dots is*

## Providing additional experimental data

Given the lack of improvement in model performance achieved by providing additional physical and chemical information to the models, we next examine providing more experimental data in the form of ecotoxicological measurements from QSAR Toolbox. This might be expected to improve the model performance if there is an expectation that the biological information about ecotoxicity is related to the Human Health Hazards biological information. However, the results in Figure 5 show that again there is no significant change in model performances over the corresponding models without the addition of the ecotoxicological information. Whilst this means that the additional information has not improved model performance, neither has the performance degraded: the addition of information that the imputation models do not find useful for predicting the ‘Human Health Hazard’ target endpoints does not reduce the model’s performance. Often modelling processes can be slowed by the desire to remove extraneous information so that models do not overfit to noisy “accidental correlations” in a dataset, but these results indicate that using this imputation approach this can be avoided, and non-informative information is ignored rather than reducing the model accuracy. This can provide a valuable time saving in modelling projects, eliminating the need of prior feature selection.



For the final analysis we combined all of the modelling data into a single dataset, the CombinedAll set described in Table 1. The performance using this data in an imputation model is compared to the baseline QSAR models in Figure 6: , demonstrating that the combination of additional data and imputation has provided statistically significant improvement on continuous endpoints over the ‘MorganFP’, ‘ToxPrintFP’, and ‘MolDesc’ QSAR models (Wilcoxon p-values 2e-5, 2e-5, and 3e-4 respectively). As discussed above, the majority of this improvement can be attributed to the imputation approach capturing relationships between different ‘Human Health Hazard’ endpoints. The ‘CombinedAll’ model does also provide a statistically significant improvement over the 'MolDesc+MorganFP+ToolBoxPhysChem+ToolBoxEcoTox' model for continuous variables (p-value 2e-2), but there is no evidence it offers an improvement for binary variables or over the 'ToxPrintFP+CompToxPhysChem+ToolBoxPhysChem+ToolBoxEcoTox' model.



# Final model performance analysis

So far, we have examined comparative performance of models trained on different datasets. Now we turn to an analysis of the overall performance of the ‘CombinedAll’ model trained on all of the data.

In Figure 7 we examine the performance of the ‘All’ model as a function of the number of data points per endpoint. We observe that for the continuous endpoints the coefficient of determination generally increases with the amount of data available (p-value for linear trend line to have a positive slope: 4.9e-2), indicating that increasing the amount of data available to learn from increases the model performance. This suggests that as more data is gathered for experimental endpoints over time the model performance on those endpoints should also increase. The slope of the linear fit trend line for the continuous endpoints suggests that adding 10 new experimental data points increases the model coefficient of determination by 0.009(5) on average (assuming the endpoints are otherwise equal), providing quantifiable evidence for the value that new experimental data brings to the computational models.

There is no evidence at the 5% confidence level that the same improvement is true for the categorical endpoints using this model.

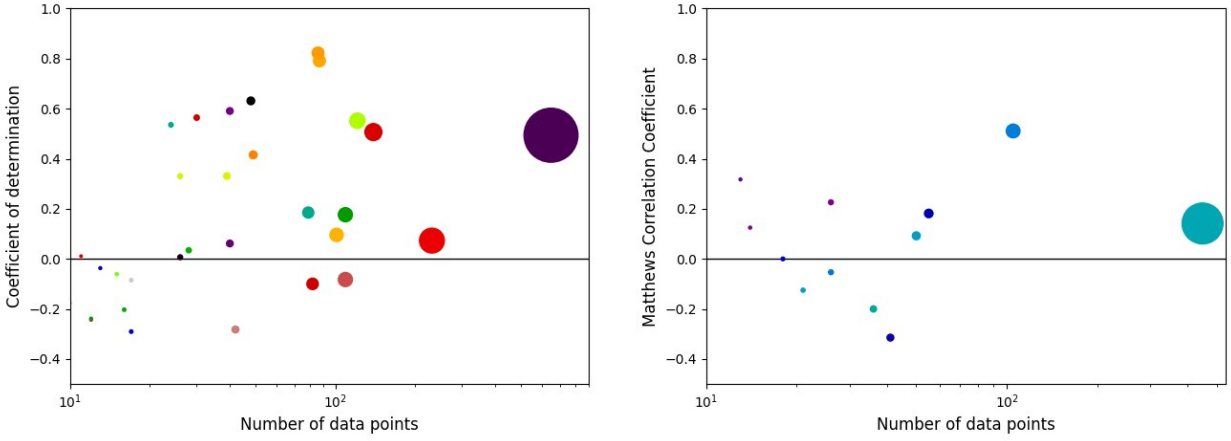
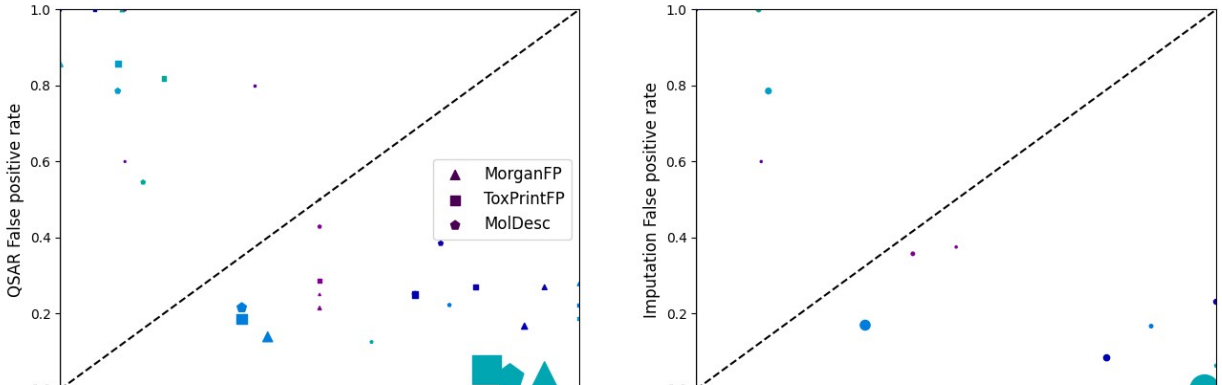



*Figure 7 performance of the 'CombinedAll' model containing all the available data as a function of the number of data points in each endpoint, for the continuous (left) and binary (right) endpoints separately.*

For toxicological endpoints, as well as overall performance, it is also important to understand whether predictive models are generally over- or under-predicting: for example, in the case of binary endpoints, whether false positives for toxicity (which might result in missed opportunities for new compounds) or false negatives (which might result in dangerous compounds being used expecting them to be safe, which is a more costly failure) are more prevalent in predictions.

Figure 8 compares the false positive rate (FPR) to the false negative rate (FNR) for the QSAR models (left) and the final ‘CombinedAll’ imputation model (right). A perfect model would have zero false positives and zero false negatives and so would appear in the bottom left of the plots. A high false positive rate (FPR) signifies that the model misclassifies a significant number of negative data points, while a high false negative rate (FNR) indicates that the model misclassifies a significant number of positive data points. From a toxicology standpoint, it is important to minimize false negatives in order to avoid underestimating relevant information related to toxicity and human health.

For all of the models we observe that the FNR is generally greater than the FPR, indicating that the models are more likely to mischaracterize a toxicologically active compound as inactive rather than a toxicologically inactive compound as active. This is the more harmful model failure mode. In order to improve this in future models the FNR could be taken as the binary variable cost function in hyperparameter optimization, rather than the more balanced Matthew’s Correlation Coefficient used above.



For real-world applications of machine learning it is also important to understand when the model is (and is not) accurate so that the model is only applied within its domain of applicability where there is an expectation of performant predictions. In Figure 99 (left) we plot the absolute deviation of the model predictions $f$ from the true values $y$ (for the continuous variables only), normalized by the standard deviation across the training data for each endpoint ($s$), against the uncertainty in each model prediction, similarly normalized. The different colors represent data from the different continuous endpoints, using the same color scheme as the previous figures. If the distribution of uncertainties was perfectly normal this plot would show a smoothed triangular structure with points mostly below the identity line and mostly in the bottom left of the plot: the tail of points towards the bottom right of the plot indicates the model is sometimes overly pessimistic, overestimating the uncertainty in its predictions. However, a linear trend line for the data in this figure has a positive slope (with p-value 1e-4), indicating that overall the model uncertainty can be used to indicate where the model is accurate: predictions with larger uncertainty are on average further from the true values. A user can then set a tolerance on the absolute deviation from true values they are comfortable with from the model and discard predictions with uncertainties that are too large.

In Figure 9 (right) we plot a 2D UMAP (Uniform Manifold Approximation and Projection) embedding (McInnes et al., 2020) of the data, compressing the complete MolDesc descriptors from RDKit into a 2D representation. Each point represents a single compound, with the color of each point corresponding to the average normalized uncertainty in the continuous predictions across all target endpoints for that compound in the final ‘CombinedAll’ model (grey points only had binary endpoints present). We observe that, in general, the most uncertain compounds are located towards the edges of the clusters of data. This suggests that the model is applicable

across multiple areas of chemical space, but within each of these areas is more confident (and hence, from Figure 9 (left) in general more accurate) towards the center of the chemical space region, as expected as most machine learning approaches are more accurate in interpolation. We note in particular that the small clusters in the bottom-left of the embedding do not show generally higher uncertainty than the rest of the compounds, and so the model does have applicability to disparate chemical types that cluster separately within the overarching chemical space. This builds confidence in the applicability of the modelling technique across a wide range of different compounds and areas of chemical space.

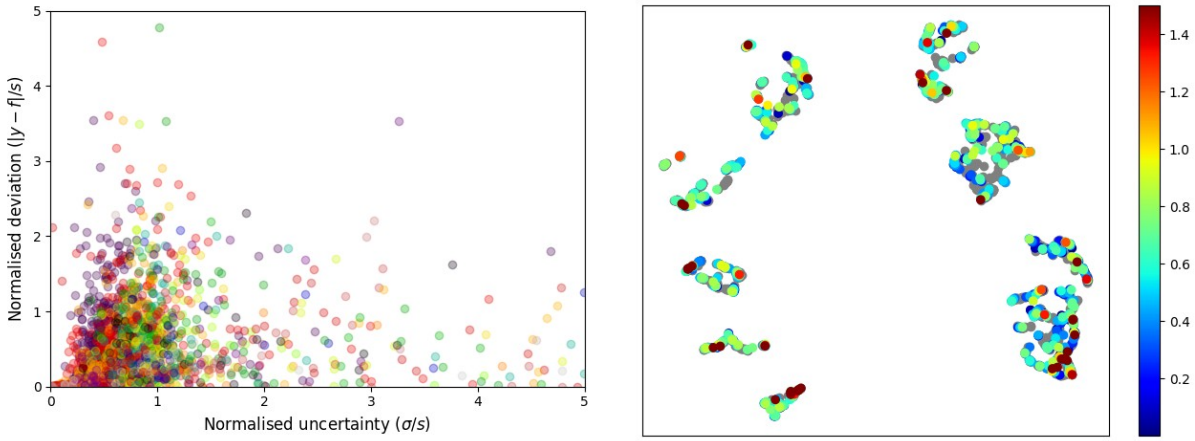



*Figure 9 the normalized absolute deviation in the 'CombinedAll' model predictions vs the uncertainty in those predictions (left); and a UMAP embedding of the data used, with each point colored by its average normalised uncertainty across all endpoints.*

To build confidence in the 'CombinedAll' imputation model it is important to understand how the model makes predictions and what information it relies on. In Figure 10 we highlight 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. We observe that the model is mostly using the variables coming from the MolDesc, CompToxPhysChem, ToolBoxPhysChem, and Human Health Hazard variables. All except the

MolDesc data is sparse, meaning that it could not be used directly with conventional QSAR methods. The MolDesc, CompToxPhysChem, and ToolBoxPhysChem data are all based on computational tools directly designed to capture chemically-informative properties of compounds, and this analysis indicates that this information is prioritized for modelling over the fingerprint data from the MorganFP or ToxPrintFP when those are also available (although Figure 2 indicates that the modelling approach is able to extract comparable information from the fingerprint data when only that is available). We do note however that the higher cardinality of the continuous MolDesc and physico-chemical data over the binary fingerprint data might bias the understanding of the most informative variables (Strobl et al., 2007)

The benefit of the imputation approach is shown by the fact that several of the Human Health Hazard endpoints are strongly used by the model: of the top 10 most used inputs, six are Human Health Hazards, including the top three being chromosome aberration endpoints based on different model species.

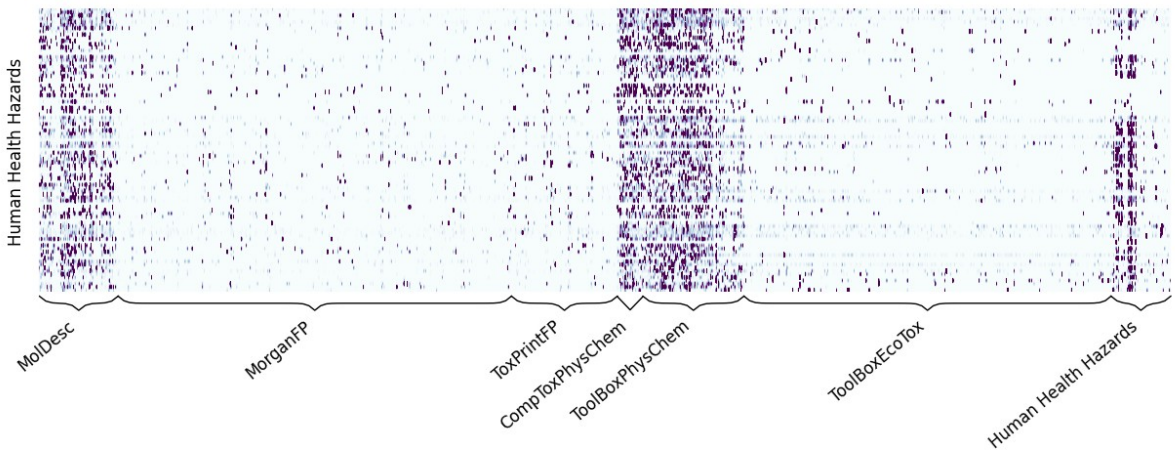



*Figure 10 A table showing which variables are used as input to the 'CombinedAll' model to predict each of the Human Health Hazard target endpoints, labelled by which dataset they came from. Input variables that were completely unused are dropped for visual clarity.*

# Conclusions

Imputation models incorporate sparse data so significantly improve QSAR modeling accuracy, demonstrated here for toxicity endpoints. Imputation models also save time, as they do not require a perfectly curated dataset and can autonomously handle irrelevant data. On the other hand, QSAR models require carefully curated data to achieve the best performance. These methods could be game changing for NAMs approaches where endpoints often depend on multiple data sources leading to a sparse database.

However, worldwide regulatory agencies will need to update their guidelines to accommodate this new family of models. When evaluating imputation modeling against the five principles established by the OECD to validate QSAR models several misalignments become evident. Firstly, imputation models could potentially be utilized for multiple endpoints, whereas the OECD recommends associating QSAR models with a defined endpoint. Additionally, the OECD advises QSAR models are developed from homogeneous datasets generated through a single protocol, whereas imputation models integrate data from various experiments conducted using different protocols.

Furthermore, the combination of large datasets from diverse sources complicates the establishment of a strict applicability domain. Finally, the complex arrangement and diverse utilization of data makes it difficult for imputation models to provide a mechanistic interpretation. Further work needs to be conducted to develop and clarify the principles needed to validate imputation models, which have the promise to serve as more relevant NAMs in the ever-increasing complexity of toxicological testing.

AUTHOR INFORMATION


### Corresponding Author

*tom@intellegens.co.uk


### Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript. ‡These authors contributed equally. (match statement to author names with a symbol)


### Funding Sources

Any funds used to support the research of the manuscript should be placed here (per journal style).


### Notes

Any additional relevant notes should be placed here.


ACKNOWLEDGMENT

(Word Style “TD_Acknowledgments”). Generally the last paragraph of the paper is the place to acknowledge people, organizations, and financing (you may state grant numbers and sponsors here). Follow the journal’s guidelines on what to include in the Acknowledgments section.


ABBREVIATIONS

QSAR, quantitative structure-activity relationship; $R^2$, coefficient of determination; MCC, Matthews correlation coefficient

# References

DiMaggio, P. A., Subramani, A., Judson, R. S., & Floudas, C. A. (2010). A Novel Framework for Predicting In Vivo Toxicities from In Vitro Data Using Optimal Methods for Dense and Sparse Matrix Reordering and Logistic Regression. *Toxicological Sciences*, *118*(1), 251–265. https://doi.org/10.1093/toxsci/kfq233

EU Commision. (2012). *Commission Implementing Regulation (EU) No 872/2012 of 1 October 2012 adopting the list of flavouring substances provided for by Regulation (EC) No 2232/96 of the European Parliament and of the Council, introducing it in Annex I to Regulation (EC) No 1334*. EUR-Lex. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32012R0872

Gwinn, M., Cowden, J., Lambert, J., Lowit, A., Wetmore, B., Scarano, L., Schappelle, S., Thomas, R., & Burman, E. (2020). *EPA Alternative Toxicity Testing Report to Congress*. 0 Bytes. https://doi.org/10.23645/EPACOMPTOX.12283958.V4

Irwin, B. W. J., Levell, J. R., Whitehead, T. M., Segall, M. D., & Conduit, G. J. (2020). Practical Applications of Deep Learning To Impute Heterogeneous Drug Discovery Data. *Journal of Chemical Information and Modeling*, *60*(6), 2848–2857. https://doi.org/10.1021/acs.jcim.0c00443

Kensert, A., Alvarsson, J., Norinder, U., & Spjuth, O. (2018). Evaluating parameters for ligand-based modeling with random forest on sparse data sets. *Journal of Cheminformatics*, *10*(1), 49. https://doi.org/10.1186/s13321-018-0304-9

Kovarich, S., Ceriani, L., Gatnik, M. F., Bassan, A., & Pavan, M. (2019). Filling Data Gaps by Read-across: A Mini Review on its Application, Developments and Challenges. *Molecular Informatics*, *38*(1800121). https://doi.org/10.1002/minf.201800121

Kwon, S., Bae, H., Jo, J., & Yoon, S. (2019). Comprehensive ensemble in QSAR prediction for drug discovery. *BMC Bioinformatics*, *20*(1), 521. https://doi.org/10.1186/s12859-019-3135-4

Mahmoud, S., Irwin, B., Chekmarev, D., Vyas, S., Kattas, J., Whitehead, T., Mansley, T., Bikker, J., Conduit, G., & Segall, M. (2021). Imputation of sensory properties using deep learning. *Journal of Computer-Aided Molecular Design*, *35*(11), 1125–1140. https://doi.org/10.1007/s10822-021-00424-3

McInnes, L., Healy, J., & Melville, J. (2020). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*.

Norinder, U., Carlsson, L., Boyer, S., & Eklund, M. (2014). Introducing Conformal Prediction in Predictive Modeling. A Transparent and Flexible Alternative to Applicability Domain Determination. *Journal of Chemical Information and Modeling*, *54*(6), 1596–1603. https://doi.org/10.1021/ci5001168

OECD. (2014). *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models*. OECD. https://doi.org/10.1787/9789264085442-en

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pradeep, P., Carlson, L. M., Judson, R., Lehmann, G. M., & Patlewicz, G. (2019). Integrating data gap filling techniques: A case study predicting TEFs for neurotoxicity TEQs to facilitate the hazard assessment of polychlorinated biphenyls. *Regulatory Toxicology and Pharmacology*, *101*, 12–23. https://doi.org/10.1016/j.yrtph.2018.10.013

*QSAR Toolbox*. (n.d.). Retrieved March 24, 2022, from https://qsartoolbox.org/

Simm, J., Arany, A., Zakeri, P., Haber, T., Wegner, J. K., Chupakhin, V., Ceulemans, H., & Moreau, Y. (2015). *Macau: Scalable Bayesian Multi-relational Factorization with Side Information using MCMC*. https://doi.org/10.48550/ARXIV.1509.04610

Simões, R. S., Maltarollo, V. G., Oliveira, P. R., & Honorio, K. M. (2018). Transfer and Multi-task Learning in QSAR Modeling: Advances and Challenges. *Frontiers in Pharmacology*, *9*, 74. https://doi.org/10.3389/fphar.2018.00074

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 25. https://doi.org/10.1186/1471-2105-8-25

Sun, X., Zhu, J., Chen, B., You, H., & Xu, H. (2022). A feature transferring workflow between data-poor compounds in various tasks. *PLOS ONE*, *17*(3), e0266088. https://doi.org/10.1371/journal.pone.0266088

Tse, E. G., Aithani, L., Anderson, M., Cardoso-Silva, J., Cincilla, G., Conduit, G. J., Galushka, M., Guan, D., Hallyburton, I., Irwin, B. W. J., Kirk, K., Lehane, A. M., Lindblom, J. C. R., Lui, R., Matthews, S., McCulloch, J., Motion, A., Ng, H. L., Öeren, M., … Todd, M. H. (2021). An Open Drug Discovery Competition: Experimental Validation of Predictive Models in a Series of Novel Antimalarials. *Journal of Medicinal Chemistry*, *64*(22), 16450–16463. https://doi.org/10.1021/acs.jmedchem.1c00313

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … Contributors, S. 1. 0. (2020). SciPy 1.0: Fundamental algorithms for scientific

computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Walter, M., Allen, L. N., de la Vega de León, A., Webb, S. J., & Gillet, V. J. (2022). Analysis of the benefits of imputation models over traditional QSAR models for toxicity prediction. *Journal of Cheminformatics*, *14*(1), 32. https://doi.org/10.1186/s13321-022-00611-w

Westmoreland, C., Bender, H. J., Doe, J. E., Jacobs, M. N., Kass, G. E. N., Madia, F., Mahony, C., Manou, I., Maxwell, G., Prieto, P., Roggeband, R., Sobanski, T., Schütte, K., Worth, A. P., Zvonar, Z., & Cronin, M. T. D. (2022). Use of New Approach Methodologies (NAMs) in regulatory decisions for chemical safety: Report from an EPAA Deep Dive Workshop. *Regulatory Toxicology and Pharmacology*, *135*, 105261. https://doi.org/10.1016/j.yrtph.2022.105261

Whitehead, T. M., Irwin, B. W. J., Hunt, P., Segall, M. D., & Conduit, G. J. (2019). Imputation of Assay Bioactivity Data Using Deep Learning. *Journal of Chemical Information and Modeling*, *59*(3), 1197–1204. https://doi.org/10.1021/acs.jcim.8b00768

Williams, A. J., Grulke, C. M., Edwards, J., McEachran, A. D., Mansouri, K., Baker, N. C., Patlewicz, G., Shah, I., Wambaugh, J. F., Judson, R. S., & Richard, A. M. (2017). The CompTox Chemistry Dashboard: A community data resource for environmental chemistry. *Journal of Cheminformatics*, *9*(1), 61. https://doi.org/10.1186/s13321-017-0247-6

Yang, C., Tarkhov, A., Marusczyk, J., Bienfait, B., Gasteiger, J., Kleinoeder, T., Magdziarz, T., Sacher, O., Schwab, C. H., Schwoebel, J., Terfloth, L., Arvidson, K., Richard, A., Worth, A., & Rathman, J. (2015). New Publicly Available Chemical Query Language, CSRML, To Support Chemotype Representations for Application to Data Mining and Modeling.

*Journal of Chemical Information and Modeling*, *55*(3), 510–528.

https://doi.org/10.1021/ci500667v