Imputation of *in vitro* toxicity data

Moritz Walter*, Antonio de la Vega de León*, Val Gillet*, Sam Webb†.
* Information School, University of Sheffield, Regent Court, 211 Portobello, Sheffield S1 4DP, UK.
† Lhasa Limited, Granary Wharf House, 2 Canal Wharf, Leeds, LS11 5PS, UK.

Introduction

Multitarget toxicity datasets are typically sparse, which means that not every compound was measured in every assay. Imputation describes the process of predicting missing values in a sparse dataset. In contrast to standard QSAR models, which are based on relations between chemical descriptors and toxicities, imputation models leverage relations between different assays to make predictions. In the present study it is investigated how imputation models compare to standard QSAR models.

Splitting Strategy

20% of toxicity labels for each assay are randomly removed for training the models and held back to evaluate their performance (test labels). The white cells of the table are filled, but no evaluation is possible, as the labels are unknown.

	A1	A2	A3	A4
C1	1	-	-	0
C2	-	1	-	1
C3	0	1	0	-
C4	-	0	1	0
C5	-	-	1	0

→

	A1	A2	A3	A4
C1	1	-	-	0
C2	-	-	-	1
C3	-	1	0	-
C4	-	0	1	-
C5	-	-	-	0

→

	A1	A2	A3	A4
C1	1	1	0	0
C2	0	0	0	1
C3	0	1	0	0
C4	0	0	1	0
C5	1	0	1	0

Train label Test label C: compound A: assay

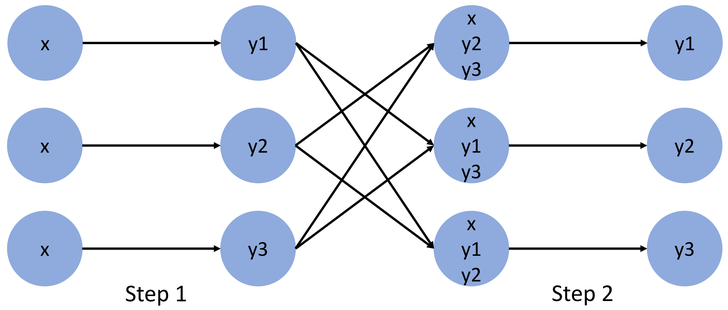
Datasets

	Ames[1]	Tox21[2]
Assay number	12	12
Assay types	6 *Salmonella typhimurium* strains ± S9-mix	7 nuclear receptors and 5 stress response pathways
Unique Compounds	6168	8090
Data density / range	40.5% / 12.4-75.0%	83.4% / 74.2-92.2%
Actives % / range	21.3% / 11.4-31.8%	6.9% / 3.0-15.6%

Imputation Techniques

1. Feature Net

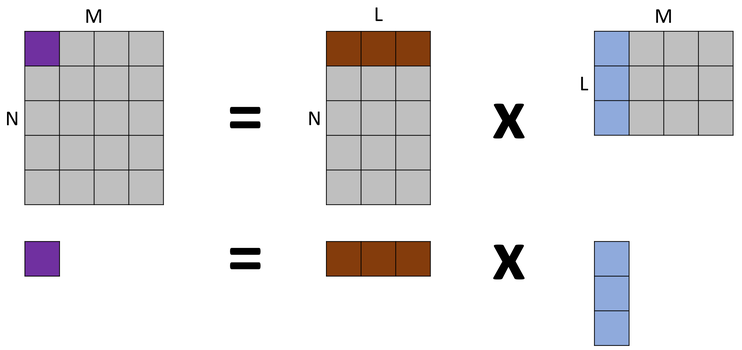
Feature Net[3] models represent a way to combine single models and can be implemented based on any supervised machine learning algorithm. Step 1: a single QSAR model is trained for each assay. Step 2: the models are re-trained using the assay labels (either predicted in Step 1 or experimentally measured) as additional features.



x: chemical descriptor, y_i: assay label

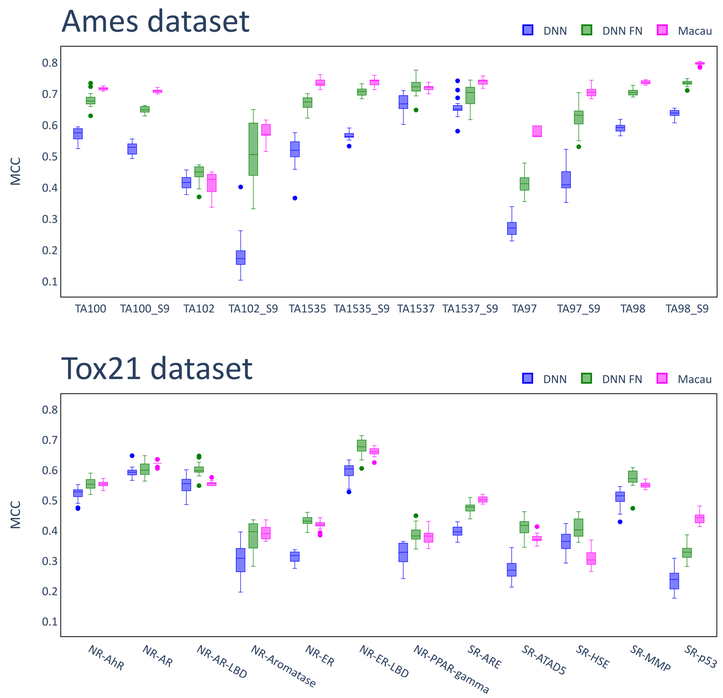
2. Macau

Macau[4] is a matrix factorization technique. The activity matrix with **M** assays and **N** compounds is modelled as the product of two smaller matrices, representing compounds and assays mapped to a **L**-dimensional latent space. The prediction of the violet cell is given by the dot product between the vectors representing the first compound (brown) and the first assay (blue). Macau allows the inclusion of side information (e.g. chemical fingerprints to describe compounds) to improve the model.



Results

The performance evaluated as MCC score of the imputation models Feature Net (based on Deep Neural Networks) and Macau are compared to Deep Neural Networks as a standard QSAR model (descriptor: ECFP4) on the Ames and Tox21 datasets. Each box sums up the results of 20 independent runs using different random seeds on the same test set.



Conclusion

These findings demonstrate that imputation approaches may provide a benefit over single task QSAR models for predicting toxicity of compounds, when data for related toxicity assays is available. Hence, imputation represents an attractive alternative to conducting additional tests for evaluating toxicity of compounds.

References

1. Benigni, R. et al. (2013) New perspectives in toxicological information management, and the role of ISSTOX databases in assessing chemical mutagenicity and carcinogenicity. *Mutagenesis*, 28(4), 401-409.
2. https://tripod.nih.gov/tox21/challenge/; downloaded: 21/11/2019.
3. Varnek, A. et al. (2009) Inductive transfer of knowledge: Application of multi-task learning and Feature Net approaches to model tissue-air partition coefficients. *J. Chem. Inf. Model.*, 49(1), 133-144.
4. Simm, J. et al. (2015) Macau: scalable Bayesian multi-relational factorization with side information using MCMC. *Arxiv:1509.04610v2*.

email: mwalter4@sheffield.ac.uk