Jiahao Yu[1,2] and David Marcus[1]
[1] *Data & Predictive Sciences, GlaxoSmithKline, Stevenage, SG1 2NY, UK*
[2] *School of Mathematics, University of Bristol, Bristol, BS8 1UG, UK*

Reference: CLR-002690
Creation Date: 11/03/24

Filling the gaps: Data Imputation Methods for Drug Discovery

For more information:
jiahao.yu@bristol.ac.uk

Introduction

Drug discovery datasets are often shown as sparse, noisy, and heterogeneous. To facilitate drug discovery projects and to ensure the effectiveness of Machine Learning (ML) algorithms and predictive models, it is necessary to find methods to fill in the gaps in this data.

Classic QSAR methods use calculated descriptors from compounds to predict assay data, as illustrated in Figure 1. Data imputation utilizes the information from measured assay data, in addition to descriptors, to make inference on missing assay data, in a multi-task setting. Figure 2 demonstrates the principle of classic QSAR modelling and data imputation.

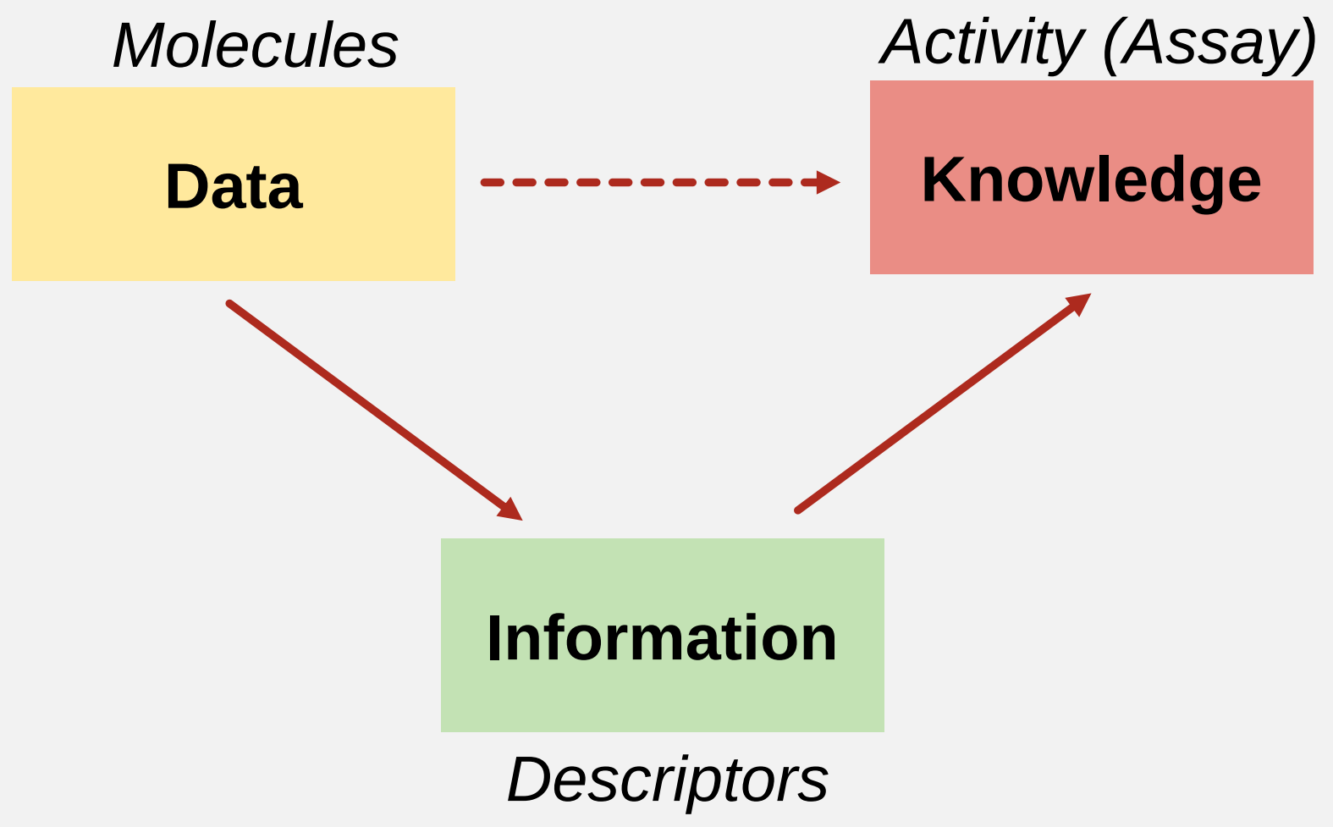

Figure 1: Classic QSAR Concept [1].

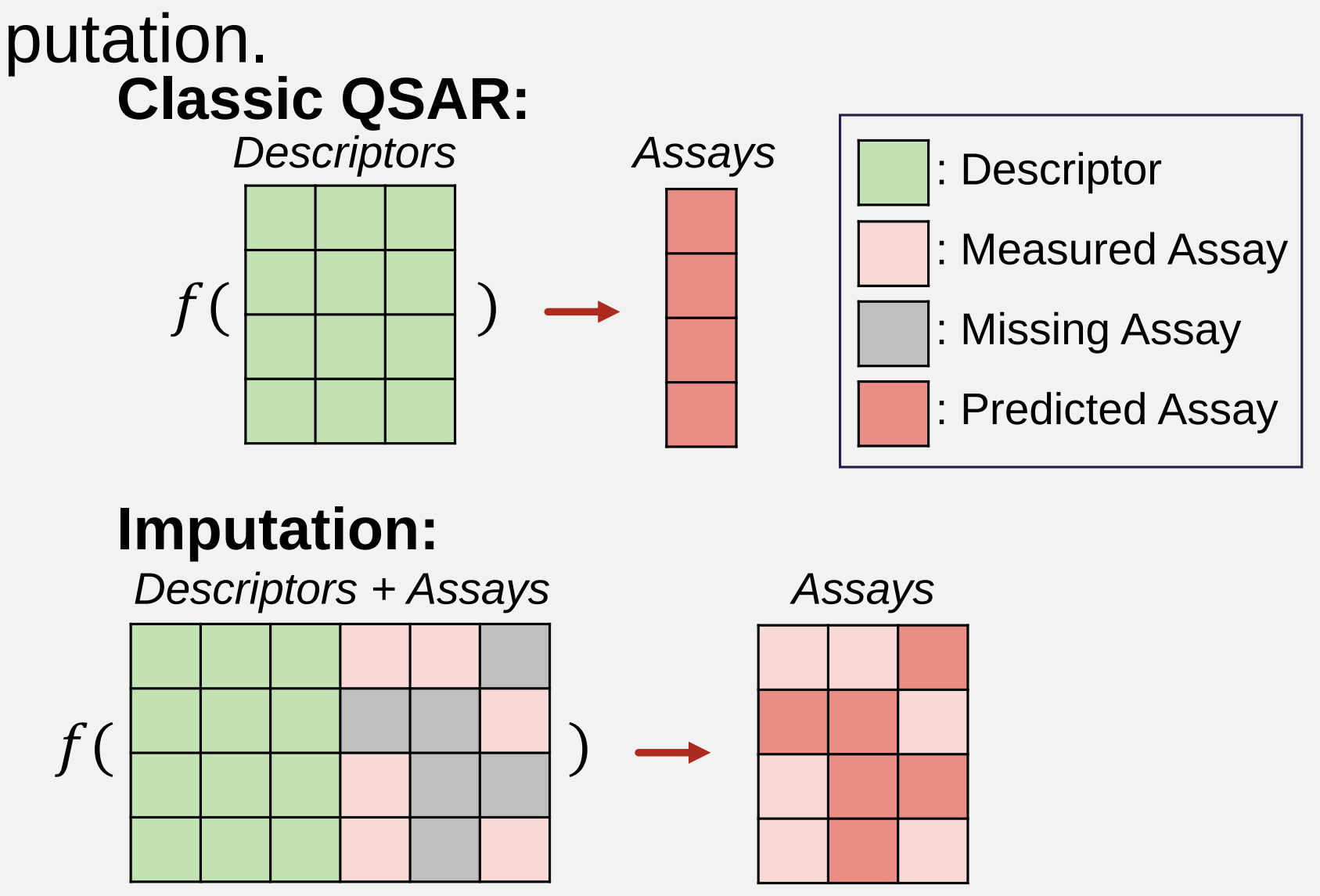

Figure 2: Illustration of classic QSAR and Imputation.

In this poster, we compare several classic and state-of-the-art methods for data imputation with classic QSAR modelling. We found that data imputation models can usually outperform classic QSAR models, however some are not suitable for data imputation in drug discovery, and some will require extensive calculation time.

Methods

Datasets

We used RDKit 2D properties and Morgan Fingerprints with radius 2 as *descriptors*. There are two types of assay data: *single* type of activity in multiple columns, and *multiple* types of activity. The data are split into training set (80%) and test set (20%). Before running experiments, all columns with zero variance are removed. The sizes of DMPK, Comp-Tox, Kinase, EXP and LD50 datasets are reduced. We summarize the datasets used in Table 1.

Problem Formulation

We model the performance of imputation models in *test sets*, in the following way, as shown in Figure 3:

- For each column of assays ($i \in \{1, 2, \cdots, \text{number of assay columns } n\}$):
  - remove data in that column ($A_i = \text{NaN}$).
  - impute all assay data, but save the imputed data of that column ($\widehat{A_i}$) only.
- Finally combine all imputed assays together ($\hat{A} = (\widehat{A_1}, \widehat{A_2}, \cdots, \widehat{A_n})$).

Selected Methods

We summarize ML methods utilized in Table 2. We experiment these methods in both classic QSAR and Imputation settings, in regression problems.

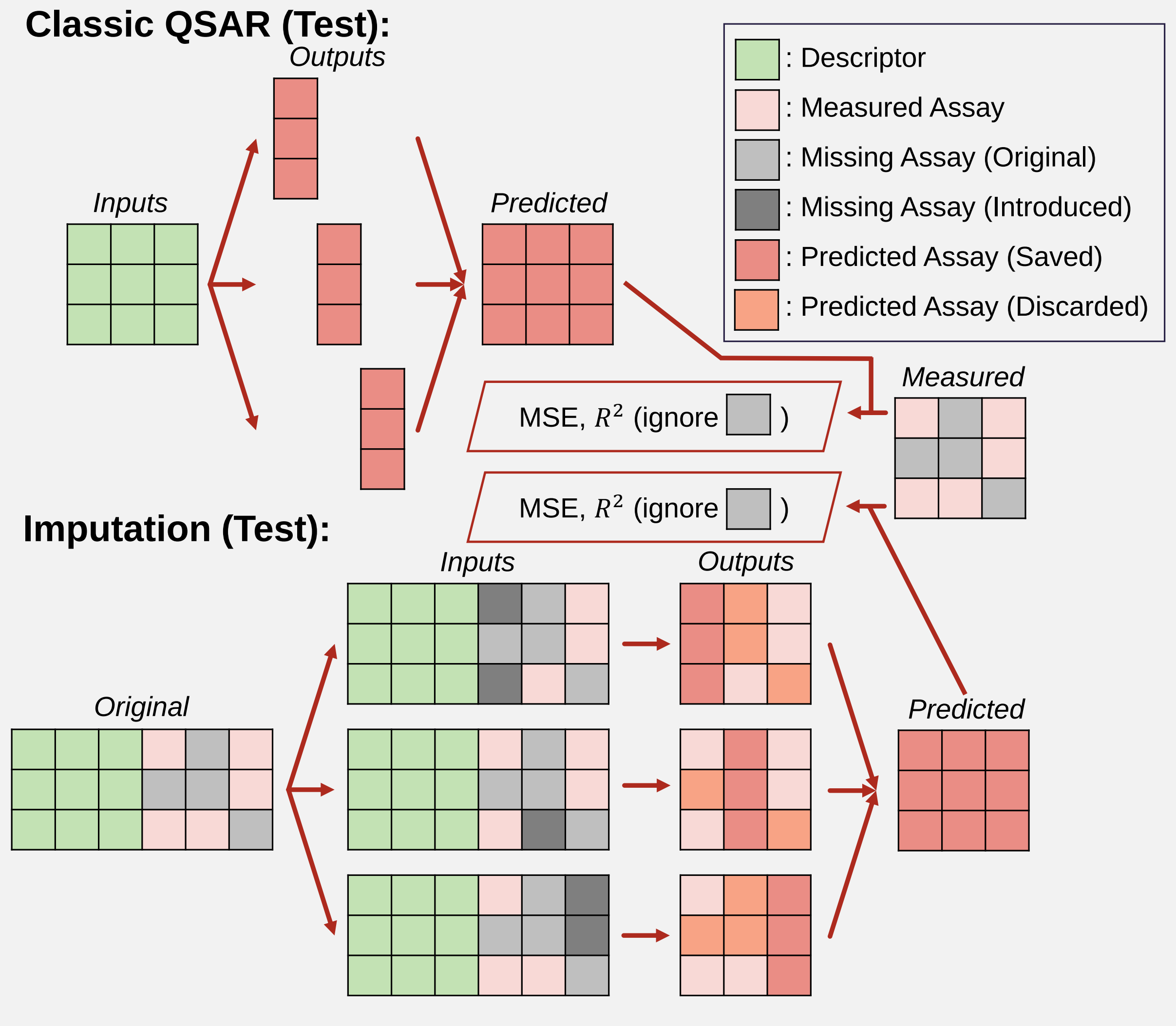

Figure 3: Illustration of classic QSAR and Imputation in test set.

Data	# Instances	# Descriptors	# Assays	Missing Rate+	Avg. abs. correlation*	Assay Type	Source
MMP-12 [2]	46	428	45	0.238	0.893	Single	GSK
DMPK	4280	2239	16	0.637	0.323	Multiple	GSK *(Proprietary)*
Comp-Tox	2154	2249	42	0.410	0.169	Multiple	GSK *(Proprietary)*
Kinase [3]	1007	2139	27	0.208	0.300	Multiple	ChEMBL
EXP	10122	2253	39	0.266	0.238	Multiple	GSK *(Proprietary)*
LD50 [4]	6396	2255	24	0.790	0.787	Multiple	ChemIDplus

+: Proportion of missing data in assays. Higher values associate with more missing assays.
*: Average of absolute values of correlation matrix of assays. Higher value represents higher correlation in assays.

Table 1: Summary of datasets.

Method	Base Method	Year	NN based?	Designed for imputation?	Uncertainty Estimation?
XGBoost	Gradient Boosting	2014	No	No	Yes
MLP	Perceptron	1958	Yes	No	Feasible
MICE [5]	Multiple Imputation	2011	No	Yes	No
pQSAR [6]	RF, PLS	2017	No	Yes	Feasible
GAIN [7]	GAN	2018	Yes	Yes	Feasible
MIDAS [8]	DAE	2022	Yes	Yes	Yes
Sinkhorn [9]	Optimal Transport	2020	Yes	Yes	Feasible
HyperImpute [10]	Model Selection	2022	Mixed	Yes	Feasible

Table 2: Summary of ML methods.

Results

We demonstrate the performance of ML methods in classic QSAR and Imputation manners in Figure 4-7. We use *Mean Square Error (MSE)* as metrics. They take the median of 2000 bootstrapped samples of normalized assays. Values with MSE > 5 are removed due to poor performance in either classic QSAR or Imputation model, or both.

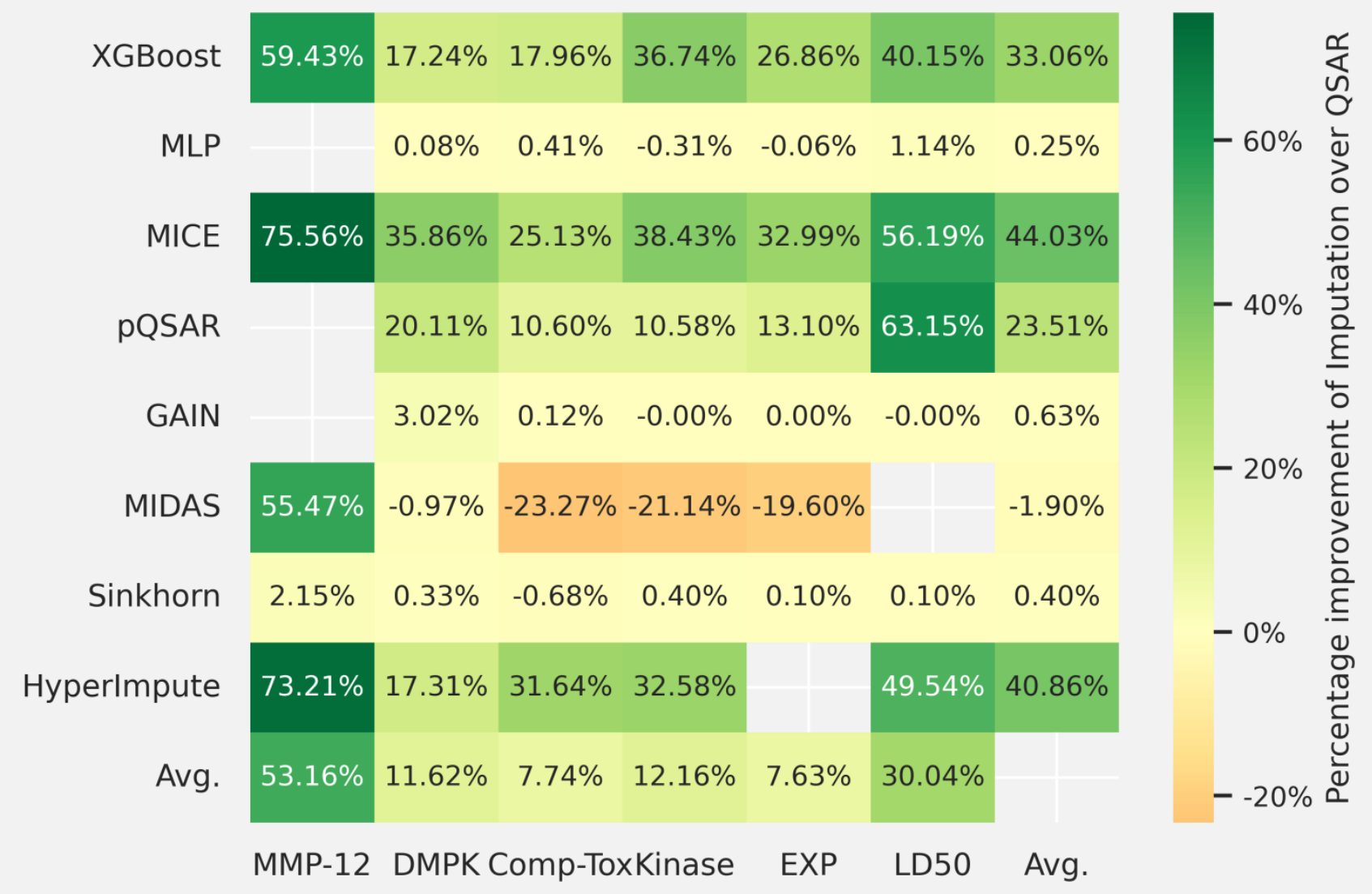

Figure 4: Percentage improvement of Imputation over QSAR.

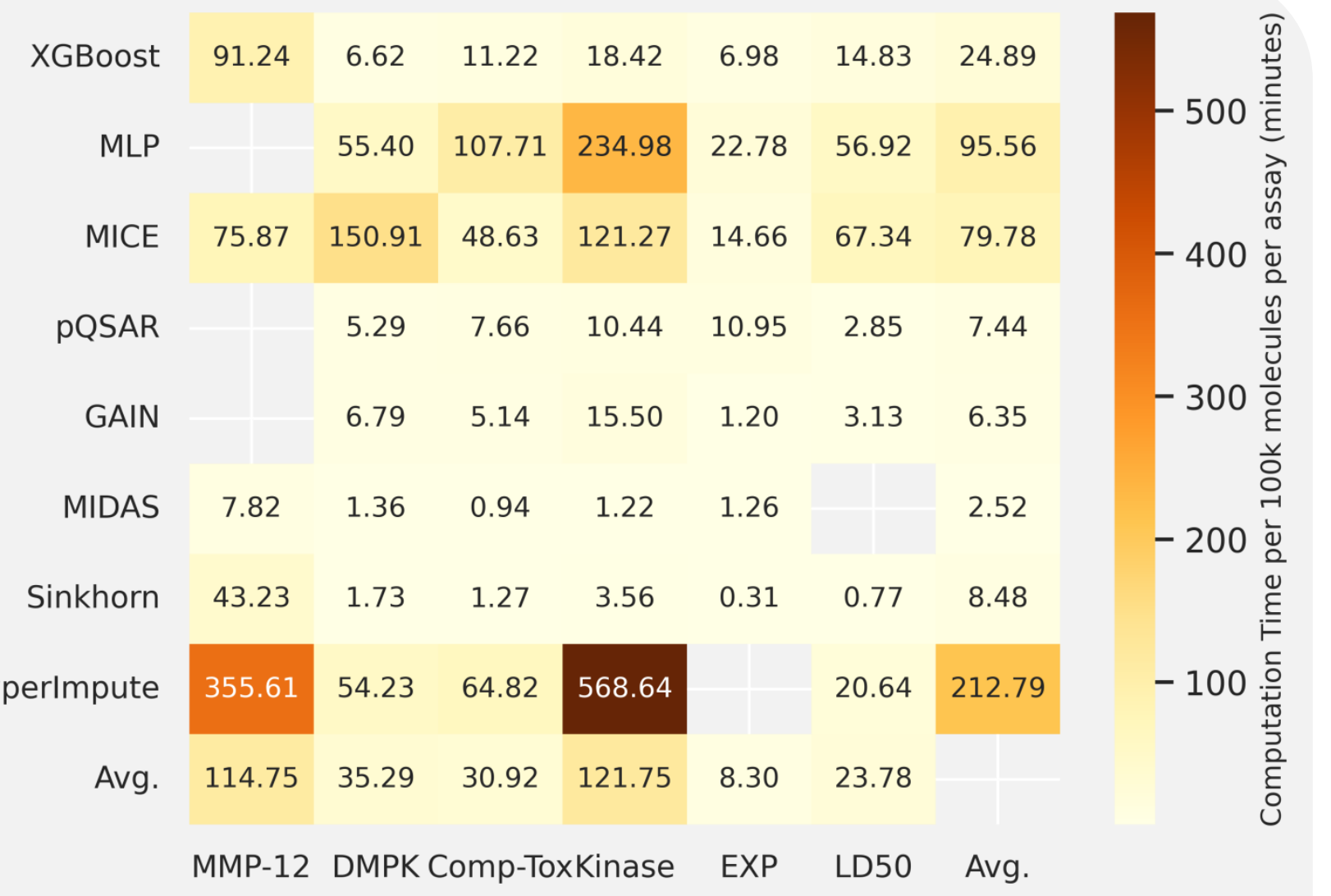

Figure 5: Computation time of Imputation.

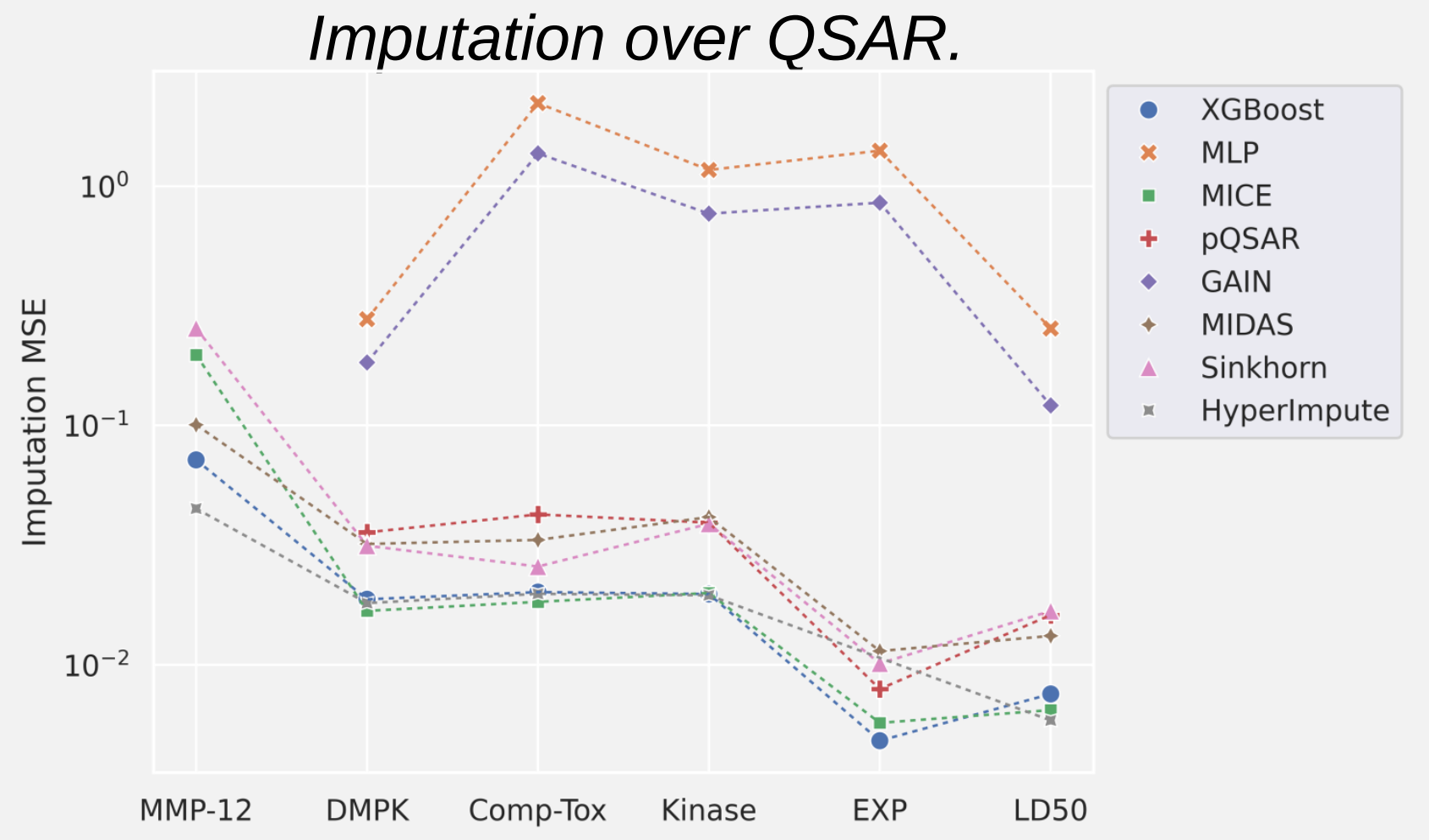

Figure 6: MSE of Imputation.

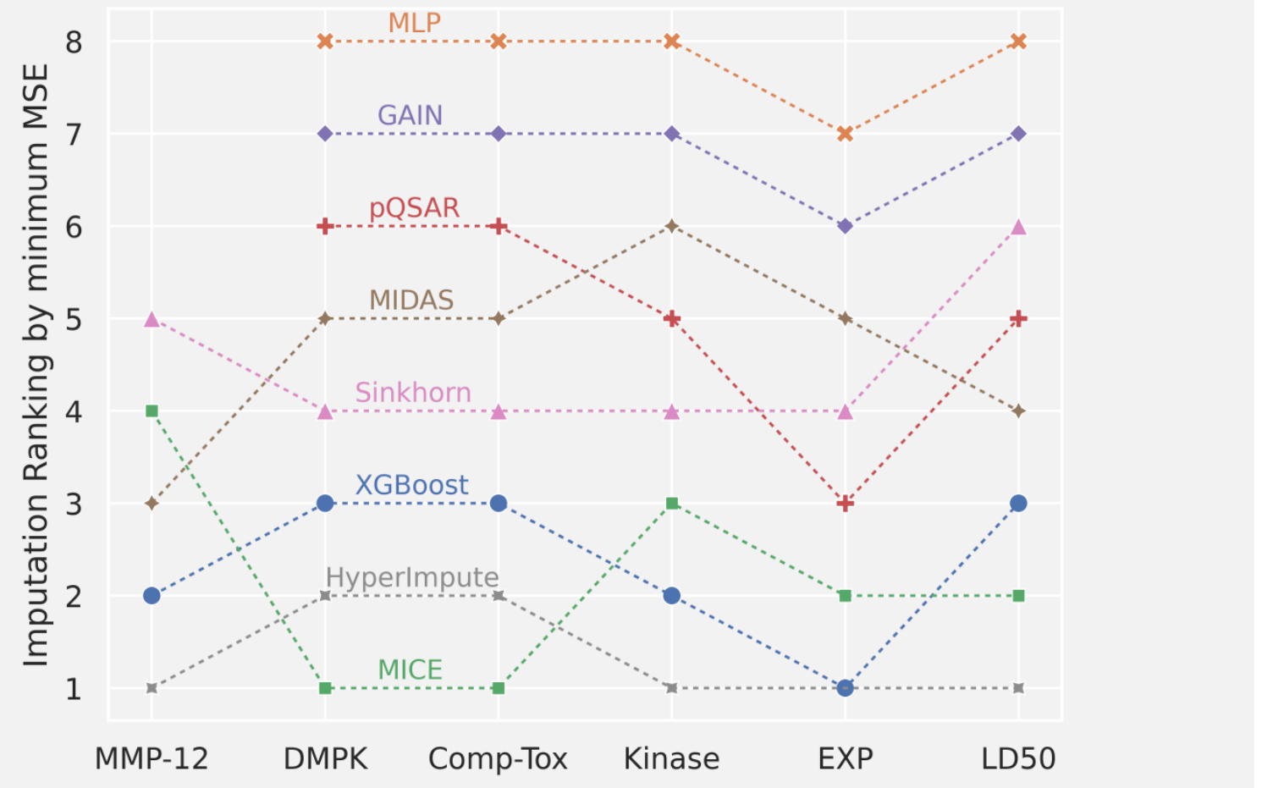

Figure 7: Ranking of Imputation by minimum MSE.

Discussion and Conclusions

- Imputation methods *outperform* classic QSAR methods in *most* cases, especially when the correlations in assays are high.
- Classic Shallow Learning ML methods *outperform* Deep Learning methods.
- Generative methods (GAIN, MIDAS) have been shown successful in other fields. Further research to integrate these to drug discovery is necessary.
- One reason for some imputation methods (e.g. GAIN) failing is that they assume MCAR scenario, which is rarely the only case in drug discovery.
- Careful structural design of NN-based models could improve the accuracy.
- Traditional statistical imputation method MICE and state-of-the-art model selection-based method HyperImpute are highly effective, but they come with high computational costs.
- Additional experiments on other types of drug discovery data are essential.
- Further research can also investigate the uncertainty of imputations.

References

(1) Gasteiger, J. and Engel, T. eds., 2006. *Chemoinformatics: a textbook*. John Wiley & Sons.
(2) Pickett, S. D.; Green, D. V.; Hunt, D. L.; Pardoe, D. A.; Hughes, I. Automated Lead Optimization of MMP-12 Inhibitors Using a Genetic Algorithm. *ACS Medicinal Chemistry Letters* **2011**, *2* (1), 28–33.
(3) Zdrazil, B.; Felix, E.; Hunter, F.; Manners, E. J.; Blackshaw, J.; Corbett, S.; de Veij, M.; Ioannidis, H.; Lopez, D. M.; Mosquera, J. F.; others. The ChEMBL Database in 2023: A Drug Discovery Platform Spanning Multiple Bioactivity Data Types and Time Periods. *Nucleic acids re.* **2024**, *52* (D1), D1180–D1192.
(4) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; others. PubChem 2023 Update. *Nucleic acids research* **2023**, *51* (D1), D1373–D1380.
(5) van Buuren, S.; Groothuis-Oudshoorn, K. Mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* **2011**, *45*, 1–67
(6) Martin, E. J.; Polyakov, V. R.; Tian, L.; Perez, R. C. Profile-QSAR 2.0: Kinase Virtual Screening Accuracy Comparable to Four-Concentration IC50s for Realistically Novel Compounds. *Journal of Chemical Information and Modeling* **2017**, *57* (8), 2077–2088.
(7) Yoon, J.; Jordon, J.; van der Schaar, M. GAIN: Missing Data Imputation Using Generative Adversarial Nets. In *Proceedings of the 35th International Conference on Machine Learning*; Dy, J., Krause, A., Eds.; Proceedings of Machine Learning Research; PMLR, 2018; Vol. 80, pp 5689–5698.
(8) Lall, R.; Robinson, T. The MIDAS Touch: Accurate and Scalable Missing-Data Imputation with Deep Learning. *Political Analysis* **2022**, *30* (2), 179–196.
(9) Muzellec, B.; Josse, J.; Boyer, C.; Cuturi, M. Missing Data Imputation Using Optimal Transport. In *Proceedings of the 37th International Conference on Machine Learning*; III, H. D., Singh, A., Eds.; Proceedings of Machine Learning Research; PMLR, 2020; Vol. 119, pp 7130–7140.
(10) Jarrett, D.; Cebere, B. C.; Liu, T.; Curth, A.; van der Schaar, M. HyperImpute: Generalized Iterative Imputation with Automatic Model Selection. In *Proceedings of the 39th International Conference on Machine Learning*; Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S., Eds.; Proceedings of Machine Learning Research; PMLR, 2022; Vol. 162, pp 9916–9937.

GSK

University of BRISTOL

Abbreviations

General Terms:
ML: Machine Learning
QSAR: Quantitative Structure-Activity Relationship
Avg. Average
Abs.: Absolute
MCAR: Missing Completely at Random

Datasets
MMP: Matrix Metalloproteinases
DMPK: Drug Metabolism and Pharmacokinetics
Comp-Tox: Computational Toxicology
EXP: Off-target Pharmacology Panel for generating alerts for early safety assessment using *in-vitro* biochemical and cellular assays
LD50: Median Lethal Dose

ML Methods:
NN: Neural Networks
MLP: Multilayer Perceptron
pQSAR: Profile-QSAR 2.0
RF: Random Forests
PLS: Partial Least Squares
MICE: Multivariate Imputation by Chained Equations
GAIN: Generative Adversarial Imputation Nets
GAN: Generative Adversarial Nets
DAE: Denoising Autoencoders
MIDAS: Multiple Imputation with Denoising Autoencoders

Metrics:
MSE: Mean Square Error

Ahead Together