Future
Drug Discovery

Special Report

For reprint orders, please contact: reprints@future-science.com

Imputation versus prediction: applications in machine learning for drug discovery

Benedict W J Irwin*,[1,2], Samar Mahmoud[1], Thomas M Whitehead[3], Gareth J Conduit[2,3] & Matthew D Segall[1]

[1] Optibrium Limited, Cambridge, CB25 9PB, UK
[2] Cavendish Laboratory, University of Cambridge, Cambridge, CB3 0HE, UK
[3] Intellegens Limited, Cambridge, CB4 3AZ, UK
*Author for correspondence: ben.irwin@optibrium.com

Imputation is a powerful statistical method that is distinct from the predictive modelling techniques more commonly used in drug discovery. Imputation uses sparse experimental data in an incomplete dataset to predict missing values by leveraging correlations between experimental assays. This contrasts with quantitative structure–activity relationship methods that use only descriptor – assay correlations. We summarize three recent imputation strategies – heterogeneous deep imputation, assay profile methods and matrix factorization – and compare these with quantitative structure–activity relationship methods, including deep learning, in drug discovery settings. We comment on the value added by imputation methods when used in an ongoing project and find that imputation produces stronger models, earlier in the project, over activity and absorption, distribution, metabolism and elimination end points.

Lay abstract: Imputation is the process of filling in the gaps in a dataset, where values have not yet been measured, using the limited data that are already present. This may appear simple when only a few values are missing but is challenging when more than half, or as much as 99% of the data are absent. Data are a precious commodity for model building and using imputation enables greater value to be extracted from the limited data available. This is particularly important in drug discovery, where the data are especially sparse and noisy. The recent application of deep learning methods for data imputation has led to significant advances in predictive power and unlocked hidden potential in drug discovery projects. In this paper, we will discuss data imputation, compare it with other modeling methods and describe some example applications in drug discovery.

First draft submitted: 25 February 2020; Accepted for publication: 2 March 2020; Published online: 6 April 2020

'Imputation' is a less-familiar term within statistics or machine learning in comparison with words such as 'regression', 'prediction' and 'classification'. In drug discovery, the latter three terms are commonly associated with quantitative structure–activity relationship (QSAR) models, which attempt to describe a property of interest as a mathematical function of some number of chemical 'descriptors' as inputs, to estimate the value of the property for an as-yet unmeasured compound. This process is portrayed in the top half of Figure 1, where the properties of interest could be biochemical or cell-based assay results, such as activities against protein targets, absorption, distribution, metabolism or elimination (ADME) end points, or physicochemical properties such as solubility.

In contrast, imputation is the process of filling in the gaps in a dataset, where values have not yet been measured, using the limited data that are already present. This process can be simple if only a handful of data points are missing but is very challenging when more than half, or even 99% of the data are absent in the first instance. Imputation is depicted in the bottom half of Figure 1, where not only descriptors are used as input to the model, but also any of the sparse data points that have already measured in the assays of interest.

Data are precious commodities for model building, and the imputation process of expanding the data that are available dramatically improves the quality of models. This not only captures correlation between descriptors and assays, but also explicitly uses the experimental data as an input, which enables the model to learn directly from

newlands
press

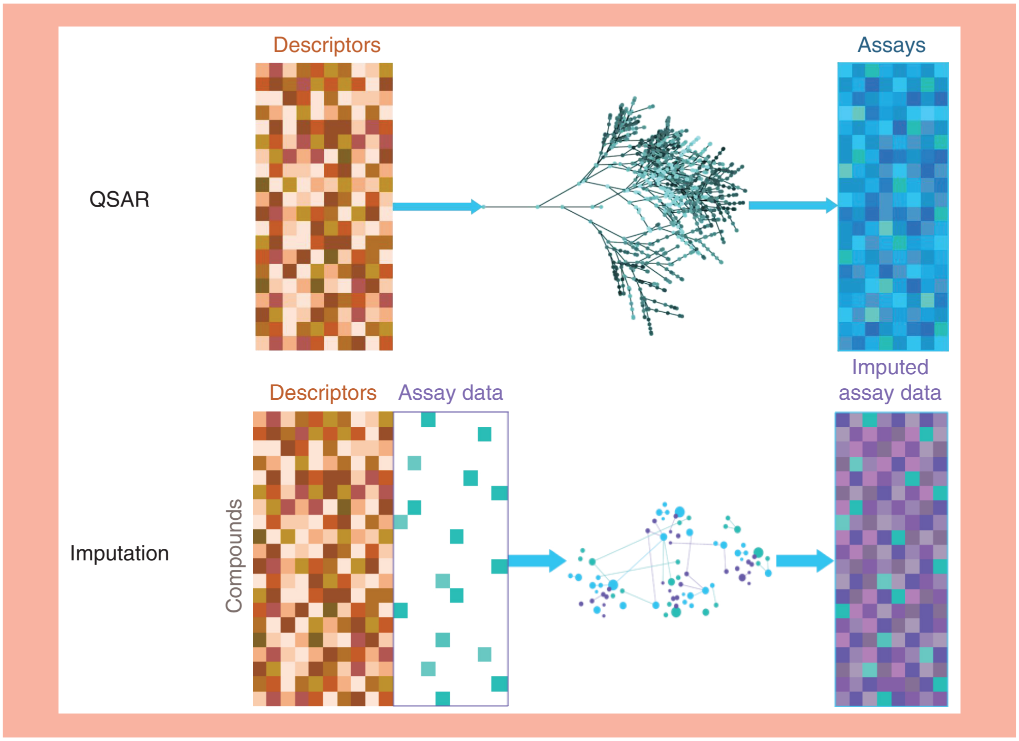

Figure 1. A comparison of quantitative structure activity regression (top) for example a random forest model, with imputation (bottom) for example the Alchemite™ deep imputation software.
QSAR: Quantitative structure–activity relationship.

assay–assay correlations [2,3]. This is particularly true in drug discovery, where the data are particularly sparse and noisy [3]. Imputation models also provide the flexibility to use newly available experimental data to strengthen predictions at any given time; as soon as a new data point is measured it can be input into the already-trained model and used to enhance the quality of further predictions for unmeasured points. Furthermore, an important requirement for successful application of models of any type is to identify when these values may be used with confidence. The recent application of deep learning methods to data imputation, along with robust uncertainty estimates, has addressed the challenge posed by spare and noisy data, led to advances in predictive power and unlocked hidden potential in drug discovery projects, earlier than equivalent QSAR models [3].

Existing strategies for imputation

There are at least three noteworthy methods for imputation in drug discovery. These create interesting comparison and are summarized below:

Heterogeneous deep imputation

Alchemite™ (Intellegens, Cambridge, UK) is a novel method which combines the power of deep learning with an imputation framework and can exploit nonlinear correlations between heterogeneous assay end points [2,3]. Sparse experimental data are input to the algorithm along with compound descriptors of any type. An initial best guess of the missing values is used, the full matrix is then reinserted into the algorithm and this process is repeated until the algorithm converges [4]. This method is one of the few that provide an uncertainty estimate for outputs by using an ensemble of predictions in the form of a distribution. This allows one to focus on the most confident and reliable results [2,3] and avoid missed opportunities [5].

Assay profile methods

Profile QSAR (pQSAR) is a two-step imputation method formed from combination of commonly used QSAR approaches. pQSAR first builds a layer of random forest models [1], then subsequently uses a linear model to look for assay–assay correlations from the individual predictions of the initial models [6]. So far, this method has focused

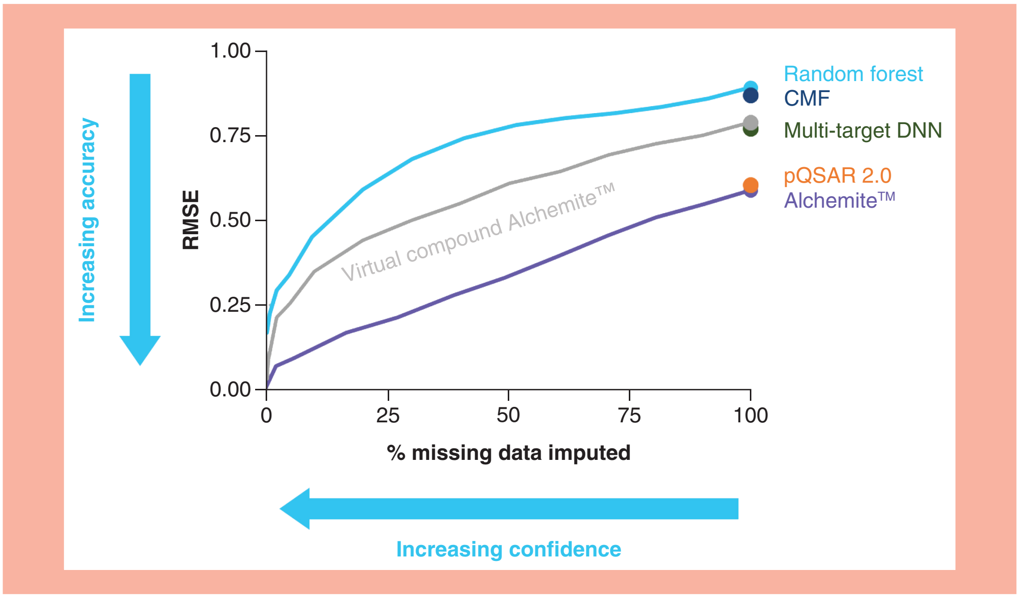

Figure 2. Performance in root-mean-square error measured on the independent test set of the Novartis benchmark ‘realistic’ split for three imputation models, random forest and deep neural network prediction models, and the Alchemite™ virtual model (where Alchemite runs as a prediction model with no input experimental data). This illustrates improvement in prediction quality (reduction in root-mean-square error) as the methods which provide uncertainty estimates alongside predictions focus on the most confident fraction of the dataset to impute/predict. CMF: Collective matrix factorization; DNN: Deep neural network; pQSAR: Profile quantitative structure–activity relationship; RMSE: Root-mean-square error.

on homogeneous datasets comprised of activity from a class of proteins, for example kinases and GPCRs, to use the linear correlations between the activities in related proteins [7].

Bayesian matrix factorization

The Macau method combines Bayesian probabilistic inference with an implementation of matrix factorization [8]. This method also takes chemical descriptors as an input, the authors refer to this as ‘high dimensional side information’. This method can handle data sets with millions of compounds and tens of millions of data points. It is inherently a linear method which may struggle with virtual compound predictions due to empty rows and columns.

Figure 2 shows a comparison of imputation and QSAR methods on a benchmarking dataset published by Novartis (Basel, Switzerland) [6], which was designed to be realistic in that the test set required extrapolation from training data. This comparison builds on that of Whitehead *et al.* [2], including the latest version of the Alchemite deep imputation method [2,3], pQSAR 2.0 [7], collective matrix factorization [9], a precursor the Macau Bayesian matrix factorization method [8], and QSAR models using random forests [1] and a multi-target deep neural network using Google Brain’s TensorFlow software [10].

Figure 2 shows the advantages of choosing an imputation method. When considering the full test set, Alchemite and pQSAR 2.0 provide the most accurate results; the lowest root-mean-square error values, which are much lower than the other methods. For methods which provide uncertainty estimates on their predictions (Alchemite and random forest) we can focus on the predictions with the greatest confidence, resulting in a reduction in root-mean-square error for these results.

The combination of imputation and effective confidence estimates mean that scientist can choose an appropriate level of uncertainty and partially fill in the dataset with sufficiently high-quality predictions.

Virtual models

An important application of QSAR models is to make predictions for ‘virtual compounds’ which have not yet been synthesized and thus have no experimental information, based solely on molecular descriptors. Deep imputation

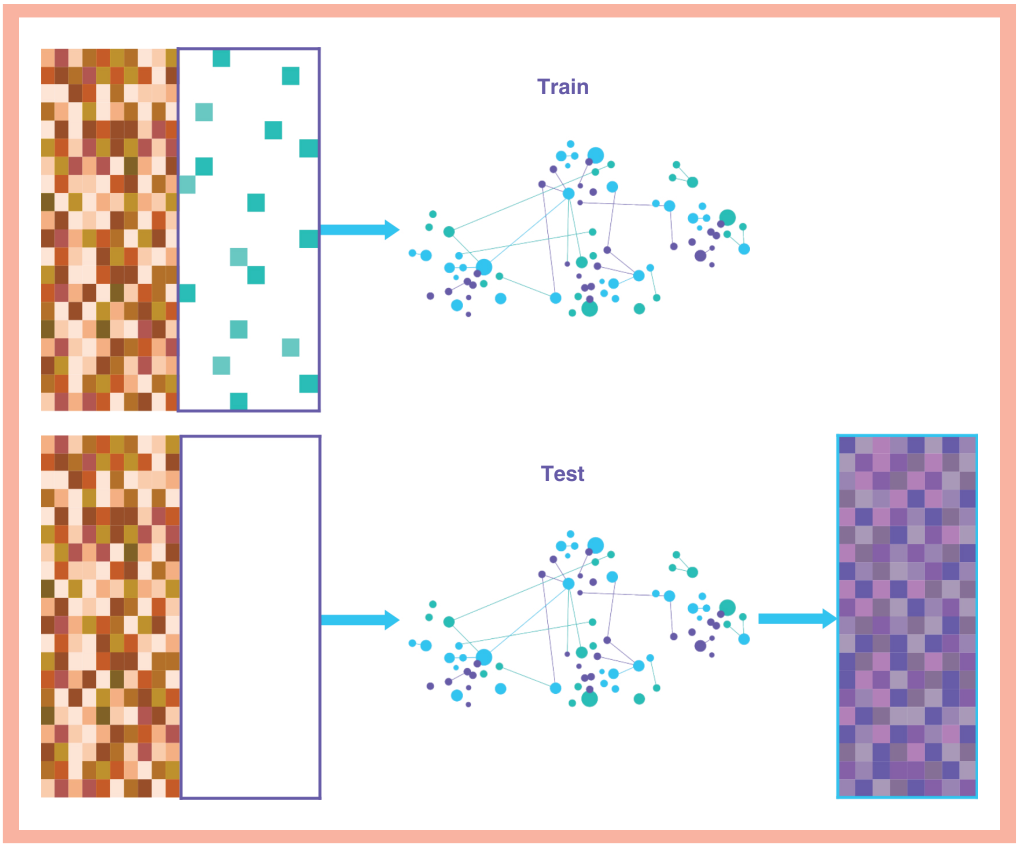

Figure 3. Example of the virtual model, compounds are trained on assay data from the training set (green squares) and no assay data are provided at test time to generate virtual predictions for all assays (purple squares).

models can easily be extended to this case by only inputting the compound descriptors for new compounds, as illustrated in Figure 3.

With no existing experimental data, the algorithm has the same information as a QSAR method at test time; however, the model can still exploit assay–assay correlations it learned in training and use these to enhance predictions. We would expect the model performance to therefore lie between a good QSAR method and the case of an imputation model with experimental information. Figure 2 shows this additional case labeled ‘Virtual compound Alchemite’. In the limit of predicting 100% of the data, the performance is comparable with a multi-target deep neural network, but the ability for Alchemite to produce an uncertainty estimate means the most confident predictions can be determined. This model clearly outperforms a random forest model with uncertainties derived from the ensemble of decision trees.

Practical application of imputation

There is an evidence that imputation adds value to drug discovery projects beyond the benchmarking data discussed above. An illustrative case study demonstrated successful application of deep imputation to heterogeneous drug discovery project data [3]. Remarkably, fewer data points than expected were needed to leverage the power of deep learning to assist in the prediction of actives. While standard deep learning models are notorious for requiring big data to be effective, the deep imputation models made confident predictions on typical project activity and ADME data – on the order of tens to hundreds of compounds [3]. The imputation models had an average coefficient of determination, $R^2 = 0.72$, where the best QSAR equivalents achieved $R^2 = 0.50$. It was also found that QSAR models completely failed to predict a cell-based activity end point, achieving a negative R^2, whereas the imputation method could make accurate predictions, $R^2 = 0.66$, by exploiting assay–assay correlations [3].

Further to that study, the value added as a project progresses is illustrated in Figure 4 by iteratively training Alchemite models on time-ordered project data, including both activity and ADME columns. Alchemite consistently

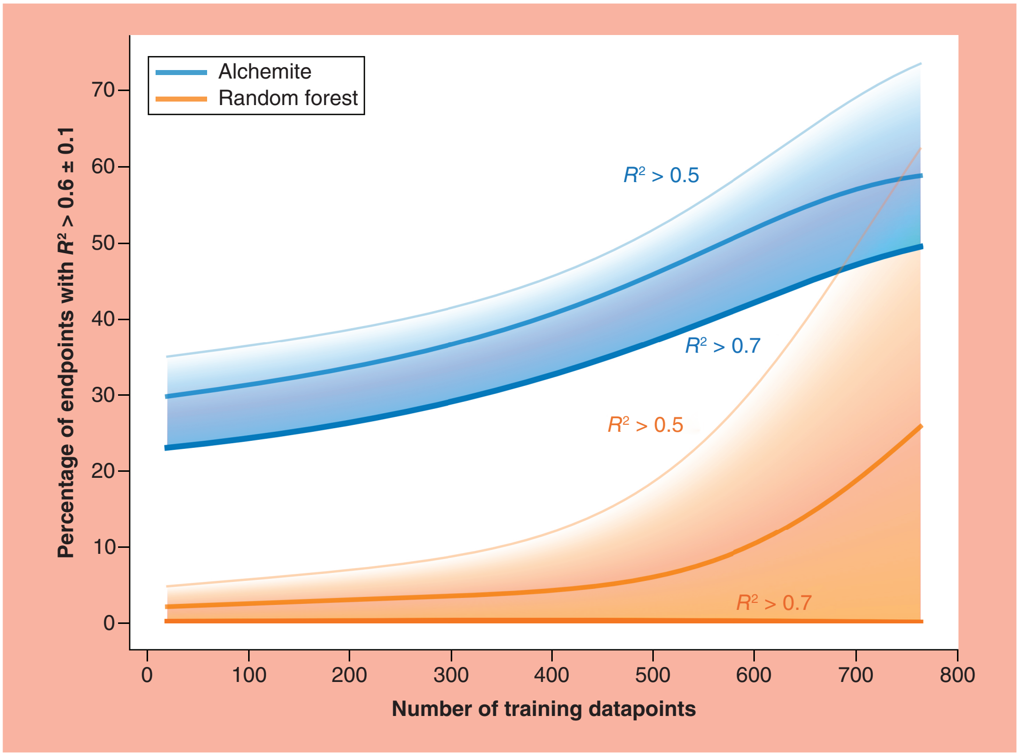

Figure 4. A performance metric across a project for a mixture of 15 drug discovery activities and absorption, distribution, metabolism or elimination end points. The percentage of end points with R^2 values above 0.5 (light), 0.6 and 0.7 (dark) is plotted against number of data points used to construct the model for both Alchemite™ deep imputation and random forest models.

outperforms random forest models. The percentage of Alchemite end points with model $R^2 > 0.6$ rises steadily, even for low numbers of training points, whereas random forest only shows a significant rise for more than 600 measured data points. On low quantities of data points, 20–30% of Alchemite end points have excellent models $R^2 > 0.7$, whereas random forest models cannot achieve this level of accuracy, even when using all of the data. Therefore, fewer data points are needed to build high quality imputation models, equating to fewer experiments, reduced cost and a shorter timeframe.

Conclusion

We have explained the concept of imputation and contrasted it with standard QSAR models. Imputation uses the known data explicitly to reconstruct the missing elements allowing assay–assay correlations and thereby gains more value from the investment in experimental measurements. One immediate and obvious advantage with imputation is the filling of the entire data matrix, which can provide as much as 100-times the original data.

Although a new method in the field of drug discovery, there is growing evidence that it can make more effective and efficient use of data in a variety of applications, to improve productivity and efficiency.

Future perspective

Imputation uses existing data resources more efficiently. An obvious application is for imputation to be used at the data source itself, whether that be a local dataset, corporate repository or large-scale data warehouse. Filling in the gaps in data sources with confident predictions will dramatically increase the wealth of information that can be mined to identify high-quality compounds. A combination of proprietary data combined with all known public domain data in a single imputed database will also give stronger predictive models for many types of end points in the chemical and biological domain.

Imputation can be used for combination of diverse types of sparse data, not only from experiments, but also computational simulations and physics-based calculations. Data can be aggregated across different resolutions and scales, with inexpensive and fast calculations assisting the prediction of moderately challenging end points which in turn assist the prediction of the hardest, slowest and most expensive data points to collect.

Any imputation method that can fulfill this would need to be proven to handle very sparse, noisy and heterogeneous data, scale to large sizes and remain robust to unrelated data sources being merged. The strongest method so far is the deep imputation method that provides uncertainty estimates in each imputed value. The flexibility of the deep learning framework will facilitate applications for novel kinds of data including time series, curve and distributional data, text data, aggregations of samples among populations and even graphical and network-based data sources.

For drug discovery, the combination of chemist and computational algorithm will enhance productivity, taking the burden of routine and intensive tasks away from the expert leaving them more freedom to be creative and focused on the human strategy elements of the design process. Imputation methods will take a leading role in achieving this goal as a foundation for further downstream AI and machine learning methods, which cannot handle incomplete data. Such algorithms could make additional statements about missing values beyond 'what number can I assign to this missing value?' and 'how confident are we in this prediction?'; we could ask questions such as 'how relevant is this missing value to my project?', 'which experiment should I perform next?' and 'can I trust this data point or should I remeasure it?'. Addressing these questions will ultimately assist a step change in efficiency and effective resource allocation.

Financial & competing interests disclosure

BWJ Irwin, S Mahmoud and MD Segall are employees of Optibrium Ltd. TM Whitehead and GJ Conduit are employees of Intellegens Ltd. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Open access

References

1. Breiman L. Random forests. *Mach. Learn.* 45, 5–32 (2001).
2. Whitehead TM, Irwin BWJ, Hunt P, Segall MD, Conduit GJ. Imputation of assay bioactivity data using deep learning. *J. Chem. Inf. Model.* 59(3), 1197–1204 (2019).
3. Irwin BWJ, Levell J, Whitehead T, Segall M, Conduit G. Practical applications of deep learning to impute drug discovery data. *J. Chem. Inf. Model.* (2020) (in press).
4. Verpoort PC, MacDonald P, Conduit GJ. Materials data validation and imputation with an artificial neural network. *Comput. Mater. Sci.* 147, 176–185 (2018).
5. Segall MD, Champness EJ. The challenges of making decisions using uncertain data. *J. Comput. Aided. Mol. Des.* 29(9), 809–816 (2015).
6. Martin EJ, Polyakov VR, Tian L, Perez RC. Profile-QSAR 2.0: kinase virtual screening accuracy comparable to four-concentration IC_{50}s for realistically novel compounds. *J. Chem. Inf. Model.* 57(8), 2077–2088 (2017).
7. Martin EJ, Polyakov VR, Zhu X-W, Tian L, Mukherjee P, Liu X. All-assay-max2 pQSAR: activity predictions as accurate as 4-concentration IC_{50}s for 8,558 novartis assays. *J. Chem. Inf. Model.* 59(10), 4450–4459 (2019).
8. Simm J, Arany A, Zakeri P *et al.* Macau: scalable bayesian factorization with high-dimensional side information using MCMC. Presented at: *IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP).* Tokyo, Japan, 25 September 2017.
9. Singh AP, Gordon GJ. Relational learning via collective matrix factorization categories and subject descriptors. Presented at: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* NV, USA, 24 August 2008.
10. Abadi M, Barham P, Chen J *et al.* TensorFlow: a system for large-scale machine learning. Presented at: *12th USENIX Symposium on Operating Systems Design and Implementation.* CA, USA, 2 November 2016.