Data Imputation Through Deep Learning

Machine learning can struggle to make accurate predictions when it comes to drug discovery; however, imputation methods are providing more accurate results for scientists allowing them to discover quality compounds

Matt Segall at Optibrium with authors from Optibrium, Intellegens, and Cavendish Laboratory, University of Cambridge, UK Scientific

Machine learning (ML) methods are routinely used in drug discovery to build models that can predict the properties of compounds directly from their chemical structure. These quantitative structure-activity relationship (QSAR) models take 'features' of chemical structures (often referred to as 'descriptors') as input to predict one or more properties, including activities against biological targets or in phenotypic assays and a broad range of absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties. However, even the most sophisticated ML methods can struggle to produce high-quality predictions, due to the limitations of drug discovery data: the number of compounds with data for any given experimental endpoint is small when compared with ML datasets in many other fields; the overlap of compounds measured in different endpoints is even smaller; and the data generated by biological assays are noisy due to experimental variability.

Imputation methods take a different approach, using the limited property data that are available as inputs to 'fill in the gaps' where measured values are not yet available. Imputation methods apply deep learning to both compound descriptors and sparse assay data, as illustrated in **Figure 1**. The resulting model 'learns' directly from correlations between experimental endpoints, in addition to relationships between structural features of compounds and the experimental data. This approach makes better use of the sparse and noisy data in drug discovery to produce more accurate predictions than QSAR models, which enables better targeting of the most promising compounds.

This approach was first demonstrated in a proof-of-concept study using a public-domain dataset, which comprised kinase activities for 13,000 compounds, measured across 159 experimental assays corresponding to different kinase targets (1). In this dataset, only 5% of the possible measured values were available. In the study, deep learning imputation outperformed a wide variety of ML methods, including the latest multi-target deep neural networks, and other imputation approaches, as illustrated in **Figure 2**. More recently, a practical application to an ongoing drug discovery project confirmed the advantages of deep learning imputation when applied to heterogeneous data, including activities measured in biochemical and phenotypic screens, and *in vitro* ADMET properties (2). In particular, in **Figure 3**, we can see the excellent performance on a complex, phenotypic endpoint that conventional QSAR models cannot predict.

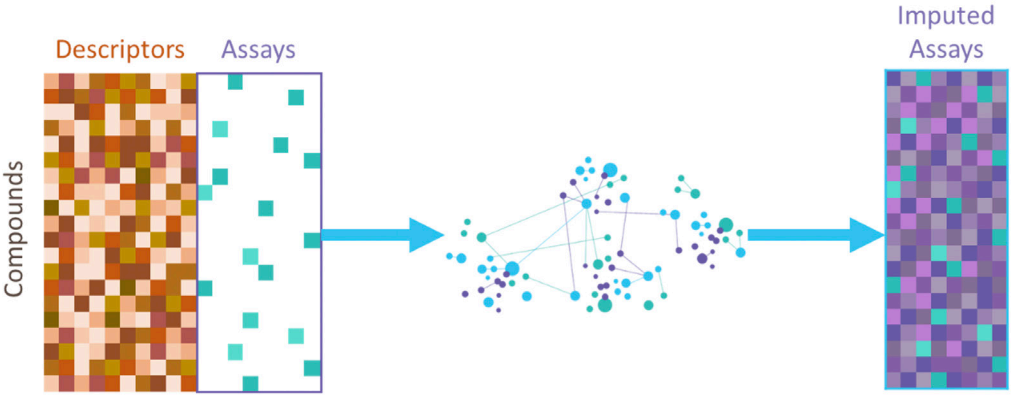


Figure 1: An illustration a deep learning imputation method. This takes compound descriptors and sparse assay data as input and imputes the missing experimental values. Compound descriptors are illustrated by orange squares in a complete matrix, assay data are shown as green squares in a sparse matrix, and imputed values as purple squares

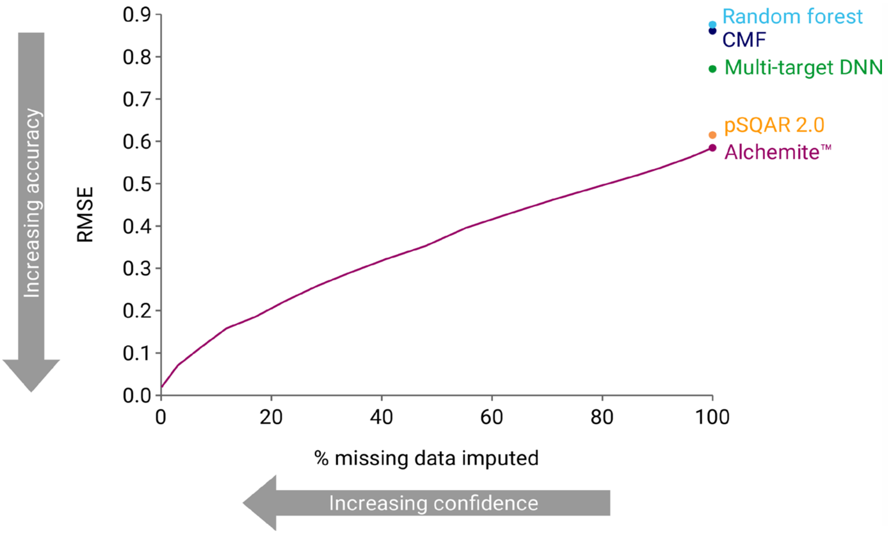


Figure 2: Performance in root mean squared error (RMSE) measured on the independent test set of kinase activity data for three imputation models – deep learning imputation, collective matrix factorisation (CMF) (6), and profile QSAR 2.0 (pQSAR 2.0) (8) – and two QSAR methods – random forest (5), and a deep neural network (Multi-target DNN) (7). This shows that deep learning imputation models achieved the most accurate predictions on the full test set. Furthermore, this illustrates that the accuracy of the models prediction increases (corresponding to a reduction in RMSE) when focusing on the most confidently predicted results

This result illustrates the advantage of directly learning the relationships between assay endpoints from sparse data. Phenotypic activities are the result of multiple factors, including target activities, cell permeability, solubility, and protein binding. It is not possible to capture these complex relationships using QSAR models, which achieved a result that is worse than random!

Deep learning imputation models also provide a probability distribution for each imputed value, focusing attention on the most confident, and hence, most accurate, results, providing the best basis for decision-making (see **Figure 2**). Analysis of the distribution of predicted values also identifies when an experimentally measured value differs significantly from the expected result. Such an outlier could represent an unexpected structure-activity relationship to guide further optimisation, or a potential experimental error to consider for retesting. Errors may be a false negative and a valuable missed opportunity.

Furthermore, as imputation methods learn about the relationships between experimental endpoints, they can be used to suggest the most important experimental data to measure and with which to predict downstream outcomes of interest with greater accuracy. This 'active learning' approach focuses experimental efforts on obtaining the critical results to choose the highest-quality compounds for progression.

A global pharmaceutical company may have millions of compounds in their collection and results from tens of thousands of experimental assays. However, typically, fewer than 1% of these potential data points will have been measured in practice, so imputing the missing values can result in up to 100x more data to store and search, corresponding to billions of data points – a true 'Big Data' challenge, we describe as the 'massive matrix'.

Addressing the Big Data Challenge

Given the magnitude and complexity of the data, several technical challenges must be addressed to deliver the full potential of a platform for data imputation:

- The platform should have access to the latest experimental data, for example, a data warehouse or electronic lab notebook

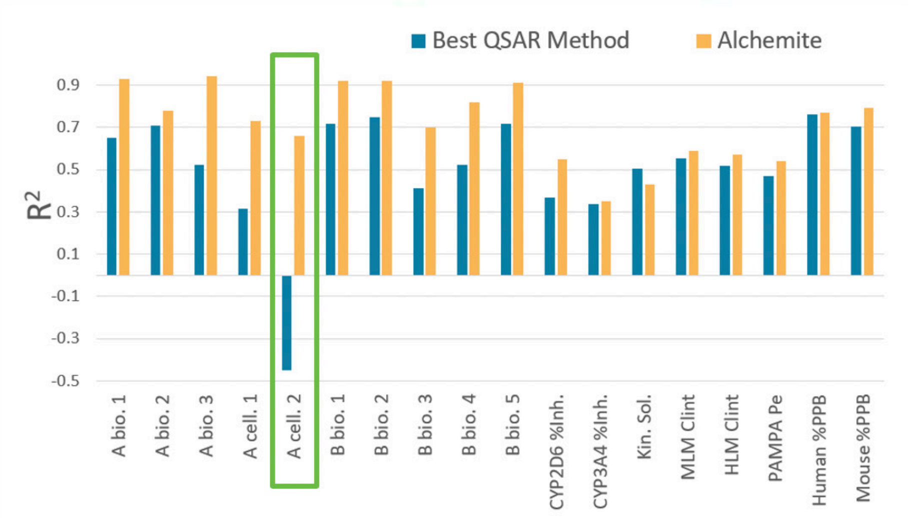


Figure 3: Results on an independent test for a project-related heterogeneous data set, including biochemical activities (A bio. 1-3 and B bio. 1-5) phenotypic cell-based assays (A cell. 1 and 2) and ADMET endpoints. The coefficient of determination (R2) is shown for the best of four QSAR methods (blue) with an Alchemite model (orange) for each of the endpoints. An R2 of one indicates a perfect prediction, zero represents random performance and a negative value is worse than random. A cell-based assay is highlighted that illustrates the ability of deep learning imputation to dramatically outperform conventional QSAR models on complex, phenotypic assays

> *This offers an even greater level of security by ensuring that the compute and storage resources are isolated from external access, benefitting from cloud providers' extensive security infrastructure, and preventing any possibility of data from different organisations being accessed.*

- Models should be updated frequently to ensure that the results are based on these latest data
- The architecture must be scalable to build models of pharma-scale datasets and handle the resulting massive matrix, which employs cloud deployment for cost-effectiveness
- Any such cloud deployment requires rigorous security to protect the intellectual property in compound structures and their associated data
- Predictions must be easily and intuitively accessible to scientists, to make quick and effective decisions on compound selection and prioritisation of experimental efforts

Figure 4 illustrates the architecture that to address these challenges. Two zones separate the handling of sensitive data on-premises (the 'blue zone') from modelling and storage of the resulting massive matrix in the cloud (the 'green zone'). In the blue zone, a 'Query Service' accesses the raw data and compound structures from the original data sources. These are pre-processed by the 'Cerella Service', to anonymise and encrypt the data while still in the blue zone. No compound structures, proprietary assay information, nor compound identifiers are passed to the green zone – even the definitions of the compound descriptors are obfuscated. This anonymisation and encryption, along with high levels of security, ensure the confidentiality of the modelling data and results.

A sparse matrix is populated with the anonymised experimental data in the green zone and used to build a deep learning imputation model. The flexible resources available on a cloud platform enable the building of models with pharma-scale datasets in hours. This process runs automatically so that the experimental data can be updated and the model rebuilt regularly, even nightly, to keep the predictions fresh. The latest model is used to impute the missing experimental values in the massive matrix.

To address the challenge of storing and efficiently searching this large volume of data, we use a distributed NoSQL document-based database (Apache Lucene) and an Elastisearch HTTP front end (3-4). A document-based database provides greater scalability than a conventional relational database and enables distribution across compute and storage servers, to provide interactive access to even the largest of datasets.

Even though the green zone components are deployed in the cloud, they can be contained in a virtual private cloud. This

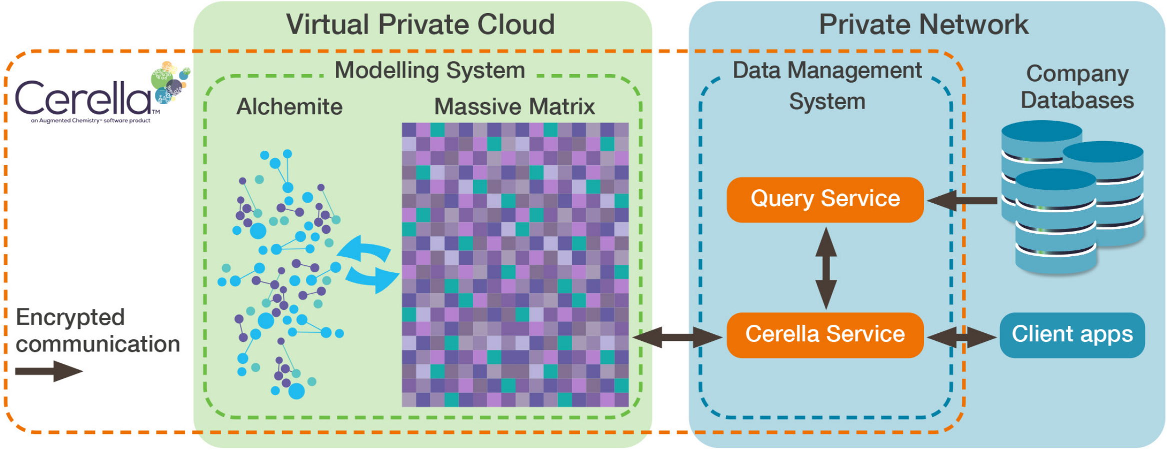
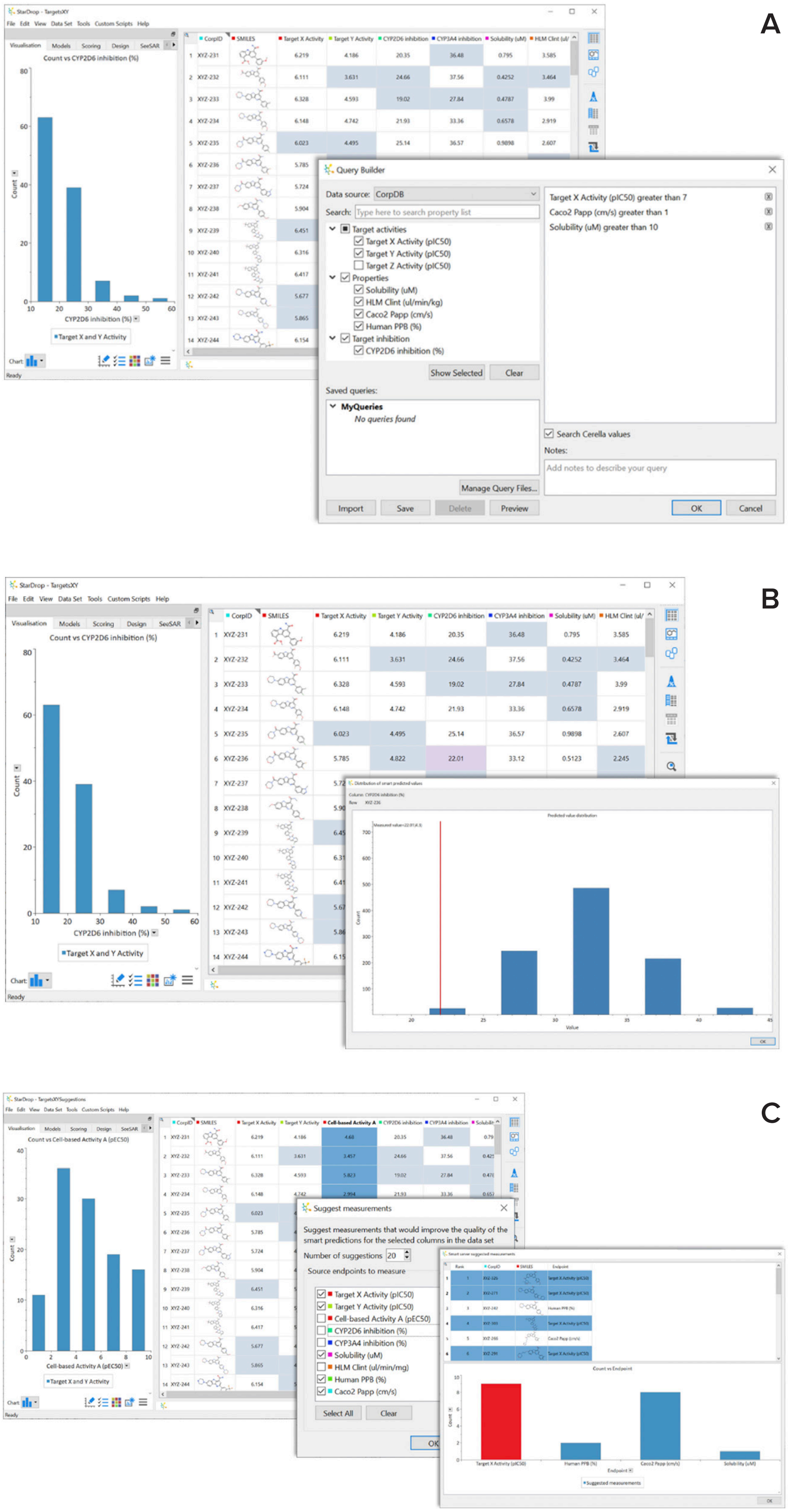


Figure 4: Schematic of the platform architecture. The 'blue zone' is hosted on-premises and manages sensitive information, such as compound structures and assay identifiers. The 'green zone' is hosted in a virtual private cloud, providing scalability for modelling, storage and searching of the 'massive matrix' of experimental and imputed data, but has no access to the most sensitive information


A
StarDrop - TargetsXY
File Edit View Data Set Tools Custom Scripts Help
Visualisation Models Scoring Design SeeSAR
Count vs CYP2D6 inhibition (%)
Count
CYP2D6 inhibition (%)
Target X and Y Activity
Chart
Ready
B
C
StarDrop - TargetsXYSuggestions
Count vs Cell-based Activity A (pEC50)
Cell-based Activity A (pEC50)

	CorpID	SMILES	Target X Activity	Target Y Activity	CYP2D6 inhibition	CYP3A4 inhibition	Solubility (uM)	HLM Clint (ul/
1	XYZ-231		6.219	4.186	20.35	36.48	0.795	3.585
2	XYZ-232		6.111	3.631	24.66	37.56	0.4252	3.464
3	XYZ-233		6.328	4.593	19.02	27.84	0.4787	3.99
4	XYZ-234		6.148	4.742	21.93	33.36	0.6578	2.919
5	XYZ-235		6.023	4.495	25.14	36.57	0.9898	2.607
6	XYZ-236		5.785					
7	XYZ-237		5.724					
8	XYZ-238		5.904					
9	XYZ-239		6.451					
10	XYZ-240		6.316					
11	XYZ-241		6.417					
12	XYZ-242		5.677					
13	XYZ-243		5.865					
14	XYZ-244		6.154					

Query Builder

Data source: CorpDB

Search: Type here to search property list

- ■ Target activities
  - ☑ Target X Activity (pIC50)
  - ☑ Target Y Activity (pIC50)
  - □ Target Z Activity (pIC50)
- ☑ Properties
  - ☑ Solubility (uM)
  - ☑ HLM Clint (ul/min/kg)
  - ☑ Caco2 Papp (cm/s)
  - ☑ Human PPB (%)
- ☑ Target inhibition
  - ☑ CYP2D6 inhibition (%)

Target X Activity (pIC50) greater than 7 ☒
Caco2 Papp (cm/s) greater than 1 ☒
Solubility (uM) greater than 10 ☒

Saved queries:

MyQueries
No queries found

☑ Search Cerella values

Notes: Add notes to describe your query

	CorpID	SMILES	Target X Activity	Target Y Activity	CYP2D6 inhibition	CYP3A4 inhibition	Solubility (uM)	HLM Clint (ul/
1	XYZ-231		6.219	4.186	20.35	36.48	0.795	3.585
2	XYZ-232		6.111	3.631	24.66	37.56	0.4252	3.464
3	XYZ-233		6.328	4.593	19.02	27.84	0.4787	3.99
4	XYZ-234		6.148	4.742	21.93	33.36	0.6578	2.919
5	XYZ-235		6.023	4.495	25.14	36.57	0.9898	2.607
6	XYZ-236		5.785	4.822	22.01	33.12	0.5123	2.245
7	XYZ-237		5.72					
8	XYZ-238		5.90					
9	XYZ-239		6.45					
10	XYZ-240		6.31					
11	XYZ-241		6.41					
12	XYZ-242		5.67					
13	XYZ-243		5.86					
14	XYZ-244		6.15					

Distribution of smart predicted values

Column: CYP2D6 inhibition (%)
Row: XYZ-236

	CorpID	SMILES	Target X Activity	Target Y Activity	Cell-based Activity A	CYP2D6 inhibition	CYP3A4 inhibition	Solubilit
1	XYZ-231		6.219	4.186	4.68	20.35	36.48	0.79
2	XYZ-232		6.111	3.631	3.457	24.66	37.56	0.425
3	XYZ-233		6.328	4.593	5.823	19.02	27.84	0.478
4	XYZ-234		6.148	4.742	2.994	21.93	33.36	0.657
5	XYZ-235		6.023					
6	XYZ-236		5.785					
7	XYZ-237		5.724					
8	XYZ-238		5.904					
9	XYZ-239		6.451					
10	XYZ-240		6.316					
11	XYZ-241		6.417					
12	XYZ-242		5.677					
13	XYZ-243		5.865					
14	XYZ-244		6.154					

Suggest measurements

Suggest measurements that would improve the quality of the smart predictions for the selected columns in the data set

Number of suggestions 20

Source endpoints to measure

- ☑ Target X Activity (pIC50)
- ☑ Target Y Activity (pIC50)
- □ Cell-based Activity A (pEC50)
- □ CYP2D6 inhibition (%)
- □ CYP3A4 inhibition (%)
- ☑ Solubility (uM)
- □ HLM Clint (ul/min/mg)
- ☑ Human PPB (%)
- ☑ Caco2 Papp (cm/s)

Smart server suggested measurements

	Rank	CorpID	SMILES	Endpoint
1	1	XYZ-326		Target X Activity (pIC50)
2	2	XYZ-271		Target X Activity (pIC50)
3	3	XYZ-242		Human PPB (%)
4	4	XYZ-303		Target X Activity (pIC50)
5	5	XYZ-266		Caco2 Papp (cm/s)
6	6	XYZ-291		Target X Activity (pIC50)

Image 1: Example workflows with results of an imputation model. (a) Querying a database for compounds with desired criteria can also return compounds with values that are imputed to meet the criteria (blue cells) as well as those that have already been experimentally measured (white cells). (b) An outlier (purple cell) can be investigated to compare the measured value (red line) with the probability distribution for the corresponding imputed value. (c) Selecting a target assay (dark blue column) for prediction and additional assays that can be performed suggests the most valuable assays and compounds to measure with which to make better predictions for the best compounds for the target assay

offers an even greater level of security by ensuring that the compute and storage resources are isolated from external access, benefitting from cloud providers' extensive security infrastructure, and preventing any possibility of data from different organisations being accessed.

The Cerella Service also handles all communication with the user in the blue zone by decrypting the results and matching them with the corresponding compound structures and assay information. A rigorous authentication system also ensures that users can only access the data for which they have permission.

Finally, the model and its result predictions must answer key drug discovery project questions in a natural way. It is not sufficient to present vast quantities of data and leave them to the scientists to analyse.

For example, as shown in **Image 1(a)**, a scientist may query a database for compounds meeting their requirements, e.g., measured activity against a target of interest, good permeability and high solubility. In addition to those found to meet these criteria based on experimental data, more compounds can be presented based on confidently imputed results. These can be prioritised for further investigation to confirm the imputed hypotheses.

Experimental results can be automatically flagged as outliers, as illustrated in **Image 1(b)**. By 'drilling down', the measured value can be compared with the imputed probability distribution for that specific value, to identify potential experimental errors or new opportunities due to false negatives.

Alternatively, we can choose an assay for which we want to make better predictions and ask which measurements would be most valuable to improve the quality of the model results. **Image 1(c)** shows a workflow that begins with selecting a cell-based assay as a target for prediction. We can then specify how other lower-cost or higher-throughput assays could be run, and the output highlights the most valuable assays and compounds to measure and provide as inputs to the model to most accurately identify the best compounds for the target assay.

Conclusions

Data imputation using deep learning is a new approach that gains more value than traditional QSAR models from experimental data, to make better predictions for compound activities and properties, and confidently guide critical decisions in the prioritisation of compounds and resources. However, this method creates new challenges in managing the volume of the resulting data, making the results easily accessible and enabling decision-makers to answer their key questions intuitively.

We have presented an approach to address these challenges, using a hybrid on-premises and cloud-based architecture, providing a best-of-both-worlds solution. The on-premises elements handle the most sensitive data, while cloud deployment provides the scalable resources required for model building and execution, and also for the storage and quick interrogation of the results.

The combination of an innovative scientific approach with modern IT infrastructure delivers new ways to guide the optimisation of high-quality compounds more efficiently.

References

1. *Whitehead T* et al, *Imputation of assay bioactivity data using deep learning,* J Chem Inf Model *59(3): pp1,179-204, 2019*
2. *Irwin B* et al, *Practical applications of deep learning to impute heterogeneous drug discovery data,* J Chem Inf Model *60(6): pp2,848-57, 2020*
3. *Visit: lucene.apache.org*
4. *Visit: www.elastic.co/elasticsearch*
5. *Visit: www.stat.berkeley.edu/~breiman/randomforest2001.pdf*
6. *Visit: www.cs.cmu.edu/~ggordon/singh-gordon-kdd-factorization.pdf*
7. *Visit: www.arxiv.org/pdf/1603.04467.pdf*
8. *Martin E* et al, *Profile-QSAR 2.0: Kinase virtual screening accuracy comparable to four-concentration IC50s for realistically novel compounds,* J Chem Inf Model *57(8): pp2,077-88, 2017*

Matthew Segall has a Master of Science in Computation from the University of Oxford, UK, and a PhD in Theoretical Physics from the University of Cambridge, UK. As Associate Director at Camitro (UK), ArQule Inc, and then Inpharmatica, he led a team developing predictive ADME models and state-of-the-art intuitive decision and visualisation tools for drug discovery and responsible for Inpharmatica's ADME business, including experimental ADME services and the StarDrop platform. Following the acquisition of Inpharmatica, Matt led a management buyout of the StarDrop business to found **Optibrium**, which continues to develop research technologies and ground-breaking artificial intelligence services to improve the efficiency and productivity of drug discovery.

info@optibrium.com

Additional authors:

Benedict Irwin, Thomas Whitehead, Samar Mahmoud, Greg Shields, Alex Elliot, Stefan-Bogdan Marcu, and Edmund Champness at Optibrium

Robert Parini at Intellegens
info@intellegens.ai

Gareth Conduit at Intellegens and Cavendish Laboratory, University of Cambridge, UK