A retrospective study into the accuracy of a deep learning method for prediction of selective activity against a broadleaf weed species

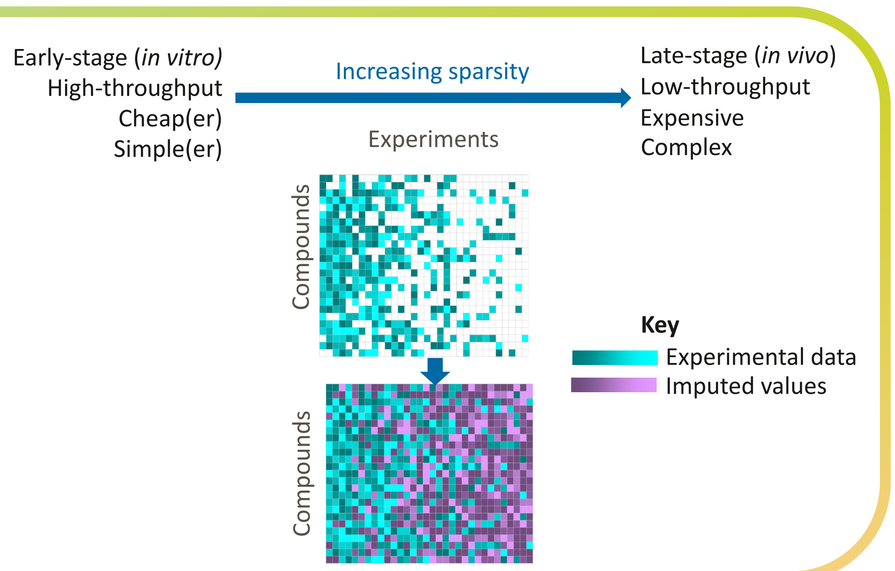
Bailey Montefiore*, Yu Tang†, Laurie Christianson†, Tamsin Mansley*, Matthew Segall*, Michael Holmes†

*Optibrium Limited, Cambridge UK † FMC Corporation, Delaware USA

The emergence of resistance and increased stringency of regulatory requirements have created a need for new agrochemicals. The long product development and increased costs mean that there is a need to look at methods to improve the success of agrochemicals in the development cycle. Machine learning methods are increasingly being applied to optimize their development. Cerella uses a deep-learning method which imputes missing data within an experimental data matrix. It accepts both molecular descriptors and sparse experimental data as input exploiting the relationships between experimentally measured endpoints [1, 2].

An ensemble of networks generates a probability distribution for each individual prediction, accounting for uncertainties in both the experimental data and any extrapolation of the training data. From this, a confidence in each prediction can be assessed.

Here, we explore how Cerella can be combined with multi-parameter optimization (MPO) to accurately prioritise compounds with an objective of controlling a broadleaf weed species to protect corn and soybean crops.



Data Set and Methods

Cerella was used to train a model of physicochemical, *in-vitro* assay, and bioactivity data against a wide range of insect, weed, and fungi species. The data spans a wide range of the development process, from high-throughput experiments to late-stage *in-vivo* ones.

Multi-parameter optimization allows us to consider multiple properties to assess a compound's likelihood of success [3]. The model predictions were combined into a scoring profile, which included a broadleaf weed species (BS1), corn, and soybean (SOYB) (Figure 1). The scoring criteria are for activity against BS1 without affecting corn and soybean crops.

Of the test set compounds, 37 have measured data. By comparing the scores for the imputed values with the scores for measured values, we can assess how accurate they were and therefore the accuracy we would expect if we were to apply the scoring profile to the next set of compounds.

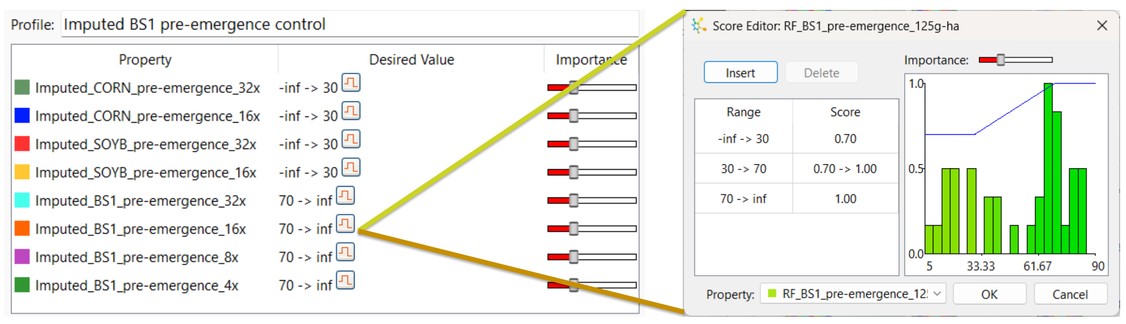


Figure 1. The scoring profile created using StarDrop to assess the compounds for activity against BS1, corn and soybean (left). An example scoring function where a value of 70 or greater is desired (right).

Results

By combining predictions into a scoring profile, we can predict which compounds are likely to be successful (Figure 2). To estimate the added benefit of deep-learning imputation, a standard quantitative structure-activity relationship (QSAR) method, Random Forest, was used as a comparison. We see a better correlation between the scores on the measured data and Cerella predictions compared to Random Forest predictions (Figures 2A and B).

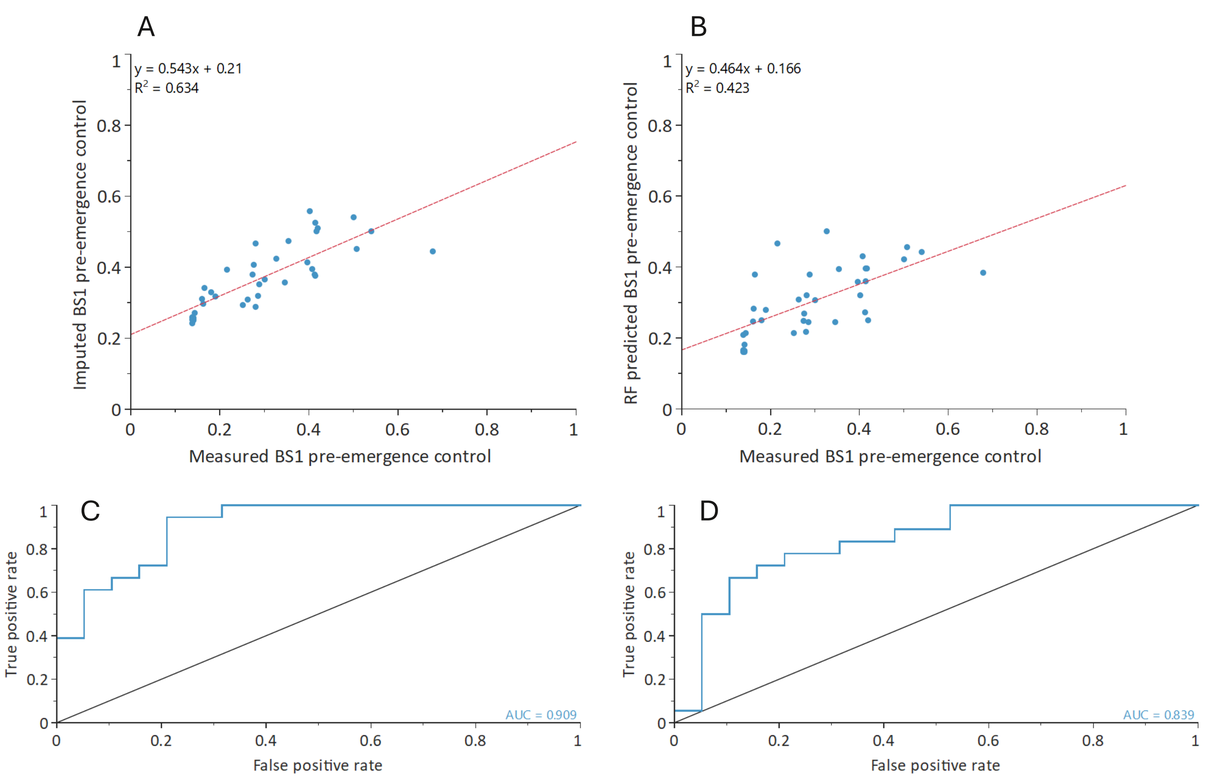


Figure 2. Accuracy of the predicted scores. Scores for Cerella predictions plotted against the scores of the measured data (A), scores for Random Forest predictions plotted against the scores of measured data (B). Receiver operating characteristic (ROC) curves showing the performance of the Cerella (C) and Random Forest predictions (D) in classifying the top-scoring 50% of compounds using measured data.

Figures 2C and D indicate a better ability of the deep-learning imputation method with MPO to differentiate between good and bad compounds over Random Forest with MPO. The area under the curve (AUC) for the Cerella predictions is 0.91 and 0.84 using Random Forest predictions.

The uncertainties assigned to each prediction can be combined into an uncertainty in the overall score, allowing us to understand how much we can confidently distinguish between compounds.

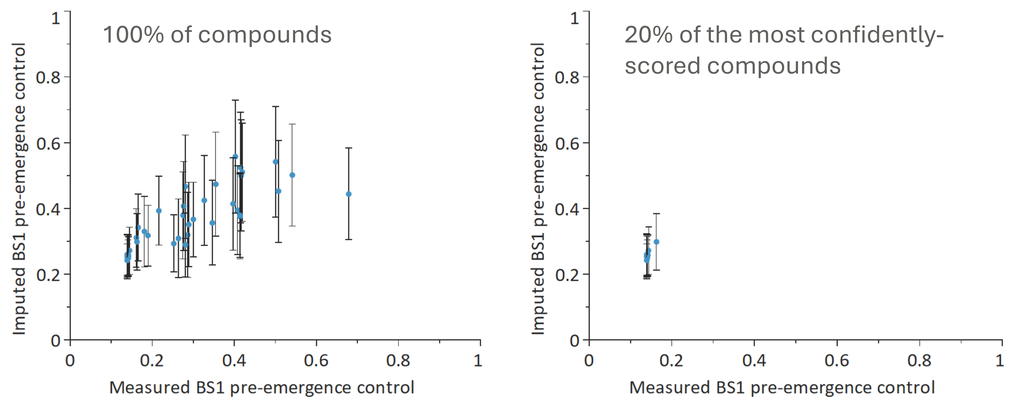


Figure 3. Correlation between the scores for predicted and measured values for all compounds, and the top 20% most confidently-scored compounds. The uncertainties in the predicted scores are shown as error bars.

The most confident scores are the low ones, and we can see that these do indeed have low measured activity (Figure 3). This indicates that we can be confident when applying this scoring profile in the future that where the scores are low with high confidence, the compounds will not meet the desired criteria for success, and we can save resources by not progressing them for further experimental testing.

Cerella allows for transparency of the inputs it has used for imputation, and this is provided at an endpoint-by-endpoint basis. The Importance Analysis allows for a comparison of these endpoints and identification of endpoints that may be most predictive of others.

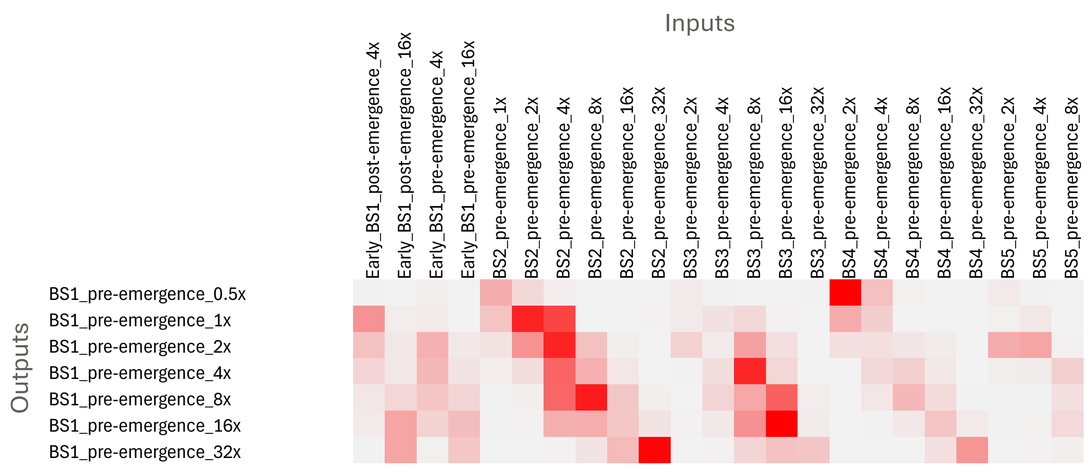


Figure 4. A heatmap showing the extent that Cerella is using measured data for other experimental endpoints in the imputation of BS1 under pre-emergence conditions (the output). Only the most informative inputs are shown.

Figure 4 shows the most informative inputs for prediction, where the deeper the red colour, the more informative it is in the imputation of the output endpoint. Cerella has identified related broadleaf species and utilized the measured data for those to inform its prediction of BS1. BS2 was particularly informative, much more so than even a less advanced BS1 screen, indicating that data collected under the same testing conditions, even though for a different species, is highly valuable.

Conclusions

We demonstrate that robust uncertainty estimates generated by Cerella enable the most accurate predictions to be identified. We can see that we could reliably eliminate compounds from the testing cascade where Cerella predicts they will not meet the success criteria with high confidence.

For the compounds assessed in this study, many had measurements for other broadleaf weed species. The ability of Cerella to utilize this data and apply the relationship it has understood has likely contributed to the greater predictive power seen when using imputation.

These results demonstrate that imputation enables more accurate selection of the best compounds through the screening cascade, thus enabling a reduction in the number of compounds required to be run on downstream, costly experiments.

Learn more on our webpage

References

[1] T. Whitehead *et al*. J. Chem Inf. Model. (2019) **59**(3) pp. 1197-1204
[2] B. Irwin *et al.* J. Chem. Inf Model. (2020) **60**(6), pp. 2848–2857
[3] Segall MD, Champness EJ. J Comput Aided Mol Des. (2015) **29**(9) pp. 809-816