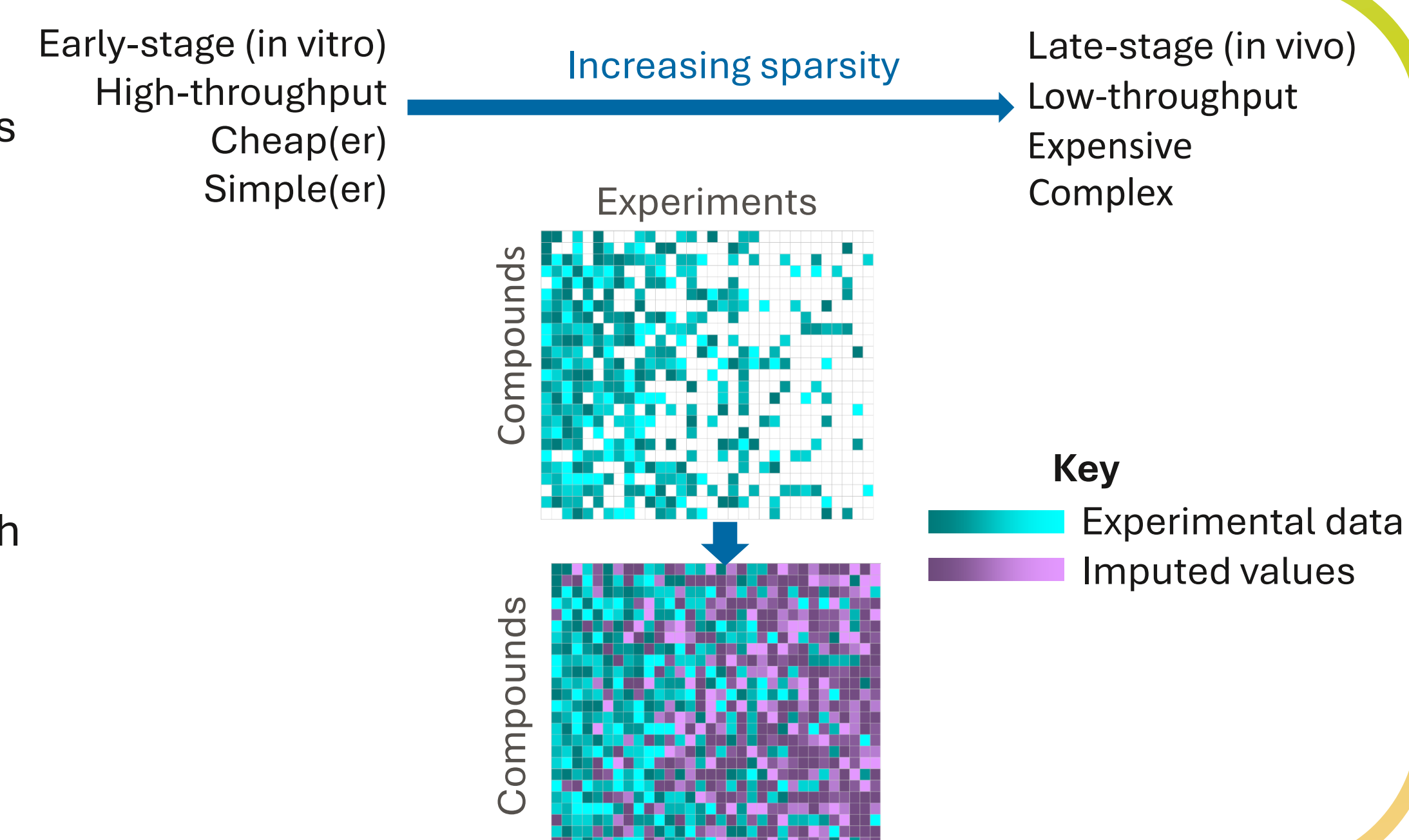


Bailey Montefiore*, Yu Tang†, Laurie Christianson†, Nathan Brown*, Michael Holmes†, Matthew Segall*
*Optibrium Limited, Cambridge UK †FMC Corporation, Delaware USA

The emergence of resistance and increased stringency of regulatory requirements have created a need for new agrochemicals. The long product development process and increased costs mean there is a need to explore methods to improve the success of agrochemicals throughout the development cycle. Machine learning methods are increasingly being applied to optimise their development. Cerella uses a deep learning method that imputes missing data in 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 in extrapolations from the training data. From this, confidence in each prediction can be assessed.

Here, we explore how Cerella can be combined with multi-parameter optimisation (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 on physicochemical and bioactivity data from a wide range of weed species, spanning the development process from early, higher-throughput experiments to late-stage in vivo experiments.

This is a retrospective proof-of-concept analysis: the training set excluded 113 compounds that were subsequently tested during a collaboration. To simulate prospective performance, predictions for these compounds were generated using only their structures and early-stage data. Within-stage correlations (early-early and late-late) were excluded from the imputation, so late-stage predictions relied only on early-stage and physicochemical data. These predictions were then validated against the late-stage data measured during the collaboration.

Results

Multi-parameter optimisation allows us to consider multiple properties to assess a compound's likelihood of success [3]. The model predictions were combined into a scoring profile that rewards activity against a broadleaf weed species (BS1) without affecting corn (CORN) or soybean (SOYB) crops, allowing us to predict which compounds are likely to be successful (Figure 1).

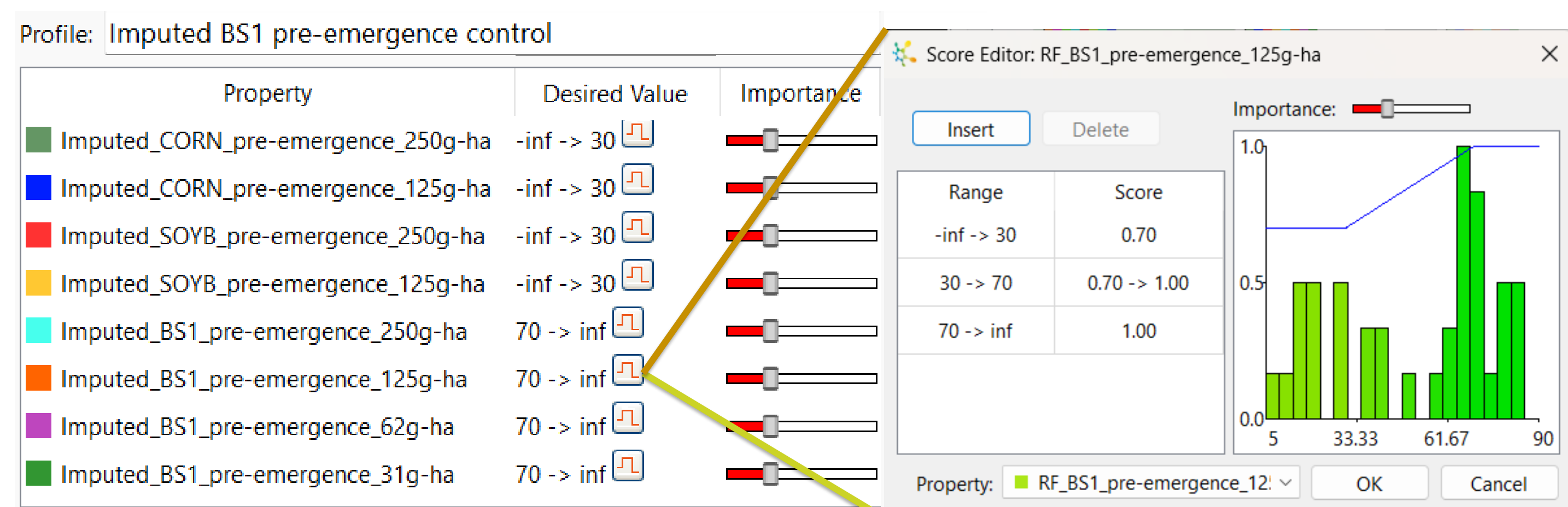


Figure 1. A multi-parameter scoring profile combines predictions across endpoints into a single, prioritisable score. The scoring profile created using StarDrop to assess the compounds for activity against BS1, corn and soybean (left). An example scoring function where a value of 70 or greater is desired (right).

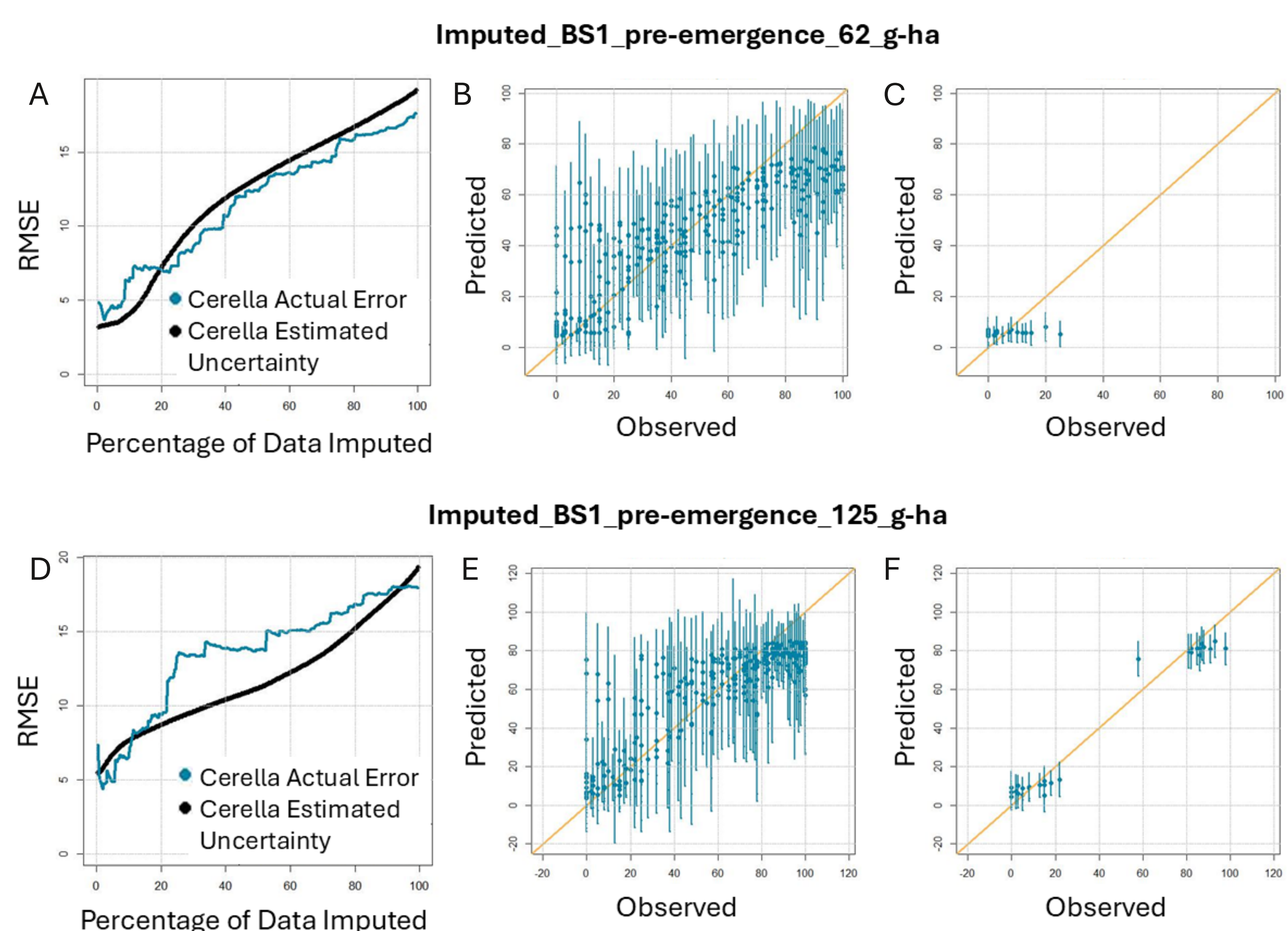


Figure 2. Cerella's uncertainty estimates are well calibrated, and predictions for the most confident compounds are the most accurate. (A, D) Root-mean-square error (RMSE) of Cerella's actual prediction error (teal) versus its estimated uncertainty (black) as compounds are cumulatively included in order of increasing predicted uncertainty. (B, E) Observed versus Cerella-predicted values for all compounds, with error bars showing predicted uncertainty. (C, F) The subset of compounds with the most confident (lowest-uncertainty) predictions.

The uncertainties assigned to each prediction can be combined into an uncertainty in the overall score, showing how confidently we can distinguish between compounds. Figure 2 shows two representative endpoints from the profile (pre-emergence control at 62 and 125 g/ha). For both, Cerella's uncertainties are well calibrated: as compounds are ranked by increasing predicted uncertainty, estimated uncertainty closely tracks the actual RMSE (Figure 2A, D), so the most confident predictions are also the most accurate. Observed versus predicted R^2 is 0.68 and 0.69, respectively (Figure 2B, E), improving further for the most confident compounds (Figure 2C, F).

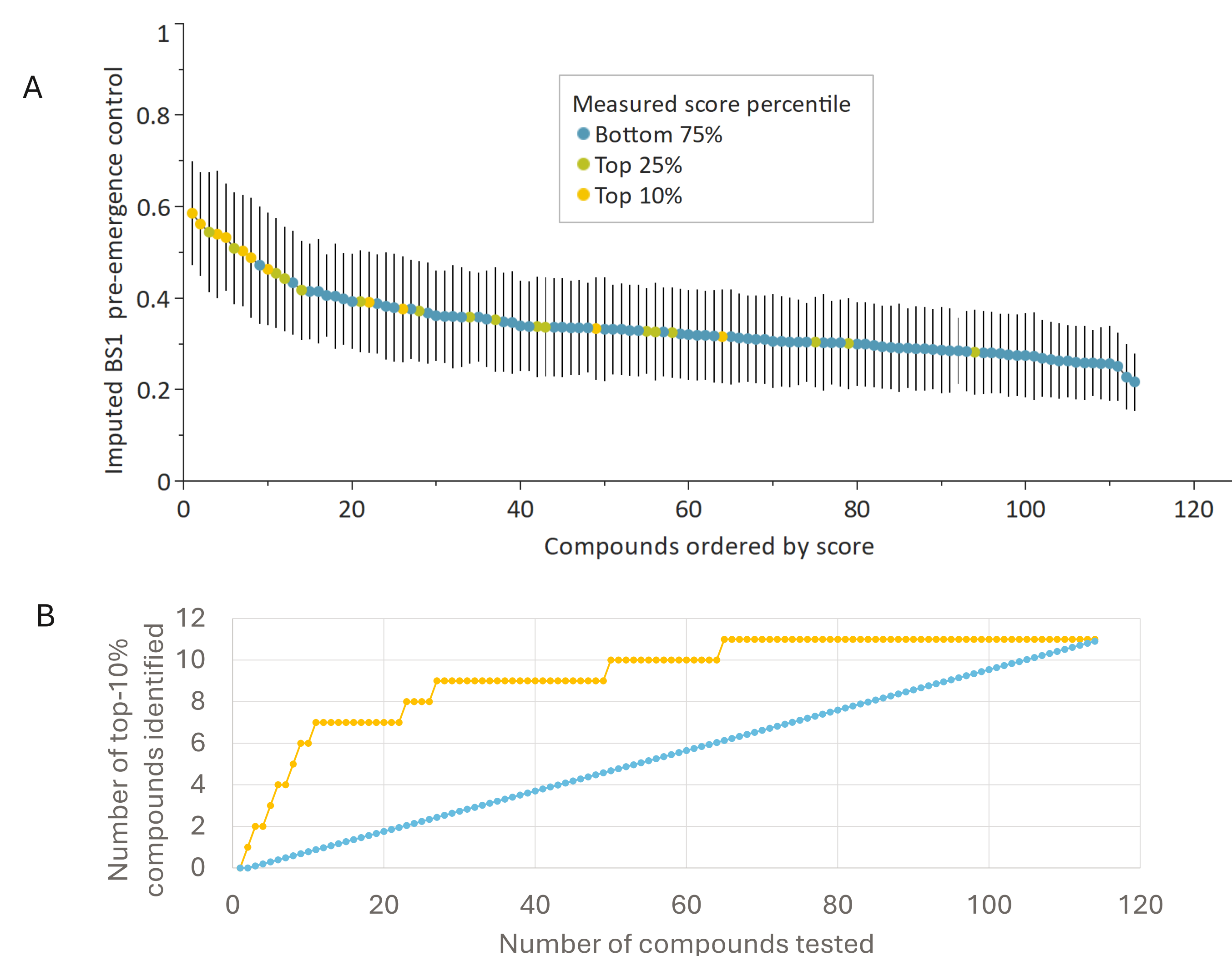


Figure 3. Prioritising compounds by predicted score identifies top performers with fewer experiments. (A) Cerella-imputed BS1 pre-emergence control scores (points) and associated uncertainty (error bars) for compounds ordered by predicted score. (B) Cumulative number of top-10% compounds identified as a function of the number of compounds tested, comparing prioritisation by predicted score (gold) with a random ordering of the same pool of compounds tested during the study period (blue).

Ranking compounds by their Cerella-predicted score achieves good prioritisation of the best-performing compounds: compounds subsequently measured in the top 10% and top 25% are concentrated among the highest-scoring compounds, becoming sparser as the predicted score decreases (Figure 3A). Testing compounds in this same order, from left to right, identified all 11 top-10% performers after just 64 of the 113 compounds evaluated, compared with needing to test nearly the entire set under a random ordering of the same compounds (Figure 3B). This represents a ~43% reduction in the compounds required to identify all top performers.

Conclusions

- Cerella's deep learning imputation learns the relationships between experimental endpoints. In this study, early-stage measurements were used to improve prediction of late-stage outcomes.
- Each prediction comes with a well-calibrated uncertainty estimate. This means compounds can be prioritised through the screening cascade with confidence, focusing resources on those most likely to succeed.
- Ranking by predicted score found all top-10% compounds after testing 64 of 113, about 43% fewer than random ordering needed.
- More accurate selection substantially reduces the number of compounds that must go on to costly downstream experiments.

References

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



Learn more
on our
webpage