Application of Deep Learning Imputation to Peptide Bioactivity and Property Prediction

Samar Mahmoud[1], Himani Tandon[1], Matthew Segall[1], Ruibo Zhang[2], Robert Thompson[3], Jeffrey Schubert[4], Dipannita Kalyani[5], Dan Sindhikara[2]

[1]R&D, Optibrium Ltd, Cambridge, United Kingdom, [2]Department of Modeling and Informatics, Merck & Co., Inc., Rahway, NJ, USA, [3]Department of Discovery Chemistry, Merck & Co., Inc., South San Francisco, CA, USA, [4]Department of Discovery Chemistry, Merck & Co., Inc., West Point, PA, USA, [5]Department of Discovery Chemistry, Merck & Co., Inc., Rahway, NJ, USA

https://doi.org/10.17952/37EPS.2024.P2267

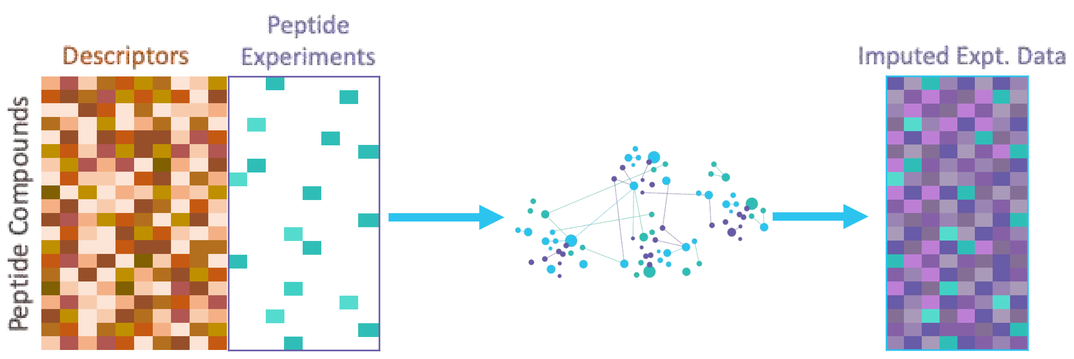
Introduction

Predicting bioactivities and properties is pivotal for advancing peptide-based therapeutics. It facilitates the rapid virtual screening of large peptide libraries and prioritisation of candidates with high target activity, stability, and permeability. However, accurate prediction is challenging due to the complex structure-activity relationships between peptide sequence, conformation and diverse biological activities, which current models often struggle to fully capture. Also, limited resources and diverse peptide libraries complicate testing, leading to sparse and noisy datasets.

In this context, we introduce the application of a deep learning platform, which employs an imputation approach to derive valuable insights from such noisy experimental data. This platform generates reliable predictions of peptide properties and provides valuable insights that can be used for peptide optimisation.

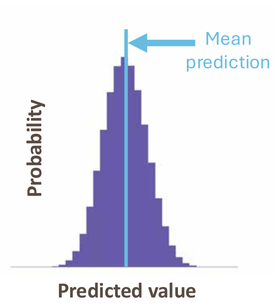
Deep Learning Imputation

Cerella™ [1,2] is a deep learning platform which accepts both molecular descriptors and sparse experimental data as inputs to impute (fill-in) the missing values, exploiting the **assay-assay correlations**, as well as structure-activity relationships (SAR).



Cerella further provides a robust estimate of the **uncertainty** in each prediction, enabling the most accurate results to be identified.

In combination, these provide **more accurate predictions** of activity and property values, enabling additional insights, such as identifying experimental outliers and the most valuable experimental data measure with which to improve a model's predictions.



Dataset and Methods

Cerella platform was applied to model two proprietary datasets of peptidic compounds, provided by Merck & Co., Inc., Rahway, NJ, USA. The peptides were mostly **macrocycles,** and **6-15 'mers** in length.

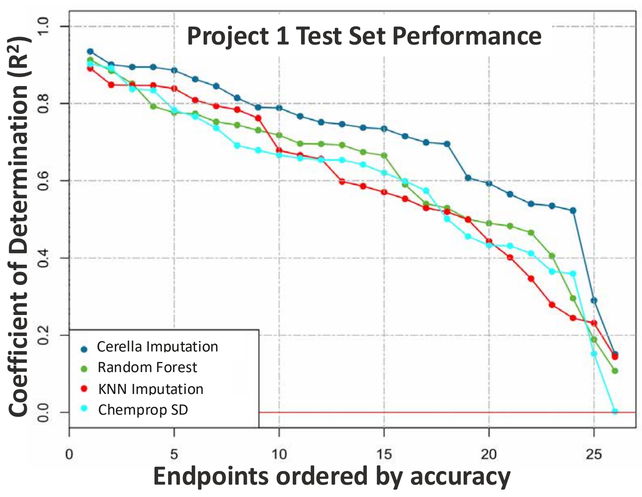
Project 1: 33 experimental endpoints and 5168 peptides

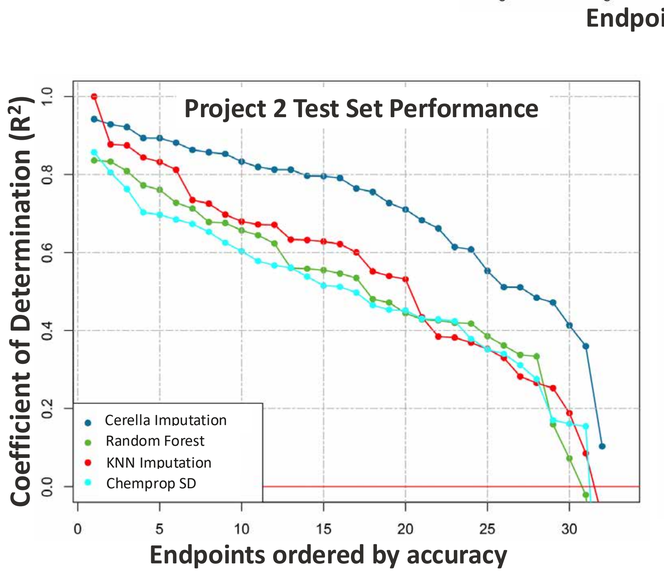
Project 2: 39 experimental endpoints and 1125 peptides

Model performances were compared with other methods including Random Forests, ChemProp, and KNN imputation.

Results

Model performance





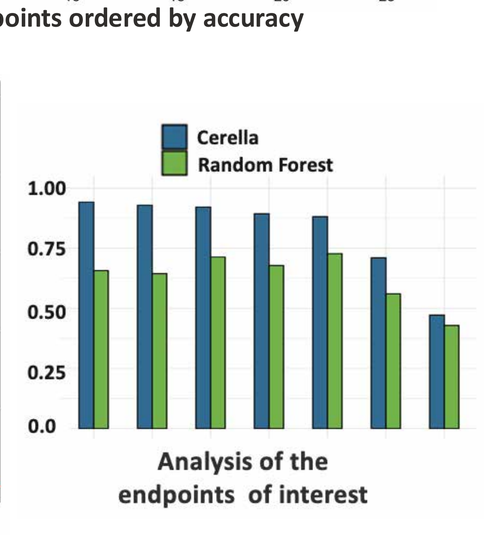


Figure 1: Imputation models outperform QSAR models for both the Projects (paired t-test p-values < 0.05), with a **substantial** improvement in **Project 2**. The bar chart illustrates the **significant** performance differences between the Imputation model and the Random Forest models for the endpoints of interest in **Project 2**. The models were built using 2D StarDrop™ descriptors. Insignificant performance differences are observed between different QSAR models (paired t-test p-values >0.05) for both Project 1 and 2

Uncertainty correlates well with accuracy

Cerella's uncertainty estimates correlate well with accuracy, highlighting the model's ability to identify accurate predictions, as shown for a peptide activity endpoints in **Project 2**. This provides **greater confidence** in decisions based on model predictions.

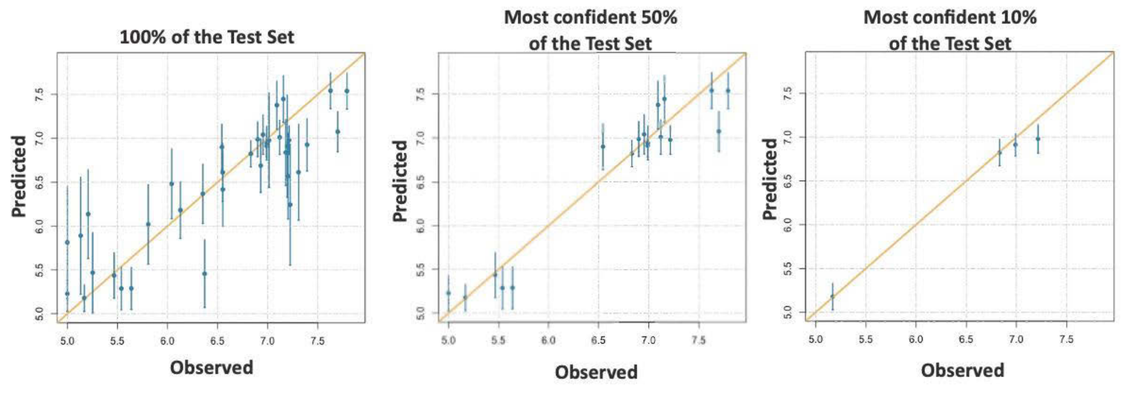


Figure 2: Correlation between uncertainty (error bars) and model accuracy: Model accuracy increases when focusing on the most confident predictions

Importance analysis guides design and optimisation

Importance analysis shows which experiments and/or descriptors contribute most to the accurate predictions of others. The heatmaps show the most informative inputs for prediction of Project 1 and 2 endpoints.

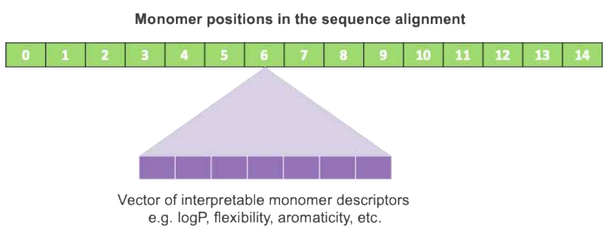


Figure 3: Sequence-based descriptors, derived from PepSeA [3] alignments using interpretable monomer descriptors, were used to enhance the imputation model.

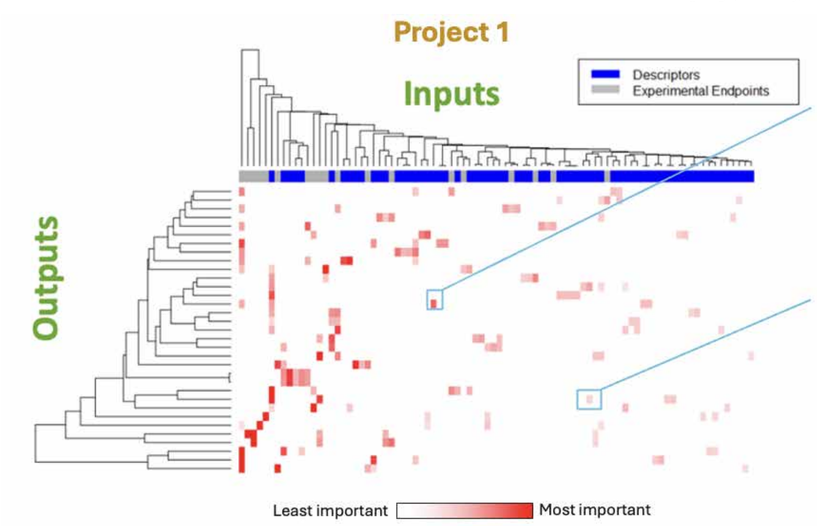


Figure 4: Experimental **Endpoint X** is influenced by the **number of aromatic atoms** at **position 4** in the peptide sequence alignment

Experimental **Endpoint Y** is informed by the monomer **logP** at **position 12** in the peptide sequence alignment

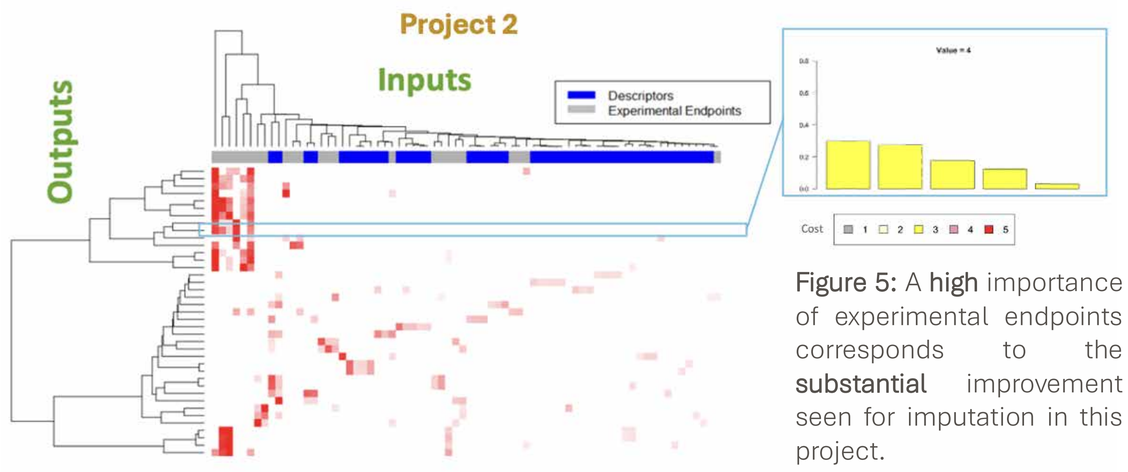


Figure 5: A **high** importance of experimental endpoints corresponds to the **substantial** improvement seen for imputation in this project.

Moderate cost experiments inform high-value outcomes, guiding the prioritisation of running the lower-cost yet informative experimental measurements to accurately impute expensive, late-stage endpoints

Conclusion

1. Cerella Imputation generates reliable predictions of peptide properties and activities by leveraging assay-assay relationships.
2. Strong correlation are confirmed between estimated uncertainties and observed accuracy of predictions.
3. Cerella guides the design of peptides and the prioritisation of experimental resources.

References

[1] B. Irwin et al. J. Chem. Inf Model. (2020) 60(6), pp. 2848–2857

[2] B. Irwin et al. Appl. AI Lett. (2021) 2(3) pp. e31-2689

[3] J. Baylon et al. J. Chem. Inf Model. (2022) 62(5) pp. 1259-1267

Learn more on our webpage