Impact of sequence representation on machine learning models of protein expression

Yuxin Shen[1], Grzegorz Kudla[2], Diego A. Oyarzún[1,3,4,*]

[1] School of Biological Sciences, University of Edinburgh, UK
[2] Institute for Genetics and Cancer, University of Edinburgh, UK
[3] School of Informatics, University of Edinburgh, UK
[4] The Alan Turing Institute, UK
[*] d.oyarzun@ed.ac.uk

Overview

Background:

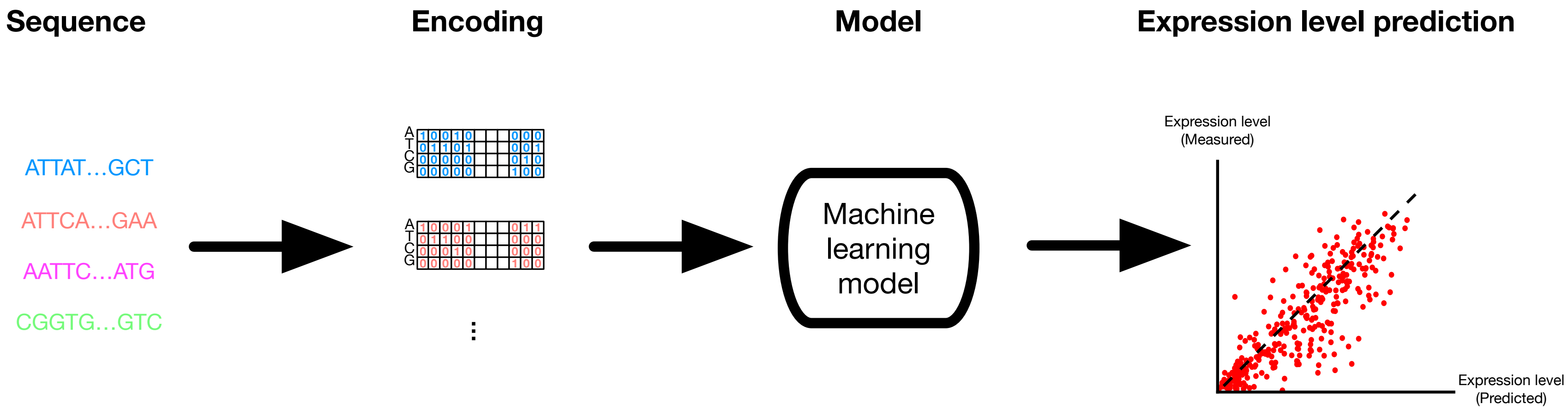
As the demand for bioproducts continues to grow, there is an increasing interest in optimizing protein expression in recombinant strains. High-throughput sequencing methods can produce sufficient data to build sequence-to-expression models using machine learning. Here, we focus on the sequence representation methods of sequence-to-expression models.

Contributions:

Our investigation involves **a comparative evaluation of one-hot encoding and mechanistic features** such as mRNA folding stability and nucleotide content. We show that models trained on mechanistic features deliver weaker local predictions compared to one-hot encoding, but provide important gains on the ability of models to predict beyond their training set. This result indicates that the DNA sequence representation is important for sequence-to-expression models along with machine learning model structures.

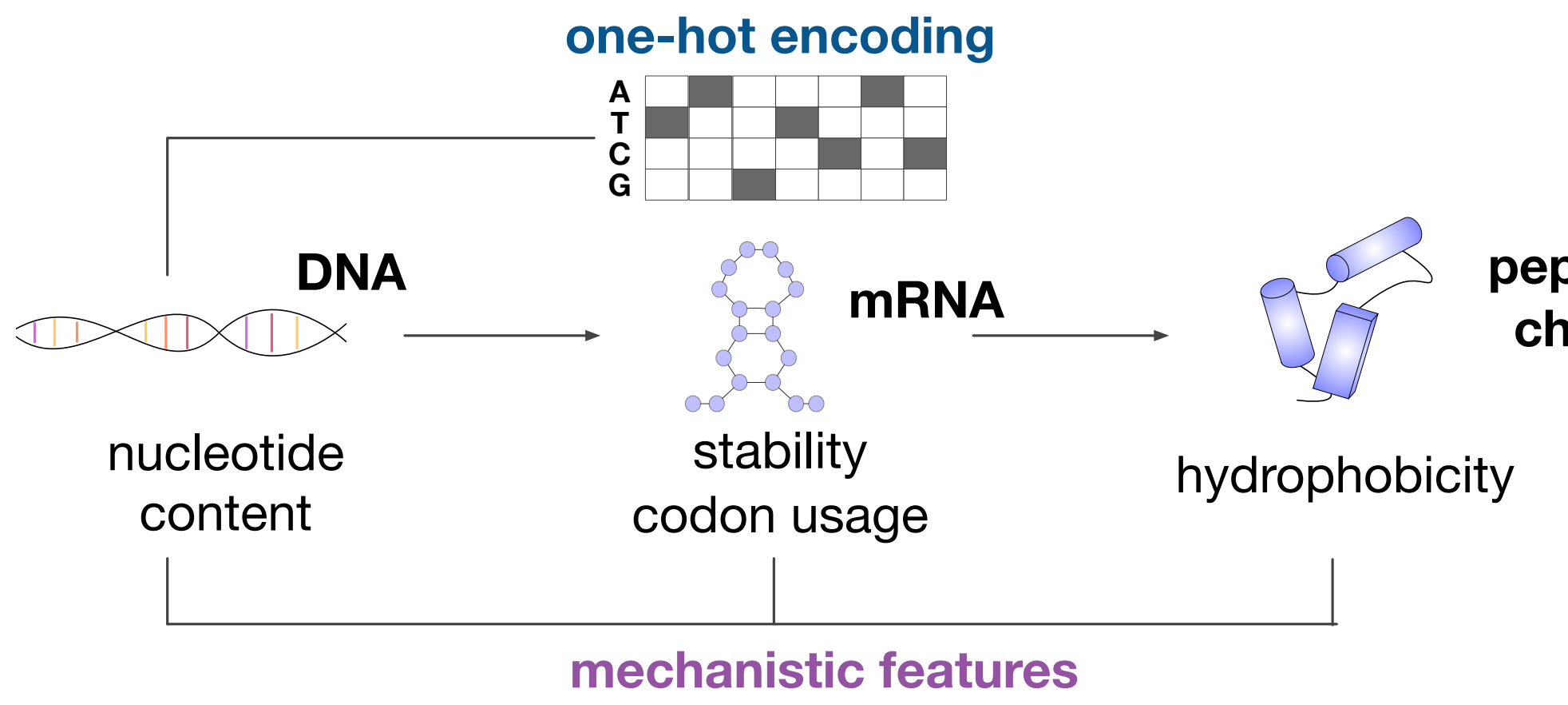
1. Sequence-to-expression model

New techniques for high-throughput screening create the large datasets that are well suited for building models to **predict protein expression in the cells from DNA sequences** using methods from machine learning and artificial intelligence (AI).

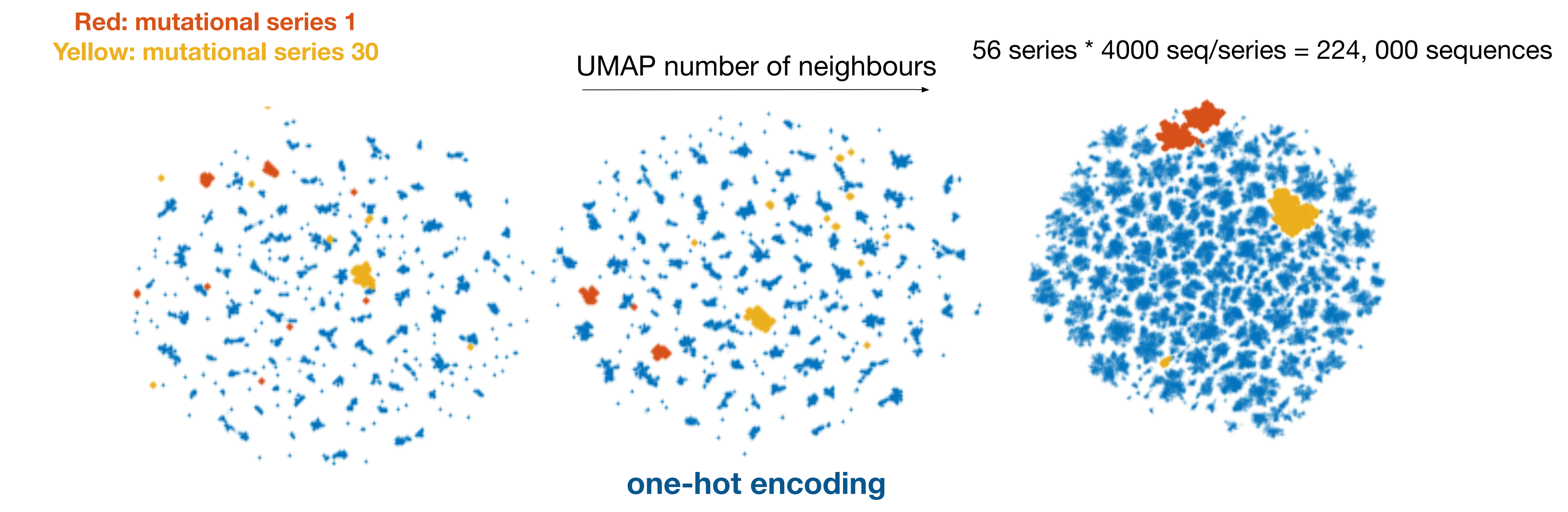


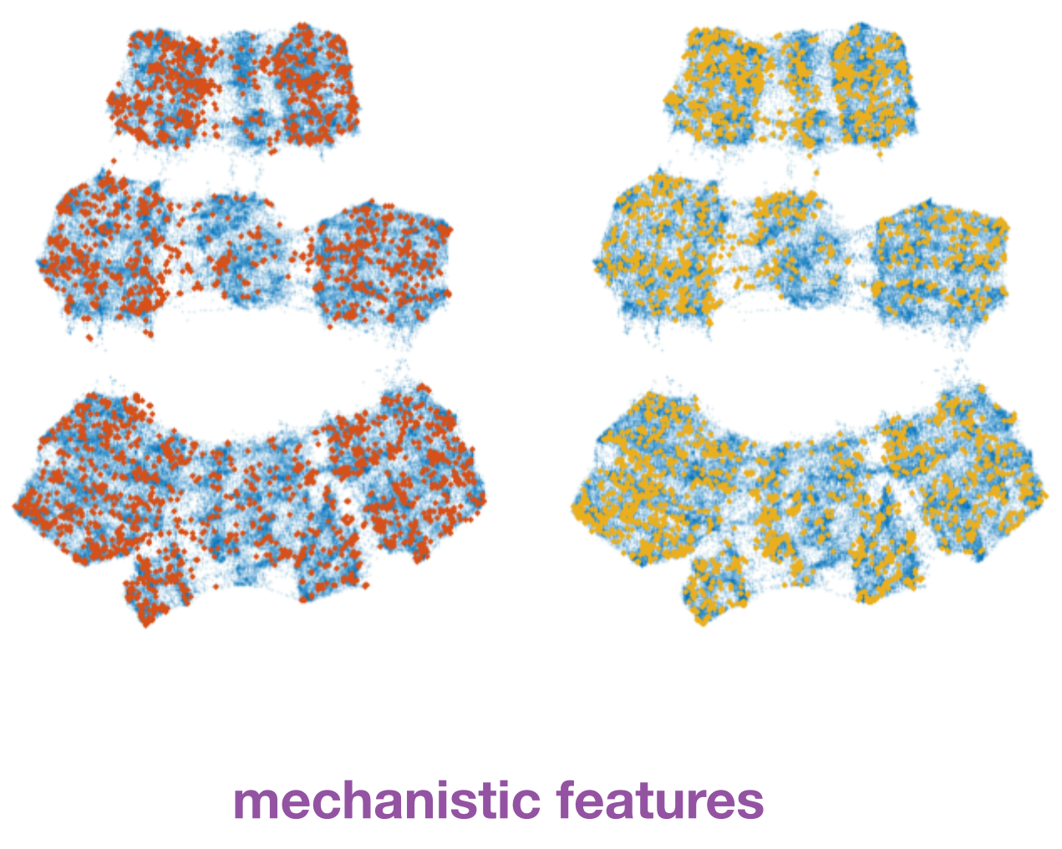
Sequence-to-expression machine learning pipeline

> In this research, we focus on the encoding method in the pipeline.



2. UMAP of feature space



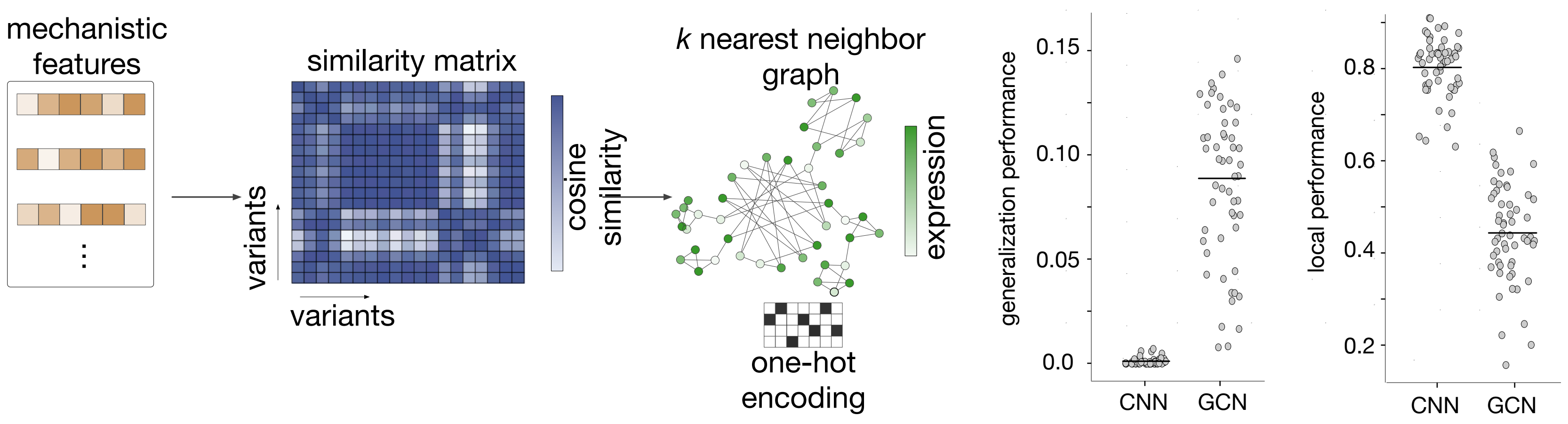


> The sequence-to-expression data in this research is 96-nt 5' CDS from the Cambray paper [2], which contains 56 mutational series.

> The UMAPs show substantial differences in **mutational series distribution** between both representations.

> Individual mutational series appear to be clustered in one-hot space; in the mechanistic feature space they tend to be **homogeneously spread** across the full library

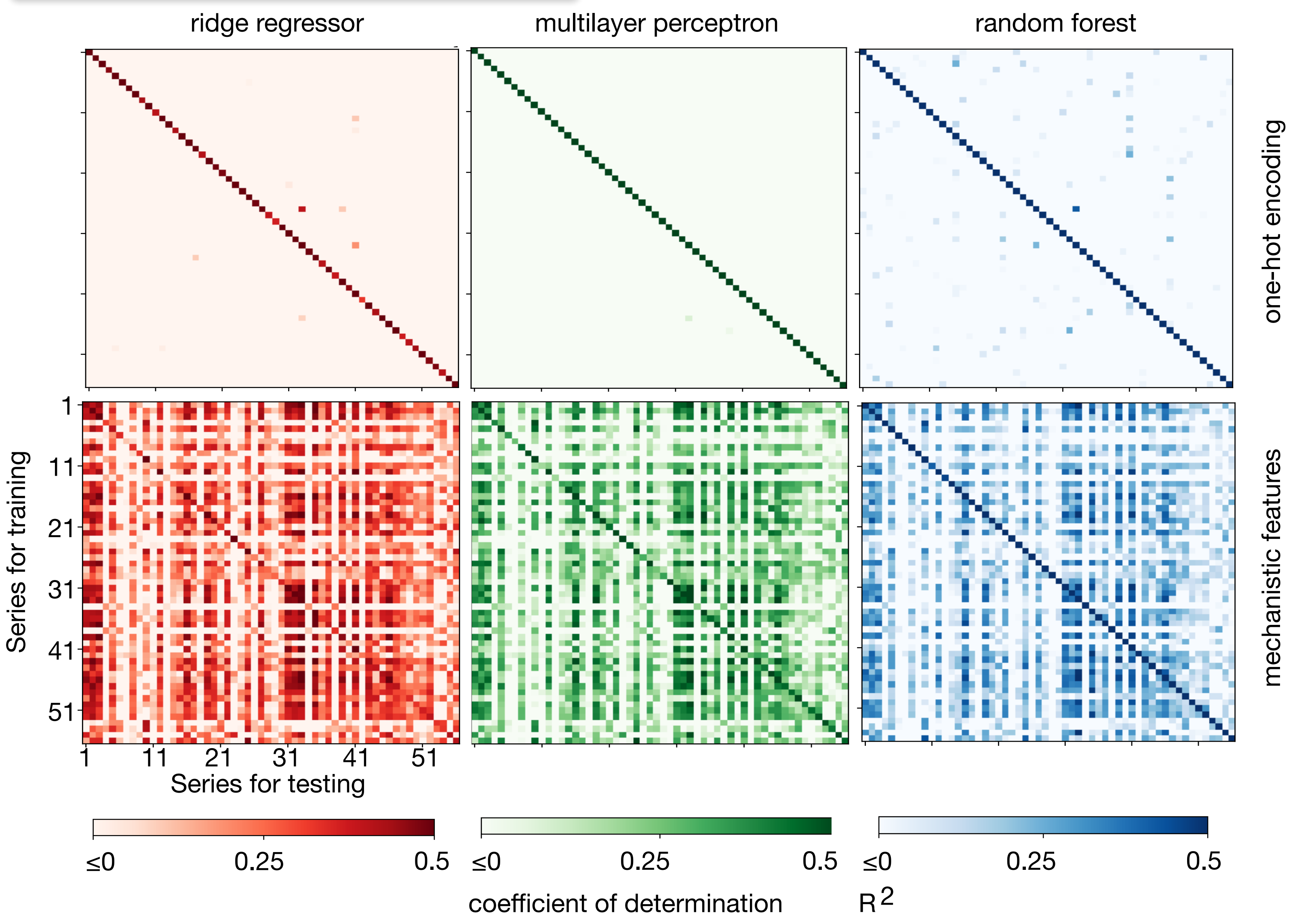
4. Geometric stacking



> Graph Neural Network (GNN) is used for geometric stacking of one-hot and mechanistic feature sets.

> GCN (a type of GNN) geometric stacking has better generalization performance compared with Convolutional Neural Network (CNN) stacking, but its local performance could decrease.

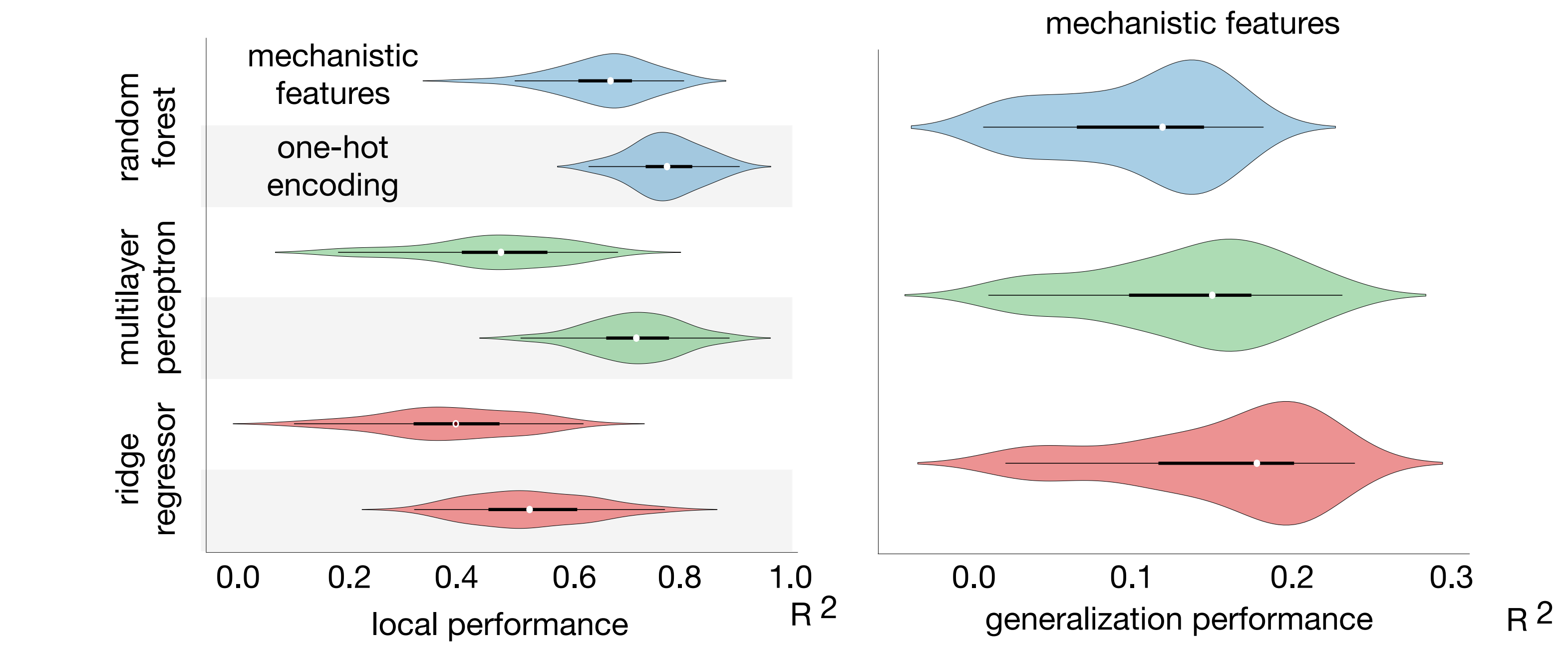
3. Performance of encodings



> Models on one-hot encoding shows almost no ability to generalize.

> Mechanistic feature encoding improves model generalization.

Quantitative comparison of model performance



> Mechanistic features improves model generalization, but affects local performance.

> Different models have different expertise for local / generalization performance.

5. Conclusion

> **Mechanistic features improves model generalization performance** compared to one-hot encoding, which shows the power of biological knowledge in modelling.

> Different sequence feature sets can be **fused together** by using model structures like **Graph Neural Network** or ensemble models to improve the performance.

> This research shows that the **DNA sequence representation** is also important for sequence-to-expression models along with machine learning model architectures.

Main references & Acknowledgements

[1] Nikolados, E. M., Wongprommoon, A., Aodha, O. M., Cambray, G., & Oyarzún, D. A. (2022). Accuracy and data efficiency in deep learning models of protein expression. *Nature Communications*, *13*(1), 7755.

[2] Cambray, G., Guimaraes, J. C., & Arkin, A. P. (2018). Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in Escherichia coli. *Nature biotechnology*, *36*(10), 1005-1015.

[3] Nikolados, E. M., & Oyarzún, D. A. (2023). Deep learning for optimization of protein expression. *Current opinion in biotechnology*, *81*, 102941.

[4] Shen, Y., Kudla, G., & Oyarzun, D. A. (2024). DNA representations and generalization performance of sequence-to-expression models. *bioRxiv*, 2024-02.

Link to our preprint

This work was supported by the UKRI Biotechnology and Biological Sciences Research Council (BBSRC) grant number BB/T00875X/1.