Deep Learning and its Application to Predicting Quantitative Phenotypes from Genomic Data

TFG by **Víctor Fernández Oliveras** / Bibliographic review / Genetics Degree

INTRODUCTION

Inside artificial intelligence (AI), **machine learning** (ML) methods are designed to **mimic human brain function**. The **data quantity and complexity** these methods can process are limited, so **deep learning** (DL), the core framework of which is **neural networks** (NN) formed of neurons or nodes, emerged as a powerful tool to handle these datasets, such as **nonlinear genotype-phenotype relations**. These predictive models are based on **supervised learning**, as they are trained with labeled samples to predict unseen data.

⚠ Not inhterently superior to ML. ⇒ **ML outperforms** in predicting **phenotypes** based on **purely additive** effects.

OBJECTIVES

- **Describe** model architecture, training, interpretability and main challenges in DL.
- **Contextualize** DL in genomics predictions.
- **Critically review** current situation and **propose** future research.
- **Create** a naive model to exemplify the concepts discussed.

METHODOLOGY

mainly → 2020-2025

Keywords: "deep learning", "neural network", "genomics", "phenotype prediction", "complex trait"...

RESULTS

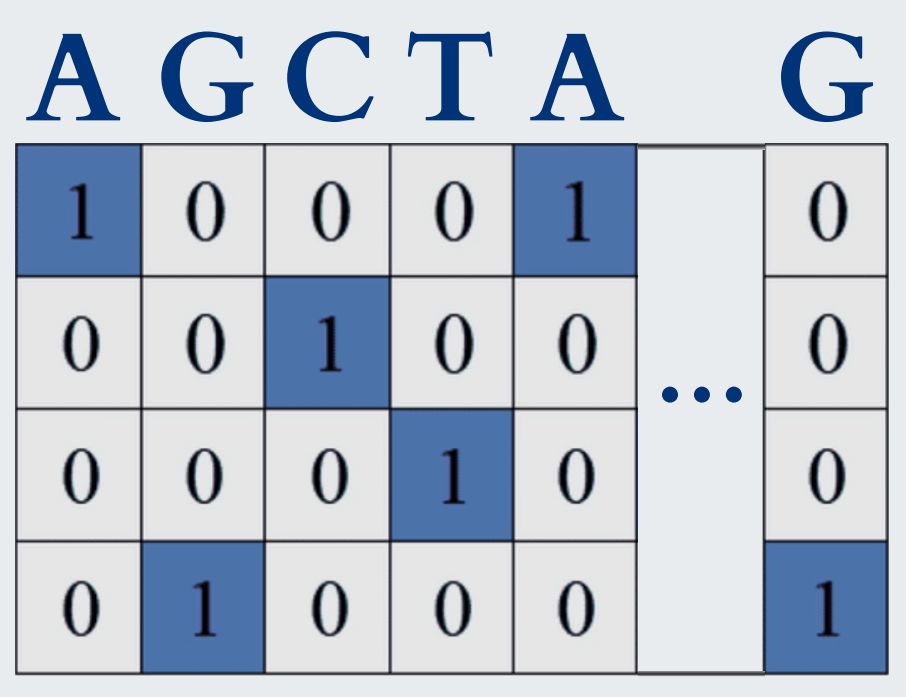
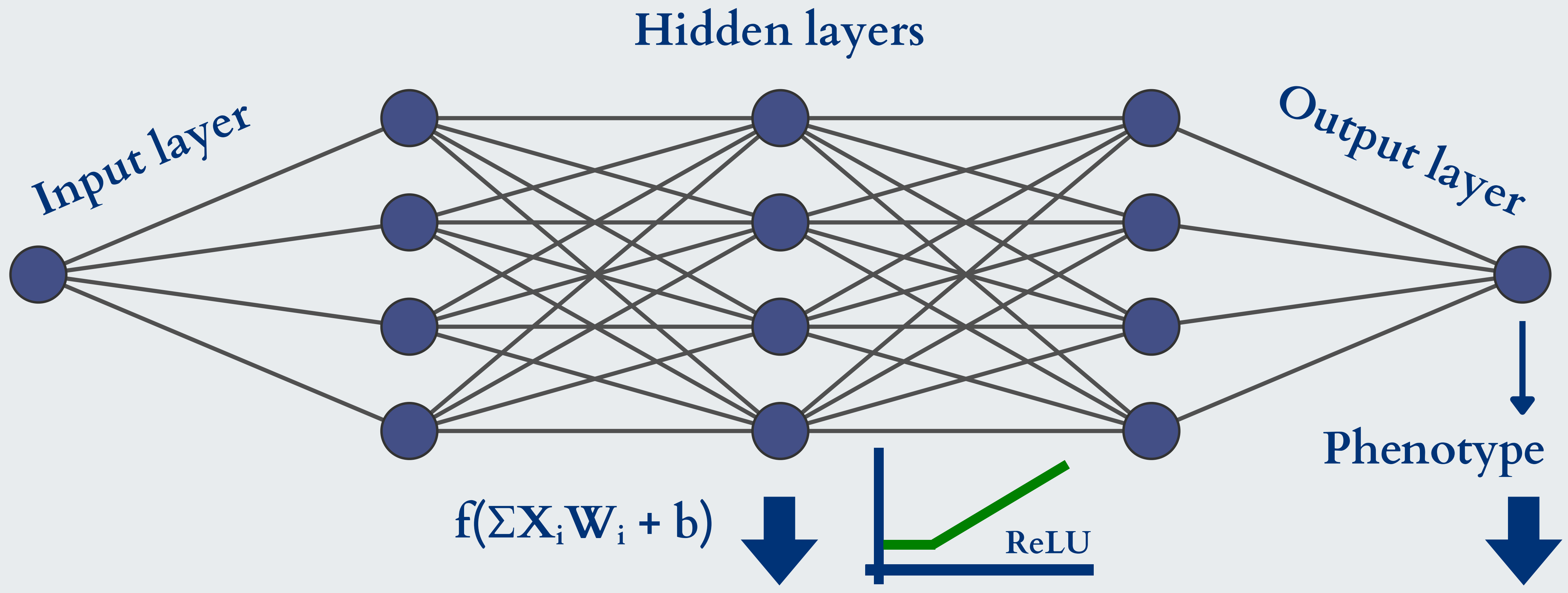


Fig 1. One-hot encoded sequence. Modified from Liu et al. (2022).



Convolutional NNs

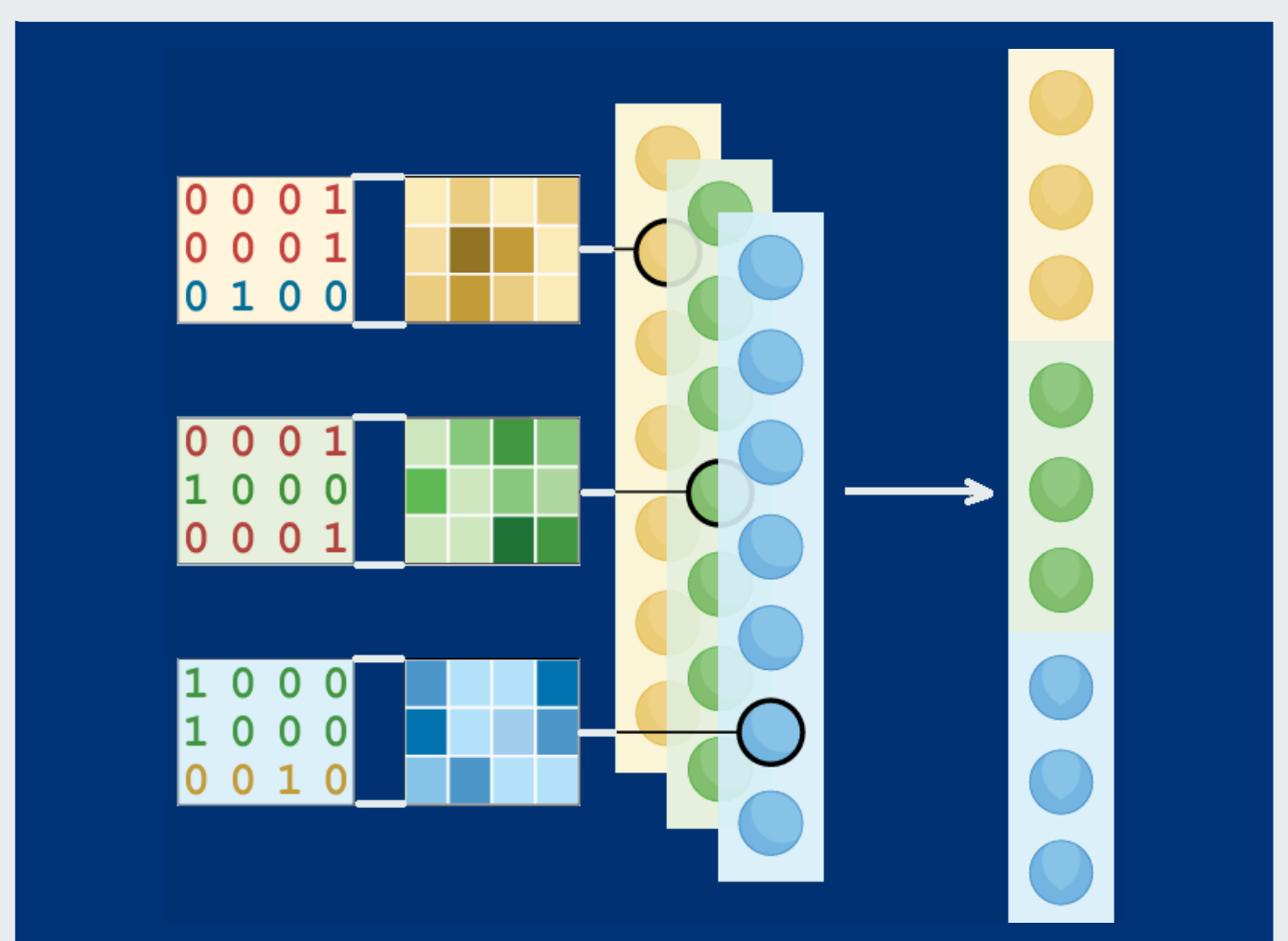


Fig 5. CNN architecture. Initial filters generate activation maps. Posterior pooling (parameter reduction) and flatting (information coupling). Modified from Novakovski et al. (2023).

Data curation

Genomic data usually contains **biases** that should be minimized, as they **impede correct training**.

↓

Distributional differences between genotypes, phenotypes or subsets.

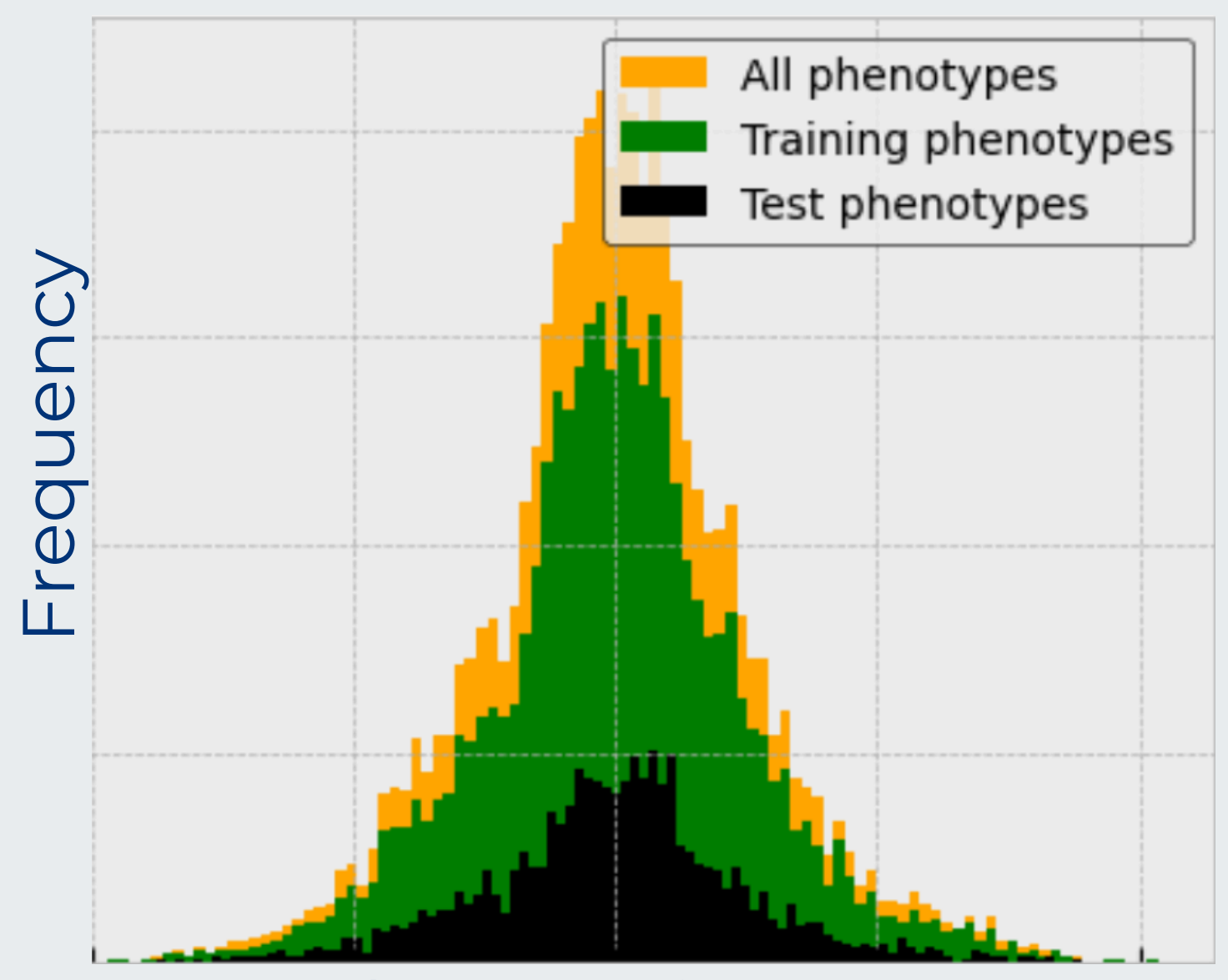


Fig 2. Distribution of phenotypic values through subdatasets.

Correlated samples

AGCTAAG
AGCTAAG
ACCTAAG

Familiar relation?

Whalen et al. (2022)

Model training

Training subset → Model preditcts a **phenotype** (Y_i') for each **genotype** (X_i). ← **Validation** subset

back-propagation

Parameter update

Calculates the **error** with the **Loss function** → **L(Y_i, Y_i')**.

Performance in the training set.

Performance in the validation set.

Evaluation

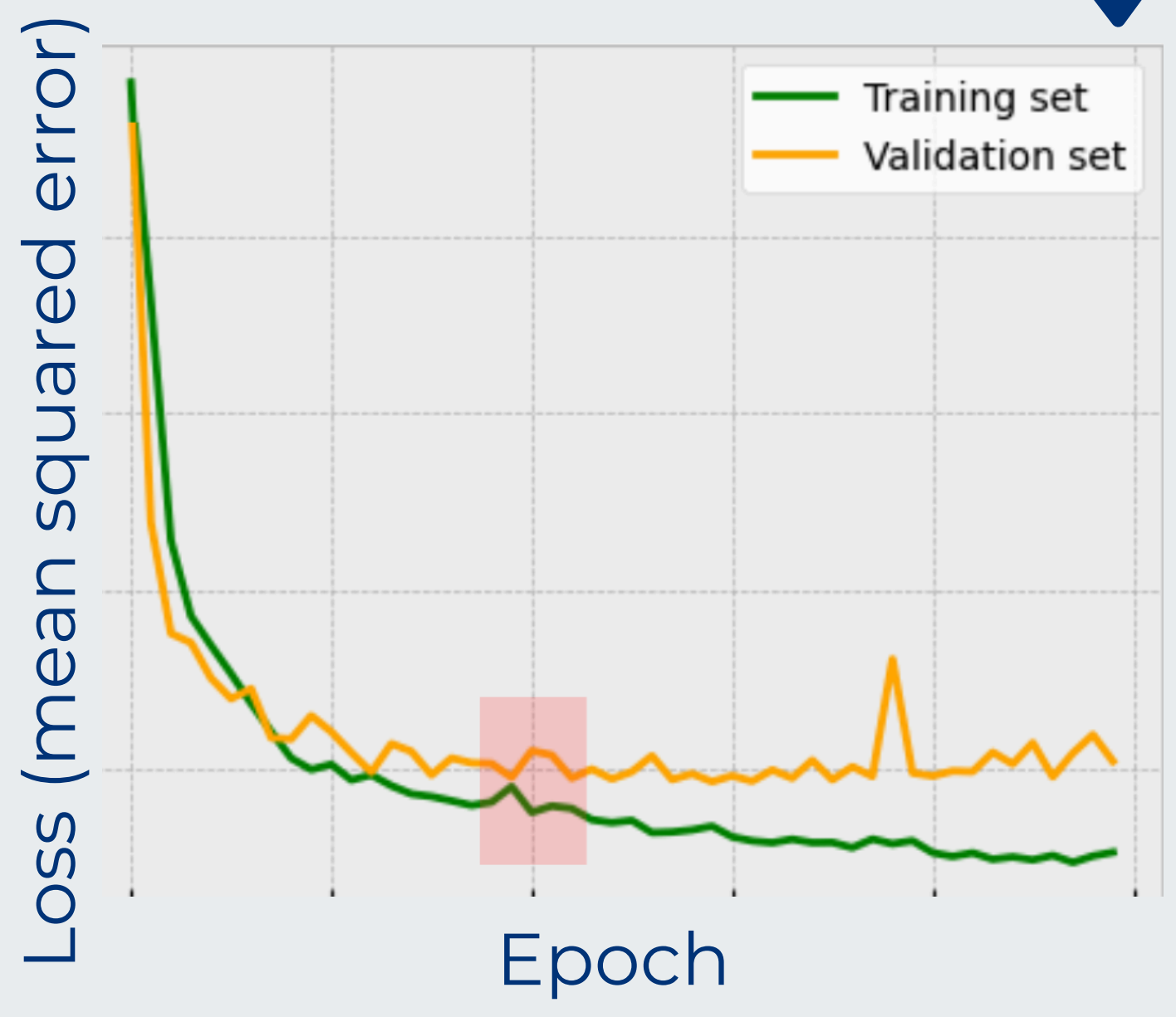


Fig 3. Loss change in subsets during training. Overfitting marked with a red square.

⚠ **Overfitting**: too many iterations can cause the model to **excessively adapt** its predictions to the **training set**.

Hyperparameters = elements decided by the modeller. → **AutoML** to perform **hyperparameter tuning**.

Zou et al. (2019)

Interpretability

DL is considered a "**Black-box**"... → explainable AI (**xAI**)

↓

Relative **importance** of dataset/model's **elements** to **prediction**-making.

Back-propagation

Perturbation-based methods.

Self-attention (epistasis)

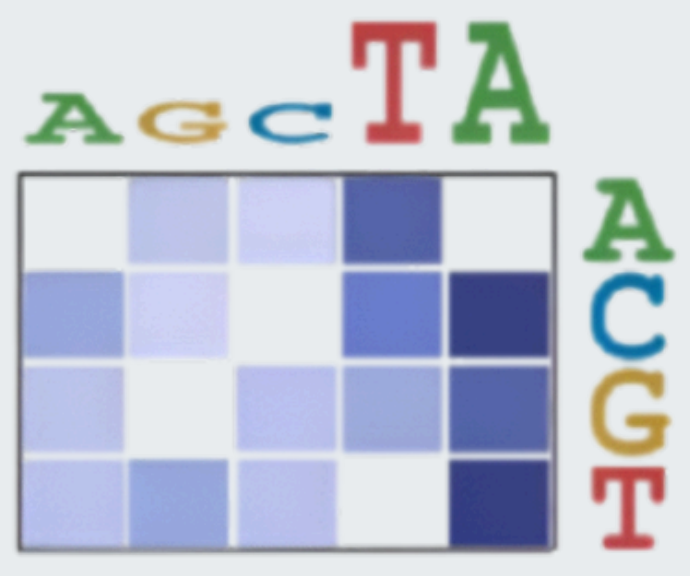


Fig 4. Attribution map. Modified from Novakovski et al. (2023)

Others:

- **Position Weighted Matrices** (**PWMs**) → motif detection.
- **Visible Neural Networks** (**VNNs**): assign genes/pathways to nodes/layers.

Novakovski et al. (2023)

DOWNSIDE

⚠

- Data is not representative of all populations.
- Models reach local extrema, not optima.
- Interpretability requires validation.
- No-free-lunch theorem → each task needs its own model → domain expertise required.

CONCLUSIONS & FUTURE RESEARCH

- **Descriptive objectives accomplished** and **contextualized in genomics prediction**: addressed architecture, training and overfitting, interpretability, data curation...
- DL is a **useful and promising tool**, but it is **not the perfect solution** for every task: ML outperforms for some problems, skewed data, sub-optimal models, difficult interpretability, needed domain knowledge...
- **In the future**, more **comparisons** between models' elements are expected **to reduce search space** in model making, as well as more **user-friendly tools** to enable the field's growth.

Simple **model** available here to **put into practice** all this theory!

MAIN REFERENCES

1. Liu, X., Xu, Y., Luo, Y., & Teng, L. (2022). Prokaryotic and eukaryotic promoters identification based on residual network transfer learning. *Bioprocess and Biosystems Engineering*, *45*(5), 955–967. https://doi.org/10.1007/s00449-022-02716-w
2. Novakovsky, G., Dexter, N., Libbrecht, M. W., Wasserman, W. W., & Mostafavi, S. (2023). Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nature Reviews Genetics*, 24(2), 125–137. https://doi.org/10.1038/s41576-022-00532-2
3. Whalen, S., Schreiber, J., Noble, W. S., & Pollard, K. S. (2022). Navigating the pitfalls of applying machine learning in genomics. *Nature Reviews Genetics*, 23(3), 169–181. https://doi.org/10.1038/s41576-021-00434-9
4. Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., & Telenti, A. (2019). A primer on deep learning in genomics. *Nature Genetics*, 51(1), 12–18. https://doi.org/10.1038/s41588-018-0295-5

UAB Universitat Autònoma de Barcelona