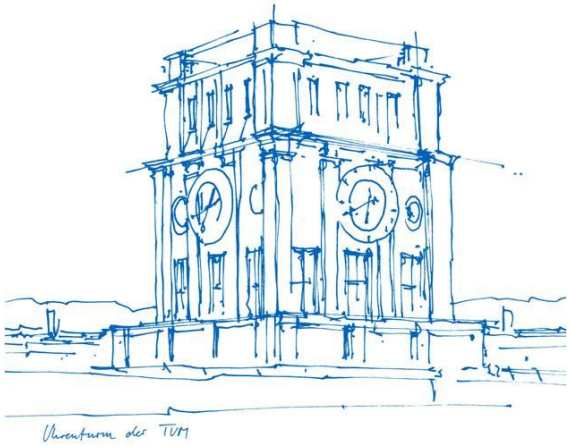
DeepSeries: Genomic Deep Learning Model for Allelic Series

Supervisors: Paolo Casale, Antonio Nappi

Members: Xi Chen, Hanane Mohaouchane, Bálint András Lassú

08.02.2024 Munich

HELMHOLTZ MUNICH



Motivation

- **Burden test:**

  Pooling variants ⟶ aggregate

  Assumption: every effect in one direction

$$\alpha + \sum_{j=1}^{J} G_j \beta_j + X'\gamma$$

*α :Intercept, **G**:Allele count, **X'**: Covariates*
β, γ : coefficients / weights

[4]

- **Allelic Series:** collection of variants in a gene with different effects on gene function

  ↑ deleteriousness = ↑ phenotypic effect

  ⟶ Therapeutic Interest

- **Problem:** Predicting deleteriousness scores for rare variants

  ⟶ Statistically challenging

Other approaches

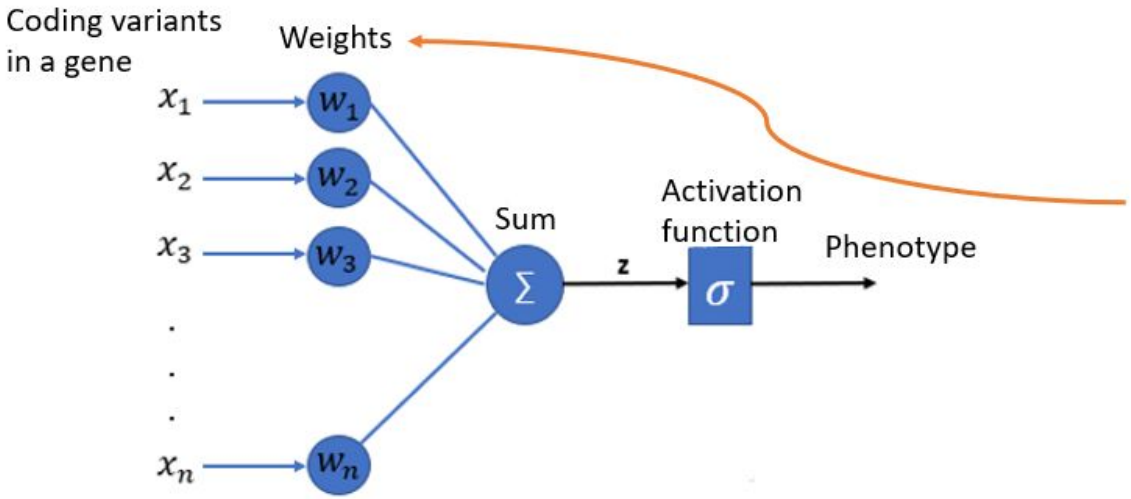
- **Sequence Kernel Association tests (SKAT):**

  Assumption: Variant effects are random based on a distribution (conflicts allowed)

- **SKAT-O:** Combine burden test and SKAT adaptively
- **COAST:** Designed specifically for allelic series identification
- **DeepRVAT:** data driven, neural network based

  - **Annotations**: CADD, PrimateAI, AlphaMissense…

Our Approach



Deleteriousness scores

$$f\left(\begin{array}{c} \text{Annotations} \\ \text{Coding Variants} \begin{bmatrix} 1 & 0 & 0.5 & 0.4 \\ 0 & 0 & 0.6 & 0.2 \\ 1 & 1 & 0.1 & 0.2 \\ 1 & 0 & 0.2 & 0.7 \end{bmatrix} \end{array}, \theta\right)$$

→ **Learn function parameters θ through backpropagation**

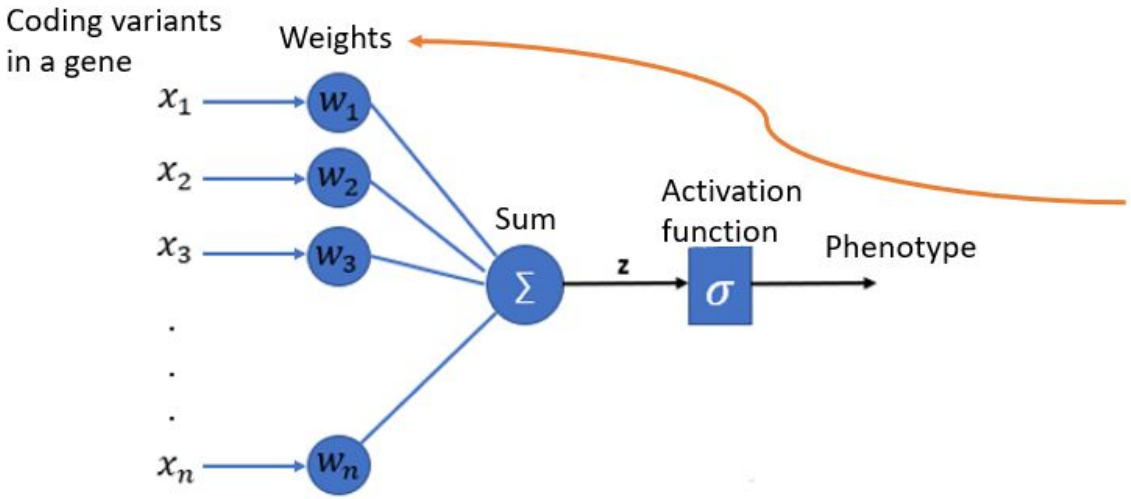
Experimental Design

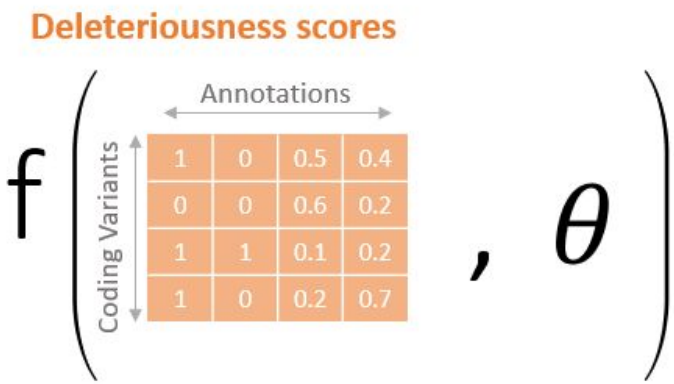
- Genetic data simulation:
  - Using realistic allelic data based on UK Biobank
    - minor allele frequency (MAF) < 1%
  - Annotations from Genebass (Karczewski et al., 2022)
    - Combined Annotation-Dependent Depletion (CADD) (Schubach et al., 2024): pathogenic score from multiple sources
    - AlphaMissense (Cheng et al., 2023): pathogenic score of missense variants
    - PrimateAI (Sundaram et al., 2018): describe rare variant effects based on comparisons between human and other primates.

Experimental Design

- Simulations: phenotype = burdenScore + noise
  - Simulate burden score from coding variants
  - Simulate phenotype from burden score
- Fit model on 50% data
- Predict on the remaining
- Test association with phenotype

Two Models to Calculate Weights





→ **Learn function parameters θ through backpropagation**

Simple Model → Linear model with 3 parameters

- $weights_{simple\_model} = f(varientType, \Theta)$

$$= \begin{array}{c} \begin{array}{ccc} \text{pLof} & \text{missense} & \text{synonymous} \end{array} \\ \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \end{array} \times \begin{array}{c} \text{theta} \\ \begin{bmatrix} \theta_{pLof} \\ \theta_{missense} \\ \theta_{synonymous} \end{bmatrix} \end{array}$$

Complex model

- $\text{weights}_{\text{complex\_model}} = \text{weights}_{\text{simple\_model}} \cdot$

$$f(AF, 1\text{-}AF, h(CADD, primateAI, alphaMissense), \Theta_2)$$

- $h(x) = \begin{cases} 1, \text{ if } x > 90\% \text{ thresh} \\ 0, \text{ others} \end{cases}$

CADD	alphaMissense	primateAI
1.9	0	0
0.5	0	0.5
1.3	0.1	0
0.7	0.91	0.3

h →

CADD	alphaMissense	primateAI
1	0	0
0	0	1
0	0	0
0	1	0

Complex model → **Non-linear model with 8 parameters**

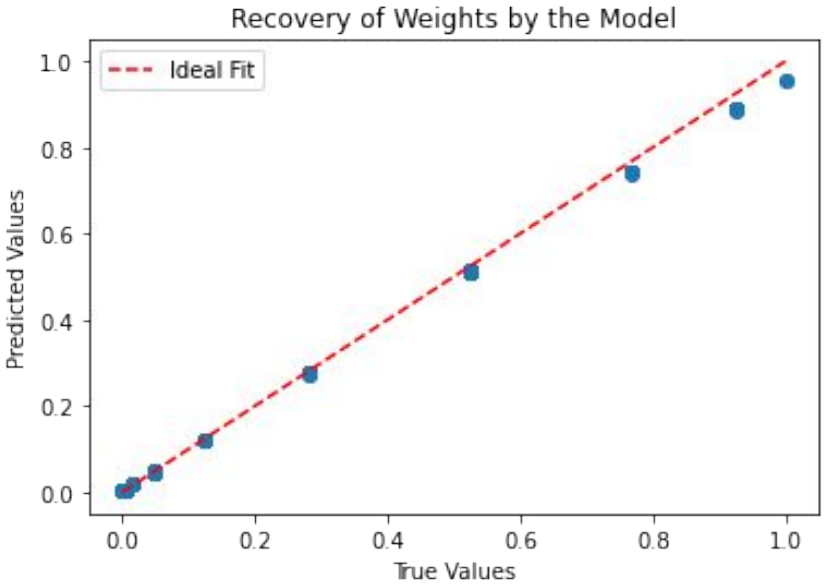
- $\text{weights}_{\text{complex\_model}} = \text{weights}_{\text{simple\_model}} \cdot f(AF, 1\text{-}AF, h(CADD, primateAI, alphaMissense), \Theta_2)$

$$= \begin{array}{c} \text{Weight} \\ \begin{bmatrix} 0.23 \\ 0.81 \\ 1.35 \\ 0.28 \end{bmatrix} \end{array} \cdot \left(\begin{array}{c} \begin{array}{ccccc} \text{AF} & \text{1-AF} & \text{CADD} & \text{alphaMissense} & \text{primateAI} \end{array} \\ \begin{bmatrix} 0.03 & 0.97 & 1 & 0 & 0 \\ 0.01 & 0.99 & 0 & 0 & 1 \\ 0.04 & 0.96 & 0 & 0 & 0 \\ 0.32 & 0.68 & 0 & 1 & 0 \end{bmatrix} \end{array} \times \begin{array}{c} \text{theta} \\ \begin{bmatrix} \theta_{AF} \\ \theta_{1\text{-}AF} \\ \theta_{CADD} \\ \theta_{alphaMissense} \\ \theta_{primateAI} \end{bmatrix} \end{array} \right)$$

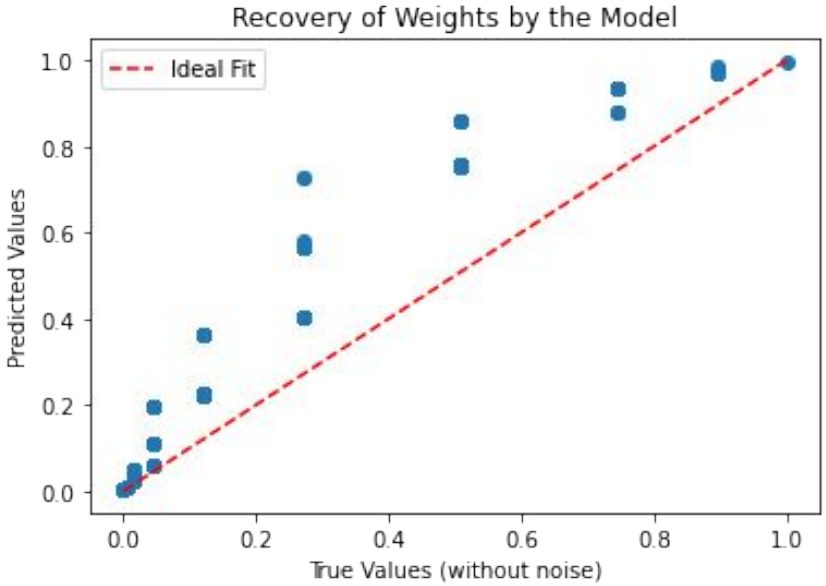
Results

Simple model recovers simulated weights for variant annotations

➔ 0% noise



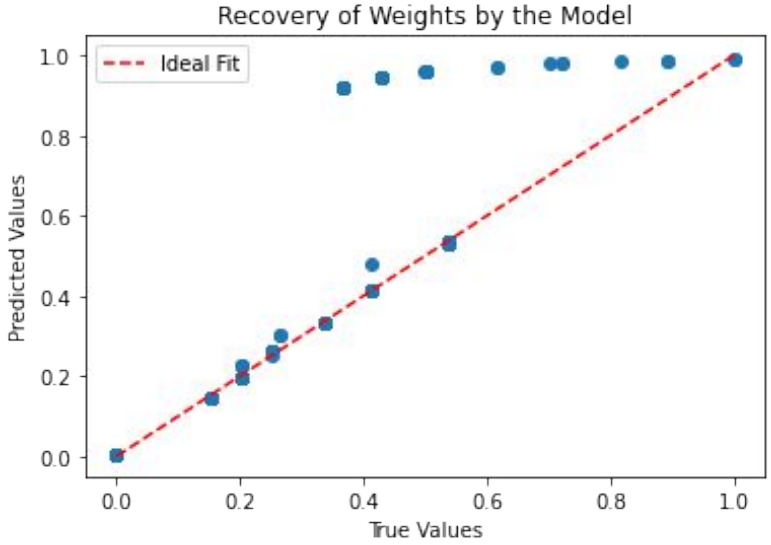
➔ 50% noise



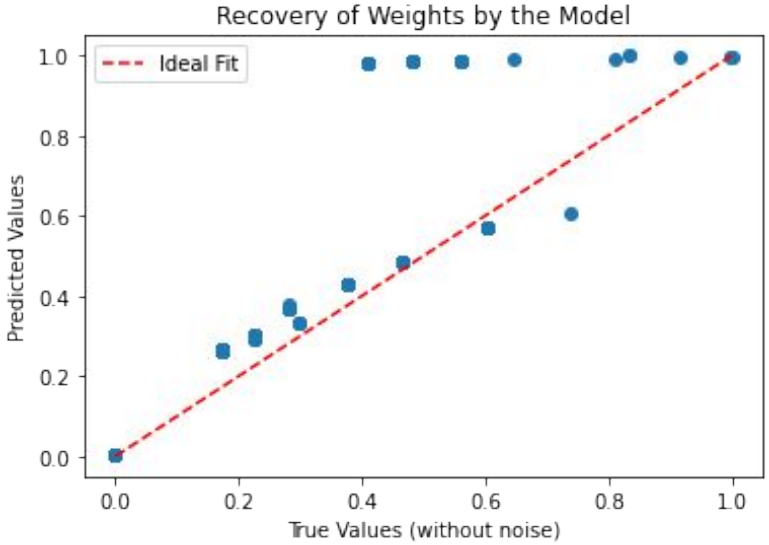
Complex model learns alternative weights for variant annotations

- Doesn't recover exact weights
- 0.99 correlation with denoised data

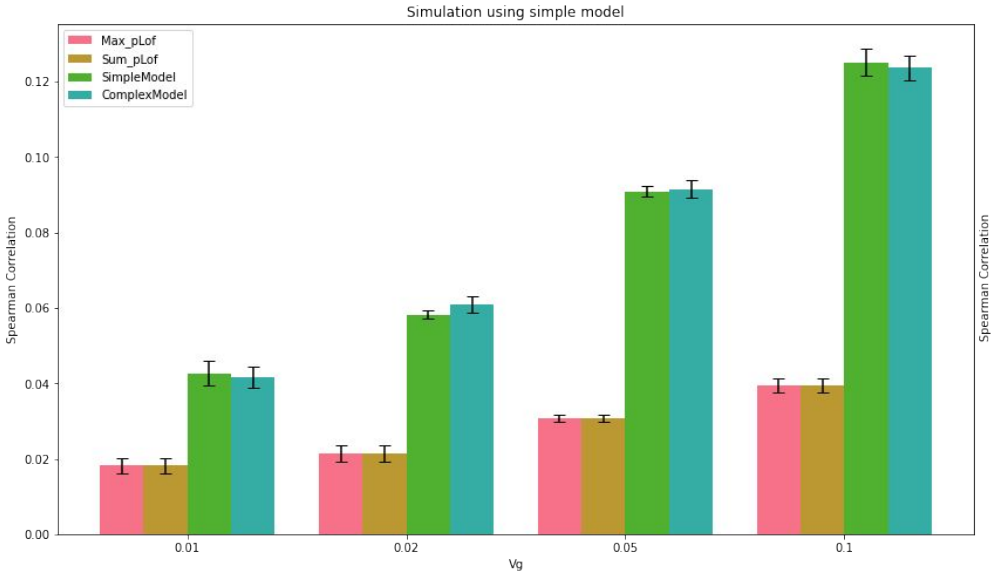
➔ 0% noise

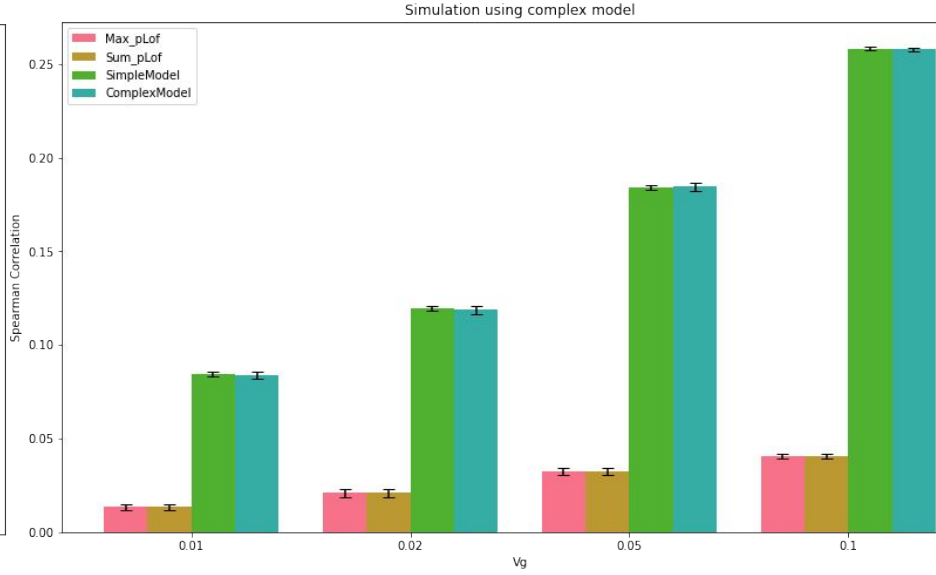


➔ 50% noise

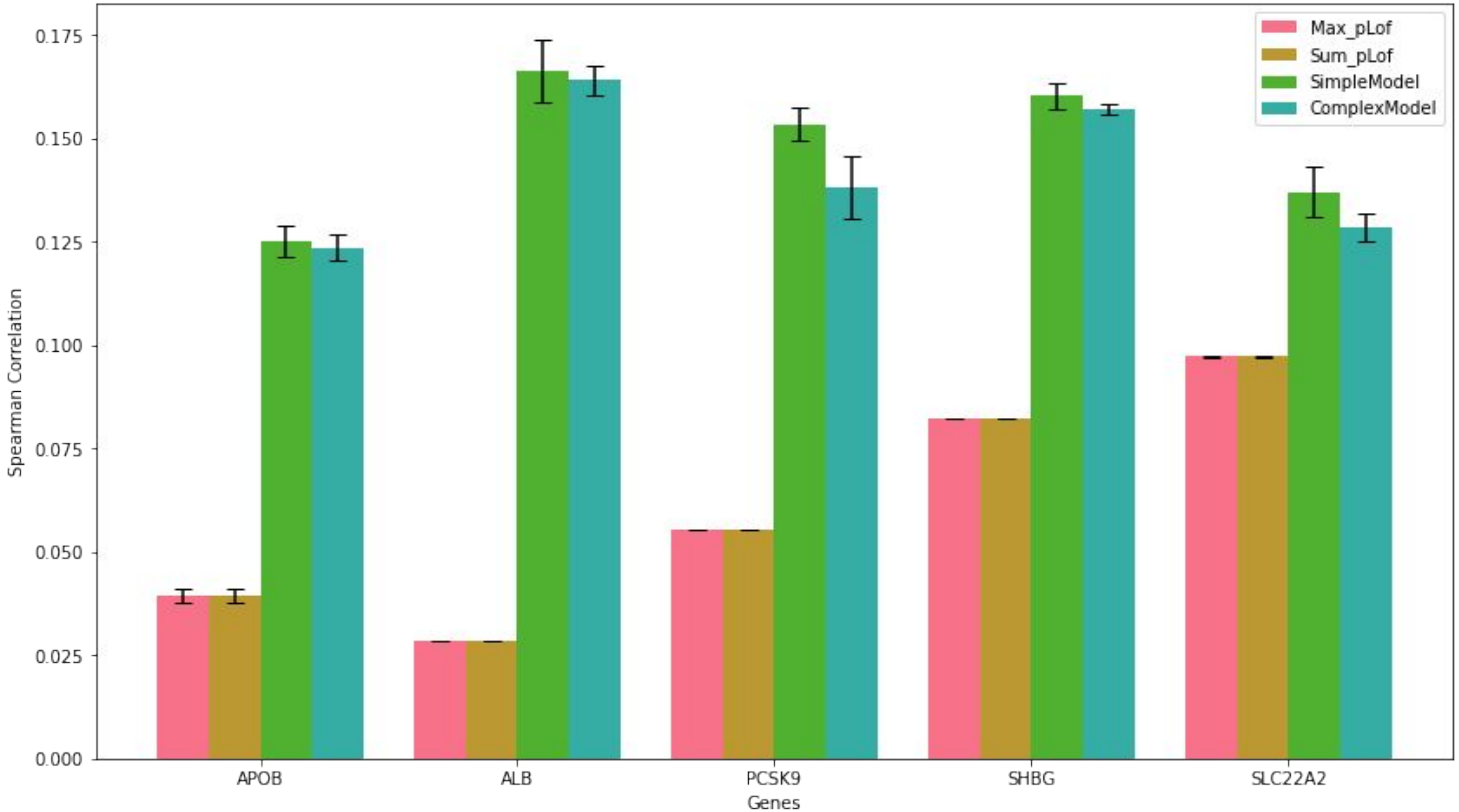


Simple and complex models outperform conventional burden tests in different simulation scenarios





Simple and complex models outperform conventional burden tests for different genes



Conclusion

- Predicting deleteriousness scores for rare variants is statistically challenging
- Our approach:
  - Calculate scores as a function of the variant annotations
  - Use gradient descent to learn the parameters of this function
- Both the simple and complex model outperform conventional burden tests for:
  - Different simulation scenarios
  - Different noise levels
  - Different genes
- Our models can learn gene-level burden scores in an interpretable manner

Reference

- [1] Karczewski, Konrad J., et al. "Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes." *Cell Genomics* 2.9 (2022).
- [2] M. Schubach, T. Maass, L. Nazaretyan, S. Röner, and M. Kircher, ‘CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions’, *Nucleic Acids Research*, vol. 52, no. D1, pp. D1143–D1154, Jan. 2024, doi: 10.1093/nar/gkad989.
- [3] J. Cheng *et al.*, ‘Accurate proteome-wide missense variant effect prediction with AlphaMissense’, *Science*, vol. 381, no. 6664, p. eadg7492, Sep. 2023, doi: 10.1126/science.adg7492.
- [4] L. Sundaram *et al.*, ‘Predicting the clinical impact of human mutation with deep neural networks’, *Nat Genet*, vol. 50, no. 8, pp. 1161–1170, Aug. 2018, doi: 10.1038/s41588-018-0167-z.
- [5] Z. R. McCaw *et al.*, ‘An allelic-series rare-variant association test for candidate-gene discovery’, *The American Journal of Human Genetics*, vol. 110, no. 8, pp. 1330–1342, Aug. 2023, doi: 10.1016/j.ajhg.2023.07.001.