RELATION

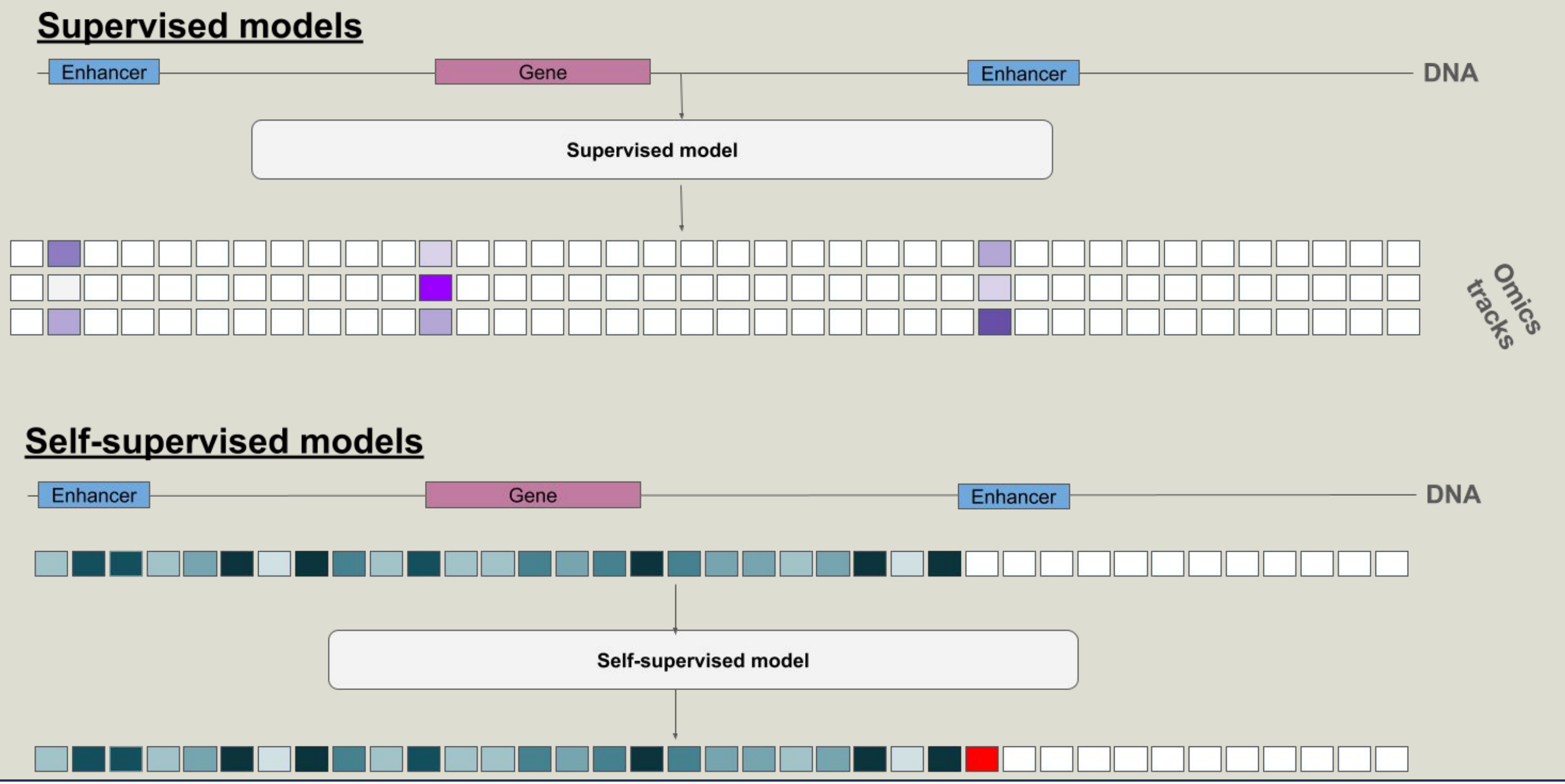
The patient is waiting

Gene-centric evaluation of causal variant prediction for DNA models

Chantriolnt-Andreas Kapourani, Alice Del Vecchio, Agnieszka Dobrowolska, Andrew Anighoro, Edith M. Hessel, Lindsay Edwards, Cristian Regep
Relation, Regent's Place, 338 Euston Road, London NW1 3BG

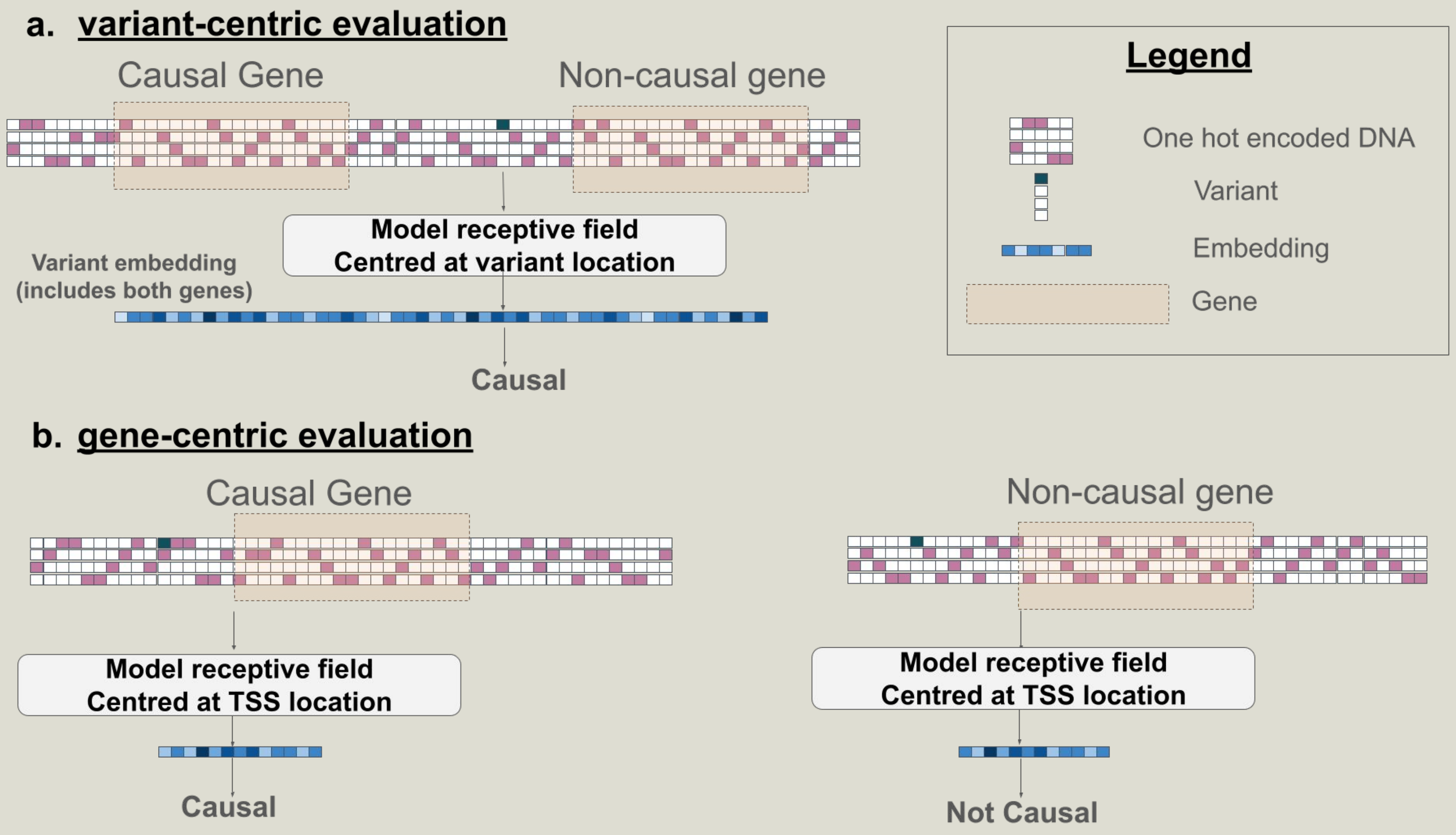
Background

- A new era of DNA models has been ushered in recently [1,3,4,5], with a particular emphasis on self-supervised models trained without the use of omics for supervision [3,4,5].
- Recent benchmarking shows that embeddings from self-supervised models can be effective in causal variant prediction [2].
- Linking non-coding variants to effector genes can lead to identifying mechanisms that drive disease but also enable discovery of novel drug targets.
- Traditional benchmarking [1,2] falls short of evaluating downstream effects and rather focuses at the variant level.



Gene-centric vs variant-centric evaluation

- In the variant-centric approach the model is evaluated on the embedding of the entire sequence.
- In a gene-centric approach the model is evaluated on a short embedding centered on the TSS of the gene.



Variant-centric benchmarking

- Causal variant prediction involves training a predictive model on top of embeddings of the reference and alternate sequences to predict whether the variant is causal.
- A variant-centric benchmark was proposed by [2], where they adapted the SuSiE finemapped variant-gene pair dataset based on GTEx first used in the Enformer model [1].
- Whether a variant can alter any gene is in large part dependent on local effect on a short regulatory sequence (e.g. the binding site of a transcription factor).
- We validated this by training a Basic CNN model using as input a one-hot encoded sequence and a much smaller receptive field compared to competing methods.

Model	Receptive field	AUC
Basic CNN	1.5 kbp	0.695
*HyenaDNA [3]	131 kbp	0.706
*Nucleotide Transformer[5]	12 kbp	0.722
*Nucleotide Transformer NTK [2]	192 kbp	0.749
*Enformer [1]	196 kbp	0.755

* Results taken from [2]

Gene-centric benchmarking

- We extended the variant-centric dataset [2] by adding examples of causal and non-causal genes.
- Reference and alternate embeddings from major models [1,3,4] were extracted as 384bp around the TSS: 3 bins for Enformer and 384 bins for HyenaDNA and Caduceus.
- Logged absolute difference of the reference and alternate embeddings was used as input to an MLP.
- The MLP was trained using a binary cross-entropy objective function to predict whether a variant-gene pair is causal or non-causal.
- Self-supervised models produce useful embeddings for this task, although a larger gap can be observed.

Model	HyenaDNA [3]	Caduceus [4]	Enformer [1]
Training receptive field	160 kbp	131 kbp	196 kbp
Inference receptive field	131 kbp	131 kbp	131 kbp
TSS embedding span	384 bp	384 bp	384 bp
AUC	0.67	0.703	0.764

Conclusions

- In the variant-centric benchmark, a simple CNN model can achieve performance close to state-of-the-art.
- Despite the fact that self-supervised models have not been trained to predict gene expression at the TSS, their embeddings can be used for causal gene prediction.
- We propose that future benchmarks should incorporate gene-centric evaluation, which is often of higher biological and drug discovery significance.

References

[1]. Avsec, Žiga, et al. "Effective gene expression prediction from sequence by integrating long-range interactions." Nature methods 18.10 (2021): 1196-1203.
[2] Kao, Chia Hsiang, et al. "Advancing dna language models: The genomics long-range benchmark." ICLR 2024 Workshop on Machine Learning for Genomics Explorations. 2024.
[3] Nguyen, Eric, et al. "Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution." Advances in neural information processing systems 36 (2024).
[4] Schiff, Yair, et al. "Caduceus: Bi-directional equivariant long-range dna sequence modeling." arXiv preprint arXiv:2403.03234 (2024).
[5] Dalla-Torre, Hugo, et al. "The nucleotide transformer: Building and evaluating robust foundation models for human genomics." BioRxiv (2023): 2023-01.