Variant effect prediction

https://doi.org/10.1038/s41588-023-01517-5

# Current approaches to genomic deep learning struggle to fully capture human genetic variation

Ziqi Tang, Shushan Toneyan & Peter K. Koo



Deep learning shows promise for predicting gene expression levels from DNA sequences. However, recent studies show that current state-of-the-art models struggle to accurately characterize expression variation from personal genomes, limiting their usefulness in personalized medicine.

Understanding the effect of genetic variability, particularly single nucleotide variants (SNVs), on phenotypic variability in health and disease is a long-standing goal in biology. Achieving this understanding would, for example, enable the tailoring of treatments based on patients' genetic makeup, taking into account their varying responses to certain drugs based on disease mechanisms. However, this task is complicated by the prevalence of non-coding genomic variation, which account for approximately 95% of all mutations[1]. The complexity arises from the intricate coordination of *cis*-regulatory elements and their influence on gene expression. In addition, linkage disequilibrium adds further complexity, making it difficult to precisely identify causal variants associated with specific phenotypes within haplotypes.

Deep learning has emerged over the past few years, demonstrating remarkable success in accurately mapping genotypes to molecular phenotypes. Deep learning is based on artificial neural networks, composed of sequential computational layers that are used to construct hierarchical representations of the data. Unlike conventional machine learning algorithms that require expert-crafted features as inputs, deep learning autonomously learns and extracts relevant features directly from the data, enabling end-to-end predictive modeling. A large-scale deep learning model, called Enformer[2], has recently showcased a powerful capability to predict gene expression, chromatin accessibility, histone marks, and transcription factor-binding sites across various tissues and cell lines using DNA sequences from the reference genome. Enformer represents a promising genomic deep learning model, owing to its ability to learn dependencies across distal regulatory elements (up to 200 kb).

Accurate predictive models, such as Enformer, serve as valuable surrogates for the experimental assays they were trained with, enabling access to an unprecedented scale of in silico perturbation experiments. For example, Enformer can probe the effects of SNVs on

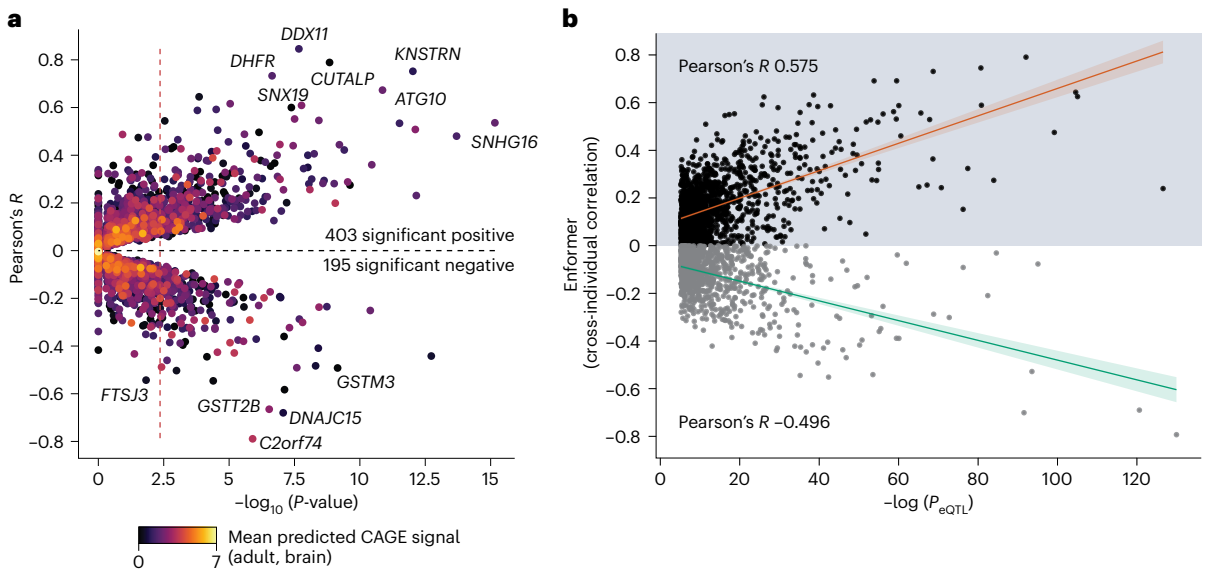


**Fig. 1 | Evaluation of Enformer in predicting individual-specific gene expression. a**,**b**, Scatter plots of cross-individual correlation between Enformer's prediction and observed expression values versus the negative $\log_{10}$-transformed *P* value of the annotated eQTL in studies by Sasse et al.[5] (**a**) and Huang et al.[6] (**b**). In **a**, the red dotted line denotes $P = 0.05$, and color indicates and Enformer's predicted mean expression for the 'CAGE, adult, brain' output track. In **b**, the lines represent the best linear fit for genes with either positive or negative cross-individual correlations. In both experiments, although the magnitude of expression could be reasonably accurately captured, the direction of effect could be reversed.

gene expression. However, to ensure the reliability of such models, they undergo extensive testing to validate the generalizability of their predictions[3,4]. Validation tests measure the correspondence between predictions with existing experimental data, spanning saturation mutagenesis within *cis*-regulatory elements via massively parallel reporter assays and enhancer–promoter interactions from CRISPR interference experiments. Although these approaches are informative, they only probe a narrow scope of what deep learning models have learned. Therefore, understanding the strengths and limitations of deep learning models such as Enformer remains an ongoing endeavor.

In two recent studies in *Nature Genetics*, Sasse et al.[5] and Huang et al.[6] investigated the effectiveness of state-of-the-art genomic deep learning models in predicting expression variation among individuals, a crucial prerequisite for their application in precision medicine and elsewhere. Sasse et al.[5] evaluated the performance of Enformer in predicting individual-specific gene expression (such as expression quantitative trait loci, eQTLs) using paired whole-genome sequencing and RNA-sequencing data from the cerebral cortex 839 individuals, specifically from the ROSMAP dataset[7]. In parallel, Huang et al.[6] expanded the comparison by incorporating three additional deep learning models – Basenji2[8], Expecto[9] and Xpresso[10]. They conducted similar experiments using expression data from the Geuvadis consortium[11], which consists of phased whole-genome sequencing and RNA-sequencing data from lymphoblastoid cell lines of 421 individuals.

Both studies conducted an initial assessment of the ability of Enformer to predict population-average gene expression using reference genome sequences as inputs. Their findings aligned with the original study, showing that Enformer achieved high prediction performance. Moreover, Huang et al.[6] demonstrated that when supplied with personal DNA as the input, Enformer's predictions, along with other deep learning models, displayed robust cross-gene correlation within each individual. These preliminary analyses highlight the proficiency of state-of-the-art deep learning models in predicting gene expression levels from sequences.

However, when evaluating the predicted expression for each gene across individuals, both studies made a surprising observation: subsets of genes exhibited strong positive correlations, whereas others displayed strong negative correlations with the observed expression (Fig. 1). This unexpected finding indicates that Enformer (and other deep learning models) face challenges in accurately identifying whether an SNV will have a positive or negative effect on gene expression, although it appeared to capture the magnitude reasonably well. In addition, both studies used a regularized linear model, similar to the transcriptome-wide association approach of PrediXcan[12], trained directly on personal genomes to establish a baseline for comparison. Even though a direct comparison cannot be made as training data was different, the PrediXcan-like model largely outperformed Enformer, indicating that common *cis*-regulatory variants are not fully captured by current deep learning models.

To gain deeper insights into the potential causes of the seemingly random direction of SNV prediction effects, both authors conducted further experiments. Huang et al.[6] explored various factors, including the distance of driver SNVs to transcription start sites, eQTL effect size, and minor allele frequency. Although some of these factors could account for the magnitude of eQTL effect sizes, they could not fully explain the direction of the effects. Interestingly, Huang et al.[6] also observed that the subset of genes with poor performance was not consistent across different deep learning models, suggesting that the random directions might be attributed to modeling noise rather than an inherent *cis*-mechanism. By contrast, Sasse et al.[5] used model interpretability to show that driver SNVs were often observed to fall outside the learned motifs. However, this analysis was not sufficiently exhaustive to draw definitive conclusions. Thus, the reasons why current genomic deep learning models struggle to predict the direction of eQTL effects remain an open question that requires further investigation.

In summary, the current paradigm of training deep learning models relying solely on the reference genome has been successful in learning 'local' molecular phenotypes, such as transcription factor binding and chromatin-accessibility sites. However, concerns about poor generalization to population-level genetic variation have been highlighted in studies by Sasse et al.[5] and Huang et al.[6]. These studies have a crucial role in uncovering the shortcomings of deep learning models that go beyond performance benchmarks. Consequently, these findings now serve as essential targets for future research, urging the need for innovative advancements to better account for population-level genetic variation.

As a possible solution, both studies propose training deep learning models on a more diverse set of genomes, exposing them to population-level genetic variation. This approach could help the models to learn a more comprehensive set of regulatory features that are missed when training only on a single reference genome. Another limitation is a lack of consideration for post-transcriptional regulation and gene regulatory networks. Incorporating these as biological priors or as part of a more comprehensive multi-modal training curriculum requires a paradigm shift away from pure sequence-based deep learning. However, such an adjustment is probably crucial in effectively addressing complex traits such as eQTLs. Furthermore, these data augmentations and the integration of biological priors could have the added benefit of improving explanations of *cis*-mechanisms from model interpretability analyses.

**Ziqi Tang, Shushan Toneyan & Peter K. Koo** ✉
Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA.
✉e-mail: koo@cshl.edu

Published online: 30 November 2023

## References

1. Leslie, R. et al. *Bioinformatics* **30**, i185–i194 (2014).
2. Avsec, Ž. et al. *Nat. Methods.* **18**, 1196–1203 (2021).
3. Toneyan, S. et al. *Nat. Mach. Intell.* **4**, 1088–1100 (2022).
4. Karollus, A. et al. *Genome Biol.* **24**, 56 (2023).
5. Sasse, A. et al. *Nat. Genet.* https://doi.org/10.1038/s41588-023-01524-6 (2023).
6. Huang, C. et al. *Nat. Genet.* https://doi.org/10.1038/s41588-023-01574-w (2023).
7. Bennett, D. A. et al. *J. Alzheimers Dis.* **64**, S161–S189 (2018).
8. Kelley, D. R. et al. *Genome Res.* **28**, 739–750 (2018).
9. Zhou, J. et al. *Nat. Genet.* **50**, 1171–1179 (2018).
10. Agarwal, V. & Shendure, J. *Cell Rep.* **31**, 107663 (2020).
11. Lappalainen, T. et al. *Nature* **501**, 506–511 (2013).
12. Amazon, E. R. et al. *Nat. Genet.* **47**, 1091–1098 (2015).

## Competing interest

The authors declare no competing interests.