Benchmarking DNA Sequence Models for Causal Regulatory Variant Prediction in Human Genetics

Gonzalo Benegas[1], Gökcen Eraslan[2,*], Yun S. Song[1,3,4,*]

[1]Computer Science Division, University of California, Berkeley
[2]Biology Research | AI Development, gRED Computational Sciences, Genentech
[3]Department of Statistics, University of California, Berkeley
[4]Center for Computational Biology, University of California, Berkeley

March 4, 2025

Abstract

Machine learning holds immense promise in biology, particularly for the challenging task of identifying causal variants for Mendelian and complex traits. Two primary approaches have emerged for this task: supervised sequence-to-function models trained on functional genomics experimental data and self-supervised DNA language models that learn evolutionary constraints on sequences. However, the field currently lacks consistently curated datasets with accurate labels, especially for non-coding variants, that are necessary to comprehensively benchmark these models and advance the field. In this work, we present TraitGym, a curated dataset of regulatory genetic variants that are either known to be causal or are strong candidates across 113 Mendelian and 83 complex traits, along with carefully constructed control variants. We frame the causal variant prediction task as a binary classification problem and benchmark various models, including functional-genomics-supervised models, self-supervised models, models that combine machine learning predictions with curated annotation features, and ensembles of these. Our results provide insights into the capabilities and limitations of different approaches for predicting the functional consequences of non-coding genetic variants. We find that alignment-based models CADD and GPN-MSA compare favorably for Mendelian traits and complex disease traits, while functional-genomics-supervised models Enformer and Borzoi perform better for complex non-disease traits. Evo2 shows substantial performance gains with scale, but still lags somewhat behind alignment-based models, struggling particularly with enhancer variants. The benchmark, including a Google Colab notebook to evaluate a model in a few minutes, is available at https://huggingface.co/datasets/songlab/TraitGym.

1 Introduction

Machine learning is increasingly transforming the fields of genomics, human genetics, and healthcare by offering new avenues for predicting the impact of genetic variants on phenotypes and by potentially improving the accuracy of trait or disease risk predictions from individual human genomes. A major challenge in these domains is determining which among millions of intercorrelated genetic variants are causal for Mendelian and complex traits, including diseases. Tackling this challenge, which has profound implications for human health, requires robust and scalable methods that can

[*]To whom correspondence should be addressed: eraslan.gokcen@gene.com, yss@berkeley.edu

decode the biological syntax of the human genome and how it drives molecular functions across different cells and tissues.

Three major classes of approaches have been developed to model DNA sequences and predict the effects of genetic variants. The first approach utilizes supervised machine learning models, commonly referred to as sequence-to-function models, which are trained to predict genome-wide functional genomics experimental data from DNA sequences (Eraslan et al., 2019); we refer to these models as *functional-genomics-supervised*. These models predict the functional effects of specific variants by assessing how changes in the DNA sequence influence experimental outcomes. The second approach involves *self-supervised* genomic language models (gLMs), such as masked or autoregressive language models, which are trained only on DNA sequences from one or multiple species without relying on experimental data (Benegas et al., 2025b). Models that utilize sequences from multiple species take advantage of evolutionary conservation to gain functional insights. Variant effects in such models are assessed by comparing the log-likelihood between the alternative and reference alleles of the variant, as well as by quantifying changes in the latent representations. Another class of methods includes *integrative* approaches, which combine machine learning predictions with curated annotation features to improve the accuracy of variant effect prediction (Schubach et al., 2024). Additionally, traditional *conservation scores* phastCons (Siepel et al., 2005) and phyloP (Pollard et al., 2010) have been strong predictors of trait-associated variants (Sullivan et al., 2023).

Despite its importance, the field currently lacks consistently processed and comprehensively curated datasets of putative causal regulatory genetic variants with reliable labels. Furthermore, there is a pressing need for establishing a common ground for systematically benchmarking state-of-the-art models based on functional-genomics-supervised, self-supervised and integrative approaches, in order to help advance the field.

In this article, we present TraitGym, a curation of two non-coding variant benchmark datasets from human genetics: one comprising causal variants for 113 Mendelian traits, and another consisting of strong causal variant candidates across 83 complex traits, along with carefully constructed control sets matching relevant summary statistics (such as minor allele frequencies, variant types, distances from transcription start sites, and linkage disequilibrium scores) of putative causal variants. We frame the task as binary classification between putatively causal and non-causal variants, allowing to evaluate several state-of-the-art functional-genomics-supervised and self-supervised models, alongside integrative methods and their ensembles. We find that alignment-based integrative and self-supervised models compare favorably for Mendelian traits and complex disease traits, while functional-genomics-supervised models do better on complex non-disease traits. The classification of variants is substantially harder for complex traits, but consistent improvement is observed by ensembling input and predicted features from different models. Additionally, we introduce a new gLM trained specifically on regulatory regions and demonstrate that it compares favorably with other alignment-free self-supervised language models.

2 Background

One of the essential quests in biology is to understand the genotype-to-phenotype relationship (Figure 1). The genotype is the genetic makeup of an organism, i.e., the set of DNA sequences composing each genome. The phenotype is the collection of observable traits of an individual, such as height or cholesterol levels. Phenotypic variance can be decomposed into components attributed to genetic and environmental factors. The influence of non-coding genetic variants on phenotype is mediated via the expression of genes in different tissues and cell types. Functional-genomics-supervised models attempt to learn the relationship between DNA sequence and gene expression,

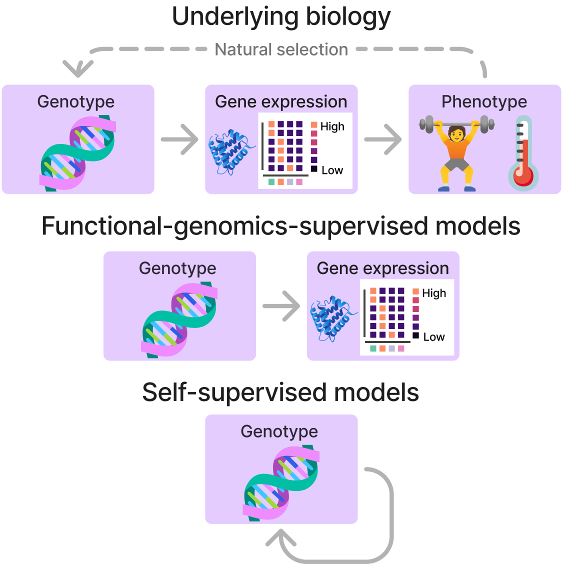

Figure 1: Genotype-to-phenotype relationship and general ML approaches for prediction.

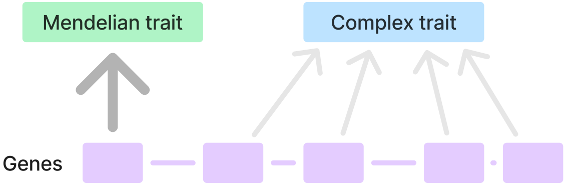

Figure 2: Mendelian vs. complex traits. A single gene typically controls a Mendelian trait, whereas a complex trait is influenced by multiple mutations across several genes, each contributing a small individual effect.

leveraging genome-wide experimental data (Eraslan et al., 2019). Natural selection closes the loop by impacting which genotypes are favored over time, based on the fitness of the phenotype on a given environment. Therefore, the space of observed DNA sequences contains rich information about the underlying biology; this is precisely the signal leveraged by self-supervised DNA language models (Benegas et al., 2025b).

The are two classes of phenotypic traits: Mendelian and complex (Figure 2). Mendelian traits, such as hemophilia, can be strongly affected by a single mutation in a single gene. On the other hand, complex traits, such as the risk to develop Alzheimer's disease, are affected by several mutations in multiple genes, each typically with a small individual effect. The fact that variants affecting Mendelian traits have larger phenotypic effect sizes than variants affecting complex traits makes the former relatively easier to predict, as they tend to have larger effects on gene expression (the signal picked up by functional-genomics-supervised models) and tend to be subject to stronger purifying selection (the signal picked up by self-supervised models).

3 Related work

Kathail et al. (2024) provide a comprehensive overview of the landscape of non-coding variant effect prediction in human genetics. GeneticsGym (Finucane et al., 2024) evaluates the prediction of causal variants for human complex traits, but limited to protein-coding variants. Dey et al. (2020)

evaluate the prediction of non-coding causal variants for human complex traits, but limited to a previous generation of functional-genomics-supervised models. A recent work (Fabiha et al., 2024) also evaluates the prediction of causal variants for complex traits, but does not cover self-supervised models nor Mendelian traits. Benegas et al. (2025a) evaluate the prediction of non-coding causal variants for human Mendelian traits, but with a much larger, non-subsampled negative set of 2.6 million variants, which makes it less practical to evaluate some of the latest, computationally expensive models.

Tang et al. (2024) and Patel et al. (2024) benchmark the ability of functional-genomics-supervised and self-supervised models to predict non-coding variant effects on gene expression and chromatin accessibility, but they cover neither Mendelian nor complex traits. BEND (Marin et al., 2024) and GV-Rep (Li et al., 2024) evaluate self-supervised models for the prediction of disease-associated variants from ClinVar (Landrum et al., 2020). While not documented, it is likely that these variants mostly cover Mendelian rather than complex diseases. Furthermore, expert-reviewed pathogenic variants in ClinVar are highly skewed towards coding and splice region variants, containing only a single promoter variant and no intergenic variants (Supplementary Table S7). Neither of these benchmarks establishes adequate baselines for this task. BEND includes a single early-generation functional-genomics-supervised model (Zhou & Troyanskaya, 2015), but does not include any conservation-based model, which are usually strong for this task (Benegas et al., 2025a). GV-Rep does not include any baseline.

Thus, TraitGym is the only benchmark of causal non-coding variant prediction for both Mendelian and complex human traits. Furthermore, it is the only available framework to evaluate both the latest functional-genomics-supervised and self-supervised models, as well as strong non-neural baselines.

4 Benchmark datasets

TraitGym consists of two curated datasets of non-coding genetic variants affecting Mendelian and complex traits (Table 1). We focus on non-coding variants since understanding their impact is a particularly important use case for DNA sequence models, compared to coding variants which are more commonly interpreted using protein sequence models. Further, we focus on single-nucleotide variants, the most common form of genetic variation, which is still challenging to interpret. Our data curation process is outlined in Figure 3 and additional details are provided in Appendix A.

Mendelian traits. Curated causal non-coding variants for 113 Mendelian diseases were collected from Online Mendelian Inheritance in Man, OMIM (Smedley et al., 2016). For additional stringency, we filtered out a small percentage of variants with minor allele frequency (MAF) greater than 0.1% in the Genome Aggregation Database, gnomAD (Chen et al., 2024). We used gnomAD common variants (MAF $>$ 5%) as controls.

Complex traits. Putative causal and control non-coding variants for 83 complex traits were

Table 1: Number of variants and traits in TraitGym.

Dataset	Putatively causal variants	Total variants	Traits
Mendelian traits	338	3,380	113
Complex traits	1,140	11,400	83

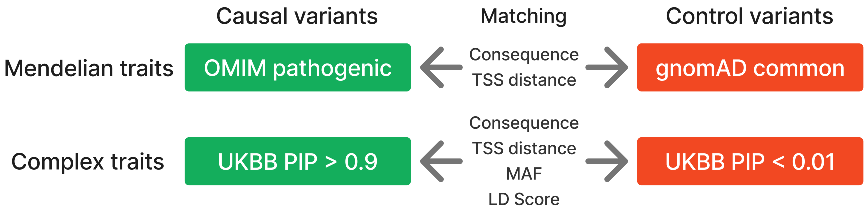

Figure 3: Matching putatively causal and control variants. Nine matched control variants are used for each putatively causal variant, within each chromosome. See the text for the details.

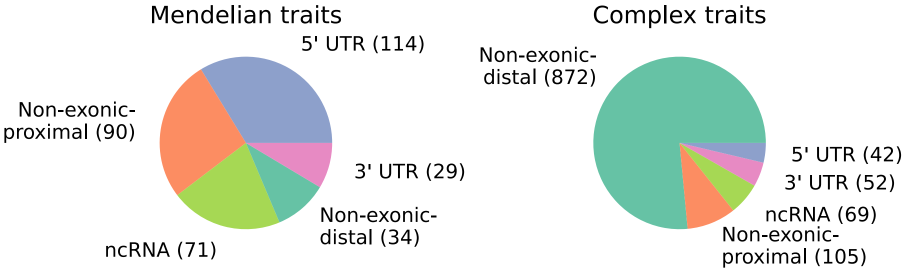

Figure 4: Distribution of consequence classes of putative causal non-coding variants.

obtained by processing statistical fine-mapping results (Kanai et al., 2021) from association studies in the UK BioBank data (Bycroft et al., 2018). Specifically, we used variants with posterior inclusion probability (PIP) in the credible set greater than 0.9 in any trait as positives and variants with PIP < 0.01 in all traits as controls. We additionally filtered the positive set to genome-wide significant variants ($p < 5 \times 10^{-8}$).

Variant type (or consequence) annotation. We annotated the consequence (e.g., intergenic, intronic, 5′ UTR, 3′ UTR, etc.) of each variant using Ensembl (McLaren et al., 2016), and refined this annotation by overlapping with candidate *cis*-regulatory elements from ENCODE (Epstein et al., 2020). Distal non-exonic variants (potential enhancers) comprise a small proportion (10%) in the Mendelian traits dataset but the vast majority (76%) in the complex traits dataset (Figure 4).

Matching positives and negatives. For each putative causal non-coding variant, we sampled 9 non-coding variants from the control set, matching chromosome, consequence, and distance to transcription start site (TSS). For complex traits, we additionally matched MAF and linkage disequilibrium (LD) score (Bulik-Sullivan et al., 2015) in the UK BioBank. We sampled only 9 controls per positive variant in order to be able to evaluate even the most computationally demanding models. However, we also provide a larger version of the dataset with millions of negative controls per positive variant, for which we evaluate a subset of the models. This expanded version of the dataset for Mendelian traits does not require any subsampling of negatives, but for complex traits we do subsample to match the MAF distribution (Finucane et al., 2024), while still keeping millions of variants.

Task definition. The task is to classify whether a variant is putatively causal for any trait or not. The input data consist of the reference and alternate allele together with the DNA sequence context. As evaluation metric, we calculate the area under the precision recall curve (AUPRC) for each chromosome (for a model trained on the remaining chromosomes), and then compute a weighted average across chromosomes based on sample size, together with a standard error estimated via bootstrapping (described in Appendix B.4). The baseline AUPRC is 0.1, which is the proportion of positives.

Table 2: Benchmarked models. Evo2 was trained with 1 M context size but utilizes a shorter context for variant effect prediction.

Model	Dependencies			Params	Context size	Extracted features	Source
	Functional genomics	Alignment	Population data				
Functional-genomics-supervised models							
Enformer	Yes	No	No	246M	196K	5,138	Avsec et al. (2021)
Sei	Yes	No	No	890M	4K	41	Chen et al. (2022)
Borzoi	Yes	No	No	186M	524K	7,617	Linder et al. (2025)
Self-supervised models							
GPN-MSA	No	Yes	No	86M	128	770	Benegas et al. (2025a)
NT	No	No	No	2.5B	6K	2,562	Dalla-Torre et al. (2024)
HyenaDNA	No	No	No	14M	160K	258	Nguyen et al. (2023)
Caduceus	No	No	No	8M	131K	514	Schiff et al. (2024)
SpeciesLM	No	No	No	97M	2K	770	Tomaz da Silva et al. (2024)
AIDO.DNA	No	No	No	7B	4K	4,354	Ellington et al. (2024)
Evo2	No	No	No	40B	8,192	8,194	Brixi et al. (2025)
GPN-Promoter	No	No	No	152M	512	1,026	This work
Integrative models							
CADD	Yes	Yes	Yes	N/A	N/A	114	Schubach et al. (2024)
Conservation scores							
phastCons	No	Yes	No	N/A	N/A	N/A	Siepel et al. (2005)
phyloP	No	Yes	No	N/A	N/A	N/A	Pollard et al. (2010)

5 Models

We benchmark functional-genomics-supervised models, self-supervised gLMs, integrative models and conservation scores (Table 2). We introduce a new gLM, called GPN-Promoter, trained using the genomes of 434 animal species, following the training objective of GPN (Benegas et al., 2023) and the ByteNet convolutional architecture (Kalchbrenner et al., 2017; Yang et al., 2024). It is only trained on promoters as an attempt to focus on regulatory regions (we would have liked to train on enhancers as well but no annotation exists for non-model organisms). In our comparisons we include SpeciesLM (Tomaz da Silva et al., 2024), which was also only trained on functional regions (2 kb upstream of start codon). LOL-EVE (Shearer et al., 2024) was also trained on promoters but is not yet publicly available. We also evaluate Evo2 (Brixi et al., 2025), the largest available gLM, trained on a dataset emphasizing exons and promoters. Additional details on models are provided in Appendix B.

We evaluate zero-shot model scores as well as ridge logistic regression classifiers (linear probing) trained using extracted features (Table 3). We use a number of folds equal to the number of chromosomes. In each fold, we test on a single chromosome using a model trained on the remaining chromosomes, and the regularization hyperparameter is chosen based on cross-validation on the training chromosomes (detailed in Appendix B.4).

Functional-genomics-supervised models. Sequence-to-function models predict activity in thousands of different functional genomic tracks, covering different assays, such as gene expression or chromatin accessibility, in different tissues and cell types. As variant effect prediction features, we calculate the norm (across spatial positions) of the predicted log-fold-change in activity between the reference and the alternate sequence, for each separate track (referred to as "ℓ_2 score" in Linder et al. (2025)). As zero-shot score, we aggregate the ℓ_2 scores of different tracks by taking their ℓ_2

Table 3: Extracted features and zero-shot scores for each model type.

Model type	Extracted features	Zero-shot score
Functional-genomics supervised (Enformer/Borzoi)	ℓ_2 scores: change in activity in each track ℓ_2 of ℓ_2 scores: aggregation of ℓ_2 scores across several tracks (all + within each assay type)	ℓ_2 of ℓ_2 scores (all tracks)
Functional-genomics supervised (Sei)	Change in sequence class scores	Max absolute change in sequence class scores
Self-supervised	LLR, abs(LLR) Embeddings inner product for each hidden dimension	LLR, abs(LLR) Embeddings inner product, ℓ_2 distance, cosine distance
Integrative	CADD input features, CADD score	CADD score
Conservation	N/A	Conservation score

norm ("ℓ_2 of ℓ_2 scores"). Sei (Chen et al., 2022) adopts a different variant scoring approach; it maps each sequence into discrete classes, such as promoters or brain-specific enhancers, and scores a variant according to how much it impacts the relative scores of different classes.

Self-supervised models. For self-supervised gLMs, a popular zero-shot score is the log-likelihood ratio (LLR) between the alternate and reference allele[1], which has been shown to reflect learned functional constraints, such as transcription factor binding sites (Benegas et al., 2023). Good results have also been obtained comparing the embeddings of the alternate and reference alleles (Dalla-Torre et al., 2024; Mendoza-Revilla et al., 2024). We evaluate these different scoring approaches for each model (Supplementary Table S8) and choose the best performing one when benchmarking against other models (Supplementary Table S9). An additional proposed zero-shot score is based on nucleotide dependencies (Tomaz da Silva et al., 2024). Due to the complexity of implementing it for different models with diverse tokenization schemes, we leave it for future work. We additionally obtain a high-dimensional featurization of a variant by calculating the inner product (across genomic positions in a given window) between contextual embeddings of the alternate and reference sequences, for each hidden dimension separately.

Integrative models. CADD (Schubach et al., 2024) is built on top of a broad set of curated annotations, including conservation, biochemical activity, population-level data as well as predictions from several machine learning models. Utilizing this rich set of input features, CADD is a logistic regression model trained to distinguish proxy-deleterious from proxy-neutral variants. The output of the model is called the CADD score, which we use as zero-shot score. In this paper, we also train our own models using the broad set of CADD *input* features, which we refer to as CADD features even though they are the input, not the output, of CADD.

Conservation scores. phastCons (Siepel et al., 2005) is a phylogenetic hidden Markov model designed to find runs of conserved positions. phyloP (Pollard et al., 2010) is a statistical test for departure from the neutral rate of substitution at a specific alignment position. We evaluate phyloP trained on 100 vertebrates (phyloP-100v) and on 241 mammals (phyloP-241m), and phastCons trained on 43 primates (phastCons-43p) (Sullivan et al., 2023).

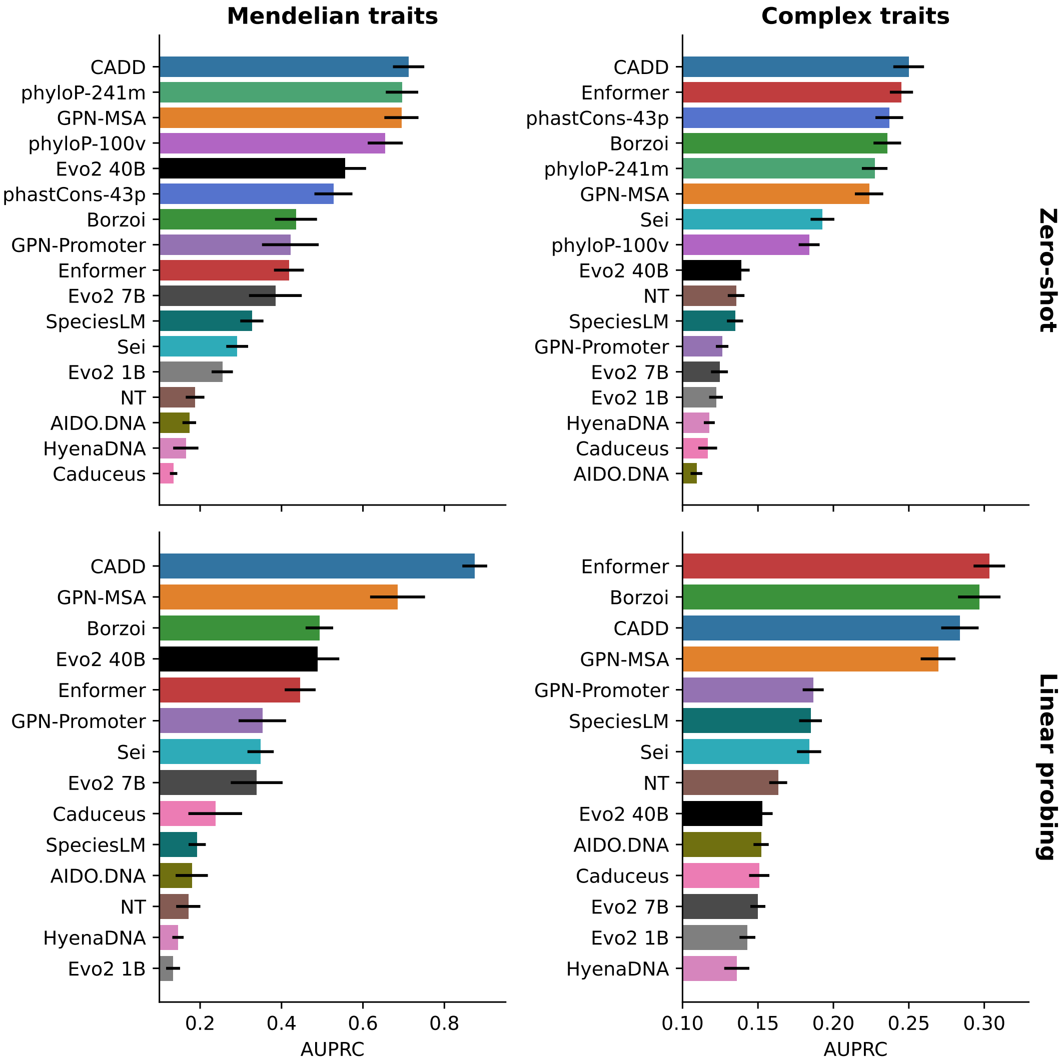

Figure 5: Results on each dataset with zero-shot and linear probing approaches. Zero-shot scores are described in Table 3. For linear probing, we use 113 CADD input features, together with the single CADD output score, while for the other models we only use output features (predicted tracks, LLR or embedding similarity).

6 Results

Mendelian traits. The best zero-shot scores are those leveraging alignments: CADD, phyloP and GPN-MSA, with phastCons a bit behind. Conservation features may be particularly useful for predicting causal variants of Mendelian traits, which are typically subject to relatively strong purifying selection. Among alignment-free models, the best-performing one is the largest Evo2 model with 40B parameters, showing notable performance gains with scale. GPN-Promoter and SpeciesLM perform relatively well, despite having a small number of parameters compared to Evo2 and being trained only on functional regions. A supervised model trained using CADD input features achieves the best performance when using linear probing (Figure 5). It can be challenging

[1] The absolute value of the LLR is more appropriate when we want scores to be invariant to which allele is the reference, as in the case of association studies.

to avoid overfitting when training high-dimensional models on this relatively small dataset, and several linear probing models perform worse than zero-shot. When using a more relaxed MAF cutoff of 1%, only 19 additional positive variants are included, resulting in very similar results (Supplementary Figure S1). Also, we explored matching negatives from the same gene rather than from the same chromosome, which required dropping many variants that could not be properly matched, but with similar overall conclusions (Supplementary Figure S2).

CADD is the only model trained on variants and its training variants overlap with around 1% of the variants in our datasets (Supplementary Table S10). However, CADD's positives and negatives are not defined based on causal variant annotations (Schubach et al., 2024), and they do not exhibit a clear association with the positive or negative sets in our datasets (Supplementary Table S10). We repeated our analysis upon removing this small amount of overlapping variants and found that the aforementioned results remain stable (Supplementary Figure S3).

Complex traits. Overall scores are much lower than in Mendelian traits (Figure 5). This is a harder task in principle since variants affecting complex traits are expected to have relatively small effect sizes. CADD, GPN-MSA and conservation scores also perform relatively well on this dataset, but Enformer and Borzoi ultimately come first when their predicted tracks are used in linear probing. GPN-Promoter and SpeciesLM perform the best among alignment-free gLMs, but only with linear probing. When using a more stringent PIP cutoff, performance generally improves, and Borzoi gains a small relative advantage (Supplementary Figure S4). When matching negatives from the same gene rather than the same chromosome, performance is lower overall, but Borzoi obtains a small advantage (Supplementary Figure S5).

While the AUPRC is generally recommended for imbalanced datasets where we are mostly interested in the positive minority class (Whalen et al., 2022), we also report the area under the receiver operating characteristic (AUROC). The main difference we see is a slight relative improvement of Enformer and Borzoi zero-shot scores (Supplementary Figure S6). We also report similar results for zero-shot scores when using a global AUPRC rather than a weighted average over chromosomes (Supplementary Figure S7).

Results on expanded datasets. We also considered expanded datasets containing millions of negative controls and evaluated the models with precomputed genome-wide zero-shot scores (CADD, GPN-MSA, phyloP and phastCons). For Mendelian traits, GPN-MSA achieves a substantial margin over other models (Figure 6), in agreement with previous benchmarks (Benegas et al., 2025a). For complex traits, CADD is the best performing model, but none of the models does very well in absolute terms (Figure 6). In the future, we hope to evaluate other models on these full datasets, but we estimate that slower models like Caduceus would take approximately 6 months of compute on an NVIDIA A100 80GB GPU.

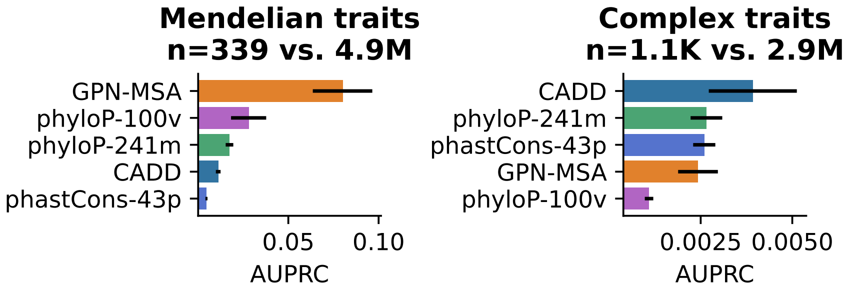

Figure 6: Results with a much larger negative set of millions of variants. The x-axis range starts at the baseline which is the proportion of positives.

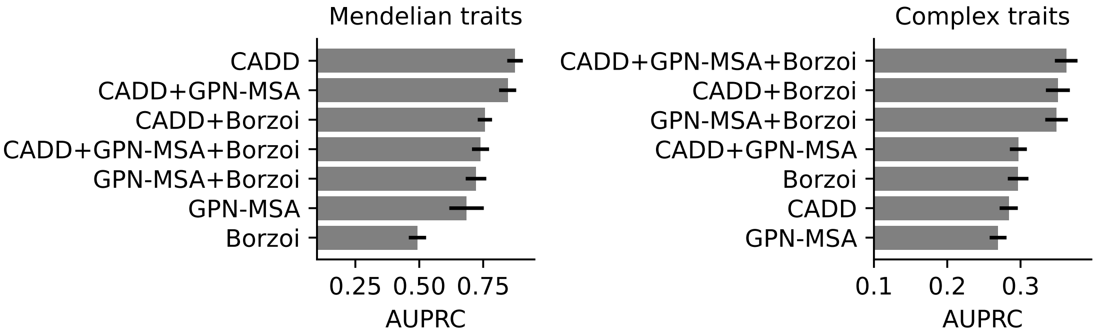

Figure 7: Results of ensembling models by training a logistic regression classifier on the concatenation of their features.

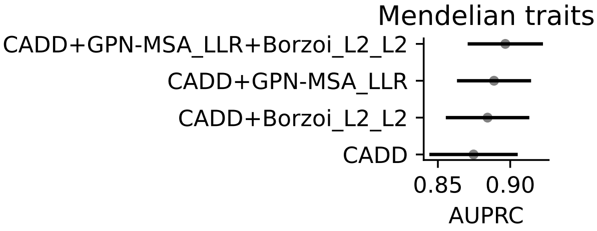

Figure 8: Results of lightweight ensembles. Full CADD features are used, but a much reduced number of features are used from other models.

Model ensembling. Given the good performance obtained by different classes of models, potentially leveraging different signals, we evaluated linear probing of combined features extracted from representative models from different classes: Borzoi (predicted tracks), GPN-MSA (latent embeddings and LLR) and CADD (input features to the model together with the single output score). The results are summarized in Figure 7. On the complex traits dataset, ensembling the three models achieves the best performance, with a particularly high jump when combining Borzoi with either of CADD or GPN-MSA. On the Mendelian traits dataset, on the other hand, ensembling the full features from different models does not improve upon CADD input features. We attribute this to the fact that (i) the room for improvement is relatively small and (ii) the dataset is small, making it easier to overfit when using high-dimensional features. We refer to the last approach as "full" feature ensembling. However, we do see small improvements when ensembling CADD with a reduced number of features from other models (LLR for GPN-MSA and "ℓ_2 of ℓ_2 scores" for Borzoi), which we refer to as "lightweight" feature ensembling (Figure 8).

Results by consequence (variant type). We also evaluated the performance stratified by variant consequence classes (Figure 9A). The most important insight here is that the advantage of ensembling for complex traits holds within each consequence class, so it is not simply that different models are experts on different consequences. Second, we note that distal (TSS distance $>$ 1 kb) non-exonic variants for complex traits (which make up the majority) are the hardest class overall. Lastly, while Borzoi performs the worst for Mendelian traits, the gap is the smallest for proximal non-exonic variants.

We then investigated the effect of scaling Evo2 models (Supplementary Figure S8). Evo2 shows substantial improvements with scale, in particular for proximal non-exonic (promoter-like) regions and ncRNA. On distal non-exonic (enhancer-like) regions, however, scaling Evo2 does not seem to help much. While Evo2's training dataset curation process emphasized exons and promoters, it did

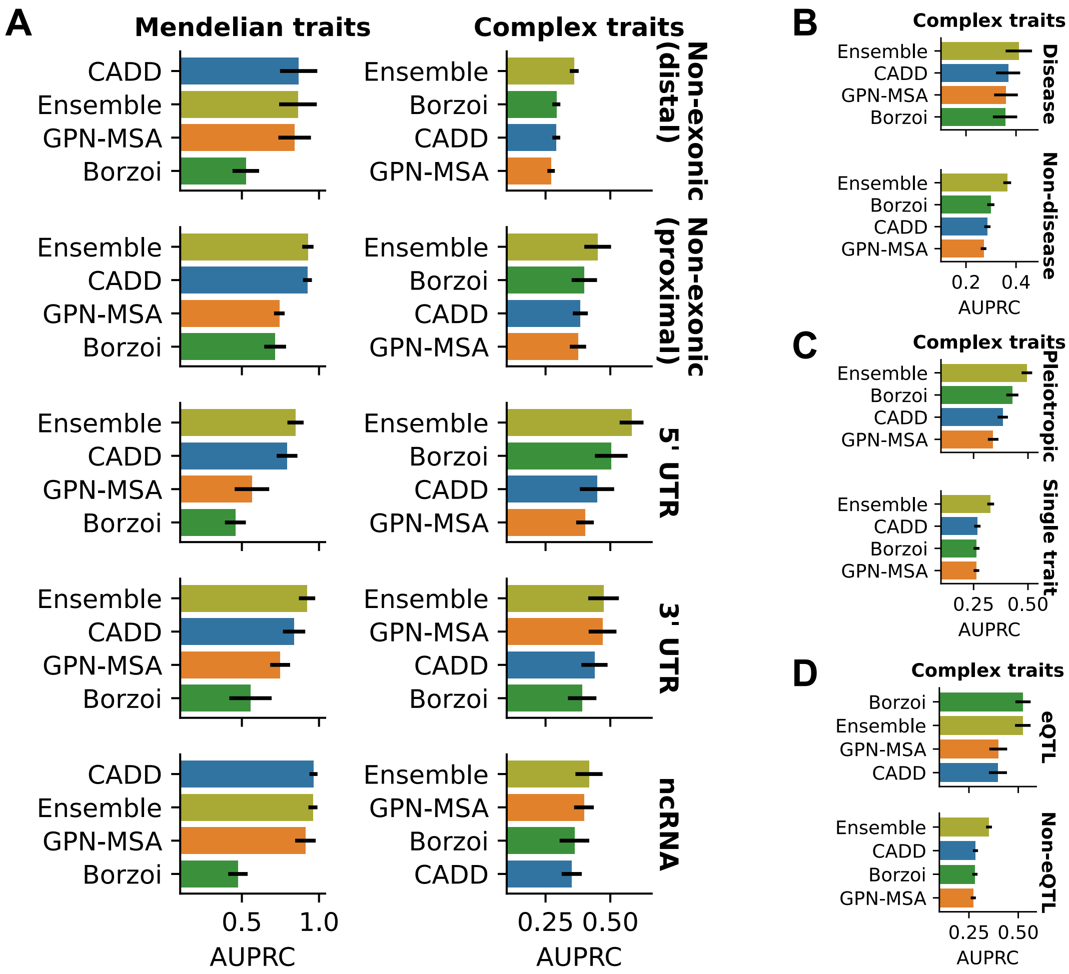

Figure 9: Stratified results. The best score is reported between zero-shot and linear probing. (A) Results by consequence (variant type). Full feature ensemble is evaluated for complex traits, but lightweight feature ensemble is evaluated for Mendelian traits. (B) Results for disease vs. non-disease complex traits. (C) Results for pleiotropic vs. non-pleiotropic variants. (D) Results for complex traits variants stratified by whether or not they overlap with fine-mapped eQTLs.

not give a special weight to enhancers, which are sparsely represented in genomes such as that of humans. To gain more insight into Evo2's behavior on enhancers, we visualized the sequence logo according to the predictions around the well-known ZRS enhancer (Supplementary Figure S9). We know that ZRS is functional from its conservation, ENCODE biochemical activity (Epstein et al., 2020) and association with polydactyly (Albuisson et al., 2011). While Evo2 predicts constraint on a nearby *LMBR1* exon, it does not predict constraint on the ZRS enhancer. In fact, after the exon, for Evo2 assigns the next highest constraint to repetitive elements. It has been previously observed that gLMs still predict a higher constraint than expected on repeats, even if they are downweighted during training (Benegas et al., 2023; Zhai et al., 2024).

We also inspected the performance of GPN-Promoter on different consequences, given that it was trained only on promoters (Supplementary Figure S10). GPN-Promoter's zero-shot scores perform better on proximal non-exonic and 5' UTR variants, which lie in the regions of the gene with the highest overlap with the model's training data (512 bp around the TSS). Except for the aforementioned classes in Mendelian traits, linear probing outperforms zero-shot scores. When

Table 4: Top CADD features in different categories. Our understanding is that official CADD feature names `ZooVerPhyloP` and `ZooPriPhyloP` can be misleading and respectively represent mammalian phyloP (Pollard et al., 2010) and primate phastCons (Siepel et al., 2005) from Zoonomia (Sullivan et al., 2023).

Dataset	Category	Feature	AUPRC	Description
Mendelian traits	Alignment	`ZooVerPhyloP`	0.673	Conservation in mammals
	Functional genomics	`EncodetotalRNA-max`	0.348	Max. RNA-seq level
	Population data	(-) `Freq100bp`	0.509	# common variants within 100bp
Complex traits	Alignment	`ZooPriPhyloP`	0.225	Conservation in primates
	Functional genomics	`EncodeDNase-max`	0.145	Max. DNase-seq level
	Population data	(-) `Freq10000bp`	0.131	# common variants within 10kb

evaluated only on promoter-like regions, GPN-Promoter (with only 152M parameters) performs comparably to the Evo2 40B model (Supplementary Figure S11). This suggests that applications involving a single genomic region require vastly fewer parameters.

Results by trait. We also report performance (Supplementary Table S11) for specific traits with sufficiently many putative causal variants and not overlapping too much with each other; specifically, traits with at least 10 causal variants and less than 10% overlap of causal variants with other traits. Ensembling wins in the majority of these traits. Among the 1,140 putative causal variants for complex traits, only 53 affect a disease trait (Supplementary Table S1). We evaluated the results stratified by disease vs. non-disease complex traits, pooled given the small sample size (Figure 9B)—for example, our dataset only contains 3 non-coding variants affecting the risk of developing Alzheimer's disease. We note that causal variants for disease traits are easier to classify overall than for non-disease traits, and that Borzoi loses the edge compared to conservation-aware CADD and GPN-MSA for disease traits. This is consistent with disease traits being under stronger selective pressures. We also noted that putative pleiotropic variants (i.e., those affecting multiple traits) are in general easier to predict, with the biggest advantage being gained by the ensemble model and Borzoi (Figure 9C).

eQTL colocalization. We found that 103 putative causal variants for complex traits (9%) overlap with fine-mapped GTEx eQTL variants (Lonsdale et al., 2013; Wang et al., 2021); we found no such overlap for Mendelian trait variants, as expected given their low allele frequencies. The low overlap of complex trait and eQTL variants is well known and Mostafavi et al. (2023) discuss several hypotheses for the cause. We found that eQTL-overlapping variants are much easier to predict than non-eQTL-overlapping variants, across all model types (Figure 9D). We also note that Borzoi achieves a wide margin compared to other models and little is gained from ensembling. We observed that eQTL-overlapping variants are enriched in exonic variants (Fisher's exact $p = 8 \times 10^{-8}$) and, among non-exonic variants, they have lower TSS distances (Mann Whitney $p = 4 \times 10^{-4}$), all of which could explain their increased predictability.

Interpreting CADD features. CADD contains informative features from three orthogonal categories: alignment, functional genomics, and population genetic data (Table 4). Conservation features are the most predictive overall. Conservation in mammals (phyloP-241m) is most predictive for Mendelian traits, whereas conservation in primates (phastCons-43p) is most predictive for complex traits, in line with recent observations (Sullivan et al., 2023).

Interpreting Borzoi features. We evaluated the performance of aggregated Borzoi scores across specific experimental assays (Figure 10). Of note, gene expression tracks (RNA and CAGE) perform

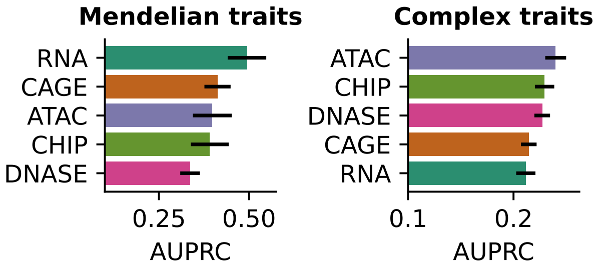

Figure 10: Results of "ℓ_2 of ℓ_2 scores" aggregating different assays (Borzoi).

Table 5: Top three tissue/cell types for different traits, ranked by the highest AUPRC of Borzoi predicted tracks from such tissue/cell type.

Trait	Tissue/cell type/cell line	AUPRC
Mendelian traits		
Beta-thalassemia	aorta	0.997
	stomach	0.988
	adrenal gland	0.986
Hemophilia B	liver	1.0
	HepG2	1.0
	hepatocyte	1.0
Hypercholesterolemia-1	CD8+ T cell	0.983
	HepG2	0.975
	CD4+ T cell	0.972
Complex traits		
Monocyte count	neutrophil	0.559
	CD14+ monocyte	0.559
	HL-60	0.559
Hemoglobin A1c	K562	0.449
	erythroblast	0.423
	hematopoietic progenitor	0.412
High density lipoprotein cholesterol	liver	0.44
	abdominal adipose tissue	0.42
	adrenal gland	0.417

the best on Mendelian traits, while epigenetic tracks (ATAC, CHIP and DNASE) perform the best on complex traits. It has been shown that models such as Borzoi tend to particularly struggle with finding causal variants affecting gene expression when these are distal as opposed to proximal (Karollus et al., 2023). In the case of distal causal variants for complex traits (which make up the majority, see Figure 4), epigenetic tracks might instead be more informative.

A key feature of functional-genomics-supervised models such as Borzoi is that their features are associated with a specific tissue or cell type, which can help interpret disease pathways as well as design therapeutics. For traits where Borzoi achieved a good performance, we inspected the tissue/cell type of the top features, and found that they are usually well aligned with previous knowledge (Table 5). For example, the top tissues for high density lipoprotein cholesterol are liver, abdominal adipose tissue and adrenal gland.

7 Discussion

TraitGym allows to benchmark DNA sequence models on the challenging task of predicting causal regulatory variants in human genetics. Alignment-based, conservation-aware models compare favorably on Mendelian traits and complex disease traits, while functional-genomics-supervised models achieve the best performance on complex non-disease traits. A reason for hope in the particularly challenging complex traits dataset is that ensembling predictions and input features from different models yields consistent improvements.

While alignment-free gLMs still lag behind alignment-based methods on causal variant prediction, substantial performance gains on Mendelian traits are achieved by scaling Evo2. The relatively good performance of GPN-Promoter and SpeciesLM, and the fact that the main weakness of Evo2 are enhancers (sparsely seen during training), suggest that functional data curation is one of the main bottlenecks for progress in genomic language modeling, as previously proposed (Tang et al., 2024; Benegas et al., 2025b; Patel et al., 2024).

CADD shows great results combining functional genomics, alignment and population data. Integrating such modalities with powerful neural networks, rather than relying on linear combinations of hand-crafted features, is another attractive avenue of research.

TraitGym was designed to have simple metrics and be tractable to run for even billion parameter models, but has limited ability to resolve subtle differences in performance. For example, the advantage of GPN-MSA over other alignment-based zero-shot scores in Mendelian traits is only visible when comparing on the expanded datasets. For more in-depth evaluation, we recommend additional benchmarks such as those in Benegas et al. (2025a), spanning millions of genome-wide variants. In the case of complex traits, partitioned LD score regression is a powerful evaluation framework as it is not limited to finemapped or even genome-wide-significant variants (Finucane et al., 2015; Dey et al., 2020; Fabiha et al., 2024).

Data, model and code availability

TraitGym is available at https://huggingface.co/datasets/songlab/TraitGym, with a leaderboard at https://huggingface.co/spaces/songlab/TraitGym-leaderboard. GPN-Promoter is available at https://huggingface.co/songlab/gpn-animal-promoter and its training dataset at https://huggingface.co/datasets/songlab/gpn-animal-promoter-dataset. Code to reproduce the analysis is available at https://github.com/songlab-cal/TraitGym.

Acknowledgments

We would like to thank Carlos Albors, Laura Gunsalus, Jerome Irudayanathan, Avantika Lal, Canal Li, Surag Nair and Chengzhong Ye for helpful discussions. We thank Rohan Sachdeva and the Innovative Genomics Institute at UC Berkeley for kindly offering computational resources to run Evo2. Protein icon was contributed by Bioicons user DBCLS and heatmap icon by Bioicons user ChenxinLi. This research is supported in part by the National Institutes of Health (grant R35-GM134922) and a grant from the Koret–UC Berkeley–Tel Aviv University Initiative in Computational Biology and Bioinformatics.

References

J Albuisson, B Isidor, M Giraud, O Pichon, T Marsaud, A David, C Le Caignec, and S Bezieau. Identification of two novel mutations in Shh long-range regulator associated with familial preaxial polydactyly. *Clinical Genetics*, 79(4):371–377, 2011.

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18 (10):1196–1203, 2021.

Gonzalo Benegas, Sanjit Singh Batra, and Yun S. Song. DNA language models are powerful predictors of genome-wide variant effects. *Proceedings of the National Academy of Sciences*, 120 (44):e2311219120, 2023.

Gonzalo Benegas, Carlos Albors, Alan J Aw, Chengzhong Ye, and Yun S Song. A DNA language model based on multispecies alignment predicts the effects of genome-wide variants. *Nature Biotechnology*, pp. 1–6, 2025a.

Gonzalo Benegas, Chengzhong Ye, Carlos Albors, Jianan Canal Li, and Yun S Song. Genomic language models: opportunities and challenges. *Trends in Genetics*, 2025b.

Christian Benner, Chris CA Spencer, Aki S Havulinna, Veikko Salomaa, Samuli Ripatti, and Matti Pirinen. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, 32(10):1493–1501, 2016.

Garyk Brixi, Matthew G Durrant, Jerome Ku, Michael Poli, Greg Brockman, Daniel Chang, Gabriel A Gonzalez, Samuel H King, David B Li, Aditi T Merchant, et al. Genome modeling and design across all domains of life with Evo 2. *bioRxiv*, pp. 2025–02, 2025.

Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47(3):291–295, 2015.

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

Kathleen M Chen, Aaron K Wong, Olga G Troyanskaya, and Jian Zhou. A sequence-based global map of regulatory activity for deciphering human genetics. *Nature Genetics*, 54(7):940–949, 2022.

Siwei Chen, Laurent C Francioli, Julia K Goodrich, Ryan L Collins, Masahiro Kanai, Qingbo Wang, Jessica Alföldi, Nicholas A Watts, Christopher Vittal, Laura D Gauthier, et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature*, 625(7993):92–100, 2024.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P de Almeida, Hassan Sirelkhatim, et al. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, pp. 1–11, 2024.

Kushal K Dey, Bryce Van de Geijn, Samuel Sungil Kim, Farhad Hormozdiari, David R Kelley, and Alkes L Price. Evaluating the informativeness of deep learning annotations for human complex diseases. *Nature Communications*, 11(1):4703, 2020.

Caleb N Ellington, Ning Sun, Nicholas Ho, Tianhua Tao, Sazan Mahbub, Dian Li, Yonghao Zhuang, Hongyi Wang, Le Song, and Eric P Xing. Accurate and general DNA representations emerge from genome foundation models at scale. *bioRxiv*, pp. 2024–12, 2024.

Charles B Epstein, Noam Shoresh, Jessika Adrian, Trupti Kawli, Carrie A Davis, Alexander Dobin, Rajinder Kaul, Jessica Halow, Eric L Van Nostrand, Peter Freese, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, 583:699, 2020.

Gökcen Eraslan, Žiga Avsec, Julien Gagneur, and Fabian J Theis. Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, 20(7):389–403, 2019.

Tabassum Fabiha, Ivy Evergreen, Soumya Kundu, Anusri Pampari, Sergey Abramov, Alexandr Boytsov, Kari Strouse, Katherine Dura, Weixiang Fang, Gaspard Kerner, et al. A consensus variant-to-function score to functionally prioritize variants for disease. *bioRxiv*, pp. 2024–11, 2024.

Hilary K Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, Kyle Farh, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics*, 47 (11):1228–1235, 2015.

Hilary K Finucane, Sophie Parsa, Jeremy Guez, Masahiro Kanai, F Kyle Satterstrom, Lethukuthula L Nkambule, Mark J Daly, Cotton Seed, and Konrad J Karczewski. Variant scoring performance across selection regimes depends on variant-to-gene and gene-to-disease components. *bioRxiv*, pp. 2024–09, 2024. URL https://www.biorxiv.org/content/10.1101/2024.09.17.613327v1.

Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aaron van den Oord, Alex Graves, and Koray Kavukcuoglu. Neural Machine Translation in Linear Time, 2017. URL https://arxiv.org/abs/1610.10099.

Masahiro Kanai, Jacob C Ulirsch, Juha Karjalainen, Mitja Kurki, Konrad J Karczewski, Eric Fauman, Qingbo S Wang, Hannah Jacobs, François Aguet, Kristin G Ardlie, et al. Insights from complex trait fine-mapping across diverse populations. *medrxiv*, pp. 2021–09, 2021. URL https://www.medrxiv.org/content/10.1101/2021.09.03.21262975v1.

Konrad J Karczewski, Rahul Gupta, Masahiro Kanai, Wenhan Lu, Kristin Tsuo, Ying Wang, Raymond K Walters, Patrick Turley, Shawneequa Callier, Nikolas Baya, et al. Pan-UK Biobank GWAS improves discovery, analysis of genetic architecture, and resolution into ancestry-enriched effects. *medRxiv*, pp. 2024–03, 2024.

Alexander Karollus, Thomas Mauermeier, and Julien Gagneur. Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome Biology*, 24(1):56, 2023.

Pooja Kathail, Ayesha Bajwa, and Nilah M Ioannidis. Leveraging genomic deep learning models for non-coding variant effect prediction. *arXiv preprint arXiv:2411.11158*, 2024.

Avantika Lal, Laura Gunsalus, Surag Nair, Tommaso Biancalani, and Gokcen Eraslan. gReLU: A comprehensive framework for DNA sequence modeling and design. *bioRxiv*, pp. 2024–09, 2024.

Melissa J Landrum, Shanmuga Chitipiralla, Garth R Brown, Chao Chen, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, Kuljeet Kaur, Chunlei Liu, et al. ClinVar: improvements to accessing data. *Nucleic Acids Research*, 48(D1):D835–D844, 2020.

Zehui Li, Vallijah Subasri, Guy-Bart Stan, Yiren Zhao, and BO WANG. GV-rep: A large-scale dataset for genetic variant representation learning. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=CW9SJyhpVt.

Johannes Linder, Divyanshi Srivastava, Han Yuan, Vikram Agarwal, and David R Kelley. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. *Nature Genetics*, pp. 1–13, 2025.

John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (GTEx) project. *Nature Genetics*, 45(6):580–585, 2013.

Frederikke Isa Marin, Felix Teufel, Marc Horlacher, Dennis Madsen, Dennis Pultz, Ole Winther, and Wouter Boomsma. BEND: Benchmarking DNA language models on biologically meaningful tasks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=uKB4cFNQFg.

William McLaren, Laurent Gil, Sarah E Hunt, Harpreet Singh Riat, Graham RS Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1):1–14, 2016.

Javier Mendoza-Revilla, Evan Trop, Liam Gonzalez, Maša Roller, Hugo Dalla-Torre, Bernardo P. de Almeida, Guillaume Richard, Jonathan Caton, Nicolas Lopez Carranza, Marcin Skwark, Alex Laterre, Karim Beguir, Thomas Pierrot, and Marie Lopez. A foundational large language model for edible plant genomes. *Communications Biology*, 7(1):835, Jul 2024. ISSN 2399-3642. doi: 10.1038/s42003-024-06465-2. URL https://doi.org/10.1038/s42003-024-06465-2.

Hakhamanesh Mostafavi, Jeffrey P Spence, Sahin Naqvi, and Jonathan K Pritchard. Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nature Genetics*, 55(11):1866–1875, 2023.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, Stefano Ermon, Christopher Ré, and Stephen Baccus. HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 43177–43201. Curran Associates, Inc., 2023.

Nuala A O'Leary, Eric Cox, J Bradley Holmes, W Ray Anderson, Robert Falk, Vichet Hem, Mirian TN Tsuchiya, Gregory D Schuler, Xuan Zhang, John Torcivia, et al. Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. *Scientific Data*, 11(1):732, 2024.

Aman Patel, Arpita Singhal, Austin Wang, Anusri Pampari, Maya Kasowski, and Anshul Kundaje. DART-eval: A comprehensive DNA language model evaluation benchmark on regulatory DNA. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=qR0x6H5WUX.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Gerardo Perez, Galt P Barber, Anna Benet-Pages, Jonathan Casper, Hiram Clawson, Mark Diekhans, Clay Fischer, Jairo Navarro Gonzalez, Angie S Hinrichs, Christopher M Lee, et al. The ucsc genome browser database: 2025 update. *Nucleic Acids Research*, 53(D1):D1243–D1249, 2025.

Katherine S Pollard, Melissa J Hubisz, Kate R Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20(1):110–121, 2010.

Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: Bi-directional equivariant long-range DNA sequence modeling. *arXiv preprint arXiv:2403.03234*, 2024. URL https://arxiv.org/abs/2403.03234.

Max Schubach, Thorben Maass, Lusiné Nazaretyan, Sebastian Röner, and Martin Kircher. CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions. *Nucleic Acids Research*, 52(D1):D1143–D1154, 2024.

Courtney Shearer, Felix Teufel, Rose Orenbuch, Daniel Ritter, Aviv Spinner, Erik Xie, Jonathan Frazer, Mafalda Dias, Pascal Notin, and Debora Marks. LOL-EVE: Predicting Promoter Variant Effects from Evolutionary Sequences. *bioRxiv*, pp. 2024–11, 2024.

Adam Siepel, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, LaDeana W Hillier, Stephen Richards, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034–1050, 2005.

Damian Smedley, Max Schubach, Julius OB Jacobsen, Sebastian Köhler, Tomasz Zemojtel, Malte Spielmann, Marten Jäger, Harry Hochheiser, Nicole L Washington, Julie A McMurry, et al. A whole-genome analysis framework for effective identification of pathogenic regulatory variants in Mendelian disease. *The American Journal of Human Genetics*, 99(3):595–606, 2016.

Patrick F Sullivan, Jennifer RS Meadows, Steven Gazal, BaDoi N Phan, Xue Li, Diane P Genereux, Michael X Dong, Matteo Bianchi, Gregory Andrews, Sharadha Sakthikumar, et al. Leveraging base-pair mammalian constraint to understand genetic variation and human disease. *Science*, 380(6643):eabn2937, 2023.

Ziqi Tang, Nirali Somia, Yiyang Yu, and Peter K Koo. Evaluating the representational power of pre-trained DNA language models for regulatory genomics. *bioRxiv*, 2024.

Pedro Tomaz da Silva, Alexander Karollus, Johannes Hingerl, Gihanna Galindez, Nils Wagner, Xavier Hernandez-Alias, Danny Incarnato, and Julien Gagneur. Nucleotide dependency analysis of DNA language models reveals genomic functional elements. *bioRxiv preprint*, pp. 2024–07, 2024. URL https://www.biorxiv.org/content/10.1101/2024.07.27.605418v1.

Gao Wang, Abhishek Sarkar, Peter Carbonetto, and Matthew Stephens. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(5):1273–1300, 2020.

Qingbo S Wang, David R Kelley, Jacob Ulirsch, Masahiro Kanai, Shuvom Sadhuka, Ran Cui, Carlos Albors, Nathan Cheng, Yukinori Okada, et al. Leveraging supervised learning for functionally informed fine-mapping of cis-eQTLs identifies an additional 20,913 putative causal eQTLs. *Nature Communications*, 12(1):3394, 2021.

Sean Whalen, Jacob Schreiber, William S Noble, and Katherine S Pollard. Navigating the pitfalls of applying machine learning in genomics. *Nature Reviews Genetics*, 23(3):169–181, 2022.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Perric Cistac, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-Art Natural Language Processing. pp. 38–45. Association for Computational Linguistics, October 2020. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Kevin K Yang, Nicolo Fusi, and Alex X Lu. Convolutions are competitive with transformers for protein sequence pretraining. *Cell Systems*, 15(3):286–294, 2024.

Jingjing Zhai, Aaron Gokaslan, Yair Schiff, Ana Berthel, Zong-Yan Liu, Wei-Yun Lai, Zachary R Miller, Armin Scheben, Michelle C Stitzer, M Cinta Romay, et al. Cross-species modeling of plant genomes at single nucleotide resolution using a pre-trained DNA language model. *bioRxiv*, 2024.

Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10):931–934, 2015.

Supplementary Material

A Datasets

A.1 Mendelian traits

Non-coding pathogenic OMIM variants were obtained from Table S6 in Smedley et al. (2016). Common variants were obtained from gnomAD (Chen et al., 2024) (version 3.1.2).

A.2 Complex traits

UK BioBank fine-mapping results (Kanai et al., 2021) were downloaded from https://www.finucanelab.org/data (version: Dec. 3rd, 2019). As recommended to increase fine-mapping accuracy (Kanai et al., 2021), we averaged the posterior inclusion probability (PIP) from FINEMAP (Benner et al., 2016) and SuSiE (Wang et al., 2020), and excluded variants where the two methods disagreed by more than 5%. Complex traits in our dataset that are considered diseases or disorders are shown in Table S1.

Table S1: Disease or disorder complex traits in our dataset.

Trait
Atrial fibrillation
Autoimmune disease (Phecode + Self-reported)
Alzheimer disease (LTFH)
Asthma
Blood clot in the lung
Breast cancer
Coronary artery disease
Colorectal cancer
Cholelithiasis
Seen doctor (GP) for nerves, anxiety, tension or depression
Blood clot in the leg
Fibroblastic disorders
Glaucoma (Phecode + Self-reported)
Hypothyroidism
Inflammatory bowel disease
Inguinal hernia
Insomnia
Migraine (Self-reported)
Prostate cancer
Type 2 diabetes
Type 2 diabetes (adjusted by BMI)

Table S2: Selected consequences in this study.

Consequence
Non-exonic
`intergenic_variant`
`upstream_gene_variant`
`downstream_gene_variant`
`intron_variant`
Exonic
`5_prime_UTR_variant`
`3_prime_UTR_variant`
`non_coding_transcript_exon_variant`

Table S3: ENCODE cCRE categories.

Category
PLS (promoter-like signature)
pELS (proximal enhancer-like signature)
dELS (distal enhancer-like signature)
DNase-H3K4me3
CTCF-only

A.3 Variant annotation

Consequences were annotated using Ensembl VEP (McLaren et al., 2016) (release 109.1), using flags `--most_severe` and `--distance 1000` (used to distinguish upstream and downstream from intergenic variants). We only kept non-coding consequences (Table S2). We discarded splice region variants, such as splice donor variants, as these were very few in number. Coding variants, as well as non-coding variants with a very high expected impact such as in splice donors, are excluded from our analysis.

We refined the annotation of non-exonic variants by checking overlap with each of five different ENCODE candidate *cis*-regulatory element (cCRE) categories (Epstein et al., 2020) (Table S3). We additionally refined the annotation if a variant overlapped not a cCRE but the 500-bp flank of a cCRE, similar to Finucane et al. (2015). When we match negative controls, we make sure to keep the exact same proportion of consequences, including the distribution of cCRE elements and their flanks. For the analysis of performance by consequence, however, we simplify the categorization of non-exonic variants into proximal (TSS dist. $\leq$ 1 kb) and distal (TSS dist. $>$ 1 kb).

TSS distance was computed with respect to protein coding transcripts only. MAF and LD scores for the UK Biobank computed by the Pan-UK Biobank initiative (Karczewski et al., 2024) were downloaded from `s3://pan-ukb-us-east-1/ld_release/UKBB.EUR.ldscore.ht`.

GTEx fine-mapping results where downloaded from `https://www.finucanelab.org/data`. We used a similar PIP cutoff of 0.9 in any tissue, combined between FINEMAP and SuSiE, to define putative causal eQTL variants.

Table S4: Global AUPRC of matched features, close to baseline (0.1).

Dataset	Feature	AUPRC
Mendelian traits	(-) TSS distance	0.115
Complex traits	(-) TSS distance	0.104
Complex traits	MAF	0.101
Complex traits	(-) LD score	0.104

A.4 Matching controls

Nine negative control variants were sampled for each positive causal variant. Chromosome and consequence were matched exactly. We matched variants with the most similar TSS distance, as well as MAF and LD score in the complex traits dataset. More precisely, we defined a vector space of (TSS distance, MAF, LD score) tuples, applied scikit-learn's robust scaler (Pedregosa et al., 2011), and selected negative variants minimizing the euclidean distance to the positive variant. Table S4 shows that the matched features have minimal predictive power, as intended. For special cases where there were not enough negative controls to match positive variants for a given chromosome and consequence, we subsampled the positive variants until we had at least nine controls per positive variant.

For the full version of the complex traits dataset, we created 100 equal-size MAF bins and subsampled the negative set until the proportion of variants in each bin was equal to that of the positive set.

B Models

B.1 Published models

We downloaded several models from Hugging Face Hub (Wolf et al., 2020) (Table S5). We downloaded Evo2 models with different sizes (`Evo2_1b_base`, `Evo2_7b`, `Evo2_40b`) using the instructions provided in the official repository. We downloaded Enformer and Borzoi from gReLU's Model Zoo (Lal et al., 2024). Sei scores were obtained via their web server: `https://hb.flatironinstitute.org/sei`. We obtained CADD v1.7 scores and annotations from `https://krishna.gs.washington.edu/download/CADD/v1.7/GRCh38/whole_genome_SNVs_inclAnno.tsv.gz`. We downloaded phyloP-100v, phyloP-241m, and phastCons-43p from the following links, respectively:
`https://hgdownload.soe.ucsc.edu/goldenPath/hg38/phyloP100way/hg38.phyloP100way.bw`,
`https://hgdownload.soe.ucsc.edu/goldenPath/hg38/cactus241way/cactus241way.phyloP.bw`,
`https://cgl.gi.ucsc.edu/data/cactus/zoonomia-2021-track-hub/hg38/phyloPPrimates.bigWig`.

B.2 Our GPN-Promoter model

GPN-Promoter was trained on 512-bp sequences centered at TSSs of protein-coding genes from reference genomes of animal species. TSS coordinates were obtained from the gene annotations available at NCBI Datasets (O'Leary et al., 2024). Species available at NCBI Datasets were subsampled, among those with gene annotations, to keep at most one per family. This resulted in 434 reference genomes. GPN-Promoter's training objective follows GPN: base-pair-level tokenization and masked language modeling of local windows of 512-bp with downweighting of repeat

Table S5: Hugging Face Hub models.

Model	Hugging Face Hub path
GPN-MSA	`songlab/gpn-msa-sapiens`
NT	`InstaDeepAI/nucleotide-transformer-2.5b-multi-species`
HyenaDNA	`LongSafari/hyenadna-medium-160k-seqlen-hf`
Caduceus	`kuleshov-group/caduceus-ps_seqlen-131k_d_model-256_n_layer-16`
SpeciesLM	`johahi/specieslm-metazoa-upstream-k6`
AIDO.DNA	`genbio-ai/AIDO.DNA-7B`

positions (soft-masked in the reference genome). GPN-Promoter's architecture follows ByteNet (Kalchbrenner et al., 2017; Yang et al., 2024), consisting of blocks alternating dilated convolutions and feed-forward layers. Hyperparameters are displayed in Table S6. Training took approximately 2 weeks using 4 NVIDIA A100 40GB GPUs.

Table S6: GPN-Promoter training hyperparameters

Window size	512
Repeat weight	0.01
Embedding dimension	1024
Slim	True
Convolutional blocks	64
Convolutional kernel size (first block)	9
Convolutional kernel size (remaining blocks)	5
Convolutional dilation schedule	$1, 2, 4, 8, 16, 32, 64, 128, 1, \ldots$
Optimizer	AdamW
Weight decay	0.01
Batch size	2048
Steps	370 K
Learning rate	10^{-3}
Learning rate warmup	1 K steps

B.3 Feature extraction

Functional-genomics-supervised models. Let $\boldsymbol{y}_i \in \mathbb{R}_+^L$ be the predicted activity for genomic track i in each of L spatial positions. The "ℓ_2 score" (Linder et al., 2025) is defined as the norm of the log-fold-change between the predicted activity for the reference vs. alternate sequences:

$$\ell_2 \text{ score}_i := \left\| \log_2\left(1 + \boldsymbol{y}_i^{(\text{alt})}\right) - \log_2\left(1 + \boldsymbol{y}_i^{(\text{ref})}\right) \right\| \tag{1}$$

We define the "ℓ_2 of ℓ_2 score" as the norm of the ℓ_2 scores across tracks in a set $\mathbb{A}$ (e.g. all genomic tracks, or all genomic tracks from the same experimental assay):

$$\ell_2 \text{ of } \ell_2 \text{ score}(\mathbb{A}) := \|(\ell_2 \text{ score}_i, i \in \mathbb{A})\| \tag{2}$$

For Sei we used the official scores provided in their web server https://hb.flatironinstitute.org/sei.

Self-supervised models. We compute the log-likelihood ratio between the reference and alternate alleles:

$$\log \frac{\mathbb{P}(\text{alt})}{\mathbb{P}(\text{ref})} \tag{3}$$

For masked language models, it can be computed from the output probabilities when the variant position is masked. For autoregressive models (HyenaDNA and Evo2), it can be computed from the likelihood of the entire reference and alternate sequences. We also compute similarity in the embedding space. Let $\boldsymbol{Z} \in \mathbb{R}^{D \times L}$ be the sequence embedding with D hidden dimensions and L spatial positions. For HyenaDNA, an autoregressive model, we take the embedding of the rightmost position (could be interpreted as $L = 1$). We compare the reference and alternate embedding using the Euclidean distance:

$$\left\| \boldsymbol{Z}^{(\text{ref})} - \boldsymbol{Z}^{(\text{alt})} \right\|_F \tag{4}$$

cosine distance:

$$1 - \frac{\langle \boldsymbol{Z}^{(\text{ref})}, \boldsymbol{Z}^{(\text{alt})} \rangle_F}{\left\| \boldsymbol{Z}^{(\text{ref})} \right\|_F \left\| \boldsymbol{Z}^{(\text{alt})} \right\|_F} \tag{5}$$

and inner product:

$$\langle \boldsymbol{Z}^{(\text{ref})}, \boldsymbol{Z}^{(\text{alt})} \rangle_F \tag{6}$$

To obtain a high-dimensional featurization of a variant we calculate the inner product separately for each individual hidden dimension d:

$$\langle \boldsymbol{Z}_{d:}^{(\text{ref})}, \boldsymbol{Z}_{d:}^{(\text{alt})} \rangle \tag{7}$$

For both functional-genomics-supervised and self-supervised models, we always average the predictions using the forward vs. reverse strand, to ensure reverse-complement invariance.

B.4 Linear probing

We train a ridge logistic regression classifier pipeline using scikit-learn (Pedregosa et al., 2011), using default arguments as much as possible (Listing 1). The pipeline starts with imputation (only relevant for CADD input features) and standardization. To choose the regularization hyperparameter, we do a grid search using group K-fold cross-validation, with the groups consisting of the training chromosomes. We use the default number (10)of grid points, but shift the range to allow for heavier regularization given that our regression setting is very high-dimensional.

We repeat the entire pipeline training on all but one chromosome and predicting on the held-out chromosome. At the end we obtain predictions for all chromosomes, but each from a separate logistic regression model. Therefore, instead of calculating a global AUPRC, we calculate the AUPRC within each chromosome, and then perform a weighted average based on sample size. To obtain a standard error, we calculate the standard deviation of the distribution of weighted means performed on 1000 bootstrap samples of chromosomes. To allow easy comparison, we also use the weighted average AUPRC to evaluate zero-shot scores, even though it is not strictly necessary.

We only evaluate zero-shot scores on the full version of the datasets. We obtain standard errors from 100 bootstrap samples within the positive and negative sets, in order to maintain the proportion of positives.

```python
from sklearn.impute import SimpleImputer
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import GroupKFold, GridSearchCV
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler

def train_logistic_regression(X, y, groups):
    pipeline = Pipeline([
        ('imputer', SimpleImputer(
            missing_values=np.nan, strategy='mean',
            keep_empty_features=True,
        )),
        ('scaler', StandardScaler()),
        ('linear', LogisticRegression(
            class_weight="balanced",
            random_state=42,
        ))
    ])
    Cs = np.logspace(-8, 0, 10)
    param_grid = {
        'linear__C': Cs,
    }
    clf = GridSearchCV(
        pipeline,
        param_grid,
        scoring="average_precision",
        cv=GroupKFold(),
        n_jobs=-1,
    )
    clf.fit(X, y, groups=groups)
    return clf
```

Listing 1: Logistic regression classifier (the default penalty is ℓ_2).

C Additional Tables and Figures

Table S7: ClinVar "Pathogenic" variant consequences (reviewed by expert panel or practice guideline). ClinVar release: 20240909.

consequence	count
stop_gained	1687
missense_variant	988
splice_donor_variant	177
splice_acceptor_variant	157
start_lost	33
splice_region_variant	23
splice_donor_5th_base_variant	22
splice_polypyrimidine_tract_variant	20
splice_donor_region_variant	13
intron_variant	6
synonymous_variant	5
stop_lost	3
3_prime_UTR_variant	1
upstream_gene_variant	1

Table S8: AUPRC for different gLM zero-shot scores. In boldface: scores within 1% of best score (for a given model).

		LLR	abs(LLR)	L2 dist.	Cosine dist.	Inner prod.
Dataset	Model					
Mendelian traits	GPN-MSA	**0.694**	0.654	0.207	0.208	0.301
	GPN-Promoter	**0.422**	0.379	0.345	0.263	0.169
	NT	0.120	0.098	**0.188**	**0.186**	**0.185**
	HyenaDNA	0.115	0.106	0.117	0.116	**0.165**
	Caduceus	0.108	0.088	**0.135**	**0.135**	**0.131**
	SpeciesLM	0.201	0.161	**0.327**	**0.325**	0.095
	AIDO.DNA	0.135	0.115	0.148	0.146	**0.173**
	Evo2 1B	0.188	0.159	**0.254**	0.118	0.165
	Evo2 7B	**0.385**	0.340	0.351	0.140	0.161
	Evo2 40B	**0.557**	0.522	0.463	0.128	0.153
Complex traits	GPN-MSA	0.212	**0.224**	0.150	0.150	0.177
	GPN-Promoter	0.112	0.110	**0.126**	**0.126**	**0.125**
	NT	0.101	0.100	0.118	0.119	**0.136**
	HyenaDNA	**0.110**	**0.111**	0.102	0.102	**0.118**
	Caduceus	0.098	0.097	**0.115**	**0.115**	**0.117**
	SpeciesLM	0.105	0.103	**0.135**	**0.133**	0.093
	AIDO.DNA	**0.106**	**0.103**	**0.102**	**0.102**	**0.109**
	Evo2 1B	0.101	0.098	**0.116**	0.105	**0.122**
	Evo2 7B	**0.118**	**0.115**	**0.124**	0.109	**0.123**
	Evo2 40B	**0.137**	**0.139**	**0.131**	0.108	0.118

Table S9: Selected zero-shot approach for each gLM.

	Mendelian traits	Complex traits
GPN-MSA	LLR	abs(LLR)
GPN-Promoter	LLR	L2 dist.
NT	L2 dist.	Inner prod.
HyenaDNA	Inner prod.	Inner prod.
Caduceus	L2 dist.	Inner prod.
SpeciesLM	L2 dist.	L2 dist.
AIDO.DNA	Inner prod.	Inner prod.
Evo 2 1B	L2 dist.	Inner prod.
Evo 2 7B	LLR	L2 dist.
Evo 2 40B	LLR	abs(LLR)

Table S10: Number of overlapping variants with CADD training set.

	CADD training positives	CADD training negatives
Mendelian traits positives	0	0
Mendelian traits negatives	18	19
Complex traits positives	8	1
Complex traits negatives	79	55

Table S11: AUPRC for selected traits (at least 10 causal variants and less than 10% overlap of causal variants with other traits). The best score is reported between zero-shot and linear probing. Full feature ensemble is evaluated for complex traits, but lightweight feature ensemble is evaluated for Mendelian traits. In boldface: scores within 1% of best score.

	Borzoi	GPN-MSA	CADD	Ensemble
Mendelian traits				
Hyperferritinemia	0.315	0.965	**0.981**	**0.985**
Beta-thalassemia	0.927	0.796	0.926	**0.955**
Pulmonary fibrosis	0.564	0.948	**1.000**	**1.000**
Hemophilia B	0.914	0.709	**1.000**	**0.991**
Cartilage-hair hypoplasia	0.594	**0.987**	0.923	0.918
Preaxial polydactyly II	0.546	0.959	**0.969**	**0.967**
Hypercholesterolemia-1	0.844	**0.974**	0.887	0.938
Dwarfism (MOPD1)	0.484	**1.000**	**1.000**	**1.000**
Complex traits				
Adult height	0.292	0.383	**0.407**	0.339
Platelet count	0.426	0.309	0.397	**0.478**
Estimated heel bone mineral density	0.308	**0.432**	0.422	0.406
Mean corpuscular volume	0.434	0.319	0.391	**0.454**
Monocyte count	**0.561**	0.404	0.375	0.535
Hemoglobin A1c	0.475	0.375	0.426	**0.517**
Albumin/Globulin ratio	0.455	0.431	0.516	**0.559**
High density lipoprotein cholesterol	0.521	0.362	0.425	**0.554**
Estimated glomerular filtration rate (cystain C)	0.457	0.456	0.421	**0.470**
Alkaline phosphatase	**0.492**	0.292	0.352	0.446
Gamma-glutamyl transferase	0.515	0.382	0.460	**0.527**
FEV1/FVC ratio	0.430	0.494	**0.505**	0.487
Pulse pressure	0.457	0.435	0.420	**0.489**
Calcium	**0.468**	0.433	0.425	0.408
Albumin	**0.615**	0.544	0.480	0.602
Body mass index	0.344	**0.514**	0.436	0.499
Balding Type 4	0.459	0.536	0.414	**0.625**
Blood clot in the leg	**0.574**	0.551	0.498	**0.565**

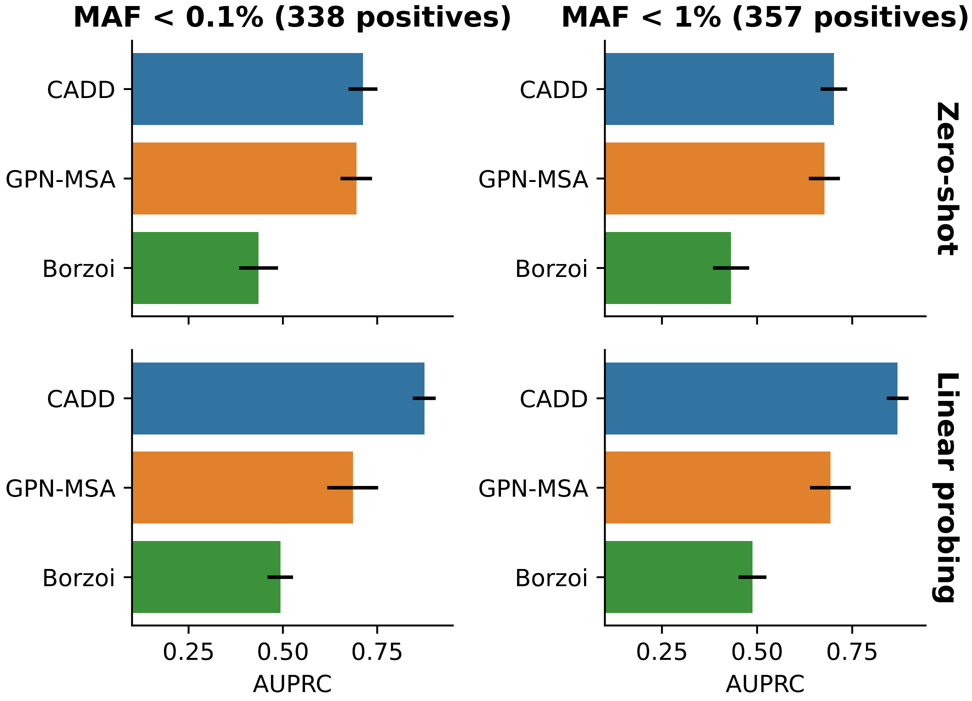

Figure S1: Ablation of MAF cutoff for positive variants in Mendelian traits dataset.

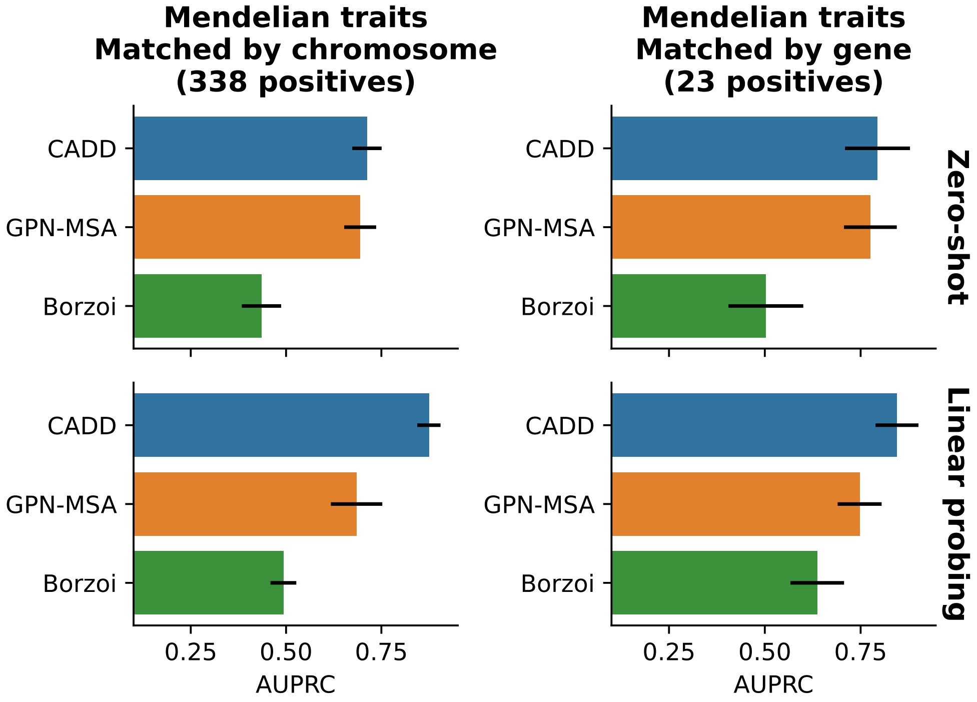

Figure S2: Mendelian traits results when positive variants are additionally matched by gene (variants that cannot be matched are dropped).

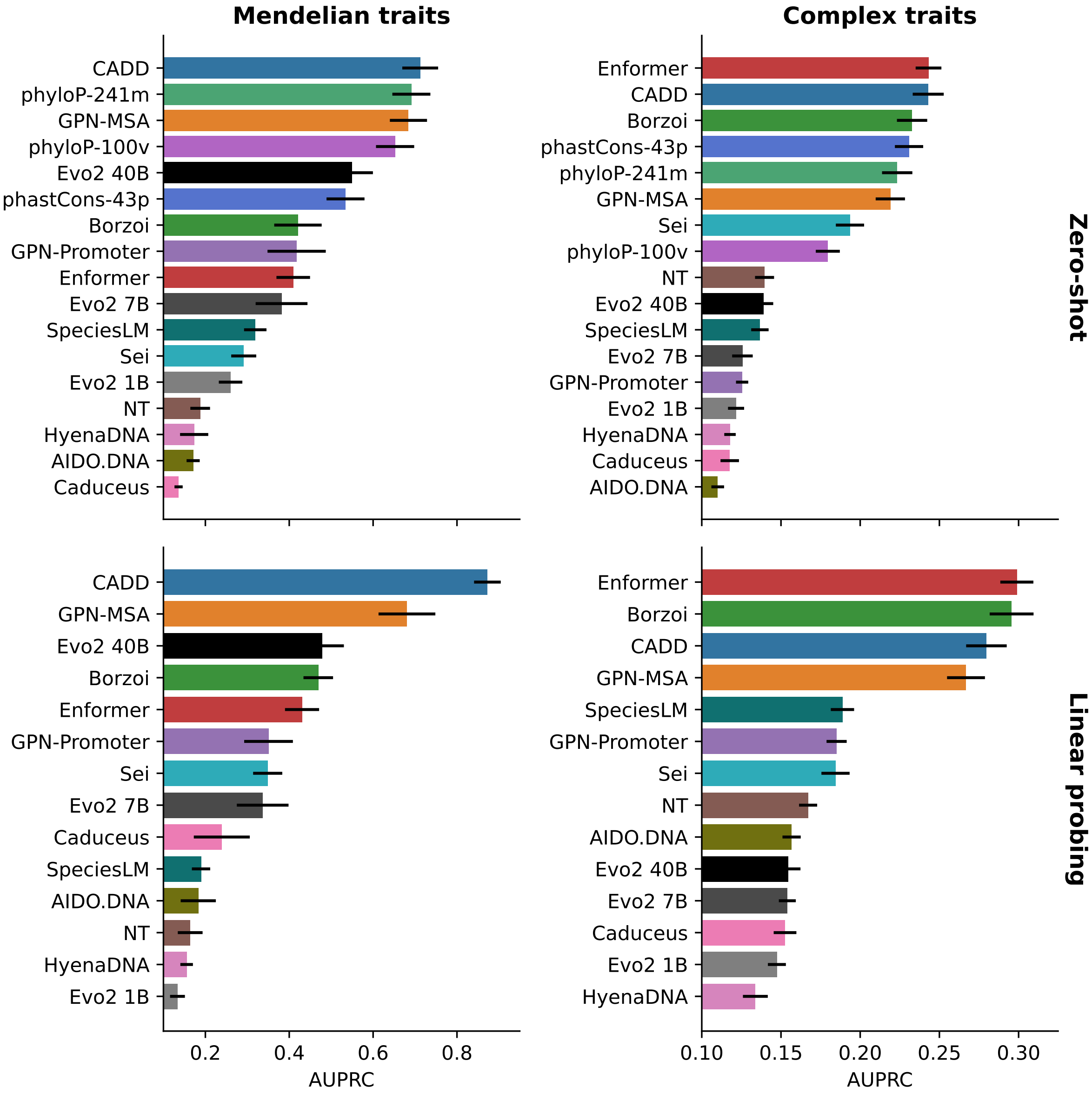

Figure S3: Results after removing a small amount of variants overlapping CADD training set.

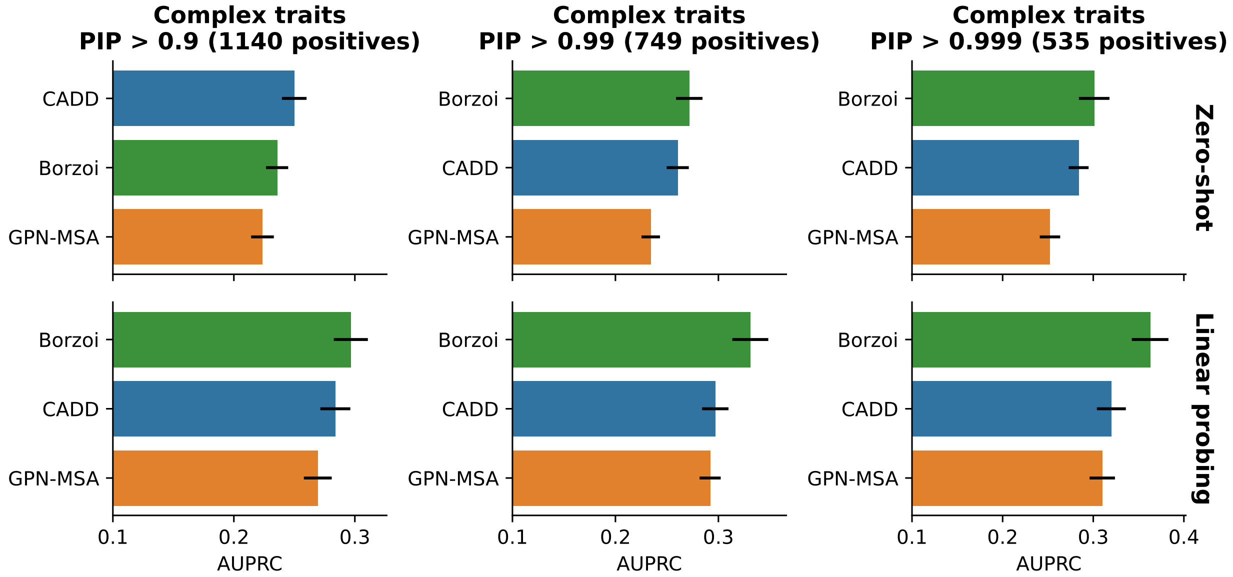

Figure S4: Results varying the PIP threshold for positive variants.

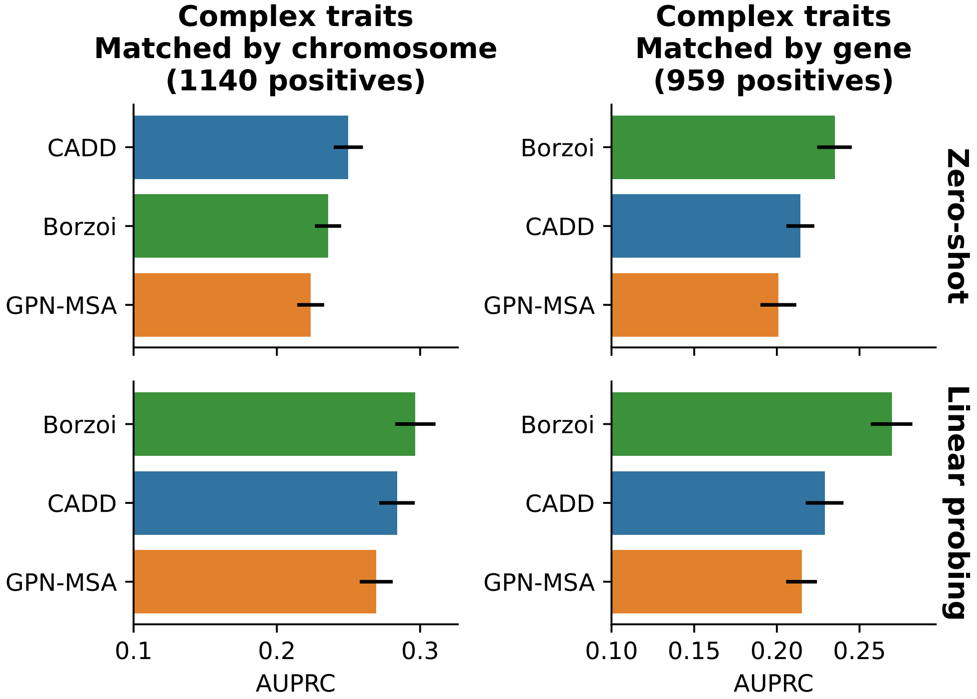

Figure S5: Complex traits results when positive variants are additionally matched by gene (variants that cannot be matched are dropped).

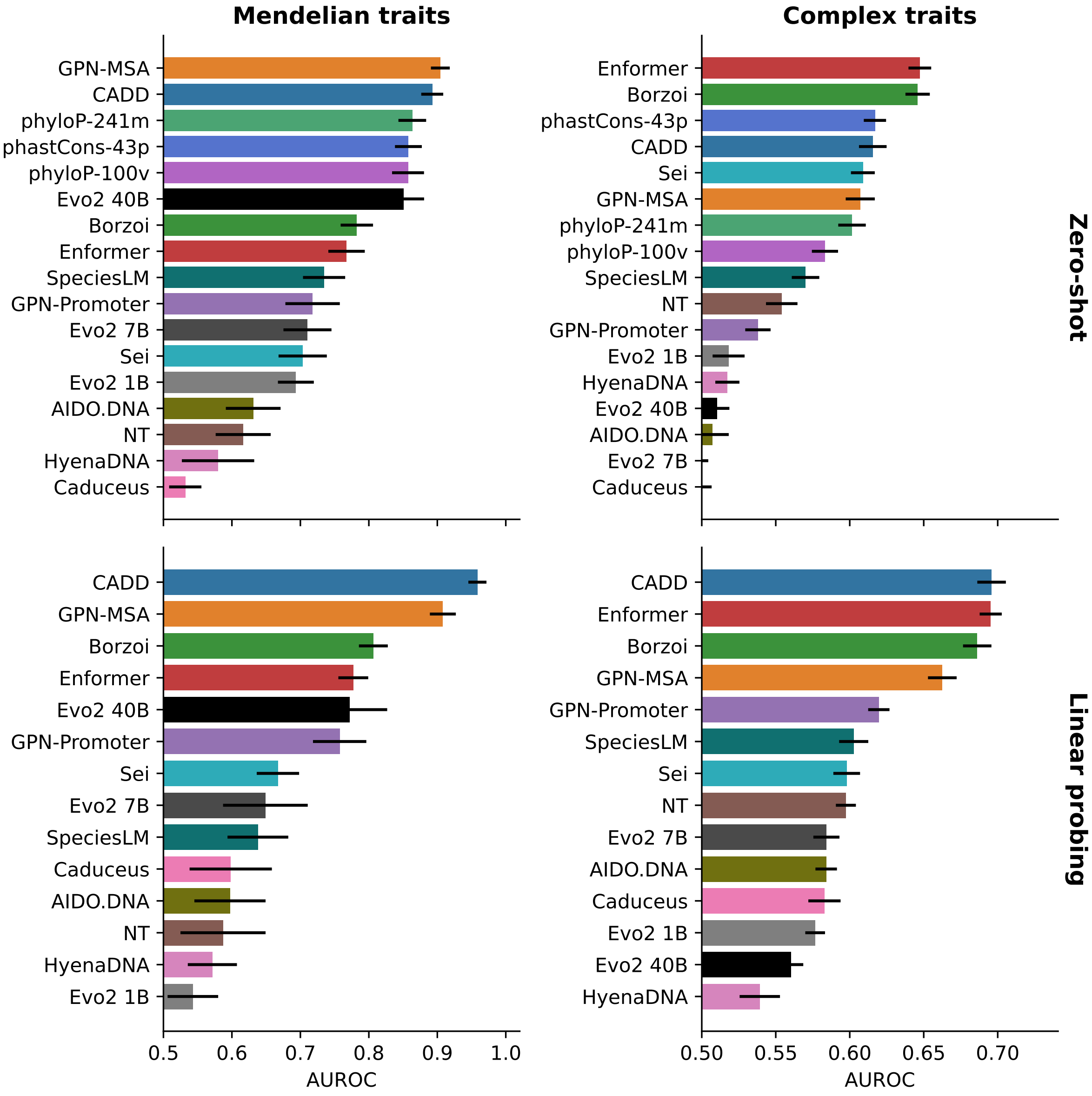

Figure S6: Results using the AUROC metric.

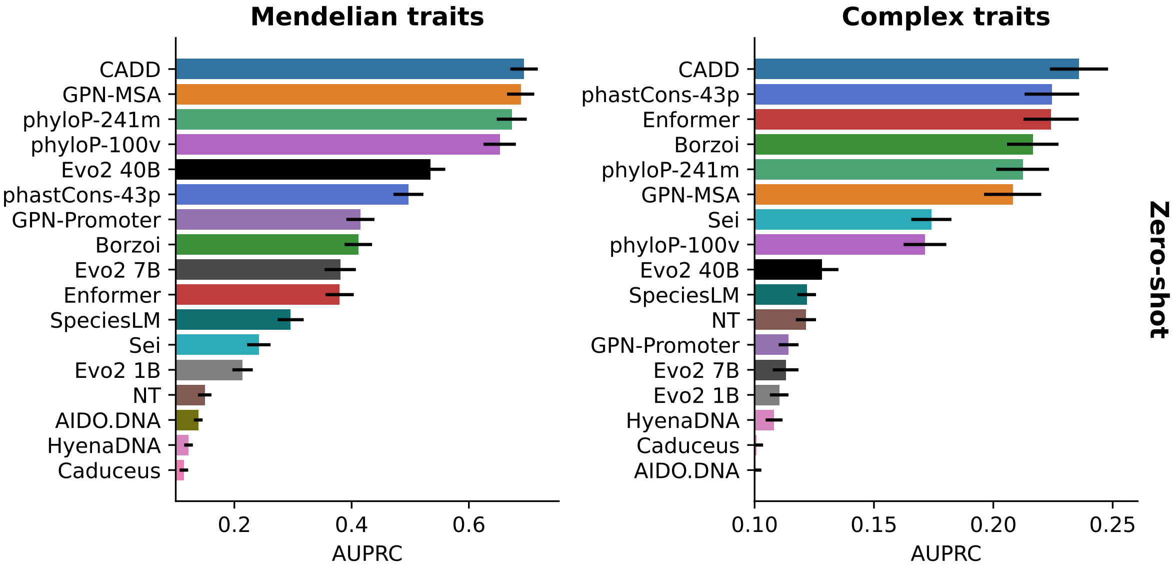

Figure S7: Results using the global AUPRC metric.

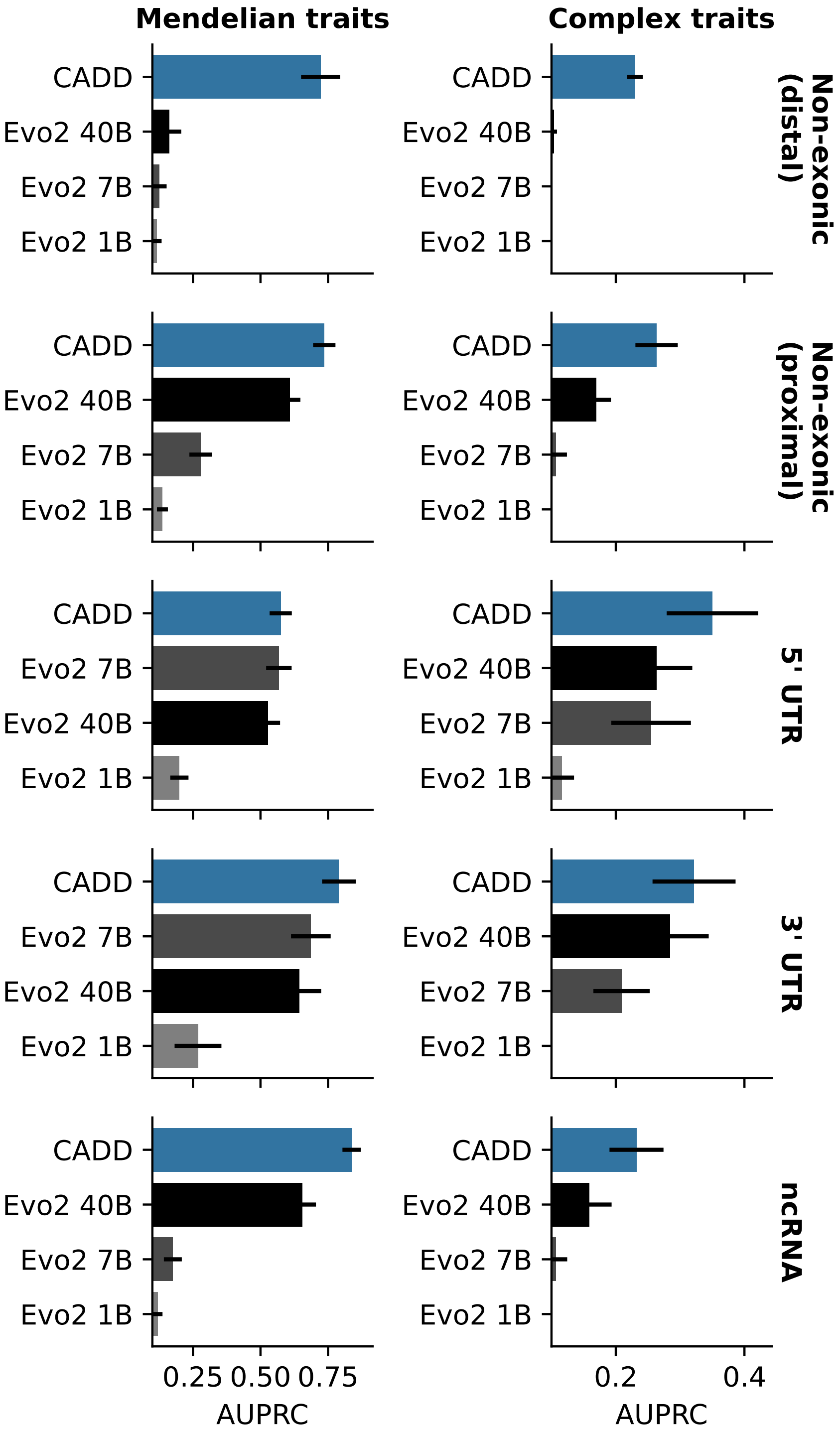

Figure S8: Evo2 zero-shot LLR results by consequence, with CADD as a baseline. LLR is used for Mendelian traits and abs(LLR) for complex traits. Metric is global AUPRC.

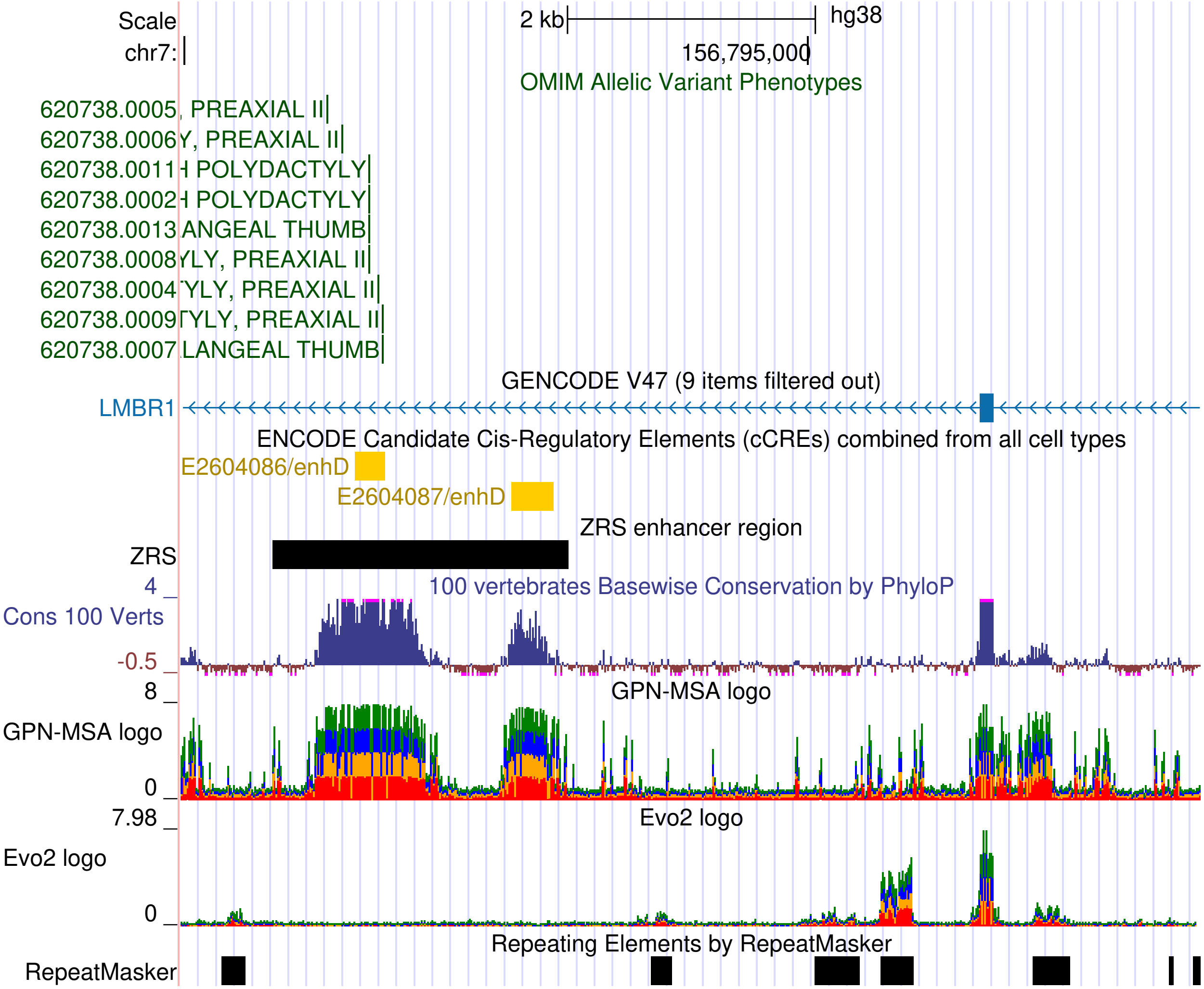
Scale
2 kb
hg38
chr7:
156,795,000
OMIM Allelic Variant Phenotypes
620738.0005 , PREAXIAL II
620738.0006 Y, PREAXIAL II
620738.0011 H POLYDACTYLY
620738.0002 H POLYDACTYLY
620738.0013 ANGEAL THUMB
620738.0008 YLY, PREAXIAL II
620738.0004 TYLY, PREAXIAL II
620738.0009 TYLY, PREAXIAL II
620738.0007 LANGEAL THUMB
GENCODE V47 (9 items filtered out)
LMBR1
ENCODE Candidate Cis-Regulatory Elements (cCREs) combined from all cell types
E2604086/enhD
E2604087/enhD
ZRS enhancer region
ZRS
4
100 vertebrates Basewise Conservation by PhyloP
Cons 100 Verts
-0.5
8
GPN-MSA logo
GPN-MSA logo
0
7.98
Evo2 logo
Evo2 logo
0
Repeating Elements by RepeatMasker
RepeatMasker

Figure S9: Evo2 40B logo plot around ZRS enhancer in the UCSC Genome Browser (Perez et al., 2025). The live genome browser session is available at https://genome.ucsc.edu/s/gbenegas/Evo2. Predicted logits were averaged between the forward and reverse strand.

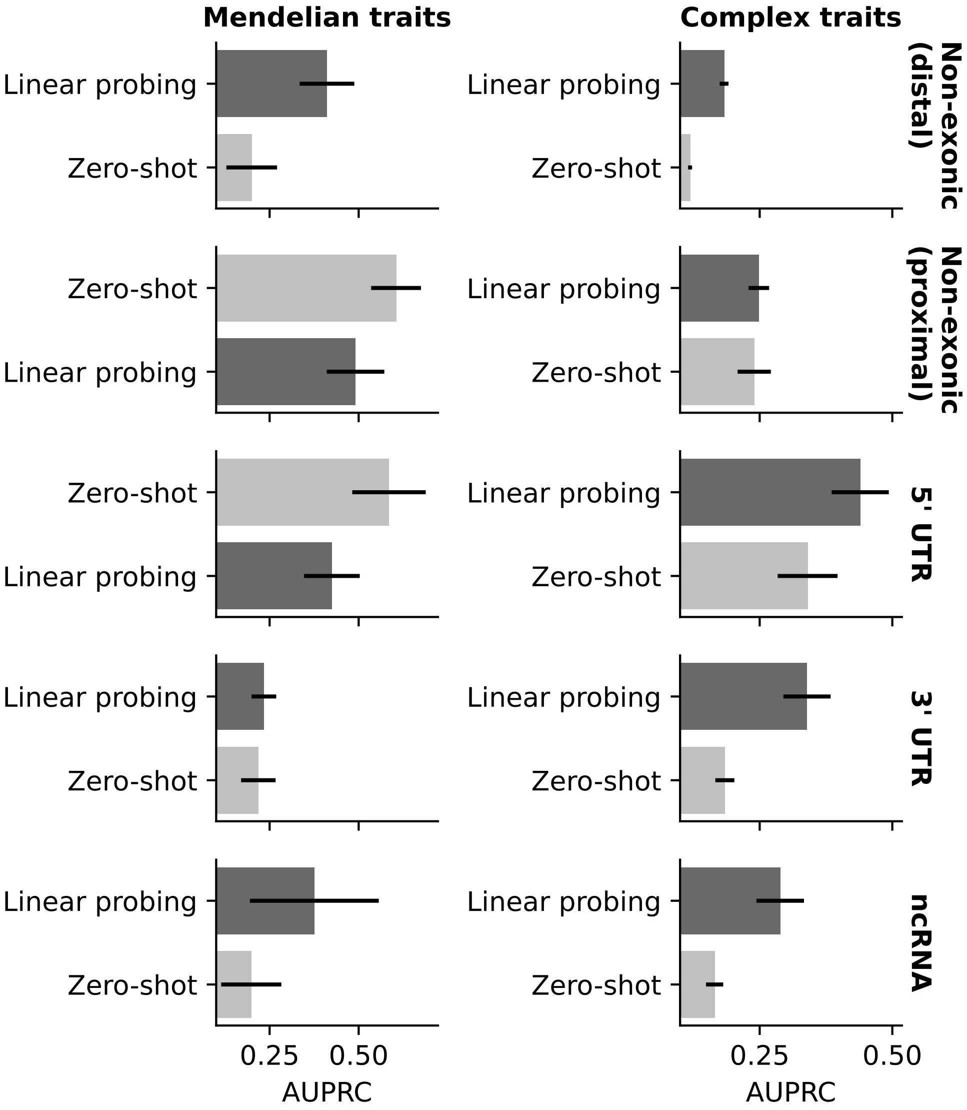

Figure S10: GPN-Promoter results by consequence.

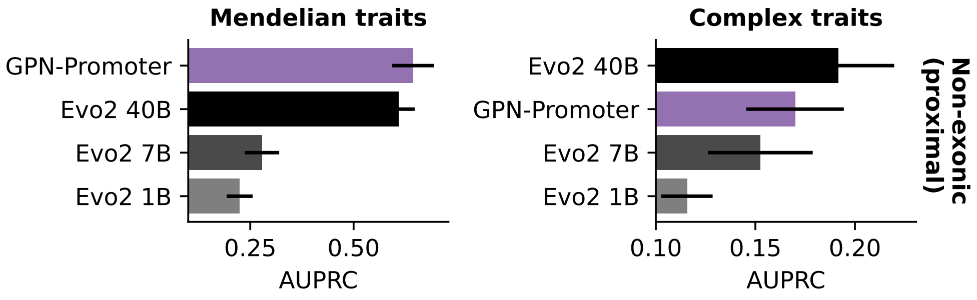

Figure S11: Global AUPRC for zero-shot results on non-exonic-proximal variants. The best performance is reported between LLR (or abs(LLR) in complex traits) and embedding Euclidean distance.