HUMAN-SPECIFIC VARIATION COMPLEMENTS BIOLOGICAL LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern biological language models are trained on vast sequence corpora representing organisms across the tree of life, enabling models to learn representations of biological sequences that generalize across a wide range of evolutionarily disparate organisms. However, for deepening understanding of human biology and disease, a human-centric evolutionary timescale is important. Human mutational constraint identifies genomic regions that have only recently been subject to selection within the human lineage and may have functional consequence for human-specific traits and diseases. We show that across modalities (DNA, RNA, and proteins), state-of-the-art biological language models capture deep evolutionary conservation but are weak proxies of human constraint. Furthermore, human constraint provides complementary information to the models tested, improving their ability to identify common functional variation with effects on both organismal and molecular phenotypes, despite the models' large scale pre-training. We additionally demonstrate that state-of-the-art MSA-based and sequence-to-function models can also be improved by explicit incorporation of metrics of human constraint. These results point to an unmet need to develop methods that model biological sequences not just across deep evolutionary divergence but also at a human-focused timescale.

1 INTRODUCTION

Models trained using self-supervised objectives on immense unstructured text corpora have rapidly become the dominant paradigm in language modeling, driving advances in both natural language processing as well as modeling of biological sequences. These biological language models (BLMs) form the foundation of advancements in open problems across modalities: genome annotation and non-coding variant effect prediction in DNA (Brixi et al., 2026; Benegas et al., 2025a); secondary and tertiary structure prediction in RNA (Penić et al., 2025; Shen et al., 2024); and structure prediction, missense variant effect prediction, and rational design in proteins (Lin et al., 2023; Brandes et al., 2023; Tran et al., 2026). A key principle underlying these models is that natural selection, or evolution, drives which biological sequences we observe and thus which sequences are used to train these models. Hence, we expect and assert (Hayes et al., 2025) that these models learn general principles of biological fitness. However, evolution acts on many timescales, so it's important to ask *what aspects of evolution do these models actually learn?*

Biological language model training efforts have focused on obtaining the largest possible sequence datasets, both to maximize the number of training tokens as well as to capture broad swathes of biology. By using large data repositories containing sequences from across the tree of life (Yates et al., 2026; The RNAcentral Consortium, 2026; The UniProt Consortium, 2025), models are able to learn biological principles shared across deep evolutionary timescales, from prokaryotic bacteria to multicellular eukaryotes, including humans (Brixi et al., 2026). However, these models are increasingly being applied to deepen our understanding of human biology, particularly in the context of human disease (Brandes et al., 2023; Benegas et al., 2025a; Cheng et al., 2023), raising the importance of more human-centric timescales. This importance is underlined by the fact that many genomic loci regulate human-specific traits and diseases, which have only recently been subject to evolutionary selection within the human lineage (Chen et al., 2024; Mangan et al., 2022; di Iulio et al., 2018). With this motivation, we believe it is important to probe how well current BLMs capture not only deep evolutionary conservation, but also mutational constraint within humans (Figure 1).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

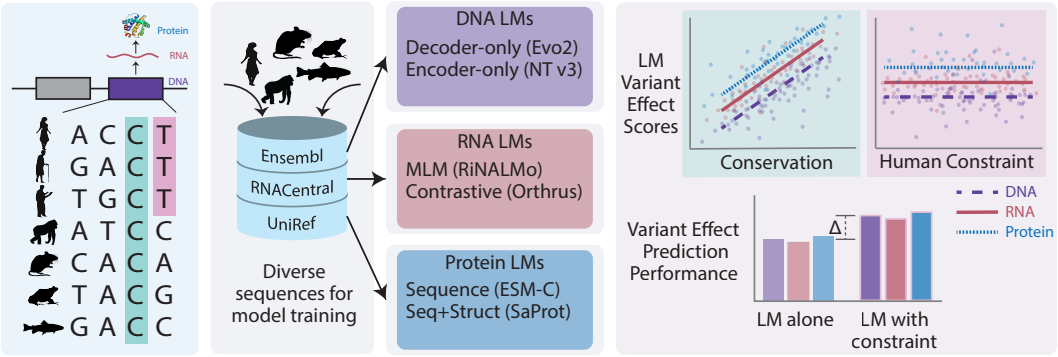

Figure 1: **(Left) Conservation at different timescales.** *Conservation* across species can indicate shared function, whereas mutational *constraint within humans* alone may indicate more recently evolved, human-specific functionality. **(Middle, Right) Biological language models are poor proxies for human constraint.** Models trained on large sequence corpora representing many diverse species produce variant effect predictions that capture broad conservation but not human mutational constraint. Explicit incorporation of human constraint improves variant effect predictions.

In this work, we ask how well biological language models capture different evolutionary timescales, dissecting whether models across DNA, RNA, and protein modalities encode both deep evolutionary conservation as well as human mutational constraint. In particular, we use phyloP scores derived from multiple alignment of 447 mammalian genomes to represent deep evolutionary conservation (Pollard et al., 2010; Armstrong et al., 2020; Genereux et al., 2020) and Gnocchi z-scores derived from whole genome sequencing of 76,156 humans to represent human mutational constraint (Chen et al., 2024). To assess a broad range of BLMs, we evaluate six state-of-the-art models, two per biological sequence modality (DNA, RNA, and proteins), with the models representing a range of architectures, training objectives, and model scales (Brixi et al., 2026; Boshar et al., 2025; Penić et al., 2025; Fradkin et al., 2026; Candido et al., 2026; Su et al., 2024). We then assess the utility of these methods for identifying functional variation in the human genome. We use ClinVar (Landrum et al., 2014) as a source of 22,796 rare, strong-effect variants causing Mendelian disease. To go beyond rare variation affecting a small number of individuals, we additionally curate 12,472 high-confidence disease- and gene expression-affecting common variants from the UK BioBank (Bycroft et al., 2018), All of Us (The All of Us Research Program Investigators, 2019), and the Genotype-Tissue Expression (GTEx) project (The GTEx Consortium, 2020). Our core findings are:

- Across modalities (DNA, RNA, and proteins), model architectures, and training methodologies, state-of-the-art BLMs capture deep evolutionary conservation but are poor proxies of human constraint
- Methods that capture deep conservation alone are sufficient for identifying strong-effect, rare variants, but failure to fully learn human constraint has real consequences for identifying common variants impacting human disease and gene expression
- Despite their large-scale pretraining and model sizes, BLMs' abilities to perform variant effect prediction are improved by explicit incorporation of constraint metrics
  - Across all BLMs tested, incorporating human constraint when predicting functional common variants results in average improvements of +11.9% and +25.6% to AUROC and AUPRC, respectively
- Additionally, recent state-of-the-art MSA- and sequence-to-function-based variant effect predictors (Ye et al., 2026; Cheng et al., 2026) display a similar bias towards deep conservation over human constraint, and similarly are improved upon by incorporating human constraint when predicting functional common variants (+4.4% AUROC, +6.6% AUPRC)

These findings point to an unmet need in the field: development of models that are effective across evolutionary timescales and thus better able to drive human-specific biological insight.

2 BACKGROUND

Predicting the phenotypic consequences of genomic variation is one of the key open problems in biology (IGVF Consortium, 2024) and has long been a major focus area for machine learning in biology. Not all genetic variants lead to disease. Therefore, distinguishing between functional and non-functional variants not only enables predicting disease risk, but also links genes to disease. Variant effect prediction methods range from supervised approaches with hand-selected features (Schubach et al., 2024), sequence-to-function models (Linder et al., 2025; Cheng et al., 2026), and self-supervised BLMs (Brandes et al., 2023; Benegas et al., 2025a). The language model approach in particular is thought to learn broad principles of biology from sequence alone (Candido et al., 2026) and has been applied across biological sequence modalities (Schiff et al., 2024; Yin et al., 2025; Aspidova et al., 2026; Lin et al., 2023). Evaluating these models' variant effect predictions commonly relies on benchmarks which emphasize rare or evolutionarily-conserved variation: ClinVar (Landrum et al., 2014) focuses on rare, strong effect, Mendelian disease-causing variants, while ProteinGym (Notin et al., 2023) and other protein fitness benchmarks (Rao et al., 2019; Dallago et al., 2021; Didi et al., 2026) assess models' abilities to predict mutational effects in proteins across a broad range of species. More recent benchmarks increasingly incorporate common human variation: TraitGym (Benegas et al., 2025b) introduced a benchmark including common trait-affecting variants and eQTL-based evaluation has become commonplace for sequence-to-function models (Avsec et al., 2021; Linder et al., 2025; Avsec et al., 2026).

In our work, we focus on using variant effect prediction tasks to ask if prevailing models capture human mutational constraint, rather than focusing on establishing a rank-ordering of models by performance on downstream tasks. We particularly focus on language models as these approaches are increasingly being applied across biological modalities, and if these models learn biological function by training on sequences sampled broadly from the tree of life, we should ask if these sequences represent the timescale that best enables improved disease prediction in humans.

3 METHODS

The core questions of our work are whether BLMs capture different evolutionary timescales, and if not, what consequences this shortcoming has for identifying functional variation within the human genome. To answer these questions, we require: (1) a set of state-of-the-art models with associated methods for extracting variant effect scores, (2) annotations of both deep evolutionary selection as well as recent human mutational constraint, and (3) sets of putatively functional and non-functional variants with varying levels of prevalence in the human population.

3.1 MODELS CONSIDERED

To perform a comprehensive evaluation and ensure the generality of our findings, we evaluate models that represent a range of architectures, training objectives and datasets, and importantly, distinct biological sequence modalities. We also include recent MSA- and sequence-to-function-based variant scoring methods both to investigate if our findings hold even beyond BLMs and also due to these methods having been shown to outperform language models for identifying functional variants in certain settings (Benegas et al., 2025b).

DNA models. We evaluate Evo2 7B (Brixi et al., 2026) and Nucleotide Transformer v3 (650M) (Boshar et al., 2025), which differ in terms of training objective (autoregressive versus bidirectional masked language modeling) and architecture (convolutional multi-hybrid versus hierarchical U-Net).

RNA models. We evaluate RiNALMo (650M) (Penić et al., 2025) and Orthrus (Fradkin et al., 2026), selected for their performance on VEP benchmarks (Shi et al., 2026). These models also represent distinct architectures (transformer- versus Mamba-based) and training objectives (masked language modeling versus contrastive learning). For Orthrus, we use the 6-track model trained with both contrastive and masked language objectives.

Protein models. We evaluate ESM-C (600M) (Candido et al., 2026) and SaProt (Su et al., 2024), providing a model that uses amino acid sequence only and a model that incorporates both amino acid sequence and tertiary protein structure.

MSA and sequence-to-function models. We use the MSA-based model GPN-Star (Ye et al., 2026), an approach that explicitly operates across evolutionary timescales both via explicitly modeling evolutionary divergence between species as well as deliberately restricting training data to species with a controlled level of divergence. We include all three versions of the model, trained using genomic sequences from vertebrates, mammals, and primates, respectively. For a sequence-to-function model, we utilized the recently released AlphaGenome Variant Impact (AVI) score (Cheng et al., 2026), which combines features from AlphaGenome's assay-level track predictions with additional features of conservation and protein loss-of-function annotations.

3.2 Datasets

3.2.1 Evolutionary annotations

Deep conservation. To capture deep evolutionary conservation, we use phyloP scores derived from a multiple alignment of 447 mammalian genomes, consisting of 247 mammalian genomes from the Zoonomia project and an additional 200 primate genomes (Pollard et al., 2010; Armstrong et al., 2020; Genereux et al., 2020; Christmas et al., 2023; Kuderna et al., 2023). These scores quantify per-base departure from the neutral substitution rate across roughly 110 million years of mammalian evolution. Positive values represent conservation under purifying selection, negative values represent acceleration relative to neutral evolution.

Human mutational constraint. To represent human mutational constraint, we use Gnocchi z-scores, which represent the deviation between the expected and observed number of mutations in 1kb windows along the human genome, calculated using 76,156 human genomes from gnomAD (Chen et al., 2024). Expected mutation rates are predicted based on trinucleotide context, DNA methylation at CpG sites, and genomic features such as GC content, repeat elements, and nucleosome density. Positive and negative values indicate stronger and weaker mutational constraint, respectively.

For clarity, we refer to phyloP scores as "Conservation" and Gnocchi z-scores as "Human constraint" in later sections.

3.2.2 Functional and non-functional human variation

Mendelian disease-causing variants. We utilize ClinVar (Landrum et al., 2014) as a dataset of variants with expert-annotated associations with Mendelian disease, or lack thereof. Beginning from the `20260208` release, we restrict to variants with high-confidence consequences by requiring a review status of 2 out of 4 stars or greater. We define functional and non-functional variants as those with classifications of "Pathogenic" or "Likely pathogenic" and "Benign" or "Likely benign", respectively.

GWAS disease-associated variants. To complement the strong effect variants in ClinVar, we curate a set of more common variants with putative functionality determined by association with complex diseases and traits from genome-wide association studies (GWAS). We obtain variants from two separate GWAS: (1) fine-mapping of 94 complex diseases and traits from 361,194 individuals from the UK Biobank (UKBB) (Kanai et al., 2021; Bycroft et al., 2018); and (2) fine-mapping of 14 quantitative traits from 199,039 individuals from the All of Us research program (Rossen et al., 2026; The All of Us Research Program Investigators, 2019). For the UKBB dataset, we remove variants associated with 9 of the 94 traits due to weak association with disease (Appendix A.1). We define functional and non-functional variants as those with posterior inclusion probabilities (PIP) greater than 0.5 and less than 0.1, respectively.

Gene-expression altering variants. We include expression quantitative trait loci (eQTLs) from the Genotype-Tissue Expression (GTEx) project (The GTEx Consortium, 2020; Wang et al., 2021) as another source of phenotype-affecting variation. Using the latest v11 release of fine-mapped cis-eQTLs, we define functional variants as those with PIP greater than 0.5 that are also the highest PIP variant within their credible set. Non-functional variants were defined as those with PIP less than 0.1 while also not the highest PIP variant within their credible set.

Variant filtering and transcript assignment. To allow comparison of biological language models across modalities, we restrict to variants contained within an exon of a protein-coding or non-coding gene. For eQTLs, this excludes variants that putatively affect gene expression by disrupting cis-regulatory elements (*e.g.* promoters or enhancers), but we retain a sufficient number of variants that affect the abundance of the biomolecules they are contained in (*e.g.* by affecting RNA stability).

Transcript overlap and predicted variant consequence was calculated with the Ensembl Variant Effect Predictor (VEP) (McLaren et al., 2016) tool with GENCODE Release 50 human gene annotations (Mudge et al., 2025). For each variant-gene pair, we select a single representative transcript, with priority given to MANE Select and Ensembl Canonical transcripts. We filter eQTLs to only those within exons of the genes whose expression they modulate. We further restrict to variants on autosomes only. For consistency across comparisons, we also remove any variants for which either evolutionary annotation was unavailable.

Counts of functional and non-functional variants per dataset are shown in Table 1 along with minor allele frequencies (MAFs) from gnomAD genomes v4.1. Full MAF distributions are shown in Appendix E.1. Detailed methods for dataset processing are available in Appendices A.1 and A.2.

Table 1: **Number of variants per functional and non-functional category per dataset.** Minor allele frequency (MAF) values are reported as median (25th, 75th percentiles). *Totals are not exactly equal to row sums due to variant overlaps across datasets.

	GWAS		eQTL		ClinVar		Total*
	High PIP	Low PIP	High PIP	Low PIP	Pathogenic	Benign	
				All variants			
N	1,655	118,213	10,817	66,229	22,796	140,353	324,073
MAF	0.083 (0.0094, 0.36)	0.19 (0.037, 0.43)	0.27 (0.11, 0.49)	0.25 (0.11, 0.48)	1.3×10^{-5} (7.0×10^{-6}, 2.6×10^{-5})	4.9×10^{-4} (2.6×10^{-5}, 0.018)	-
				Protein-coding missense			
N	683	12,116	820	4,669	9,885	24,304	47,752
MAF	0.014 (0.0032, 0.097)	0.045 (0.0024, 0.28)	0.23 (0.079, 0.46)	0.25 (0.10, 0.47)	1.3×10^{-5} (7.0×10^{-6}, 3.3×10^{-5})	0.0049 (4.8×10^{-4}, 0.023)	-

3.3 Experimental setup

Variant scoring. For each model, we generate variant effect scores as the difference in model likelihood between the reference and alternate alleles, such that more positive values indicate a substitution that is considered more "unlikely" or "disruptive". Scores are generated zero-shot, *i.e.* without any task-specific fine-tuning, and are generated according to the original authors' approaches wherever possible. Detailed scoring methods for each model are available in Appendix B.1.

Assessing conservation and constraint within models. We ask whether the considered models capture deep conservation and human constraint using two approaches. First, we quantify the correlation between model effect scores and evolutionary annotations across all the variant loci in our dataset using Pearson r and Spearman ρ, directly measuring whether model scores track with mutational constraint at different timescales. Second, we measure whether model scores are able to separate highly and poorly constrained loci. We partition our dataset into loci with extreme values of conservation or constraint, defined as loci with scores in the top or bottom quartile of each annotation's distribution, and use AUROC to quantify the ability of each model's scores to separate the extreme-valued loci.

Incorporating evolutionary annotations with model scores. We explore whether conservation- or constraint-based metrics show any complementarity with model scores for variant effect prediction by fitting logistic regression models with feature combinations and interactions using `scikit-learn` (Pedregosa et al., 2011). Models use 2nd-degree polynomial terms and are trained and evaluated using stratified 5-fold cross-validation with consistent folds across all models. Improvements are assessed via AUROC difference relative to the corresponding base model, estimated using a paired bootstrap with 1000 samples. Please see Appendix B.2 for detailed methods.

For all experiments, we show protein model performance on protein-coding missense variants only and DNA and RNA model performance on all variants in our dataset. See Appendix Table E.1 for DNA and RNA model performance on protein-coding missense variants only.

4 EXPERIMENTS

4.1 ASSESSING CONSERVATION AND HUMAN CONSTRAINT WITHIN MODELS

We first ask: **are BLMs able to capture both deep evolutionary conservation as well as human mutational constraint?** Across our dataset of 324,073 functional and non-functional variants (47,752 for protein models), we find that all models' scores correlate highly with deep conservation and show very low correlation with human mutational constraint, with median Pearson r and Spearman ρ values of 0.32/0.39 and 0.059/0.061 for conservation and constraint across models (Figure 2). When tasked with separating loci with conservation or constraint values in the extremes of their respective distributions, we find that models are able to distinguish high versus low conservation (mean AUROC $= 0.78$) but are only weakly able to distinguish high versus low human constraint (mean AUROC $= 0.55$) (Table 2). We also verify that phyloP scores and Gnocchi z-scores are uncorrelated (Appendix E.3, D). Taken together, these results support our first finding:

Takeaway 1: a diverse set of state-of-the-art BLMs—including different model architectures, training methodologies, and underlying biological modalities—are all highly correlated with deep conservation but are poor proxies of human-specific mutational constraint.

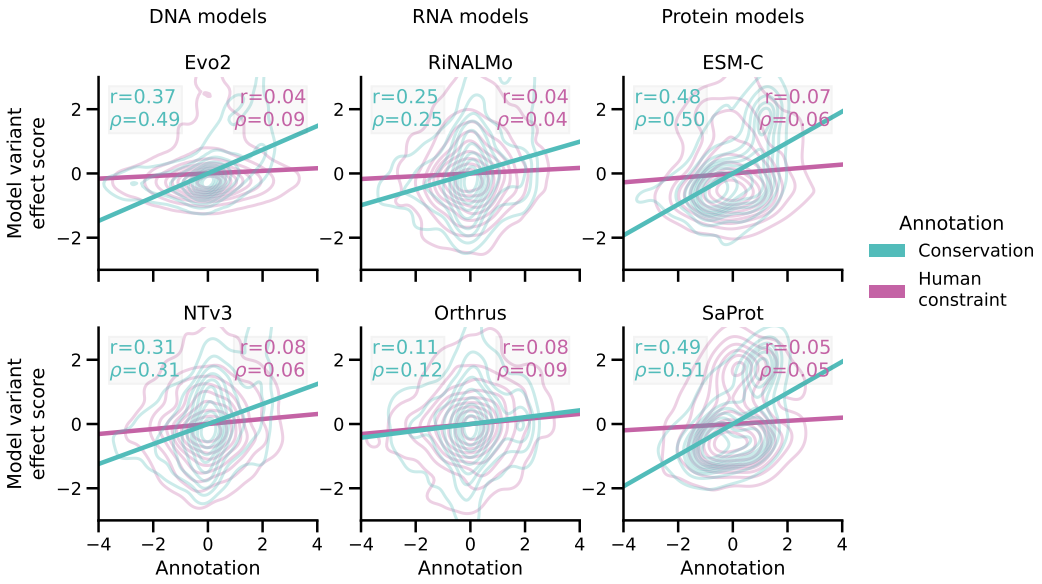

Figure 2: **Model scores correlated with evolutionary annotations.** Density of individual variants are indicated by contours. Lines display linear fits of annotation value to model score. Plotted values are z-scores for consistency with correlations, plots in original scale are in Appendix Figure E.2.

Table 2: **Model ability to separate highly conserved or human constrained loci.** Values shown are AUROCs for separating loci with either phyloP or Gnocchi scores in the top or bottom quartile of all loci, respectively.

Annotation	DNA		RNA		Protein		Average
	Evo2	NTv3	RiNALMo	Orthrus	ESM-C	SaProt	
Conservation	0.838	0.718	0.678	0.578	0.863	0.865	0.778
Human constrained	0.573	0.549	0.533	0.566	0.544	0.531	0.547

4.2 BLM AND EVOLUTIONARY ANNOTATION VARIANT EFFECT PREDICTION PERFORMANCE

Given that all evaluated models poorly captured human mutational constraint, our next question was: **what are the consequences of poor modeling of human constraint on identifying both rare and common functional variants?** Table 3 shows the ability of each evolutionary annotation and model score to distinguish functional variants across our three datasets. As it can be unclear whether the reference or alternate allele is more deleterious for GWAS and eQTL variants, we test both the raw zero-shot variant effect score as well as its absolute value, and report the higher of the two (Appendix Table E.2). Performance for strong-effect variants from ClinVar is generally high (median AUROC $= 0.892$), with some models such as Evo2, ESM-C, and SaProt achieving AUROCs of > 0.9. Performance for the common variants in the GWAS and eQTL datasets is much lower (GWAS median AUROC $= 0.579$; eQTL median AUROC $= 0.513$). AUPRC values show similar trends (Appendix Table E.2).

These results further reinforce our first finding, that deep conservation is the primary signal used by these models for variant effect prediction, in two ways: 1) model performance on datasets largely tracks the performance of phyloP, and 2) the models that displayed the highest correlation with phyloP scores in Figure 2 also display the highest performance on the ClinVar and GWAS datasets. Furthermore, the strong performance on ClinVar variants and weak performance on common functional variants leads to our second takeaway:

> **Takeaway 2**: assessing models using strong-effect rare variants gives a biased view that conservation alone is sufficient for VEP. Assessing on functional common variants shows that poor modeling of human constraint results in models missing true functional variants.

Table 3: **Model and annotation variant effect prediction performance.** Values shown are AUROC for distinguishing functional and non-functional variants across datasets. *Protein models evaluated only on missense variants.

Dataset	Conservation	Human constraint	DNA		RNA		Protein*	
			Evo2	NTv3	RiNALMo	Orthrus	ESM-C	SaProt
ClinVar	0.886	0.479	0.978	0.855	0.781	0.590	0.929	0.936
GWAS	0.695	0.593	0.671	0.538	0.508	0.503	0.619	0.631
eQTL	0.525	0.573	0.513	0.506	0.506	0.512	0.532	0.535

4.3 COMPLEMENTARITY OF EVOLUTIONARY ANNOTATIONS FOR VARIANT EFFECT PREDICTION

A natural next question to ask is: **can incorporating metrics of human mutational constraint improve model performance on variant effect prediction?** This could be true both for datasets where constraint is a distinguishing signal (*i.e.* eQTLs) and also for datasets where conservation is a useful signal (*i.e.* GWAS) but the possibility of complementarity remains (Table 3).

Indeed, when combining model scores with evolutionary annotations in a simple interaction model, we find that adding Gnocchi scores consistently improves variant effect prediction on the common variant datasets, albeit with stronger effects for DNA and RNA models (Figure 3). Mirroring the univariate results, improvements are largest on the eQTL dataset, although improvements for DNA and RNA models for GWAS variants are statistically significant as well (Appendix E.6). Surprisingly, we also find that incorporating phyloP improves the tested models' performance on both ClinVar and the GWAS dataset despite many of the models' strong correlation with phyloP scores (Appendix Figure E.6). Similarly positive, statistically significant improvements are seen in AUPRC (Appendix Table E.4). Improvements relative to human constraint alone show complementarity with the best performing models (Appendix E.9).

> **Takeaway 3**: despite training on trillions of biological tokens and containing up to billions of parameters (Brixi et al., 2026; Candido et al., 2026), prevailing BLMs' variant effect predictions can be improved by simple combination with explicit metrics of human constraint.

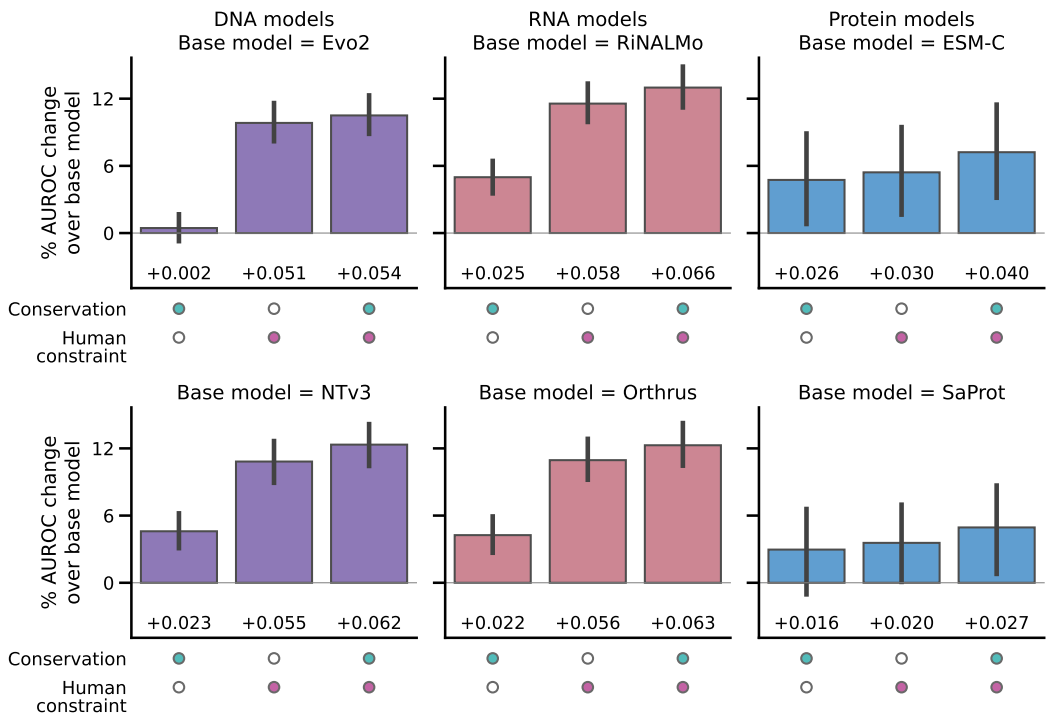

Figure 3: **Models are complemented by evolutionary annotations.** Performance assessed on the combined eQTL and GWAS datasets. Percent AUROC improvements are relative to the base model. Lines indicate 95% percentile intervals estimated via paired bootstrap. Absolute AUROC changes are listed below bars. Colored pips indicate which annotations are included in each combined model.

4.4 Extension to MSA and sequence-to-function models

Finally, as MSA- and sequence-to-function-based methods are also strong approaches for identifying common functional variants in the human genome, we asked: **do our previous findings extend to MSA and sequence-to-function models?** To this end, we performed similar analyses using the recently released GPN-Star and AlphaGenome Variant Impact (AVI) approaches (Cheng et al., 2026; Ye et al., 2026). We found that all of the additional models similarly showed high correlation with conservation and low-to-moderate correlation with human constraint (Figure 4, top; Appendix Table E.3). GPN-Star in particular shows the highest correlation with constraint out of all the models tested, possibly due to its explicit modeling of evolutionary divergence between species. Interestingly, we observe no meaningful trends with their increasingly human-centric models (*i.e.* vertebrate $\rightarrow$ mammal $\rightarrow$ primate). AVI scores show much higher phyloP correlation than other models, likely due to two of its constituent features (mammalian phyloP and vertebrate PhastCons) closely matching our selected metric of conservation.

Table 4: **MSA and sequence-to-function variant effect prediction performance.** Values are as in Table 3. We include Evo2 for reference as the best performing language model.

Dataset	Conservation	Human constraint	Evo2	GPN-Star			AlphaGenome VI
				V	M	P	
ClinVar	0.886	0.479	0.978	0.961	0.961	0.928	0.992
GWAS	0.695	0.593	0.671	0.716	0.696	0.685	0.743
eQTL	0.525	0.573	0.513	0.535	0.536	0.532	0.597

While both methods display strong performance for classifying functional variants in our datasets (Table 4), we still find that incorporation of explicit metrics of human constraint consistently leads to a modest, but statistically significant, improvement (Figure 4, bottom; Appendix Table E.4).

Takeaway 4: human mutational constraint should be incorporated into variant effect prediction methods, as even simple inclusion methods result in performance gains for both BLMs and other state-of-the-art approaches.

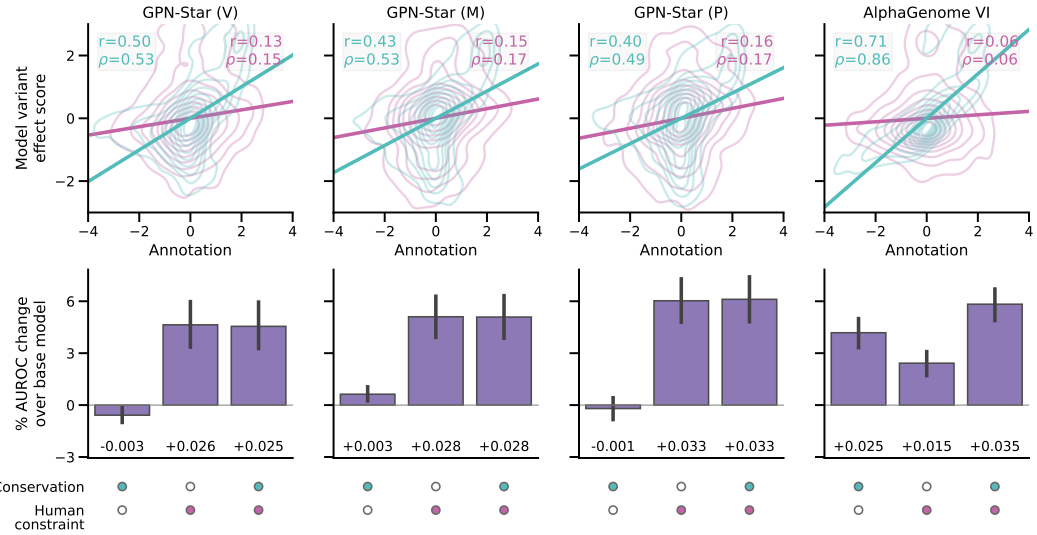

Figure 4: **MSA and sequence-to-function models similarly benefit from incorporation of human constraint.** Top: Model scores correlated with evolutionary annotations, as in Figure 2. Bottom: Percent AUROC improvement for distinguishing functional GWAS and eQTL variants when model scores are complemented with annotations, as in Figure 3.

5 Discussion

Here we demonstrate that prominent biological language models predict evolutionary conservation well across biomolecular modalities (DNA, RNA, protein), but do not adequately learn sequence constraint within humans. These selective properties propagate into variant effect prediction, as shown by high prediction accuracy for highly conserved, rare (median MAF $\approx 1\times10^{-5}$) Mendelian disease-causing variants (median AUROC $= 0.89$), whereas common (median MAF ≈ 0.08), less conserved functional variants are poorly predicted (median AUROC $= 0.52$). Importantly, common variant prediction by these models can be improved by incorporating human constraint, indicating that adding more recent evolutionary history of *Homo sapiens* might improve models' performance specifically for predicting common variant pathogenicity and molecular effects. These gains may be particularly pronounced for poorly conserved regions in the human genome, such as the $\approx 36,000$ long non-coding RNA genes.

Some limitations of our study restrict the generality of our findings and point to promising future directions. To enable comparison across modalities, we restricted to variants within transcribed regions of the genome; future work looking at non-transcribed regions with DNA-only models may be insightful due to evidence that some non-transcribed regulatory regions have undergone recent human-specific selection (Mangan et al., 2022). Additionally, our method of incorporating human constraint with model scores was purposefully simple, yet still resulted in improved variant effect prediction. Future work may garner further improvements with more complex methods such as post-training or reinforcement learning-based alignment (Widatalla et al., 2026).

In summary, our results expose a current shortcoming of state-of-the-art variant effect prediction methods in humans and show that incorporating prior knowledge of human-specific variation can improve insights in poorly conserved regions of the human genome.

AI USE STATEMENT

Large language models were used to assist in writing code used for running analyses. No language models or other generative AI tools were involved in initial drafts of the manuscript or figure generation. Large language models were used to edit specific sections of text for brevity and clarity.

ETHICS STATEMENT

This work involves the analysis of publicly available genomic sequence and genetic variation from established databases (GENCODE, UK Biobank, All of Us). All data used in this study are derived from previously published, publicly-available sources and no novel data generation was performed that involved human subjects, animal experiments, or the generation of new biological data that require ethical oversight. Our work assesses whether existing self-supervised machine learning models are able to learn human mutational constraint. Addressing these shortcomings could contribute to future models' advancement in drug discovery and biotechnology. We encourage responsible use of our methods and datasets, which are publicly available, to promote scientific reproducibility and advancement.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our work, we provide the following:

1. All code for our analyses, including variant curation, model scoring, collation of results, and evaluation, is available at `https://anonymous.4open.science/r/beyond-conservation/`.
2. We provide comprehensive methods details, including detailed dataset filtering steps and model scoring methods (Appendix A.1, A.2, B.1).
3. We provide download links for all external datafiles used in our study in the associated appendices.
4. All experiments can be reproduced using 1 NVIDIA A100 GPU.

REFERENCES

Joel Armstrong, Glenn Hickey, Mark Diekhans, Ian T. Fiddes, Adam M. Novak, Alden Deran, Qi Fang, Duo Xie, Shaohong Feng, Josefin Stiller, Diane Genereux, Jeremy Johnson, Voichita Dana Marinescu, Jessica Alföldi, Robert S. Harris, Kerstin Lindblad-Toh, David Haussler, Elinor Karlsson, Erich D. Jarvis, Guojie Zhang, and Benedict Paten. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature*, 587(7833):246–251, 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2871-y.

Alena Aspidova, Yuri Kuratov, Artem Shadskiy, Mikhail Burtsev, and Veniamin Fishman. Back to BERT in 2026: ModernGENA as a Strong, Efficient Baseline for DNA Foundation Models. In *ICLR 2026 Workshop on Foundation Models for Science: Real-World Impact and Science-First Design*, 2026.

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. ISSN 1548-7105. doi: 10.1038/s41592-021-01252-x.

Žiga Avsec, Natasha Latysheva, Jun Cheng, Guido Novati, Kyle R. Taylor, Tom Ward, Clare Bycroft, Lauren Nicolaisen, Eirini Arvaniti, Joshua Pan, Raina Thomas, Vincent Dutordoir, Matteo Perino, Soham De, Alexander Karollus, Adam Gayoso, Toby Sargeant, Anne Mottram, Lai Hong Wong, Pavol Drotár, Adam Kosiorek, Andrew Senior, Richard Tanburn, Taylor Applebaum, Souradeep Basu, Demis Hassabis, and Pushmeet Kohli. Advancing regulatory variant effect prediction with AlphaGenome. *Nature*, 649(8099):1206–1218, 2026. ISSN 1476-4687. doi: 10.1038/s41586-025-10014-0.

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*, 43(12):1960–1965, 2025a. ISSN 1546-1696. doi: 10.1038/s41587-024-02511-w.

Gonzalo Benegas, Gokcen Eraslan, and Yun S. Song. Benchmarking DNA Sequence Models for Causal Regulatory Variant Prediction in Human Genetics, 2025b.

Damian Bertoni, Maxim Tsenkov, Paulyna Magana, Sreenath Nair, Ivanna Pidruchna, Marcelo Querino Lima Afonso, Adam Midlik, Urmila Paramval, Dare Lawal, Ahsan Tanweer, Meera Last, Risha Patel, Agata Laydon, Dariusz Lasecki, Nick Dietrich, Hamish Tomlinson, Augustin Žídek, Tim Green, Oleg Kovalevskiy, Andy Lau, Shaun Kandathil, Nicola Bordin, Ian Sillitoe, Milot Mirdita, David Jones, Christine Orengo, Martin Steinegger, Jennifer R Fleming, and Sameer Velankar. AlphaFold Protein Structure Database 2025: A redesigned interface and updated structural coverage. *Nucleic Acids Research*, 54(D1):D358–D362, 2026. ISSN 1362-4962. doi: 10.1093/nar/gkaf1226.

Sam Boshar, Benjamin Evans, Ziqi Tang, Armand Picard, Yanis Adel, Franziska K. Lorbeer, Chandana Rajesh, Tristan Karch, Shawn Sidbon, David Emms, Javier Mendoza-Revilla, Fatimah Al-Ani, Evan Seitz, Yair Schiff, Yohan Bornachot, Ariana Hernandez, Marie Lopez, Alexandre Laterre, Karim Beguir, Peter Koo, Volodymyr Kuleshov, Alexander Stark, Bernardo P. de Almeida, and Thomas Pierrot. A foundational model for joint sequence-function multi-species modeling at scale for long-range genomic prediction, 2025. ISSN 2692-8205.

Nadav Brandes, Grant Goldman, Charlotte H. Wang, Chun Jimmie Ye, and Vasilis Ntranos. Genome-wide prediction of disease variant effects with a deep protein language model. *Nature Genetics*, 55(9):1512–1522, 2023. ISSN 1546-1718. doi: 10.1038/s41588-023-01465-0.

Garyk Brixi, Matthew G. Durrant, Jerome Ku, Mohsen Naghipourfar, Michael Poli, Gwanggyu Sun, Greg Brockman, Daniel Chang, Alison Fanton, Gabriel A. Gonzalez, Samuel H. King, David B. Li, Aditi T. Merchant, Eric Nguyen, Chiara Ricci-Tam, David W. Romero, Jonathan C. Schmok, Ali Taghibakhshi, Anton Vorontsov, Brandon Yang, Myra Deng, Liv Gorton, Nam Nguyen, Nicholas K. Wang, Michael T. Pearce, Elana Simon, Etowah Adams, Zachary J. Amador, Euan A. Ashley, Stephen A. Baccus, Haoyu Dai, Steven Dillmann, Stefano Ermon, Daniel Guo, Michael H. Herschl, Rajesh Ilango, Ken Janik, Amy X. Lu, Reshma Mehta, Mohammad R. K. Mofrad, Madelena Y. Ng, Jaspreet Pannu, Christopher Ré, John St. John, Jeremy Sullivan, Joseph Tey, Ben Viggiano, Kevin Zhu, Greg Zynda, Daniel Balsam, Patrick Collison, Anthony B. Costa, Tina Hernandez-Boussard, Eric Ho, Ming-Yu Liu, Thomas McGrath, Kimberly Powell, Sudarshan Pinglay, Dave P. Burke, Hani Goodarzi, Patrick D. Hsu, and Brian L. Hie. Genome modelling and design across all domains of life with Evo 2. *Nature*, 652(8112):1349–1361, 2026. ISSN 1476-4687. doi: 10.1038/s41586-026-10176-5.

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018. ISSN 1476-4687. doi: 10.1038/s41586-018-0579-z.

Salvatore Candido, Thomas Hayes, Alexander Derry, Roshan Rao, Zeming Lin, Robert Verkuil, Bryan Z. Wu, Jin Sub Lee, Elise S. Bruguera, Jehan A. Keval, Mykhailo Kopylov, John E. Pak, Wesley Wu, Neil Thomas, Samson Mataraso, Alvin Hsu, Ashton C. Trotman-Grant, Kilian Fatras, Allan dos Santos Costa, Rohil Badkundri, Halil Akın, Deniz Oktay, Jonathan Deaton, Elizabeth Montabana, Hrishita Sitwala, Yue Yu, Marius Wiggert, Dylan Alexander Carlin, Anthony W. Goering, Tomasz Blazejewski, McCullen Sandora, Michael Hla, Tina Z. Jia, Leon H. Kloker, Nicholas J. Sofroniew, Masatoshi Uehara, Jassi Pannu, Sharrol Bachas, Daniel S. Liu, Tom Sercu, and Alexander Rives. Language Modeling Materializes a World Model of Protein Biology, 2026. ISSN 2692-8205.

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, Timothy Poterba, Michael W. Wilson, Yekaterina Tarasova, William Phu, Riley Grant, Mary T. Yohannes, Zan Koenig, Yossi Farjoun, Eric Banks, Stacey Donnelly, Stacey Gabriel, Namrata Gupta, Steven

Ferriera, Charlotte Tolonen, Sam Novod, Louis Bergelson, David Roazen, Valentin Ruano-Rubio, Miguel Covarrubias, Christopher Llanwarne, Nikelle Petrillo, Gordon Wade, Thibault Jeandet, Ruchi Munshi, Kathleen Tibbetts, Anne O'Donnell-Luria, Matthew Solomonson, Cotton Seed, Alicia R. Martin, Michael E. Talkowski, Heidi L. Rehm, Mark J. Daly, Grace Tiao, Benjamin M. Neale, Daniel G. MacArthur, and Konrad J. Karczewski. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature*, 625(7993):92–100, 2024. ISSN 1476-4687. doi: 10.1038/s41586-023-06045-0.

Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, Rosalia G. Schneider, Andrew W. Senior, John Jumper, Demis Hassabis, Pushmeet Kohli, and Žiga Avsec. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, 381(6664): eadg7492, 2023. doi: 10.1126/science.adg7492.

Jun Cheng, Kyle R. Taylor, Lauren Nicolaisen, Joshua Pan, Clare Bycroft, Matteo Perino, Tom Ward, Gareth Hawkes, Laura E. Covill, Melanie Weilert, Raina W. Thomas, Natasha Latysheva, Maile J. Hirschmann, Xi Dawn Chen, Robin N. Beaumont, V. Kartik Chundru, Michael N. Weedon, Simon Bourdareau, Hoyin Chu, Dhavanthi Hariharan, Thais Kagohara, Lucas Tenório, Yosuke Ushigome, Courtney A. Shearer, Barbara Ikica, Ada Fang, Mouad Naciri, Victoria Johnston, Richard Green, Lai Hong Wong, Vincent Dutordoir, Anne Mottram, Adam Gayoso, Eirini Arvaniti, Guido Novati, Heidi L. Rehm, Fei Chen, Caleb A. Lareau, Caroline F. Wright, Anne O'Donnell-Luria, Julia Zeitlinger, Pushmeet Kohli, and Žiga Avsec. AlphaGenome Atlas: in silico mutagenesis of the entire human genome improves prioritization and interpretation of noncoding variants. Technical report, Google DeepMind, September 2026. Preprint.

Matthew J. Christmas, Irene M. Kaplow, Diane P. Genereux, Michael X. Dong, Graham M. Hughes, Xue Li, Patrick F. Sullivan, Allyson G. Hindle, Gregory Andrews, Joel C. Armstrong, Matteo Bianchi, Ana M. Breit, Mark Diekhans, Cornelia Fanter, Nicole M. Foley, Daniel B. Goodman, Linda Goodman, Kathleen C. Keough, Bogdan Kirilenko, Amanda Kowalczyk, Colleen Lawless, Abigail L. Lind, Jennifer R. S. Meadows, Lucas R. Moreira, Ruby W. Redlich, Louise Ryan, Ross Swofford, Alejandro Valenzuela, Franziska Wagner, Ola Wallerman, Ashley R. Brown, Joana Damas, Kaili Fan, John Gatesy, Jenna Grimshaw, Jeremy Johnson, Sergey V. Kozyrev, Alyssa J. Lawler, Voichita D. Marinescu, Kathleen M. Morrill, Austin Osmanski, Nicole S. Paulat, BaDoi N. Phan, Steven K. Reilly, Daniel E. Schäffer, Cynthia Steiner, Megan A. Supple, Aryn P. Wilder, Morgan E. Wirthlin, James R. Xue, Zoonomia Consortium, Bruce W. Birren, Steven Gazal, Robert M. Hubley, Klaus-Peter Koepfli, Tomas Marques-Bonet, Wynn K. Meyer, Martin Nweeia, Pardis C. Sabeti, Beth Shapiro, Arian F. A. Smit, Mark S. Springer, Emma C. Teeling, Zhiping Weng, Michael Hiller, Danielle L. Levesque, Harris A. Lewin, William J. Murphy, Arcadi Navarro, Benedict Paten, Katherine S. Pollard, David A. Ray, Irina Ruf, Oliver A. Ryder, Andreas R. Pfenning, Kerstin Lindblad-Toh, and Elinor K. Karlsson. Evolutionary constraint and innovation across hundreds of placental mammals. *Science*, 380(6643):eabn3943, 2023. doi: 10.1126/science.abn3943.

Christian Dallago, Jody Mou, Kadina Johnston, Bruce Wittmann, Nicholas Bhattacharya, Samuel Goldman, Ali Madani, and Kevin Yang. FLIP: Benchmark tasks in fitness landscape inference for proteins. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1, 2021.

Julia di Iulio, Istvan Bartha, Emily H. M. Wong, Hung-Chun Yu, Victor Lavrenko, Dongchan Yang, Inkyung Jung, Michael A. Hicks, Naisha Shah, Ewen F. Kirkness, Martin M. Fabani, William H. Biggs, Bing Ren, J. Craig Venter, and Amalio Telenti. The human noncoding genome defined by genetic diversity. *Nature Genetics*, 50(3):333–337, 2018. ISSN 1546-1718. doi: 10.1038/s41588-018-0062-7.

Kieran Didi, Sarah Alamdari, Alex Xijie Lu, Bruce James Wittmann, Kadina E. Johnston, Ava P. Amini, Ali Madani, Maya Czeneszew, Christian Dallago, and Kevin K. Yang. FLIP2: Expanding Protein Fitness Landscape Benchmarks for Real-World Machine Learning Applications. In *Forty-Third International Conference on Machine Learning*, 2026.

Philip Fradkin, Ruian "Ian" Shi, Taykhoom Dalal, Keren Isaev, Brendan J. Frey, Leo J. Lee, Quaid Morris, and Bo Wang. Orthrus: Toward evolutionary and functional RNA foundation models. *Nature Methods*, 23(5):935–945, 2026. ISSN 1548-7105. doi: 10.1038/s41592-026-03064-3.

Diane P. Genereux, Aitor Serres, Joel Armstrong, Jeremy Johnson, Voichita D. Marinescu, Eva Murén, David Juan, Gill Bejerano, Nicholas R. Casewell, Leona G. Chemnick, Joana Damas, Federica Di Palma, Mark Diekhans, Ian T. Fiddes, Manuel Garber, Vadim N. Gladyshev, Linda Goodman, Wilfried Haerty, Marlys L. Houck, Robert Hubley, Teemu Kivioja, Klaus-Peter Koepfli, Lukas F. K. Kuderna, Eric S. Lander, Jennifer R. S. Meadows, William J. Murphy, Will Nash, Hyun Ji Noh, Martin Nweeia, Andreas R. Pfenning, Katherine S. Pollard, David A. Ray, Beth Shapiro, Arian F. A. Smit, Mark S. Springer, Cynthia C. Steiner, Ross Swofford, Jussi Taipale, Emma C. Teeling, Jason Turner-Maier, Jessica Alfoldi, Bruce Birren, Oliver A. Ryder, Harris A. Lewin, Benedict Paten, Tomas Marques-Bonet, Kerstin Lindblad-Toh, Elinor K. Karlsson, and Zoonomia Consortium. A comparative genomics multitool for scientific discovery and conservation. *Nature*, 587(7833):240–245, 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2876-6.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf A. Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025. doi: 10.1126/science.ads0018.

IGVF Consortium. Deciphering the impact of genomic variation on function. *Nature*, 633(8028): 47–57, 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07510-0.

Masahiro Kanai, Jacob C. Ulirsch, Juha Karjalainen, Mitja Kurki, Konrad J. Karczewski, Eric Fauman, Qingbo S. Wang, Hannah Jacobs, François Aguet, Kristin G. Ardlie, Nurlan Kerimov, Kaur Alasoo, Christian Benner, Kazuyoshi Ishigaki, Saori Sakaue, Steven Reilly, The BioBank Japan Project, FinnGen, Yoichiro Kamatani, Koichi Matsuda, Aarno Palotie, Benjamin M. Neale, Ryan Tewhey, Pardis C. Sabeti, Yukinori Okada, Mark J. Daly, and Hilary K. Finucane. Insights from complex trait fine-mapping across diverse populations, 2021.

Lukas F. K. Kuderna, Hong Gao, Mareike C. Janiak, Martin Kuhlwilm, Joseph D. Orkin, Thomas Bataillon, Shivakumara Manu, Alejandro Valenzuela, Juraj Bergman, Marjolaine Rousselle, Felipe Ennes Silva, Lidia Agueda, Julie Blanc, Marta Gut, Dorien de Vries, Ian Goodhead, R. Alan Harris, Muthuswamy Raveendran, Axel Jensen, Idrissa S. Chuma, Julie E. Horvath, Christina Hvilsom, David Juan, Peter Frandsen, Joshua G. Schraiber, Fabiano R. de Melo, Fabrício Bertuol, Hazel Byrne, Iracilda Sampaio, Izeni Farias, João Valsecchi, Malu Messias, Maria N. F. da Silva, Mihir Trivedi, Rogerio Rossi, Tomas Hrbek, Nicole Andriaholinirina, Clément J. Rabarivola, Alphonse Zaramody, Clifford J. Jolly, Jane Phillips-Conroy, Gregory Wilkerson, Christian Abee, Joe H. Simmons, Eduardo Fernandez-Duque, Sree Kanthaswamy, Fekadu Shiferaw, Dongdong Wu, Long Zhou, Yong Shao, Guojie Zhang, Julius D. Keyyu, Sascha Knauf, Minh D. Le, Esther Lizano, Stefan Merker, Arcadi Navarro, Tilo Nadler, Chiea Chuen Khor, Jessica Lee, Patrick Tan, Weng Khong Lim, Andrew C. Kitchener, Dietmar Zinner, Ivo Gut, Amanda D. Melin, Katerina Guschanski, Mikkel Heide Schierup, Robin M. D. Beck, Govindhaswamy Umapathy, Christian Roos, Jean P. Boubli, Jeffrey Rogers, Kyle Kai-How Farh, and Tomas Marques Bonet. A global catalog of whole-genome diversity from 233 primate species. *Science*, 380(6648):906–913, 2023. doi: 10.1126/science.abn7829.

Melissa J. Landrum, Jennifer M. Lee, George R. Riley, Wonhee Jang, Wendy S. Rubinstein, Deanna M. Church, and Donna R. Maglott. ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 42(D1):D980–D985, 2014. ISSN 0305-1048. doi: 10.1093/nar/gkt1113.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

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. ISSN 1546-1718. doi: 10.1038/s41588-024-02053-6.

Riley J. Mangan, Fernando C. Alsina, Federica Mosti, Jesús Emiliano Sotelo-Fonseca, Daniel A. Snellings, Eric H. Au, Juliana Carvalho, Laya Sathyan, Graham D. Johnson, Timothy E. Reddy, Debra L. Silver, and Craig B. Lowe. Adaptive sequence divergence forged new neurodevelopmental enhancers in humans. *Cell*, 185(24):4587–4603.e23, 2022. ISSN 0092-8674. doi: 10.1016/j.cell.2022.10.016.

William McLaren, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1):122, 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-0974-4.

Jonathan M Mudge, Sílvia Carbonell-Sala, Mark Diekhans, Jose Gonzalez Martinez, Toby Hunt, Irwin Jungreis, Jane E Loveland, Carme Arnan, If Barnes, Ruth Bennett, Andrew Berry, Alexandra Bignell, Daniel Cerdán-Vélez, Kelly Cochran, Lucas T Cortés, Claire Davidson, Sarah Donaldson, Cagatay Dursun, Reham Fatima, Matthew Hardy, Prajna Hebbar, Zoe Hollis, Benjamin T James, Yunzhe Jiang, Rory Johnson, Gazaldeep Kaur, Mike Kay, Riley J Mangan, Miguel Maquedano, Laura Martínez Gómez, Nourhen Mathlouthi, Ryan Merritt, Pengyu Ni, Emilio Palumbo, Tamara Perteghella, Fernando Pozo, Shriya Raj, Cristina Sisu, Emily Steed, Dulika Sumathipala, Marie-Marthe Suner, Barbara Uszczynska-Ratajczak, Elizabeth Wass, Yucheng T Yang, Dingyao Zhang, Robert D Finn, Mark Gerstein, Roderic Guigó, Tim J P Hubbard, Manolis Kellis, Anshul Kundaje, Benedict Paten, Michael L Tress, Ewan Birney, Fergal J Martin, and Adam Frankish. GENCODE 2025: Reference gene annotation for human and mouse. *Nucleic Acids Research*, 53 (D1):D966–D975, 2025. ISSN 1362-4962. doi: 10.1093/nar/gkae1078.

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Yarin Gal, and Debora Marks. ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design. *Advances in Neural Information Processing Systems*, 36:64331–64379, 2023. doi: 10.52202/075280-2810.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12 (85):2825–2830, 2011. ISSN 1533-7928.

Rafael Josip Penić, Tin Vlašić, Roland G. Huber, Yue Wan, and Mile Šikić. RiNALMo: General-purpose RNA language models can generalize well on structure prediction tasks. *Nature Communications*, 16(1):5671, 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-60872-5.

Lon Phan, Hua Zhang, Qiang Wang, Ricardo Villamarin, Tim Hefferon, Aravinthan Ramanathan, and Brandi Kattman. The evolution of dbSNP: 25 years of impact in genomic research. *Nucleic Acids Research*, 53(D1):D925–D931, 2025. ISSN 1362-4962. doi: 10.1093/nar/gkae977.

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. ISSN 1088-9051. doi: 10.1101/gr.097857.109.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S. Song. Evaluating Protein Transfer Learning with TAPE, 2019.

Jordan Rossen, Huwenbo Shi, Benjamin J. Strober, Martin Jinye Zhang, Masahiro Kanai, Zachary R. McCaw, Liming Liang, Omer Weissbrod, and Alkes L. Price. MultiSuSiE improves multi-ancestry fine-mapping in All of Us whole-genome sequencing data. *Nature Genetics*, 58(1): 67–76, 2026. ISSN 1546-1718. doi: 10.1038/s41588-025-02450-5.

Yair Schiff, Chia Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: Bi-Directional Equivariant Long-Range DNA Sequence Modeling. In *Forty-First International Conference on Machine Learning*, 2024.

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. ISSN 0305-1048. doi: 10.1093/nar/gkad989.

Tao Shen, Zhihang Hu, Siqi Sun, Di Liu, Felix Wong, Jiuming Wang, Jiayang Chen, Yixuan Wang, Liang Hong, Jin Xiao, Liangzhen Zheng, Tejas Krishnamoorthi, Irwin King, Sheng Wang, Peng Yin, James J. Collins, and Yu Li. Accurate RNA 3D structure prediction using a language model-based deep learning approach. *Nature Methods*, 21(12):2287–2298, 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02487-0.

Ruian Shi, Taykhoom Dalal, Philip Fradkin, Divya Koyyalagunta, Simran Chhabria, Andrew J. Jung, Cyrus L. Tam, Defne Ceyhan, Jessica Lin, Kaitlin U. Laverty, Ilyes Baali, Bo Wang, and Quaid Morris. mRNABench: A curated benchmark for mature mRNA property and function prediction. In *Learning Meaningful Representations of Life (LMRL) Workshop at ICLR 2026*, 2026.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. SaProt: Protein Language Modeling with Structure-aware Vocabulary. In *The Twelfth International Conference on Learning Representations*, 2024.

The All of Us Research Program Investigators. The "All of Us" research program. *New England Journal of Medicine*, 381(7):668–676, 2019. doi: 10.1056/NEJMsr1809937. URL `https://doi.org/10.1056/NEJMsr1809937`.

The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, 2020. doi: 10.1126/science.aaz1776.

The RNAcentral Consortium. RNAcentral in 2026: Genes and literature integration. *Nucleic Acids Research*, 54(D1):D303–D313, 2026. ISSN 1362-4962. doi: 10.1093/nar/gkaf1329.

The UniProt Consortium. UniProt: The Universal Protein Knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, 2025. ISSN 1362-4962. doi: 10.1093/nar/gkae1010.

Vincent Q. Tran, Matthew Nemeth, Liam J. Bartie, Sita S. Chandrasekaran, Alison Fanton, Hyungseok C. Moon, Brian L. Hie, Silvana Konermann, and Patrick D. Hsu. Rapid directed evolution guided by protein language models and epistatic interactions. *Science*, 392(6798): eaea1820, 2026. doi: 10.1126/science.aea1820.

Michel van Kempen, Stephanie S. Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L. M. Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*, 42(2):243–246, 2024. ISSN 1546-1696. doi: 10.1038/s41587-023-01773-0.

Qingbo S. Wang, David R. Kelley, Jacob Ulirsch, Masahiro Kanai, Shuvom Sadhuka, Ran Cui, Carlos Albors, Nathan Cheng, Yukinori Okada, Francois Aguet, Kristin G. Ardlie, Daniel G. MacArthur, and Hilary K. Finucane. 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. ISSN 2041-1723. doi: 10.1038/s41467-021-23134-8.

Talal Widatalla, Ashir A. Borah, Samuel H. King, Claudia L. Driscoll, Rafael Rafailov, and Brian L. Hie. Aligning protein-generative models to experimental fitness with ProteinDPO. *Nature Methods*, 23(9):1805–1813, 2026. ISSN 1548-7105. doi: 10.1038/s41592-026-03137-3.

Andrew D Yates, Olanrewaju Austine-Orimoloye, Andrey G Azov, Matthieu Barba, If Barnes, Vianey Paola Barrera-Enriquez, Arne Becker, Ruth Bennett, Andrew Berry, Jyothish Bhai, Simarpreet Kaur Bhurji, Paulo R Branco Lins, Lucy Brooks, Shashank Budhanuru Ramaraju, Lahcen I Campbell, Manuel Carbajo Martinez, Jack Carpenter, Mehrnaz Charkhchi, Lucas A Cortes, Claire Davidson, Suzanna Dickson, Kamalkumar Dodiya, Sarah Donaldson, Bilal El Houdaigui, Tamara El Naboulsi, Aine Fairbrother-Browne, Oluwadamilare Falola, Reham Fatima, Jose Gonzalez Martinez, Tatiana Gurbich, Holly Hall, Matthew Hardy, Zoe Hollis, Toby Hunt, Mike Kay, Vinay Kaikala, Anna Lazar, Diana Lemos, Disha Lodha, Nourhen Mathlouthi, Gabriela A Merino, Ryan Merritt, Louisse Paola Mirabueno, Aleena Mushtaq, Syed Nakib Hossain, José G Pérez-Silva, Ivana Piližota, Daniel Poppleton, Irina Prosovetskaia, Shriya Raj, Ahamed Imran Abdul Salam, Shradha Saraf, Swati Sinha, Botond Sipos, Vasily Sitnik, Marie-Marthe Suner, Likhitha Surapaneni, Jack A S Tierney, David Urbina-Gómez, Andres Veidenberg, Thomas A Walsh, Jamie M Allen, Jorge Alvarez-Jarreta, Jitender Cheema, Jorge Batista da Rocha, Nishadi H De Silva, Francesca Floriana Tricomi, Stefano Giorgetti, Garth R

Ilsley, Jon Keatley, Jane E Loveland, Jonathan M Mudge, Guy Naamati, John Tate, Natalie L Willhoft, Andrea Winterbottom, Bethany R Flint, Adam Frankish, Leanne Haggerty, Sarah E Hunt, Emily L Clark, Sarah C Dyer, Mallory A Freeberg, Fergal J Martin, and Robert D Finn. Ensembl 2026. *Nucleic Acids Research*, 54(D1):D1053–D1060, 2026. ISSN 1362-4962. doi: 10.1093/nar/gkaf1239.

Chengzhong Ye, Gonzalo Benegas, Carlos Albors, Jianan Canal Li, Sebastian Prillo, Peter D. Fields, Brian Clarke, and Yun S. Song. Predicting genome-wide functional constraints with GPN-Star. *Nature*, pp. 1–10, 2026. ISSN 1476-4687. doi: 10.1038/s41586-026-11005-5.

Weijie Yin, Zhaoyu Zhang, Shuo Zhang, Liang He, Ruiyang Zhang, Rui Jiang, Gan Liu, Jingyi Wang, Xuegong Zhang, Tao Qin, and Zhen Xie. ERNIE-RNA: An RNA language model with structure-enhanced representations. *Nature Communications*, 16(1):10076, 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-64972-0.

A Dataset curation

A.1 Variant datasets

ClinVar variants. The VCF of all ClinVar variants on GRCh38 was downloaded directly from ClinVar's FTP server. The `20260208` release was used, which is now available at `https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/archive_2.0/2026/clinvar_20260208.vcf.gz`

GWAS disease-associated variants. Fine-mapped UKBB variants provided by the original authors in hg19 coordinates were downloaded from `https://www.finucanelab.org/data`. Lift-over to GRCh38 was accomplished via direct mapping of dbSNP rsID (Phan et al., 2025) with confirmation that the reference allele in GRCh38 matched expectation. This process retained 1,679,687 / 1,688,534 variants. We additionally remove variants associated with the following 9 traits: `[``Age_at_Menarche'', ``Age_at_Menopause'', ``Balding_Type4'', ``Guilty_Feelings'', ``Morning_Person'', ``Risk_Taking'', ``Sensitivity'', ``Tense'', ``Worry_Too_Long'']`.

Fine-mapped trait-associated variants from All of Us were downloaded from the authors' Zenodo repository (`https://zenodo.org/records/17370173`), from which the `pips.tsv` file was used.

Gene-expression altering variants. Fine-mapped cis-eQTLs were sourced from the v11 release of the Genotype Tissue Expression (GTEx) project. We used the `GTEx_Analysis_v11_eQTL_SuSiE.tar` file directly from GTEx (`https://gtexportal.org/home/downloads/adult-gtex/qtl`).

Genome sequence and annotations. We used GENCODE Release 50 (GRCh38.p14) for all gene and transcript annotations. The following files were downloaded from the official GENCODE website at `https://www.gencodegenes.org/human/`:

- Comprehensive gene annotation (reference chromosomes only): `gencode.v50.annotation.gtf.gz`
- Transcript sequences: `gencode.v50.transcripts.fa.gz`
- Protein-coding transcript translation sequences: `gencode.v50.pc_transcripts.fa.gz`
- Genome sequence (GRCh38.p14): `GRCh38.p14.genome.fa.gz`

Any of our analyses utilizing gene annotations or sequences of transcripts, proteins, or the human genome used the above files.

Variant filtering and transcript assignment. We use the Ensembl Variant Effect Predictor (VEP) (McLaren et al., 2016) to obtain affected transcripts and predicted consequences for all variants in our dataset, with GENCODE Release 50 comprehensive gene annotations (`gencode.v50.annotation.gtf.gz`) (Mudge et al., 2025). To obtain the final set of variant-transcript pairs, we perform the following filters:

1. Remove variant-transcript pairs where the variant falls outside of an exon: remove any rows where the predicted consequence is one of [``downstream_gene_variant'', ``upstream_gene_variant'', ``intron_variant'', ``intergenic_variant'']
2. Retain protein-coding or long non-coding RNAs (lncRNA) only: remove any rows where the transcript biotype is not in [``lncRNA'', ``protein_coding'']
3. Remove variants that overlap both a protein-coding gene and a lncRNA
4. For each variant-gene pair, we select a single representative transcript with the following prioritization: MANE Select, Ensembl Canonical, transcript with highest transcript support level (TSL). Ties are broken deterministically

A.2 Evolutionary annotations

phyloP scores. phyloP scores were obtained via the bigWig file for the `cons447way` track for hg38 from the UCSC Genome Browser (`https://hgdownload.soe.ucsc.edu/goldenPath/hg38/phyloP447way/hg38.phyloP447way.bw`).

Gnocchi z-scores. Gnocchi scores were obtained from the gnomAD portal under the "Genomic Constraint" section of "v3 Downloads" (`https://gnomad.broadinstitute.org/data#v3-genomic-constraint`). We used the "QCed genomic constraint by 1kb regions" scores (`constraint_z_genome_1kb.qc.download.txt.gz`).

B Variant scoring and complementarity assessments

B.1 Variant scoring with biological language models

Here we provide additional details for each language model utilized in our analyses.

Evo2. Variant scoring with Evo2 (Brixi et al., 2026) was performed by extracting an 8,192 bp window centered on each variant and calculating the ratio of the likelihoods of the full sequence with either the reference or alternate allele. In particular, for a variant at position p, we extract an 8,192 bp window of GRCh38 centered on p and compute the difference in mean per-token autoregressive log-likelihood between the reference and alternate windows,

$$\Delta(v) \;=\; \frac{1}{L}\sum_{i=1}^{L}\log P_\theta\big(x_i^{\text{ref}} \mid x_{<i}^{\text{ref}}\big) \;-\; \frac{1}{L}\sum_{i=1}^{L}\log P_\theta\big(x_i^{\text{alt}} \mid x_{<i}^{\text{alt}}\big)\,, \tag{1}$$

with $L = 8{,}192$. Because Evo2 is a causal model, a variant influences only tokens at or after its own position; we therefore score each window on both strands and average the forward and reverse-complement values, so that sequence context on both sides of the variant contributes. We restrict to single-nucleotide variants.

Nucleotide Transformer v3. Variant scoring with Nucleotide Transformer v3 was performed by extracting an 8,192 bp window centered on the variant, replacing the variant position with a `[MASK]` token, and computing the log-likelihood ratio of the reference and alternate allele.

The authors of Nucleotide Transformer v3 make several checkpoints, including "post-trained" models that have undergone additional training to perform sequence-to-function tasks. Since our focus is on the language model component of each model, we use the authors' pre-trained checkpoint which has only been trained with the masked language modeling task. In particular, we utilize the `NTv3_650M_pre` checkpoint (`https://huggingface.co/InstaDeepAI/NTv3_650M_pre`).

RiNALMo. As RiNALMo (Penić et al., 2025) has 1,024-nucleotide context window, transcripts longer than the context are subset to a 1,024 nt window centered on the variant. No strand averaging is applied, as the transcript is already strand-correct. To produce scores, we replace the variant position with the `[MASK]` token and compute the log-likelihood ratio of the reference and alternate allele. We download RiNALMo weights from the authors' original Zenodo repository (`https://zenodo.org/records/15043668`).

Orthrus. Orthrus scoring is performed similarly to RiNALMo in that we use the selected transcript sequence from GENCODE v50 and replace the variant position with the `[MASK]` token to compute

the log-likelihood ratio of the reference and alternate allele. For Orthrus, we use the full transcript sequence due to the longer context window of the model.

Orthrus is originally trained using a contrastive objective in which sequences corresponding to alternative splicing isoforms from the same gene or corresponding to orthologs across species form positive pairs. In follow-up work, the authors additionally trained a version of their model with an additional masked language modeling (MLM) objective, which was shown to have improved performance for variant effect prediction (Shi et al., 2026). We utilize the 6-track MLM version of their model for our analyses (`https://huggingface.co/antichronology/orthrus-mlm-6-track`).

ESM-C. ESM-C weights were sourced directly from the author's HuggingFace space (`https://huggingface.co/biohub/ESMC-600M`).

SaProt. SaProt is a protein language model that utilizes both amino acid sequences as well as protein structure as represented by FoldSeek (van Kempen et al., 2024) tokens. To obtain complete structures for proteins covered by our variant datasets, we utilized structures from AlphaFold DB(Bertoni et al., 2026). For each transcript covered by a variant in our dataset, we performed the following steps to obtain a corresponding structure:

1. Map the corresponding transcript's Ensembl transcript ID (`ENST...`) to an Ensembl protein ID (`ENSP...`)
2. Map the Ensembl protein ID to a UniProt ID via Ensembl BioMart (Yates et al., 2026)
3. Query the UniProt API to identify protein isoforms (`https://rest.uniprot.org/uniprotkb/{accession}.json`)
4. For each isoform returned, query the AlphaFold DB API to obtain the isoform's structure and sequence (`https://alphafold.ebi.ac.uk/api/prediction/{accession}`)
5. If AlphaFold DB returns an isoform where the sequence is an exact match to the original Ensembl translated protein-coding sequence, download and use the `.pdb` file for that structure.

If a transcript fails at any of these steps (e.g. no mapping to UniProt ID, no AlphaFold DB isoform with exact sequence match), we remove that transcript from consideration for both ESM-C and SaProt.

Given the `.pdb` files for structures from above, FoldSeek tokens were generated with the command `foldseek structureto3didescriptor --chain-name-mode 1`. As recommended by SaProt authors, FoldSeek tokens were masked for positions with AlphaFold pLDDT less than 70.

SaProt weights were sourced directly from the author's HuggingFace space (`https://huggingface.co/westlake-repl/SaProt_650M_AF2`).

For both SaProt and ESM-C, we use the GENCODE protein-coding transcript translation sequences for each transcript as the input amino acid sequence. Variants are scored by replacing the mutated amino acid with the `[MASK]` token and computing the log-likelihood ratio of the reference and alternate amino acid, where SaProt amino acid likelihoods are computed by summing over all possible structure tokens.

For proteins longer than 1,024 amino acids, we use the scoring methodology from Brandes et al. (2023). Briefly, long proteins are tiled with 1,024 amino acid windows with approximately 50% overlap. If a variant appears in multiple overlapping windows, it is scored within each context and per-amino-acid scores are then averaged using a sigmoid weighting scheme that downweights positions near artificial boundaries created by window edges.

GPN-Star. Precomputed scores for the vertebrate, mammal, and primate versions of GPN-Star were downloaded from HuggingFace (`https://huggingface.co/datasets/songlab/gpn-star-scores`) and `llr_calibrated` scores were used directly.

AlphaGenome Variant Impact. Precomputed AVI scores were downloaded directly from the webpage provided by the authors (`https://deepmind.google.com/science/alphagenome/downloads`, "AVI SNV scores") and used without modification.

B.2 FEATURE COMPLEMENTARITY ASSESSMENTS

Logistic regressions for feature complementarity assessments were fit in `scikit-learn` with the following parameters:

```
from sklearn.linear_model import  LogisticRegressionCV
from sklearn.pipeline import make_pipeline
from sklearn.preprocessing import PolynomialFeatures, StandardScaler

clf = make_pipeline(
    PolynomialFeatures(degree=2, include_bias=False),
    StandardScaler(),
    LogisticRegressionCV(
        l1_ratios=(0.0,),
        scoring="neg_log_loss",
        use_legacy_attributes=False,
        n_jobs=8,
    )
)
```

to obtain scores for all samples, an outer stratified 5-fold cross-validation was performed with the following parameters:

```
groups = variant_data["variant_id"].to_numpy()

cv = StratifiedGroupKFold(
    n_splits=5,
    shuffle=True,
    random_state=42,
)
splits = list(cv.split(
  variant_data,
  variant_data["label"],
  groups,
))
```

with the same splits used across all assessed feature combinations. For creating splits, grouping is performed using variant identifiers, such that if a variant affects multiple phenotypes, all rows corresponding to that variant are assigned to the same split.

Percentile intervals for AUROC improvements relative to the base model were calculated using a paired bootstrap as follows. Let $\{(y_i, \hat{y}_i^{\text{base}}, \hat{y}_i^{\text{oth}})\}_{i=1}^{N}$ denote the true labels and predictions from the base and comparison models. For each bootstrap replicate $b = 1, \ldots, B$, we sample N indices with replacement,

$$I_b = \{i_1, \ldots, i_N\}, \qquad i_j \sim \text{Uniform}\{1, \ldots, N\}.$$

Using the same bootstrap sample for both models, we compute

$$A_{\text{base}}^{(b)} = \text{AUROC}\left(\mathbf{y}_{I_b}, \hat{\mathbf{y}}_{I_b}^{\text{base}}\right), \qquad A_{\text{oth}}^{(b)} = \text{AUROC}\left(\mathbf{y}_{I_b}, \hat{\mathbf{y}}_{I_b}^{\text{oth}}\right).$$

The bootstrap AUROC difference is then

$$\Delta A^{(b)} = A_{\text{oth}}^{(b)} - A_{\text{base}}^{(b)},$$

and the corresponding percent change relative to the reference model is

$$\Delta A_{\%}^{(b)} = 100 \frac{A_{\text{oth}}^{(b)} - A_{\text{base}}^{(b)}}{A_{\text{base}}^{(b)}}.$$

The resulting bootstrap distributions $\{\Delta A^{(b)}\}_{b=1}^{B}$ are used to quantify uncertainty in the difference in model performance. Similar to grouping for cross-validation, bootstrapping samples over variants rather than individual rows of variant-phenotype associations.

To ensure fair comparison with the non-linear feature combination models, the base model scores are similarly calculated using a logistic regression with polynomial features rather than solely using the raw score itself.

C AlphaGenome assessment of eQTL dataset

Our eQTL dataset differs from standard eQTL datasets in that we only select eQTLs contained within exons of transcribed genes, and further require that the eQTL modulates the expression of the gene in which it's contained. Due to this difference, we wanted to ensure that models' poor performance on our dataset was driven by model failures rather than noise or extreme difficulty of the dataset. AlphaGenome has been shown to be a strong predictor of eQTLs due to its direct modeling of RNA-seq data (Avsec et al., 2026), so we used AlphaGenome to generate predictions for our eQTL dataset as a positive control comparison.

Input construction. AlphaGenome was run at a fixed input length of 524,288 bp. For each variant we defined the smallest interval containing both the variant and the entire target gene body, and placed a fixed-length window centered on that interval, shifted where necessary to remain within chromosome boundaries.

Every window therefore spans exactly 524,288 bp and contains both the variant and the complete gene. Associations whose required interval exceeded the window were excluded (1,028 of 78,420, 1.31%; 80 high-PIP, or 0.72% of high-PIP associations); excluded cases are almost entirely very long genes (median gene length 722 kb versus 38 kb for retained cases). Interval length was the only exclusion criterion applied.

Due to the lower throughput of AlphaGenome predictions with large sequences and the large number of negatives (low-PIP candidate eQTLs) in our dataset, negatives were randomly downsampled to a 1:1 ratio, giving 22,056 scored associations (11,028 high-PIP, 11,028 low-PIP) over 21,988 distinct variants, 12,368 genes and 50 tissues. Reference and alternate sequences for each association differ at exactly one of 524,288 positions and were evaluated in the same batch.

Scoring. We used the AlphaGenome `all_folds` checkpoint, requesting the RNA-seq output head at 1 bp resolution. Each association was read out from the single unstranded GTEx RNA-seq track matching its tissue, identically for reference and alternate sequences.

Predicted signal was averaged over the target gene's exons, with the final variant effect score taken as $\log_2(\text{alt}) - \log_2(\text{ref})$. Using this score resulted in an AUROC of 0.726.

D phyloP and Gnocchi correlation analyses

Rationale. Two possible confounders exist for our results comparing model scores to annotations (Figure 2) and showing lack of correlation between phyloP and Gnocchi within our dataset (Appendix Figure E.3):

1. phyloP and Gnocchi are calculated at two different resolutions: phyloP is a basepair-level metric while Gnocchi is calculated in 1kb windows
2. Our analyses use the variant loci we've selected for downstream variant effect prediction analyses. Other genomic windows may show different results

To ensure that neither of the above confounders affect our results, we undertook the following analysis.

Methods. As a representative model score, we used GPN-Star scores due to their availability as precomputed scores for nearly all possible single nucleotide changes in the human genome, for all three GPN-Star models tested. We use the same 447-way mammalian phyloP scores and Gnocchi z-scores as used in the main text.

To examine aggregation differences, we compared means over the observed variant loci in our dataset with means over all available scores in Gnocchi's associated 1kb window containing that variant. Full-window means included all scored bases for phyloP and all scored alternate substitutions for GPN-Star, typically 3,000 SNVs per window. We additionally evaluated full-window means in a sample of 20,000 1kb windows drawn uniformly from the set of 1kb windows made available by Gnocchi. We computed Spearman correlations with Gnocchi, excluding missing values for each comparison.

Results. We show the results of these analyses in Table D.1. We find that computing correlations at our selected loci produces results largely similar to using uniformly drawn genomic windows (Table D.1). We find that averaging model scores or phyloP values within a 1kb window before comparing to Gnocchi values increases correlations, but not to a degree that substantially changes our results. Importantly, analyses showing improvement to model variant effect predictions upon incorporation of Gnocchi scores remain unaffected, as these analyses do not rely on the spatial resolution of Gnocchi scores.

Table D.1: **Gnocchi correlations under different aggregation schemes.** Values shown are Spearman ρ of Gnocchi values and the score in each column, aggregated as specified in the "Setting".

Setting	Unit	phyloP	GPN-Star (V)	GPN-Star (M)	GPN-Star (P)
Observed variant scores	variant	0.029	0.152	0.168	0.172
Full-window means at represented windows	window	0.172	0.246	0.218	0.246
Full-window means at uniformly sampled genomic windows	window	0.060	0.170	0.109	0.192

E Additional figures and tables

E.1 MAF and phyloP values by dataset

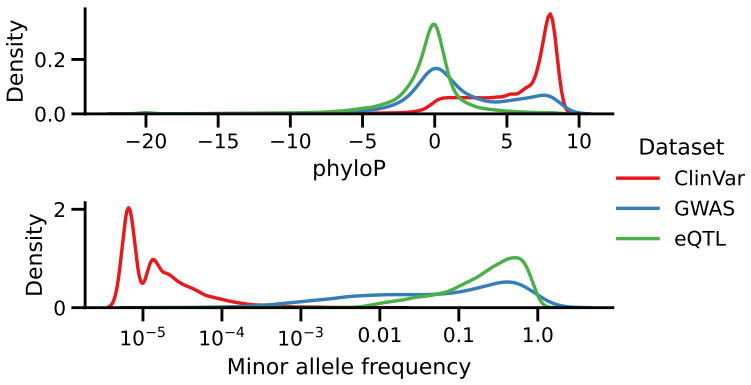

Figure E.1: **Minor allele frequencies and phyloP values of functional variants within each dataset.**

E.2 Model correlation with evolutionary annotations

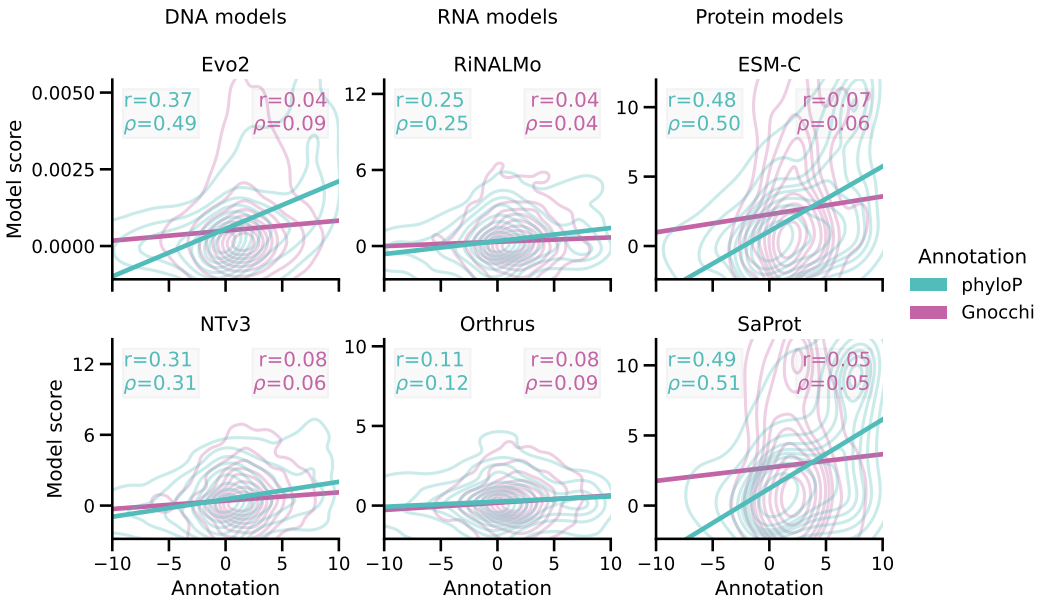

Figure E.2: **Model scores correlated with evolutionary annotations.** Data is exactly as in Figure 2, except here shown in the original data space rather than as z-scores.

E.3 PHYLOP CORRELATION WITH GNOCCHI

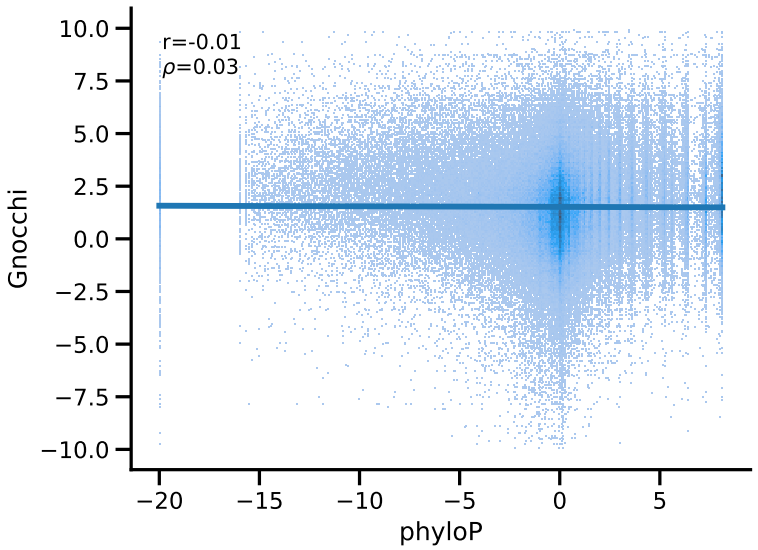

Figure E.3: **Correlation of phyloP and Gnocchi z-scores.** Across all loci in our dataset, phyloP and Gnocchi show no correlation.

E.4 VARIANT EFFECT PREDICTION PERFORMANCE ON MISSENSE VARIANTS ONLY

Table E.1: AUROC for distinguishing functional and non-functional variants across datasets, restricted to the same set of protein-coding missense variants for all models.

Dataset	Conservation	Human constraint	DNA		RNA		Protein	
			Evo2	NTv3	RiNALMo	Orthrus	ESM-C	SaProt
ClinVar	0.842	0.535	0.852	0.785	0.581	0.567	0.929	0.936
GWAS	0.659	0.550	0.639	0.544	0.501	0.498	0.619	0.632
eQTL	0.521	0.590	0.533	0.485	0.489	0.488	0.500	0.504

E.5 VARIANT EFFECT PREDICTION WITH RAW AND ABSOLUTE ZERO-SHOT VARIANT EFFECT SCORES

Table E.2: Variant effect prediction performance per model and dataset for both raw zero-shot variant effect scores ("Signed") and absolute values of zero-shot variant effect scores ("Unsigned"). Bold indicates the better signed or unsigned value for each metric within each row.

		Signed		Unsigned	
Dataset	Score	AUROC	AUPRC	AUROC	AUPRC
ClinVar	Conservation	**0.886**	**0.625**	0.760	0.305
	Human constraint	**0.479**	**0.149**	0.472	0.147
	Evo2	**0.978**	**0.893**	0.977	0.888
	NTv3	**0.855**	**0.702**	0.832	0.679
	RiNALMo	**0.781**	**0.628**	0.756	0.608
	Orthrus	**0.590**	**0.298**	0.503	0.273
	ESM-C	**0.929**	**0.894**	0.926	0.891
	SaProt	**0.936**	**0.897**	0.934	0.895
	GPN-Star (V)	**0.961**	**0.903**	0.954	0.894
	GPN-Star (M)	**0.961**	**0.888**	0.951	0.844
	GPN-Star (P)	**0.928**	**0.827**	0.912	0.767
	AlphaGenome VI	**0.992**	**0.970**	0.990	0.966
GWAS	Conservation	**0.695**	**0.056**	0.657	0.029
	Human constraint	**0.593**	**0.021**	0.576	0.019
	Evo2	0.657	0.049	**0.671**	**0.050**
	NTv3	0.522	**0.018**	**0.538**	0.017
	RiNALMo	0.500	**0.014**	**0.508**	0.014
	Orthrus	0.493	0.014	**0.503**	**0.014**
	ESM-C	**0.619**	**0.113**	0.613	0.112
	SaProt	**0.631**	**0.112**	0.629	0.110
	GPN-Star (V)	0.688	**0.064**	**0.716**	0.061
	GPN-Star (M)	0.690	**0.068**	**0.696**	0.043
	GPN-Star (P)	0.682	**0.062**	**0.685**	0.037
	AlphaGenome VI	**0.743**	**0.072**	0.733	0.069
eQTL	Conservation	0.508	**0.140**	**0.525**	0.136
	Human constraint	**0.573**	**0.168**	0.561	0.154
	Evo2	**0.513**	**0.135**	0.494	0.130
	NTv3	**0.506**	**0.125**	0.492	0.120
	RiNALMo	**0.506**	**0.125**	0.494	0.120
	Orthrus	**0.512**	**0.127**	0.490	0.119
	ESM-C	0.500	0.145	**0.532**	**0.158**
	SaProt	0.504	0.147	**0.535**	**0.157**
	GPN-Star (V)	**0.535**	**0.155**	0.522	0.145
	GPN-Star (M)	**0.536**	**0.157**	0.511	0.134
	GPN-Star (P)	**0.532**	**0.151**	0.508	0.132
	AlphaGenome VI	**0.597**	**0.196**	0.586	0.186

E.6 Feature complementarity assessment by dataset

Here we show results analogous to those shown in Figure 3 but broken down by dataset rather than considering both common variant datasets together (GWAS disease-associated variants and eQTLs).

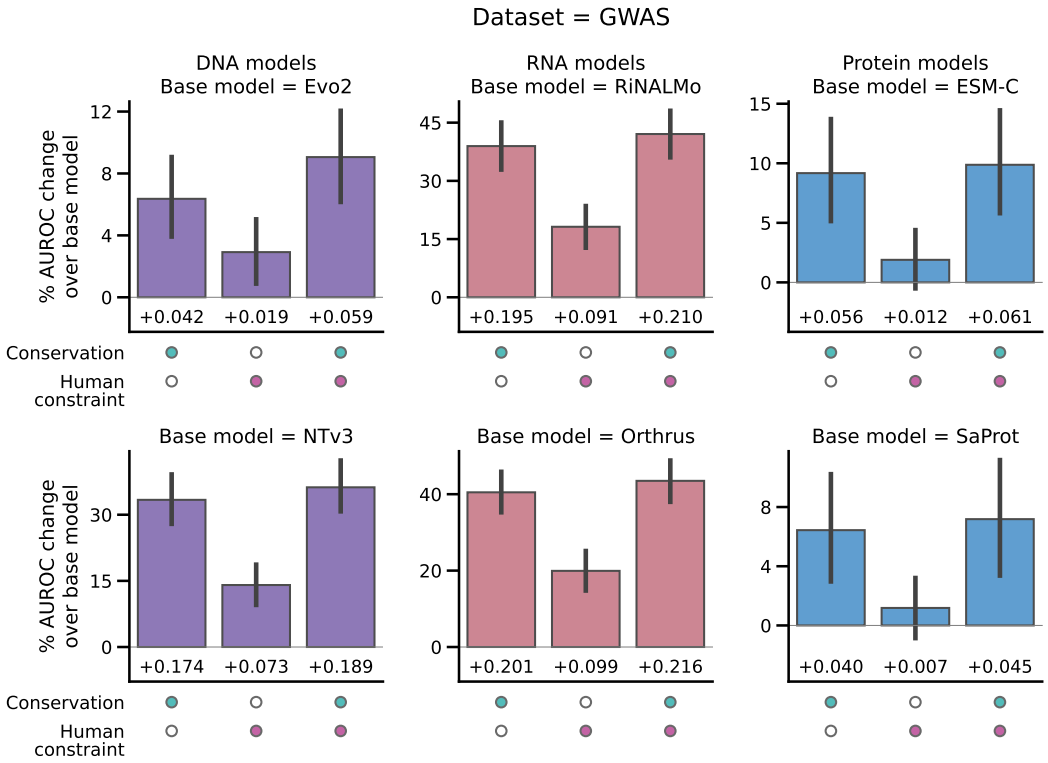

Figure E.4: **Effect of feature combinations on GWAS variant effect predictions.**

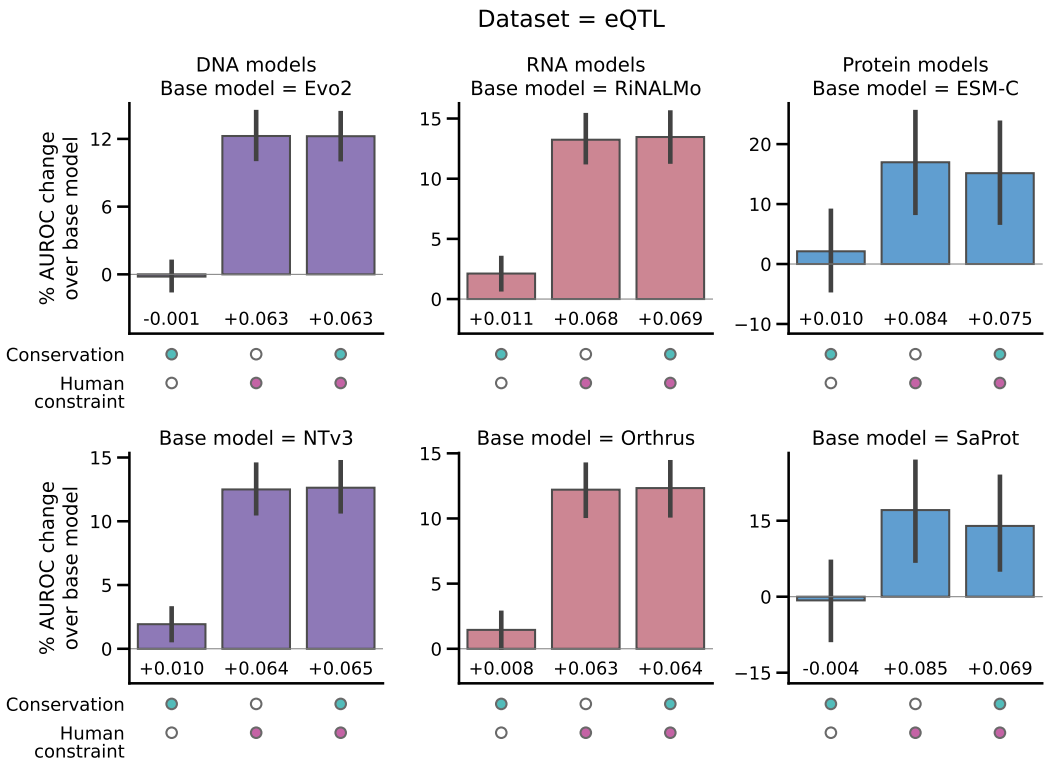

Figure E.5: **Effect of feature combinations on eQTL variant effect predictions.**

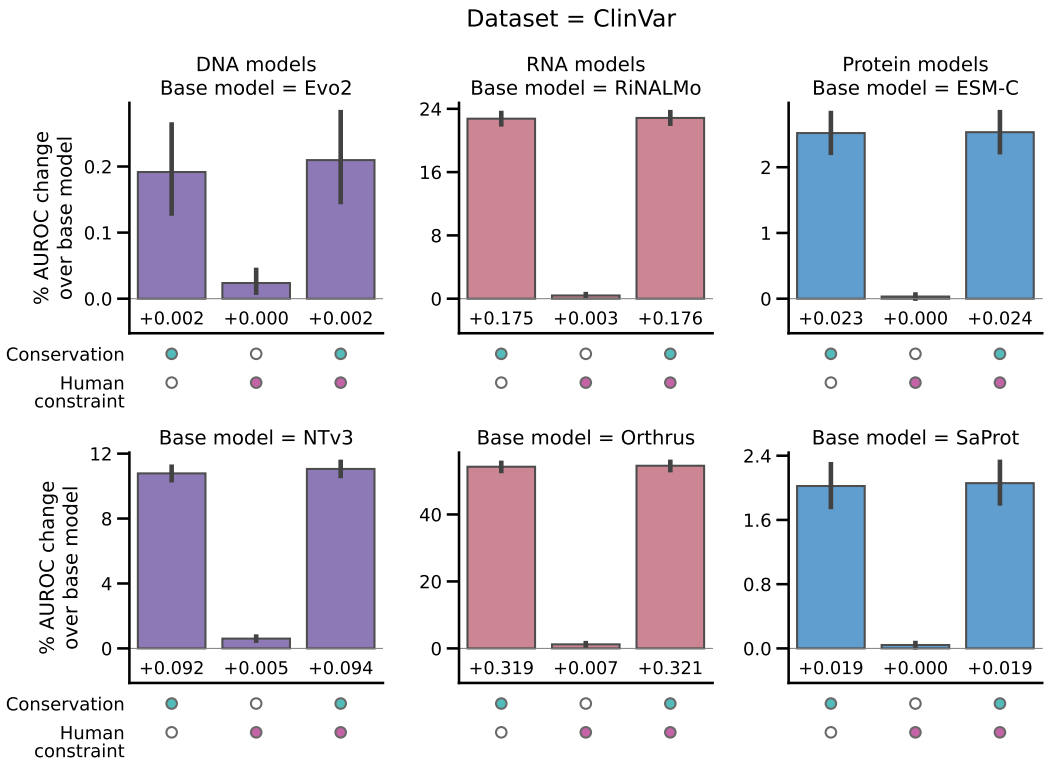

Figure E.6: **Effect of feature combinations on ClinVar variant effect predictions.**

E.7 GPN-STAR AND ALPHAGENOME VI EXTREME-VALUED LOCI CLASSIFICATION

Table E.3: **GPN-Star and AlphaGenome VI ability to separate highly conserved or human constrained loci.** Values shownare AUROCs for separating loci with either phyloP or Gnocchi scores in the top or bottom quartileof all loci, respectively.

	GPN-Star			AlphaGenome VI
Annotation	V	M	P	
Conservation	0.879	0.882	0.858	0.988
Human constrained	0.617	0.630	0.632	0.546

E.8 FULL VARIANT EFFECT PERFORMANCE IMPROVEMENTS

Table E.4: **Variant effect prediction improvements relative to each base model alone.** Values in brackets give the 95% percentile interval estimated from paired bootstrapping. Bold indicates a positive range that does not overlap zero.

			AUROC improvement (%)		AUPRC improvement (%)	
Dataset	Modality	Model	+ Gnocchi	+ Gnocchi + phyloP	+ Gnocchi	+ Gnocchi + phyloP
GWAS+eQTL	DNA	Evo2	**+9.83 [8.20, 11.61]**	**+10.50 [8.86, 12.29]**	**+13.82 [10.51, 17.48]**	**+17.29 [14.02, 21.13]**
	DNA	NTv3	**+10.83 [8.93, 12.65]**	**+12.34 [10.43, 14.17]**	**+24.23 [20.03, 28.17]**	**+30.59 [25.87, 35.40]**
	RNA	RiNALMo	**+11.56 [9.92, 13.35]**	**+12.99 [11.22, 14.86]**	**+25.56 [21.59, 29.64]**	**+31.87 [27.54, 36.51]**
	RNA	Orthrus	**+10.95 [9.20, 12.85]**	**+12.28 [10.46, 14.26]**	**+22.34 [18.08, 26.69]**	**+28.57 [23.85, 33.77]**
	Protein	ESM-C	**+5.42 [1.65, 9.46]**	**+7.22 [3.15, 11.47]**	+4.25 [-0.73, 9.70]	**+7.84 [1.80, 14.35]**
	Protein	SaProt	**+3.56 [0.08, 6.97]**	**+4.94 [0.79, 8.68]**	+1.20 [-3.64, 6.52]	+5.02 [-1.88, 11.82]
	Other	GPN-Star (V)	**+4.64 [3.35, 5.98]**	**+4.55 [3.27, 5.95]**	**+6.86 [4.75, 9.15]**	**+6.81 [4.59, 9.12]**
	Other	GPN-Star (M)	**+5.10 [3.92, 6.29]**	**+5.09 [3.87, 6.32]**	**+6.44 [4.36, 8.68]**	**+6.29 [4.07, 8.50]**
	Other	GPN-Star (P)	**+6.03 [4.79, 7.29]**	**+6.11 [4.81, 7.41]**	**+7.33 [5.08, 9.64]**	**+7.75 [5.52, 10.13]**
	Other	AlphaGenome VI	**+2.42 [1.71, 3.09]**	**+5.83 [4.89, 6.71]**	**+2.38 [1.02, 3.79]**	**+12.37 [9.64, 15.01]**
GWAS	DNA	Evo2	**+2.92 [0.88, 5.04]**	**+9.06 [6.16, 12.05]**	+0.30 [-3.81, 4.51]	**+22.19 [11.09, 33.71]**
	DNA	NTv3	**+14.06 [9.55, 18.67]**	**+36.22 [30.77, 42.29]**	**+30.82 [17.33, 44.76]**	**+235.78 [193.31, 284.93]**
	RNA	RiNALMo	**+18.17 [12.78, 23.51]**	**+42.07 [36.06, 48.05]**	**+47.64 [32.07, 64.16]**	**+292.93 [235.50, 350.73]**
	RNA	Orthrus	**+19.93 [14.78, 25.13]**	**+43.49 [38.01, 48.81]**	**+55.26 [39.53, 73.08]**	**+307.98 [256.77, 365.67]**
	Protein	ESM-C	+1.89 [-0.50, 4.39]	**+9.87 [5.81, 14.44]**	+0.19 [-3.70, 4.39]	**+9.43 [0.77, 19.00]**
	Protein	SaProt	+1.18 [-0.86, 3.21]	**+7.18 [3.37, 11.17]**	+0.29 [-4.18, 4.81]	+9.08 [-0.11, 18.74]
	Other	GPN-Star (V)	**+2.08 [0.49, 3.87]**	**+2.24 [0.61, 3.98]**	+0.82 [-1.09, 2.96]	+1.29 [-2.58, 6.39]
	Other	GPN-Star (M)	**+1.46 [0.32, 2.64]**	**+1.45 [0.14, 2.68]**	+0.50 [-1.12, 2.21]	-0.75 [-3.12, 1.44]
	Other	GPN-Star (P)	**+1.52 [0.37, 2.69]**	**+2.29 [0.84, 3.79]**	+0.93 [-0.81, 2.82]	+3.50 [-0.19, 7.31]
	Other	AlphaGenome VI	**+1.40 [0.64, 2.18]**	**+1.56 [0.67, 2.47]**	+1.16 [-1.26, 3.74]	+1.32 [-3.30, 5.93]
eQTL	DNA	Evo2	**+12.26 [10.23, 14.37]**	**+12.24 [10.21, 14.28]**	**+25.69 [21.31, 30.38]**	**+26.90 [22.77, 31.38]**
	DNA	NTv3	**+12.50 [10.63, 14.44]**	**+12.63 [10.78, 14.63]**	**+30.82 [26.12, 35.56]**	**+32.49 [27.59, 37.28]**
	RNA	RiNALMo	**+13.24 [11.38, 15.28]**	**+13.46 [11.43, 15.50]**	**+31.91 [27.08, 36.40]**	**+33.72 [28.69, 38.78]**
	RNA	Orthrus	**+12.21 [10.21, 14.13]**	**+12.33 [10.24, 14.31]**	**+29.95 [25.44, 34.99]**	**+31.59 [26.88, 36.73]**
	Protein	ESM-C	**+16.97 [8.55, 25.33]**	**+15.15 [6.91, 23.56]**	**+25.55 [8.22, 44.12]**	**+20.52 [4.52, 37.80]**
	Protein	SaProt	**+17.10 [7.14, 26.62]**	**+13.97 [5.38, 23.66]**	**+29.41 [11.73, 48.35]**	**+22.25 [5.52, 39.16]**
	Other	GPN-Star (V)	**+7.62 [6.12, 9.15]**	**+7.69 [6.17, 9.30]**	**+13.95 [10.94, 17.17]**	**+13.91 [10.83, 17.13]**
	Other	GPN-Star (M)	**+8.16 [6.74, 9.56]**	**+8.28 [6.99, 9.68]**	**+13.81 [10.70, 17.14]**	**+14.04 [10.79, 17.36]**
	Other	GPN-Star (P)	**+8.81 [7.35, 10.41]**	**+8.76 [7.28, 10.38]**	**+15.35 [12.13, 18.98]**	**+15.48 [12.30, 19.07]**
	Other	AlphaGenome VI	**+3.88 [2.92, 4.93]**	**+7.41 [6.18, 8.62]**	**+5.97 [4.09, 8.14]**	**+14.58 [11.89, 17.39]**
ClinVar	DNA	Evo2	**+0.02 [0.01, 0.04]**	**+0.21 [0.15, 0.28]**	+0.02 [-0.06, 0.10]	**+0.27 [0.10, 0.46]**
	DNA	NTv3	**+0.60 [0.48, 0.72]**	**+11.06 [10.63, 11.49]**	**+1.10 [0.95, 1.26]**	**+18.42 [17.68, 19.18]**
	RNA	RiNALMo	**+0.41 [0.24, 0.57]**	**+22.85 [22.13, 23.57]**	**+0.51 [0.41, 0.63]**	**+32.40 [31.28, 33.56]**
	RNA	Orthrus	**+1.21 [0.85, 1.56]**	**+54.55 [53.26, 55.69]**	**+2.11 [1.69, 2.49]**	**+131.18 [126.50, 135.54]**
	Protein	ESM-C	**+0.03 [0.00, 0.06]**	**+2.53 [2.23, 2.84]**	+0.04 [-0.01, 0.10]	**+2.15 [1.88, 2.41]**
	Protein	SaProt	**+0.04 [0.02, 0.07]**	**+2.06 [1.81, 2.32]**	**+0.22 [0.11, 0.33]**	**+2.05 [1.79, 2.30]**
	Other	GPN-Star (V)	**+0.31 [0.27, 0.35]**	**+0.57 [0.52, 0.63]**	**+0.19 [0.15, 0.24]**	**+0.61 [0.53, 0.69]**
	Other	GPN-Star (M)	**+0.30 [0.26, 0.35]**	**+0.48 [0.43, 0.52]**	**+0.33 [0.27, 0.39]**	**+0.71 [0.61, 0.80]**
	Other	GPN-Star (P)	**+0.24 [0.19, 0.28]**	**+1.93 [1.79, 2.07]**	**+0.27 [0.20, 0.33]**	**+1.41 [1.25, 1.57]**
	Other	AlphaGenome VI	+0.00 [-0.00, 0.01]	**+0.04 [0.03, 0.05]**	+0.04 [-0.00, 0.09]	**+0.16 [0.08, 0.24]**

E.9 COMPLEMENTARITY ASSESSMENT RELATIVE TO HUMAN CONSTRAINT ALONE

In Figure 3 and 4, we show improvements over biological language models and other state-of-the-art VEP methods (GPN-Star, AlphaGenome VI) when combined with human constraint metrics. In Table E.5, we show improvements relative to human constraint (Gnocchi) alone. We see that model scores and human constraint metrics are complementary when the associated model alone was able to perform well, *e.g.* GPN-Star and AlphaGenome VI for GWAS and eQTL. Put another way, complementarity is evident even among the highest-performing models, precisely where it may offer the greatest potential for further state-of-the-art improvements.

Table E.5: **Variant effect prediction improvements relative to human constraint alone.** Values shown are percent change relative to human constraint (Gnocchi) performance on the same dataset. Values in brackets give the 95% percentile interval estimated from paired bootstrapping. Bold indicates a positive range that does not overlap zero, *i.e.* a bold value in the "Base model + Gnocchi" column indicates a dataset in which the associated model provides complementary information over Gnocchi for variant effect prediction.

			AUROC improvement (%)		AUPRC improvement (%)	
Dataset	Modality	Model	Base model	Base model + Gnocchi	Base model	Base model + Gnocchi
GWAS+eQTL	DNA	Evo2	-8.07 [-9.72, -6.48]	**+0.96 [0.56, 1.38]**	-9.16 [-12.74, -5.57]	**+3.36 [1.93, 4.95]**
	DNA	NTv3	-9.39 [-11.02, -7.59]	**+0.41 [0.07, 0.72]**	-19.28 [-21.97, -16.33]	+0.25 [-0.22, 0.74]
	RNA	RiNALMo	-9.71 [-11.32, -8.13]	**+0.72 [0.29, 1.14]**	-19.88 [-22.57, -17.12]	**+0.57 [0.01, 1.12]**
	RNA	Orthrus	-8.59 [-10.34, -6.94]	**+1.41 [0.99, 1.82]**	-17.24 [-20.20, -14.10]	**+1.22 [0.64, 1.79]**
	Protein	ESM-C	+1.27 [-4.11, 7.07]	**+6.71 [3.68, 9.88]**	**+16.44 [2.47, 30.85]**	**+21.24 [9.97, 33.08]**
	Protein	SaProt	+1.90 [-3.80, 7.86]	**+5.49 [2.08, 8.65]**	**+14.63 [1.61, 28.10]**	**+15.86 [6.30, 25.86]**
	Other	GPN-Star (V)	-0.87 [-2.60, 0.88]	**+3.72 [2.94, 4.52]**	**+5.78 [1.23, 10.13]**	**+13.02 [9.91, 16.21]**
	Other	GPN-Star (M)	-1.48 [-3.18, 0.21]	**+3.54 [2.68, 4.40]**	**+6.11 [1.32, 10.52]**	**+12.92 [10.00, 15.97]**
	Other	GPN-Star (P)	-3.08 [-4.77, -1.46]	**+2.76 [1.98, 3.59]**	+2.14 [-2.19, 6.35]	**+9.60 [6.94, 12.24]**
	Other	AlphaGenome VI	**+7.22 [5.41, 8.99]**	**+9.82 [8.49, 11.08]**	**+26.83 [21.31, 32.41]**	**+29.83 [25.23, 34.59]**
GWAS	DNA	Evo2	**+11.00 [6.16, 15.60]**	**+14.22 [11.15, 17.39]**	**+128.81 [92.77, 168.11]**	**+129.22 [98.33, 163.02]**
	DNA	NTv3	-11.76 [-15.67, -7.44]	+0.59 [-0.59, 1.86]	-18.88 [-29.73, -6.17]	+5.73 [-0.56, 13.26]
	RNA	RiNALMo	-15.37 [-19.20, -11.26]	-0.05 [-0.55, 0.39]	-32.32 [-39.50, -24.54]	-0.39 [-1.30, 0.53]
	RNA	Orthrus	-16.10 [-19.74, -12.12]	+0.57 [-0.16, 1.23]	-34.77 [-41.72, -27.59]	+0.98 [-0.17, 2.19]
	Protein	ESM-C	**+13.62 [5.57, 22.56]**	**+15.73 [9.36, 22.73]**	**+58.03 [30.71, 88.64]**	**+58.14 [33.94, 83.23]**
	Protein	SaProt	**+16.21 [8.31, 24.89]**	**+17.55 [11.08, 24.71]**	**+57.26 [32.50, 87.59]**	**+57.52 [35.66, 83.84]**
	Other	GPN-Star (V)	**+20.81 [16.37, 25.37]**	**+23.32 [19.84, 26.60]**	**+215.02 [169.44, 271.89]**	**+217.44 [175.00, 269.00]**
	Other	GPN-Star (M)	**+22.76 [18.29, 27.29]**	**+24.54 [20.93, 28.40]**	**+237.60 [187.21, 299.79]**	**+239.15 [191.17, 296.37]**
	Other	GPN-Star (P)	**+21.06 [16.46, 25.85]**	**+22.88 [19.21, 26.79]**	**+207.96 [162.84, 262.65]**	**+210.67 [167.96, 262.60]**
	Other	AlphaGenome VI	**+25.63 [21.15, 30.39]**	**+27.38 [23.53, 31.61]**	**+234.47 [186.83, 288.65]**	**+238.18 [195.31, 289.25]**
eQTL	DNA	Evo2	-10.46 [-12.28, -8.65]	**+0.50 [0.19, 0.83]**	-19.79 [-23.00, -16.55]	**+0.78 [0.04, 1.62]**
	DNA	NTv3	-10.12 [-11.85, -8.30]	**+1.11 [0.63, 1.58]**	-22.85 [-25.65, -19.98]	**+0.90 [0.26, 1.47]**
	RNA	RiNALMo	-10.65 [-12.46, -8.86]	**+1.17 [0.69, 1.63]**	-23.57 [-26.34, -20.66]	**+0.79 [0.19, 1.35]**
	RNA	Orthrus	-9.32 [-11.13, -7.46]	**+1.75 [1.27, 2.25]**	-22.09 [-25.14, -19.06]	**+1.20 [0.53, 1.96]**
	Protein	ESM-C	-14.22 [-20.93, -6.40]	+0.18 [-1.87, 2.36]	-16.38 [-29.62, 0.10]	+4.36 [-1.92, 11.83]
	Protein	SaProt	-14.71 [-22.07, -6.74]	-0.31 [-2.17, 1.65]	-22.11 [-33.41, -8.32]	+0.24 [-4.64, 5.66]
	Other	GPN-Star (V)	-4.91 [-6.63, -3.10]	**+2.32 [1.60, 3.05]**	-6.53 [-10.50, -2.35]	**+6.47 [4.32, 8.91]**
	Other	GPN-Star (M)	-5.60 [-7.31, -3.88]	**+2.10 [1.32, 2.93]**	-5.97 [-9.91, -1.88]	**+6.99 [4.91, 9.23]**
	Other	GPN-Star (P)	-6.81 [-8.59, -5.11]	**+1.39 [0.72, 2.09]**	-9.85 [-13.60, -6.07]	**+3.95 [2.33, 5.71]**
	Other	AlphaGenome VI	**+4.25 [2.33, 6.08]**	**+8.29 [7.06, 9.51]**	**+16.12 [10.83, 21.63]**	**+23.02 [19.14, 27.37]**
ClinVar	DNA	Evo2	**+86.48 [84.81, 88.34]**	**+86.52 [84.84, 88.37]**	**+439.58 [428.78, 451.51]**	**+439.67 [428.67, 451.53]**
	DNA	NTv3	**+62.99 [61.42, 64.82]**	**+63.97 [62.47, 65.64]**	**+324.26 [316.05, 333.49]**	**+328.94 [320.75, 337.90]**
	RNA	RiNALMo	**+47.13 [45.66, 48.73]**	**+47.73 [46.24, 49.36]**	**+278.21 [270.47, 286.56]**	**+280.15 [272.36, 288.67]**
	RNA	Orthrus	**+12.14 [10.81, 13.58]**	**+13.50 [12.30, 14.78]**	**+80.11 [75.49, 85.02]**	**+83.91 [79.35, 88.44]**
	Protein	ESM-C	**+73.97 [71.23, 76.93]**	**+74.03 [71.32, 76.97]**	**+159.89 [153.32, 167.28]**	**+160.01 [153.43, 167.39]**
	Protein	SaProt	**+75.26 [72.51, 78.16]**	**+75.34 [72.57, 78.23]**	**+160.60 [154.10, 167.83]**	**+161.18 [154.85, 168.51]**
	Other	GPN-Star (V)	**+83.41 [81.68, 85.25]**	**+83.97 [82.28, 85.82]**	**+445.77 [434.81, 458.13]**	**+446.83 [435.87, 459.12]**
	Other	GPN-Star (M)	**+83.34 [81.59, 85.20]**	**+83.89 [82.18, 85.73]**	**+436.84 [426.31, 448.61]**	**+438.60 [428.05, 450.08]**
	Other	GPN-Star (P)	**+77.00 [75.30, 78.89]**	**+77.42 [75.79, 79.29]**	**+400.02 [389.71, 410.83]**	**+401.35 [391.02, 412.27]**
	Other	AlphaGenome VI	**+89.24 [87.58, 91.12]**	**+89.24 [87.59, 91.12]**	**+486.24 [474.83, 498.79]**	**+486.48 [474.81, 499.11]**

E.10 MSA AND SEQUENCE-TO-FUNCTION COMPLEMENTARITY ASSESSMENT BY DATASET

The following plots show the same feature complementarity models tested in Figure 4 (i.e. GPN-Star models and AlphaGenome VI), broken down by dataset and with improvements quantified with both AUROC and AUPRC.

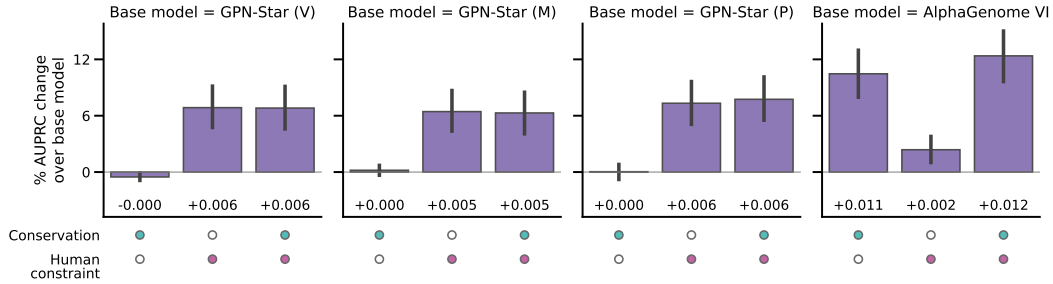

Figure E.7: **Effect of feature combinations on combined GWAS and eQTL variant effect predictions (AUPRC).**

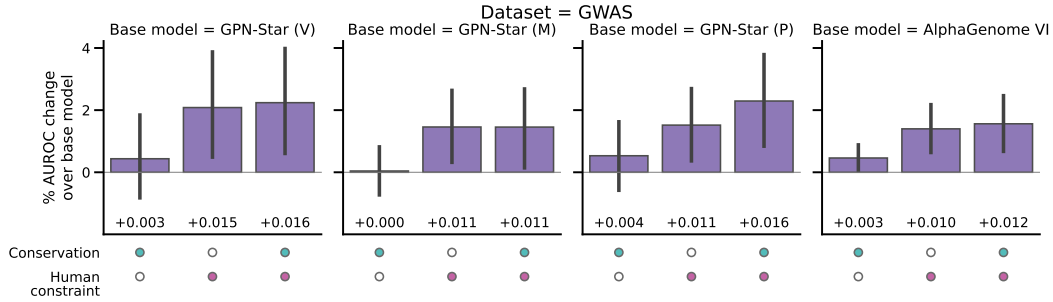

Figure E.8: **Effect of feature combinations on GWAS variant effect predictions.**

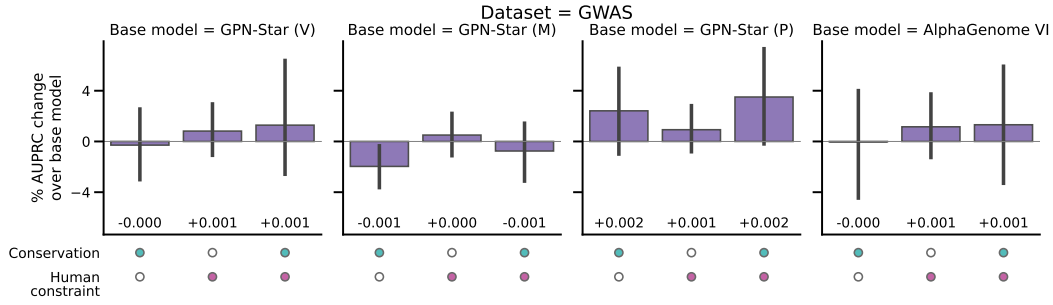

Figure E.9: **Effect of feature combinations on GWAS variant effect predictions (AUPRC).**

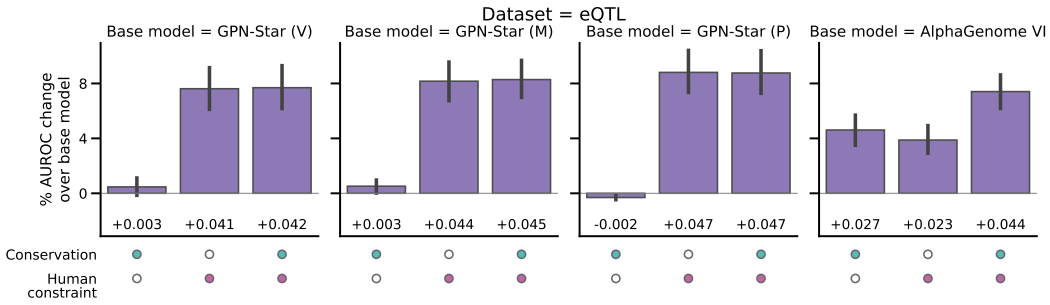

Figure E.10: **Effect of feature combinations on eQTL variant effect predictions.**

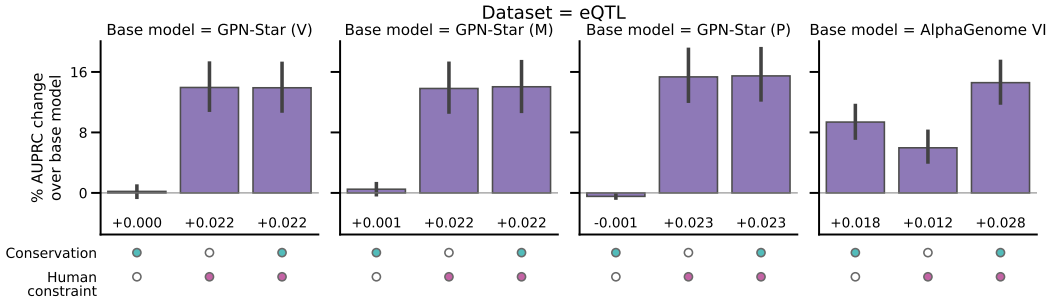

Figure E.11: **Effect of feature combinations on eQTL variant effect predictions (AUPRC).**

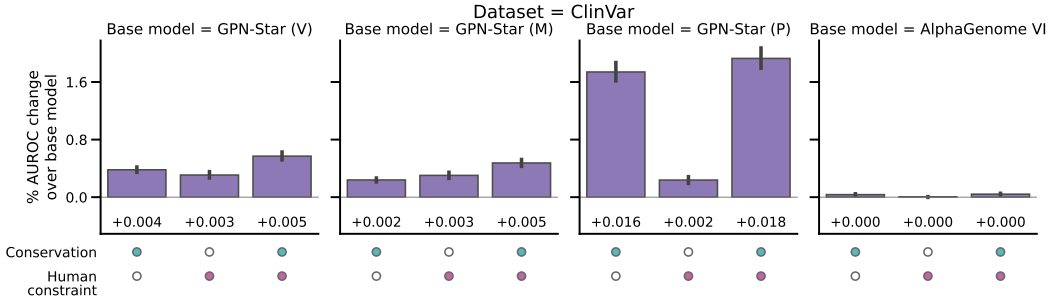

Figure E.12: **Effect of feature combinations on ClinVar variant effect predictions.**

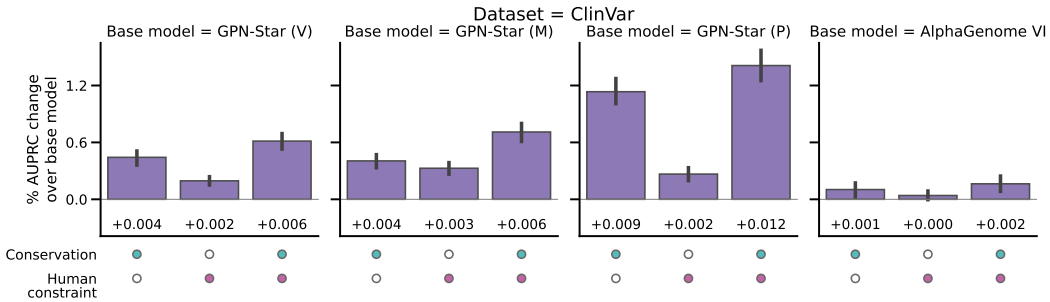

Figure E.13: **Effect of feature combinations on ClinVar variant effect predictions (AUPRC).**

E.11 ABSOLUTE PERFORMANCE VALUES OF FEATURE COMBINATION MODELS

Table E.6 provides the exhaustive set of performance metrics of all tested models and feature combinations on each dataset. Importantly, brackets show 95% percentile intervals from bootstrapping for individual model performance. For assessing performance differences *between models*, the paired bootstrap values from Appendix Tables E.4 and E.5 should be used instead.

Table E.6: Absolute AUROC and AUPRC for each model and combination of scores. Brackets give 95% percentile intervals for bootstrapped absolute performance values.

Dataset	Modality	Base model	Score combination	AUROC	AUPRC
GWAS+eQTL	DNA	Evo2	Base model	0.514 [0.507, 0.521]	0.072 [0.070, 0.075]
			+ phyloP	0.516 [0.509, 0.523]	0.075 [0.072, 0.079]
			+ Gnocchi	0.564 [0.558, 0.571]	0.082 [0.079, 0.085]
			+ Gnocchi, phyloP	0.568 [0.561, 0.575]	0.085 [0.082, 0.088]
			phyloP alone	0.519 [0.512, 0.526]	0.074 [0.071, 0.077]
			Gnocchi alone	0.559 [0.552, 0.566]	0.080 [0.077, 0.083]
GWAS+eQTL	DNA	NTv3	Base model	0.507 [0.500, 0.514]	0.064 [0.062, 0.066]
			+ phyloP	0.530 [0.523, 0.537]	0.074 [0.072, 0.077]
			+ Gnocchi	0.561 [0.555, 0.568]	0.080 [0.077, 0.083]
			+ Gnocchi, phyloP	0.569 [0.562, 0.576]	0.084 [0.081, 0.087]
			phyloP alone	0.519 [0.512, 0.526]	0.074 [0.071, 0.077]
			Gnocchi alone	0.559 [0.552, 0.566]	0.080 [0.077, 0.083]
GWAS+eQTL	RNA	RiNALMo	Base model	0.505 [0.498, 0.511]	0.064 [0.062, 0.066]
			+ phyloP	0.530 [0.522, 0.537]	0.075 [0.072, 0.078]
			+ Gnocchi	0.563 [0.556, 0.570]	0.080 [0.077, 0.083]
			+ Gnocchi, phyloP	0.570 [0.563, 0.577]	0.084 [0.081, 0.087]
			phyloP alone	0.519 [0.512, 0.526]	0.074 [0.071, 0.077]
			Gnocchi alone	0.559 [0.552, 0.566]	0.080 [0.077, 0.083]
GWAS+eQTL	RNA	Orthrus	Base model	0.511 [0.504, 0.517]	0.066 [0.064, 0.068]
			+ phyloP	0.533 [0.525, 0.540]	0.077 [0.074, 0.080]
			+ Gnocchi	0.567 [0.560, 0.573]	0.081 [0.078, 0.084]
			+ Gnocchi, phyloP	0.574 [0.567, 0.581]	0.085 [0.082, 0.088]
			phyloP alone	0.519 [0.512, 0.526]	0.074 [0.071, 0.077]
			Gnocchi alone	0.559 [0.552, 0.566]	0.080 [0.077, 0.083]
GWAS+eQTL	Protein	ESM-C	Base model	0.550 [0.528, 0.571]	0.121 [0.107, 0.136]
			+ phyloP	0.576 [0.552, 0.598]	0.130 [0.117, 0.145]
			+ Gnocchi	0.579 [0.558, 0.601]	0.126 [0.112, 0.142]
			+ Gnocchi, phyloP	0.589 [0.567, 0.611]	0.130 [0.117, 0.146]
			phyloP alone	0.564 [0.540, 0.586]	0.117 [0.105, 0.128]
			Gnocchi alone	0.543 [0.522, 0.565]	0.104 [0.094, 0.115]
GWAS+eQTL	Protein	SaProt	Base model	0.553 [0.532, 0.575]	0.119 [0.107, 0.134]
			+ phyloP	0.569 [0.547, 0.592]	0.126 [0.113, 0.141]
			+ Gnocchi	0.573 [0.552, 0.593]	0.120 [0.108, 0.135]
			+ Gnocchi, phyloP	0.580 [0.558, 0.603]	0.125 [0.112, 0.140]
			phyloP alone	0.564 [0.540, 0.586]	0.117 [0.105, 0.128]
			Gnocchi alone	0.543 [0.522, 0.565]	0.104 [0.094, 0.115]
GWAS+eQTL	Other	GPN-Star (V)	Base model	0.554 [0.546, 0.561]	0.084 [0.081, 0.088]
			+ phyloP	0.551 [0.543, 0.558]	0.084 [0.080, 0.088]
			+ Gnocchi	0.580 [0.573, 0.587]	0.090 [0.086, 0.094]
			+ Gnocchi, phyloP	0.579 [0.572, 0.587]	0.090 [0.086, 0.094]
			phyloP alone	0.519 [0.512, 0.526]	0.074 [0.071, 0.077]
			Gnocchi alone	0.559 [0.552, 0.566]	0.080 [0.077, 0.083]
GWAS+eQTL	Other	GPN-Star (M)	Base model	0.551 [0.544, 0.558]	0.085 [0.081, 0.088]
			+ phyloP	0.554 [0.547, 0.562]	0.085 [0.081, 0.088]
			+ Gnocchi	0.579 [0.572, 0.587]	0.090 [0.087, 0.094]
			+ Gnocchi, phyloP	0.579 [0.572, 0.586]	0.090 [0.086, 0.093]
			phyloP alone	0.519 [0.512, 0.526]	0.074 [0.071, 0.077]
			Gnocchi alone	0.559 [0.552, 0.566]	0.080 [0.077, 0.083]
GWAS+eQTL	Other	GPN-Star (P)	Base model	0.542 [0.535, 0.550]	0.081 [0.078, 0.085]
			+ phyloP	0.541 [0.533, 0.548]	0.081 [0.078, 0.085]
			+ Gnocchi	0.574 [0.568, 0.582]	0.087 [0.084, 0.091]
			+ Gnocchi, phyloP	0.575 [0.568, 0.583]	0.088 [0.084, 0.091]
			phyloP alone	0.519 [0.512, 0.526]	0.074 [0.071, 0.077]
			Gnocchi alone	0.559 [0.552, 0.566]	0.080 [0.077, 0.083]
GWAS+eQTL	Other	AlphaGenome VI	Base model	0.599 [0.592, 0.606]	0.101 [0.096, 0.106]
			+ phyloP	0.624 [0.617, 0.631]	0.112 [0.107, 0.116]
			+ Gnocchi	0.614 [0.607, 0.621]	0.103 [0.099, 0.108]
			+ Gnocchi, phyloP	0.634 [0.627, 0.641]	0.114 [0.109, 0.118]

Dataset	Modality	Base model	Score combination	AUROC	AUPRC
			phyloP alone	0.519 [0.512, 0.526]	0.074 [0.071, 0.077]
			Gnocchi alone	0.559 [0.552, 0.566]	0.080 [0.077, 0.083]
GWAS	DNA	Evo2	Base model	0.656 [0.637, 0.676]	0.048 [0.042, 0.056]
			+ phyloP	0.698 [0.679, 0.717]	0.060 [0.052, 0.069]
			+ Gnocchi	0.676 [0.657, 0.694]	0.048 [0.042, 0.055]
			+ Gnocchi, phyloP	0.716 [0.699, 0.735]	0.059 [0.051, 0.067]
			phyloP alone	0.694 [0.675, 0.712]	0.056 [0.049, 0.065]
			Gnocchi alone	0.592 [0.572, 0.609]	0.021 [0.019, 0.023]
GWAS	DNA	NTv3	Base model	0.522 [0.505, 0.540]	0.017 [0.015, 0.019]
			+ phyloP	0.696 [0.677, 0.714]	0.059 [0.050, 0.070]
			+ Gnocchi	0.595 [0.577, 0.613]	0.022 [0.020, 0.025]
			+ Gnocchi, phyloP	0.711 [0.693, 0.729]	0.057 [0.049, 0.066]
			phyloP alone	0.694 [0.675, 0.712]	0.056 [0.049, 0.065]
			Gnocchi alone	0.592 [0.572, 0.609]	0.021 [0.019, 0.023]
GWAS	RNA	RiNALMo	Base model	0.500 [0.483, 0.517]	0.014 [0.013, 0.016]
			+ phyloP	0.695 [0.676, 0.714]	0.056 [0.048, 0.064]
			+ Gnocchi	0.591 [0.573, 0.609]	0.021 [0.019, 0.023]
			+ Gnocchi, phyloP	0.711 [0.692, 0.730]	0.056 [0.049, 0.064]
			phyloP alone	0.694 [0.675, 0.712]	0.056 [0.049, 0.065]
			Gnocchi alone	0.592 [0.572, 0.609]	0.021 [0.019, 0.023]
GWAS	RNA	Orthrus	Base model	0.496 [0.481, 0.513]	0.014 [0.013, 0.015]
			+ phyloP	0.697 [0.678, 0.715]	0.058 [0.050, 0.068]
			+ Gnocchi	0.595 [0.576, 0.613]	0.021 [0.019, 0.024]
			+ Gnocchi, phyloP	0.712 [0.693, 0.731]	0.056 [0.049, 0.064]
			phyloP alone	0.694 [0.675, 0.712]	0.056 [0.049, 0.065]
			Gnocchi alone	0.592 [0.572, 0.609]	0.021 [0.019, 0.023]
GWAS	Protein	ESM-C	Base model	0.616 [0.585, 0.646]	0.114 [0.095, 0.135]
			+ phyloP	0.672 [0.645, 0.701]	0.127 [0.107, 0.148]
			+ Gnocchi	0.627 [0.598, 0.656]	0.114 [0.096, 0.135]
			+ Gnocchi, phyloP	0.676 [0.649, 0.705]	0.124 [0.106, 0.145]
			phyloP alone	0.656 [0.629, 0.684]	0.108 [0.094, 0.123]
			Gnocchi alone	0.542 [0.515, 0.571]	0.072 [0.063, 0.083]
GWAS	Protein	SaProt	Base model	0.630 [0.602, 0.659]	0.113 [0.095, 0.134]
			+ phyloP	0.670 [0.641, 0.699]	0.126 [0.108, 0.148]
			+ Gnocchi	0.637 [0.609, 0.667]	0.113 [0.097, 0.133]
			+ Gnocchi, phyloP	0.675 [0.647, 0.703]	0.123 [0.106, 0.143]
			phyloP alone	0.656 [0.629, 0.684]	0.108 [0.094, 0.123]
			Gnocchi alone	0.542 [0.515, 0.571]	0.072 [0.063, 0.083]
GWAS	Other	GPN-Star (V)	Base model	0.714 [0.694, 0.732]	0.066 [0.057, 0.077]
			+ phyloP	0.718 [0.699, 0.735]	0.066 [0.058, 0.076]
			+ Gnocchi	0.729 [0.710, 0.748]	0.067 [0.058, 0.077]
			+ Gnocchi, phyloP	0.730 [0.711, 0.748]	0.067 [0.058, 0.077]
			phyloP alone	0.694 [0.675, 0.712]	0.056 [0.049, 0.065]
			Gnocchi alone	0.592 [0.572, 0.609]	0.021 [0.019, 0.023]
GWAS	Other	GPN-Star (M)	Base model	0.726 [0.706, 0.744]	0.071 [0.061, 0.082]
			+ phyloP	0.726 [0.707, 0.744]	0.070 [0.059, 0.080]
			+ Gnocchi	0.737 [0.716, 0.755]	0.072 [0.061, 0.082]
			+ Gnocchi, phyloP	0.737 [0.717, 0.754]	0.071 [0.060, 0.081]
			phyloP alone	0.694 [0.675, 0.712]	0.056 [0.049, 0.065]
			Gnocchi alone	0.592 [0.572, 0.609]	0.021 [0.019, 0.023]
GWAS	Other	GPN-Star (P)	Base model	0.716 [0.697, 0.734]	0.065 [0.056, 0.075]
			+ phyloP	0.720 [0.702, 0.738]	0.067 [0.057, 0.077]
			+ Gnocchi	0.727 [0.707, 0.745]	0.066 [0.056, 0.075]
			+ Gnocchi, phyloP	0.732 [0.713, 0.750]	0.067 [0.058, 0.077]
			phyloP alone	0.694 [0.675, 0.712]	0.056 [0.049, 0.065]
			Gnocchi alone	0.592 [0.572, 0.609]	0.021 [0.019, 0.023]
GWAS	Other	AlphaGenome VI	Base model	0.743 [0.724, 0.761]	0.071 [0.062, 0.081]
			+ phyloP	0.746 [0.727, 0.763]	0.070 [0.062, 0.081]
			+ Gnocchi	0.753 [0.736, 0.771]	0.071 [0.062, 0.082]
			+ Gnocchi, phyloP	0.755 [0.736, 0.771]	0.071 [0.062, 0.082]
			phyloP alone	0.694 [0.675, 0.712]	0.056 [0.049, 0.065]
			Gnocchi alone	0.592 [0.572, 0.609]	0.021 [0.019, 0.023]
eQTL	DNA	Evo2	Base model	0.512 [0.505, 0.520]	0.135 [0.130, 0.140]
			+ phyloP	0.511 [0.504, 0.520]	0.140 [0.135, 0.145]
			+ Gnocchi	0.575 [0.567, 0.583]	0.169 [0.163, 0.176]
			+ Gnocchi, phyloP	0.575 [0.567, 0.583]	0.171 [0.165, 0.177]
			phyloP alone	0.506 [0.498, 0.514]	0.139 [0.134, 0.144]
			Gnocchi alone	0.572 [0.564, 0.580]	0.168 [0.162, 0.175]
eQTL	DNA	NTv3	Base model	0.514 [0.507, 0.522]	0.130 [0.126, 0.134]
			+ phyloP	0.524 [0.517, 0.532]	0.141 [0.136, 0.146]

Dataset	Modality	Base model	Score combination	AUROC	AUPRC
			+ Gnocchi	0.579 [0.571, 0.587]	0.170 [0.164, 0.176]
			+ Gnocchi, phyloP	0.579 [0.572, 0.587]	0.172 [0.165, 0.178]
			phyloP alone	0.506 [0.498, 0.514]	0.139 [0.134, 0.144]
			Gnocchi alone	0.572 [0.564, 0.580]	0.168 [0.162, 0.175]
eQTL	RNA	RiNALMo	Base model	0.511 [0.504, 0.518]	0.128 [0.125, 0.133]
			+ phyloP	0.522 [0.514, 0.530]	0.141 [0.136, 0.147]
			+ Gnocchi	0.579 [0.571, 0.587]	0.169 [0.163, 0.176]
			+ Gnocchi, phyloP	0.580 [0.573, 0.588]	0.172 [0.166, 0.178]
			phyloP alone	0.506 [0.498, 0.514]	0.139 [0.134, 0.144]
			Gnocchi alone	0.572 [0.564, 0.580]	0.168 [0.162, 0.175]
eQTL	RNA	Orthrus	Base model	0.519 [0.512, 0.527]	0.131 [0.127, 0.135]
			+ phyloP	0.526 [0.518, 0.534]	0.143 [0.137, 0.148]
			+ Gnocchi	0.582 [0.575, 0.590]	0.170 [0.164, 0.177]
			+ Gnocchi, phyloP	0.583 [0.575, 0.591]	0.172 [0.166, 0.179]
			phyloP alone	0.506 [0.498, 0.514]	0.139 [0.134, 0.144]
			Gnocchi alone	0.572 [0.564, 0.580]	0.168 [0.162, 0.175]
eQTL	Protein	ESM-C	Base model	0.501 [0.467, 0.534]	0.150 [0.127, 0.177]
			+ phyloP	0.511 [0.476, 0.544]	0.157 [0.133, 0.184]
			+ Gnocchi	0.585 [0.550, 0.620]	0.188 [0.159, 0.219]
			+ Gnocchi, phyloP	0.576 [0.542, 0.607]	0.180 [0.153, 0.211]
			phyloP alone	0.507 [0.475, 0.541]	0.144 [0.126, 0.166]
			Gnocchi alone	0.584 [0.550, 0.619]	0.180 [0.153, 0.208]
eQTL	Protein	SaProt	Base model	0.498 [0.464, 0.532]	0.140 [0.121, 0.160]
			+ phyloP	0.494 [0.460, 0.525]	0.140 [0.121, 0.161]
			+ Gnocchi	0.582 [0.547, 0.616]	0.180 [0.153, 0.208]
			+ Gnocchi, phyloP	0.567 [0.532, 0.598]	0.171 [0.146, 0.195]
			phyloP alone	0.507 [0.475, 0.541]	0.144 [0.126, 0.166]
			Gnocchi alone	0.584 [0.550, 0.619]	0.180 [0.153, 0.208]
eQTL	Other	GPN-Star (V)	Base model	0.544 [0.536, 0.552]	0.157 [0.151, 0.164]
			+ phyloP	0.547 [0.538, 0.555]	0.157 [0.151, 0.164]
			+ Gnocchi	0.586 [0.578, 0.594]	0.179 [0.172, 0.186]
			+ Gnocchi, phyloP	0.586 [0.578, 0.594]	0.179 [0.172, 0.186]
			phyloP alone	0.506 [0.498, 0.514]	0.139 [0.134, 0.144]
			Gnocchi alone	0.572 [0.564, 0.580]	0.168 [0.162, 0.175]
eQTL	Other	GPN-Star (M)	Base model	0.540 [0.533, 0.548]	0.158 [0.152, 0.165]
			+ phyloP	0.543 [0.535, 0.551]	0.159 [0.153, 0.165]
			+ Gnocchi	0.584 [0.576, 0.592]	0.180 [0.173, 0.187]
			+ Gnocchi, phyloP	0.585 [0.577, 0.593]	0.180 [0.173, 0.187]
			phyloP alone	0.506 [0.498, 0.514]	0.139 [0.134, 0.144]
			Gnocchi alone	0.572 [0.564, 0.580]	0.168 [0.162, 0.175]
eQTL	Other	GPN-Star (P)	Base model	0.533 [0.526, 0.541]	0.152 [0.146, 0.157]
			+ phyloP	0.532 [0.524, 0.539]	0.151 [0.145, 0.157]
			+ Gnocchi	0.580 [0.572, 0.588]	0.175 [0.168, 0.182]
			+ Gnocchi, phyloP	0.580 [0.572, 0.588]	0.175 [0.168, 0.182]
			phyloP alone	0.506 [0.498, 0.514]	0.139 [0.134, 0.144]
			Gnocchi alone	0.572 [0.564, 0.580]	0.168 [0.162, 0.175]
eQTL	Other	AlphaGenome VI	Base model	0.597 [0.589, 0.605]	0.195 [0.188, 0.204]
			+ phyloP	0.624 [0.617, 0.631]	0.213 [0.205, 0.222]
			+ Gnocchi	0.620 [0.612, 0.627]	0.207 [0.198, 0.216]
			+ Gnocchi, phyloP	0.641 [0.633, 0.648]	0.224 [0.214, 0.233]
			phyloP alone	0.506 [0.498, 0.514]	0.139 [0.134, 0.144]
			Gnocchi alone	0.572 [0.564, 0.580]	0.168 [0.162, 0.175]
ClinVar	DNA	Evo2	Base model	0.977 [0.976, 0.979]	0.892 [0.888, 0.897]
			+ phyloP	0.979 [0.978, 0.980]	0.894 [0.890, 0.899]
			+ Gnocchi	0.978 [0.977, 0.979]	0.893 [0.888, 0.897]
			+ Gnocchi, phyloP	0.980 [0.979, 0.980]	0.895 [0.890, 0.899]
			phyloP alone	0.886 [0.883, 0.889]	0.626 [0.617, 0.634]
			Gnocchi alone	0.524 [0.519, 0.529]	0.165 [0.162, 0.169]
ClinVar	DNA	NTv3	Base model	0.854 [0.851, 0.858]	0.702 [0.695, 0.708]
			+ phyloP	0.946 [0.945, 0.948]	0.824 [0.818, 0.829]
			+ Gnocchi	0.860 [0.856, 0.863]	0.709 [0.703, 0.716]
			+ Gnocchi, phyloP	0.949 [0.947, 0.951]	0.831 [0.825, 0.836]
			phyloP alone	0.886 [0.883, 0.889]	0.626 [0.617, 0.634]
			Gnocchi alone	0.524 [0.519, 0.529]	0.165 [0.162, 0.169]
ClinVar	RNA	RiNALMo	Base model	0.771 [0.767, 0.776]	0.626 [0.619, 0.633]
			+ phyloP	0.947 [0.945, 0.948]	0.825 [0.820, 0.830]
			+ Gnocchi	0.774 [0.770, 0.779]	0.629 [0.622, 0.636]
			+ Gnocchi, phyloP	0.947 [0.946, 0.949]	0.828 [0.823, 0.833]
			phyloP alone	0.886 [0.883, 0.889]	0.626 [0.617, 0.634]
			Gnocchi alone	0.524 [0.519, 0.529]	0.165 [0.162, 0.169]

Dataset	Modality	Base model	Score combination	AUROC	AUPRC
ClinVar	RNA	Orthrus	Base model	0.588 [0.584, 0.593]	0.298 [0.291, 0.305]
			+ phyloP	0.907 [0.904, 0.909]	0.682 [0.675, 0.689]
			+ Gnocchi	0.595 [0.590, 0.600]	0.304 [0.298, 0.311]
			+ Gnocchi, phyloP	0.908 [0.906, 0.911]	0.689 [0.681, 0.696]
			phyloP alone	0.886 [0.883, 0.889]	0.626 [0.617, 0.634]
			Gnocchi alone	0.524 [0.519, 0.529]	0.165 [0.162, 0.169]
ClinVar	Protein	ESM-C	Base model	0.929 [0.925, 0.933]	0.894 [0.887, 0.901]
			+ phyloP	0.952 [0.949, 0.955]	0.913 [0.907, 0.919]
			+ Gnocchi	0.929 [0.925, 0.934]	0.894 [0.888, 0.901]
			+ Gnocchi, phyloP	0.952 [0.949, 0.956]	0.913 [0.907, 0.919]
			phyloP alone	0.841 [0.836, 0.847]	0.651 [0.637, 0.664]
			Gnocchi alone	0.534 [0.525, 0.542]	0.344 [0.334, 0.353]
ClinVar	Protein	SaProt	Base model	0.936 [0.932, 0.940]	0.896 [0.889, 0.904]
			+ phyloP	0.955 [0.951, 0.958]	0.913 [0.906, 0.920]
			+ Gnocchi	0.936 [0.932, 0.940]	0.898 [0.891, 0.905]
			+ Gnocchi, phyloP	0.955 [0.952, 0.958]	0.915 [0.908, 0.921]
			phyloP alone	0.841 [0.836, 0.847]	0.651 [0.637, 0.664]
			Gnocchi alone	0.534 [0.525, 0.542]	0.344 [0.334, 0.353]
ClinVar	Other	GPN-Star (V)	Base model	0.961 [0.959, 0.963]	0.903 [0.899, 0.907]
			+ phyloP	0.965 [0.963, 0.967]	0.907 [0.903, 0.910]
			+ Gnocchi	0.964 [0.962, 0.966]	0.904 [0.901, 0.908]
			+ Gnocchi, phyloP	0.967 [0.965, 0.969]	0.908 [0.905, 0.912]
			phyloP alone	0.886 [0.883, 0.889]	0.626 [0.617, 0.634]
			Gnocchi alone	0.524 [0.519, 0.529]	0.165 [0.162, 0.169]
ClinVar	Other	GPN-Star (M)	Base model	0.961 [0.959, 0.963]	0.888 [0.884, 0.892]
			+ phyloP	0.963 [0.962, 0.965]	0.892 [0.887, 0.896]
			+ Gnocchi	0.964 [0.962, 0.966]	0.891 [0.887, 0.895]
			+ Gnocchi, phyloP	0.966 [0.964, 0.967]	0.894 [0.890, 0.898]
			phyloP alone	0.886 [0.883, 0.889]	0.626 [0.617, 0.634]
			Gnocchi alone	0.524 [0.519, 0.529]	0.165 [0.162, 0.169]
ClinVar	Other	GPN-Star (P)	Base model	0.928 [0.925, 0.931]	0.827 [0.822, 0.832]
			+ phyloP	0.944 [0.942, 0.946]	0.836 [0.831, 0.842]
			+ Gnocchi	0.930 [0.927, 0.933]	0.829 [0.824, 0.835]
			+ Gnocchi, phyloP	0.946 [0.943, 0.948]	0.839 [0.834, 0.844]
			phyloP alone	0.886 [0.883, 0.889]	0.626 [0.617, 0.634]
			Gnocchi alone	0.524 [0.519, 0.529]	0.165 [0.162, 0.169]
ClinVar	Other	AlphaGenome VI	Base model	0.992 [0.991, 0.993]	0.970 [0.967, 0.972]
			+ phyloP	0.992 [0.992, 0.993]	0.971 [0.968, 0.973]
			+ Gnocchi	0.992 [0.991, 0.993]	0.970 [0.967, 0.972]
			+ Gnocchi, phyloP	0.992 [0.992, 0.993]	0.971 [0.969, 0.973]
			phyloP alone	0.886 [0.883, 0.889]	0.626 [0.617, 0.634]
			Gnocchi alone	0.524 [0.519, 0.529]	0.165 [0.162, 0.169]