GENOMIC HETEROGENEITY INFLATES THE PERFORMANCE OF VARIANT PATHOGENICITY PREDICTIONS

Baiyu Lu[*]
Courant Institute of Mathematical Sciences, New York University
New York, NY, USA
baiyu.lu@nyu.edu

Xueshen Liu[*]
Courant Institute of Mathematical Sciences, New York University
New York, NY, USA
xl5279@nyu.edu

Po-Yu Lin[*]
Center for Human Genetics & Genomics, New York University
Department of Neurology & Genomic Medicine, National Cheng Kung University Hospital
College of Medicine, National Cheng Kung University
Tainan, Taiwan
Po-Yu.Lin@nyulangone.org

Nadav Brandes[†]
Center for Human Genetics & Genomics, New York University
New York, NY, USA
nadav.brandes@nyulangone.org

ABSTRACT

Recent studies have reported unprecedented accuracy predicting pathogenic variants across the genome, including in noncoding regions, using large AI models trained on vast genomic data. We present a comprehensive evaluation of these frontier models, showing that performance is inflated by differences in the prevalence of pathogenic variants across genomic contexts. We identify the best-performing models for each variant type and establish a benchmark to guide future progress.

1 BACKGROUND

Mendelian diseases arising from single pathogenic variants account for most of the $> 7{,}000$ known rare conditions that collectively affect 3–6% of the population The Lancet Global, H. (2024). However, current tests provide a clear genetic diagnosis for only $\sim$50% of cases Montano et al. (2022). This gap largely reflects our limited understanding of variant effects, especially those involving regulatory genomic functions which remain underdiagnosed Ellingford et al. (2022).

To help close this gap, numerous computational methods have been developed over the past 25 years to predict whether variants are pathogenic or benign. In recent years, a major advance came with large-scale artificial intelligence (AI) models trained on massive datasets of protein or DNA sequences (Table 1) Brixi et al. (2025); Benegas et al. (2025); Avsec et al. (2025); Albors et al. (2025); Dalla-Torre et al. (2024); Zhou et al. (2023); Lin et al. (2023); Gao et al. (2023); Cheng et al. (2023); Rives et al. (2021); Meier et al. (2021). These models not only improved prediction accuracy but, thanks to their largely unsupervised training, also gained greater generality. This

[*]Equal contribution.
[†]Corresponding author.

Table 1: Frontier sequence-based models evaluated in this study

MODEL	MODALITY	PARAMS	CTX	INPUT	DATA	TASK
Evo2	DNA	7B	8192	seq.	cross-sp. seq.[§]	autoreg. LM
AlphaGenome	DNA	450M	1M	seq.	func. geno.[¶]	predict annot.[‡]
GPN-MSA	DNA	86M	128	MSA	cross-sp. seq.	masked LM
PhyloGPN	DNA	83M	481	seq.	cross-sp. seq.	phylo sub. probs
DNA-BERT2	DNA	117M	512[†]	seq.	cross-sp. seq.	masked LM
Nucleotide Trans.	DNA	2.5B	6000	seq.	cross-sp. seq.	masked LM
ESM-1b	protein	650M	1022	seq.	cross-sp. seq.	masked LM
ESM-1v	protein	650M	1022	seq.	cross-sp. seq.	masked LM
ESM-2	protein	650M	1022	seq.	cross-sp. seq.	masked LM
AlphaMissense	protein	unsp.	256	seq., MSA	cross-sp./human AF	class. freq./abs.[*]
PrimateAI-3D	protein	unsp.	unsp.	3D, MSA	cross-sp./primates AF	class. freq./abs.[*]

Notes: seq.: sequence; func. geno.: functional genomics; cross-sp.: cross-species; phylo sub.: phylogenetic substitution; AF: allele frequency; 3D: three-dimensional protein structure; autoreg.: autoregressive; MSA: multiple sequence alignment. [§]cross-sp. seq. = cross-species pretraining on the OpenGenome2 corpus; [¶]func. geno. = functional genomics signals (e.g., RNA-seq, ChIP-seq). [†]Ctx = 512 corresponds to models trained on context lengths of 128–700; [‡]predict annot. = predicting 7,058 genomic annotations; [*]class. freq./abs. = classifying mutations as frequent or absent in humans and primates.

approach has proven particularly valuable for regulatory regions, where only a limited number of variants are well characterized and can be used as prediction targets for supervised learning.

However, prior studies evaluating these models for pathogenicity prediction have each used different benchmarks and evaluation metrics, making it challenging to directly compare them. Furthermore, existing evaluations have measured performance on highly heterogeneous variant groups, a challenge we refer to as genomic heterogeneity (comprising diverse categories including coding, noncoding, splice site, and UTR variants, and others). For example, some studies have reported overall accuracy across all noncoding variants, without noting that a substantial portion of labeled variants are at canonical splice sites, which are almost always pathogenic and trivial to detect. As a result, it remains unclear whether these models simply separate broad variant categories (e.g., canonical splice sites vs. other intronic regions) or can also distinguish pathogenic from benign variants of the same type.

To address these gaps, we systematically benchmarked frontier sequence-based AI models (Table 1) at variant pathogenicity prediction and studied their strengths and weaknesses across variant types.

2 METHOD

CLINVAR BENCHMARK PREPARATION

We downloaded the ClinVar Landrum et al. (2016) variant summary file (release: 2025-03-31) and extracted 259,600 single-nucleotide variants (SNVs) mapped to the GRCh38 genome with definitive pathogenic or benign label (Figure 1). Each variant was annotated using two complementary strategies.

First, variants were mapped to transcripts according to the MANE RefSeq annotation database (v1.4 GFF file) Morales et al. (2022); O'Leary et al. (2016). Second, transcript information and amino-acid changes were also extracted directly from the HGVS nomenclature provided by ClinVar, providing an independent annotation source. Variants were categorized as noncoding if neither annotation indicated a protein-coding region.

Genomic region categories were assigned from either source when available, including start-loss, stop-loss, canonical splice site (defined as intron variants within 1–2 nt of exon boundaries), and RNA genes. Missense, synonymous, and stop-gain categories were taken from the ClinVar HGVS annotations, while the 5′ UTR, 3′ UTRs, and non-splice intron categories were derived exclusively from MANE annotations. The final benchmark and code for generating it are available online (see Data and Code Availability).

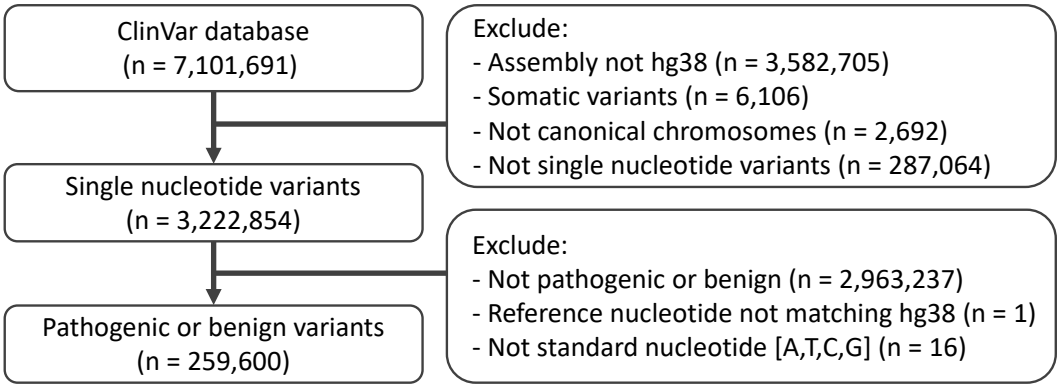


Figure 1: Pipeline for generating the ClinVar benchmark.

Calculating and Comparing AUROC Scores

We evaluated how well models distinguish between pathogenic and benign variants within each variant type using the area under the receiver operating characteristic curve (AUROC). AUROC has a straightforward probabilistic interpretation: it is the probability that a randomly chosen pathogenic variant receives a more damaging score than a randomly chosen benign variant. For example, an AUROC of 0.9 means that in 90% of such pairs, the pathogenic variant is ranked above the benign one.

This makes AUROC an intuitive measure of a model's ability to separate the two clinical labels. AUROC is also insensitive to class imbalance, since it conditions on the selection of one pathogenic and one benign variant, regardless of their prevalence. This is critical in our study, as variant-type categories differ dramatically in pathogenicity prevalence. For example, 99.3% of canonical splice variants are pathogenic compared to 1.6% of all other intron variants. Despite this disparity, the AUROC metric allows direct comparison between these variant types.

In addition, AUROC evaluates the full ranking of predictions without requiring calibration of the scores or selection of a cutoff, allowing models to be compared on equal footing. Other metrics, such as accuracy and the area under the precision-recall curve (AUPRC), are sensitive to class balance and therefore unsuitable for this study.

For models in which more negative scores indicate higher pathogenicity (Evo2, GPN-MSA, PhyloGPN, and the ESM models), we multiplied the scores by -1 so that higher values consistently indicate greater pathogenicity. We then computed the AUROC using these transformed scores and the pathogenicity labels (pathogenic = 1, benign = 0).

For each variant group (all variants, noncoding variants, or a specific variant type), variants with missing values in any of the compared models were excluded to ensure direct comparison on the same variant set. The only exception was stop-gain mutations, for which only 19,795 out of the 42,408 variants had PrimateAI-3D scores.

Variant Sequence Construction

For DNA models, we extracted sequences from the GRCh38 human genome, centered around each variant position, with sequence length determined by the model's input context (Table 1). For protein models, we used the full amino-acid sequence of the annotated protein isoform. If the protein exceeded the model's context length, we applied a sliding window approach, subdividing the sequence into overlapping windows of 1,022 aa with $>$511 aa overlaps Brandes et al. (2023). The final score was computed as a weighted sum of the scores across the relevant segments.

SCORING STOP-GAIN VARIANTS BY PROTEIN MODELS

Protein models do not directly provide predictions for stop-gain variants. To estimate their impact, we used the most deleterious prediction among possible missense variants within the truncated protein region (between the introduced stop codon and the end of the protein). For the ESM models, we considered all possible missense mutations, whereas for AlphaMissense and PrimateAI-3D we only considered missense mutations that can result from a single-nucleotide substitution between two codons (as only these scores were available).

EVO2

We used the `evo2_7b_base` model with 8,192 bp context length, downloaded from the Evo2 GitHub repository (`https://github.com/ArcInstitute/evo2`) Nguyen et al. (2024). Variant scores were computed using the built-in `model.score_sequences` function over the reference and alternative sequences separately to obtain the log-likelihood of each sequence. The variant effect score was calculated as the log-likelihood of the alternative sequence minus that of the reference sequence.

GPN-MSA

We used precomputed GPN-MSA variant scores downloaded from the authors' dataset (`https://huggingface.co/datasets/songlab/gpn-msa-hg38-scores/resolve/main/scores.tsv.bgz`) Benegas et al. (2025), which provides predicted effect scores for all possible SNVs in the GRCh38 genome.

PHYLOGPN

We obtained prediction scores from PhyloGPN using the code provided in the model's official GitHub repository (`https://github.com/songlab-cal/gpn/blob/main/examples/phylogpn/basic_example.ipynb`) Albors et al. (2025).

NUCLEOTIDE TRANSFORMER

We used the Nucleotide Transformer 2.5B Multi-Species model available on HuggingFace (`https://huggingface.co/InstaDeepAI/nucleotide-transformer-2.5b-multi-species`) Dalla-Torre et al. (2024) with a context length of 6,000 bp. As suggested by the authors, reference and alternate sequences were processed independently, and the L2 distance between their last hidden layer embeddings at the [CLS] token was used as the variant effect score.

DNABERT2

We used the DNABERT2–117M model via HuggingFace (`https://huggingface.co/zhihan1996/DNABERT-2-117M`) Zhou et al. (2023). Because DNABERT2 does not use single-nucleotide tokens, we could not directly compute log-likelihood ratios from its logits. Instead, we followed the same method used for the Nucleotide Transformer. Specifically, we used a context length of 512 bp (typically corresponding to fewer than 128 tokens under byte-pair encoding), consistent with the sequence lengths the model was trained on. Reference and alternate sequences were processed independently, and the L2 distance between their final [CLS] embeddings was taken as the variant effect score.

ESM MODELS

We followed the same pipeline described in our previous work Brandes et al. (2023). For ESM-1b, we used the `esm1b_t33_650M_UR50S` model from the official ESM GitHub repository (`https://github.com/facebookresearch/esm`) Rives et al. (2021). For ESM-1v, we used `esm1v_t33_650M_UR90S_1`, the first of the five ensemble models Meier et al. (2021). For ESM-2, we used `esm2_t33_650M_UR50D` Lin et al. (2023). All ESM models used a 1,022 aa

context length. For each variant, the reference sequence was used as input, and the variant effect score was calculated as the difference in logits (log-likelihood scores) between the alternative and reference residues.

ALPHAMISSENSE

We downloaded precomputed AlphaMissense variant effect scores from the authors' dataset (`https://alphamissense.hegelab.org/`) Cheng et al. (2023), which provides predictions for all possible missense variants in the human genome. There were 87 variants with more than one AlphaMissense score, which we excluded from our missense variant analysis.

For stop-gain variants, we mapped the RefSeq IDs from our ClinVar dataset to UniProt IDs in the AlphaMissense precomputed score files using the MyGene API O'Leary et al. (2016); The UniProt Consortium (2024); Wu et al. (2013). We then collected all missense scores within the UniProt sequence at or downstream of the stop-gain position and selected the most deleterious score among them.

PHYLOP

We downloaded precomputed PhyloP conservation scores based on the 100-way vertebrate alignment under the GRCh38/hg38 assembly from the UCSC Genome Browser Pollard et al. (2010); Kent et al. (2002).

PRIMATEAI-3D

We downloaded precomputed PrimateAI-3D variant effect scores from the official model website (`https://primateai3d.basespace.illumina.com/`) Gao et al. (2023). For stop-gain variants, we excluded 22,613 variants whose ClinVar isoform (identified by RefSeq ID) did not match any isoform in the PrimateAI-3D dataset, leaving 19,795 variants successfully mapped and used in our evaluation.

ALPHAGENOME

The AlphaGenome model outputs 5,930 human and 1,128 mouse genomic annotations across 11 modalities, including gene expression, splicing, chromatin states, and chromatin contact maps. We used the model via the official API (`https://deepmind.google.com/science/alphagenome/`) Avsec et al. (2025) with a 1 Mbp input context and organism set to human, enabling outputs from all modalities.

To calculate variant effect scores, we used the recommended implementation, which converts each output to quantiles and assigns each variant the maximum percentile observed across all outputs. Taking the maximum percentile across all outputs—whether or not they are relevant to a given variant type—likely introduced noise, as some outputs are more informative than others for predicting pathogenicity.

For splice variants, we also considered the splice-variant scorer defined in the original publication, which takes a weighted sum of raw scores from three splice-related modalities (splice site usage, splice site, and splice junction). We further tested scores based on each of these modalities individually, maximizing either raw scores or quantiles (Figure 3a). We found that the variant scorer reported in the original publication performs better (AUROC $= 0.708$) than the composite score based on all modalities (AUROC $= 0.614$), but that an even simpler scorer based solely on raw splice site usage performs best (AUROC $= 0.779$). Further work is needed to establish the optimal variant effect prediction strategy for AlphaGenome across different variant types.

3 RESULTS

We curated a benchmark comprising all single-nucleotide variants (SNVs) in ClinVar Landrum et al. (2016) labeled as definitively pathogenic or benign, grouped by the types of genomic regions they affect. We benchmarked six DNA-sequence AI models (Evo2 Brixi et al. (2025), AlphaGenome Avsec

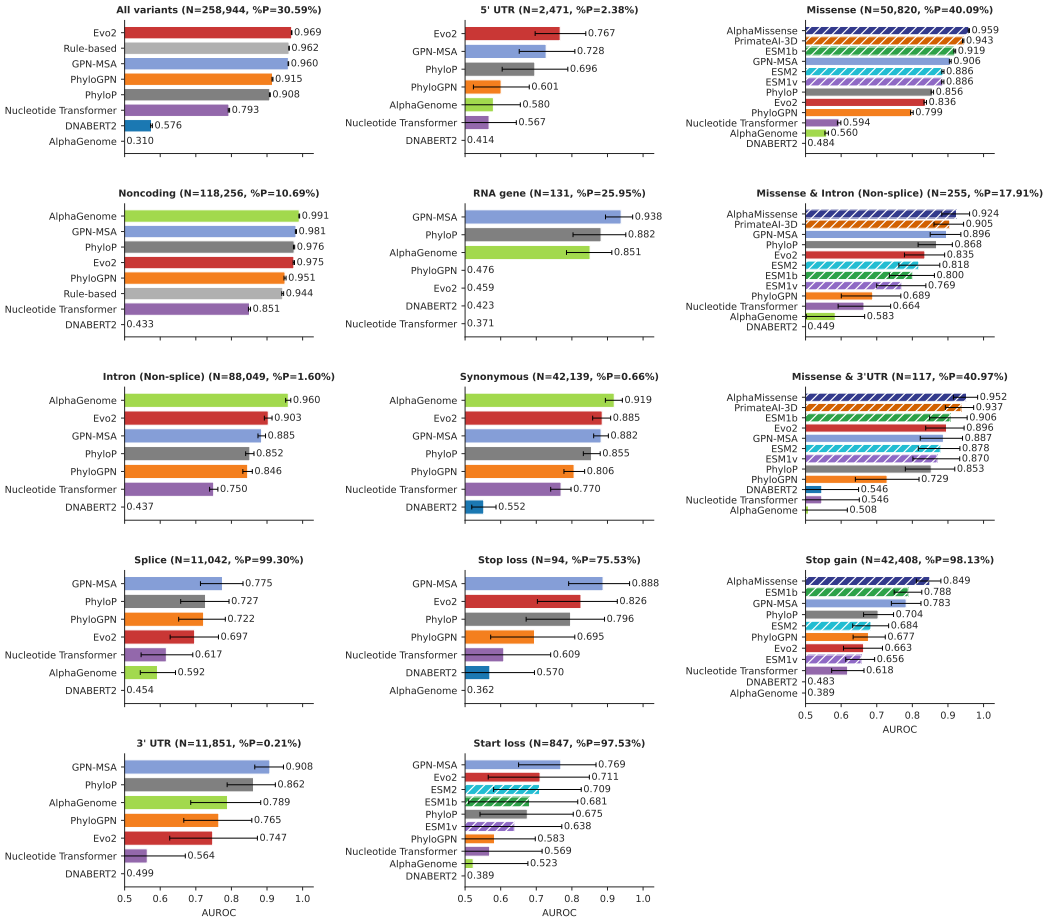


Figure 2: Pathogenicity prediction performance of frontier sequence-based models across variant types.

Note: Evaluation and comparison of DNA and protein sequence AI models for their capacity to distinguish between pathogenic and benign variants across variant types were measured using the area under the receiver operating characteristic curve (AUROC). Error bars denote 95% confidence intervals estimated by stratified bootstrap resampling (1,000 iterations) within each variant group. The percentage of pathogenic variants in each group (%P) indicates the prevalence of pathogenicity. Some variant groups were defined by multiple annotated effects (e.g., both missense and 3′ UTR, depending on transcript annotation). DNA models are shown as solid bars, and protein models as dashed bars. *The evaluation of PrimateAI-3D on stop-gain variants includes only 19,795 variants (see Methods).

et al. (2025), GPN-MSA Benegas et al. (2025), PhyloGPN Albors et al. (2025), Nucleotide Transformer Dalla-Torre et al. (2024), and DNABERT2 Zhou et al. (2023)) and five protein-sequence AI models (ESM-1b Rives et al. (2021); Brandes et al. (2023), ESM-1v Meier et al. (2021), ESM-2 Lin et al. (2023), AlphaMissense Cheng et al. (2023), and PrimateAI-3D Gao et al. (2023)) (Table 1). We also included PhyloP Pollard et al. (2010), a simple and widely used phylogenetic model of conservation scores per genomic position, as a baseline (Figure 2).

All DNA models showed dramatically reduced performance when evaluated within specific variant types. For example, Evo2 appears to perform very well across all noncoding variants (AUROC $= 0.975$), but drops sharply over specific types of noncoding effects (e.g., 0.903 for non-splice intron variants, 0.697 for canonical splice sites). To illustrate the impact of heterogeneous variant groups, we examined the distributions of Evo2 scores across canonical splice and 5′ UTR variants (Figure 3b). The distribution of Evo2 scores for 5′ UTR variants is shifted towards less damaging

prediction for both pathogenic and benign variants, consistent with their much lower prevalence of pathogenicity compared to splice variants (99.3% vs. 2.4%). As a result, merging these two subgroups artificially inflates performance, because pathogenic variants are disproportionately splice variants that receive more damaging scores.

Overall, the strong performance of most models observed in the composite groups (all variants and noncoding variants) primarily reflects variant-type heterogeneity. To account for this, we constructed a rule-based baseline assigning each variant a score equal to the frequency of pathogenic variants within its type. This information alone yields strong separation between clinical labels across all variants (AUROC $= 0.962$) and noncoding variants (AUROC $= 0.944$; Figure 2).

We also observe substantial performance variability across different types of noncoding variants, with no single model consistently outperforming the others (Figure 2). For non-splice intronic variants, which comprise 74% of noncoding variants in our benchmark, AlphaGenome performed best (AUROC $= 0.960$), followed by Evo2 (0.903) and GPN-MSA (0.885), both of which also exceeded PhyloP (0.852). For canonical splice variants, on the other hand, performance dropped across all models, with only GPN-MSA (0.775) surpassing PhyloP (0.727), while AlphaGenome dropped to 0.592. In $3'$ UTR variants, only GPN-MSA (0.908) outperformed PhyloP (0.862). For $5'$ UTR variants, Evo2 (0.767) and GPN-MSA (0.728) surpassed PhyloP (0.696). For RNA gene variants, GPN-MSA (0.938) again exceeded PhyloP (0.882), but this result is less reliable due to the small sample size ($n = 131$, of which 34 are pathogenic).

When evaluating performance over coding variants, we also considered protein models, which generally outperformed DNA models (Figure 2). For missense variants, the best performers were AlphaMissense (AUROC $= 0.959$) and PrimateAI-3D (0.943), followed by ESM-1b (0.919) and other ESM models. DNA models lagged behind, with GPN-MSA (0.906) coming closest. Interestingly, for variants that are missense in one transcript but noncoding in another ($3'$ UTR or non-splice intronic regions), protein models showed reduced performance, bringing DNA models closer, though still slightly behind.

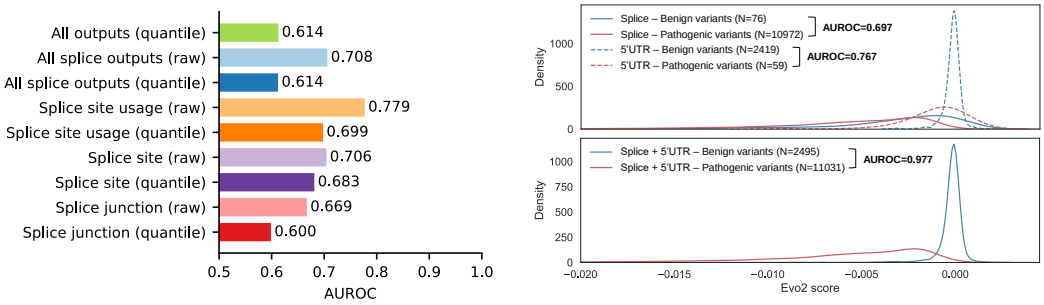
For coding variants that are likely to involve DNA- or transcript-level effects beyond amino-acid changes, such as start-loss and stop-gain variants, both DNA and protein models performed worse relative to missense variants. Stop-gain variants were scored by protein models by taking the missense score predicted as most deleterious between the mutation site and the end of the protein (see Methods) Brandes et al. (2023). This approach worked surprisingly well: AlphaMissense (0.849), PrimateAI-3D (0.817), and ESM-1b (0.788) all outperformed GPN-MSA (0.783) and other DNA models.

Start-loss variants were treated similarly to missense variants, leading to GPN-MSA (0.769) and Evo2 (0.711) slightly outperforming the ESM models (0.709), although this result is based on only 21 benign variants. Only DNA models were applicable to synonymous and stop-loss variants. For synonymous variants, AlphaGenome performed best (AUROC $= 0.919$), followed by Evo2 (0.885) and GPN-MSA (0.882), both exceeding PhyloP (0.855). For stop-loss variants ($n = 71$ pathogenic, $n = 23$ benign), GPN-MSA (0.888) and Evo2 (0.826) also surpassed PhyloP (0.796).

4 Discussion

The metric we used (AUROC) allows direct comparison across variant types despite widely different rates of pathogenicity, ensuring that the observed performance differences can be attributed to model capabilities (see Methods). Our analysis reveals substantial variation in the performance of all models across variant types. No single model was universally optimal, although GPN-MSA and AlphaMissense emerged as the most robust DNA and protein models, respectively, consistently surpassing PhyloP across all variant types. Evo2 and AlphaGenome also led in some contexts but showed marked drops for other variant types. Notably, AlphaGenome is highly sensitive to the specific method used to extract scores from the model, which may be partly responsible for its unstable performance (see Methods). These results suggest a promising meta-prediction approach, where variants are first stratified by type, followed by prediction using the best-performing model within each group.

Our analysis also reveals that combining heterogeneous variant types inflates overall performance, as models can easily infer and exploit variant-type pathogenicity priors, even with limited predictive



(a) Performance of different AlphaGenome outputs on splice variants.

(b) Merging splice and 5' UTR variants artificially improves the separation between clinical labels achieved by Evo2.

Figure 3: Splice-variant performance and Evo2 score distributions.

Note: (a) Prediction performance (AUROC) over splice variants (N=11,047, %P=99.32%) using different AlphaGenome outputs. In addition to the composite score integrating all AlphaGenome outputs (as in Figure 2), we also considered specific splicing-related outputs (splice site usage, splice site, and splice junction) and a composite score integrating them, using either raw scores or quantiles (see Methods). (b) Distributions of Evo2 scores across pathogenic and benign variants and across splice and 5' UTR variants (top) and the union of these two groups (bottom).

power within subgroups. These findings underscore the necessity of variant-type-specific benchmarking and of choosing models with respect to specific variant types. To help future evaluations meet this standard, we have released our benchmarks and code (see Data and Code Availability).

For certain variant types, the state of the art is already highly reliable (AUROC > 0.9). This includes missense (AlphaMissense), synonymous (AlphaGenome), non-splice intron (AlphaGenome), $3'$ UTR (GPN-MSA), and RNA gene variants (GPN-MSA). By contrast, stop-gain, start-loss, stop-loss, splice, and $5'$ UTR variants remain difficult for current sequence-based models.

The superior performance of GPN-MSA, AlphaMissense, and PrimateAI-3D likely stems from their use of multiple sequence alignment (MSA) inputs, which provide rich evolutionary information. However, reliance on sequence homology makes these models difficult to run locally and limits their generality beyond point mutations. Language models, in contrast, tend to be more general and are also applicable to indels Brandes et al. (2023). PhyloGPN takes an interesting middle path, using MSA during training but sequence only during inference. However, it consistently underperforms both the PhyloP baseline and its relative method GPN-MSA. The authors of PhyloGPN note that it relies on a relatively simple evolutionary model (F81) Felsenstein (1981), which may partly explain these results.

A common concern is the interaction between model pretraining and evaluation benchmarks derived from ClinVar Landrum et al. (2016). We emphasize that while these models are trained on cross-species sequences, with some incorporating signals like allele frequencies (e.g., AlphaMissense, PrimateAI-3D) or functional genomics (e.g., AlphaGenome), none utilize ClinVar pathogenicity labels, ensuring no direct data leakage. However, pretraining inevitably encodes coarse priors, such as evolutionary constraint, which can inflate performance when distinct variant types are aggregated. Herein lies the distinct value of our study: we stratify evaluation by genomic heterogeneity. This approach isolates within-type discriminative ability, ensuring that performance metrics reflect genuine pathogenicity prediction rather than the exploitation of broad, pre-learned priors.

This study has several limitations. First, we restricted our analysis to single-nucleotide variants (SNVs). Second, our evaluation was constrained by the scope of ClinVar, which lacks definitive clinical labels for certain variant types, such as those in promoters or trans-regulatory elements. Third, our analysis of stop-gain variants did not account for nonsense-mediated decay (NMD). Specifically, protein models are expected to perform better on variants not expected to result in NMD (e.g., according to the "50 bp rule") Brandes et al. (2023), while DNA models may hold greater potential to capture NMD. However, this likely requires context spanning multiple exons, whereas most DNA models have a relatively short context length (e.g., 8,192 bp for Evo2; Table 1).

5 Data and Code Availability

Our full benchmark, which includes the ClinVar labels, genomic annotations, and model scores of each variant, is provided as (Figure 1). Our source code, which includes the calculation and extraction of predictions from each of the evaluated models, performance calculations (AUROC), and the generation of presented plots, is available as an open-source project at `https://github.com/Brandes-Lab/VEP-eval`.

References

C. Albors, J. C. Li, G. Benegas, C. Ye, and Y. S. Song. A dna language model based on phylogenetic generalization. In *Proceedings of the International Conference on Research in Computational Molecular Biology*, pp. 99–117. Springer, 2025.

Ž. Avsec et al. Alphagenome: advancing regulatory variant effect prediction with a unified dna sequence model. *bioRxiv*, 2025. doi: 10.1101/2025.06.25.661532. URL https://doi.org/10.1101/2025.06.25.661532.

G. Benegas, C. Albors, A. J. Aw, C. Ye, and Y. S. Song. A dna language model based on multispecies alignment predicts the effects of genome-wide variants. *Nature Biotechnology*, 2025. doi: 10.1038/s41587-024-02511-w. URL https://doi.org/10.1038/s41587-024-02511-w.

N. Brandes, G. Goldman, C. H. Wang, C. J. Ye, and V. Ntranos. Genome-wide prediction of disease variant effects with a deep protein language model. *Nature Genetics*, 55:1512–1522, 2023. doi: 10.1038/s41588-023-01465-0. URL https://doi.org/10.1038/s41588-023-01465-0.

G. Brixi et al. Genome modeling and design across all domains of life with evo 2. *bioRxiv*, 2025. doi: 10.1101/2025.02.18.638918. URL https://doi.org/10.1101/2025.02.18.638918.

J. Cheng et al. Accurate proteome-wide missense variant effect prediction with alphamissense. *Science*, 381:eadg7492, 2023. doi: 10.1126/science.adg7492. URL https://doi.org/10.1126/science.adg7492.

H. Dalla-Torre et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 2024. doi: 10.1038/s41592-024-02523-z. URL https://doi.org/10.1038/s41592-024-02523-z.

J. M. Ellingford et al. Recommendations for clinical interpretation of variants found in non-coding regions of the genome. *Genome Medicine*, 14:73, 2022. doi: 10.1186/s13073-022-01073-3. URL https://doi.org/10.1186/s13073-022-01073-3.

J. Felsenstein. Evolutionary trees from dna sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981. doi: 10.1007/BF01734359. URL https://doi.org/10.1007/BF01734359.

H. Gao et al. The landscape of tolerated genetic variation in humans and primates. *Science*, 380, 2023. doi: 10.1126/science.abn8197. URL https://doi.org/10.1126/science.abn8197.

W. J. Kent et al. The human genome browser at ucsc. *Genome Research*, 12:996–1006, 2002. doi: 10.1101/gr.229102. URL https://doi.org/10.1101/gr.229102.

M. J. Landrum et al. Clinvar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 44:D862–D868, 2016. doi: 10.1093/nar/gkv1222. URL https://doi.org/10.1093/nar/gkv1222.

Z. Lin et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379:1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://doi.org/10.1126/science.ade2574.

J. Meier et al. Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv*, 2021. doi: 10.1101/2021.07.09.450648. URL https://doi.org/10.1101/2021.07.09.450648.

C. Montano et al. Diagnosis and discovery: Insights from the nih undiagnosed diseases program. *Journal of Inherited Metabolic Disease*, 45:907–918, 2022. doi: 10.1002/jimd.12506. URL https://doi.org/10.1002/jimd.12506.

J. Morales et al. A joint ncbi and embl-ebi transcript set for clinical genomics and research. *Nature*, 604:310–315, 2022. doi: 10.1038/s41586-022-04558-8. URL https://doi.org/10.1038/s41586-022-04558-8.

E. Nguyen et al. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386:eado9336, 2024. doi: 10.1126/science.ado9336. URL https://doi.org/10.1126/science.ado9336.

N. A. O'Leary et al. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44:D733–745, 2016. doi: 10.1093/nar/gkv1189. URL https://doi.org/10.1093/nar/gkv1189.

K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, and A. Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20:110–121, 2010. doi: 10.1101/gr.097857.109. URL https://doi.org/10.1101/gr.097857.109.

A. Rives et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences of the USA*, 118, 2021. doi: 10.1073/pnas.2016239118. URL https://doi.org/10.1073/pnas.2016239118.

The Lancet Global, H. The landscape for rare diseases in 2024. *The Lancet Global Health*, 12: e341, 2024. doi: 10.1016/S2214-109X(24)00056-1. URL https://doi.org/10.1016/S2214-109X(24)00056-1.

The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 2024. doi: 10.1093/nar/gkae1010. URL https://doi.org/10.1093/nar/gkae1010.

C. Wu, I. Macleod, and A. I. Su. Biogps and mygene.info: organizing online, gene-centric information. *Nucleic Acids Research*, 41:D561–565, 2013. doi: 10.1093/nar/gks1114. URL https://doi.org/10.1093/nar/gks1114.

Z. Zhou et al. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*, 2023.