

A Survey of Machine Learning Approaches for Genomic Variant Effect Prediction, with a Proposed Tri-Modal Framework for Variants of Uncertain Significance

Aakanksha A Dutt (1RF23IS001), Inchara B (1RF23IS039), Jayati Gupta (1RF23IS043), Malepati Tanay (1RF23IS052)

Department of Information Science and Engineering, RV Institute of Technology and Management, Bengaluru, Karnataka, India Affiliated to
Visvesvaraya Technological University, Belagavi

Guide: Prof. Sunita J, Department of Information Science and Engineering, RVITM

Abstract - Determining whether a genetic variant is harmful is one of the central open problems in clinical genomics. Sequencing throughput has grown far faster than the human expertise available to interpret its output, and the result is a growing backlog of variants that carry no actionable clinical label. Roughly two out of every five variants reported by clinical diagnostic tests in protein-coding regions are recorded as Variants of Uncertain Significance, and only a small fraction of these are ever reclassified. This survey reviews the machine learning methods developed for variant effect prediction over the past decade, organised into six families: alignment and conservation scores, feature-engineered ensemble classifiers, protein language models, DNA language models, graph neural network approaches, and structure-informed or phenotype-conditioned predictors. For each family we examine the representation used, the supervision signal, the treatment of class imbalance, the evaluation protocol, and the computational cost. Three recurring weaknesses emerge. First, most published systems are formulated as binary classifiers over confidently labelled variants and discard the uncertain class entirely, which removes the majority of clinically difficult cases from evaluation. Second, random row-level train and test splitting remains common, permitting gene-level memorisation and inflating reported accuracy. Third, the highest-performing recent systems depend on very large foundation models or on multi-source biomedical knowledge graphs whose compute requirements place them out of reach of ordinary laboratory hardware. We then present GenoHybrid-VEP, a proposed framework designed against these findings. It couples a one-dimensional convolutional encoder over a 101 base-pair genomic context window, a dense encoder over clinical and population biomarkers, and a graph convolutional encoder over variant co-occurrence relations, fusing the three into a single latent representation for three-way classification. Variants of Uncertain Significance are retained as a first-class label, training uses gene-grouped splitting with balanced class weighting, and epistemic uncertainty is quantified through Monte Carlo dropout so that low-confidence predictions can be flagged rather than silently reported. The framework is described here as an architectural proposal; empirical validation is reserved for subsequent work.

Keywords: variant effect prediction, variants of uncertain significance, ClinVar, multimodal deep learning, late fusion, graph convolutional networks, class imbalance, epistemic uncertainty, Monte Carlo dropout, clinical genomics

1. INTRODUCTION

A human genome differs from the reference assembly at several million positions. The overwhelming majority of these differences are harmless. A small number are not, and identifying which ones matter is the task that stands between a sequencing report and a clinical decision. This is the variant effect prediction problem.

The problem is not new, but its shape has changed. When sequencing was expensive, variant interpretation was a bottleneck of cost. Now that sequencing is cheap, interpretation is a bottleneck of expertise. Curated databases such as ClinVar aggregate assertions submitted by clinical laboratories worldwide, and while this has produced an invaluable public resource, it has also made the scale of the unresolved backlog visible. A large share of ClinVar entries carry the label Variant of Uncertain Significance, meaning the submitting laboratory could find neither sufficient evidence for pathogenicity nor sufficient evidence for benignity. These variants are not a rounding error in the dataset. They are its dominant class, and they are precisely the cases where a clinician needs help.

Computational prediction has been proposed as a partial answer since the early 2000s. The methods have evolved through several distinct phases, summarised in Figure 1. Early tools scored variants by evolutionary conservation. Later systems replaced hand-built conservation scores with supervised classifiers trained on dozens of annotated features. More recently, self-supervised language models trained on protein or DNA sequence have been used to produce learned representations, and graph neural networks have been used to inject biological relational structure that pure sequence models cannot see.

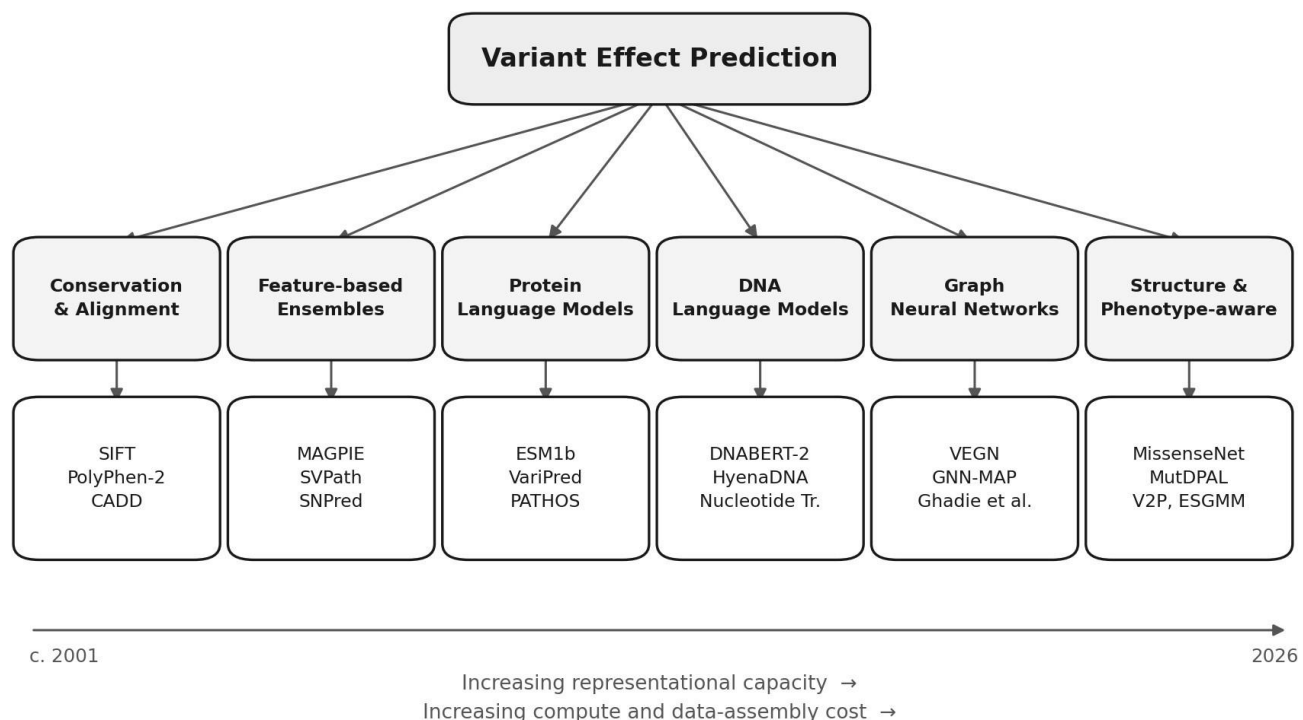


Figure 1. Taxonomy of variant effect prediction methods, with representative systems in each family. Capability has increased from left to right, but so has the compute and data-assembly cost required to run the method.

Progress has been real. Reported accuracies on benchmark splits have climbed steadily. But there is a gap between reported accuracy and clinical usefulness, and this survey argues that the gap has three specific causes.

The first is task formulation. A very large share of published work reduces the problem to a two-class decision between pathogenic and benign, using only variants with confident labels and discarding everything uncertain. This is methodologically convenient and clinically hollow. A model that never sees an uncertain variant during training has nothing useful to say when one arrives, and a benchmark that excludes them measures performance on the subset of the problem that was never difficult.

The second is evaluation protocol. Variants cluster heavily by gene. A random split across rows will place variants from the same gene in both the training and the test partition, and a model can then score well by learning that variants in a given gene tend to be pathogenic, without learning anything about variant mechanics at all. Reported numbers under random splitting are therefore not comparable to numbers under gene-grouped splitting, and the literature is not consistent about which is used.

The third is deployment feasibility. The best-performing recent systems require multiple GPUs with tens of gigabytes of memory, terabyte-scale annotation databases, or knowledge graphs assembled from a dozen separate biomedical sources. For a large well-funded genomics centre this is acceptable. For a hospital laboratory or a university research group it is not, and a method that cannot be run is not a method that can be adopted.

This paper is organised as follows. Section 2 states the problem formally and describes the data sources and evaluation metrics used across the field. Section 3 surveys the six method families. Section 4 compares them and consolidates the research gaps. Section 5 presents GenoHybrid-VEP. Section 6 outlines the evaluation protocol we intend to apply. Section 7 concludes.

2. PROBLEM FORMULATION AND BACKGROUND

2.1 The prediction task

A variant is described by a chromosome, a genomic position on a stated reference assembly, a reference allele, and an alternate allele. Given this description together with any available annotation, the task is to assign a clinical significance label.

The binary formulation defines the label space as {Benign, Pathogenic}. The three-class formulation used in this work

defines it as {Benign, Pathogenic, Variant of Uncertain Significance}. The distinction is not cosmetic. In the binary case the decision boundary separates two categories that are, by construction, well separated in the evidence available to the submitting laboratories. In the three-class case the model must additionally learn to recognise the region of feature space in which evidence is insufficient, which is a fundamentally different and harder learning problem.

A third formulation, disease-specific prediction, treats pathogenicity as a property of a variant and disease pair rather than of the variant alone. This is

clinically more precise, since a variant may be causative for one condition and irrelevant to another, but it requires that the disease of interest already exists in the model's knowledge base and therefore cannot be applied to variants in genes with no established disease association.

2.2 Data sources

Table 1 summarises the resources used across the surveyed literature and indicates which of them the proposed framework consumes.

Table 1. Data resources for variant effect prediction.

Resource	Content	Role in interpretation	Used in this work
ClinVar [3]	Clinical assertions, review status, submitter counts, variant origin, gene	Primary source of supervised labels and clinical metadata	Yes, labels and biomarkers
gnomAD [37]	Population allele frequencies across large reference cohorts	High frequency in healthy populations is strong evidence for benignity	Yes, allele frequency
PhyloP / PhastCons / GERP	Per-position evolutionary conservation scores	Conserved positions are more likely to be functionally constrained	Yes, PhyloP and PhastCons
GRCh38 reference assembly	Reference nucleotide sequence	Provides local sequence context around the variant position	Yes, 101 bp window
Ensembl VEP [38]	Predicted molecular consequence of a variant	Distinguishes missense, synonymous, splice-region and other classes	Yes, categorical feature
PDB / AlphaFold	Experimental and predicted protein structures	Enables reasoning about folding, burial and interaction interfaces	No, uneven coverage
Deep mutational scanning	Direct functional measurements per variant	Ground truth independent of clinical assertion bias	No, limited gene coverage

The ClinVar review status deserves particular attention because it is ordinal rather than categorical. It ranges from unreviewed submissions through single-submitter and multiple-submitter criteria-based assertions up to expert panel review and practice guideline status, and this ordering carries real information about assertion reliability. Several of the surveyed methods use it as a feature; several others discard it.

2.3 Evaluation metrics

Accuracy alone is misleading under the class imbalance characteristic of this domain, because a classifier that predicts the majority class for every input can report a high figure while being clinically worthless. The field therefore reports balanced accuracy, macro-averaged F1, area under the receiver operating characteristic curve, and Matthews correlation coefficient. Sensitivity and negative predictive value carry particular weight in a clinical screening context, since the cost of missing a pathogenic variant is asymmetric with the cost of a false alarm.

Calibration is reported far less often than it should be. A model that outputs a confidence of 0.9 should be correct approximately ninety percent of the time when it does so. Expected calibration error measures the deviation from this property, and in a domain where a prediction feeds into a clinical decision, an uncalibrated confidence score is arguably worse than no confidence score at all.

3. LITERATURE SURVEY

Table 2 consolidates the surveyed work before the discussion that follows.

Table 2. Summary of surveyed methods.

Ref	Method	Year	Core technique	Data	Reported performance	Key limitation
[6]	PolyPhen-2	2013	Bayesian classifier over alignment and structural features	Alignments, PDB	Widely benchmarked	Continuous score; user must set threshold
[5]	CADD	2019	Classifier separating simulated from surviving variants	Genome-wide	Widely benchmarked	No clinical context

[8]	SVPath	2022	Gradient boosted trees	Exon structural variants	Not directly comparable	Annotation-pipeline bound
[7]	MAGPIE	2024	Supervised ML over multi-category features	Multiple variant types	Not directly comparable	Binary formulation
[9]	SNPred	2023	Ensemble of SNV predictors	ClinVar SNVs	Competitive on SNVs	Documents ClinVar evaluation difficulty
[10]	Brandes et al. (ESM1b)	2023	Protein language model, genome-wide scoring	UniProt, ClinVar	Genome-wide coverage	Missense only
[11]	VariPred	2024	PLM log-likelihood plus residue embeddings	ClinVar	MCC 0.714, AUC-ROC 0.928	Missense only
[12]	PATHOS	2026	ESM-C and Ankh2 embeddings plus biological features	Clinical and ClinVar sets	MCC 0.591 clinical, 0.826 ClinVar	Missense only; no gain from fine-tuning
[16]	Sayed et al.	2024	Benchmark of genomic foundation models	SNVs	Benchmark study	Finds FMs do not clearly outperform
[17]	VEGN	2021	GNN over gene and variant graph	ClinVar	Not directly comparable	Binary formulation
[18]	Zhang et al.	2022	Graph attention with co-evolution edge weights	Missense variants	Competitive	Missense only
[22]	GNN-MAP	2025	GNN over variant similarity graph, six annotation modalities	ClinVar, orthogonal sets	Strong on rare and imbalanced sets	Binary formulation
[2]	Ghadie et al. (base paper)	2025	BioBERT and DNA-LM embeddings, GCN plus NN decoder over 12-type knowledge graph	ClinVar missense, 42,335 variants	85.6% balanced accuracy, 90.5% sensitivity, 89.8% NPV	Disease-specific only; VUS excluded; multi-GPU
[25]	MissenseNet	2024	ShuffleNet backbone with SE module and AlphaFold2 features	gnomAD, ClinVar	91.82% accuracy, AUC 0.9701	Depends on AlphaFold2 accuracy; weak on disordered regions
[26]	MutDPAL	2025	Two Bio-LLMs, cross-attention disease-protein association	Transmembrane proteins	15-category classification	Domain restricted to membrane proteins
[27]	V2P	2025	Multi-task model conditioned on HPO phenotype classes	SNVs and indels, genome-wide	Outperformed compared methods	Requires phenotype ontology mapping
[28]	ESGMM	2024	MSA Transformer with global-local Gaussian mixture, PSO	2,715 clinical proteins	85.7 to 96.6% ACC across four proteins	Unsupervised binary; excludes uncertain cases
[29]	Bromberg et al.	2024	Review of ML methods for variant effect prediction	Survey	Survey	Notes no standard definition of variant effect

3.1 Alignment and conservation based scores

The first generation of predictors reasoned from evolutionary evidence. SIFT scores a substitution by how well the alternate residue is tolerated at that position across a multiple sequence alignment of homologous proteins. PolyPhen-2 combines sequence alignment evidence with a small number of structural and annotation features under a Bayesian classifier [6]. CADD takes a different route, training a classifier to separate simulated de novo variants from variants that have survived in the human lineage, and producing a single integrated deleteriousness score [5].

These tools remain in routine clinical use and are named in the interpretation guidelines issued by the American College of Medical Genetics [39]. Their durability rests on three properties: they are fast, they are interpretable in terms a geneticist recognises, and they require no training data specific to the gene under examination.

Their limitations are equally clear. They operate on evolutionary signal alone and are blind to clinical context, so they cannot distinguish a germline variant from a somatic one, and they cannot weigh the quality of the clinical evidence already accumulated for a variant. They also produce a continuous score rather than a class, which pushes the burden of threshold selection onto the user, and thresholds that work in one gene often fail in another.

3.2 Feature-engineered ensemble classifiers

The second generation replaced hand-designed scores with supervised learning over large engineered feature vectors. MAGPIE applies a machine learning approach across multiple variant types using features spanning epigenomics, functional effect, splicing effect, population frequency, biochemical properties and conservation [7]. SVPath applies gradient boosted trees to exon structural variants after annotation with genomic, proteomic and epigenomic features [8]. SNPred is an ensemble predictor for single nucleotide variants which, in addition to reporting competitive performance, documents the difficulties involved in using ClinVar itself as an evaluation standard [9].

The strength of this family is that tree ensembles handle heterogeneous tabular features well, tolerate missing values

gracefully, and offer feature importance measures that support interpretation. Training is fast and the compute requirement is modest.

The weaknesses are structural. Tree models treat the reference and alternate alleles as independent categorical splits and therefore cannot represent the substitution as a single relational object, nor can they see the sequence context surrounding the variant position. Performance is also bounded by the quality and completeness of the annotation pipeline, which is where most of the actual engineering effort in these systems goes.

3.3 Protein language models

Protein language models trained by self-supervision on very large sequence corpora learn representations that encode structural and functional constraints without being told about them. Brandes et al. demonstrated genome-wide variant effect prediction from a deep protein language model, showing that the likelihood a model assigns to an amino acid at a position correlates with functional tolerance at that position [10]. VariPred showed that combining log-likelihood ratios with residue embeddings from a protein language model is competitive without any additional biological features, reporting a Matthews correlation coefficient of 0.714 and an area under the receiver operating characteristic curve of 0.928 on a ClinVar test set [11]. PATHOS extends this direction, integrating embeddings from ESM-C and Ankh2 with conventional features such as allele frequency and phylogenetically derived mutation probabilities, and reports a Matthews correlation coefficient of 0.591 on a curated clinical dataset while outperforming sixty-five competing predictors [12]. Notably, PATHOS reports that fine-tuning the protein language model did not improve performance over the original embeddings, which is a useful counterweight to the assumption that task-specific fine-tuning is always worth its cost.

The appeal of this family is that a single pretrained model generalises across all proteins, including those with few characterised homologs, and that the representation is learned rather than designed. The costs are model size, which for the larger checkpoints runs to hundreds of millions or billions of parameters, and the opacity of the resulting representation. There is also a scope restriction that is easy to overlook: protein language models operate on protein sequence and therefore apply only to coding variants that change a residue. Synonymous, intronic, splice-region and regulatory variants are outside their domain entirely.

3.4 DNA language models

DNA language models apply the same self-supervised paradigm directly to nucleotide sequence, removing the coding-region restriction. DNABERT-2 provides an efficient multi-species foundation model [13]. HyenaDNA achieves single-nucleotide resolution over context windows extending to a million tokens [14]. The Nucleotide Transformer offers models pretrained across hundreds of genomes from a wide range of species [15].

The evidence on whether this helps for variant pathogenicity specifically is more equivocal than the general enthusiasm for foundation models would suggest. Sayeed et al. benchmarked genomic foundation models on single nucleotide variant pathogenicity and found that they did not necessarily outperform simpler alternatives on this particular task [16]. Ghadie et al. report a consistent finding: in their experiments, embedding a variant's type, location and alleles through a general biomedical text model achieved performance comparable to embeddings drawn from dedicated DNA language models, and a neural network operating on raw DNA language model embeddings alone was unable to predict non-disease-specific pathogenicity at better than chance level [2]. They also observe that fine-tuning the DNA language model on the pathogenicity task specifically produced a substantial improvement, which suggests the pretraining objective and the clinical objective are less aligned than assumed.

Context window size emerges as a subtle design variable. A larger window can in principle capture long-range regulatory context, but it also admits noise, and Ghadie et al. report that reducing context length was important for successful fine-tuning. Longer is not automatically better, a finding that directly informs the window choice in Section 5.3.

3.5 Graph neural network approaches

Graph methods encode the observation that variants are not independent. They occur in genes, genes participate in pathways, pathways connect to phenotypes, and pathogenicity propagates along these relations.

VEGN applies graph neural networks over a joint graph of genes and variants [17]. Zhang et al. use a graph attention network in which nodes carry amino acid predictive features and edges are weighted by feature co-evolution strength [18]. Wang et al. map atomic connectivity in three-dimensional protein structures into residue-level network representations with structural, topological and biophysical node attributes [19]. Kamada et al. work at the pathway level to address variants of uncertain significance specifically [20]. Hou combines protein sequence representations from an Evolutionary Scale Modeling backbone with a graph neural network for nonsynonymous mutations [21].

GNN-MAP is the closest published analogue to the graph component proposed here. It constructs a graph based on similarity between variants rather than on an external biomedical ontology, integrates multimodal annotations spanning six categories, and reports strong performance across multiple variant types including rare variants and highly imbalanced datasets [22]. The similarity-graph construction is significant because it avoids the dependency on curated biomedical knowledge bases that limits the coverage of ontology-based approaches.

The most comprehensive recent instance, and the base reference for the present work, is Ghadie et al. [2]. They construct a heterogeneous biomedical knowledge graph containing more than 129,000 nodes across twelve entity types including genes, proteins, variants, diseases, drugs, phenotypes, pathways, molecular functions, biological processes, cellular components, exposures and anatomy, connected by more than eight million edges across thirty relation types [23]. Node features are initialised from a biomedical language model applied to textual descriptions [24], variant nodes additionally receive DNA language model embeddings, and a two-layer graph convolutional network feeds a three-layer neural decoder that predicts pathogenicity for a variant and disease pair. Reported balanced accuracy reaches approximately 85.6 percent with fine-tuned variant embeddings, at a sensitivity of 90.5 percent and negative predictive value of 89.8 percent.

Two of their ablation findings are of direct methodological interest. First, pruning the full twelve-type graph down to five node types, namely variants, genes, proteins, pathways and diseases, left performance essentially unchanged. The higher-level clinical and environmental entities contributed little, which implies that most of the achievable relational signal lives in a much smaller and much cheaper graph. Second, balancing benign and pathogenic variants within each gene during training reduced reported accuracy, because it removed the model's ability to exploit the skewed within-gene label distribution. The lower number is the more honest one.

The limitations they acknowledge are equally instructive. Their approach is disease-specific and therefore cannot handle variants in genes with no established disease link. Their evaluation is restricted to missense variants labelled confidently as benign or pathogenic, with all uncertain-significance variants and all conflicting-interpretation variants discarded. And the pipeline requires an assembled multi-source knowledge graph plus GPU resources at the level of an A100 with forty gigabytes of memory, rising to four such GPUs for the fine-tuning stage.

3.6 Structure-informed, phenotype-conditioned and assay-grounded methods

MissenseNet integrates AlphaFold2-derived structural features with a ShuffleNet-based backbone augmented by an encoder-decoder structure and a squeeze-and-excitation module, reporting an accuracy of 91.82 percent and an area under the receiver operating characteristic curve of 0.9701 on its evaluation set [25]. Its own conclusion notes the dependency this creates: the structural features are only as reliable as the AlphaFold2 predictions behind them, and intrinsically disordered regions remain a weak point.

MutDPAL takes a domain-specialisation approach, using two pretrained biological large language models to predict pathogenicity in transmembrane proteins and further classify pathogenic mutations into fifteen disease categories through cross-attention between protein and disease representations [26]. Performance within the transmembrane domain is strong, at the cost of generality across the proteome.

V2P conditions its predictions on top-level Human Phenotype Ontology categories, producing a multi-task multi-output model that predicts pathogenicity for single nucleotide variants and small insertions and deletions across the genome while simultaneously predicting the disease phenotype class [27]. This is a different answer to the same problem Ghadie et al. address through disease-specific edges, and it has the advantage of covering variant types beyond missense.

ESGMM combines a Transformer-based protein language model over multiple sequence alignments with a global-local Gaussian mixture model, and is notable for treating uncertainty explicitly through predictive entropy [28]. Its authors report that excluding high-uncertainty variants raises accuracy by between seven and twelve percentage points across four clinical proteins, which quantifies something the rest of the field mostly leaves implicit: a large share of residual error is concentrated in cases the model itself could flag as unreliable if asked to. This finding is a direct motivation for the uncertainty quantification described in Section 5.6.

On the assay side, deep mutational scanning provides functional measurements independent of clinical assertion bias. The constraint is economic rather than technical: generating such data is expensive, coverage is limited to a small number of genes, and reuse across genes is limited.

Broad reviews converge on a consistent conclusion. Bromberg et al., surveying machine learning methods for variant effect prediction, find that unsupervised deep learning methods now perform as well as or better than established supervised ones on some tasks, but that performance varies substantially depending on the evaluation metric and the type of variant effect being predicted, and that the field still lacks a standardised definition of what variant effect even means [29]. That last point deserves emphasis, because it means cross-paper comparisons are less meaningful than their tables suggest.

4. COMPARATIVE ANALYSIS AND RESEARCH GAPS

4.1 Comparison across method families

Table 3 compares the six families along the dimensions that matter for adoption.

Table 3. Comparison of variant effect prediction method families.

Family	Representation	Handles VUS	Coverage	Typical compute	Principal weakness
Conservation and alignment	Evolutionary scores	No	Genome-wide	Negligible	No clinical context; user must set thresholds
Feature-engineered ensembles	Annotated tabular features	Rarely	Depends on annotation pipeline	Low, CPU sufficient	Cannot represent sequence context; annotation-bound
Protein language models	Learned residue embeddings	No	Coding, missense only	High, large GPU	Excludes non-coding; opaque; heavy
DNA language models	Learned nucleotide embeddings	No	Genome-wide	High, large GPU	Weak transfer to pathogenicity without fine-tuning
Graph neural networks	Relational node embeddings	No	Limited to graph-covered genes	High, multi-GPU plus graph assembly	Cannot handle genes absent from the graph

Structure and phenotype informed	Folding, interface or phenotype features	No	Structure or ontology dependent	Moderate to high	Uneven coverage; external dependency
----------------------------------	--	----	---------------------------------	------------------	--------------------------------------

Four observations follow.

The uncertain class is systematically excluded. Across all six families, the standard evaluation protocol discards variants of uncertain significance, along with conflicting interpretations and likely-pathogenic or likely-benign intermediate labels. This is defensible as a modelling choice, since the labels are noisy by definition. It is not defensible as a claim about clinical utility, because it removes the majority class and the entire population of cases a clinician would actually want help with. A model evaluated only on confidently labelled variants is being tested on the problem that was already solved. ESGMM is a partial exception, in that it quantifies uncertainty rather than ignoring it, but it still excludes uncertain variants from the final comparison rather than predicting them.

Evaluation rigour varies and is not always stated. Gene-grouped splitting and random row-level splitting produce systematically different numbers, and the gap is large. Papers that do not state which protocol was used cannot be meaningfully compared against papers that do. The finding of Ghadie et al. that per-gene balancing lowered their reported accuracy is a useful reference point, since it quantifies how much of a headline number can be attributable to within-gene label skew. Bromberg et al.'s observation that the field lacks a shared definition of variant effect compounds the problem.

Capability and accessibility are moving in opposite directions. The trend line runs from single-CPU conservation scores through modest tree ensembles to multi-GPU foundation models over multi-source knowledge graphs. Accuracy has improved along this line, but so has the barrier to entry. Two findings suggest the added complexity is not always buying proportionate accuracy: the five-node-type graph performing on par with the twelve-node-type graph in Ghadie et al., and PATHOS reporting no gain from fine-tuning its protein language models.

Confidence is reported without being calibrated. Almost every system emits a probability. Very few report expected calibration error or any other calibration diagnostic. ESGMM is the clearest exception and its results argue that the omission is costly. In a screening context where the output informs a triage decision, an overconfident wrong answer is more damaging than a hedged one, and the field's near-silence on calibration is a substantive gap rather than a presentational one.

4.2 Consolidated research gaps

Table 4 states the gaps and maps each to the component of the proposed framework that addresses it.

Table 4. Research gaps and corresponding design responses.

ID	Gap	Evidence from the survey	Design response (Section)
G1	Three-class prediction is under-addressed; VUS is the dominant class yet is excluded from nearly all evaluations	All six families in Table 3	Three-class output head, VUS retained as a label (5.4)
G2	Single-modality bottlenecks; sequence, clinical and relational evidence are rarely combined coherently	Sections 3.1 to 3.5	Tri-modal encoders with late fusion (5.4)
G3	Naive imbalance handling produces the zero-trap; high accuracy with near-zero minority F1	ClinVar class distribution; [9]	Balanced class weighting; SMOTE rejected (5.5)
G4	Leakage-resistant evaluation is inconsistently applied	[2] per-gene balancing result; [29]	GroupKFold on GeneID; both splits reported (5.5, 6)
G5	Compute requirements exclude most practitioners	[2] requires up to four 40 GB GPUs	Single commodity GPU target (5.2, 5.4)
G6	Confidence scores are emitted but not calibrated	[28] uncertainty exclusion gains of 7 to 12 points	Monte Carlo dropout, ECE reported (5.6, 6)

5. Proposed Framework: GenoHybrid-VEP

We outline GenoHybrid-VEP as a direct response to G1 through G6. What follows is an architectural specification. No performance claims are made in this paper; empirical validation is deferred to a companion work.

5.1 Design principles

The framework rests on four commitments. Uncertainty is a label, not a filter, so variants of uncertain significance are retained as a third target class and are additionally accompanied by a quantified epistemic uncertainty estimate. Modality separation is preserved until representations are learned, so sequence, biomarker and relational evidence are processed by dedicated encoders and fused only at the latent level. Evaluation is grouped by gene so that reported performance reflects generalisation to unseen genes. And the pipeline is constrained to train and serve on a single commodity GPU.

5.2 Data pipeline

The source is the ClinVar release on the GRCh38 assembly. The raw archive is processed chunk-by-chunk and stored as columnar Parquet with compression, which reduces a multi-gigabyte text corpus to roughly sixty megabytes on disk and allows individual feature columns to be loaded without reading the full record. All coordinates are normalised to GRCh38, which eliminates the assembly ambiguity that affects legacy ClinVar records mapped to earlier builds.

Sequence context is retrieved from a local GRCh38 FASTA store indexed for random access through pyfaidx. This reference store is the one substantial local asset in the pipeline; everything downstream of it is small. Population frequency and conservation annotations are resolved through external APIs with responses cached locally, so that repeated queries against the same variant incur no additional network cost.

Figure 2 shows the end-to-end execution lifecycle from ingestion to diagnostic output.

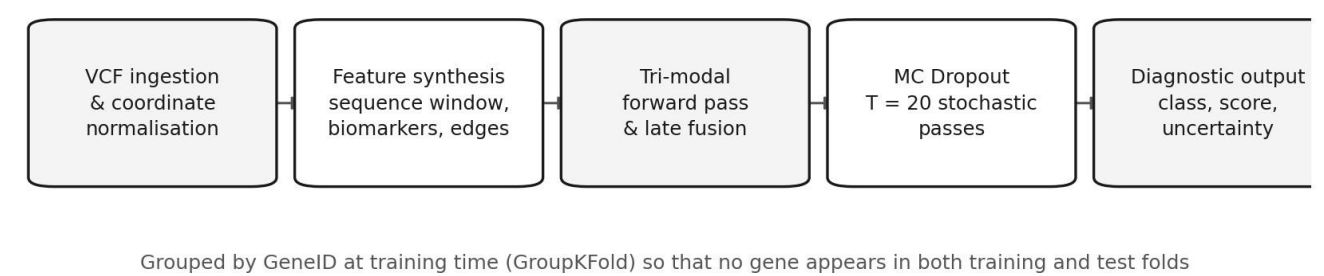


Figure 2. End-to-end data lifecycle. Grouping by GeneID at training time is what makes the reported test performance a measure of generalisation to unseen genes rather than of gene-level recall.

5.3 Feature construction

Three modalities are constructed per variant, detailed in Table 5.

Table 5. Input features by modality.

Modality	Feature	Encoding	Rationale
Sequential	Reference window POS – 50 to POS + 50	One-hot across A, C, G, T; tensor (4, 101)	Captures local sequence context, splice-adjacent signal and neighbourhood composition
Biomarker	gnomAD allele frequency	Continuous, scaled	High population frequency is strong evidence for benignity
Biomarker	PhyloP score	Continuous, scaled	Per-position evolutionary constraint
Biomarker	PhastCons score	Continuous, scaled	Conserved-element membership probability
Biomarker	Variant origin	One-hot: germline, somatic, other	Inheritance mode changes the interpretive frame
Biomarker	ClinVar review status	Ordinal, unreviewed through practice guideline	Converts assertion stars into an evidence-quality weight

Biomarker	Molecular consequence	One-hot	Distinguishes missense, synonymous, splice-region and others
Biomarker	Gene prior	Laplace-smoothed pathogenic-to-total ratio	Summarises gene-level mutational intolerance; neutral default for unseen genes
Relational	Shared-gene and proximity edges	Sparse edge index tensor	Encodes non-independence between neighbouring variants

The window length is a deliberate choice informed by Section 3.4: a wider context admits noise that degrades discrimination between variants in close proximity, and the reported difficulty of fine-tuning at very long context lengths argues against maximising window size on principle. A 101 base-pair window captures local context without the memory cost of long-range models. Section 6 describes the ablation that tests this reasoning rather than assuming it.

The Laplace smoothing on the gene prior is what allows the model to accept variants in genes ClinVar has never recorded. An unseen gene receives a neutral value of 0.5 and the prediction is then driven by the sequence and biomarker evidence, rather than failing outright. This directly addresses the coverage limitation that ontology-based graph methods acknowledge.

5.4 Tri-modal encoder and fusion

Figure 3 shows the network. Table 6 gives the layer-by-layer specification.

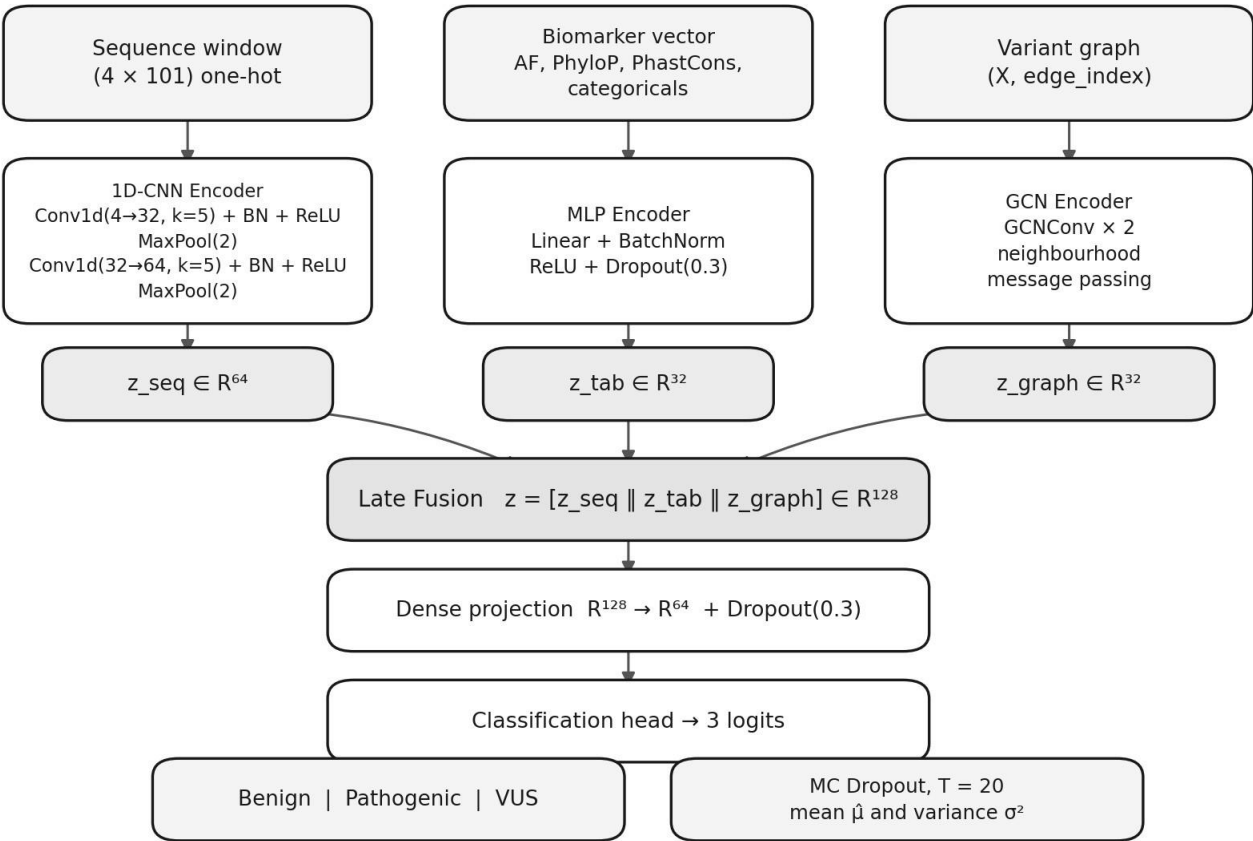


Figure 3. Tri-modal convergence and fusion architecture. Each modality is encoded independently before the three latent vectors are concatenated, which prevents the dense layers from having to model unrelated input domains simultaneously.

Table 6. Layer-by-layer specification.

Component	Layer	Configuration	Output
Sequence encoder	Conv block 1	Conv1d(4 → 32, k=5, pad=2), BatchNorm1d, ReLU, MaxPool1d(2)	(32, 50)
Sequence encoder	Conv block 2	Conv1d(32 → 64, k=5, pad=2), BatchNorm1d, ReLU, MaxPool1d(2)	$z_{seq} \in \mathbb{R}^{64}$
Biomarker encoder	Dense stack	Linear, BatchNorm1d, ReLU, Dropout(0.3)	$z_{tab} \in \mathbb{R}^{32}$
Graph encoder	GCNConv × 2	Message passing over shared-gene and proximity edges	$z_{graph} \in \mathbb{R}^{32}$
Fusion	Concatenation	$[z_{seq} \parallel z_{tab} \parallel z_{graph}]$	$z \in \mathbb{R}^{128}$

Head	Dense projection	Linear(128 → 64), BatchNorm1d, ReLU, Dropout(0.3)	$\mathbb{R}^{64}$
Head	Output	Linear(64 → 3)	3 logits

The kernel size of 5 reflects the scale at which local nucleotide context is informative, and pooling provides limited translation tolerance so that the encoder is not brittle to exact positioning within the window. Node states in the graph encoder are initialised from the fused output of the first two encoders, so that the graph refines an existing prediction rather than constructing one from scratch.

Late fusion is chosen over early fusion because concatenating raw one-hot nucleotide indicators with continuous biomarker floats before any representation learning forces a single dense stack to model unrelated domains simultaneously, degrading the spatial pattern extraction the convolution exists to perform. Tree ensembles are set aside for the sequence modality because they fragment a 101-position window into independent categorical splits, discarding positional structure entirely. Large genomic transformers are set aside for this stage on the grounds discussed in Section 3.4, namely that the reported benefit for single nucleotide variant pathogenicity is equivocal while the parameter and memory cost is not.

5.5 Training strategy

Splitting is performed with GroupKFold on GeneID, so that all variants of a given gene fall entirely within one fold. The test partition therefore contains only genes the model has never encountered, and reported performance measures generalisation rather than gene-level recall.

Class imbalance is handled through balanced class weighting inside the cross-entropy objective, with weights derived from inverse class frequency. Synthetic minority oversampling was considered and rejected: SMOTE generates new samples by interpolating between existing ones, which is coherent for continuous features but produces biologically meaningless values when applied to one-hot nucleotide channels and categorical clinical annotations [32]. A synthetic variant that is 0.4 adenine and 0.6 guanine at a reference position does not correspond to anything, and training on such samples corrupts precisely the categorical structure the encoders are meant to learn. Reweighting the loss achieves the same rebalancing objective without fabricating data. Weighted focal loss [31], which additionally down-weights easy examples to concentrate gradient on the ambiguous boundary between uncertain and pathogenic, is retained as an alternative to be evaluated in the loss ablation described in Section 6.

Optimisation uses AdamW with decoupled weight decay [33] under a cosine annealing schedule.

5.6 Uncertainty quantification

Dropout at rate 0.3 is retained at inference time and $T = 20$ stochastic forward passes are performed per variant, following the Monte Carlo dropout approximation to Bayesian inference [30]. For an input x , the predicted class distribution and the epistemic uncertainty are

$$\hat{\mu} = (1/T) \sum \hat{y}^{(t)} \text{ and } \sigma^2 = (1/T) \sum (\hat{y}^{(t)} - \hat{\mu})^2$$

with the sum taken over the T stochastic passes. The mean gives the prediction and the variance is surfaced to the user rather than discarded.

The motivation is empirical as much as theoretical. ESGMM reports accuracy gains of seven to twelve percentage points from excluding high-uncertainty predictions, which indicates that a substantial share of residual error is concentrated in cases the model can identify as unreliable. In a triage setting the correct response to such a case is to route it to human review, not to emit a confident label.

5.7 Deployment architecture

The system follows a three-tier decoupled architecture, shown in Figure 4.

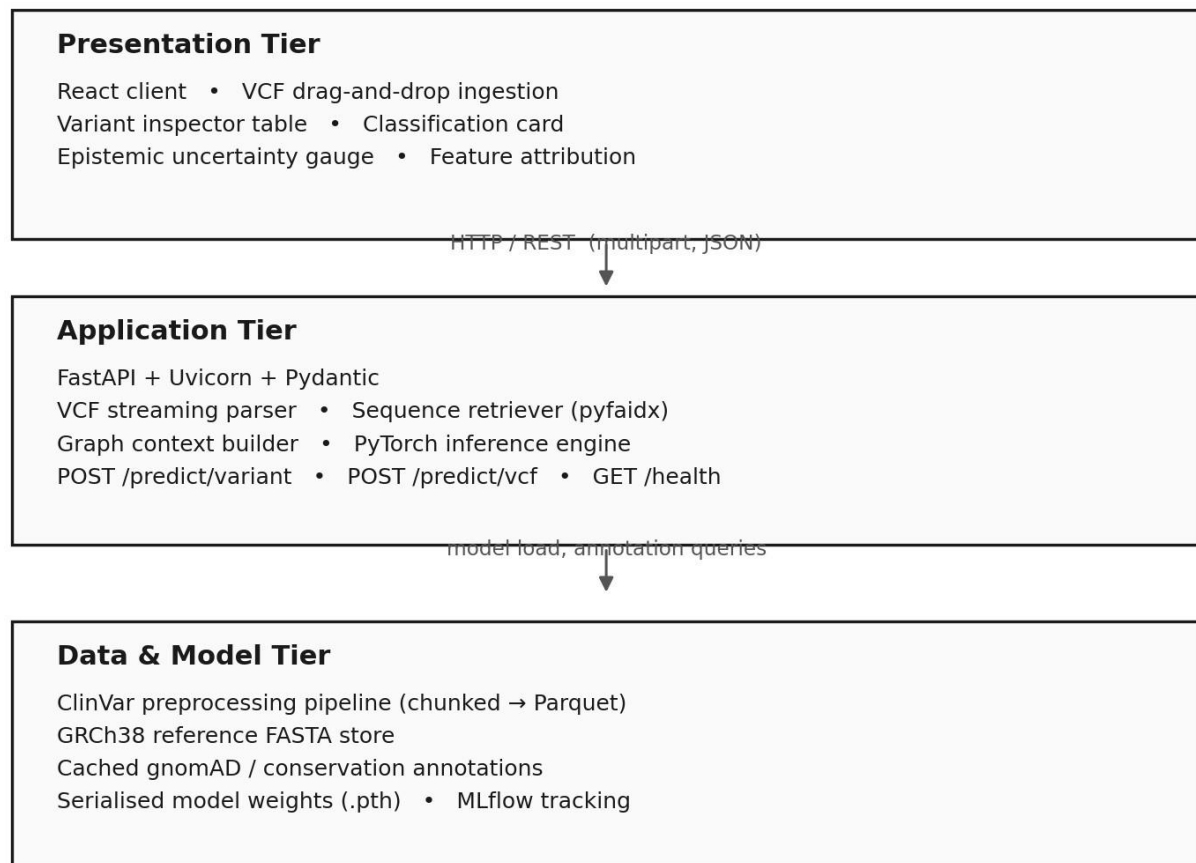


Figure 4. Three-tier system topology. Decoupling the tiers allows the inference engine to be scaled or replaced without changes to the client, and keeps the reference and annotation stores behind a single service boundary.

The presentation tier is a React client providing drag-and-drop ingestion of VCF files alongside manual single-variant entry, a variant inspector table sortable by predicted pathogenicity, a colour-coded classification card showing the full probability distribution, an epistemic uncertainty gauge surfacing the Monte Carlo variance, and a feature attribution panel identifying the biomarkers contributing most to each prediction.

The application tier is a FastAPI service under an ASGI server with Pydantic validation schemas. It exposes a single-variant prediction endpoint, an asynchronous batch endpoint for VCF files, and a health endpoint reporting loaded weights and inference device state. A streaming VCF parser extracts chromosome, position, reference and alternate fields while filtering header comments and malformed records; a sequence retriever queries the local FASTA through pyfaidx; and a graph service constructs edge tensors from the parsed batch.

The data and model tier holds the ClinVar preprocessing pipeline, the GRCh38 FASTA store, the annotation cache and the serialised model weights. The full stack is containerised for reproducibility, with experiment tracking through MLflow.

5.8 Implementation status

At the time of writing, the presentation tier and the first two encoders, sequential and biomarker, are implemented along with the fusion layer and the classification head. The relational graph encoder is specified but not yet integrated. Reporting this explicitly matters for interpreting Section 6: the ablation sequence described there is a plan, not a completed set of experiments.

6. PROPOSED EVALUATION PROTOCOL

We state the intended protocol in advance, since a substantial part of the critique in Section 4 concerns evaluation practice rather than modelling.

Reported metrics will be balanced accuracy, macro-averaged F1, per-class precision and recall, area under the receiver operating characteristic curve computed one-versus-rest, and Matthews correlation coefficient. Expected calibration

error and reliability diagrams will be reported for every configuration.

Results will be reported under both gene-grouped and random splitting, so that the magnitude of the leakage effect is visible rather than assumed. A temporal split, training on variants submitted up to a cutoff date and testing on later submissions, will be reported as a proxy for forward deployment.

Table 7 lists the planned ablations.

Table 7. Planned ablation studies.

Ablation	Variants compared	Question answered
Modality	Sequence only; biomarker only; sequence + biomarker; full tri-modal	What does each modality contribute?
Graph	With and without GCN refinement	Does relational context resolve boundary cases?
Window size	51, 101 and 201 base pairs	Is the 101 bp choice justified or merely convenient?
Loss function	Unweighted CE; balanced-weighted CE; weighted focal	How large is the zero-trap effect?
Split protocol	GroupKFold on GeneID vs random row split	How much apparent accuracy comes from gene memorisation?
Task formulation	Binary vs three-class	What does retaining VUS actually cost?

Baselines will include established clinical tools and a gradient boosted tree over the same biomarker features, so that the contribution of the architecture is separated from the contribution of the feature set.

7. CONCLUSION

Variant effect prediction has advanced from conservation heuristics through engineered ensembles to foundation models and relational graphs, and reported accuracy has improved along the way. This survey argues that the improvement has been partly measured against a task that was narrowed to make it tractable. Excluding variants of uncertain significance removes the majority class and the clinically difficult cases; random splitting permits gene-level memorisation; and the compute requirements of the leading systems restrict them to well-resourced institutions.

GenoHybrid-VEP is proposed as a response shaped by these findings rather than by architectural ambition. It retains uncertainty as a predictable class and quantifies it explicitly, separates sequence, biomarker and relational modalities until their representations are learned, trains under an objective built for skewed data without fabricating synthetic biological samples, evaluates under gene-grouped splitting, and constrains itself to hardware an ordinary laboratory already owns. Whether these choices translate into competitive performance is an empirical question this paper does not answer, and the evaluation protocol in Section 6 is stated in advance precisely so that the answer, when it arrives, is interpretable.

REFERENCES

1. Liu, Y., Yeung, W.S.B., Chiu, P.C.N., Cao, D. Computational approaches for predicting variant impact: an overview from resources, principles to applications. *Frontiers in Genetics*, 13:981005, 2022. doi:10.3389/fgene.2022.981005

2. Ghadie, M., Sardaar, S., Trakadis, Y. Disease-specific prediction of missense variant pathogenicity with DNA language models and graph neural networks. *Bioengineering*, 12(10):1098, 2025. doi:10.3390/bioengineering12101098

3. Landrum, M.J., Lee, J.M., Benson, M., et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 44(D1):D862–D868, 2016. doi:10.1093/nar/gkv1222

4. Chen, E., Facio, F.M., Aradhya, K.W., et al. Rates and classification of variants of uncertain significance in hereditary disease genetic testing. *JAMA Network Open*, 6:e2339571, 2023.

5. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, 47:D886–D894, 2019.

6. Adzhubei, I., Jordan, D.M., Sunyaev, S.R. Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols in Human Genetics*, 76(1):7.20.1–7.20.41, 2013. doi:10.1002/0471142905.hg0720s76

7. Liu, Y., Zhang, T., You, N., Wu, S., Shen, N. MAGPIE: accurate pathogenic prediction for multiple variant types using machine learning approach. *Genome Medicine*, 16:3, 2024. doi:10.1186/s13073-023-01274-4

8. Yang, Y., Wang, X., Zhou, D., Wei, D.Q., Peng, S. SVPath: an accurate pipeline for predicting the pathogenicity of human exon structural variants. *Briefings in Bioinformatics*, 23:bbac014, 2022.

9. Molotkov, I., Koboldt, D.C., Artomov, M. SNPred outperforms other ensemble-based SNV pathogenicity predictors and elucidates the challenges of using ClinVar for evaluation of variant classification quality. *medRxiv*, 2023. doi:10.1101/2023.09.07.23295192

10. Brandes, N., Goldman, G., Wang, C.H., Ye, C.J., Ntranos, V. 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

11. Lin, W., Wells, J., Wang, Z., Orengo, C., Martin, A.C.R. Enhancing missense variant pathogenicity prediction with protein language models using VariPred. *Scientific Reports*, 14, 2024. doi:10.1038/s41598-024-51489-7
12. Radjasandirane, R., Cretin, G., Diharce, J., de Brevern, A.G., Gelly, J.C. PATHOS: predicting variant pathogenicity by combining protein language models and biological features. *Artificial Intelligence in the Life Sciences*, 100165, 2026. doi:10.1016/j.ailsci.2026.100165
13. Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R., Liu, H. DNABERT-2: efficient foundation model and benchmark for multi-species genome. *arXiv:2306.15006*, 2024.
14. Nguyen, E., Poli, M., Faizi, M., et al. HyenaDNA: long-range genomic sequence modeling at single nucleotide resolution. *Advances in Neural Information Processing Systems*, 36:43177–43201, 2023.
15. Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., et al. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22:287–297, 2025.
16. Sayeed, M.A., Aldarmaki, H., Ben Amor, B. Gene pathogenicity prediction using genomic foundation models. *AAAI Symposium on Clinical Foundation Models*, 2024.
17. Cheng, J., Lawrence, C., Niepert, M. VEGN: variant effect prediction with graph neural networks. *arXiv:2106.13642*, 2021.
18. Zhang, H., Xu, M.S., Fan, X., Chung, W.K., Shen, Y. Predicting functional effect of missense variants using graph attention neural networks. *Nature Machine Intelligence*, 4:1017–1028, 2022.
19. Wang, B., Ye, B., Xu, L., Liang, J. AlphaGMut: a rationale-guided alpha shape graph neural network to evaluate mutation effects. *arXiv:2406.09159*, 2024.
20. Kamada, M., Takagi, A., Kojima, R., et al. Network-based pathogenicity prediction for variants of uncertain significance. *bioRxiv*, 2021. doi:10.1101/2021.07.15.452566
21. Hou, X. GraphnsSNPs: accurate prediction of disease-associated nsSNPs by utilizing graph neural network. *Proceedings of the 4th International Conference on Bioinformatics and Intelligent Computing*, 331–336, 2024.
22. Yu, H., He, G., Wang, W., Qin, S., Wang, Y., Bai, M., Shu, K., Pu, D. A graph neural network approach for accurate prediction of pathogenicity in multi-type variants. *Briefings in Bioinformatics*, 26(2):bbaf151, 2025. doi:10.1093/bib/bbaf151
23. Chandak, P., Huang, K., Zitnik, M. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10:67, 2023.
24. Lee, J., Yoon, W., Kim, S., et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36:1234–1240, 2020.
25. Liu, J., Chen, Y., Huang, K., Guan, X. Enhancing missense variant pathogenicity prediction with MissenseNet: integrating structural insights and ShuffleNet-based deep learning techniques. *Biomolecules*, 14(9):1105, 2024. doi:10.3390/biom14091105
26. Cao, L., Quan, L., Chen, Q., et al. Prediction of pathogenic mutations in human transmembrane proteins and their associated diseases via utilizing pre-trained Bio-LLMs. *Communications Biology*, 8, 2025. doi:10.1038/s42003-025-08452-7
27. Stein, D., Kars, M.E., Milisavljevic, B., Mort, M., Stenson, P.D., Casanova, J.L., Cooper, D.N., Boisson, B., Zhang, P., Schlessinger, A., Itan, Y. Expanding the utility of variant effect predictions with phenotype-specific models. *Nature Communications*, 16:11113, 2025. doi:10.1038/s41467-025-66607-w
28. Liu, J., Zhang, C., Liu, X., Zhang, Y., Liu, J., Guan, X. Variant pathogenicity prediction based on the ESGMM algorithm. *Discover Computing*, 27:49, 2024. doi:10.1007/s10791-024-09487-w
29. Bromberg, Y., Prabakaran, R., Kabir, A., Shehu, A. Variant effect prediction in the age of machine learning. *Cold Spring Harbor Perspectives in Biology*, 16(7):a041467, 2024. doi:10.1101/cshperspect.a041467
30. Gal, Y., Ghahramani, Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. *Proceedings of the 33rd International Conference on Machine Learning*, 1050–1059, 2016.
31. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P. Focal loss for dense object detection. *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988, 2017.
32. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.
33. Loshchilov, I., Hutter, F. Decoupled weight decay regularization. *International Conference on Learning Representations*, 2019.
34. Kipf, T.N., Welling, M. Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations*, 2017.
35. Paszke, A., Gross, S., Massa, F., et al. PyTorch: an imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32:8026–8037, 2019.
36. Fey, M., Lenssen, J.E. Fast graph representation learning with PyTorch Geometric. *arXiv:1903.02428*, 2019.
37. Karczewski, K.J., Francioli, L.C., Tiao, G., et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581:434–443, 2020.
38. McLaren, W., Gil, L., Hunt, S.E., et al. The Ensembl Variant Effect Predictor. *Genome Biology*, 17:122, 2016.
39. Richards, S., Aziz, N., Bale, S., et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, 17:405–424, 2015.