Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

RABANT: Lottery Ticket Compression of Genomic Transformers for Clinically Oriented Variant Pathogenicity Classification

Anirudha M

anirudham.cs24@rvce.edu.in

R.V. College of Engineering

Ashwin Acharya

R.V. College of Engineering

Rajashree Shettar

R.V. College of Engineering

Badarinath K B

R.V. College of Engineering

Article

Keywords: deep model compression, lottery ticket hypothesis, magnitude pruning, genomic transformer, DNA language model, variant pathogenicity prediction

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11639937/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

RABANT: Lottery Ticket Compression of Genomic Transformers for Clinically Oriented Variant Pathogenicity Classification

Anirudha M[1*], Ashwin Acharya[1], Rajashree Shettar[1], K. Badari Nath[1]

[1*]Department of Computer Science and Engineering, R V College of Engineering, Bengaluru, 560059, Karnataka, India.

*Corresponding author(s). E-mail(s): anirudham.cs24@rvce.edu.in;
Contributing authors: ashwinacharya.cs24@rvce.edu.in;
rajashreeshettar@rvce.edu.in; badarinath.kb@rvce.edu.in;

Abstract

Genomic foundation models improve variant effect prediction, but their size limits clinical deployment; each variant requires two forward passes, one per allele. We present RABANT (Rewinding-Aware Backbone Sparsification for Adaptive Nucleotide Transformers), a Lottery Ticket framework that prunes the pre-trained Nucleotide Transformer v2-100M, adapted as a dual-sequence reference-alternate classifier, and rewinds surviving weights before sparse fine-tuning. We compare four importance criteria (magnitude, EMA gradient, diagonal Fisher information, weight movement) on a leakage-audited benchmark of 217,026 variants from ClinVar, dbSNP, gnomAD and cBioPortal. Exploratory results identify magnitude as the only criterion near dense performance at useful sparsity and reveal a sharp boundary between 60% and 70% sparsity. In confirmatory five-fold evaluation on three unseen test sets, RABANT at 50% sparsity (48 million active parameters) showed no significant AUROC difference from the dense model (**88.13 ± 0.35** vs. **88.34 ± 0.33**), while increasing precision and specificity. Against a fully fine-tuned native NT-v2-50M of similar active size, RABANT improved AUROC by 1.44 percentage points and specificity by 6.09 points. Pruning a large pre-trained genomic transformer can therefore be preferable to training a smaller native model under limited inference resources.

Keywords: deep model compression, lottery ticket hypothesis, magnitude pruning, genomic transformer, DNA language model, variant pathogenicity prediction

1 Introduction

A single clinical whole-genome sequencing run yields four to six million candidate variants, of which only a small fraction carry curated evidence sufficient for confident interpretation [1]. The ACMG/AMP framework standardised variant classification by combining population frequency, molecular consequence, functional evidence and clinical observation, but remains dependent on curated evidence that is incomplete across many loci and populations [2]. ClinVar, dbSNP, gnomAD and cBioPortal complement one another with clinical, population-frequency and somatic-cancer evidence, yet represent fundamentally distinct annotation regimes [3–10]. Computational pathogenicity prediction is therefore a deployment-critical component of clinical pipelines, particularly in resource-constrained settings that require on-premises inference [11–12].

DNA language models have improved on classical variant-scoring tools by learning representations directly from nucleotide context rather than from pre-computed annotations. Their cost, however, scales with parameter count, and the dual-sequence evaluation required for variant effect prediction (one forward pass for the reference allele and one for the alternate allele) doubles the per-variant burden. A pipeline scoring millions of candidates per run consequently faces a compute budget that is difficult to satisfy on local hardware.

Neural network pruning offers a principled route to reducing that cost. The Lottery Ticket Hypothesis (LTH) states that dense networks contain sparse subnetworks which, trained from suitable initial conditions, match full-model accuracy [13]. Through RABANT, we apply the LTH to a pre-trained genomic transformer. We further explore whether a pruned large model is preferable to a natively smaller one, under an evaluation protocol strong enough to support clinically oriented claims. Two methodological considerations are particularly important when evaluating pruning methods: accounting for variability across training runs, and ensuring that smaller-architecture baselines are trained under comparable protocols. Single-run point estimates cannot separate a genuine effect from seed variance, and smaller-architecture baselines trained under weaker protocols lead to unfair comparisons with the pruned models they are evaluated against. RABANT addresses both concerns.

To summarise, our contributions are as follows.

1. We formalise Lottery Ticket subnetwork selection for a pre-trained genomic transformer as a constrained combinatorial optimisation problem. We relate the four evaluated importance criteria (absolute weight magnitude, exponential-moving-average gradient, diagonal Fisher information and weight movement) to a unified second-order expansion of the loss increment induced by parameter removal, and show that magnitude pruning is the minimax-optimal choice when curvature is unknown. This provides an analytical basis for comparing the criteria.
2. We adopt a two-phase experimental design that separates model selection from statistical confirmation. An exploratory phase determines the importance criterion and sparsity operating point, after which a confirmatory phase evaluates only the selected configuration using five-fold cross-validation, accompanied by dispersion estimates, confidence intervals and hypothesis tests.

3. We demonstrate that RABANT at 50% sparsity retains performance comparable to the dense baseline, with the 95% confidence interval for the AUROC difference bounding the observed degradation at 0.71 percentage points (pp).
4. We benchmark RABANT against a native NT-v2-50M model fine-tuned across all twelve transformer layers, which provides a stronger comparison than a partially unfrozen baseline. RABANT retains higher AUROC than this baseline and also achieves higher precision and specificity at the default decision threshold.

An overview of the dual-sequence classifier and the RABANT compression pipeline is shown in Fig. 1.

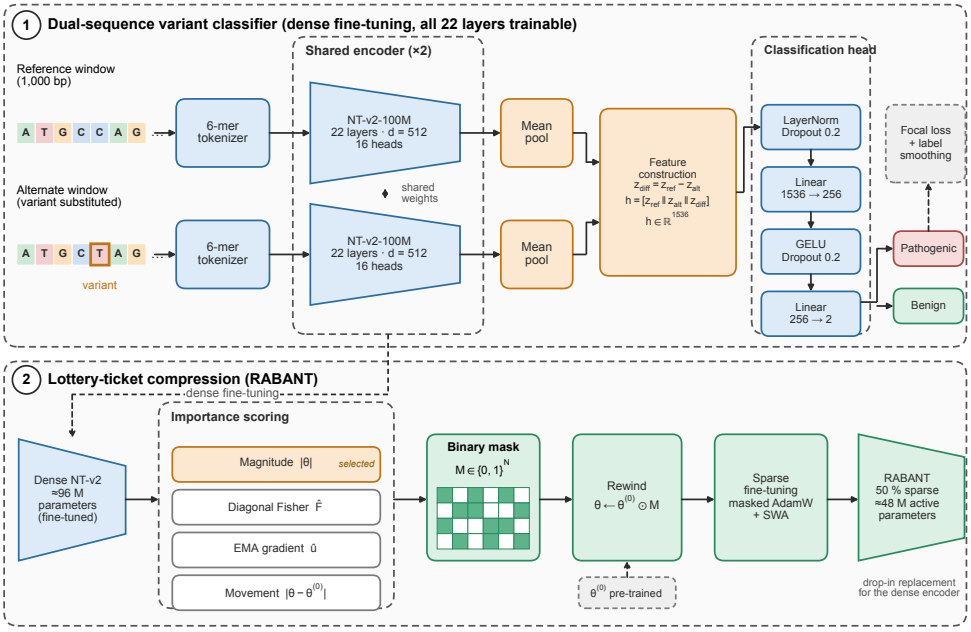


Fig. 1 Overview of the dual-sequence variant classifier and the RABANT compression pipeline. (1) Reference and alternate 1,000-bp windows (alternate carries the substituted variant) are tokenised with a 6-mer tokenizer and passed through a shared NT-v2-100M encoder. Mean-pooled embeddings $\bar{\mathbf{z}}_{\text{ref}}$ and $\bar{\mathbf{z}}_{\text{alt}}$ and their difference $\mathbf{z}_{\text{diff}}$ are concatenated into $\mathbf{h} \in \mathbb{R}^{1536}$ and passed to a classification head that outputs pathogenic and benign logits, trained with focal loss and label smoothing. **(2)** The fine-tuned dense model (approximately 96 M parameters) is scored by 4 importance criteria, of which magnitude is selected. A binary mask M is built, the surviving weights are rewound to the pre-trained values $\boldsymbol{\theta}^{(0)}$, and the masked network is fine-tuned, producing the 50% sparse RABANT model with approximately 48 M active parameters.

2 Related Work

2.1 Computational variant effect prediction

Early variant effect predictors combined evolutionary conservation with functional annotation. SIFT and PolyPhen-2 assess amino-acid tolerance from alignments [14–15]; MutationTaster2 and FATHMM extend this to disease- and cancer-associated substitutions [16–17]; CADD and REVEL integrate large annotation sets into a single score [18–19]; DANN and Eigen add representation learning to CADD-style scoring [20–21]; PrimateAI exploits primate-lineage variation [22]; and AlphaMissense combines structural and evolutionary representations for markedly improved proteome-wide accuracy [23]. A recent benchmark of 65 such predictors found that variant difficulty spans a wide spectrum, with a substantial residual set that no current tool resolves reliably, and that difficulty correlates with structural and functional protein properties [24]. Semantic-embedding approaches over curated annotation text have also been shown to support rapid pathogenicity prediction [25]. The shared limitation of annotation-driven methods is that their features are pre-computed and hand-engineered, under-representing splice-region, codon-neighbourhood and regulatory-motif effects that are visible only in raw nucleotide context.

2.2 Genomic language models

Sequence-native deep learning removed that dependency. DeepSEA, Basset, DanQ, ExPecto and Enformer predict chromatin and regulatory signal directly from sequence [26–30]. DNA foundation models then transferred masked language modelling to nucleotide sequence: DNABERT and DNABERT-2 introduced BERT-style multi-species pre-training [31–32]; the Nucleotide Transformer scaled masked language modelling across hundreds of species [33–34]; HyenaDNA and Caduceus extended context length with reverse-complement-equivariant modelling [35–36]; GPN showed that DNA language-model likelihoods are themselves strong zero-shot variant-effect predictors [37]; and Evo scaled genomic modelling toward whole-genome design [38]. These models capture context and motif regularities that fixed-feature predictors cannot, but their parameter counts are precisely what makes on-premises clinical deployment difficult. Efficient compression of these models is therefore a promising route to deployment.

2.3 Network pruning and the Lottery Ticket Hypothesis

Pruning has a long history grounded in second-order sensitivity analysis. Optimal Brain Damage and Optimal Brain Surgeon estimate the loss increase from deleting a parameter using curvature information [39–40]; later work replaced curvature with magnitude or gradient-based proxies for tractability [41–42]. The Lottery Ticket Hypothesis reframed pruning as a search for trainable sparse subnetworks rather than a post-hoc compression step [13], and subsequent work established that rewinding surviving weights to early checkpoints stabilises sparse training at scale [43], that winning tickets transfer across datasets and optimisers [44], and that such subnetworks exist natively inside large pre-trained language models, where magnitude masks

overlap strongly with high-performing subnetworks without iterative search [45–46]. Movement pruning, SparseGPT and Wanda compress large transformers using fine-tuning- or activation-aware criteria [47–49], and broad surveys catalogue the resulting trade-offs [50].

2.4 The gap addressed here

To summarise the current fragmented landscape, variant-effect predictors have grown steadily more accurate, yet remain anchored to pre-computed, hand-engineered features that miss raw-sequence context (§ 2.1). Genomic language models resolve that limitation by learning directly from nucleotide sequence, but their large parameter counts create a practical barrier to on-premises clinical deployment (§ 2.2). Network pruning offers a principled route to compression; however, the vast majority of pruning research targets vision and natural-language models (§ 2.3), and its application to biological sequence transformers is comparatively less studied. These three lines of work have advanced largely in parallel, and their intersection exposes three concrete gaps that motivate the present study.

No shared analytical frame for pruning criteria. Importance criteria such as absolute magnitude, diagonal Fisher information, EMA gradient and weight movement are routinely compared on leader-board benchmarks [41–42, 47], yet the comparisons are purely empirical: they reveal *that* one criterion outperforms another at a given sparsity level but not *why* a criterion that is competitive at low sparsity collapses at higher sparsity. Without a common analytical framework relating these criteria to local loss sensitivity, practitioners have no principled basis for selecting among them when compressing a new model class such as a genomic transformer.

Absence of clinically oriented evaluation rigour. Compression results for biological sequence models are seldom validated under cross-validation with dispersion estimates. Clinical variant labels carry inherent annotation uncertainty, and evaluation requires strict control of cross-database leakage [51]; single-split reporting, which dominates the pruning literature, cannot surface the variance that such label noise introduces. This highlights the need to subject a pruned genomic language model to held-out, source-specific five-fold cross-validation with mean ± standard-deviation reporting.

Missing fair small-model baseline. The question that matters most operationally is whether it is better to prune a large pre-trained model or simply to train a smaller one. This comparison is seldom conducted fairly: the smaller baseline is typically used off-the-shelf or trained with a limited budget, which risks overstating the advantage of pruning. Unless the native small model is fine-tuned to its own full potential and evaluated on the same unseen test partitions and under the same protocol, any claimed benefit of pruning over direct deployment remains confounded.

We address all three gaps within a single study. The four criteria are placed within a common Taylor-based analytical framework, which allows us to identify which components of the local loss expansion each criterion retains, approximates, or omits as sparsity increases. We adopt a two-phase experimental design: an exploratory selection phase followed by a confirmatory stage that is five-fold cross-validated on held-out, source-specific test sets, thereby providing the dispersion estimates that a clinically

oriented evaluation requires. Finally, we fully fine-tune the native small-model baseline using the identical dataset, classification head architecture and unseen test partitions, and evaluate it under the same five-fold cross-validation protocol. Together, these elements provide an end-to-end, theoretically grounded evaluation framework for pruning a genomic language model under clinically oriented evaluation constraints.

3 Results

3.1 Experimental design

All results follow a two-phase protocol that separates configuration selection from performance estimation. The *exploratory phase* compared the four importance criteria and swept sparsity from 10% to 90% using single-run training on a fixed validation split; its purpose was solely to select an importance criterion and an operating point, and its numbers are reported as such. The *confirmatory phase* then evaluated only the selected configuration (magnitude pruning at 50% sparsity), together with the dense baseline and the native small-model baseline, under five-fold cross-validation on three held-out, source-specific unseen test sets. All statistical claims in this paper derive exclusively from the confirmatory phase. Unless stated otherwise, confirmatory values are mean ± standard deviation across five folds, and comparisons use Welch's t-test on fold-level estimates with Cohen's d as effect size. The study design is summarised in Fig. 2.

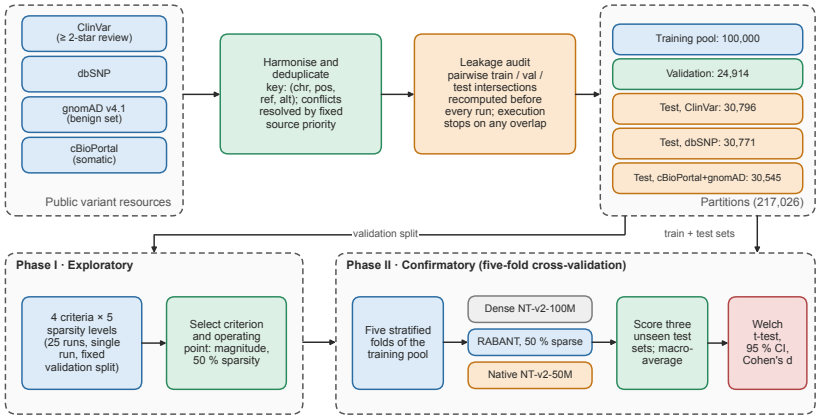


Fig. 2 Study design. Variants from 4 public resources are deduplicated and audited for leakage, then partitioned into a training pool (100,000 variants), a validation set (24,914) and three source-specific unseen test sets (30,796, 30,771 and 30,545). Phase I (exploratory) used the fixed validation split to compare four importance criteria at five sparsity levels (20 single runs, plus 5 rank-fusion runs) and to select the criterion and operating point. Phase II (confirmatory) evaluated the dense (uncompressed), native and RABANT models in five-fold cross-validation on the training pool, scored the three unseen test sets, and compared fold-level macro-averages with Welch's t-test, 95% confidence intervals and Cohen's d.

3.2 Selecting the importance criterion

Each of the four criteria was evaluated independently at five sparsity levels (20 pruning runs; Table 1 and Fig. 3a), and five further runs evaluated the rank-fusion score of Section 3.6, giving 25 pruning runs in total. Absolute magnitude was the only criterion that remained close to the dense baseline across the clinically useful range, staying within 0.07 pp at 50% sparsity and degrading mildly to 88.08% at 60%. Diagonal Fisher information was competitive at low sparsity (marginally exceeding magnitude at 10%, 88.71% vs. 88.50%) but degraded sharply beyond 30%, falling to 84.36% at 50% and 75.87% at 60%. The two criteria derived purely from the fine-tuning trajectory collapsed much earlier: EMA gradient fell from 87.84% at 10% to 69.07% at 50%, and weight movement from 86.57% to 71.86% over the same range.

Table 1 Exploratory phase: validation AUROC (%) by importance criterion and sparsity level, single-run. These results were used only to select the criterion and operating point; all confirmatory claims derive from the five-fold evaluation in Tables 3–5. Bold indicates the best criterion per column; † marks the selected operating point.

Criterion	10%	20%	30%	50%†	60%
Absolute magnitude	88.50	**89.18**	**88.93**	**89.35**	**88.08**
Diagonal Fisher	**88.71**	88.63	88.30	84.36	75.87
EMA gradient	87.84	83.43	74.00	69.07	68.95
Weight movement	86.57	79.92	75.78	71.86	71.90
Dense baseline (uncompressed)	89.42	89.42	89.42	89.42	89.42

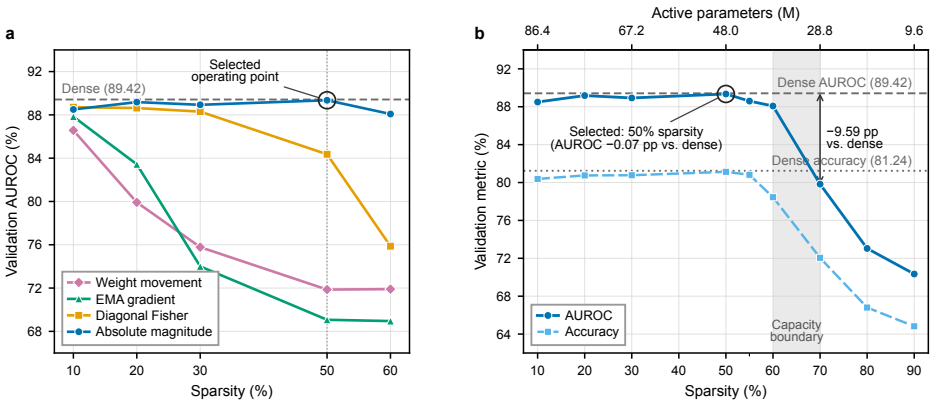


Fig. 3 Exploratory phase: choice of importance criterion and compression boundary. a, Validation AUROC of the four importance criteria at 10 to 60% sparsity (single runs). The dashed line is the dense baseline (89.42%) and the circle marks the selected operating point (magnitude, 50% sparsity). **b**, Magnitude pruning from 10 to 90% sparsity: validation AUROC and accuracy against sparsity, with the corresponding number of active parameters (M) on the upper axis. The shaded band marks the capacity boundary between 60% and 70% sparsity; at 70% the AUROC is 9.59 pp below the dense baseline.

The observed ordering is consistent with the analytical framework in Section 5.4, which explains the robustness of magnitude pruning under uncertainty, while the relative ordering among Fisher, EMA gradient and weight movement has been obtained empirically. Magnitude carries information accumulated over pre-training and operates directly on the weight values, whereas the diagonal Fisher estimate measures downstream curvature around the fine-tuned solution. EMA gradient measures recent downstream gradient behaviour and movement measures parameter displacement during fine-tuning. These three criteria are estimated from the downstream objective alone and cannot reconstruct the structure encoded during multi-species pre-training; once a substantial fraction of parameters must be removed, they plausibly begin deleting weights that are large in magnitude and functionally central, and performance collapses. Fisher's intermediate behaviour is consistent with its diagonal approximation: it captures per-parameter curvature but discards the off-diagonal coupling between attention and feed-forward blocks, an omission that becomes material only when many parameters are removed simultaneously. To test directly how the criteria differ in what they remove, we analysed the pruned weights of the dense fine-tuned model at 50% sparsity (Fig. 4).

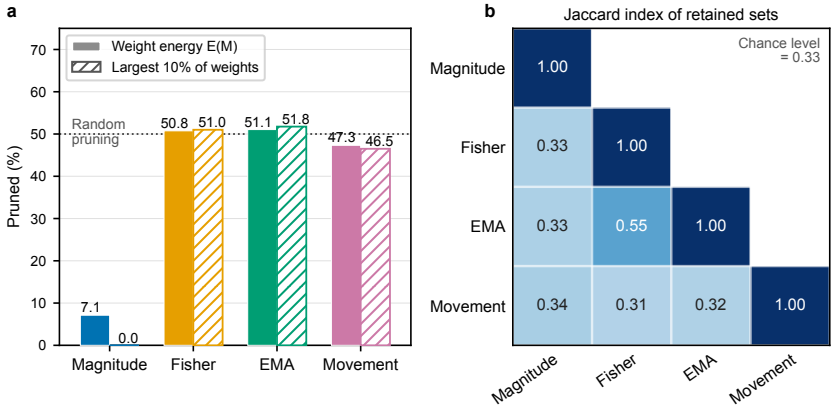


Fig. 4 Structure of the pruned weights at 50% sparsity for the four importance criteria (dense fine-tuned model). a, Percentage of the total weight energy $E(M)$ (solid bars) and of the largest 10% of weights (hatched bars) removed by each criterion. The dotted line marks the 50% removed by random pruning. By construction, magnitude pruning removes none of the largest 10% of weights (0.0%). **b**, Pairwise Jaccard index between the retained sets of each criterion over the prunable weights. Off-diagonal values near the chance level (1/3) show that the magnitude mask is statistically independent of the other three masks; only the Fisher and EMA gradient masks overlap more strongly. Together, these panels show that, among the four criteria, only magnitude pruning preserves the largest weights, whereas the other three remove about as much weight energy as random pruning.

The masks produced by the criteria differ structurally from the magnitude mask (Fig. 4), in a way that is consistent with their weaker downstream performance. While magnitude pruning strictly removes the lowest-magnitude weights, the Fisher, EMA gradient, and movement criteria select weights across the entire magnitude range,

behaving almost like random pruning. Consequently, they strip away 50.8%, 51.1% and 47.3% of the total weight energy $E(M)$, and roughly half of the largest 10% of weights (51.0%, 51.8% and 46.5%, respectively; Fig. 4a). Magnitude pruning, by contrast, removes only 7.1% of the total weight energy and none of the largest 10% of weights. Their pruned energy is therefore about 6.7 to 7.2 times that of magnitude pruning, in agreement with the ordering $E(M) \geq E(M^{\mathrm{mag}})$ of Proposition 1.

The retained sets are also nearly unrelated (Fig. 4b). The Jaccard index between the magnitude mask and each of the other three masks (0.33 to 0.34) equals the value expected for two independent random masks at 50% sparsity (1/3), whereas the Fisher and EMA masks overlap more strongly (0.55). This indicates that, with respect to weight size, the alternative criteria select subnetworks that are no more related to the magnitude mask than random masks would be, and that they discard a large part of the high-energy weights that magnitude pruning preserves. The analysis separates magnitude from the other three criteria; it does not by itself explain the ordering among Fisher, EMA gradient and weight movement, whose masks remove similar amounts of weight energy. The exact values are given in Supplementary Tables S3 and S4.

3.3 Locating the compression boundary

A finer sweep of magnitude pruning from 10% to 90% (Table 2 and Fig. 3b) revealed two regimes separated by an abrupt boundary. Validation AUROC stayed within 1.5 pp of dense through 60% sparsity, then fell sharply: from 88.08% at 60% to 79.83% at 70% (a drop of 8.25 pp, or $\Delta = -9.59$ pp relative to dense), and further to 70.34% at 90%. We selected 50% sparsity as the operating point, where AUROC was 89.35% ($\Delta = -0.07$ pp) at 48.0 M active parameters.

The plateau-then-cliff shape locates an architecture-specific capacity limit: up to roughly 60% sparsity the retained subnetwork preserves the features required to rank pathogenicity, beyond which it can no longer represent the reference-alternate perturbation the classification head depends on. The structural analysis above (Fig. 4) suggests a possible reason why this boundary is specific to magnitude pruning. Because magnitude pruning leaves the high-energy weights of the network intact, it plausibly preserves the representational capacity of the dense model up to the cliff. By contrast, the alternative criteria remove roughly half of the high-energy weights at 50% sparsity, which is consistent with, but does not by itself establish, their poorer performance at higher sparsity.

3.4 Compression bounds AUROC loss to at most 0.71 pp

The confirmatory evaluation compared three models across the three unseen test sets under five-fold cross-validation: the dense NT-v2-100M, RABANT (the 50%-sparse magnitude-pruned NT-v2-100M), and a natively trained NT-v2-50M (Table 3). On macro-averaged AUROC, RABANT scored 88.13 ± 0.35 against the dense model's 88.34 ± 0.33, a difference of -0.21 pp ($t = -0.99$, $p = 0.354$). The 95% confidence interval on the difference, $[-0.71, +0.28]$ pp, is the more informative statement: at

Table 2 Exploratory phase: magnitude pruning across sparsity levels, single-run validation. Active parameters $N_{\text{active}} \approx (1-s)N$ with $N \approx 96\,\text{M}$ are nominal values: embeddings, layer-normalisation parameters and biases (about 3.2 M parameters) are never pruned, so the exact number of non-zero parameters is slightly higher (about 49.6 M at 50% sparsity; Supplementary Table S1). Bold marks the selected operating point.

Sparsity s	Active params (M)	Accuracy (%)	AUROC (%)	ΔAUROC (pp)
Dense (uncompressed)	96.0	81.24	89.42	–
10%	86.4	80.38	88.50	−0.92
20%	76.8	80.75	89.18	−0.24
30%	67.2	80.77	88.93	−0.49
50%	**48.0**	**81.13**	**89.35**	**−0.07**
55%	43.2	80.82	88.60	−0.82
60%	38.4	78.45	88.08	−1.34
70%	28.8	72.05	79.83	−9.59
80%	19.2	66.80	73.04	−16.38
90%	9.6	64.82	70.34	−19.08

95% confidence, halving the active parameter count costs at most 0.71 pp of macro-averaged AUROC, and the interval includes zero. The same holds per source, where no RABANT-versus-dense difference reached significance (ClinVar $p = 0.135$, dbSNP $p = 0.938$, cBioPortal+gnomAD $p = 0.633$). All comparisons use Welch's t-test.

Table 3 Confirmatory phase: unseen-test AUROC (%) under five-fold cross-validation, mean ± standard deviation. Δ_{dense} and Δ_{native} give the difference of RABANT against the dense and native baselines in percentage points. Significance from Welch's t-test on fold-level estimates: $^{*}p < 0.05$, $^{***}p < 0.001$, ns = not significant.

Test set	Dense NT-v2-100M (uncompressed)	Native NT-v2-50M	**RABANT 48M (Ours)**	Δ_{dense}	Δ_{native}
ClinVar	92.83 ± 0.25	90.97 ± 0.38	92.62 ± 0.08	−0.21 ns	$+1.65^{***}$
dbSNP	82.72 ± 1.98	81.55 ± 2.02	82.62 ± 1.98	−0.10 ns	+1.07 ns
cBioPortal+gnomAD	89.47 ± 0.98	87.54 ± 0.95	89.15 ± 1.09	−0.33 ns	$+1.60^{*}$
Macro-average	$\mathbf{88.34 \pm 0.33}$	$\mathbf{86.69 \pm 0.43}$	$\mathbf{88.13 \pm 0.35}$	−0.21 ns	$\mathbf{+1.44}^{***}$

Fig. 5 shows the validation AUROC of the three models over the training epochs in each fold. All three models reach their peak validation AUROC within the first four to five epochs and decline afterwards, which motivates early stopping on validation AUROC. In every fold, RABANT attains a peak comparable to that of the dense model (about 88 to 89%) and above that of the native model (about 87%). RABANT also starts from a higher first-epoch validation AUROC (about 83.5 to 84%, against 76 to 78% for the other two models), partly reflecting the smaller effective batch size of its sparse stage (Section 5.8), which doubles the number of optimiser updates per epoch.

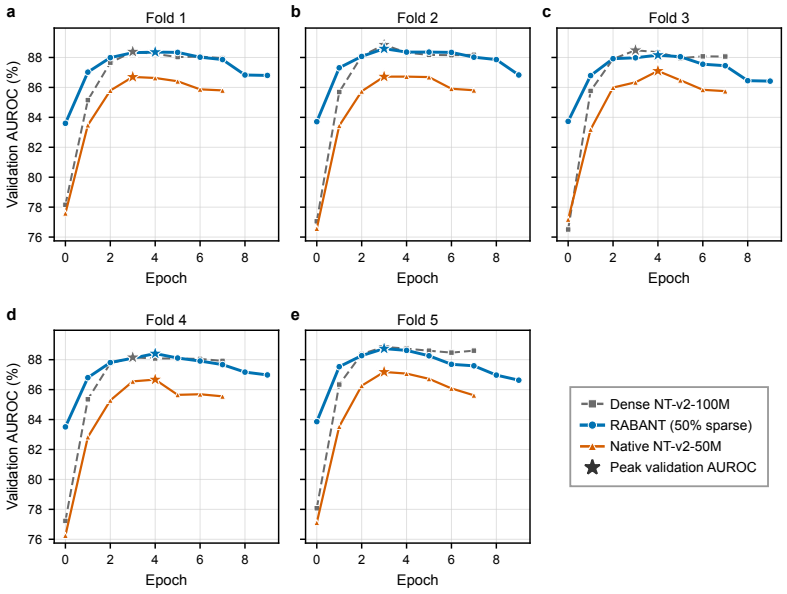


Fig. 5 Validation AUROC against epoch in each cross-validation fold. Panels **a** to **e** show folds 1 to 5 for the dense NT-v2-100M (uncompressed), native NT-v2-50M and RABANT 48M (Ours) models. For RABANT, epochs count the sparse fine-tuning stage. Stars mark the peak validation AUROC of each run.

3.5 RABANT outperforms a fully fine-tuned native NT-v2-50M model

The operationally decisive comparison is against a natively smaller transformer of comparable active size. To avoid understating that baseline, the NT-v2-50M was fine-tuned across all twelve of its transformer layers (without partial unfreezing), using the same dataset, classification head, loss, optimiser family and evaluation procedure as the NT-v2-100M model. Across five folds and the three unseen test sets (92,112 variants in total) it reached 86.69 ± 0.43 AUROC with 77.59 ± 0.72 accuracy (Table 4).

Table 4 Confirmatory phase: macro-averaged unseen-test metrics across the three holdout sets, reported as mean ± standard deviation over five folds.

Metric	Dense NT-v2-100M (uncompressed)	Native NT-v2-50M	**RABANT 48M (Ours)**
AUROC (%)	88.34 ± 0.33	86.69 ± 0.43	$\mathbf{88.13 \pm 0.35}$
Accuracy (%)	79.26 ± 0.39	77.59 ± 0.72	$\mathbf{78.93 \pm 0.49}$
Precision (%)	78.95 ± 0.99	77.96 ± 2.48	$\mathbf{82.11 \pm 1.73}$
Specificity (%)	77.00 ± 3.84	76.69 ± 4.83	$\mathbf{82.77 \pm 2.74}$
Active parameters	96M	∼50M	**48M**
Transformer layers	22	12	**22**

Even against this stronger baseline, RABANT retained a clear advantage (Tables 3 and 4). Macro-averaged unseen AUROC was 88.13 ± 0.35 (RABANT) versus 86.69 ± 0.43 (native NT-v2-50M model), a gain of 1.44 pp that is statistically significant (Welch's $t = 5.77$, $p = 0.0005$) with a 95% confidence interval of $[+0.86, +2.02]$ pp that excludes zero by a comfortable margin. The advantage was consistent in direction across all three sources and reached significance on ClinVar (+1.65 pp, $p = 0.0005$) and cBioPortal+gnomAD (+1.60 pp, $p = 0.039$); on dbSNP the +1.07 pp gain did not reach significance ($p = 0.422$), a consequence of that set's high fold-to-fold variance, which exceeds 1.9 standard deviation points for every model.

The gain is larger still in specificity. Macro-averaged specificity was 82.77 ± 2.74 for RABANT against 76.69 ± 4.83 for the native NT-v2-50M, an advantage of 6.09 pp ($p = 0.048$), accompanied by a 4.15 pp gain in precision ($p = 0.018$). RABANT therefore retains discriminative performance while shifting the balance of outcomes at the default decision threshold toward fewer false-positive pathogenic calls, as reflected in its higher precision and specificity. The four macro-averaged metrics of the three models are compared in Fig. 6, and the differences with their confidence intervals are shown in Table 5 and Fig. 7.

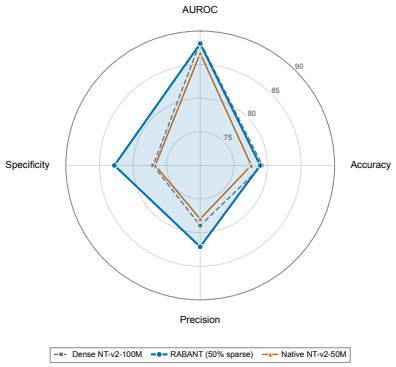


Fig. 6 Macro-averaged unseen-test metrics of the three models. Radar plot of AUROC, accuracy, precision and specificity (%, mean over five folds and three test sets; Table 4) for the dense NT-v2-100M (uncompressed), native NT-v2-50M and RABANT 48M (Ours) models. The radial axis spans 70 to 90%.

Table 5 Confirmatory phase: Welch's t-test on macro-averaged unseen-test metrics with five folds per model. Differences are RABANT minus comparator, in percentage points; d is Cohen's d.

Metric	Difference	95% CI	t	p	d
RABANT 48M (Ours) *vs. dense NT-v2-100M (uncompressed)*					
AUROC	−0.21	[−0.71, +0.28]	−0.99	0.354	−0.62
Accuracy	−0.33	[−0.98, +0.32]	−1.19	0.270	−0.75
Precision	+3.17	[+1.01, +5.32]	+3.54	0.011	+2.24
Specificity	+5.78	[+0.82, +10.74]	+2.74	0.028	+1.73
RABANT 48M (Ours) *vs. native NT-v2-50M*					
AUROC	+1.44	[+0.86, +2.02]	+5.77	0.0005	+3.65
Accuracy	+1.34	[+0.42, +2.26]	+3.45	0.011	+2.18
Precision	+4.15	[+0.96, +7.34]	+3.06	0.018	+1.94
Specificity	+6.09	[+0.08, +12.09]	+2.45	0.048	+1.55

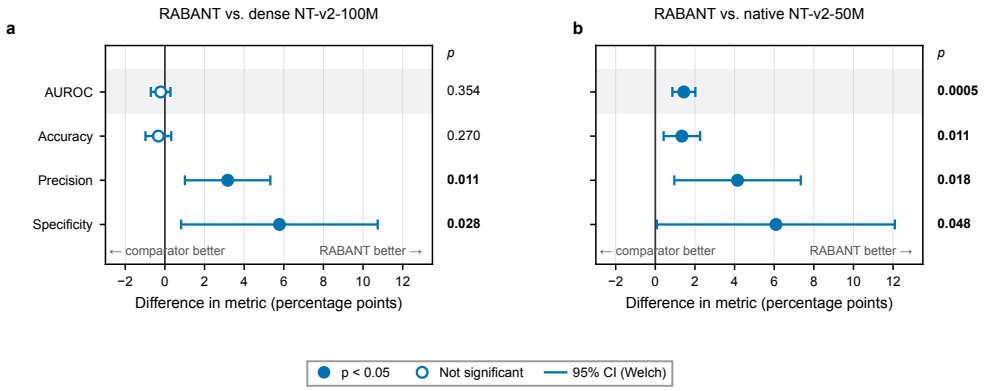


Fig. 7 Differences between RABANT and the two baselines. Differences in macro-averaged metrics (percentage points) between RABANT and **a**, the dense NT-v2-100M (uncompressed) and **b**, the native NT-v2-50M. Markers show mean differences and horizontal bars the 95% confidence intervals from Welch's t-test (Table 5); filled markers denote $p < 0.05$ and open markers non-significant differences. Two-sided p-values are given on the right (bold if $p < 0.05$); no adjustment for multiple comparisons was applied.

3.6 Rank fusion as a consistency check

As an auxiliary check on the exploratory selection, we computed a Dynamic Rank Fusion (DRF) score that weights the four criteria by a temperature-scaled softmax over their observed validation AUROC (Section 5.7). Because the weights derive from measured performance rather than an assumed ordering, the fusion summarises the evidence of Table 1 on a common scale: weight is assigned in proportion to demonstrated performance. The fusion concentrated on magnitude at every sparsity level, assigning it 0.902 of the total weight at 50% sparsity and 0.987 at 60%, and the fused model did not exceed magnitude alone at the selected operating point (89.09% vs. 89.35%). We therefore report DRF only as corroboration that magnitude is the optimal pruning signal; the weighting it induces is an empirical artefact of the observed AUROC values and carries no independent theoretical justification. The weights for all sparsity levels are given in Supplementary Table S5.

4 Discussion

4.1 Why magnitude survives compression and gradient signals do not

The criteria evaluated here differ in which part of a model's history they observe, and Section 5.4 makes that difference explicit. All four can be related to the loss increment $\Delta\mathcal{L}_i$ incurred by deleting parameter i: magnitude retains the weight factor of that increment, whereas the other three estimate its sensitivity factors from the downstream objective over different horizons. EMA gradient and weight movement are estimated over a fine-tuning trajectory of order 10^3 steps. Absolute magnitude, by

contrast, is a statistic of the weight itself, and therefore summarises the cumulative outcome of masked-language-model optimisation across 850 species.

For a model fine-tuned with a backbone learning rate of 5×10^{-6} over a handful of epochs, the parameters are expected to remain close to the pre-trained values: $\boldsymbol{\theta} \approx \boldsymbol{\theta}^{(0)}$. The fine-tuning gradient consequently describes sensitivity of the *downstream* objective in a small neighbourhood of a point that was chosen by a vastly longer and broader optimisation. A subnetwork must, however, continue to support the pre-trained representation, not merely the downstream decision boundary. This asymmetry is consistent with the collapse pattern in Table 1: at low sparsity, any criterion that avoids deleting clearly important weights suffices, and all four perform within a few points of each other; as sparsity rises, the trajectory-based criteria begin removing high-magnitude parameters that encode conserved sequence structure but happen to carry small downstream gradients, and performance falls away. Fisher curvature occupies a middle position for a related reason: it estimates curvature of the downstream loss but, through the squared-gradient expectation, aggregates over many mini-batches and therefore reflects a broader neighbourhood than an instantaneous gradient, while still discarding off-diagonal structure. The pruned-weight analysis (Fig. 4) is consistent with this reading at 50% sparsity, the only level analysed: at that sparsity, the three criteria derived from the downstream objective remove roughly seven times as much weight energy as magnitude pruning.

This reasoning suggests a general principle that our data support but cannot by themselves establish: the longer and more diverse the pre-training, the more reliable magnitude becomes as a pruning criterion, and the less informative short downstream gradients are.

4.2 Sparse depth retention versus architectural miniaturisation

The comparison against a fully fine-tuned NT-v2-50M isolates a question that arises whenever a smaller model is wanted: prune a large one, or train a small one? Our evidence favours pruning by a more modest margin than partial-fine-tuning baselines would suggest. Granting the small model its full capacity (all twelve transformer layers) provides more scope to adapt to all four dataset sources for better generalisability. RABANT nevertheless outperforms the native NT-v2-50M model on all four metrics reported in Table 4.

The mechanism we propose is depth retention. The native model has twelve transformer layers and must learn the task from a shallower hierarchy; RABANT retains all twenty-two layers and removes parameters *within* them. These are structurally different operations: architectural reduction removes representational stages, whereas pruning thins stages that remain. Because the layers of a genomic language model plausibly encode a progression from local nucleotide syntax through motif composition to longer-range context, removing stages is more damaging than thinning them. The higher specificity of RABANT is consistent with this interpretation, although as a threshold-dependent metric it is weaker evidence than AUROC.

4.3 Clinical implications of the precision and specificity shift

Relative to the dense model, RABANT retains discriminative performance, with no significant AUROC difference, while improving precision and specificity at the default decision threshold (Table 4). The shift we observe is therefore one in which performance is preserved as measured by AUROC while the proportion of benign variants correctly identified (specificity) and the proportion of pathogenic calls that are correct (precision) both increase. Threshold-free and threshold-dependent metrics answer different questions: AUROC characterises discrimination across all thresholds, whereas precision and specificity describe behaviour at the operating point that a deployed system would actually use. Reporting both is a general recommendation for pruning studies in clinical machine learning.

For confirmatory variant interpretation, the higher precision and specificity of RABANT reduce the false-positive burden that dominates curation workload. Deployments with different trade-offs between false-positive and false-negative errors should select the decision threshold on local validation data before use. In every case, halving the active parameters reduces the cost of the two forward passes each variant requires, which matters most in the privacy-sensitive on-premises settings where sending patient genomic data to external compute is not acceptable [11].

4.4 Limitations and future work

The pruning studied here is unstructured: individual weights are zeroed within otherwise dense matrices, so the reduction in active parameters does not by itself translate into wall-clock speed-up without sparse-kernel support. Structured or hardware-aligned sparsity, such as removing attention heads, feed-forward channels or entire layers, or enforcing N:M sparsity patterns, could convert parameter reductions into measurable throughput gains and is a natural direction for future work on target hardware. The evaluation can also be extended to other DNA language models to examine generalisability beyond the Nucleotide Transformer family. Assessing performance across ancestral backgrounds [12] and prospective clinical validation are further directions for this work.

5 Methods

5.1 Problem formulation

The goal is to find a sparse subnetwork of a trained classifier that retains the full network's predictive performance at a fraction of its parameter count. We formalise this as a constrained optimisation problem over binary masks. The notation is summarised in Table 9.

Let $\mathcal{D} = \{(\mathbf{x}_{\text{ref}}^{(j)}, \mathbf{x}_{\text{alt}}^{(j)}, y^{(j)})\}_{j=1}^{m}$ denote a variant dataset of m examples. Each example consists of two genomic sequences: $\mathbf{x}_{\text{ref}}$, the reference allele context, and $\mathbf{x}_{\text{alt}}$, the corresponding alternate allele context surrounding a variant site. The label $y \in \{0, 1\}$ indicates whether the variant is pathogenic ($y = 1$) or benign ($y = 0$). Let $f_{\boldsymbol{\theta}}$ be the dual-sequence classifier with parameters $\boldsymbol{\theta} \in \mathbb{R}^N$. The empirical risk over the dataset

is

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{m}\sum_{j=1}^{m} \ell\big(f_{\boldsymbol{\theta}}(\mathbf{x}_{\mathrm{ref}}^{(j)}, \mathbf{x}_{\mathrm{alt}}^{(j)}),\, y^{(j)}\big), \tag{1}$$

where ℓ denotes the per-example classification loss, specified in Section 5.3.

Pruning amounts to selecting which parameters to retain and which to discard. A binary mask $\mathbf{M} \in \{0,1\}^N$, one entry per parameter, encodes this selection: a parameter θ_i is kept if $M_i = 1$ and zeroed out if $M_i = 0$. The pruned subnetwork therefore has parameters $\boldsymbol{\theta} \odot \mathbf{M}$, where $\odot$ denotes element-wise multiplication. At a target sparsity s (the fraction of parameters removed), the optimal mask is the solution to

$$\mathbf{M}^{\star} = \underset{\mathbf{M}\in\{0,1\}^N,\ \|\mathbf{M}\|_0 \le (1-s)N}{\arg\min} \mathcal{L}(\boldsymbol{\theta} \odot \mathbf{M}). \tag{2}$$

In words, Equation (2) asks: among all subsets of at most $(1-s)N$ parameters, which subset minimises the loss? This is a cardinality-constrained subset-selection problem. For the Nucleotide Transformer v2-100M, where $N \approx 9.6 \times 10^7$, the search space at $s = 0.5$ contains $\binom{N}{N/2}$ candidate masks, which is a combinatorially intractable quantity. Every practical pruning method therefore replaces Equation (2) with a tractable surrogate: instead of solving the intractable problem directly, it uses a practical approximation criterion that assigns each parameter an importance score, and then retains the highest-scoring parameters. The different criteria compared in this work differ precisely in how they construct that surrogate. In practice the cardinality constraint is applied to the prunable subset of parameters defined in Section 5.4.

5.2 Multi-source benchmark and leakage control

Variants were collected from four public resources (Table 6) to reflect the heterogeneity of variant interpretation, spanning clinical assertions, population frequency and somatic cancer data. Because the resources differ in biological origin and in how a pathogenic or benign label is defined, label assignment is described separately for each resource. Pathogenic variants were coded as 1 and benign variants as 0.

ClinVar.

ClinVar variants were labelled from the curated clinical-significance assertion [3–4]. Pathogenic and likely pathogenic variants (including pathogenic/likely pathogenic and low-penetrance or risk-factor pathogenic assertions) were labelled pathogenic, and benign and likely benign variants were labelled benign. Variants of uncertain significance, variants with conflicting interpretations and other categories (for example drug response or not provided) were excluded. Only single-nucleotide variants with at least one review star (that is, at least one submitter providing assertion criteria) were retained. Pathogenic variants were restricted to protein-altering consequences (missense, stop-gained, stop-lost, start-lost and canonical splice-site variants), and benign variants to missense, synonymous, stop-retained and splice-region consequences. The two classes were balanced at 37,500 variants each, with pathogenic variants sampled preferentially from higher review tiers.

Table 6 Label assignment and variant counts by resource. Counts are the numbers of variants available for each class before consolidation (P, pathogenic; B, benign). CLNSIG is the ClinVar clinical-significance field; MAF is the minor allele frequency.

Resource	Origin	Pathogenic label	Benign label	Variants (P/B)
ClinVar	Germline	Pathogenic or likely pathogenic (CLNSIG); at least one review star	Benign or likely benign (CLNSIG); at least one review star	37,498 / 37,500
dbSNP	Germline	ClinVar cross-reference (CLNSIG 4 or 5)	Common SNP (MAF $\geq$ 0.01 in at least one population) without pathogenic ClinVar annotation	31,398 / 31,398
gnomAD v4.1	Germline (population)	Not used	Global allele frequency $\geq$ 0.01	0 / 54,691
cBioPortal	Somatic	Non-synonymous coding mutation (84 studies)	Not used	63,149 / 0

dbSNP.

dbSNP contributed two distinct sub-sources. Pathogenic labels derive from ClinVar cross-references recorded in dbSNP (CLNSIG values 4, likely pathogenic, or 5, pathogenic, including concordant multi-submitter entries). Benign labels were assigned to common variants flagged by dbSNP (minor allele frequency of at least 0.01 in at least one population) that carry no pathogenic ClinVar annotation. The pathogenic class is predominantly missense (65.6%) and stop-gained (33.8%) variants, whereas the benign class is divided almost evenly between missense (50.2%) and synonymous (49.8%) variants. The pathogenic labels in dbSNP are therefore not independent of ClinVar, because both rest on the same underlying clinical assertions; dbSNP broadens the variant coverage rather than providing an independent source of clinical evidence.

gnomAD.

gnomAD v4.1 (807,162 individuals) [6] contributed only benign variants. All variants with a global allele frequency of at least 0.01 (median 0.11) were labelled benign, on the premise that variants common in the general population are under near-neutral selection; this choice follows the allele-frequency evidence used for benign classification in the ACMG/AMP framework [2]. The set comprises synonymous (47.0%), missense (45.3%) and splice-region (6.5%) variants, with the remaining 1.2% being splice-donor, stop-gained, start-lost, stop-lost and stop-retained variants. Because gnomAD is a germline population resource, these labels reflect allele frequency and not functional annotation.

cBioPortal.

Somatic variants were drawn from 84 cancer genomics studies hosted on cBioPortal [7–9]: 79 studies of The Cancer Genome Atlas (TCGA; 58,689 variants, 92.9%) and five independent cohorts (MSK-IMPACT, METABRIC breast cancer, SU2C prostate cancer, DFCI melanoma and OncoSG lung adenocarcinoma; 4,460 variants, 7.1%). Non-synonymous coding mutations (missense, nonsense, splice-site, translation-start-site, nonstop, frameshift and in-frame mutations) were labelled pathogenic; the

retained variants are predominantly missense (91.5%). Silent mutations were not retained, so cBioPortal contributes pathogenic variants only. This labelling uses mutation consequence as a proxy for pathogenicity: a somatic mutation observed in a tumour is not necessarily a cancer driver, and the label is therefore a functional-impact proxy and not a clinical pathogenicity assertion.

Conflict resolution.

Each variant was identified by the key chromosome, position, reference allele and alternate allele. When the same variant appeared in more than one resource, the label of the highest-priority resource was retained, using the fixed order ClinVar > dbSNP (ClinVar cross-referenced) > gnomAD > dbSNP (common) > cBioPortal. This rule always yields a single label, so no variant was discarded because of an unresolved conflict.

Partitions and leakage control.

After harmonisation, deduplication and class balancing, the training benchmark contained 100,000 variants. A validation set of 24,914 variants was drawn from the remaining pool, and three independent source-specific unseen held-out test sets (ClinVar, dbSNP, and cBioPortal combined with gnomAD) were constructed with strict variant-level deduplication against training and validation, totalling 92,112 variants (Table 7). Leakage control was enforced by deduplicating on the tuple (chromosome, position, reference allele, alternate allele) before splitting, and by an automated audit that recomputed pairwise train/validation/test intersections before every run and terminated execution on any overlap [51]. The fixed validation set was used in the exploratory phase, whereas the confirmatory phase used five-fold cross-validation within the training partition and the three unseen test sets.

Table 7 Composition of the multi-source genomic variant benchmark. Unseen test sets were held out after strict variant-level deduplication by chromosome, position, reference allele and alternate allele.

Partition	Samples	Pathogenic	Benign
Training	100,000	50,000	50,000
Validation	24,914	12,422	12,492
Test: ClinVar	30,796	15,398	15,398
Test: dbSNP	30,771	15,398	15,373
Test: cBioPortal+gnomAD	30,545	15,148	15,397
Total	**217,026**	**108,366**	**108,660**

Sequence representation.

Each variant was represented as a paired reference–alternate sequence: for a variant at coordinate v, a 1,000-bp reference window $\mathbf{x}_{\text{ref}}$ was extracted from GRCh38/hg38

centred on v, and $\mathbf{x}_{\text{alt}}$ was obtained by substituting the alternate allele at the centre. This forces the model to evaluate the sequence perturbation rather than a single mutated sequence in isolation. Ambiguous bases were tolerated up to 5% of a window and resolved by random resampling over $\{A, C, G, T\}$; windows exceeding this tolerance or falling outside genome bounds were discarded.

Scope of the labels.

Three properties of the benchmark should be kept in mind when interpreting the results. First, the training data combine germline variants (ClinVar, dbSNP, gnomAD) and somatic variants (cBioPortal), so that the binary pathogenic/benign label covers both germline disease-causing variants and somatic mutations of cancer genomes. Second, the cBioPortal and gnomAD test set is coupled to the data source: every pathogenic variant is somatic and every benign variant is germline, so performance on this set may partly reflect the distinction between somatic and germline mutational context, and it is best read as an assessment of cross-source generalisation. The ClinVar test set, whose two classes are drawn from the same annotation framework, provides the most direct evidence for clinical pathogenicity discrimination, and the dbSNP test set is intermediate because its classes come from different sub-sources of one database. Third, the classes differ in consequence composition in every resource (for example, stop-gained variants are almost exclusively pathogenic and synonymous variants almost exclusively benign in dbSNP), so part of the signal that a sequence model can learn reflects variant consequence type.

5.3 Dual-sequence architecture and training objective

The backbone is the Nucleotide Transformer v2-100M multi-species checkpoint, a transformer encoder pre-trained on 850 species with a masked-language-modelling objective [33–34]: 22 layers, hidden dimension $d_h = 512$, 16 attention heads per layer, feed-forward dimension 2,048, and a 6-mer vocabulary of approximately 4,107 tokens.

Reference and alternate sequences are passed independently through the shared backbone and their hidden states mean-pooled across token positions,

$$\bar{\mathbf{z}}_{\text{ref}} = \frac{1}{n_{\text{tok}}} \sum_{t=1}^{n_{\text{tok}}} \mathbf{z}_{\text{ref},t}, \qquad \bar{\mathbf{z}}_{\text{alt}} = \frac{1}{n_{\text{tok}}} \sum_{t=1}^{n_{\text{tok}}} \mathbf{z}_{\text{alt},t}, \tag{3}$$

where n_{tok} is the token count and $\mathbf{z}_{\text{ref},t}, \mathbf{z}_{\text{alt},t} \in \mathbb{R}^{d_h}$ are per-token contextual embeddings. Mean pooling is preferred here to a single classification token because pathogenicity evidence may be distributed across the local context (splice-region, regulatory and codon-neighbourhood effects) rather than concentrated at one position. The two embeddings and an explicit mutation-difference term are finally concatenated,

$$\mathbf{h} = [\bar{\mathbf{z}}_{\text{ref}} \,\|\, \bar{\mathbf{z}}_{\text{alt}} \,\|\, (\bar{\mathbf{z}}_{\text{ref}} - \bar{\mathbf{z}}_{\text{alt}})] \in \mathbb{R}^{3d_h}, \tag{4}$$

giving a 1,536-dimensional vector whose difference term represents the variant's perturbation directly rather than requiring the head to infer it. Classification makes use

of a compact two-layer head with layer normalisation, dropout ($p = 0.2$) and GELU,

$$\mathbf{o} = \mathrm{Linear}_{256\to 2}\left(\mathrm{Dropout}_{0.2}\left[\mathrm{GELU}\left(\mathrm{Linear}_{1536\to 256}\left(\mathrm{Dropout}_{0.2}\left[\mathrm{LayerNorm}(\mathbf{h})\right]\right)\right)\right]\right), \tag{5}$$

where $\mathbf{o} \in \mathbb{R}^2$ denotes the unnormalised class logits. Both linear layers are initialised using Xavier-uniform initialisation.

The objective combines focal re-weighting [52] with label smoothing. For class label y_c, predicted probability p_c, smoothing ϵ and focusing parameter γ,

$$\ell_{\mathrm{focal}} = -\sum_{c=0}^{1}(1-p_c)^{\gamma}\left[(1-\epsilon)y_c + \frac{\epsilon}{2}\right]\log(p_c). \tag{6}$$

The factor $(1-p_c)^{\gamma}$ down-weights examples the model already classifies confidently, concentrating gradient signal on ambiguous variants, while the smoothing term prevents over-confident fitting to labels that themselves carry curation uncertainty. For dense training $\gamma = 1.5$, $\epsilon = 0.05$, up to 20 epochs with early stopping on validation AUROC (patience 4), all 22 layers and input embeddings unfrozen, AdamW [53] with differential learning rates ($\eta_{\mathrm{backbone}} = 5\times10^{-6}$, $\eta_{\mathrm{head}} = 5\times10^{-4}$), weight decay 0.01, 15% linear warm-up then cosine annealing, batch size 32 with 4-step accumulation and gradient clipping at 1.0. For sparse fine-tuning $\gamma = 2.0$, $\epsilon = 0.08$.

5.4 Analysis of the pruning criteria

Fig. 8 summarises the analytical framework developed in this section.

Prunable set and sparsity.

Let $\mathcal{P} \subseteq \{1,\dots,N\}$ index the prunable parameters, namely the two-dimensional weight matrices of the attention, feed-forward and classification-head layers, and let $N_{\mathcal{P}} = |\mathcal{P}|$. Layer-normalisation parameters, biases and input embeddings are always retained, so $M_i = 1$ for every $i \notin \mathcal{P}$. The target sparsity s is defined on the prunable set: exactly $K_s = \lfloor sN_{\mathcal{P}} \rfloor$ coordinates are removed. The overall fraction of removed parameters is therefore $s_{\mathrm{tot}} = K_s/N$, which is marginally smaller than s because retained parameters lie outside $\mathcal{P}$.

Decomposition of the loss increment.

Let $\mathcal{Z} = \{i \in \mathcal{P} : M_i = 0\}$ be the pruned set and $\delta = -\boldsymbol{\theta} \odot (\mathbf{1} - M)$ the deletion vector, so that $\boldsymbol{\theta} \odot M = \boldsymbol{\theta} + \delta$. Write $g = \nabla\mathcal{L}(\boldsymbol{\theta})$ and $H = \nabla^2\mathcal{L}(\boldsymbol{\theta})$. A second-order Taylor expansion about $\boldsymbol{\theta}$ gives

$$\mathcal{L}(\boldsymbol{\theta} \odot M) - \mathcal{L}(\boldsymbol{\theta}) \;=\; \sum_{i\in\mathcal{Z}} \Delta\mathcal{L}_i \;+\; \mathcal{I}(M) \;+\; R_3(M), \tag{7}$$

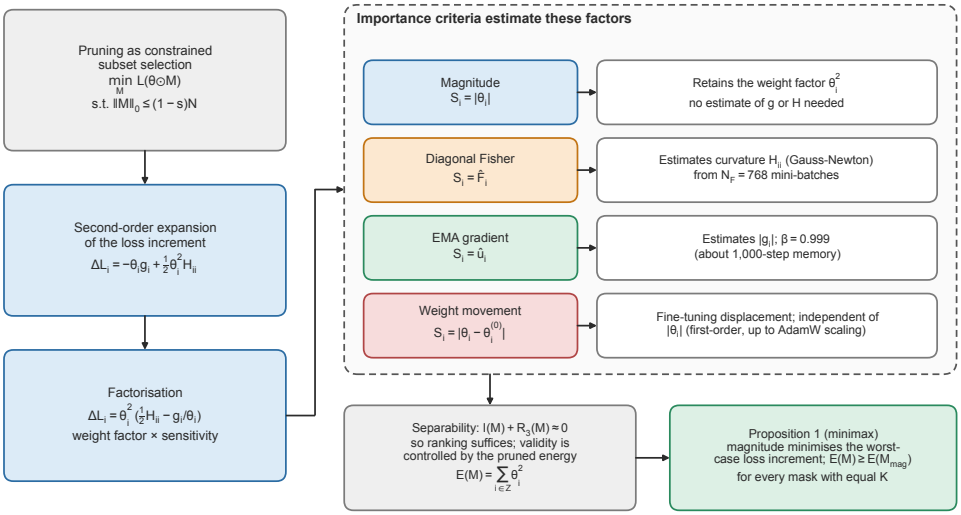


Fig. 8 Analytical framework for the importance criteria. Pruning is posed as constrained subset selection. The loss increment from removing parameter i is expanded to second order and factorised into a weight factor and a sensitivity factor. Magnitude retains the weight factor, whereas the Fisher, EMA gradient and movement criteria estimate sensitivity terms from the downstream objective. The separability approximation is controlled by the pruned energy $E(M)$, and magnitude pruning, which minimises $E(M)$, is minimax optimal (Proposition 1).

with the per-parameter increment and the interaction term

$$\Delta\mathcal{L}_i = -\theta_i g_i + \tfrac{1}{2}\theta_i^2 H_{ii}, \qquad \mathcal{I}(M) = \tfrac{1}{2}\sum_{\substack{i,j\in\mathcal{Z}\\ i\neq j}} \theta_i\theta_j H_{ij}. \tag{8}$$

If the Hessian is L_3-Lipschitz along the segment from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}\odot M$, the remainder satisfies $|R_3(M)| \le \frac{L_3}{6}E(M)^{3/2}$, where

$$E(M) = \|\delta\|_2^2 = \sum_{i\in\mathcal{Z}} \theta_i^2 \tag{9}$$

is the pruned weight energy. Writing $H_{\mathcal{Z}}^{\mathrm{off}}$ for the off-diagonal part of the principal submatrix of H indexed by $\mathcal{Z}$, the interaction term obeys

$$|\mathcal{I}(M)| \;\le\; \tfrac{1}{2}\left\|H_{\mathcal{Z}}^{\mathrm{off}}\right\|_2 E(M). \tag{10}$$

Saliency-based pruning rests on the separability assumption that $\mathcal{I}(M) + R_3(M)$ is negligible, so that the loss increase is the sum of independent per-parameter increments and the cardinality-constrained problem of Eq. (2) reduces to ranking. Equations (7)–(10) show what governs the validity of that assumption: the approximation error is

controlled by the pruned energy $E(M)$ and not by the number of pruned coordinates. It therefore deteriorates as sparsity grows and larger weights are removed.

Magnitude pruning is minimax optimal.

Magnitude pruning requires no estimate of g or H. The following result states precisely when it is the best available choice.

Proposition 1 (Minimax optimality of magnitude pruning)**.** *Fix $K \in \{0, \dots, N_{\mathcal{P}}\}$ and constants $G, \Lambda \geq 0$ with $G + \Lambda > 0$. Let*

$$\mathcal{U}(G, \Lambda) = \big\{(g, H) : \ g \in \mathbb{R}^N, \ \|g\|_2 \leq G, \ H = H^\top \in \mathbb{R}^{N \times N}, \ \|H\|_2 \leq \Lambda\big\}$$

and, for a mask M with $|\mathcal{Z}| = K$, let $Q(M; g, H) = g^\top \delta + \frac{1}{2}\delta^\top H \delta$ be the second-order loss increment. Then

$$\max_{(g,H) \in \mathcal{U}(G,\Lambda)} Q(M; g, H) \ = \ G\sqrt{E(M)} + \tfrac{1}{2}\Lambda E(M), \tag{11}$$

and the mask minimising this worst-case increment removes the K coordinates of $\mathcal{P}$ with smallest $|\theta_i|$. Moreover, every mask M with $|\mathcal{Z}| = K$ satisfies $E(M) \geq E(M^{\mathrm{mag}})$, where M^{mag} is the magnitude mask.

Proof By the Cauchy–Schwarz inequality $g^\top \delta \leq \|g\|_2 \|\delta\|_2 \leq G\sqrt{E(M)}$, with equality at $g = G\delta/\|\delta\|_2$ (any g if $\delta = 0$). Since $\|H\|_2 \leq \Lambda$, $\delta^\top H \delta \leq \Lambda \|\delta\|_2^2$, with equality at $H = \Lambda I$, which belongs to $\mathcal{U}(G, \Lambda)$. The two maximisers are chosen independently, so both bounds are attained simultaneously, which proves Eq. (11). The map $E \mapsto G\sqrt{E} + \frac{1}{2}\Lambda E$ is strictly increasing on $[0, \infty)$ because $G + \Lambda > 0$, so minimising the worst case over masks is equivalent to minimising $E(M) = \sum_{i \in \mathcal{Z}} \theta_i^2$ over all $\mathcal{Z} \subseteq \mathcal{P}$ with $|\mathcal{Z}| = K$. A sum of K squares drawn without replacement from $\{\theta_i^2\}_{i \in \mathcal{P}}$ is smallest when the K smallest squares are chosen, that is, when $\mathcal{Z}$ holds the K smallest values of $|\theta_i|$. The final claim is the same minimisation. □

Remark 1. If curvature is isotropic, $H_{ii} = \lambda$ for all $i \in \mathcal{Z}$, and the gradient vanishes, then $\Delta\mathcal{L}_i = \frac{1}{2}\lambda\theta_i^2$ and ranking by $\Delta\mathcal{L}_i$ coincides with ranking by $|\theta_i|$. Proposition 1 removes the need for either assumption: when the curvature structure is unknown, magnitude is the minimax choice. It also controls the interaction bound in Eq. (10), since no other mask with the same K has a smaller $E(M)$.

Remark 2 (Scope). The proposition concerns a second-order surrogate over an uncertainty set. It does not assert that magnitude minimises the realised loss, and when H is known or can be estimated reliably, informed criteria such as Optimal Brain Surgeon can do better. In a network with $N \approx 10^8$ parameters, the diagonal Fisher estimate from $N_F = 768$ mini-batches is noisy and discards all off-diagonal curvature, whereas magnitude carries no estimation error.

The remaining criteria as sensitivity and trajectory surrogates.

Factorising Eq. (8) as $\Delta\mathcal{L}_i = \theta_i^2\big(\frac{1}{2}H_{ii} - g_i/\theta_i\big)$ separates a weight factor from sensitivity factors. Magnitude retains the weight factor θ_i^2. The other three criteria estimate

sensitivity from data.

$$S_i^{\mathrm{mag}} = |\theta_i|, \tag{12}$$

$$S_i^{\mathrm{fisher}} = \hat{F}_i = \frac{1}{N_F}\sum_{b=1}^{N_F}\Big(\frac{\partial \mathcal{L}^{(b)}}{\partial \theta_i}\Big)^2, \tag{13}$$

$$S_i^{\mathrm{ema}} = \hat{u}_i^{(t_{\mathrm{end}})}, \quad u_i^{(t)} = \beta u_i^{(t-1)} + (1-\beta)\Big|\frac{\partial \mathcal{L}}{\partial \theta_i}\Big|^{(t)}, \quad \hat{u}_i^{(t)} = \frac{u_i^{(t)}}{1-\beta^t}, \tag{14}$$

$$S_i^{\mathrm{mov}} = \big|\theta_i^{(t_{\mathrm{end}})} - \theta_i^{(0)}\big|. \tag{15}$$

Here $\beta = 0.999$, giving an effective memory of $1/(1-\beta) = 1000$ steps, and t_{end} is the final step of the scoring trajectory. For the Fisher score, the Fisher information matrix coincides with the Gauss–Newton approximation to the Hessian for a negative-log-likelihood loss, and its diagonal is estimated by squared gradients. The scores in Eqs. (13) and (14) estimate H_{ii} and $|g_i|$ respectively, but, as implemented, omit the weight factor. The Taylor-consistent forms are

$$S_i^{\mathrm{OBD}} = \tfrac{1}{2}\theta_i^2 \hat{F}_i, \qquad S_i^{\mathrm{T1}} = |\theta_i|\,\hat{u}_i^{(t_{\mathrm{end}})}. \tag{16}$$

The movement score is independent of $|\theta_i|$ altogether. Under SGD the displacement equals $-\eta\sum_t g_i^{(t)}$, but under AdamW it is a sum of per-coordinate normalised steps, so Eq. (15) measures first-order importance only up to that normalisation.

5.5 Mask construction

For a single criterion the mask is obtained by thresholding the raw score at an order statistic within each prunable weight matrix. Let $\mathcal{P}_m \subseteq \mathcal{P}$ index the entries of the m-th matrix, let $K_s^{(m)} = \lfloor s\,|\mathcal{P}_m| \rfloor$, and let $q_s^{(m)}$ be the $K_s^{(m)}$-th smallest value of $\{S_i\}_{i\in\mathcal{P}_m}$. Then

$$M_i = \mathbb{1}[S_i > q_s^{(m)}] \;\; (i \in \mathcal{P}_m), \qquad M_i = 1 \;\; (i \notin \mathcal{P}), \tag{17}$$

with ties broken by parameter index. Empirical percentile ranks,

$$r_i^{(k)} = \frac{1}{N_{\mathcal{P}}}\sum_{j\in\mathcal{P}} \mathbb{1}\big[S_j^{(k)} \le S_i^{(k)}\big], \tag{18}$$

are monotone transformations and leave single-criterion masks unchanged. They are needed only to place heterogeneous scores on a common scale for rank fusion (Section 5.7). The threshold is therefore applied per matrix rather than as a single global threshold over $\mathcal{P}$, so that exactly 50% of the weights of every matrix are removed at $s = 0.5$; this, and the agreement of the magnitude mask with per-matrix thresholding of the dense weights, is verified in Supplementary Table S2. Per-matrix thresholds avoid comparing scores whose scales differ between layer types, which matters in particular for the classification head: its weights are trained from a random initialisation with a learning rate two orders of magnitude larger, so their magnitude

scale differs from the backbone. The minimisation of $E(M)$ in Proposition 1 separates across matrices, so for the magnitude criterion the result holds under these per-matrix removal counts, with $\mathcal{Z}$ containing the $K_s^{(m)}$ smallest values of $|\theta_i|$ in each matrix (proof in the Supplementary Note).

Mask diagnostics.

Two quantities are used to compare the masks of different criteria at the same sparsity (Fig. 4). First, the pruned energy fraction $E(M)/\|\boldsymbol{\theta}_{\mathcal{P}}\|_2^2$ (Eq. (9)) and the fraction of the largest 10% of weights (ranked by magnitude within each matrix) that are removed measure how much of the weight mass a mask discards; random pruning at 50% sparsity removes 50% of both. Second, the Jaccard index between the retained sets of two masks,

$$J(M, M') = \frac{|\{i \in \mathcal{P} : M_i = M'_i = 1\}|}{|\{i \in \mathcal{P} : M_i = 1 \text{ or } M'_i = 1\}|}, \tag{19}$$

measures their agreement. For two independent random masks that each retain a fraction ρ of the weights, the expected value is $\rho/(2-\rho)$, which equals $1/3$ at 50% sparsity and serves as the chance level.

5.6 Rewinding and masked optimisation

After mask construction, the surviving weights are reset to their initial values,

$$\theta_i \leftarrow \theta_i^{(0)} M_i. \tag{20}$$

For backbone coordinates $\boldsymbol{\theta}^{(0)}$ is the pre-trained checkpoint. For the classification head it is the random initialisation drawn before dense fine-tuning. Sparse fine-tuning then solves

$$\boldsymbol{\theta}^\star = \arg\min_{\boldsymbol{\theta} \in \mathcal{S}_M} \mathcal{L}(\boldsymbol{\theta}), \qquad \mathcal{S}_M = \{\boldsymbol{\theta} \in \mathbb{R}^N : \theta_i = 0 \ \ \forall i \text{ with } M_i = 0\}, \tag{21}$$

that is, optimisation restricted to the linear coordinate subspace of retained parameters. For plain gradient descent this is projected gradient descent, $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta\, M \odot \nabla\mathcal{L}(\boldsymbol{\theta}^{(t)})$. Because the optimiser used is AdamW, the following statement is needed to justify masking only the gradient.

Lemma 1 (Zero preservation under masked AdamW)**.** *Let* $\tilde{g}^{(t)} = M \odot \nabla\mathcal{L}(\boldsymbol{\theta}^{(t)})$ *be the masked gradient and let AdamW update its first and second moment estimates* $\mathbf{a}$ *and* $\mathbf{b}$ *and the parameters as*

$$\mathbf{a} \leftarrow \beta_1\mathbf{a} + (1-\beta_1)\tilde{g}, \quad \mathbf{b} \leftarrow \beta_2\mathbf{b} + (1-\beta_2)\tilde{g}^2, \quad \boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta\Big(\frac{\hat{\mathbf{a}}}{\sqrt{\hat{\mathbf{b}}} + \epsilon_{\text{Adam}}} + \lambda_{\text{wd}}\boldsymbol{\theta}\Big).$$

If $\theta_i^{(0)} = 0$ *and* $a_i = b_i = 0$ *at initialisation for every* i *with* $M_i = 0$*, then* $\theta_i^{(t)} = 0$ *for all* t *and all such* i*.*

Proof For a masked coordinate $\tilde{g}_i^{(t)} = 0$ at every step, so a_i and b_i remain zero. The adaptive term equals $0/(0 + \epsilon_{\text{Adam}}) = 0$, and the decoupled weight-decay term equals $\lambda_{\text{wd}}\theta_i = 0$ by induction. Hence $\theta_i^{(t+1)} = \theta_i^{(t)} = 0$. □

Masking the gradient therefore suffices to keep the iterate in $\mathcal{S}_M$ for the entire run. Because $\mathcal{S}_M$ is a linear subspace, any average of iterates also lies in $\mathcal{S}_M$, so stochastic weight averaging preserves the mask exactly.

Sparse fine-tuning uses learning-rate rewinding: the AdamW state and the cosine schedule with warm-up are re-initialised at step zero instead of continuing the dense schedule. A running average of end-of-epoch weights is accumulated on the CPU over the final 60% of the epoch budget $E_{\max}$,

$$\bar{\boldsymbol{\theta}}^{\text{SWA}} = \frac{1}{|\mathcal{T}|}\sum_{e\in\mathcal{T}} \boldsymbol{\theta}^{(e)}, \qquad \mathcal{T} = \{e : \ e \geq 0.4E_{\max}\}, \tag{22}$$

and the averaged and best single checkpoints are compared on validation AUROC, retaining the higher. The complete procedure is given in Algorithm 1.

5.7 Dynamic Rank Fusion (auxiliary diagnostic)

Dynamic Rank Fusion (DRF) is used only as a consistency check on the exploratory selection (Section 3.6) and is not a pruning criterion. For criterion k and sparsity $s \in \Sigma = \{10, 20, 30, 50, 60\}\%$ let $A_k(s)$ denote its validation AUROC expressed in percentage points. A blended performance score combines the sparsity-specific value with the mean over all sparsity levels,

$$\tilde{A}_k(s) = 0.6\, A_k(s) + 0.4\, \bar{A}_k, \qquad \bar{A}_k = \frac{1}{|\Sigma|}\sum_{s'\in\Sigma} A_k(s'), \tag{23}$$

and the fusion weights and hybrid score are

$$\alpha_k(s) = \frac{\exp\big(\tilde{A}_k(s)/T\big)}{\sum_j \exp\big(\tilde{A}_j(s)/T\big)}, \qquad S_i^{\text{hybrid}}(s) = \sum_k \alpha_k(s)\, r_i^{(k)}, \qquad T = 2.0. \tag{24}$$

Algorithm 1 RABANT: magnitude-based Lottery Ticket compression of the dense genomic transformer

Require: Pre-trained checkpoint $\boldsymbol{\theta}^{(0)}$; data $D_{\text{train}}, D_{\text{val}}$; target sparsity s; prunable set $\mathcal{P}$

Ensure: Sparse model $\boldsymbol{\theta}^{\star}$, binary mask M, non-zero count nnz

Dense fine-tuning. Attach the dual-sequence head, unfreeze all 22 layers, and train on D_{train} with ℓ_{focal} to obtain $\boldsymbol{\theta}$

Scoring. $S_i \leftarrow |\theta_i|$ for $i \in \mathcal{P}$ ▷ Eq. (12)

Mask. For each prunable matrix $\mathcal{P}_m$: $K_s^{(m)} \leftarrow \lfloor s|\mathcal{P}_m| \rfloor$; $q_s^{(m)} \leftarrow K_s^{(m)}$-th smallest S_i in $\mathcal{P}_m$; $M_i \leftarrow \mathbb{1}[S_i > q_s^{(m)}]$ for $i \in \mathcal{P}_m$; $M_i \leftarrow 1$ otherwise ▷ Eq. (17)

Rewind. $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^{(0)} \odot M$ ▷ Eq. (20)

Reset. Initialise AdamW state and cosine warm-up schedule at step 0

for $e = 1, \ldots, E_{\max}$ **until** early stopping (patience 5 on validation AUROC) **do**

  for each mini-batch $B \subset D_{\text{train}}$ **do**

    $\tilde{g} \leftarrow M \odot \nabla \ell_{\text{focal}}(\boldsymbol{\theta}; B)$; take one AdamW step with $\tilde{g}$ ▷ Lemma 1

  end for

  Evaluate validation AUROC; update $\boldsymbol{\theta}^{\text{best}}$

  if $e \geq \lceil 0.4 E_{\max} \rceil$ **then** update $\bar{\boldsymbol{\theta}}^{\text{SWA}}$ ▷ Eq. (22)

  end if

end for

$\boldsymbol{\theta}^{\star} \leftarrow \arg\max_{\boldsymbol{\theta} \in \{\boldsymbol{\theta}^{\text{best}}, \bar{\boldsymbol{\theta}}^{\text{SWA}}\}} \text{AUROC}_{\text{val}}(\boldsymbol{\theta})$

return $\boldsymbol{\theta}^{\star}$, M, nnz $= \|\boldsymbol{\theta}^{\star}\|_0$

The temperature is $T = 2.0$ AUROC percentage points, equivalently 0.02 when AUROC is expressed as a fraction. The weights reported in the Results (for example 0.902 for magnitude at $s = 50\%$ and 0.987 at $s = 60\%$) are reproduced by Eqs. (23)–(24) with these settings. DRF is only used as an auxiliary diagnostic, with no trainable parameters.

5.8 Native small-model baseline and training protocols

To separate sparse depth retention from raw parameter count, a natively smaller NT-v2-50M (12 layers, hidden dimension 512, 16 attention heads, approximately 50 M parameters) was fine-tuned with all twelve transformer layers unfrozen, using the same dataset, classification head, loss family, optimiser family and evaluation procedure as the dense NT-v2-100M. Table 8 lists the complete training configuration of the three models. The sparse stage of RABANT uses its own hyperparameters, selected from preliminary runs on the validation set, whereas the native NT-v2-50M used the same hyperparameters as the dense NT-v2-100M.

Implementation and seeding.

All experiments were run on a single NVIDIA A100 GPU using PyTorch. A master seed of 42 initialised the PyTorch CPU and CUDA random generators and the NumPy generator, and cuDNN benchmark mode was enabled because all inputs have a fixed length of 1,000 bp. In cross-validation, the seed of fold $f \in \{0, \ldots, 4\}$ was $42 + 1000f$

Table 8 Training configuration of the three confirmatory models.

Setting	Dense NT-v2-100M (uncompressed)	Native NT-v2-50M	**RABANT 48M (Ours)**
Sparsity	0%	0%	50%
Initialisation	Pre-trained	Pre-trained NT-v2-50M	Pre-trained, rewound ($\boldsymbol{\theta}^{(0)} \odot M$)
Trainable transformer layers	22	12	22 (masked)
Attention heads	16	16	16
Maximum epochs	20	20	15
Early-stopping patience (validation AUROC)	4	4	5
Backbone / head learning rate	5×10^{-6} / 5×10^{-4}	5×10^{-6} / 5×10^{-4}	4×10^{-6} / 4×10^{-4}
Weight decay	0.01	0.01	0.01
Warm-up and schedule	15% linear, cosine	15%, cosine	12%, cosine
Batch (physical × accumulation)	$32 \times 4 = 128$	effective 128	$8 \times 8 = 64$
Gradient clipping (norm)	1.0	1.0	1.0
Focal γ / smoothing ϵ	1.5 / 0.05	1.5 / 0.05	2.0 / 0.08
Weight averaging	None	None	SWA from $0.4E_{\max}$
Checkpoint rule	Best validation AUROC	Best validation AUROC	Higher of best and SWA

(42, 1042, 2042, 3042 and 4042), applied before model instantiation, so that each fold used an independently drawn Xavier-uniform initialisation of the classification-head weights, whereas the backbone weights were always loaded from the same pre-trained NT-v2 checkpoint. The complete five-fold procedure is given in Algorithm 2.

Algorithm 2 Confirmatory five-fold protocol

Require: Training partition D_{pool} (100,000 variants, Table 7), stratified by class into $F = 5$ folds of approximately 20,000 variants; fixed unseen test sets $D^{(1)}_{\text{test}}, D^{(2)}_{\text{test}}, D^{(3)}_{\text{test}}$ (ClinVar, dbSNP, cBioPortal+gnomAD); selected configuration (magnitude criterion, $s = 0.5$)

Ensure: Fold-level metrics for the dense, RABANT and native models

for $f = 0, \ldots, F-1$ **do**
  Set fold f as the validation set D^f_{va} and the remaining four folds as the training set D^f_{tr}
  Train the dense model on D^f_{tr}, early stopping on D^f_{va}, giving $\boldsymbol{\theta}^f_{\text{d}}$
  Run Algorithm 1 on the same split, reusing $\boldsymbol{\theta}^f_{\text{d}}$ in its first step, giving $\boldsymbol{\theta}^f_{\text{r}}$ and a fold-specific mask M^f
  Train the native model on $(D^f_{\text{tr}}, D^f_{\text{va}})$, giving $\boldsymbol{\theta}^f_{\text{n}}$
  for each model $q \in \{\text{d}, \text{r}, \text{n}\}$ and each source $j \in \{1, 2, 3\}$ **do**
    Score $D^{(j)}_{\text{test}}$ with $\boldsymbol{\theta}^f_q$ and compute metrics $\varrho^f_{q,j}$ at decision threshold $c = 0.5$
  end for
  $\varrho^f_q \leftarrow \frac{1}{3} \sum_j \varrho^f_{q,j}$ ▷ macro-average over sources
end for
return $\{\varrho^f_q\}$, the fold-level estimates compared in Section 5.10

5.9 Evaluation metrics

Let $\mathbf{o}_i \in \mathbb{R}^2$ denote the logits of variant i (Eq. (5)) and $p_i = \mathrm{softmax}(\mathbf{o}_i)_1$ the predicted probability of the pathogenic class. At decision threshold c a variant is called pathogenic when $p_i \geq c$. Threshold-dependent metrics are computed at $c = 0.5$ from the counts of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN), and are treated as secondary metrics:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \tag{25}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{26}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{27}$$

The threshold-free primary endpoint is the area under the receiver operating characteristic curve, estimated as the Mann–Whitney probability that a pathogenic variant outscores a benign one,

$$\widehat{\text{AUROC}} = \frac{1}{n^+ n^-} \sum_{i:\, y_i = 1} \sum_{j:\, y_j = 0} \Big(\mathbb{1}[p_i > p_j] + \tfrac{1}{2} \mathbb{1}[p_i = p_j] \Big), \tag{28}$$

where n^+ and n^- are the numbers of pathogenic and benign variants. Macro-averages are unweighted means over the three source-specific test sets.

5.10 Statistical analysis

Each model was trained on five folds, giving five fold-level estimates per metric, each a macro-average over the three unseen test sets (or the value on a single source for per-source comparisons). Two models were compared with Welch's unequal-variance t-test on these fold-level estimates. For sample means μ_1, μ_2, standard deviations σ_1, σ_2 and sample sizes $n_1 = n_2 = 5$, the test statistic and the Welch–Satterthwaite degrees of freedom are

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}, \qquad \nu = \frac{\left(\sigma_1^2/n_1 + \sigma_2^2/n_2\right)^2}{\frac{(\sigma_1^2/n_1)^2}{n_1 - 1} + \frac{(\sigma_2^2/n_2)^2}{n_2 - 1}}, \tag{29}$$

and the 95% confidence interval for the difference of means is

$$(\mu_1 - \mu_2) \;\pm\; t_{0.975,\,\nu} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}. \tag{30}$$

The effect size is Cohen's d,

$$d = \frac{\mu_1 - \mu_2}{s_p}, \qquad s_p = \sqrt{\frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n_1 + n_2 - 2}}. \tag{31}$$

Differences are reported as RABANT minus the comparator, p-values are two-sided with significance level 0.05, and no adjustment for multiple comparisons was applied. Because the five folds share training data, fold-level estimates are not strictly independent, and the resulting p-values and intervals should be regarded as approximate. Further statistical details are given in the Supplementary Methods.

Table 9 Summary of notation.

Symbol	Meaning
$\mathcal{D}$, m	Variant dataset and its size
$\mathbf{x}_{\text{ref}}, \mathbf{x}_{\text{alt}}, y$	Reference and alternate sequence windows; label (1 pathogenic, 0 benign)
d_h, n_{tok}	Hidden dimension (512); number of tokens in a sequence
$\boldsymbol{\theta}, \boldsymbol{\theta}^{(0)}$	Parameters after dense fine-tuning; initial parameters (pre-trained checkpoint, head random initialisation)
N, $\mathcal{P}$, $N_{\mathcal{P}}$	Total parameters; prunable index set; its size
M, $\mathcal{Z}$	Binary mask; pruned set $\{i \in \mathcal{P} : M_i = 0\}$
s, K_s, s_{tot}	Target sparsity on $\mathcal{P}$; number of removed coordinates $\lfloor sN_{\mathcal{P}} \rfloor$; overall sparsity K_s/N
$\mathcal{P}_m$, $K_s^{(m)}$, $q_s^{(m)}$	Entries of the m-th prunable matrix; removal count and score threshold applied to that matrix
g, H	Gradient and Hessian of the loss at $\boldsymbol{\theta}$
δ	Deletion vector $-\boldsymbol{\theta} \odot (\mathbf{1} - M)$
$\Delta\mathcal{L}_i$, $\mathcal{I}(M)$	Per-parameter loss increment; interaction term of Eq. (8)
$E(M)$	Pruned weight energy, Eq. (9)
$S_i^{(k)}$, $r_i^{(k)}$	Score and percentile rank of parameter i under criterion k
$\hat{F}_i$, N_F	Diagonal Fisher estimate; number of mini-batches used to compute it
$\hat{u}_i$, β, t_{end}	EMA gradient estimate; EMA decay (0.999); final step of the scoring trajectory
$\alpha_k(s)$, T, Σ	DRF weight of criterion k; DRF temperature (2.0 percentage points); set of sparsity levels
$J(M, M')$	Jaccard overlap of two masks
$\mathbf{o}$, p_i, c	Two-dimensional vector of unnormalised class logits; pathogenic-class probability; decision threshold
F, f	Number of folds (5); fold index
μ_j, σ_j, n_j, ν, d	Sample mean, standard deviation and size of model j; Welch degrees of freedom; Cohen's d

Supplementary information. Supplementary Information accompanies this paper.

Acknowledgements. The authors thank Dr. Divita Kulshrestha, who was affiliated with Accenture Solutions Pvt. Ltd. at the time of the discussions, for her inputs on the biological context of the study and the datasets used. Her discussions helped inform the interpretation of the biological aspects of this research. The authors also acknowledge ClinVar, dbSNP, gnomAD and cBioPortal for maintaining public variant resources.

Declarations

- **Funding:** This research received no specific grant from any public, commercial or not-for-profit funding agency.
- **Competing interests:** The authors declare no competing interests.

- **Ethics approval and consent to participate:** Not applicable. This study used only publicly available, de-identified variant records and involved no human participants, human tissue or animals.
- **Consent for publication:** Not applicable.
- **Data availability:** The variant records analysed in this study derive from public resources: ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), dbSNP (https://www.ncbi.nlm.nih.gov/snp/), gnomAD v4.1 (https://gnomad.broadinstitute.org/) and cBioPortal (https://www.cbioportal.org/). Reference sequences were extracted from the GRCh38/hg38 human genome assembly (https://hgdownload.soe.ucsc.edu/goldenPath/hg38/). The pre-trained Nucleotide Transformer v2 checkpoints are available from the InstaDeepAI organisation on Hugging Face (https://huggingface.co/InstaDeepAI/nucleotide-transformer-v2-100m-multi-species). Additional data supporting the findings are included in the article and its Supplementary Information.
- **Materials availability:** Not applicable.
- **Code availability:** The code for training, pruning, evaluation and figure generation is available at https://github.com/anirudham986/dnai-gpu1/tree/main.
- **Author contributions:** A.M. conceived the study, developed the methodology and software, performed the experiments and the formal analysis, wrote the original draft and revised the manuscript. A.A. contributed to the investigation, validation and software, and reviewed and edited the manuscript. R.S. and K.B.N. supervised the work, administered the project, and reviewed and edited the manuscript. All authors reviewed the manuscript.

References

[1] Stephens, Z. D. *et al.* Big data: astronomical or genomical? *PLoS Biol.* **13**, e1002195 (2015).

[2] Richards, 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. *Genet. Med.* **17**, 405–424 (2015).

[3] Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).

[4] Harrison, S. M., Biesecker, L. G. & Rehm, H. L. Using ClinVar to improve public data on the significance of clinical variants. *Nucleic Acids Res.* **47**, D903–D911 (2019).

[5] Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

[6] Chen, S. *et al.* A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).

[7] Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, pl1 (2013).

[8] Cerami, E. *et al.* The cBio Cancer Genomics Portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).

[9] Zehir, A. *et al.* Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,945 patients. *Nat. Med.* **23**, 703–713 (2017).

[10] Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

[11] Price, W. N. & Cohen, I. G. Privacy in the age of medical big data. *Nat. Med.* **25**, 37–43 (2019).

[12] Sirugo, G., Williams, S. M. & Tishkoff, S. A. The missing diversity in human genetic studies. *Cell* **177**, 26–31 (2019).

[13] Frankle, J. & Carbin, M. The lottery ticket hypothesis: finding sparse, trainable neural networks. *ICLR* (2019).

[14] Ng, P. C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).

[15] Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).

[16] Schwarz, J. M. *et al.* MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods* **11**, 361–362 (2014).

[17] Shihab, H. A. *et al.* Predicting the functional consequences of cancer-associated amino acid substitutions. *Bioinformatics* **29**, 1504–1510 (2013).

[18] Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).

[19] Ioannidis, N. M. *et al.* REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016).

[20] Quang, D., Chen, Y. & Xie, X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31**, 761–763 (2015).

[21] Ionita-Laza, I. *et al.* A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.* **48**, 214–220 (2016).

[22] Sundaram, L. *et al.* Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* **50**, 1161–1170 (2018).

[23] Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).

[24] Radjasandirane, R., Diharce, J., Gelly, J. & de Brevern, A. G. Assessment of variant effect predictors unveils variants difficulty as a critical performance indicator. *Genomics* (2025).

[25] Wu, J., Gai, X. *et al.* Semantic embedding of variant effect annotations enables rapid and accurate pathogenicity prediction with VUS.Life. *Scientific Reports* (2026).

[26] Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).

[27] Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **26**, 990–999 (2016).

[28] Quang, D. & Xie, X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of dna sequences. *Nucleic Acids Res.* **44**, e107 (2016).

[29] Zhou, J. *et al.* Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.* **50**, 1171–1179 (2018).

[30] Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).

[31] Ji, Y. *et al.* DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* **37**, 2112–2120 (2021).

[32] Zhou, Z. *et al.* DNABERT-2: efficient foundation model and benchmark for multi-species genome. *ICLR* (2024).

[33] Dalla-Torre, H. *et al.* Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nat. Methods* (2024).

[34] Vaswani, A. *et al.* Attention is all you need. *NeurIPS* **30** (2017).

[35] Nguyen, E. *et al.* HyenaDNA: long-range genomic sequence modeling at single nucleotide resolution. *NeurIPS* **36** (2024).

[36] Schiff, Y. *et al.* Caduceus: bi-directional equivariant long-range DNA sequence modeling. *arXiv:2403.03234* (2024).

[37] Benegas, G., Batra, S. S. & Song, Y. S. DNA language models are powerful predictors of genome-wide variant effects. *Proc. Natl. Acad. Sci. USA* **120**, e2311219120

(2023).

[38] Nguyen, E. *et al.* Sequence modeling and design from molecular to genome scale with Evo. *Science* **386**, eado9336 (2024).

[39] LeCun, Y., Denker, J. S. & Solla, S. A. Optimal brain damage. *NeurIPS* **2**, 598–605 (1990).

[40] Hassibi, B. & Stork, D. G. Second order derivatives for network pruning: optimal brain surgeon. *NeurIPS* **5**, 164–171 (1993).

[41] Molchanov, P. *et al.* Pruning convolutional neural networks for resource efficient inference. *ICLR* (2017).

[42] Han, S., Mao, H. & Dally, W. J. Deep compression: compressing deep neural networks with pruning, trained quantization and Huffman coding. *ICLR* (2016).

[43] Frankle, J. *et al.* Linear mode connectivity and the lottery ticket hypothesis. *ICML* 3259–3269 (2020).

[44] Morcos, A. S. *et al.* One ticket to win them all: generalizing lottery ticket initializations across datasets and optimizers. *NeurIPS* **32** (2019).

[45] Chen, T. *et al.* The lottery ticket hypothesis for pre-trained BERT networks. *NeurIPS* **33**, 15834–15846 (2020).

[46] Zhang, Y. *et al.* Lottery jackpots exist in pre-trained models. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 14662–14672 (2023).

[47] Sanh, V., Wolf, T. & Rush, A. M. Movement pruning: adaptive sparsity by fine-tuning. *NeurIPS* **33** (2020).

[48] Frantar, E. & Alistarh, D. SparseGPT: massive language models can be accurately pruned in one-shot. *ICML* (2023).

[49] Sun, M. *et al.* A simple and effective pruning approach for large language models. *ICLR* (2024).

[50] Hoefler, T. *et al.* Sparsity in deep learning: pruning and growth for efficient inference and training in neural networks. *J. Mach. Learn. Res.* **22**, 1–124 (2021).

[51] Whalen, S. *et al.* Navigating the pitfalls of applying machine learning in genomics. *Nat. Rev. Genet.* **23**, 169–181 (2022).

[52] Lin, T.-Y. *et al.* Focal loss for dense object detection. *ICCV* 2980–2988 (2017).

[53] Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. *ICLR* (2019).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- RABANTSRSupplementary.zip