
Evaluating the Ranking Layer of Rare-Disease Decision Support: Bayesian Uncertainty Under Temporal and Patient Shift

Bonaventure F. P. Dossou^{1,3*} and Mame Diarra Toure^{2*}

¹School of Computer Science, McGill University

²Department of Mathematics and Statistics, McGill University

³Mila Quebec AI Institute

*Equal Contribution

Abstract

Whether a variant ranker transfers across patients, and whether its uncertainty flags unfamiliar genes, are questions that a single solved proband cannot answer. They are also properties of the ranking layer that a generative clinical assistant inherits and cannot verify on its own, because a fluent explanation can make an incorrect ranking appear more credible. We study this gap through the Rare Disease, Real Kid MVA Hackathon 2026 and three post hoc generalization benchmarks. A phenotype-guided whole-genome pipeline processed 5.08 million normalized variants and ranked the accepted compound-heterozygous *BUB1B* pair first without a disease-specific bonus, achieving 100 Rank Points and F-max 1.0. We compare deterministic baselines, deep ensembles, singular Bayesian neural networks (SBNNs), Bayes-by-Backprop (BBB), and wider Bayesian architectures. On 313,171 temporal ClinVar records, a wide rank-50 SBNN achieves the best simulated top-1 ranking (84.35%) and F-max (0.9235), while the ensemble gives the strongest unseen-gene OOD detection (AUROC 0.7485). On 719 deduplicated Phenopacket cases whose causal variants were absent from ClinVar 2023, HPO conditioning raises dense BBB top-1 from 34.1% to 58.9%, but provides no aggregate improvement in the subgroup whose causal genes are absent from the phenotype reference. On 275,739 VarPB spike-in trials across 107 real-genome exonic backgrounds, a predictor-augmented SBNN reaches 86.63% Top-100, compared with 84.82% for REVEL and 69.72% for AlphaMissense. These results motivate reporting molecular scores, model uncertainty and phenotype-reference coverage separately.

1 Introduction

Whole-genome sequencing can yield millions of variants, while a Mendelian diagnosis may depend on one variant or an inheritance-consistent pair. Prioritization must combine molecular consequence, population rarity, phenotype compatibility, inheritance, and sequencing validity as annotations and gene-disease knowledge evolve. Exomiser and LIRICAL integrate these evidence types, while MAVERICK and AI-MARRVEL use learned prioritization models [8, 17, 24, 27]. aiDIVA additionally combines evidence-based and learned scores with language-model refinement [2]. These systems motivate evaluating the reliability of an upstream ranker before a generative component explains or acts on its output. Random database splits can retain stable gene and annotation patterns, while a model’s confidence need not identify unfamiliar genes or unsupported patient hypotheses. We therefore distinguish an independently submitted genome-wide challenge result from post hoc tests

of temporal molecular ranking, patient transfer, and epistemic uncertainty. Our contribution is an empirical comparison of Bayesian posterior families and deterministic baselines, including a negative pair-learning result. We examine what each evaluation establishes and where biological evidence, phenotype coverage, and model uncertainty remain distinct. The external patient analysis also tests phenotype corruption and disease novelty. On a fixed 719-case candidate cohort, temporal-consensus top-1 changes from 59.81% with the original HPO terms to 58.55% after removing 40% of the terms. Performance remains much lower for genes and diseases absent from the phenotype reference (Appendix I, Figure 6).

We further evaluate full-background ranking in VarPB, crossing 2,577 causal variants with exonic backgrounds from 107 real genomes [7]. The original small SBNN reaches only 0.20% Top-100. A revised SBNN combines the molecular features with REVEL and AlphaMissense scores and learns pairwise ordering. It reaches 86.63% Top-100, compared with 84.82% for REVEL and 69.72% for AlphaMissense, while REVEL retains higher Top-1 and MRR. This result extends the evaluation beyond compact candidate sets; it tests integration of pretrained predictor scores, rather than an independent replacement for those predictors. The full-background results are shown in Figure 7; the revised-model pipeline is shown in Figure 8 in Appendix J.

2 Challenge and genome-wide diagnosis

Rare Disease, Real Kid MVA Hackathon 2026 Sage Bionetworks organized the Rare Disease, Real Kid MVA Hackathon 2026 using genomic and clinical data from a child with mosaic variegated aneuploidy (MVA), a rare chromosome-instability disorder [25, 26]. Track 1 asks participants to identify and rank causal variants or variant pairs from supplied FASTQ, VCF, and phenotype data. Automated scoring uses Rank Points and F-max, adapted from CAGI Rare Genomes methodology [28]. Our deterministic submission placed the accepted pair at candidate rank 1 and received 100 Rank Points and F-max 1.0. Learned adapters were subsequently evaluated using a local full-match and allele-set scoring audit.

Figure 1 shows the system and evaluation boundary; Figure 2 shows the candidate funnel and challenge scores. Table 1 gives the clinical evidence and strict shortlist. Scoring, biological background and processing details are retained in Appendices H.1, H.2, H.4 and H.6.

Table 1: Clinical evidence and the complete strict compound-heterozygous shortlist. Population frequencies were rechecked in gnomAD v4.1.1.

Allele	Consequence	Read QC	Population evidence	Clinical / computational evidence
chr15:40209701 T>G	p.Leu737Ter	DP 46; GQ 99; AB .543	120/1,614,020; 7.435×10^{-5} ; 0 hom.	AF ClinVar P/LP, 2 stars; CADD 38; LOFTEE HC pLoF
chr15:40220612 T>G	p.Asn1002Lys	DP 28; GQ 99; AB .464	1/1,614,226; AF 6.195×10^{-7} ; 0 hom.	Exact allele unclassified; same-residue VUS; CADD 26.4; REVEL .472; PolyPhen .997

Rank	Gene	Pair score	EPCR	Phase	Main evidence / limitation
1	<i>BUB1B</i>	.8458	.950624	unknown	pathogenic LoF + ultra-rare coding partner
2	<i>LRP4</i>	.7069	.827678	unknown	alternative hypothesis
3	<i>DNAH10</i>	.6957	.811000	unknown	alternative hypothesis
4	<i>FCGBP</i>	.6495	.730034	trans	read-backed trans
5	<i>PKD1</i>	.6450	.721057	trans	read-backed trans
6	<i>TXNDC11</i>	.6181	.664048	unknown	alternative hypothesis

The appendix provides the detailed robustness, alignment QC, annotation-resource and allele-record tables (Tables 5, 6, 7, 8, 9, 10, 11, 12). Dataset provenance and the complete adapter comparison are in Tables 13 and 14; Table 15 maps the report material to the paper.

3 Bayesian ranking and generalization methods

3.1 Initial pair-level Bayesian learning

We first tested whether learned pair scores could replace the explicit evidence rule. Phenopacket Store 0.1.27 supplied 9,739 solved cases, including 959 two-variant cases [6, 19]. After annotation

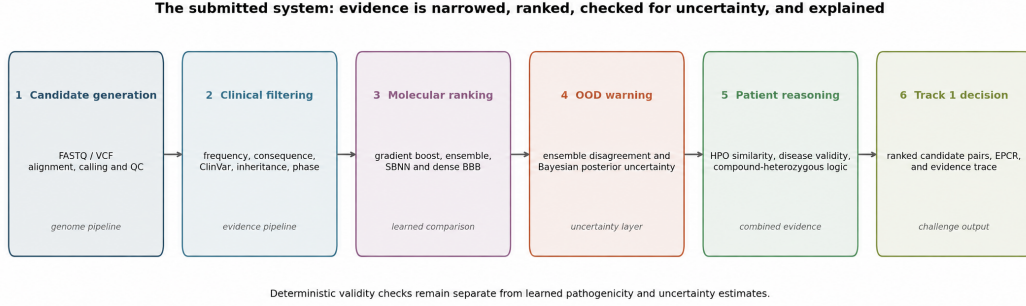


Figure 1: System and evaluation boundary. ‘Deterministic genome processing supplies candidates passing the specified filters; learned models rank molecular evidence and quantify uncertainty; phenotype and inheritance information provide additional context. Candidate validity, ranking, and uncertainty are tested as separate claims.

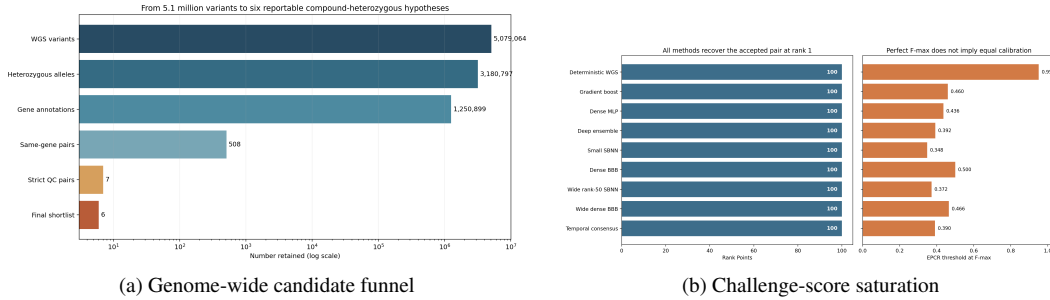


Figure 2: The upstream pipeline narrows 5.08 million variants to six reportable pairs (a). Once this biologically constrained candidate set is available, every transferred model recovers the accepted pair at rank 1, yet the EPCR threshold that isolates it varies substantially (b).

and decoy requirements, gene-, publication-, and case-disjoint splits contained 425 training, 44 validation, and 89 development cases; *BUB1B* was excluded. Each case contained one causal pair and 22 decoys. The models used 15 molecular/phenotype features exported from the upstream pipeline. See Appendix H.8.

The first Singular Bayesian Neural Network (SBNN) [31] used widths 32-16 and ranks (8, 8, 1) on the same features. This development experiment tests whether flexible learning improves the fixed evidence rule rather than assuming that it must.

3.2 Temporal molecular benchmark

The 26 December 2023 ClinVar release supplied 1,084,930 unambiguous benign/pathogenic labels. Balanced training contains 60,000 variants, including 30,000 pathogenic variants from 5,043 genes; validation contains 15,000 variants from 1,209 genes disjoint from the training genes. The temporal pool contains 626,287 Variation IDs absent in 2023 and present by 29 August 2026, including 117,147 pathogenic records across 14,715 genes. Even-ID temporal records are used for model selection, including the compact SBNN, dense BBB and wide SBNN searches and consensus weights; 313,171 odd-ID records form the reported evaluation partition. Historical gene-disjoint validation controls early stopping in the Bayesian training routines. All models receive 50 molecular features covering allele properties, population frequency and missingness, dbSNP, origin and review indicators, consequences, and three training-only gene-history features. Clinical-significance labels and disease names are not direct inputs. Classification metrics use all odd-ID records. Ranking uses 2,000 simulated cases, each containing one pathogenic and 99 benign temporal variants. OOD is defined as a gene absent from molecular training. The temporal task uses conditionally independent Bernoulli labels $y_i \in \{0, 1\}$ with $p_i(\theta) = \sigma(f_\theta(x_i))$, where θ includes all weights or factors and biases. Its likelihood is $p(D \mid \theta) = \prod_i p_i(\theta)^{y_i} [1 - p_i(\theta)]^{1-y_i}$, and the per-record data-fit loss

is $\ell_{\text{BCE}}(y, p) = -y \log p - (1 - y) \log(1 - p)$. The learned probabilities are thus used to rank candidates.

3.3 Bayesian models and posterior inference

We compare histogram gradient boosting, a dense 50-32-16-1 MLP, five independently trained MLPs [15], SBNNs [31], full-weight mean-field Bayes-by-backprop (BBB) [1], wider 200-100 architectures, and a consensus. Even-ID temporal selection by average precision chose probability averaging weights of 0.30 for gradient boosting, 0.20 for the ensemble, and 0.50 for the compact SBNN: $p_{\text{cons}} = 0.30p_{\text{GB}} + 0.20p_{\text{ens}} + 0.50p_{\text{SBNN}}$. Here each component is the predicted pathogenic-class probability. For a layer with $W \in \mathbb{R}^{m \times n}$, the SBNN parameterizes the weight matrix through rank- r latent factors,

$$W = T(A, B) = AB^\top, \quad A \in \mathbb{R}^{m \times r}, \quad B \in \mathbb{R}^{n \times r}. \quad (1)$$

Here m and n are the input and output widths, respectively, and the row-vector layer map is $x \mapsto xW + b$ with $b \in \mathbb{R}^n$. We require $1 \leq r \leq \min(m, n)$. See Appendix H.9.

The compact SBNN search covered 375 combinations of rank, KL weight, factor posterior-scale initialization multiplier, and broad-component prior scale σ_1 , while dense BBB used 75 configurations. Wider SBNNs considered ranks 8, 16, 32, and 50, as well as the maximum admissible factor rank. The selected wide SBNN has architecture 50–200–100–1 and layerwise factor ranks (50, 50, 1). Only its middle layer is strictly rank constrained relative to its maximum matrix rank. It contains 55,804 variational scalars, compared with 60,802 for matched wide dense BBB, a reduction of 8.22%. Random and SVD-based mean initializations were evaluated. Appendix D reports the complete hyperparameter grids, feature definitions, and benchmark construction.

3.4 Posterior prediction and uncertainty

For a fitted distribution $q_\phi(\theta)$ over all model parameters, the approximate posterior predictive distribution is $p_q(y | x) = \mathbb{E}_{q_\phi}[p(y | x, \theta)]$. We estimate it by $\bar{p}_M(y | x) = M^{-1} \sum_{m=1}^M p(y | x, \theta^{(m)})$ using independent draws $\theta^{(m)} \sim q_\phi$. For conditional logit, prediction instead integrates $p(y_i = j | X_i, w)$ over the Laplace approximation, conditioning on the complete candidate set X_i . A five-member deep ensemble averages five independently fitted predictive distributions; its members are not assumed to be posterior draws.

For a categorical distribution p , define $H(p) = -\sum_y p(y) \log p(y)$ in nats, with $0 \log 0 = 0$. The mutual information under the fitted distribution is

$$I_q(Y; \theta | x) = H(p_q(\cdot | x)) - \mathbb{E}_{q_\phi}[H(p(\cdot | x, \theta))]. \quad (2)$$

Writing $p_m(\cdot | x) = p(\cdot | x, \theta^{(m)})$, we use $\hat{I}_{q, M}(x) = H(\bar{p}_M(\cdot | x)) - \frac{1}{M} \sum_{m=1}^M H(p_m(\cdot | x))$. This is a finite-sample estimator under the fitted distribution, not exact mutual information under the true posterior. For an ensemble, the analogous expression measures disagreement with the model-member index. Under independent sampling, the Monte Carlo estimator converges as M increases; increasing M does not repair model misspecification or a deficient variational approximation. See Appendix H.10 and H.11.

3.5 External patient-level benchmark with a ClinVar temporal variant split

We retain Phenopacket cases with one or two exact GRCh38 causal variants only when every causal coordinate was absent from raw 2023 ClinVar and pathogenic in 2026. Deduplication by exact causal-variant signature leaves 719 retained cases from 265 genes: 655 single-variant and 64 two-variant cases. Each case contains one causal hypothesis and nine temporally new benign hypotheses, matched by type/consequence where possible. For a two-allele hypothesis the implemented molecular aggregation is the geometric mean $s_{\text{pair}}(v_1, v_2) = \sqrt{\bar{p}(v_1)\bar{p}(v_2)}$, which penalizes a which reduces the score relative to the stronger allele and penalizes imbalance relative to arithmetic averaging; it is not a joint pathogenicity probability. For this patient benchmark, the three molecular gene-history features are set to the implementation’s neutral values (0, 0.5, 1) before applying the fitted scaler; the molecular models are not retrained. The HPO layer uses the 9 October 2023 ontology. The reference pool comprises 2,231 retained cases whose causal variants were all present in the raw 2023

ClinVar snapshot. For HPO weight selection, 1,388 cases form the phenotype reference and 843 form the weight-selection set. After selecting the combination parameters, the evaluation phenotype matcher is built from all 2,231 reference-pool cases, recomputing its information content on that pool. Case publication and diagnosis dates are not used as inclusion criteria; the temporal boundary concerns causal-variant presence in ClinVar, not availability of case phenotypes before 2023. The phenotype score is a normalized symmetric best-match Resnik score, maximized over reference cases for the candidate gene[22]. We combine the molecular score and phenotype score by adding β times the latter to the molecular logit, using a selected constant c_{miss} for missing gene scores. The weight β and missing-score constant c_{miss} are selected separately for each molecular model on the weight-selection cases; the 719 evaluation cases are not used in that selection. This transformation is used for ranking and is not asserted to be a calibrated posterior conditional on HPO evidence. Five runs resample benign decoys in both the weight-selection and evaluation sets and reselect HPO combination parameters on the weight-selection cases. Their variation reflects candidate composition and parameter reselection, not independent patient cohorts or repeated molecular-model training.

4 Results

4.1 Challenge saturation and negative pair learning

All twelve learned adapters recovered the accepted pair at rank 1 with 100 Rank Points and F-max 1.0 in the local audit (Appendix D.5). With only six strict hypotheses, this agreement cannot distinguish transferability, and threshold-maximized F-max does not establish calibration. On the separate 89-case development benchmark, the fixed evidence rule achieved 70.8% top-1 and 83.1% top-5, compared with 58.4%/87.6% for conditional logit and 56.2%/87.6% for the first SBNN, using stable-sort ties. Learned models improved shortlist retention but reduced first-position accuracy. The experiment does not isolate the cause and does not justify replacing the fixed evidence rule with these learners.

4.2 Temporal ranking and unseen-gene detection

Table 2 shows that strong molecular classification and useful shift detection select different models. Gradient boosting already reaches 81.50% simulated top-1 and PR-AUC 0.9674. Wide rank-50 SBNN improves top-1 by 2.85 points, F-max by 0.0205, and NLL from 0.1924 to 0.1232. Its 84.35% top-1 exceeds matched wide dense BBB by 3.00 points, although the latter has slightly higher top-5. Width alone therefore does not explain the ranking gain; factorization, regularization, and optimization remain plausible contributors. The consensus reaches 83.65% top-1 but does not surpass the wide SBNN. The ensemble has lower top-1 (80.25%) but the strongest unseen-gene OOD AUROC (0.7485), compared with 0.6284 for wide SBNN. This measures discrimination under the specified shift, not a calibrated probability that a gene is unfamiliar. The compact SBNN sweep raised OOD AUROC from 0.6020 to 0.6231 with ranks (8, 4, 1), KL weight $\lambda = 0.01$, factor posterior-scale initialization multiplier 0.1, and broad-component prior scale $\sigma_1 = 0.3$. Configurations exceeding 0.70 by degrading classification toward chance were rejected: uncertainty quality cannot be optimized independently of task performance. Dense BBB selected $\rho_{\text{init}} = -4$, KL weight $\lambda = 0.1$, and broad-component prior scale $\sigma_1 = 0.3$, using 4,354 variational scalars versus 10,885 fitted ensemble weights. Wide rank-50 SBNN achieved 99.12% balanced accuracy at 80% coverage, but its OOD discrimination remained weaker than the ensemble. Figure 3 shows the corresponding results.

4.3 Patient transfer and phenotype coverage

The external benchmark changes the unit of evaluation from individual molecular records to patient hypotheses. Molecular-only top-1 ranges from 29.2% to 34.2%, above the 10% random baseline but far below the challenge result. The tasks differ in case mix and candidate construction, so the decrease is not a contradiction. Adding pre-cutoff HPO evidence improves every model by 23.1-26.6 points (Table 3). Dense BBB reaches $58.9 \pm 1.0\%$ overall; wide dense BBB leads the 64 two-variant cases at 68.1%, and the consensus leads HPO top-5 at 88.6%. Available phenotype evidence contributes more to this patient task than the modest differences among molecular architectures. Figure 4 shows the corresponding results. Figure 5 shows the corresponding results.

Table 2: Odd-ID temporal ClinVar holdout. Ranking simulations contain one pathogenic and 99 benign variants. OOD AUROC tests detection of genes absent from 2023 molecular training. Top- k uses the midrank convention defined in the text. PR-AUC denotes non-interpolated average precision, computed using average_precision_score.

Model	PR-AUC	F-max	NLL	Top-1	Top-5	OOD AUC
Gradient boost	.9674	.9030	.1924	81.50	94.05	-
Deep ensemble	.9688	.9095	.1245	80.25	95.00	.7485
Small SBNN	.9691	.9173	.1332	82.10	94.30	.6020 [†]
Consensus	.9731	.9181	.1358	83.65	94.80	-
Dense BBB	.9729	.9202	.1245	83.00	94.55	.6665
Wide rank-50 SBNN	.9736	.9235	.1232	84.35	94.85	.6284
Wide dense BBB	.9718	.9200	.1276	81.35	95.05	.6567

[†]The compact-SBNN OOD value was computed on the full temporal set and is context rather than a perfectly matched odd-ID estimate.

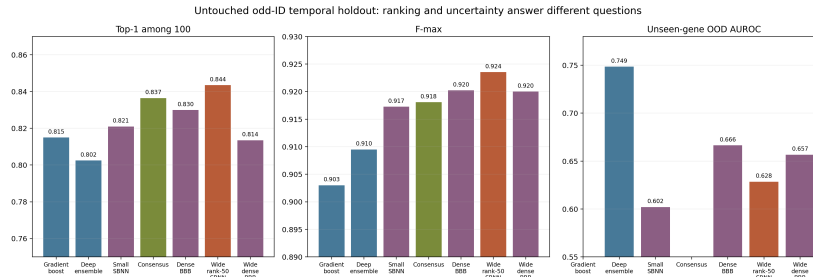


Figure 3: Temporal ranking and unseen-gene uncertainty answer different questions. The wide rank-50 SBNN gives the best Top-1 and F-max; independently trained ensemble members give the strongest OOD separation.

Table 3: Patient-level ranking under a ClinVar temporal variant split over five decoy-resampling runs with HPO parameter reselection. Values are mean percentages; $\pm$ denotes sample SD across runs. Gain is in percentage points and is computed from unrounded means. Top- k uses the midrank convention defined in the text.

Model	Molecular	+ HPO	Gain	HPO Top-5	Two-var. Top-1
Gradient boost	30.6 $\pm$ 1.3	56.5 $\pm$ 0.6	+25.9	86.5	66.2
Dense MLP	29.2 $\pm$ 0.8	55.9 $\pm$ 2.0	+26.6	84.3	65.9
Deep ensemble	30.6 $\pm$ 0.7	54.2 $\pm$ 1.9	+23.6	85.0	64.1
Small SBNN	32.9 $\pm$ 1.0	55.9 $\pm$ 1.9	+23.1	85.3	61.6
Dense BBB	34.1 $\pm$ 1.0	58.9 $\pm$ 1.0	+24.8	88.5	65.6
Wide rank-50 SBNN	32.8 $\pm$ 1.2	58.2 $\pm$ 1.1	+25.4	86.6	64.7
Wide dense BBB	34.2 $\pm$ 1.4	58.6 $\pm$ 1.4	+24.4	85.9	68.1
Temporal consensus	34.0 $\pm$ 0.5	57.3 $\pm$ 1.7	+23.3	88.6	64.4

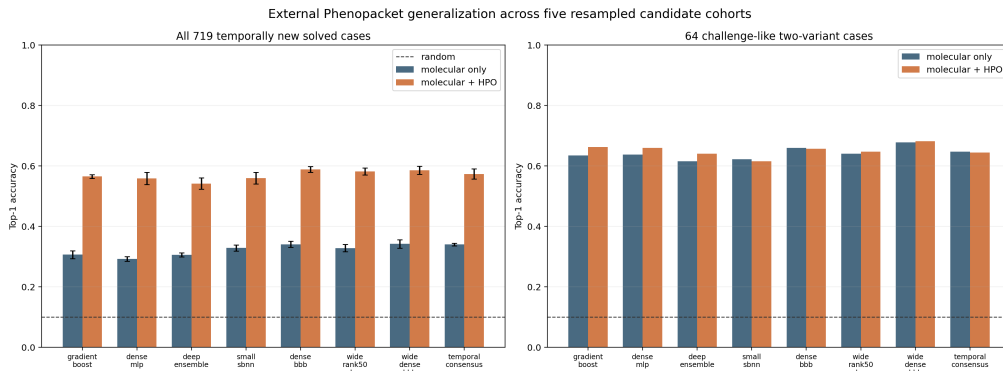
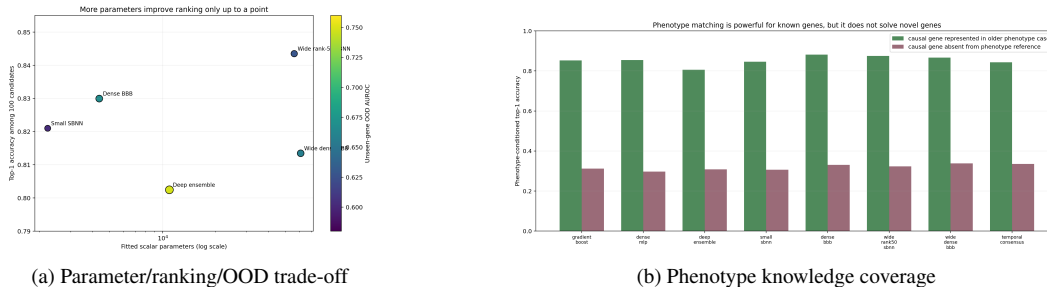


Figure 4: External patient-level ranking. Error bars in the left panel are SDs over five benign-decoy cohorts. The right panel isolates the 64 two-variant cases closest to the challenge structure.



(a) Parameter/ranking/OOD trade-off (b) Phenotype knowledge coverage
Figure 5: More Bayesian capacity improves temporal ranking only up to a point and does not reproduce ensemble OOD behavior (a). Parameter counts denote fitted scalar parameters, not measured runtime or memory use. Phenotype matching is powerful for genes represented in older cases but remains near molecular-only performance for unseen genes (b).

The gain is concentrated in patients with historical gene-phenotype evidence. For dense BBB, known-gene top-1 rises from 34.7% to 88.1% across 337 cases, while unseen-gene performance changes from 33.6% to 33.1% across 382 cases. The missing-gene prior limits the penalty for absent evidence but cannot create a missing association. Aggregate performance therefore depends on knowledge coverage, not just the molecular model. Relative to molecular-only ranking, HPO-augmented dense BBB gains 175–194 correct Top-1 cases and loses 2–6 per decoy-resampling run. Two-sided exact McNemar p values range from approximately 1.79×10^{-52} to 3.03×10^{-44} across the 5 runs. These tests compare paired Top-1 outcomes within each run and assume independent case-level discordances.

4.4 Uncertainty behavior

The mean across-draw standard deviation of predicted pathogenic-class probability rises from 0.00157 to 0.00403 between molecular-training-represented and molecular-training-unseen genes for the tuned SBNN, and from 0.01179 to 0.03569 for the ensemble. Both respond to shift, but absolute uncertainty magnitudes are not directly comparable across differently parameterized predictors, so OOD AUROC is more directly relevant to the stated detection task. The wide-SBNN selective result shows high balanced accuracy on retained examples at 80% coverage, not that the retained cases are clinically safe or that the same threshold transfers to other populations. NLL, ranking, OOD discrimination, and selective accuracy must remain separate evaluation targets.

4.5 Phenotype corruption and disease novelty

We froze each model’s HPO mixing coefficient after older-case validation and created three additional versions of the same 719-patient test: removal of 20% of observed HPO terms, removal of 40%, and addition of 20% unrelated terms. No coefficient was retuned on a corrupted test. On the fixed candidate cohort, temporal-consensus top-1 changed from 59.81% with the original HPO terms to 58.83%, 58.55%, and 59.39%, respectively. Wide rank-50 SBNN changed from 57.44% to 57.16%, 56.33%, and 56.88%. Moderate phenotype corruption therefore causes limited degradation in this compact setting.

Disease novelty remains difficult. Among 390 test cases whose disease identifier was absent from the older reference, temporal consensus reached 34.36% top-1; among the 382 unseen-gene cases it reached 33.77%, compared with 89.32% on 337 seen-gene cases. The robustness result therefore does not erase the central limitation: ontology matching transfers well when older knowledge covers the causal gene or disease, but does not solve gene discovery.

The case-level evaluator records the causal hypothesis independently from the prediction table. If upstream filtering removes the truth, reciprocal rank is set to zero and end-to-end Top- k is a failure. In the current compact experiment, however, the causal hypothesis is inserted by construction, so the observed candidate recall of 1.0 is only a pipeline-integrity check. Figure 6 shows the fixed-cohort corruption results; Appendix I gives the full protocol.

4.6 Full-background generalization beyond Track 1

After achieving the maximum Track 1 score, we asked whether the trained ranking models could generalize beyond the challenge child to independent rare-disease cases. We therefore evaluated their ability to prioritize known causal variants in VarPB against exonic candidate backgrounds derived from 107 real genomes. Neither the challenge child nor the VarPB causal answers were used for model training. This experiment complements the compact Phenopacket analysis by testing the models in a substantially larger and more realistic candidate space, allowing us to distinguish success on one solved case from transferable variant-ranking performance. Table 4 reports the full-background comparison.

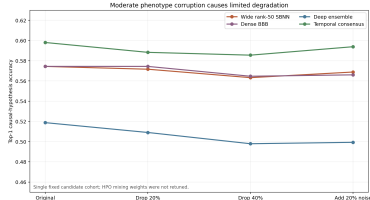


Figure 6: Top-1 robustness after fixed phenotype corruption. This is one fixed candidate cohort, and the HPO weights remain those selected on older validation cases.

Table 4: Locked VarPB v3 full-background generalization results. All methods retain complete score-file coverage.

Method	Median rank	Top-1	Top-20	Top-100	MRR
Augmented					
pairwise SBNN	14	7.26%	65.70%	86.63%	0.1596
REVEL	8.5	12.48%	65.90%	84.82%	0.2516
AlphaMissense	29	8.84%	44.24%	69.72%	0.1632
Original small					
SBNN	522	0.00%	0.00%	0.20%	0.0021
Wide rank-50 SBNN	1,562	0.00%	0.00%	0.08%	0.0011

The new result is competitive for clinically useful shortlist recall, not uniformly superior. Its Top-20 is within 0.20 percentage points of REVEL and its Top-100 is 1.81 points higher, while its Top-1 and MRR remain lower. It substantially exceeds AlphaMissense at Top-20 and Top-100 but is slightly weaker at Top-1 and MRR. Because REVEL and AlphaMissense are inputs, the contribution is Bayesian evidence integration and uncertainty-aware ranking, not independent rediscovery of their learned biology. VarPB still lacks native HPO and pedigree data and therefore remains a molecular-ranking benchmark rather than a complete clinical diagnostic cohort. Figure 7 shows the full-background ranking results. See Appendix J for the extended experiment.

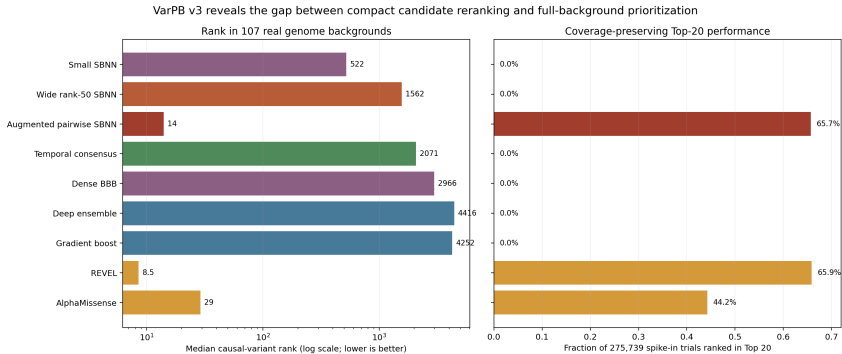


Figure 7: Full-background ranking across 275,739 spike-in trials. Lower median rank is better. The augmented SBNN essentially matches REVEL at Top-20 and exceeds it at Top-100, but REVEL remains stronger at the top of the list.

The model changes are detailed in Table 16 in Appendix J. Figure 8 in that appendix shows the training and evaluation workflow; Figure 9 reports fitted-parameter counts and available uncertainty outputs.

5 Discussion

Interpreting the Bayesian trade-off The model comparison is consistent with independently optimized ensemble members exploring more diverse functions than samples from the tested variational

posterior. This could explain why additional SBNN capacity improves ranking without producing equally strong unseen-gene warnings. It is not a general theorem about low-rank inference: prior specification, variational approximation, initialization, optimization, and compute budget are confounded across model families. The parameter-count advantage is consequently a representation-efficiency result, not proof of superior computational efficiency or posterior quality. The failed high-ODD configurations reveal a more general selection problem. A model can distinguish shift by becoming uninformative, so OOD AUROC alone is insufficient. A defensible selection procedure would constrain predictive performance and assess uncertainty on a separate shift-validation set before testing a coverage-risk policy. Our experiments establish this failure mode but do not validate a combined ranker-and-clinical-abstention system.

Evidence coverage and decision support The fitted variational distribution describes parameter uncertainty within its model and training objective; the reported uncertainty scores do not directly verify variant calls, phase or phenotype-reference coverage. A ranker cannot select a hypothesis omitted by candidate generation, and sampling its fitted parameters does not add missing reference associations. Unresolved phase may motivate segregation or long-range phasing, whereas absent phenotype-reference evidence may motivate additional clinical or functional investigation. The ensemble’s stronger molecular-unseen-gene AUROC motivates reporting a separately evaluated uncertainty signal alongside ranking scores. Separately, 382 evaluation cases had causal genes absent from the final phenotype reference. A proposed assistant could report that limitation explicitly without equating it with absence of all gene–disease knowledge. We did not evaluate a generative assistant, its explanations, or the clinical benefit of these proposed reporting behaviors.

6 Conclusion

A single solved proband is a meaningful result, but it does not indicate which ranker to trust on the next patient. We therefore paired the genome-wide challenge solution with a temporal molecular benchmark and an external patient-level benchmark, treating ranking quality, posterior uncertainty, and phenotype evidence as separate claims. The components diverge under that treatment: low-rank Bayesian posteriors give the best temporal ranking, independently trained ensembles give the strongest unfamiliar-gene detection, and phenotype conditioning helps substantially where a gene-disease relation already exists and almost not at all where it does not. No single model was best on all three. Posterior sampling does not add variants or phenotype associations absent from the modeled inputs. Keeping these signals distinct, rather than collapsing them into one score, is what allows a layer above the ranker to report what remains unresolved instead of asserting a diagnosis.

7 Limitations

Clinical evidence: The accepted *BUB1B* pair remains unphased across the two variants and p.Asn1002Lys is not independently established as pathogenic, so the leaderboard result is a strong causal hypothesis rather than a completed clinical diagnosis.

Benchmark scope: The 719-patient benchmark uses constructed ten-candidate sets, not complete genomes, and does not measure candidate-generation recall or clinical diagnostic sensitivity. Temporal database appearance reflects curation as well as biology; later annotations and phenotype associations may introduce hindsight despite historical training labels and ontology versions. The phenotype reference is assembled from the available Phenopacket release without a publication-date filter; a 2023 ontology does not establish that its case-level associations were available in 2023. The patient benchmark therefore does not simulate a fully time-frozen phenotype knowledge base. Deduplication retains one case per exact causal-variant signature and can collapse distinct patients sharing that signature. Architecture selection was not independently sealed from every reported evaluation. Balanced molecular training does not establish calibration at the prevalence of the temporal or clinical populations. Review, origin and database-presence features can encode ascertainment or curation patterns even when clinical-significance labels are excluded. Absence of a Variation ID does not by itself rule out equivalent historical alleles under a different representation. OOD is limited to unseen genes, and the two-variant subset is small. These limitations prevent claims of prospective deployment performance or universal Bayesian superiority.

Uncertainty and reproducibility: Appendices A to G retain the clinical audit, reproducibility workflow, model design, complete limitations, and a section-by-section coverage map.

References

- [1] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1613–1622. PMLR, 2015. URL <https://proceedings.mlr.press/v37/blundell115.html>.
- [2] Dominic Bocek, Lucia Laugwitz, Marc Sturm, Daniela Bezdán, Axel Gschwind, Tobias B. Haack, and Stephan Ossowski. aiDIVA—hybrid AI for rare disease diagnostics using evidence-based, machine learning and language models. *npj Genomic Medicine*, 11(1):49, 2026. doi: 10.1038/s41525-026-00611-x. URL <https://doi.org/10.1038/s41525-026-00611-x>.
- [3] Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, Rosalia G. Schneider, Andrew W. Senior, John Jumper, Demis Hassabis, Pushmeet Kohli, and Žiga Avsec. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, 381(6664):eadg7492, 2023. doi: 10.1126/science.adg7492. URL <https://doi.org/10.1126/science.adg7492>.
- [4] Clinical Genome Resource. BUB1B dosage sensitivity curation. ClinGen Gene Dosage Sensitivity Map, 2026. URL <https://search.clinicalgenome.org/kb/gene-dosage/HGNC:1149>. Recessive disease evidence and no evidence for haploinsufficiency; accessed September 2026.
- [5] Petr Danecek, James K. Bonfield, Jennifer Liddle, et al. Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2):giab008, 2021. doi: 10.1093/gigascience/giab008.
- [6] Daniel Danis et al. A corpus of GA4GH phenopackets: Case-level phenotyping for genomic diagnostics and discovery. *Human Genetics and Genomics Advances*, 6(1):100371, 2025. doi: 10.1016/j.xhgg.2024.100371. URL <https://doi.org/10.1016/j.xhgg.2024.100371>.
- [7] Matt Danzi. The variant prioritization benchmark (VarPB), 2023. URL <https://doi.org/10.5281/zenodo.10291114>.
- [8] Matt C. Danzi, Maïke F. Dohrn, Sarah Fazal, Danique Beijer, Adriana P. Rebelo, Vivian Cintra, and Stephan Züchner. Deep structured learning for variant prioritization in mendelian diseases. *Nature Communications*, 14(1):4167, 2023. doi: 10.1038/s41467-023-39306-7. URL <https://doi.org/10.1038/s41467-023-39306-7>.
- [9] Michael A. Gargano, Nicolas Matentzoglou, Ben Coleman, et al. The human phenotype ontology in 2024: phenotypes around the world. *Nucleic Acids Research*, 52(D1):D1333–D1346, 2024. doi: 10.1093/nar/gkad1005.
- [10] Genome Aggregation Database. gnomAD v4.1.1 variant records for 15-40209701-t-g and 15-40220612-t-g (GRCh38). Broad Institute gnomAD browser, 2026. Database records consulted for post-ranking clinical verification.
- [11] Sandra Hanks, Kim Coleman, Sarah Reid, Alberto Plaja, Helen Firth, David FitzPatrick, Alexa Kidd, Károly Méhes, Richard Nash, Nathaniel Robin, Nora Shannon, John Tolmie, John Swansbury, Alexandre Irrthum, Jenny Douglas, and Nazneen Rahman. Constitutional aneuploidy and cancer predisposition caused by biallelic mutations in BUB1B. *Nature Genetics*, 36(11):1159–1161, 2004. doi: 10.1038/ng1449. URL <https://doi.org/10.1038/ng1449>.
- [12] Nilah M. Ioannidis et al. REVEL: An ensemble method for predicting the pathogenicity of rare missense variants. *The American Journal of Human Genetics*, 99(4):877–885, 2016. doi: 10.1016/j.ajhg.2016.08.016. URL <https://doi.org/10.1016/j.ajhg.2016.08.016>.
- [13] Kishore Jaganathan et al. Predicting splicing from primary sequence with deep learning. *Cell*, 176(3):535–548.e24, 2019. doi: 10.1016/j.cell.2018.12.015. URL <https://doi.org/10.1016/j.cell.2018.12.015>.

- [14] Martin Kircher, Daniela M. Witten, Preti Jain, Brian J. O’Roak, Gregory M. Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3):310–315, 2014. doi: 10.1038/ng.2892. URL <https://doi.org/10.1038/ng.2892>.
- [15] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30*, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.
- [16] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009. doi: 10.1093/bioinformatics/btp324.
- [17] Dongxue Mao, Chaozhong Liu, Linhua Wang, Rami Al-Ouran, Cole Deisseroth, Sasidhar Pasupuleti, Seon Young Kim, Lucian Li, Jill A. Rosenfeld, Linyan Meng, Lindsay C. Burrage, Michael F. Wangler, Shinya Yamamoto, Undiagnosed Diseases Network, Michael Santana, Victor Perez, Priyank Shukla, Christine M. Eng, Brendan Lee, Bo Yuan, Fan Xia, Hugo J. Bellen, Pengfei Liu, and Zhandong Liu. AI-MARRVEL—a knowledge-driven AI system for diagnosing mendelian disorders. *NEJM AI*, 1(5):AIoa2300009, 2024. doi: 10.1056/AIoa2300009. URL <https://doi.org/10.1056/AIoa2300009>.
- [18] Marcel Martin, Peter Ebert, and Tobias Marschall. Read-based phasing and analysis of phased variants with WhatsHap. In *Haplotyping: Methods and Protocols*, volume 2590 of *Methods in Molecular Biology*, pages 127–138. Springer, 2023. doi: 10.1007/978-1-0716-2819-5_8.
- [19] Monarch Initiative. Phenopacket Store, release 0.1.27. GitHub release archive, 2026. URL <https://github.com/monarch-initiative/phenopacket-store/releases/tag/0.1.27>. Dataset version used in this study; accessed 7 September 2026.
- [20] National Center for Biotechnology Information. ClinVar record vcv000533901 for BUB1B c.2210t>g (p.leu737ter). NCBI ClinVar, 2026. Database record consulted for post-ranking clinical verification.
- [21] Kevin A. Pagel, Ryan Kim, Kevin Moad, et al. Integrated informatics analysis of cancer-related variants. *JCO Clinical Cancer Informatics*, 4:310–317, 2020. doi: 10.1200/CCI.19.00132.
- [22] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453, 1995.
- [23] Thomas Rio Frio, Josée Lavoie, Nancy Hamel, Felipe C. Geyer, Yaël B. Kushner, David J. Novak, Landon Wark, Cristian Capelli, Jorge S. Reis-Filho, Sabine Mai, Tomi Pastinen, Marc D. Tischkowitz, Victoria A. Marcus, and William D. Foulkes. Homozygous BUB1B mutation and susceptibility to gastrointestinal neoplasia. *New England Journal of Medicine*, 363(27):2628–2637, 2010. doi: 10.1056/NEJMoa1006565. URL <https://doi.org/10.1056/NEJMoa1006565>.
- [24] Peter N. Robinson, Vida Ravanmehr, Julius O. B. Jacobsen, Daniel Danis, Xingmin Aaron Zhang, Leigh C. Carmody, Michael A. Gargano, Courtney L. Thaxton, Guy Karlebach, Justin Reese, Manuel Holtgrewe, Sebastian Köhler, Julie A. McMurphy, Melissa A. Haendel, and Damian Smedley. Interpretable clinical genomics with a likelihood ratio paradigm. *The American Journal of Human Genetics*, 107(3):403–417, 2020. doi: 10.1016/j.ajhg.2020.06.021. URL <https://doi.org/10.1016/j.ajhg.2020.06.021>.
- [25] Sage Bionetworks. Rare disease, real kid: MVA hackathon 2026. Hugging Face Space, 2026. URL <https://huggingface.co/spaces/SageBio/rare-disease-real-kid-mva-hackathon-2026>. Challenge rules and scoring; accessed 7 September 2026.
- [26] Sage Bionetworks. Rare disease, real kid: MVA hackathon 2026 dataset. Hugging Face dataset repository, 2026. URL <https://huggingface.co/datasets/SageBio/mva-hackathon-2026-data>. Gated CC BY 4.0 dataset; accessed by an authorized participant.

- [27] Damian Smedley, Julius O. B. Jacobsen, Marten Jäger, Sebastian Köhler, Manuel Holtgrewe, Max Schubach, Enrico Siragusa, Tomasz Zemojtel, Orion J. Buske, Nicole L. Washington, William P. Bone, Melissa A. Haendel, and Peter N. Robinson. Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nature Protocols*, 10(12):2004–2015, 2015. doi: 10.1038/nprot.2015.124. URL <https://doi.org/10.1038/nprot.2015.124>.
- [28] Sarah L. Stenton et al. Critical assessment of variant prioritization methods for rare disease diagnosis within the Rare Genomes Project. *Human Genomics*, 18(1):44, 2024. doi: 10.1186/s40246-024-00604-w. URL <https://doi.org/10.1186/s40246-024-00604-w>.
- [29] Saskia J. E. Suijkerbuijk, Maria H. J. van Osch, Frank L. Bos, Sandra Hanks, Nazneen Rahman, and Geert J. P. L. Kops. Molecular causes for BUBR1 dysfunction in the human cancer predisposition syndrome mosaic variegated aneuploidy. *Cancer Research*, 70(12):4891–4900, 2010. doi: 10.1158/0008-5472.CAN-09-4319. URL <https://doi.org/10.1158/0008-5472.CAN-09-4319>.
- [30] Sean V. Tavtigian, Marc S. Greenblatt, Steven M. Harrison, Robert L. Nussbaum, Snehit A. Prabhu, Kenneth M. Boucher, and Leslie G. Biesecker. Modeling the ACMG/AMP variant classification guidelines as a bayesian classification framework. *Genetics in Medicine*, 20(9):1054–1060, 2018. doi: 10.1038/gim.2017.210. URL <https://doi.org/10.1038/gim.2017.210>.
- [31] Mame Diarra Toure and David A. Stephens. Singular bayesian neural networks. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL <https://arxiv.org/abs/2602.00387>. ICML 2026; preprint arXiv:2602.00387.

A Complete genome-wide audit and submission provenance

This appendix retains the full technical and clinical evidence trail from the original analysis report. It is deliberately more detailed than the nine-page main text. The challenge case used the complete supplied whole-genome VCF, four paired-end FASTQ lanes, and the supplied phenotype document from the gated Rare Disease, Real Kid MVA Hackathon resource [25, 26]. The reference build was GRCh38, with the exact reconstructed reference representation

GCA_000001405.15_GRCh38_no_alt_analysis_set_plus_hs38d1_maskedGRC_exclusions_v2_no_chr.fasta.

Validation against the VCF sequence dictionary checked 2,580 contigs against 2,580 FASTA contigs, with zero missing contigs and zero length mismatches.

A.1 Submission identity, discovery timeline, and scoring boundary

The organizer-confirmed submission was the genome-wide output generated before any disease-specific MVA reranking. The original report records a participant-prefixed repository filename for this artifact; that identifying prefix is withheld during double-blind review. Its contents were identical to the file uploaded to the leaderboard, with only the repository filename normalized. The accepted pair therefore emerged from the general genome-wide ranking rather than from a manually inserted *BUB1B* or MVA-gene bonus. A later MVA-specific reranker was used only for post-submission sensitivity and interpretation and is not treated as independent validation.

The conservative allele score used in the submitted analysis was

$$0.36 s_{\text{consequence}} + 0.26 s_{\text{rarity}} + 0.20 s_{\text{QC}} + 0.10 s_{\text{ClinVar}} + 0.08 s_{\text{predictor}}, \quad (3)$$

where the predictor composite used raw REVEL, normalized CADD PHRED, and raw SpliceAI delta when available. The final analysis explicitly used raw REVEL rather than REVEL rankscore. For the leading *BUB1B* pair, the gene prior was 0.0, the phenotype score was 0.6641, the pair score was 0.8458, and EPCR was 0.950624.

A.2 Challenge scoring definitions

The main text summarizes the evaluator in one sentence; the explicit definitions are as follows. Let G be the accepted allele set and S_i the allele set in submitted row i . The full-match rank is the first row with $S_i = G$; rank 1 earns 100 Rank Points, and partial-allele credit is possible. At each EPCR threshold t , the evaluator retains the predicted alleles and computes

$$P(t) = \frac{TP(t)}{TP(t) + FP(t)}, \quad R(t) = \frac{TP(t)}{TP(t) + FN(t)}, \quad (4)$$

$$F_1(t) = \frac{2P(t)R(t)}{P(t) + R(t)}, \quad F_{\max} = \max_t F_1(t). \quad (5)$$

Because $F_{\max}$ is maximized over thresholds, a value of 1.0 on a single proband records that some threshold isolates the accepted pair; it is not a statement about probability calibration.

Table 5: Exact pre-submission robustness sequence.

Stage	Main change	Effect on leading result
Canonical annotation join	Reconciled VCF and OpenCRAVAT allele representations	3,180,796 of 3,180,797 heterozygous alleles matched
Conservative predictor revision	Replaced REVEL rankscore with raw REVEL and separated CADD evidence	<i>BUB1B</i> remained rank 1
Primary-contig ranking	Moved unplaced and alternate-contig pairs to an audit tier	<i>BUB1B</i> remained rank 1
Strict germline QC	PASS, DP ≥ 10 , GQ ≥ 20 , allele balance 0.20–0.80, AF ≤ 0.005 when observed	<i>BUB1B</i> remained rank 1
Phase-aware exclusion	Removed a <i>ZNF619</i> pair with direct read-backed cis evidence	<i>BUB1B</i> remained rank 1

A.3 Alignment, variant, phasing, and annotation audits

All supplied reads were aligned without genome-wide downsampling or candidate-gene restriction using a BWA-compatible alignment path [16]; SAM/BAM and VCF utilities followed standard SAMtools/BCFtools conventions [5]. The custom alignment path did not preserve the SAM proper-pair (0x2) bit, but mate coordinates, R1/R2 identity, insert sizes, and paired-end relationships were retained and independently audited.

Table 6: Complete alignment and paired-end QC reported for the challenge genome.

Metric	Value	Metric	Value
Total reads	1,076,022,862	Primary reads	1,071,835,104
Primary mapped reads	1,067,608,257	Primary mapping rate	99.61%
Read 1	535,917,552	Read 2	535,917,552
Singletons	0.34%	Primary duplicates	137,812,223
Mate mapped, 1M-pair audit	99.651%	Same chromosome	98.662%
Non-zero template length	97.964%	Median template length	430 bp
Template-length IQR	325–558 bp		

Normalization produced 5,079,064 variants, including 3,180,797 heterozygous variants. WhatsHap read-backed phasing [18] phased 2,754,502 heterozygous variants (86.60%) into 439,579 phase sets; the largest contained 6,891 heterozygous variants. Phase was interpreted conservatively: same phase set plus opposite haplotypes was evidence for trans; same phase set plus the same haplotype was evidence for cis; different phase sets or unphased calls were recorded as unknown rather than cis. The two leading *BUB1B* variants were not connected by a read-backed phase block.

OpenCRAVAT [21] was run locally with ClinVar, gnomAD v2, gnomAD v3, CADD, REVEL, and SpliceAI. Canonical chromosome/position/ref/alt normalization reconciled allele representations across VCF and annotation tools. Of 3,180,797 heterozygous alleles, 3,180,796 matched functional annotations, leaving one unmatched allele; 1,250,899 candidate gene annotations were produced.

Table 7: Installed annotation resources recorded during the analysis.

Module	Installed release observed	Approximate local size
ClinVar	2026.07.01	1.7 GB
gnomAD	v2.1	17.4 GB
gnomAD3	v3	44.0 GB
REVEL	v4.3a resource	7.0 GB
SpliceAI	v4.0 resource	1.8 GB
CADD	v1.6	86.6 GB

A.4 Phenotype evidence and candidate construction

The complete phenotype constellation used for HPO-based scoring is the term set listed in the genome-wide processing subsection of the main text: rhabdomyosarcoma, nephrocalcinosis, short stature, failure to thrive, muscle atrophy, prematurity, growth restriction, and parental recurrent miscarriages. No single phenotype was used as a diagnostic shortcut. Candidate compound-heterozygous pairs were constructed from heterozygous variants affecting the same gene. There were 508 initial same-gene pairs: 131 on primary chromosomes and 377 on non-primary or alternate contigs retained for audit. Strict submission QC left seven primary-chromosome pairs; direct cis evidence removed *ZNF619*, yielding the six reportable pairs shown in Table 1.

The complete *BUB1B* heterozygous set was also reviewed internally. Only p.Leu737Ter and p.Asn1002Lys were strong coding candidates. Remaining heterozygous *BUB1B* variants were intronic and had substantially weaker functional evidence; several also had population frequencies incompatible with an ultra-rare recessive hypothesis. Thus p.Asn1002Lys was the strongest plausible second *BUB1B* allele in the patient despite its unresolved pathogenicity.

Table 8: Robustness of the leading pair to scoring and filtering revisions.

Analysis	<i>BUB1B</i> pair score	Rank
Initial canonical-join ranking	0.8541	1
Conservative raw-REVEL ranking	0.8458	1
Strict primary-contig/QC ranking used for Submission 1	0.8458	1

The exact submitted primary-hypothesis schema was:

```

PROBAND01
chr15:40209701 T>G
chr15:40220612 T>G
BUB1B
EPCR = 0.950624
finding_type = primary

```

B Complete BUB1B clinical evidence audit

The pair is clinically asymmetric: p.Leu737Ter has an established loss-of-function interpretation, whereas p.Asn1002Lys remains the weaker link. The paper therefore keeps sequencing validity, molecular prediction, population rarity, clinical classification, and phase as separate evidence classes.

B.1 p.Leu737Ter: sequence, annotation, population, and ClinVar detail

The first allele is chr15:40209701 T>G, NM_001211.6:c.2210T>G, p.Leu737Ter, on the MANE Select transcript ENST00000287598.11. DP was 46, GQ 99, and allele balance 0.543. CADD PHRED was 38; SpliceAI delta was 0.0300, Pangolin -0.0300 , and phyloP 2.37. LOFTEE classified the variant as a high-confidence predicted loss-of-function allele on both annotated transcripts. The raw pileup contained REF forward/reverse counts 13/9 and ALT 14/12; ALT median MAPQ was 60 and median base quality 37.

Identifiers retained in the audit were dbSNP rs759242053, ClinVar Variation ID 533901 (VCV000533901.9), ClinGen Allele Registry CA7475986, and canonical SPDI

NC_000015.10:40209700:T>G. VEP on the MANE Select transcript reproduced c.2210T>G, p.Leu737Ter, and the CADD PHRED 38.0 annotation.

Table 9: Direct gnomAD v4.1.1 verification for p.Leu737Ter.

Callset	Filter	Allele count	Allele number	Allele frequency
Exomes	PASS	115	1,461,846	7.867×10^{-5}
Genomes	PASS	5	152,174	3.286×10^{-5}
Total	–	120	1,614,020	7.435×10^{-5}

No homozygotes were observed in 1,614,020 alleles. The gnomAD grpmax filtering allele frequency at 95% confidence was 8.341×10^{-5} ; the highest observed ancestry-group frequency was 9.830×10^{-5} . Both callsets carried PASS filters and the variant page showed no quality warning in the audit. These values are compatible with a rare recessive carrier allele. Using the global allele frequency as a rough carrier-frequency scale gives an expected biallelic frequency on the order of 5×10^{-9} , consistent with an exceptionally rare syndrome; the report nevertheless treats rarity as supportive rather than as proof of pathogenicity. A current Labcorp ClinVar submission note separately cautions that population-frequency quality metrics at this position were considered unreliable in its evaluation. Because that submitter comment and the direct v4.1.1 browser audit are not the same evidence source, we retain the directly recorded counts but do not use population frequency as an independent pathogenicity argument.

The aggregate ClinVar classification was Pathogenic/Likely pathogenic, criteria provided, multiple submitters, no conflicts (two stars), last evaluated 9 October 2024 [20]. The disease annotation includes Mosaic variegated aneuploidy syndrome 1.

Table 10: ClinVar submissions underlying the p.Leu737Ter aggregate.

Submitter	Classification	Condition / accession	Last evaluated
Labcorp Genetics (formerly Invitae)	Pathogenic	MVA syndrome 1; SCV000762865.8	9 Oct 2024
GeneDx	Likely pathogenic	not provided; SCV004031025.1	17 Aug 2023

The corresponding RCV accessions were RCV000641226.9 and RCV003324778.1. The variant first appeared in ClinVar on 28 May 2018; the report recorded the most recent submission as 25 February 2025. Both submitters cite Hanks et al. (PMID 15475955) [11] and Rio Frio et al. (PMID 21190457) [23] as mechanism-level evidence for pathogenic BUB1B loss of function rather than as reports of this exact allele. ClinGen dosage curation records the recessive disease association and no evidence for haploinsufficiency, consistent with the compound-heterozygous disease model [4].

Regional genomic constraint in the surrounding 1 kb (15:40209000–40210000) was $Z = 1.79$ with observed/expected 0.88. The report therefore treats local constraint as mildly supportive at most, not as independent pathogenic evidence.

B.2 p.Asn1002Lys: sequence, annotation, population, and ClinVar detail

The second allele is chr15:40220612 T>G, NM_001211.6:c.3006T>G, p.Asn1002Lys, missense. DP was 28, GQ 99, and allele balance 0.464. CADD PHRED was 26.4, REVEL 0.472, PolyPhen max 0.997, SpliceAI delta 0.0200, Pangolin 0.0100, and phyloP 3.69. The raw pileup contained REF forward/reverse counts 5/11 and ALT 5/8; ALT median MAPQ was 60 and median base quality 37. VEP on the MANE Select transcript ENST00000287598.11 independently reproduced c.3006T>G and p.Asn1002Lys together with the reported CADD and REVEL values.

The installed gnomAD v2/v3 modules returned no population frequency. Direct gnomAD v4.1.1 verification showed that the allele is not absent from population data but is a singleton introduced by the larger v4 exome callset:

No homozygotes were observed. The highest ancestry-group frequency was 8.474×10^{-7} . The single carrier was male (XY). A grpmax filtering allele frequency at 95% confidence was not computable

Table 11: Direct gnomAD v4.1.1 verification for p.Asn1002Lys.

Callset	Filter	Allele count	Allele number	Allele frequency
Exomes	PASS	1	1,461,878	6.841×10^{-7}
Genomes	no variant	0	152,348	0
Total	–	1	1,614,226	6.195×10^{-7}

for a singleton. The large allele number also shows that the site was callable across essentially the whole cohort. The difference between the older ranking annotation and v4.1.1 is therefore a database-version effect rather than a representation or coverage failure.

A live ClinVar positional query used `BUB1B[gene] AND 40220612[chrpos38]; chr15:40209701 T>G` was queried first as a positive control and returned Variation ID 533901. The target position returned two records, neither corresponding to c.3006T>G:

Table 12: ClinVar records at the p.Asn1002 position.

Variant	Protein change	Consequence	Germline classification
NM_001211.6:c.3006T>C	p.Asn1002=	synonymous	Likely benign, one star
NM_001211.6:c.3006T>A	p.Asn1002Lys	missense	Uncertain significance, one star

Thus the exact c.3006T>G allele is unclassified, but the amino-acid substitution is not wholly unobserved. Because the same-residue c.3006T>A submission is a VUS rather than a pathogenic assertion, ACMG/AMP criterion PS1 does not apply [30]. The VUS is a single-submitter assertion under the broad condition label “Inborn genetic diseases”, so it is evidence of insufficient classification rather than evidence of benignity. The synonymous c.3006T>C record is not informative for the missense allele.

Regional constraint for 15:40220000–40221000 was $Z = -0.8$ with observed/expected 1.06, so local constraint does not support pathogenicity. The report therefore describes p.Asn1002Lys as a strong candidate missense allele, not as independently proven pathogenic.

B.3 Highest-value missing evidence

Three follow-up experiments were identified, in increasing order of effort. First, parental sequencing would establish segregation and whether the alleles are in trans. Second, long-read sequencing or targeted long-range PCR across the 10,911-bp interval would directly phase the alleles without parental samples. Third, a BubR1 kinetochore-function or chromosome-mis-segregation assay in patient-derived cells would add functional evidence that computational predictors cannot supply. Any future clinical submission of patient-derived evidence to ClinVar would require appropriate family and challenge-organizer agreement under the data-sharing terms.

C Full reproducibility workflow for the genome-wide submission

The following commands preserve the original cluster workflow and all computational flags. User-specific filesystem paths and the participant-prefixed submission filename are replaced by anonymous placeholders for review. Raw challenge files are gated and are not included in the public artifact. The pipeline discovers the supplied VCF, paired-end FASTQs, and clinical phenotype document under DATA.

C.1 Paths and dependencies

```

REPO=/path/to/anonymous_repository
DATA=/path/to/authorized_challenge_data
SCRATCH=/path/to/scratch
RUN="$DATA/mva_track1_run"
cd "$REPO"

```

```
python -m pip install -r requirements.txt
mkdir -p "$REPO/results"
python --version > "$REPO/results/python_version.txt"
python -m pip freeze | sort > "$REPO/results/pip_freeze.txt"
```

C.2 Scratch configuration and OpenCRAVAT resources

The OpenCRAVAT modules used in the analysis occupied roughly 160 GB and were placed on scratch storage.

```
export PIP_CACHE_DIR="$SCRATCH/cache/pip"
export XDG_CACHE_HOME="$SCRATCH/cache"
export TMPDIR="$SCRATCH/tmp"
export SSL_CERT_FILE="$(python -m certifi)"
mkdir -p "$PIP_CACHE_DIR" "$XDG_CACHE_HOME" "$TMPDIR" \
    "$SCRATCH/opencravat/modules"
oc config md "$SCRATCH/opencravat/modules"
oc module install-base
oc module install clinvar gnomad gnomad3 revel spliceai cadd
oc module ls > "$REPO/results/opencravat_modules.txt"
```

C.3 Stage-by-stage execution

```
# Inspect all supplied files
python mva_track1_pipeline.py inspect \
    --data-dir "$DATA" --verbose

# Extract phenotype and compute HPO gene scores
python mva_track1_pipeline.py phenotype \
    --data-dir "$DATA" --verbose

# Reconstruct and validate the matching GRCh38 reference
python mva_track1_pipeline.py prepare-reference \
    --data-dir "$DATA" --verbose

# Align all four FASTQ lanes; shard settings limit worker lifetime, not data
python mva_track1_pipeline.py align \
    --data-dir "$DATA" \
    --threads 6 \
    --align-batch-pairs 2000 \
    --align-shard-pairs 2000000 \
    --bwa-chunk-bases 10000000 \
    --verbose

# Normalize the supplied VCF
python mva_track1_pipeline.py normalize \
    --data-dir "$DATA" --threads 6 --verbose

# Read-backed phasing
python mva_track1_pipeline.py phase \
    --data-dir "$DATA" --threads 6 --verbose

# Local functional and population annotation
python mva_track1_pipeline.py annotate \
    --data-dir "$DATA" --threads 6 \
    --oc-annotators clinvar gnomad gnomad3 revel spliceai cadd \
    --verbose

# Rank variants and construct compound-heterozygous pairs
python mva_track1_pipeline.py rank \
    --data-dir "$DATA" --threads 6 --verbose

# Export exact Track 1 schema
python mva_track1_pipeline.py submit \
    --data-dir "$DATA" \
    --proband-id PROBAND01 \
```

```
--top-n 6 \
--submission-name anonymous_genomewide-hpo-comphet.csv \
--verbose
```

The phenotype stage downloaded HPO resources under `$DATA/resources/hpo/` on first use and wrote genome-wide phenotype scores under `$RUN/phenotype/`. Reference preparation reconstructed and indexed the exact GRCh38 representation validated against all 2,580 VCF contigs.

The important ranking outputs were

```
$RUN/ranking/variant_candidates.tsv
$RUN/ranking/compound_het_pairs.tsv
$RUN/ranking/compound_het_pairs_nonprimary_audit.tsv
$RUN/ranking/scoring_audit_v12.json
$RUN/ranking/annotation_join_audit.json
$RUN/ranking/anonymous_genomewide-hpo-comphet.csv
```

The main BAM QC report was `$RUN/alignment/flagstat.txt`. The internal sequencing sample name was `WGS_EX2312012`, so `PROBAND01` was supplied explicitly at export time.

An intentional rebuild of the alignment used:

```
python mva_track1_pipeline.py align \
--data-dir "$DATA" \
--threads 6 \
--align-batch-pairs 2000 \
--align-shard-pairs 2000000 \
--bwa-chunk-bases 10000000 \
--force --verbose
```

The generated submission could be copied into the repository with:

```
cp "$RUN/ranking/anonymous_genomewide-hpo-comphet.csv" \
"$REPO/anonymous_genomewide-hpo-comphet.csv"
```

After dependencies and annotation resources were installed, the equivalent one-command workflow was:

```
python mva_track1_pipeline.py all \
--data-dir "$DATA" \
--threads 6 \
--align-batch-pairs 2000 \
--align-shard-pairs 2000000 \
--bwa-chunk-bases 10000000 \
--oc-annotators clinvar gnomad reval spliceai cadd \
--proband-id PROBAND01 \
--top-n 6 \
--submission-name anonymous_genomewide-hpo-comphet.csv \
--verbose
```

Aggregate non-sensitive audit outputs were copied for reproducibility:

```
mkdir -p "$REPO/results"
cp "$RUN/ranking/scoring_audit_v12.json" \
"$REPO/results/scoring_audit.json"
cp "$RUN/ranking/annotation_join_audit.json" \
"$REPO/results/annotation_join_audit.json"
cp "$RUN/alignment/flagstat.txt" \
"$REPO/results/alignment_flagstat.txt"
```

The reproducible artifact contains `mva_track1_pipeline.py`; `requirements.txt`; `results/pip_freeze.txt` and `results/python_version.txt`; reference reconstruction/validation configuration; phenotype/HPO scoring; OpenCRAVAT ingestion and canonical allele joining; phasing, ranking, QC, pair construction, and submission generation; `results/scoring_audit.json`; `results/annotation_join_audit.json`; `results/alignment_flagstat.txt`; and the submitted CSV. No patient-specific training data were available or assumed. External resources were annotation/prior evidence rather than labels for supervised training on this patient.

The artifact does not include raw FASTQs, the full or normalized VCF, BAM/BAI, phased VCF, the supplied clinical document, the OpenCRAVAT database, credentials, or full patient-level candidate tables. The identifying public repository URLs from the original report are withheld from the anonymous review manuscript and should be restored only in the de-anonymized camera-ready version.

D Complete learned-model design and provenance

D.1 Component and contribution boundaries in anonymous form

For double-blind review, personal names and identifying repository URLs are omitted. The substantive division of work from the original report is preserved as follows. Author A developed and executed the genome-wide workflow, including input discovery, reference preparation, alignment, normalization, phasing, annotation, HPO gene scoring, variant scoring, pair construction, QC, and challenge export. Author B developed the Bayesian ranking and generalization extension, including the shared feature contract, linear Bayesian ranker, SBNN and BBB models, ensemble and gradient-boosting comparisons, hyperparameter studies, temporal ClinVar benchmark, Phenopacket/HPO benchmark, challenge-adaptor verification, uncertainty analysis, and visualization. The diagnosis, modeling results, interpretation, and final outcome are joint work. This anonymized statement should be replaced by the named contribution statement in a camera-ready version.

The first pair models consume phenotype score; minimum and maximum consequence severity; minimum and mean rarity; maximum population AF and missingness fraction; minimum and maximum normalized CADD and SpliceAI plus availability indicators; phenotype-by-consequence and phenotype-by-rarity interactions; and the same pair identity, phase exclusion, and strict/broad candidate definitions used by the deterministic pipeline. The temporal ClinVar models are a distinct molecular layer with 50 features derived from ClinVar records and training-only gene summaries. When transferred back to Track 1 they score the alleles in the genome-wide candidate tables, combine allele probabilities geometrically, and multiply by the phenotype score. They do not replace candidate generation, annotation, phase logic, or inheritance filtering.

The Bayesian extension had three explicit objectives: (1) preserve Track 1 compatibility, (2) test molecular generalization to variants unavailable at training time, and (3) test patient generalization and uncertainty on other solved disorders.

D.2 Public datasets, local acquisition, and hashes

Table 13: Full dataset provenance retained from the report.

Dataset	Local acquisition / scale	Use and boundary
Compact challenge ranking export	<code>ranking.zip</code> ; annotated candidate tables only	Lets learned rankers consume the same strict/broad candidate universe without moving the gated WGS data
Phenopacket Store 0.1.27	Versioned release archive, 19 MB in the original local acquisition [19]	Solved patients, exact GRCh38 causal variants, genes, source groups, and HPO terms; used in pair and temporal-patient benchmarks
Phenopacket annotation cache	VEP-derived features for benchmark hypotheses	Consequence, population frequency, CADD, SpliceAI for causal and benign decoys; not full patient genomes
ClinVar 26 Dec 2023	<code>clinvar_20231226.vcf.gz</code> , archived GRCh38 VCF, 87 MB	Training labels and training-only gene history; 2,340,600 total Variation IDs and 1,084,930 unambiguous labels
ClinVar 29 Aug 2026	<code>clinvar_20260829.vcf.gz</code> , archived GRCh38 VCF, 184 MB	Temporal test; 626,287 Variation IDs absent in 2023
HPO 9 Oct 2023	<code>v2023-10-09/hp.obo</code> , 9.2 MB	Frozen ontology for temporal patient benchmark
Challenge evaluator	locally reproduced public logic	Used only to verify CSV behavior; private ground-truth file was not copied

The report recorded truncated checksums for compact public inputs: 2023 ClinVar 510c54e3fd04b078...cb63786; 2026 ClinVar e72f5619f003c80c...6ec91f9; Phenopacket Store 0.1.27 d0c70005bb09b870...edf205a; HPO 2023-10-09 d3c1280f2780628c...e6cd4e52. Full hashes were retained in the project artifacts.

D.3 Bayesian conditional-logit pair ranker

The first learned pair benchmark loaded 9,739 solved Phenopacket cases, including 959 reported two-variant cases. After annotation and decoy requirements, 425 cases were used for training, 44 for validation, and 89 for development testing. Genes, source publications, and case identifiers were disjoint across splits, and *BUBIB* was excluded. Each case had 23 constructed candidate pairs. For candidate feature vector x_{ij} ,

$$p(y_i = j \mid X_i, w) = \frac{\exp(w^\top x_{ij})}{\sum_k \exp(w^\top x_{ik})}. \quad (6)$$

A zero-centered Gaussian prior regularized w ; MAP was optimized and a Laplace approximation around the optimum supplied a Gaussian posterior. Prior scale was selected by validation NLL. On the 89-case development set, the fixed evidence rule achieved 70.8% Top-1 / 83.1% Top-5, the Bayesian posterior mean 58.4% / 87.6%, and the first SBNN 56.2% / 87.6%. The negative result motivated moving the learned comparison to a larger variant-level task rather than replacing explicit biological structure.

D.4 Low-rank SBNN, dense BBB, and architecture search

The first pair SBNN used hidden widths 32 and 16 with factor ranks (8, 8, 1). Both random and SVD-based posterior-mean initialization were tested. SBNN weight matrices use $W = AB^\top$ with Gaussian variational factors. The main text gives the implemented factor-space objective; the corresponding weight-space form it approximates is

$$\mathcal{L} = \mathbb{E}_{q(W)}[-\log p(y \mid X, W)] + \lambda \text{KL}[q(W) \parallel p(W)], \quad (7)$$

following Toure and Stephens [31]. Sampling is reparameterized, for example $A = \mu_A + \text{softplus}(\rho_A) \odot \epsilon_A$ with $\epsilon_A \sim \mathcal{N}(0, I)$, so the expected loss is differentiable. Gaussian factor priors give analytic KL terms; other priors would require Monte Carlo estimation.

For the temporal task, the compact SBNN factorial grid contained 375 configurations: ranks (8, 4, 1), (8, 8, 1), (16, 8, 1), (16, 16, 1), and (32, 16, 1); KL weights 0.001, 0.01, 0.1, 1, 3; posterior-scale multipliers 0.1, 0.3, 1, 3, 10; and broad prior standard deviations 0.3, 1, 3. The best confirmed compact configuration was (8, 4, 1), KL 0.01, posterior multiplier 0.1, prior 0.3, with unseen-gene OOD AUROC 0.6231 versus 0.6020 for the original compact model. Configurations exceeding roughly 0.70 OOD AUROC only by collapsing pathogenicity classification toward chance were rejected.

Dense Bayes by Backprop assigned an independent Gaussian mean and ρ to every weight and bias [1]. Its 75-cell grid crossed $\rho \in \{-8, -6, -5, -4, -3\}$, KL weight $\in \{0.001, 0.01, 0.1, 1, 3\}$, and prior standard deviation $\in \{0.3, 1, 3\}$. The selected setting used $\rho = -4$, KL 0.1, and prior 0.3. Compact dense BBB used 4,354 variational scalars versus 10,885 fitted weights across the five-network compact ensemble.

Wider models used hidden widths 200 and 100. SBNN ranks 8, 16, 32, 50, and an algebraically full factorization were crossed with KL and posterior-scale settings. The selected rank-50 SBNN used 55,804 variational scalars, about 8% fewer than the matching wide dense BBB. It achieved the best temporal Top-1 (84.35%), PR-AUC (0.9736), F-max (0.9235), and 80%-coverage balanced accuracy (99.12%), while its OOD AUROC remained 0.6284. Increasing rank and width improved the primary task without reproducing the independently optimized function-space diversity of the ensemble.

D.5 Complete Track 1 adapter comparison

The main text reports that every compatible learned adapter recovers the accepted pair at rank 1. Table 14 gives the per-method result. Only the deterministic WGS submission was organizer-confirmed; all other rows were produced with locally reproduced evaluator logic on the same genome-wide candidate tables. The uniformity of the column is the finding: once the upstream workflow has reduced 5.08 million variants to six strict hypotheses, this benchmark no longer separates the models.

Table 14: Complete Track 1 comparison. The deterministic WGS submission was organizer-confirmed; learned adapters were evaluated locally. All methods recover the accepted pair at rank 1.

Method	Rows	Rank	RP/F-max	Method	Rows	Rank	RP/F-max
WGS pipeline	6	1	100/1.000	Dense MLP	6	1	100/1.000
Linear Bayes, strict	6	1	100/1.000	Deep ensemble	6	1	100/1.000
Linear Bayes, broad	10	1	100/1.000	Small SBNN	6	1	100/1.000
Pair SBNN, strict	6	1	100/1.000	Dense BBB	6	1	100/1.000
Pair SBNN, broad	10	1	100/1.000	Wide rank-50 SBNN	6	1	100/1.000
Gradient boost	6	1	100/1.000	Wide dense BBB	6	1	100/1.000
Temporal consensus	6	1	100/1.000				

D.6 Temporal ClinVar benchmark details

The 2023 historical release supplied 1,084,930 unambiguous benign/pathogenic labels. The balanced 60,000-variant training set contained 30,000 pathogenic variants from 5,043 genes. The 15,000-variant validation set was gene-disjoint and contained 1,209 genes. The temporal test contained 626,287 newly appearing Variation IDs, including 117,147 pathogenic variants across 14,715 genes. Even IDs selected the consensus mixture; 313,171 odd IDs remained untouched. The selected consensus weight was 0.30 gradient boost, 0.20 deep ensemble, 0.50 compact SBNN.

The 50 molecular features contained reference and alternate allele length, variant type, transition status, substitution class, population frequency and missingness, dbSNP availability, germline/inherited/de-novo origin flags, review-level indicators, molecular consequence categories, and training-only gene count, pathogenic rate, and unseen-gene indicator. No disease name or clinical-significance label was used as input. For the ranking simulation, 2,000 pathogenic temporal variants were each placed among 99 benign temporal variants sampled with a fixed seed. PR-AUC and F-max were computed across all 313,171 odd-ID records; NLL measured probability quality; OOD AUROC tested whether each model’s stochastic uncertainty score distinguished genes absent from 2023 training.

D.7 Patient-level temporal benchmark details

A Phenopacket hypothesis entered the final test only if it contained one or two exact GRCh38 causal variants, every causal coordinate was absent from raw 2023 ClinVar, and those variants were pathogenic in the 2026 snapshot. Repeated causal hypotheses were deduplicated. The final set contained 719 post-2023 cases from 265 genes: 655 single-variant and 64 two-variant cases. Each patient had one causal hypothesis and nine temporally new benign ClinVar hypotheses. Single-variant decoys were matched on variant type and consequence group when possible; two-variant decoys consisted of two benign alleles in one gene. Pair molecular scores were geometric means of allele probabilities.

The HPO layer used only the 9 October 2023 ontology. Older solved cases were split into 1,388 phenotype-reference cases and 843 weight-selection cases. For candidate gene g and query phenotype set Q , phenotype support was the best symmetric Resnik similarity to an older case for g [22]. With $IC(h) = -\log P(h)$, that similarity is

$$R(a, b) = \max_{h \in \text{Anc}(a) \cap \text{Anc}(b)} IC(h),$$

$$S(Q, T) = \frac{1}{2} \left[\frac{1}{|Q|} \sum_{a \in Q} \max_{b \in T} R(a, b) + \frac{1}{|T|} \sum_{b \in T} \max_{a \in Q} R(a, b) \right]. \quad (8)$$

Each model selected its molecular/HPO coefficient and missing-gene prior on the 843 older cases; the 719 post-2023 patients were untouched. Five independently resampled benign candidate cohorts measured sensitivity to decoy composition; the reported SDs therefore reflect candidate resampling rather than five independent patient cohorts.

For dense BBB, HPO conditioning improved Top-1 by 23.5–26.1 percentage points in every cohort. Each paired cohort gained 175–194 correct cases while losing only 2–6, with exact McNemar p values of approximately 10^{-44} to 10^{-53} . Of the 719 test cases, 337 causal genes appeared in older

phenotype-reference cases and 382 did not. Dense BBB rose from 34.7% to 88.1% Top-1 on known-gene cases but changed from 33.6% to 33.1% on unseen-gene cases. The missing-gene prior prevents an excessive penalty when no phenotype reference exists but does not invent novel gene-disease knowledge.

D.8 Three probability concepts and uncertainty interpretation

The study distinguishes (i) deterministic EPCR, a monotone ranking confidence for the challenge output; (ii) posterior predictive probabilities from Bayesian weight distributions; and (iii) clinical posterior evidence under ACMG/AMP. They are not interchangeable. For stochastic predictions p_m with mean $\bar{p}$, predictive entropy is $H[\bar{p}]$ and epistemic mutual information is

$$I[y, W \mid x, D] = H[\bar{p}] - \frac{1}{M} \sum_{m=1}^M H[p_m]. \quad (9)$$

For binary probabilities the law of total variance gives the complementary decomposition

$$\bar{p}(1 - \bar{p}) = \frac{1}{M} \sum_m p_m(1 - p_m) + \frac{1}{M} \sum_m (p_m - \bar{p})^2, \quad (10)$$

separating the mean within-sample variance from the dispersion across samples. The tuned low-rank SBNN’s mean predictive SD rose from 0.00157 on represented genes to 0.00403 on unseen genes; the deep ensemble rose from 0.01179 to 0.03569. Both respond to shift, but the ensemble’s function-space dispersion is much larger. More Monte Carlo draws estimate a given posterior more accurately; they do not create the distinct optimized modes of independently trained ensemble members.

D.9 Selective prediction definition

The main text describes selective balanced accuracy in words; the explicit definition is as follows. For uncertainty score u_i , let $\mathcal{R}_c$ contain the least-uncertain fraction c of evaluated examples. In binary evaluation,

$$\text{BA}(c) = \frac{1}{2} \left[\frac{\sum_{i \in \mathcal{R}_c} \mathbf{1}[y_i = 1, \hat{y}_i = 1]}{\sum_{i \in \mathcal{R}_c} \mathbf{1}[y_i = 1]} + \frac{\sum_{i \in \mathcal{R}_c} \mathbf{1}[y_i = 0, \hat{y}_i = 0]}{\sum_{i \in \mathcal{R}_c} \mathbf{1}[y_i = 0]} \right], \quad (11)$$

with coverage $|\mathcal{R}_c|/N$ and both class denominators required to be nonzero. The reported 99.12% value for the wide rank-50 SBNN is $\text{BA}(0.8)$ on the odd-ID temporal partition. It is a retrospective operating point on that partition, not a validated clinical referral threshold, and the coverage level was not selected against an independent shift-validation set.

D.10 Comparison with established rare-disease systems

The original report placed the experiments in the context of hybrid rare-disease systems. Exomiser combines frequency, predicted pathogenicity, pedigree, and phenotype evidence [27]; LIRICAL uses interpretable phenotype likelihood ratios with genotype information [24]; MAVERICK is an ensemble of transformer-based neural models with structured variant evidence and inheritance-aware outputs [8]; and AI-MARRVEL combines learned predictions from millions of variants with expert-engineered knowledge features [17]. In the CAGI Rare Genomes Project, leading teams recovered up to 13 of 14 solved test families in their top five while still requiring clinical review [28]. The design here follows the same layered principle at smaller scale: candidate validity, molecular pathogenicity, phenotype, inheritance, phase, and uncertainty remain distinct evidence sources.

E Complete limitations, established claims, and next steps

E.1 Clinical and genome-wide limitations

1. Parental genomes were not provided, preventing direct segregation analysis.
2. Short-read phasing could not establish whether the two *BUB1B* alleles are in trans.
3. Computational prediction cannot independently establish pathogenicity of p.Asn1002Lys.

4. Structural variants, complex repeat-associated events, and other variant classes may require specialized analysis beyond the supplied small-variant VCF.
5. The ranking used gnomAD v2.1/v3. Direct v4.1.1 verification later showed p.Asn1002Lys to be a singleton rather than absent; both states are far below the ranking threshold, so the rank is unchanged.
6. The exact c.3006T>G allele has no ClinVar record, while c.3006T>A produces the same amino-acid substitution and is a one-star VUS; ACMG/AMP PS1 therefore does not apply.
7. Local constraint around p.Asn1002Lys is unremarkable (o/e 1.06), so constraint provides no independent support.
8. Local constraint around p.Leu737Ter is only mild (o/e 0.88; $Z = 1.79$), below a conventional constrained-region threshold.
9. The p.Leu737Ter ClinVar classification is a two-star aggregate rather than an expert-panel or practice-guideline assertion.
10. Extreme rarity is supportive but not itself proof of pathogenicity.
11. The ranking system is an evidence-prioritization tool and is not a substitute for formal clinical molecular interpretation.

E.2 Generalization and uncertainty limitations

The external 58.9% patient result is not full-exome or full-genome diagnostic sensitivity. Candidate-generation recall has not been measured on an independent patient cohort. The submitted EPCR is not clinically calibrated across patients. The patient benchmark uses literature-derived Phenopackets and compact ClinVar-derived decoys. Temporal ClinVar entry time is partly a curation/submission process rather than a pure biological clock. OOD is defined only by unseen gene; ancestry, sequencing platform, phenotype capture, variant mechanism, and annotation shift remain untested. No Bayesian model in the study surpasses the deep ensemble on unseen-gene OOD detection.

E.3 Claims established by executed artifacts

The executed artifacts establish that: (1) the uploaded deterministic pipeline found the accepted *BUB1B* pair at rank 1 and achieved 100 Rank Points / F-max 1.0; (2) every compatible learned adapter also ranks the pair first under locally reproduced evaluator logic; (3) the temporal corpus contains more than one million historical labels and 626,287 newly appearing test IDs; (4) the wide rank-50 SBNN gives the best tested temporal simulated ranking and F-max; (5) deep-ensemble disagreement is the best tested unseen-gene detector; (6) HPO conditioning selected only on older cases materially improves compact post-2023 patient ranking; and (7) tests, seeds, model artifacts, and metric files are retained. The artifacts do not establish functional pathogenicity of p.Asn1002Lys, trans phase of the two *BUB1B* alleles, full-exome clinical sensitivity, or universal superiority of Bayesian uncertainty.

F AI assistance, data handling, and reproducibility boundary

Generative-AI assistance was used in the original workflow to help implement and debug analysis code, reason through aggregate intermediate outputs, inspect failure modes, refine the ranking pipeline, and draft portions of the report. The researchers defined the research direction, hypotheses, analyses, and acceptance criteria; all computation was executed by the researchers on an institutional compute cluster (name withheld for review). No candidate call or scientific conclusion was accepted solely because a model suggested it.

The categories shared with the model were source code, configuration and error traces; aggregate QC summaries such as alignment totals, variant/heterozygous/phasing counts and annotation-join counts; ranking/scoring audit summaries; coordinates, consequences, and public annotations for the small number of top candidate pairs; and the HPO terms reported in the manuscript. The raw FASTQs, supplied/full/normalized/phased VCFs, BAM/BAI files, supplied clinical document, and full patient-level candidate tables were never uploaded to or made accessible to a third-party model service. The gated challenge data are not redistributed. Reproduction of the genome-wide stage requires authorized access; the temporal ClinVar/Phenopacket experiments use public versioned

resources. The aggregate material shared for interpretation was a subset of the report/submission-level information released under the Hackathon terms; no gated patient-level file was redistributed.

G Coverage audit from the original report to this paper

Table 15 is the explicit completeness audit requested during manuscript preparation. It maps every substantive section of the original report to the main paper or appendix. Repeated prose, shell comments, and navigation text are not duplicated verbatim, but every scientific result, evidence class, experimental boundary, figure, major table, limitation, and reproducibility stage is represented. Material that was condensed in the main text for length, including the challenge scoring equations, the per-method Track 1 comparison, and the selective-prediction definition, is retained in full in this appendix and is marked as such below. Author names and identifying repository URLs are intentionally anonymized for double-blind review; their substantive contribution and reproducibility content are retained.

Table 15: Report-to-paper coverage map.

Original report element	Location in this paper
TL;DR, abstract, system overview	Abstract, Introduction, Figure 1
Submission 1 status and maximum score	Challenge and genome-wide diagnosis; Results, challenge saturation; Table 14; Appendix A
Challenge scoring metrics (Rank Points, F-max)	Summarized in Challenge and genome-wide diagnosis; full definitions in Appendix A.2
Scientific contribution at a glance	Introduction; Figure 2; Appendix A
Discovery versus post-submission validation	Genome-wide processing and candidate construction; Appendix A
Summary, data, build validation	Genome-wide processing and candidate construction; Appendix A
Clinical phenotype/HPO terms	Genome-wide processing and candidate construction; Appendix A.4
Alignment and paired-end QC	Genome-wide processing and candidate construction; Appendix A
Variant normalization	Genome-wide processing and candidate construction; Appendix A
Read-backed phasing and phase interpretation	Genome-wide processing and candidate construction; Appendix A
Functional/population annotation and canonical join	Genome-wide processing and candidate construction; Appendix A
Phenotype-guided gene scoring	Genome-wide processing and candidate construction; Appendix A.4
Variant scoring formula and raw-REVEL correction	Genome-wide processing and candidate construction; Appendix A
Candidate pair construction and strict QC	Genome-wide processing and candidate construction; Figure 2a; Appendix A
Leading p.Leu737Ter allele	Table 1; Appendix B
p.Leu737Ter gnomAD verification	Appendix B
p.Leu737Ter ClinVar submissions and mechanism evidence	Appendix B
p.Leu737Ter LOF-TEE/VEP/constraint detail	Appendix B
Leading p.Asn1002Lys allele	Table 1; Appendix B
p.Asn1002Lys gnomAD singleton verification	Appendix B

Original report element	Location in this paper
p.Asn1002Lys direct ClinVar positional query and same-residue VUS	Appendix B
p.Asn1002Lys VEP/predictors/constraint	Appendix B
Internal <i>BUB1B</i> variant comparison	Appendix A.4
Final ranked six pairs	Table 1
Differential diagnosis	Table 1; Leading hypothesis subsection
Robustness scores	Appendix A
Submitted primary hypothesis/schema	Appendix A
Full reproducibility workflow	Appendix C
Repository contents and gated-data boundary	Appendices C and F
AI assistance disclosure	Limitations, uncertainty and reproducibility paragraph; Appendix F
Open question and proposed segregation/phasing/functional follow-up	Leading hypothesis subsection, “What remains to be demonstrated”; Appendix B
Complete Part I limitations	Limitations; Appendix E
Team contributions/model boundaries	Appendix D, anonymized for review
Objectives of Bayesian layer	Bayesian ranking and generalization methods; Appendix D
Training-data provenance and checksums	Table 13; Appendix D
Bayesian conditional-logit pair ranker	Initial pair-level Bayesian learning; Results, negative pair learning; Appendix D
First low-rank SBNN pair model	Initial pair-level Bayesian learning; Appendix D
Temporal ClinVar experiment and 50 features	Temporal molecular benchmark; Table 2; Appendix D
Factorial SBNN sweep	Bayesian models and posterior inference; Results; Appendix D
Dense BBB grid	Bayesian models and posterior inference; Results; Appendix D
Wide SBNN/dense models and parameter counts	Figure 5a; Appendix D
Complete Track 1 model comparison	Summarized in Results, challenge saturation; full table in Appendix D.5
Temporal table interpretation	Results, temporal ranking; Figure 3
External patient benchmark construction	External patient-level temporal benchmark; Appendix D
External patient results	Table 3; Figure 4
Known-gene versus unseen-gene behavior	Results, patient transfer; Figure 5b; Appendix D
Calibration, MI, and meaning of Bayesian probability	Posterior prediction and uncertainty; Appendix D
Selective prediction and coverage-risk reporting	Summarized in Ranking, shift detection, and selective prediction; full definition in Appendix D.9
Comparison with production systems	Introduction; Discussion; Appendix D
Established results and remaining gaps	Conclusion; Appendix E
Original conclusion	Conclusion
All seven report figures	Main paper only: Figures 1, 2, 3, 4, and 5 contain all seven original image assets

Original report element	Location in this paper
References and external resources	Main bibliography plus added verified references

H Material relocated from the main text

H.1 Local challenge scoring

Let G be the accepted allele set, S_i the allele set in row i , and e_i its submitted EPCR score. In our local full-match audit, rows are sorted by decreasing EPCR and the full-match rank is the first index with $S_i = G$; candidate rank 1 earns 100 Rank Points. For threshold t , define $P_t = \bigcup_{i:e_i \geq t} S_i$. Precision and recall are $|P_t \cap G|/|P_t|$ and $|P_t \cap G|/|G|$, respectively, with zero precision for an empty prediction set. The audit reports $F_{\max} = \max_t F_1(t)$, taking $F_1 = 0$ when precision and recall are both zero. A threshold isolating the accepted pair yields F-max 1.0 for this proband without establishing cross-patient calibration.

H.2 Biological background

H.3 Biological background: chromosome segregation and MVA

Chromosome segregation and MVA: Human somatic cells normally contain 46 chromosomes in 23 pairs. Before cell division, each chromosome is duplicated into two sister chromatids, and during mitosis, spindle microtubules attach to the kinetochores and pull the chromatids toward opposite daughter cells. The spindle assembly checkpoint delays their separation until attachment is sufficiently secure, so failure of this surveillance can leave a daughter cell with an extra or missing chromosome, a state called aneuploidy. In MVA, recurrent segregation errors produce a mosaic of cell populations with gains or losses of different chromosomes. The resulting chromosomal instability is associated with growth restriction, microcephaly, developmental abnormalities, and predisposition to childhood malignancies, but the phenotype varies, and no individual clinical feature is diagnostic. MVA is genetically heterogeneous: established causes include *BUB1B*, *CEP57*, and *TRIP13*, which affect different components of chromosome-segregation machinery [11, 29].

Why BUB1B is biologically plausible: *BUB1B* encodes BubR1, a central spindle-checkpoint protein that helps prevent premature chromosome separation and contributes to stable kinetochore-microtubule attachments. Pathogenic reductions in BubR1 abundance or function can weaken checkpoint signalling and increase segregation errors. Foundational studies identified biallelic truncating and missense *BUB1B* variants in MVA families, including patients with embryonal rhabdomyosarcoma. Experiments in patient-derived cells subsequently demonstrated impaired checkpoint activity, chromosome-alignment defects, and reduced BubR1 abundance; restoring BubR1 rescued checkpoint function [11, 29]. These studies establish the gene-disease mechanism, but do not establish that every rare missense variant in *BUB1B* is pathogenic.

Recessive inheritance and phase: For a typical autosomal recessive disorder, disease-causing changes must affect both inherited copies of a gene. Two heterozygous variants form a compound-heterozygous genotype only if they are in *trans*, on separately inherited chromosomes; in *cis* they affect the same copy and may leave the other intact. A variant call establishes neither phase nor the functional effect of an allele.

H.4 Genome-wide processing and candidate construction

H.5 Genome-wide processing and candidate construction

The input comprised four paired-end FASTQ lanes, a WGS VCF, and a phenotype document. The reconstructed GRCh38 reference matched all 2,580 VCF contigs. Alignment processed 1,076,022,862 reads, with 99.61% primary mapping. Normalization yielded 5,079,064 variants and 3,180,797 heterozygous calls. WhatsHap phased 2,754,502 heterozygotes (86.60%) into 439,579 phase sets

[18]. A phase set is a group of variants whose relative chromosome assignment is supported by sequencing reads; variants in different sets need not have a known relative phase.

A canonical chromosome/position/reference/alternate-allele join matched 3,180,796 heterozygous alleles to OpenCRAVAT annotations containing ClinVar, gnomAD, CADD, REVEL, and SpliceAI evidence [12–14, 21]. ClinVar records clinical interpretations, gnomAD supplies population frequencies, and computational predictors estimate molecular consequences. HPO terms encode the patient’s clinical findings in a standardized ontology [9]. The phenotype included rhabdomyosarcoma, nephrocalcinosis, short stature, failure to thrive, muscle atrophy, prematurity, growth restriction, and parental recurrent miscarriages. No *BUB1B* or MVA-gene bonus or whitelist was inserted. The conservative upstream variant score was

$$s_v = 0.36s_{\text{consequence}} + 0.26s_{\text{rarity}} + 0.20s_{\text{QC}} + 0.10s_{\text{ClinVar}} + 0.08s_{\text{predictor}}. \quad (12)$$

The predictor term is the arithmetic mean of available raw REVEL, clip(CADD PHRED/40, 0, 1), and maximum available SpliceAI delta scores; it is zero when all three are missing. The coefficients define an engineered ranking score, not fitted clinical posterior probabilities. For a pair (v_1, v_2) , let $\bar{s} = (s_{v_1} + s_{v_2})/2$, $s_{\min} = \min(s_{v_1}, s_{v_2})$, and h be its gene’s phenotype score. The base pair score is $a = 0.55\bar{s} + 0.25s_{\min} + 0.15h + 0.05s_{\text{phase}}$, where s_{phase} is 1 for read-supported trans, 0 for read-supported cis, and 0.5 for unknown phase. Before filtering, cis pairs receive a multiplicative factor of 0.20. The score then receives a capped 0.04 bonus if both consequence scores are at least 0.72, followed by a capped 0.03 bonus if each allele’s AF is missing or at most 10^{-3} ; each cap is 1. These are generic evidence bonuses, not disease-specific bonuses. The raw EPCR proxy is $\text{clip}(\sigma((s_{\text{pair}} - 0.55)/0.10), 0.001, 0.999)$. After sorting, the implementation enforces descending submitted EPCR values by capping each value below the previous value by 10^{-6} , with a lower bound of 0.001. Same-gene heterozygous pairs were restricted to primary chromosomes, PASS, depth (DP) at least 10, genotype quality (GQ) at least 20, allele balance 0.20-0.80, and AF at most 0.005 when observed. Within a shared phase set, read-backed same-haplotype pairs were excluded as cis and opposite haplotypes supported trans, and different phase sets remained unknown. Of 508 initial pairs, 131 were on primary chromosomes, and seven survived strict QC. Removing one read-backed cis *ZNF619* pair left six reportable hypotheses. The uploaded result preceded disease-specific reranking, which is treated only as post-submission sensitivity analysis. Appendix A gives the full preprocessing, alignment, phasing, and phenotype audits.

See Figure 2 in the main text.

H.6 Leading hypothesis and evidence

H.7 Leading hypothesis: evidence and unresolved questions

The leading hypothesis is a *BUB1B* compound-heterozygous pair, p.Leu737Ter/p.Asn1002Lys, with score 0.8458 and EPCR 0.950624. The first is a stop-gain variant, expected to truncate the protein or trigger degradation of the abnormal transcript; the second is a missense variant, whose functional effect cannot be inferred from the amino-acid substitution alone. Read support, population evidence, and computational scores for both alleles are given in Table 1.

The truncating allele: The GRCh38 variant chr15:40209701 T>G, c.2210T>G is annotated as a high-confidence loss-of-function variant by LOFTEE and carries two-star Pathogenic/Likely pathogenic ClinVar evidence. Direct gnomAD v4.1.1 verification identified 120 alternate alleles across 1,614,020 variants, with no observed homozygotes. These observations jointly provide strong molecular and clinical support, although individual annotations are not independent pieces of evidence [10, 20].

The missense allele: The GRCh38 variant chr15:40220612 T>G, c.3006T>G, p.Asn1002Lys is a singleton in gnomAD v4.1.1 (1/1,614,226), with no homozygotes. The exact nucleotide change has no ClinVar classification in the consulted record; a different change, c.3006T>A, producing the same amino-acid substitution, has a one-star VUS assertion. A same-residue assertion cannot be transferred automatically to a different nucleotide change, and computational predictions do not establish the variant’s effect on BubR1 expression, stability, or function [30].

What remains to be demonstrated: The two alleles are 10,911 bp apart and are not connected by a read-backed phase block. If they are in trans and the missense allele substantially impairs BubR1, the

combination would fit the established recessive mechanism. If they are in cis, or the missense allele is functionally benign, the interpretation changes. Parental segregation or long-range phasing would address chromosome assignment; RNA/protein studies and patient-derived cell assays could assess the missense effect, BubR1 abundance, spindle-checkpoint activity, and chromosome missegregation. Small-variant analysis also does not exhaust structural or complex variation. The observed clinical findings support the hypothesis, but neither a high ranking score nor a matching syndrome substitutes for these tests. Appendix B gives the complete evidence audit for both alleles.

See Table 1 in the main text.

H.8 Conditional-logit definitions

For case i , let $X_i \in \mathbb{R}^{K_i \times 15}$ contain candidate feature rows $x_{ij}^\top$, where $K_i = 23$, $j \in \{1, \dots, K_i\}$, and y_i indexes the accepted pair. With $w \in \mathbb{R}^{15}$, conditional logit assigns

$$p(y_i = j \mid X_i, w) = \frac{\exp(w^\top x_{ij})}{\sum_{k=1}^{K_i} \exp(w^\top x_{ik})}. \quad (13)$$

This normalizes scores within a candidate set rather than classifying each pair independently. A zero-centred Gaussian prior regularizes w . With $w \sim \mathcal{N}(0, \tau^2 I)$, $\tau > 0$, and summed categorical log-likelihood $\ell(w) = \sum_i \log p(y_i \mid X_i, w)$ under conditional independence across cases, the MAP estimate minimizes $-\ell(w) + \|w\|_2^2 / (2\tau^2)$. The Laplace approximation $q_L(w) = \mathcal{N}(\hat{w}, H^{-1})$ takes H to be the Hessian of that regularized objective at $\hat{w}$. The proper Gaussian prior contributes $\tau^{-2} I$ to H . The conditional-logit prior standard deviation τ is selected by validation NLL.

H.9 Variational distributions, priors and objective

Independent mean-field Gaussian variational distributions are placed on the two factors and the bias:

$$q_\phi(A, B, b) = q_{\phi_A}(A) q_{\phi_B}(B) q_{\phi_b}(b),$$

$$q_{\phi_Z}(Z) = \prod_{u \in \mathcal{I}_Z} \mathcal{N}(Z_u \mid \mu_u^Z, \text{softplus}(\rho_u^Z)^2), \quad Z \in \{A, B, b\}, \quad (14)$$

where $\mathcal{I}_Z$ indexes the scalar entries of Z and $\text{softplus}(t) = \log(1 + e^t)$. In $\mathcal{N}(z \mid \mu, \sigma^2)$, the parameter following the mean is the variance. Thus the standard deviation of each variational Gaussian is $\text{softplus}(\rho_u^Z) > 0$. The factorization in Equation 14 is an independence assumption in the latent factor space; it is not a factorization of the posterior over the entries of W .

Instead, the variational distribution on W is induced by the map $T(A, B) = AB^\top$. In pushforward notation,

$$q_\phi^W = T_\#(q_{\phi_A} \otimes q_{\phi_B}). \quad (15)$$

Consequently, although the entries of A and B have independent Gaussian variational distributions, the induced entries of W are generally dependent and non-Gaussian. When $r < \min(m, n)$, q_ϕ^W is supported on the set of matrices with rank at most r and is singular with respect to Lebesgue measure on the full weight space. This differs from full-weight mean-field BBB, which assigns an independent Gaussian approximation directly to each entry of a dense weight matrix.

For this parameterization, an SBNN layer has exactly $2r(m+n) + 2n$ variational location and scale parameters, compared with $2mn + 2n$ for dense BBB. A layer uses fewer variational scalars only when $r(m+n) < mn$. For $r < \min(m, n)$, factorization constrains the weight matrix itself, rather than adding a low-rank covariance correction to a full-dimensional weight distribution. For each factor or bias array Z , the prior is

$$p_Z(Z) = \prod_{u \in \mathcal{I}_Z} [\pi \mathcal{N}(Z_u \mid 0, \sigma_1^2) + (1 - \pi) \mathcal{N}(Z_u \mid 0, \sigma_2^2)]. \quad (16)$$

The implementation uses $\pi = 0.5$ and $\sigma_2 = e^{-6}$; σ_1 is the configurable broad-component scale. Dense BBB uses the same scalar prior family directly on weights and biases. The KL is estimated using sampled log-density differences, rather than a Gaussian-to-Gaussian closed-form KL.

Let $\theta = \{A_\ell, B_\ell, b_\ell\}_{\ell=1}^L$ collect all factor and bias parameters, with independent variational factors and priors across layers. Training uses the KL-weighted objective

$$\mathcal{J}_\lambda(\phi) = -\mathbb{E}_{q_\phi(\theta)} \log p(D \mid \theta) + \lambda \text{KL}(q_\phi(\theta) \parallel p(\theta)). \quad (17)$$

The KL decomposes as a sum over layers and their factor and bias distributions. At $\lambda = 1$, this is the ordinary negative ELBO; other values define a KL-weighted variational objective.

For the molecular benchmark, a minibatch $\mathcal{B}$ of the N_{tr} training records uses one sampled network $\theta^{(s)} \sim q_\phi$ and the estimator

$$\begin{aligned} \hat{\mathcal{J}}_{\lambda,t} = & \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \ell_{\text{BCE}}(y_i, p_i(\theta^{(s)})) \\ & + \frac{a_t \lambda}{N_{\text{tr}}} \left[\log q_\phi(\theta^{(s)}) - \log p(\theta^{(s)}) \right]. \end{aligned} \quad (18)$$

Here $\ell_{\text{BCE}}(y, p) = -y \log p - (1 - y) \log(1 - p)$, and $a_t = \min(1, t/10)$ at epoch $t \geq 1$ in the temporal SBNN and BBB implementations. Once warmup is complete, this estimates $\mathcal{J}_\lambda/N_{\text{tr}}$. The initial pair-level SBNN uses a categorical data-fit term instead of binary cross-entropy. The factor-space KL is defined using the factor-space prior above. For a single layer, the corresponding induced weight prior is $p^W = T_\#(p_A \otimes p_B)$. The data-processing inequality gives

$$\text{KL}(q_\phi^W \| p^W) \leq \text{KL}(q_{\phi_A} \otimes q_{\phi_B} \| p_A \otimes p_B). \quad (19)$$

Equality need not hold. If $r < \min(m, n)$, the induced weight distribution is singular with respect to Lebesgue measure.

H.10 Predictive standard deviation

In the temporal molecular experiment, predictive standard deviation refers to variation in the sampled pathogenic-class probabilities:

$$u_i^{\text{SD}} = \left[\frac{1}{M} \sum_{m=1}^M (p_m(y = 1 | x_i) - \bar{p}_M(y = 1 | x_i))^2 \right]^{1/2}. \quad (20)$$

This uses divisor M and measures variation across parameter samples, not the standard deviation of the Bernoulli outcome.

H.11 Metric definitions

H.12 Ranking, shift detection, and selective prediction

For case i , let C_i denote its candidate set, s_{ij} the score of candidate j , and j_i^* the accepted hypothesis. In the temporal molecular and patient benchmarks, we assign the accepted hypothesis the midrank

$$\begin{aligned} r_i = & 1 + \sum_{\substack{j \in C_i \\ j \neq j_i^*}} \mathbf{1}\{s_{ij} > s_{i,j_i^*}\} \\ & + \frac{1}{2} \sum_{\substack{j \in C_i \\ j \neq j_i^*}} \mathbf{1}\{s_{ij} = s_{i,j_i^*}\}. \end{aligned} \quad (21)$$

For N_c evaluated cases, we report

$$\text{Top-}k = \frac{1}{N_c} \sum_{i=1}^{N_c} \mathbf{1}\{r_i \leq k\}. \quad (22)$$

Thus a tie for first is not counted as a Top-1 success. The initial pair-level experiment instead uses descending stable sorting, breaking ties by candidate order. The molecular benchmark uses one pathogenic and 99 benign variants per simulated case; the patient benchmark uses one causal and nine benign hypotheses.

NLL, Brier score, and ECE assess probability quality, whereas OOD AUROC measures how well an uncertainty score separates a specified shift. For N evaluated binary records, let $\bar{p}_i$ denote the

averaged predictive probability of the pathogenic class. We compute

$$\begin{aligned} \text{NLL} &= -\frac{1}{N} \sum_{i=1}^N [y_i \log \bar{p}_i + (1 - y_i) \log(1 - \bar{p}_i)], \\ \text{Brier} &= \frac{1}{N} \sum_{i=1}^N (\bar{p}_i - y_i)^2. \end{aligned} \quad (23)$$

Evaluation NLL uses the averaged predictive probability, rather than averaging the NLLs of individual parameter samples. Brier uses the binary scalar convention.

For 15 equal-width bins B_b of pathogenic-class probability, we compute

$$\text{ECE}_{15} = \sum_{b: |B_b| > 0} \frac{|B_b|}{N} \left| \frac{1}{|B_b|} \sum_{i \in B_b} y_i - \frac{1}{|B_b|} \sum_{i \in B_b} \bar{p}_i \right|. \quad (24)$$

Bins are left-closed and right-open, except that the final bin includes 1. Empty bins contribute zero. This is positive-class ECE, not top-label ECE. NLL and Brier are proper scoring rules; ECE is a bin-dependent calibration summary. Molecular $F_{\max}$ is the maximum record-level F1 over distinct predictive-score thresholds:

$$F_{\max} = \max_{t \in \{\bar{p}_i\}_{i=1}^N} \frac{2TP(t)}{2TP(t) + FP(t) + FN(t)}. \quad (25)$$

Here a record is predicted positive if and only if $\bar{p}_i \geq t$. All records tied at a threshold are included together. We assign F1 zero when its denominator is zero. OOD AUROC uses absence from the molecular training gene set as the positive class and larger mutual information as stronger OOD evidence, with half credit for tied uncertainty scores:

$$\text{AUROC}_{\text{OOD}} = P(u_{\text{unseen}} > u_{\text{seen}}) + \frac{1}{2}P(u_{\text{unseen}} = u_{\text{seen}}), \quad (26)$$

where the probabilities refer to independently drawn unseen-gene and seen-gene records from the evaluated populations. Let $u_i = \hat{I}_{q,M}(x_i)$ denote estimated mutual information, or its ensemble analogue. Let $\mathcal{R}_c$ contain the $\text{round}(cN)$ records with smallest u_i , using the implementation’s ordering for ties. With $\hat{y}_i = \mathbf{1}\{\bar{p}_i \geq 0.5\}$, selective balanced accuracy is

$$\begin{aligned} \text{BA}(\mathcal{R}_c) &= \frac{1}{2} \left[\frac{\sum_{i \in \mathcal{R}_c} \mathbf{1}\{y_i = 1, \hat{y}_i = 1\}}{\sum_{i \in \mathcal{R}_c} \mathbf{1}\{y_i = 1\}} \right. \\ &\quad \left. + \frac{\sum_{i \in \mathcal{R}_c} \mathbf{1}\{y_i = 0, \hat{y}_i = 0\}}{\sum_{i \in \mathcal{R}_c} \mathbf{1}\{y_i = 0\}} \right], \end{aligned} \quad (27)$$

provided both denominators are nonzero; realized coverage is $|\mathcal{R}_c|/N$. The reported 99.12% value refers to balanced accuracy at 80% coverage, not to an independently validated clinical referral threshold. Challenge EPCR, molecular predictive probability, and a clinical ACMG/AMP posterior are not interchangeable [30].

See Figure 1 in the main text.

See Figure 3 in the main text.

See Figure 4 in the main text.

See Figure 5 in the main text.

I External patient-level generalization

I.1 Construction without test leakage

We next used Phenopacket Store cases with one or two exact GRCh38 causal variants. A causal hypothesis entered the test only when every causal coordinate was absent from the raw 2023 ClinVar

snapshot and pathogenic in the 2026 snapshot. Repeated causal hypotheses were deduplicated. This retained 719 post-2023 cases from 265 genes: 655 single-variant and 64 two-variant cases.

Each case had one causal hypothesis and nine temporally new benign ClinVar hypotheses. Single-variant decoys were matched on variant type and consequence group when possible. Two-variant decoys contained two benign alleles in one gene. The molecular score of a pair was the geometric mean of its allele probabilities. Ten candidates imply a 10% random top-1 baseline.

The HPO layer used only the 9 October 2023 ontology. Older solved cases were split into 1,388 phenotype-reference cases and 843 weight-selection cases. The validation split included familiar genes and genes entirely removed from the reference. For candidate gene g and patient phenotype set Q , the score was the best symmetric Resnik similarity between Q and an older case for g . Each model selected a molecular/HPO coefficient and missing-gene prior on the 843 older cases; the 719 post-2023 patients remained untouched.

Five independent benign candidate cohorts tested sensitivity to decoy composition. Reported means and sample standard deviations therefore reflect candidate resampling, not five independent patient cohorts.

I.2 Results

See Table 3 in the main text.

See Figure 4 in the main text.

For dense BBB, HPO conditioning improved top-1 accuracy by 23.5–26.1 percentage points in every cohort. Each paired comparison gained 175–194 cases while losing only 2–6, with exact McNemar p -values between approximately 10^{-44} and 10^{-53} . The effect is statistically clear within this benchmark.

I.3 Known-gene and unseen-gene behavior

Of the 719 test cases, 337 had a causal gene represented in older phenotype cases and 382 did not. Dense BBB rose from 34.7% to 88.1% top-1 on the known-gene subset, but stayed near baseline on unseen genes (33.6% to 33.1%). The tuned missing-gene prior prevents a catastrophic penalty when phenotype knowledge is absent, but it does not invent knowledge about novel genes.

See Figure 5 in the main text.

I.4 Phenotype corruption and disease novelty

We froze each model’s HPO mixing coefficient after older-case validation and created three additional versions of the same 719-patient test: removal of 20% of observed HPO terms, removal of 40%, and addition of 20% unrelated terms. No coefficient was retuned on a corrupted test. On the fixed candidate cohort, temporal-consensus top-1 changed from 59.81% with the original HPO terms to 58.83%, 58.55%, and 59.39%, respectively. Wide rank-50 SBNN changed from 57.44% to 57.16%, 56.33%, and 56.88%. Moderate phenotype corruption therefore causes limited degradation in this compact setting.

See Figure 6 in the main text.

Disease novelty remains difficult. Among 390 test cases whose disease identifier was absent from the older reference, temporal consensus reached 34.36% top-1; among the 382 unseen-gene cases it reached 33.77%, compared with 89.32% on 337 seen-gene cases. The robustness result therefore does not erase the central limitation: ontology matching transfers well when older knowledge covers the causal gene or disease, but does not solve gene discovery.

The case-level evaluator records the causal hypothesis independently from the prediction table. If upstream filtering removes the truth, reciprocal rank is set to zero and end-to-end Top- k is a failure. In the current compact experiment, however, the causal hypothesis is inserted by construction, so the observed candidate recall of 1.0 is only a pipeline-integrity check. It is not an estimate of full-genome filtering sensitivity.

J Full-background generalization beyond Track 1

J.1 Motivation and evaluation question

After achieving the maximum Track 1 score, we asked whether the trained ranking models could generalize beyond the challenge child to independent rare-disease cases. We therefore evaluated their ability to prioritize known causal variants in VarPB against exonic candidate backgrounds derived from 107 real genomes. Neither the challenge child nor the VarPB causal answers were used for model training. This experiment complements the compact Phenopacket analysis by testing the models in a substantially larger and more realistic candidate space, allowing us to distinguish success on one solved case from transferable variant-ranking performance.

J.2 From the initial stress test to the revised model

We downloaded the 96.7 MB VarPB background and the two small 2023 causal tables, not the 624 MB convenience archive or precomputed method matrices [7]. A disk-backed SQLite index reduced 17,872,513 sample rows to 891,753 unique background variants. Adding the causal tables produced a union of 894,327 hg19 variants. OpenCRAVAT mapped 893,268; the exporter retained 1,059 failures as explicit missing values, including two causal variants, so all 2,577 causal keys remained evaluable.

The generalization study had two stages. We first applied the submission-era models without changing their training or features. Each method was evaluated on every causal variant crossed with all 107 backgrounds, giving 275,739 trials per method. ClinVar review and training-gene history were neutralized. The original small SBNN reached median rank 522, MRR 0.00213, and Top-100 0.20%; the wide rank-50 SBNN reached median rank 1,562. These results did not end the experiment. They showed that a feature contract designed for compact candidate reranking did not contain enough protein-level information to prioritize causal missense variants against a full exonic background.

We then evaluated REVEL and AlphaMissense as external pretrained comparators. Missing scores were set to zero rather than dropped, preserving all 894,327 keys. REVEL reached median rank 8.5 and Top-20 65.90%; AlphaMissense reached median rank 29 and Top-20 44.24%. This comparison identified the missing information and motivated a revised SBNN that combines the original molecular descriptors with both predictor scores and explicit missingness indicators. The final augmented model, described below, recovered competitive full-background recall while retaining a Bayesian posterior over its learned ranking function.

J.3 Predictor-augmented Bayesian ranking

The revised model is still an SBNN; it is not REVEL or AlphaMissense followed by a conventional dense classifier. The exact ClinVar GRCh38 archive from 26 December 2023 was downloaded from NCBI and matched the SHA-256 already recorded for the temporal experiment. A deterministic filter retained unambiguous pathogenic/likely-pathogenic or benign/likely-benign, biallelic SNVs whose molecular consequence contains `missense_variant`. From 2,340,600 input records, this produced 146,810 labeled missense SNVs: 54,548 pathogenic and 92,262 benign. The same REVEL and AlphaMissense modules annotated 146,786; 24 mapper failures remained as a zero value plus a separate missingness flag. REVEL and AlphaMissense were fixed feature generators: their parameters were neither updated nor treated as Bayesian.

The augmented vector has 54 inputs: the original 50 sequence, consequence, allele-frequency, review, origin, substitution, and training-only gene-history features, followed by the REVEL score, a REVEL-missing indicator, the AlphaMissense score, and an AlphaMissense-missing indicator. The previous wide SBNN used 50 inputs, a balanced cap of 30,000 variants per class, and pointwise pathogenic-versus-benign binary cross-entropy. The revised model uses the complete 146,810-variant labeled missense set and optimizes pairwise ordering. Genes were split 80/20 before fitting, yielding 117,700 training and 29,110 validation variants. VarPB and the challenge child were absent from training, validation, and hyperparameter selection.

For layer ℓ , the model places independent mean-field Gaussian variational posteriors on two low-rank factors and the bias. For $Z \in \{A, B, b\}$, a reparameterized posterior draw is

$$\begin{aligned} Z_\ell^{(m)} &= \mu_{Z,\ell} + \text{softplus}(\rho_{Z,\ell}) \odot \epsilon_{Z,\ell}^{(m)}, \quad \epsilon_{Z,\ell}^{(m)} \sim \mathcal{N}(0, I), \\ W_\ell^{(m)} &= A_\ell^{(m)} B_\ell^{(m)\top}, \quad \ell = 1, 2, 3. \end{aligned} \quad (28)$$

Here $\odot$ denotes elementwise multiplication. The joint latent posterior is

$$q_\phi(A_\ell, B_\ell, b_\ell) = q_\phi(A_\ell) q_\phi(B_\ell) q_\phi(b_\ell).$$

This is an independence assumption among the latent factors and bias. It must not be read as a density for W_ℓ . Instead, $q_\phi^{W_\ell}$ is the pushforward of $q_\phi(A_\ell) q_\phi(B_\ell)$ under $(A_\ell, B_\ell) \mapsto A_\ell B_\ell^\top$, exactly as in Equation 15. This induced distribution is generally non-Gaussian, couples entries of W_ℓ through shared factors, and is singular when the chosen rank is below both layer dimensions. Equation 28 is therefore where the SBNN enters: the sampled dense weight matrix is constructed from sampled low-rank factors rather than receiving an entrywise mean-field posterior. The architecture is 54-200-100-1 with ReLU hidden layers and ranks (50, 50, 1). Its output $s_\phi(x)$ is a scalar pathogenicity logit.

Training uses Bayesian pairwise RankNet. For every minibatch, a pathogenic missense variant x_i^+ is paired with a benign missense variant x_i^- . Both pass through the same posterior draw, and the optimized objective is

$$\begin{aligned} \mathcal{L}(\phi) &= \frac{1}{B} \sum_{i=1}^B \log \left(1 + \exp[-(s_\phi^{(m)}(x_i^+) - s_\phi^{(m)}(x_i^-))] \right) \\ &+ \lambda \sum_{\ell=1}^3 [\text{KL}(q_\phi(A_\ell) \| p(A_\ell)) + \text{KL}(q_\phi(B_\ell) \| p(B_\ell)) \\ &\quad + \text{KL}(q_\phi(b_\ell) \| p(b_\ell))] . \end{aligned} \quad (29)$$

Here, “pairwise” refers to a pathogenic-versus-benign training pair; it does not mean that the neural network consumes a compound-heterozygous allele pair as one input. Using one draw for both training members ensures that the likelihood learns their ordering rather than differences caused by two unrelated posterior samples. The prior is a zero-mean Gaussian scale mixture with mixing probability 0.5, a tunable wide scale, and a narrow scale $\exp(-6)$. The complete $3 \times 2 \times 2 \times 2$ screen varied ranks $\{(16,16,1), (32,32,1), (50,50,1)\}$, KL weights $\{0.001, 0.01\}$, posterior initialization scales $\{0.05, 0.1\}$, and prior scales $\{0.3, 1.0\}$. Widths, gene split, training budget, and selection metrics stayed fixed. The selected model used ranks (50, 50, 1), KL weight 0.001, posterior scale 0.1, and prior scale 1.0. After full retraining, gene-disjoint validation reached Top-1 62.15%, Top-10 96.15%, MRR 0.7449, and mean rank 2.684 among 100 candidates.

Table 16: What changed between the submission-era wide SBNN and the predictor-augmented SBNN.

Component	Submission-era wide SBNN	Predictor-augmented SBNN
Input	50 ClinVar-derived molecular features	Same 50 features plus REVEL, AlphaMissense, and two missingness flags
Training cases	Balanced cap of 30,000 variants per class	All 146,810 eligible labeled missense variants
Objective	Pointwise binary cross-entropy plus KL	Shared-posterior-draw RankNet loss plus factorwise KL
Architecture	50-200-100-1, ranks (50, 50, 1)	54-200-100-1, ranks (50, 50, 1)
Selected regularization	KL 0.01; posterior scale 0.1; prior scale 0.3	KL 0.001; posterior scale 0.1; prior scale 1.0
Purpose	General molecular pathogenicity baseline	Full-background causal-variant prioritization with posterior uncertainty

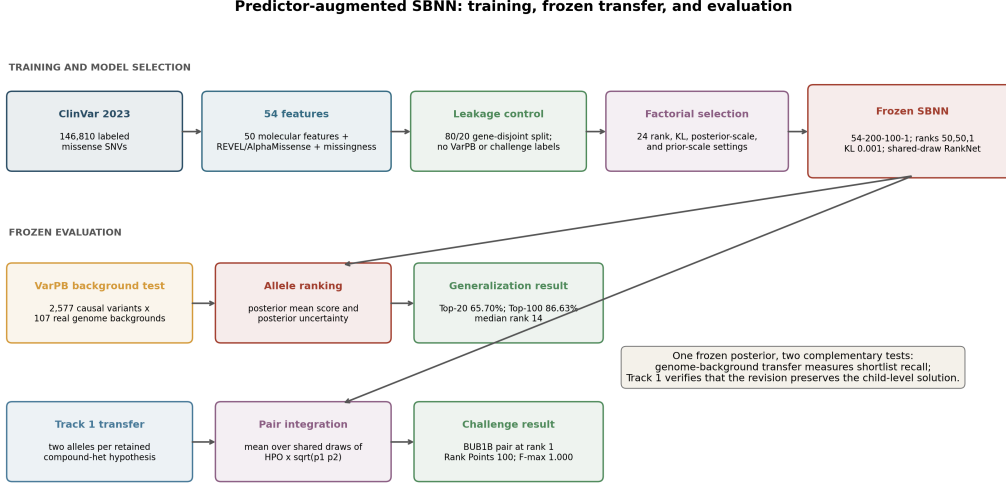


Figure 8: Complete predictor-augmented SBNN workflow. Model selection uses only pre-2023 ClinVar with a gene-disjoint validation split. The frozen posterior is then evaluated independently on VarPB full-genome backgrounds and on the Track 1 compound-heterozygous shortlist.

See Table 4 in the main text.

See Figure 7 in the main text.

The new result is competitive for clinically useful shortlist recall, not uniformly superior. Its Top-20 is within 0.20 percentage points of REVEL and its Top-100 is 1.81 points higher, while its Top-1 and MRR remain lower. It substantially exceeds AlphaMissense at Top-20 and Top-100 but is slightly weaker at Top-1 and MRR. Because REVEL and AlphaMissense are inputs, the contribution is Bayesian evidence integration and uncertainty-aware ranking, not independent rediscovery of their learned biology. VarPB still lacks native HPO and pedigree data and therefore remains a molecular-ranking benchmark rather than a complete clinical diagnostic cohort.

J.4 Transfer of the augmented model to the challenge child

We finally froze the same augmented checkpoint and applied it to the Track 1 candidates. This was a transfer evaluation, not another training or selection stage. Each candidate allele received the same 54-feature representation used in the VarPB experiment. For posterior draw m , the score of compound-heterozygous hypothesis j was

$$e_j^{(m)} = h_j \sqrt{\text{sigmoid}(s_\phi^{(m)}(x_{j1})) \text{sigmoid}(s_\phi^{(m)}(x_{j2}))}, \quad \text{EPCR}_j = \frac{1}{M} \sum_{m=1}^M e_j^{(m)}, \quad (30)$$

where h_j is Bonaventure’s HPO phenotype-similarity score. The same weight draw was used for both alleles, and $M = 128$. Missing predictor values were represented by zero together with the corresponding missingness flag; among the 12 alleles in the strict six-pair shortlist, REVEL and AlphaMissense each covered nine.

The accepted *BUBIB* pair ranked first in both the strict and broad candidate sets. Under the locally reproduced official evaluator, the strict six-row file achieved Rank Points 100 and F-max 1.000, with an F-max threshold of 0.4332283540. The posterior standard deviation of the top pair’s integrated score was 0.00788. Thus the model developed for stronger external transfer preserves the maximum child-level result achieved by every earlier adapter. This result is locally official-equivalent; unlike the selected submission-era wide SBNN, this augmented CSV has not been uploaded to the live evaluator.

J.5 Scalability and uncertainty of the augmented SBNN

For the architecture-matched 54-200-100-1 comparison, the augmented rank-50 SBNN contains 56,204 variational scalars. A dense BBB with identical widths contains 62,402, while five deterministic networks contain 156,005 stored weights. The SBNN therefore uses 9.9% fewer fitted scalars than dense BBB and 64.0% fewer than the five-network ensemble. With Adam, parameters plus two optimizer arrays require 168,612 scalars for the SBNN and 187,206 for dense BBB, excluding activations. Ensemble members can be trained sequentially to lower peak optimizer memory, but all five point estimates are retained at inference. The SBNN checkpoint, which also contains the scaler and gene statistics, is 670,999 bytes.

The modest 9.9% dense-BBB saving should not be overstated: rank 50 is close to the smaller hidden width. The structural advantage appears as width grows. A dense BBB square layer requires approximately $2d^2$ variational scalars, whereas a fixed-rank SBNN requires approximately $4rd$; training gradients and optimizer state follow the same scaling. Posterior Monte Carlo inference still requires repeated forward passes and therefore trades compute for uncertainty estimates.

REVEL and AlphaMissense return deterministic pathogenicity scores in their standard outputs [3, 12]. They do not natively return posterior samples, predictive standard deviation, or epistemic mutual information. The SBNN produces these quantities from one learned posterior. This is an architectural capability, not automatic evidence of good calibration: the augmented model’s gene-disjoint validation mean mutual information was 0.000118 and mean predictive standard deviation was 0.00337, while VarPB supplies no OOD label with which to validate them. A separate distribution-shift benchmark is required before claiming superior uncertainty or referral performance.

Finally, the count above describes the added Bayesian ranker, not the complete annotation stack. Because REVEL and AlphaMissense are inputs, the augmented system inherits their upstream compute and storage. The defensible scalability claim is that evidence integration and posterior uncertainty add only a compact, rank-controlled model, not that the full pipeline replaces its external predictors.

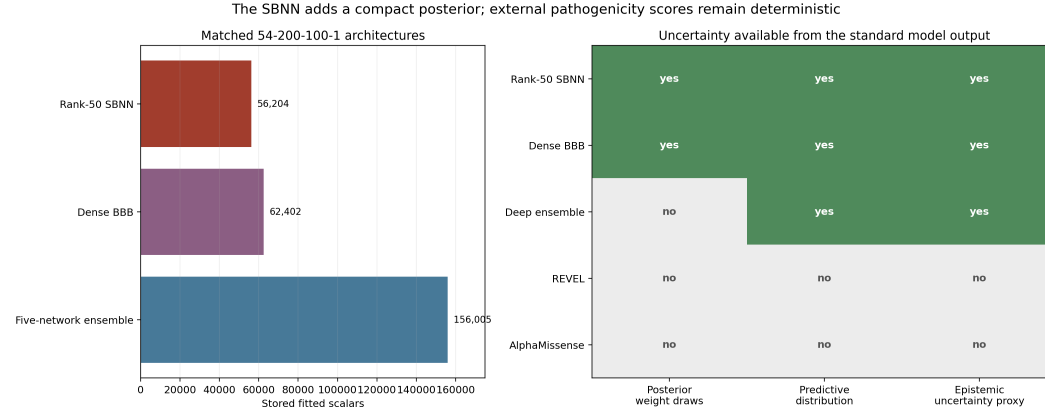


Figure 9: Left: fitted-scalar counts for architecture-matched rankers. Right: uncertainty quantities available directly from each standard model output. Deep ensembles provide member-disagreement uncertainty but require multiple separately fitted networks; REVEL and AlphaMissense provide deterministic scores rather than posterior uncertainty.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [\[Yes\]](#)

Justification: The abstract and Introduction state the challenge result, the temporal and patient-level generalization results, and explicitly scope the work as support for a modular

design rather than a claim of clinical diagnostic sensitivity. Section 4.1 states that the challenge benchmark is saturated and cannot distinguish the compared models.

Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The Limitations section covers unresolved phase and pathogenicity for the leading pair, the constructed ten-candidate benchmark, curation-driven hindsight in temporal labels, unsealed architecture selection, and the small two-variant subset. Appendix E gives the complete itemized list.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate “Limitations” section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren’t acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper presents no theorems or formal results. The equations define the scoring rule, the variational objective, and the evaluation metrics, and are stated rather than proved.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Appendix C gives the stage-by-stage workflow with all computational flags, and Appendix D gives the hyperparameter grids, selected configurations, feature contract, dataset versions with checksums, and benchmark construction. The genome-wide stage requires authorized access to the gated challenge data; the temporal and patient-level experiments use public versioned resources and are reproducible without it.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The repository is withheld during double-blind review and the gated challenge data cannot be redistributed under the hackathon data-sharing terms. All other inputs are public versioned releases identified by date, filename, and checksum in Appendix D, with the full command sequence in Appendix C.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://neurips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://neurips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Section 3 and Appendix D give the split construction, the even/odd Variation ID partition, the 50-feature contract, the full search grids, the selected hyperparameters, and parameter counts for every compared model.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Patient-level results report sample standard deviations over five resampled benign-decoy cohorts, and paired exact McNemar tests are reported for the HPO comparison. The paper states that these resamplings concern the same patients, so the variability reflects candidate composition rather than independent clinical cohorts.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Appendix C reports the institutional cluster setting, thread counts, alignment sharding parameters, and the approximate 160 GB of annotation resources required. Total wall-clock time and accelerator hours for the model sweeps are not reported.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The work uses challenge data under the organizers' terms, does not redistribute gated patient material, and states that the ranking system is an evidence-prioritization tool rather than a substitute for clinical molecular interpretation.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Section 5.2 and the Limitations section address the risk of treating a neural score as a clinical posterior, the danger of presenting missing evidence as established, and the boundary between prioritization and diagnosis. The paper states that the uncertainty scores are not clinically calibrated referral probabilities and that the selective-prediction result is not a validated decision policy.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: Appendix F specifies which categories of material were and were not shared with third-party model services, and the released artifact excludes raw reads, VCF/BAM files, the supplied clinical document, and full patient-level candidate tables.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: ClinVar, gnomAD, HPO, Phenopacket Store, OpenCRAVAT, CADD, REVEL, SpliceAI, and WhatsHap are cited with the specific versions and release dates used, as listed in Appendix D. The gated challenge data are used under the hackathon terms and are not redistributed.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: No new dataset or model is released with this submission. The analysis pipeline is documented in Appendix C and withheld during double-blind review.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The work involves no crowdsourcing and no data collection from human participants by the authors. The patient data were supplied by the challenge organizers under their existing governance framework, and the patient-level benchmark uses previously published Phenopacket cases.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No new human-subjects research was conducted by the authors. The challenge data were released to participants under the organizers' consent and data-sharing framework, and the paper notes that any future clinical submission of patient-derived evidence would require appropriate family and organizer agreement.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not a component of the reported methods; assistance was limited to code implementation, debugging, and drafting. The paper nonetheless discloses this use in the Limitations section and Appendix F.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.