AnomalyModifier: Suppressor Modifier Discovery in Familial Hypercholesterolemia via One-Class Anomaly Detection

Inpyo Hong [1] **Joohyun Han** [1] **Min Kang** [1] **Doyeon Ha** [1]

Abstract

Target identification is the principal bottleneck in developing therapies for rare monogenic diseases. Unaffected carriers, who harbor a known causal variant yet do not develop disease, offer a direct route: their resilience has been partly attributed to suppressor modifier variants that attenuate the causal effect, and pinpointing such modifiers directly nominates new therapeutic targets. We propose **AnomalyModifier**, a one-class anomaly detection framework that treats co-occurring *(causal, non-suppressor)* variant pairs from patient samples as in-distribution and flags out-of-distribution query pairs as suppressor modifier candidates. AnomalyModifier is trained on patient variant pairs with a hypersphere-regularized autoencoder (AE) objective; the hypersphere center and radius, inspired by Deep SVDD, are updated by deterministic rules decoupled from the encoder gradient. From a 1,121,951-pair deduplicated patient corpus (897,559 used for training) for familial hypercholesterolemia (FH) type 1 (OMIM 143890, LDLR causal gene), the model retrieves approved drug targets near the top: **PCSK9 at rank 2, ANGPTL3 at rank 58, MTTP at rank 384, and APOB at rank 101** (4-gene mean rank 136, gene-level AUROC = 0.993) on a synthetic loss-of-function (LoF) benchmark over 19,292 protein-coding genes. On 3,130 ClinVar-curated clinically validated variants, the model achieves **variant-level AUROC = 0.930** (Cohen's $d = +1.94$).

1. Introduction

In rare monogenic diseases, individuals carrying the same causal variant sometimes show attenuated or no symptoms (Cooper et al., 2013; Chen et al., 2016). Identifying suppressor modifier variants in other genes that account for such resilience is a direct route to new therapeutic targets: naturally occurring loss-of-function variants in PCSK9 attenuate the hypercholesterolemia phenotype caused by LDLR variants (Cohen et al., 2006), motivating the clinical development of PCSK9 inhibitors such as evolocumab (Raal et al., 2015).

However, validated suppressor labels are essentially absent, ruling out a supervised formulation; moreover, a variant's effect must be evaluated jointly with its paired causal variant rather than in isolation. We therefore reformulate the task as one-class anomaly detection: every co-occurring *(causal, modifier-candidate)* pair from a patient cohort is guaranteed to be *non-suppressor* by phenotype and constitutes the in-distribution training data, and we hypothesize that pairs that rescue the causal phenotype lie out-of-distribution (Appendices A and B).

Contributions. (1) We propose **AnomalyModifier**, a pair-level one-class anomaly detection framework for suppressor modifier discovery. (2) We introduce decoupled training dynamics in which the encoder is trained by a hypersphere-regularized AE objective over patient variant pairs, while the SVDD updates only the hypersphere center and radius in a gradient-decoupled manner. (3) Anomaly scores are defined at the variant-pair level, enabling variant-level assessment of suppressor modifiers. (4) On a two-stage evaluation protocol (synthetic LoF over 19,292 protein-coding genes and 3,130 ClinVar clinically validated variants), AnomalyModifier retrieves all four drug-validated FH targets within the top-500 (PCSK9 at rank 2, 4-gene mean rank 136), outperforming both supervised gene-feature classifiers and label-free baselines.

2. Method

2.1. Problem formulation

Each patient sample s has a variant set V_s, in which the causal variant $c \in V_s$ is determined a priori. The training data is constructed as

$$\mathcal{D} = \big\{(c_s, m_s) \;\big|\; s \in \text{patients},\ m_s \in V_s \setminus \{c_s\}\big\}.$$

[1]Department of Artificial Intelligence, Research and Development Center, 3billion, Inc., Seoul, Republic of Korea. Correspondence to: Doyeon Ha <biologysaves@gmail.com>.

Accepted at the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026)

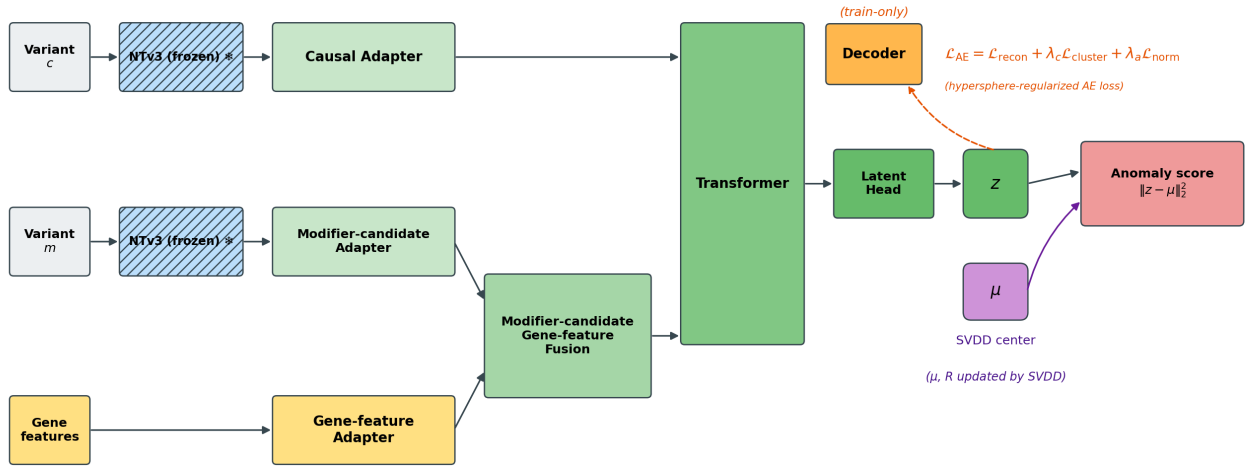

Figure 1. AnomalyModifier architecture. The causal and modifier-candidate variants are encoded by a frozen NTv3, refined by per-stream Adapter blocks, and, for the modifier-candidate stream, fused with eight disease-aware gene features. The Causal Adapter output and the fusion layer output enter a Transformer and a latent head, producing z. Dashed = train-only (training objective in Section 2.3; anomaly score $\|z-\mu\|_2^2$).

Here m_s denotes a *modifier-candidate* variant: a co-occurring variant whose suppressor status is unknown a priori. By construction, every m_s drawn from the patient cohort is non-suppressor by phenotype. Training pairs are produced by an in-house pipeline (Appendix B).

2.2. Model architecture

The overall architecture is illustrated in Figure 1.

- **DNA language model encoder.** A variant sequence in a $\pm 1{,}000$ bp window around the CPRA (chromosome-position-reference-alternate) coordinate is fed to a DNA language model (DLM), and a mean-pooled 768-d variant embedding is obtained. The DLM is the Nucleotide Transformer v3 (NTv3) 100M_post model (Boshar et al., 2025), kept frozen during training (Appendix C).
- **Causal / Modifier-candidate Adapter.** Each stream applies Linear $\rightarrow$ LN $\rightarrow$ GELU $\rightarrow$ Dropout to its NTv3 embedding.
- **Gene-feature Adapter.** Eight disease-aware features of the modifier-candidate gene (Appendix D) are projected to the variant-embedding dimension.
- **Modifier-candidate & Gene-feature Fusion.** The projected gene features are concatenated with the Modifier-candidate Adapter output and linearly mixed.
- **Transformer.** Each input is augmented with a segment embedding (causal=0, modifier-candidate=1) and passed through a Transformer (Vaswani et al., 2017) stack (Appendix E).
- **Latent head.** A projection produces the 32-d latent representation z.

2.3. Training objective: autoencoder-driven encoder + decoupled SVDD

All trainable parameters of AnomalyModifier are updated by the hypersphere-regularized AE loss over patient variant pairs. The hypersphere center μ and radius R are not learnable parameters but buffers, updated by separate deterministic rules (Appendix F).

Hypersphere-regularized AE loss.

$$\mathcal{L}_{\mathrm{AE}}(z,h) = \mathcal{L}_{\mathrm{recon}} + \lambda_c\, \mathcal{L}_{\mathrm{cluster}} + \lambda_a\, \mathcal{L}_{\mathrm{norm}} \tag{1}$$

The three terms are defined as follows and serve distinct roles.

$$\mathcal{L}_{\mathrm{recon}} = \left\|\mathrm{Dec}(z) - \mathrm{sg}(h)\right\|_2^2 \tag{2}$$

$$\mathcal{L}_{\mathrm{cluster}} = \|z - \bar{z}\|_2^2 \tag{3}$$

$$\mathcal{L}_{\mathrm{norm}} = (\|z\|_2 - \tau)^2 \tag{4}$$

Here h is the Transformer output, $\mathrm{sg}(\cdot)$ is stop-gradient, and $\bar{z}$ is the batch mean.

- $\mathcal{L}_{\mathrm{recon}}$ (*information preservation*): forces the 32-d latent z alone to carry enough information to recover the Transformer output h.

- $\mathcal{L}_{\text{cluster}}$ (*patient latent aggregation*): pulls patient latents toward the batch mean $\bar{z}$, an empirical estimator of μ (batch size 2,048; variance $\mathcal{O}(1/B)$). This concentrates the patient distribution around μ, so the inference-time score $\|z-\mu\|_2^2$ remains small for in-distribution pairs and grows for out-of-distribution ones.
- $\mathcal{L}_{\text{norm}}$ (*scale anchoring*): anchors the latent norm near τ to prevent trivial collapse such as $z \to 0$.

Anomaly score. The anomaly score for a variant pair (c, m) is the squared latent distance to the center μ (the patient mean; Appendix G):

$$s(c,m) = \left\|z(c,m) - \mu\right\|_2^2. \tag{5}$$

μ and R are non-learnable buffers: μ is fixed to the training-set patient mean after warmup, and R is the $(1-\nu)$-quantile of patient distances, refreshed each epoch. R^2 is used only as a calibrated threshold ($\mathbb{P}(s \le R^2) \ge 1-\nu$ by construction) delimiting the normal region; it does not appear in s.

Hypersphere updates. After a short warmup, μ is initialized once as the mean of patient latents over the full training set, computed in a single eval-mode pass under `no_grad`, and then frozen. At the start of every post-warmup epoch, the μ-distances of all patient latents are aggregated under `no_grad`, and R is set to the $(1-\nu)$ quantile.

Inference. The anomaly score is $\|z-\mu\|_2^2$. The variant-pair scores within the same gene are averaged to produce a gene-level score.

3. Experimental Setup

Target disease. Familial hypercholesterolemia type 1 (OMIM 143890, LDLR causal gene). We use 1,121,951 deduplicated patient variant pairs (train 897,559 / val 112,196 / test 112,196). All causal variants from LDLR patients are used during training; for evaluation consistency, the top three LDLR causal variants in our in-house diagnostic cohort (`19-11113375-C-G`, `19-11105429-G-A`, `19-11102741-G-A`) are used as fixed evaluation causal variants.

(a) Synthetic LoF benchmark. Suppressor modifier variants typically act by reducing or abolishing gene function; we therefore probe each candidate gene with synthetic LoF variants to simulate the canonical suppressor modifier mechanism. Positive = {PCSK9, ANGPTL3, MTTP, APOB}; these four genes are selected because each is the protein target of an FDA-approved LDL-lowering drug for FH or homozygous FH (HoFH): evolocumab (Raal et al., 2015), evinacumab (Raal et al., 2020), lomitapide (Cuchel et al., 2013), and mipomersen (Raal et al., 2010), respectively. Negative = the remaining GENCODE v49 (Frankish et al., 2023) protein-coding genes. For each gene, we deterministically generate 10 evenly-spaced positions within the first 25% of the CDS (skipping the start codon), and at each position apply both a codon-aware stop_gained substitution and a 1-bp frameshift deletion (up to 20 variants per gene; Figure 2). Pairing each variant with the three LDLR causal variants yields 748,815 (causal, modifier-candidate) pairs after codon-boundary filtering (some positions cannot yield a valid stop_gained). Gene-level scores are the per-gene mean of these pair scores; ranks are reported over 19,292 genes after excluding HGNC-unmapped ENSG-only IDs and readthrough fusion symbols.

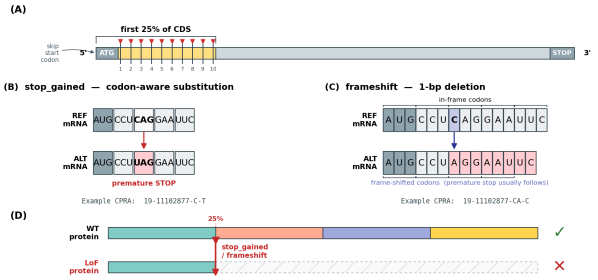

Figure 2. Synthetic LoF benchmark variant design. (A) 10 evenly-spaced CDS offsets within the first 25% of the coding sequence (start codon skipped). (B) At each offset, a codon-aware substitution that yields a premature stop codon (TAA / TAG / TGA). (C) At each offset, a 1-bp frameshift deletion. (D) Truncation at the first 25% disables the protein.

(b) Clinically validated variants benchmark. Variants are curated from ClinVar (Landrum et al., 2020): from the 15-gene LDL-pathway panel (Appendix H), hypoBA / sitosterolemia / Wolman / CESD / LDL-lowering variants form the positive set ($n = 2{,}311$); from 30 non-LDL housekeeping or single-organ disease genes, (likely) pathogenic missense/LoF variants form the negative set ($n = 821$, per-gene cap 50). Variant-level scores are averaged across the three LDLR causal variants.

4. Experiments

4.1. Results on synthetic LoF

AnomalyModifier places **PCSK9 at rank 2, ANGPTL3 at rank 58, MTTP at rank 384, and APOB at rank 101** (4-gene mean rank 136, gene-level AUROC = 0.993, variant-pair AUROC = 0.974; Table 1 and Figure 3; Appendices I and J). The two supervised gene-feature classifiers (LightGBM, GaussianNB) approach but do not match either the mean rank or the gene-level AUROC, and they require LDL

panel labels that AnomalyModifier does not. Their variant-pair AUROC is left blank in Table 1 because they cannot be measured at the variant level. The two label-free single-modality baselines (Linear PCA1, NTv3 cosine concat-pair) trail AnomalyModifier by 1–2 orders of magnitude in mean rank, indicating that neither gene features nor sequence embeddings alone recover the LDL signal. Top-K analysis further surfaces ANGPTL8 at rank 5, a clinically validated LDL-lowering target absent from the 15-gene panel, indicating broader pathway-level recovery (Appendix J).

Table 1. Synthetic LoF results. Gene-level ranks are over 19,292 genes; *Gene AUROC* is computed from those ranks; *Pair AUROC* and AP are at the variant-pair level. Pair AUROC is left blank for methods that cannot be measured at the variant level. Controls: *LightGBM* (Ke et al., 2017) and *GaussianNB* are supervised on the LDL-pathway panel (Appendix H); *Linear PCA1* and *NTv3 cosine concat-pair* are unsupervised gene-feature and sequence-only baselines.

	PCSK9	ANGPTL3	MTTP	APOB	4-gene mean	Gene AUROC	Pair AUROC	AP
AnomalyModifier	**2**	**58**	384	**101**	**136**	**0.993**	**0.974**	**0.0931**
LightGBM (sup.)	261	431	**164**	189	261	0.987	—	0.0096
GaussianNB (sup.)	29	576	475	532	403	0.979	—	0.0060
Linear PCA1	7,675	3,469	4,785	6,529	5,615	0.709	—	0.0005
NTv3 cosine concat-pair	1,575	7,092	11,328	15,163	8,790	0.544	0.544	0.0003

4.2. Results on clinically validated variants

AnomalyModifier achieves **variant-level AUROC = 0.930** (Cohen's $d = +1.94$; Table 2) on the 9,390 paired evaluations (Appendix H).

Table 2. Clinically validated variants evaluation. "Pair" columns are at the variant-pair level over 9,390 pairs (3,130 variants $\times$ 3 LDLR causals); "Var." columns average the 3 causal pair scores per variant ($n = 3{,}130$). d is Cohen's d. LightGBM and GaussianNB are omitted: their supervision panels (Appendix H) overlap with the ClinVar evaluation panels and any reported numbers would reflect label memorization.

	Pair AUROC	Pair AP	Pair d	Var. AUROC	Var. AP	Var. d
AnomalyModifier	**0.900**	**0.9320**	**+1.71**	**0.930**	**0.9390**	**+1.94**
Linear PCA1	0.564	0.8500	+0.58	0.564	0.8500	+0.58
NTv3 cosine concat-pair	0.506	0.7170	−0.29	0.506	0.7180	−0.29

Clinically validated LDL-lowering candidates appear at the top of AnomalyModifier's ranking: PCSK9 hypoBA variants reach rank 1 (median 410) and APOB hypoBA variants reach rank 11 (Appendix K). The two label-free controls collapse to near-random separation (AUROC 0.506–0.564, $|d| \leq 0.58$).

4.3. Modality ablation

To assess the individual contribution of each input modality, we retrained AnomalyModifier with either the DLM or the gene features removed under the same data and hyperparameters (Table 3).

As shown in Table 3: (i) Removing either the DLM or the gene features severely degrades performance (4-gene mean rank 4,858 to 10,957). The full model combining both signals achieves a 4-gene mean rank of 136, roughly 36 to 81$\times$ better. (ii) The DLM-only variant partially recovers PCSK9 (322) but fails on MTTP / APOB; sequence alone cannot reproduce the LDL-pathway signal. (iii) The gene-feat-only variant partially handles ANGPTL3 (435) and MTTP (298) but fails on PCSK9 / APOB. Combining both components is required to retrieve all four positives together.

Table 3. Synthetic LoF results when each modality is removed. Columns are gene-level ranks (cf. Section 3).

Variant	DLM	gene-feat	PCSK9	ANGPTL3	MTTP	APOB	4-gene mean
Best	✓	✓	**2**	**58**	384	**101**	**136**
Gene-feat only	–	✓	12,187	435	**298**	6,512	4,858
DLM only	✓	–	322	9,465	16,594	17,446	10,957

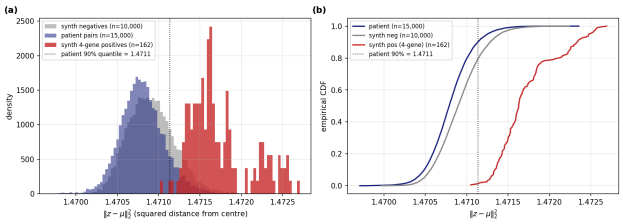

Figure 3. Distribution of the SVDD-center squared distance $\|z - \mu\|^2$ in the 32-d latent. Patients (purple) and synthetic negatives (gray) almost overlap around the mean, while the 4-gene synthetic positives (red) lie at distances exceeding the 90% patient quantile. (a) histogram, (b) ECDF. Positive pairs are cleanly separated from the bulk of the patient distribution.

5. Conclusion

AnomalyModifier reformulates suppressor modifier discovery as one-class anomaly detection over patient variant pairs, sidestepping the absence of validated suppressor labels. Decoupling the encoder gradient from the hypersphere center and radius yields stable training. On familial hypercholesterolemia type 1, the model retrieves all four drug-validated LDL-lowering targets (PCSK9, ANGPTL3, MTTP, APOB) within the top-500 of 19,292 genes (4-gene mean rank 136, gene-level AUROC = 0.993) on the synthetic LoF benchmark, and achieves variant-level AUROC = 0.930 (Cohen's $d = +1.94$) on 3,130 ClinVar clinically validated variants, outperforming supervised classifiers and dominating label-free baselines without any suppressor or LDL-panel labels. Beyond the seeded positives, top-K analysis recovers broader LDL-pathway content, including the clinically validated LDL-lowering target ANGPTL8 at rank 5 (Appendix J). The framework nominates candidate suppressor modifiers without curated gene panels, and is applicable to other monogenic diseases for which such panels are unavailable (Appendix L).

Impact Statement

Rare monogenic diseases collectively affect millions of patients worldwide, yet most lack approved disease-modifying therapies because target identification remains the principal bottleneck. AnomalyModifier addresses this gap by computationally nominating suppressor modifier candidates from patient variant pairs alone, without curated suppressor or pathway labels. Translating the resulting candidates into experimentally testable hypotheses, analogous to the PCSK9 program in familial hypercholesterolemia, could accelerate target identification for disorders that currently lack established therapies.

The model outputs computational hypotheses, not clinical recommendations. Anomaly scores measure distance from the patient distribution rather than suppressor identity, so high-ranking candidates may include non-suppressor biology such as rare gain-of-function variants, comorbid pathogenic variants, or sequence outliers (Appendix L); independent functional validation and replication in external cohorts are required before any diagnostic or therapeutic use. Because the training corpus is drawn from a single-disease cohort (OMIM 143890) and the gene-feature pipeline is anchored to the LDLR causal gene, predictions can inherit cohort-specific ascertainment biases and should not be extrapolated to other populations, ancestries, or diseases without retraining and re-validation.

The training and evaluation sets contain only de-identified variant coordinates and aggregated annotations; no direct identifiers or raw sequencing reads are used. To support reproducibility, code and the trained model will be released publicly (Appendix M); the in-house patient corpus remains governed by data-use restrictions. Any downstream clinical deployment should follow established regulatory frameworks for Software as a Medical Device (SaMD), with appropriate data governance and bias auditing.

References

Aleksander, S. A., Balhoff, J., Carbon, S., et al. The Gene Ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, 2023.

Boshar, S., Evans, B., Tang, Z., et al. A foundational model for joint sequence-function multi-species modeling at scale for long-range genomic prediction. *bioRxiv*, 2025. doi: 10.64898/2025.12.22.695963.

Chen, R., Shi, L., Hakenberg, J., et al. Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. *Nature Biotechnology*, 34(5):531–538, 2016.

Cirulli, E. T. and Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics*, 11(6):415–425, 2010.

Cohen, J. C., Boerwinkle, E., Mosley, T. H., and Hobbs, H. H. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *New England Journal of Medicine*, 354(12):1264–1272, 2006.

Cooper, D. N., Krawczak, M., Polychronakos, C., et al. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Human Genetics*, 132(10):1077–1130, 2013.

Cuchel, M., Meagher, E. A., du Toit Theron, H., et al. Efficacy and safety of a microsomal triglyceride transfer protein inhibitor in patients with homozygous familial hypercholesterolaemia: a single-arm, open-label, phase 3 study. *Lancet*, 381(9860):40–46, 2013.

Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., et al. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2):287–297, 2024.

Frankish, A., Carbonell-Sala, S., Diekhans, M., et al. GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Research*, 51(D1):D942–D949, 2023.

Gaudet, D., Gonciarz, M., Shen, X., et al. Targeting the angiopoietin-like protein 3/8 complex with a monoclonal antibody in patients with mixed hyperlipidemia: a phase 1 trial. *Nature Medicine*, 31(8):2632–2639, 2025.

Gong, D., Liu, L., Le, V., Saha, B., Mansour, M. R., Venkatesh, S., and van den Hengel, A. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1705–1714, 2019.

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

Hojjati, H. and Armanfard, N. DASVDD: Deep autoencoding support vector data descriptor for anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 36(8):3739–3750, 2024.

Ji, Y., Zhou, Z., Liu, H., and Davuluri, R. V. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.

Karczewski, K. J., Francioli, L. C., Tiao, G., et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443, 2020.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3146–3154, 2017.

Kolberg, L., Raudvere, U., Kuzmin, I., et al. g:Profiler: interoperable web service for functional enrichment analysis and gene identifier mapping. *Nucleic Acids Research*, 51(W1):W207–W212, 2023.

Landrum, M. J., Chitipiralla, S., Brown, G. R., et al. ClinVar: improvements to accessing data. *Nucleic Acids Research*, 48(D1):D835–D844, 2020.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

McInnes, L., Healy, J., and Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

McLaren, W., Gil, L., Hunt, S. E., et al. The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1):122, 2016.

Meyers, R. M., Bryan, J. G., McFarland, J. M., et al. Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. *Nature Genetics*, 49(12):1779–1784, 2017.

Milacic, M., Beavers, D., Conley, P., et al. The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Research*, 52(D1):D672–D678, 2024.

Nguyen, E., Poli, M., Faizi, M., et al. HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Raal, F. J., Santos, R. D., Blom, D. J., et al. Mipomersen, an apolipoprotein B synthesis inhibitor, for lowering of LDL cholesterol concentrations in patients with homozygous familial hypercholesterolaemia: a randomised, double-blind, placebo-controlled trial. *Lancet*, 375(9719):998–1006, 2010.

Raal, F. J., Honarpour, N., Blom, D. J., et al. Inhibition of PCSK9 with evolocumab in homozygous familial hypercholesterolaemia (TESLA Part B): a randomised, double-blind, placebo-controlled trial. *Lancet*, 385(9965):341–350, 2015.

Raal, F. J., Rosenson, R. S., Reeskamp, L. F., et al. Evinacumab for homozygous familial hypercholesterolemia. *New England Journal of Medicine*, 383(8):711–720, 2020.

Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., and Kloft, M. Deep one-class classification. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *PMLR*, pp. 4393–4402, 2018.

Sakurada, M. and Yairi, T. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*, pp. 4–11. ACM, 2014.

Szklarczyk, D., Kirsch, R., Koutrouli, M., et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008, 2017.

Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R. V., and Liu, H. DNABERT-2: Efficient foundation model and benchmark for multi-species genomes. In *International Conference on Learning Representations (ICLR)*, 2024.

Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., and Chen, H. Deep autoencoding Gaussian mixture model for unsupervised anomaly detection. In *International Conference on Learning Representations (ICLR)*, 2018.

Appendix

A. Related Work

DNA language models. Large-scale models pretrained on genomic sequences, such as the Nucleotide Transformer family (Dalla-Torre et al., 2024; Boshar et al., 2025), DNABERT (Ji et al., 2021), DNABERT-2 (Zhou et al., 2024), and HyenaDNA (Nguyen et al., 2023), have been used for variant effect prediction. This work uses the InstaDeepAI NTv3 100M_post model (Boshar et al., 2025), frozen, as a variant embedding extractor.

One-class anomaly detection. Autoencoder-based detectors score samples by reconstruction error (Sakurada & Yairi, 2014) or jointly model the latent code and residual with a Gaussian mixture (Zong et al., 2018), but can over-generalize and reconstruct anomalies faithfully (Gong et al., 2019). Deep SVDD (Ruff et al., 2018) instead minimizes the volume of an enclosing latent hypersphere, and DASVDD (Hojjati & Armanfard, 2024) jointly trains an autoencoder with the SVDD objective and scores by reconstruction plus center distance. We decouple the SVDD loss from the encoder gradient and train the encoder with a hypersphere-regularized AE objective over patient variant pairs (Section 2.3); the anomaly score uses only the latent distance, sidestepping the AE over-generalization issue (Appendix G).

Suppressor modifier discovery. Existing work has focused on SNV-level epistasis analysis or PRS-based approaches (Cirulli & Goldstein, 2010). Approaches that jointly leverage sequence context and the paired distribution from a patient cohort remain rare.

B. Patient variant-pair construction

Training pairs are produced by an in-house pipeline operating on a ClickHouse-backed clinical genomics warehouse. For a given OMIM disease, the pipeline (i) resolves disease metadata (inheritance mode, associated genes, required genotype) from an internal disease registry; (ii) assembles the causal-variant set from ClinVar Pathogenic / Likely-pathogenic records (Landrum et al., 2020) restricted to the disease's associated genes (ClinVar review status ≥ 1 gold star; GRCh38 CPRA representation, where CPRA denotes chromosome-position-reference-alternate; benign and conflicting-classification records excluded) together with in-house clinically confirmed variants; (iii) annotates each candidate with the Ensembl Variant Effect Predictor (VEP) (McLaren et al., 2016) (RefSeq canonical `NM_` transcript prioritized via `VEP --canonical`; consequence and impact), gnomAD and in-house allele frequencies (WES / WGS-stratified and merged), and gene-level features with respect to each of the disease's associated genes: Reactome / GO BP pathway Jaccard indices and STRING shortest-path protein–protein interaction (PPI) distances drawn from pre-computed reference tables; (iv) emits one row per (disease, causal, co-occurring) variant triple with its zygosity, supporting samples, and full annotation.

C. DLM ablation: comparison across encoder models

To assess the contribution of the NTv3 100M_post used by AnomalyModifier, we trained two variants under the same data and hyperparameters with only the DLM swapped to DNABERT-2 (Zhou et al., 2024) (117M) and NTv3 8M_pre, keeping the same architecture. Both comparison models are evaluated on the same 4-gene benchmark (Table 4).

Table 4. Synthetic LoF results across DLMs. Rank columns are gene-level ranks (cf. Section 3); *Gene AUROC* is computed from those ranks; *Pair AUROC* and AP are at the variant-pair level.

Encoder	PCSK9	ANGPTL3	MTTP	APOB	4-gene mean	Gene AUROC	Pair AUROC	AP
NTv3 100M_post (AnomalyModifier)	**2**	**58**	**384**	**101**	**136**	**0.993**	**0.974**	**0.0931**
DNABERT-2 (117M)	4,286	6,079	5,285	4,416	5,017	0.740	0.637	0.0003
NTv3 8M_pre	18,188	18,685	18,256	18,401	18,383	0.047	0.046	0.0001

NTv3 100M_post achieves roughly 37$\times$ better 4-gene mean rank than the similarly sized DNABERT-2, while the NTv3 model without post-training fails to produce useful predictions.

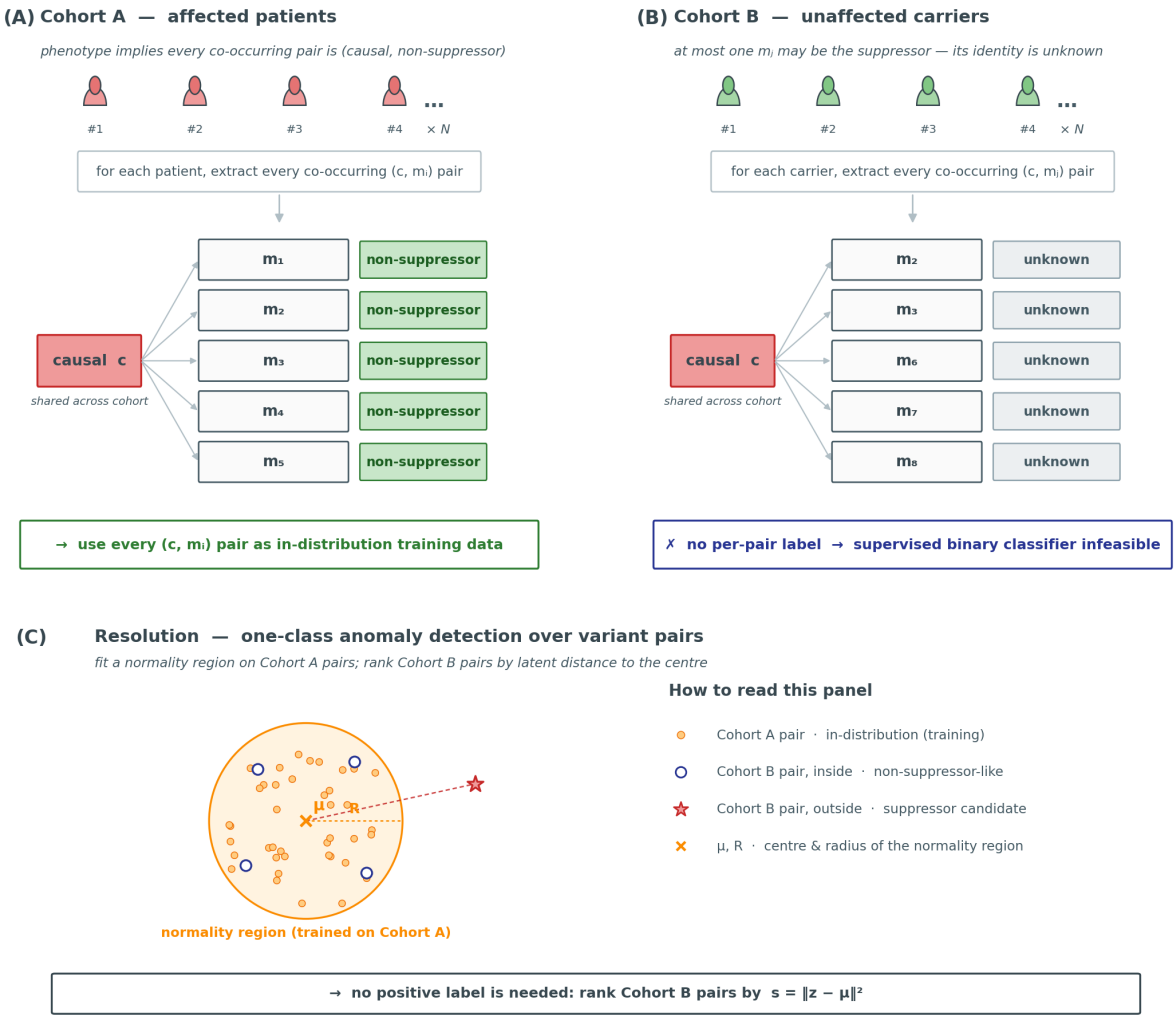

Figure 4. Data asymmetry that motivates the one-class formulation. **(A)** Cohort A (affected patients): every co-occurring pair is *non-suppressor* by phenotype and is used as in-distribution training data. **(B)** Cohort B (unaffected carriers): at most one co-occurring variant may be the suppressor, but its identity is unknown at the pair level, so a supervised label cannot be assigned. **(C)** A normality region (center μ, radius R) is fit on Cohort A pairs in latent space; query pairs are scored by $s = \|z - \mu\|_2^2$, with out-of-distribution pairs flagged as suppressor candidates.

D. Disease-aware gene features

#	feature	source	meaning	class
1	`ppi_distance`	STRING v12 (score $\geq$ 700)	shortest PPI-graph distance to LDLR	LDLR-relative
2	`reactome_shared`	Reactome	number of pathways shared with LDLR	LDLR-relative
3	`go_similarity`	GOA human	GO BP Jaccard similarity to LDLR	LDLR-relative
4	`pli`	gnomAD v4.1.1	probability of LoF intolerance	gene-intrinsic
5	`loeuf`	gnomAD v4.1.1	upper bound of LoF observed/expected	gene-intrinsic
6	`depmap_score`	DepMap CRISPR	mean gene-effect (essentiality)	gene-intrinsic
7	`tissue_breadth`	GTEx v11	number of tissues with TPM $\geq$ 1	gene-intrinsic
8	`mean_log_tpm`	GTEx v11	log-TPM tissue mean	gene-intrinsic

Sources. STRING (Szklarczyk et al., 2023), Reactome (Milacic et al., 2024), Gene Ontology (Aleksander et al., 2023), gnomAD (Karczewski et al., 2020), DepMap CRISPR (CERES) (Meyers et al., 2017), and GTEx (GTEx Consortium, 2020).

E. Transformer layer ablation

We trained 7 models with the number of Transformer layers $\in \{1, 2, 4, 5, 6, 8, 10\}$ under the same data and hyperparameters (Table 5, Figure 5).

Table 5. Layer ablation. Columns are gene-level ranks (cf. Section 3).

Layers	PCSK9	ANGPTL3	MTTP	APOB	4-gene mean
1	469	509	7,194	4,724	3,224
2	203	169	2,341	518	808
4	3	**25**	370	171	142
5	**1**	59	**243**	417	180
6	**1**	146	370	196	178
8	3	128	1,403	**43**	394
10	2	58	384	101	**136**

- Layers 1–2 underfit (4-gene mean rank 808 to 3,224).
- The 10-layer configuration achieves the lowest 4-gene mean rank of 136 across the entire grid, balancing PCSK9 (rank 2) and APOB (rank 101), and is adopted as the main configuration.

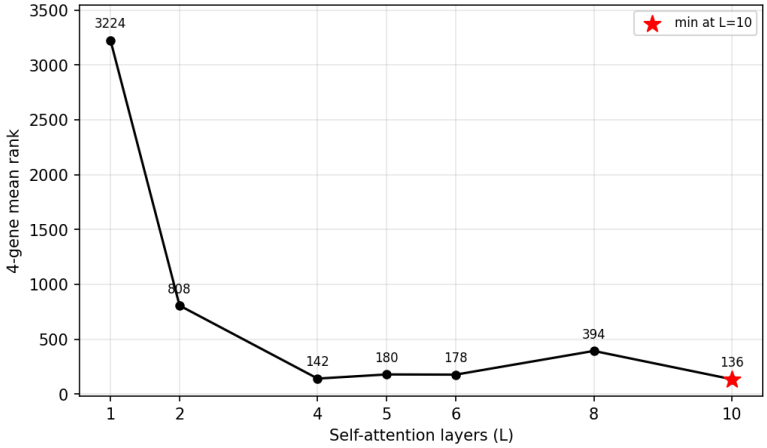

Figure 5. Transformer layer ablation. 4-gene mean rank.

F. Decoupled SVDD vs. Standard SVDD

To validate the effectiveness of our core training methodology (decoupled SVDD with hypersphere-regularized AE), we trained a standard Deep SVDD with an identical architecture on the same data (Table 6).

Table 6. Decoupled SVDD (ours) vs. Standard SVDD on synthetic LoF. Columns are gene-level ranks (cf. Section 3).

Variant	PCSK9	ANGPTL3	MTTP	APOB	4-gene mean
Decoupled SVDD	**2**	**58**	384	**101**	**136**
Standard SVDD	58	234	**332**	399	256

Decoupled SVDD achieves roughly 2× better 4-gene mean rank than standard SVDD on the synthetic LoF benchmark.

G. Implementation details

Encoder configuration.

- DNA sequence window: $\pm 1{,}000$ bp around the CPRA coordinate ($2{,}000$ bp total) with the SNV/indel applied.
- DNA language model: NTv3 100M_post; mean-pooled 768-d embeddings pre-extracted.
- Causal / Modifier-candidate Adapter: $\mathrm{Linear}(768 \to 768) \to \mathrm{LN} \to \mathrm{GELU} \to \mathrm{Dropout}(0.2)$.
- Gene-feature Adapter: 8-d feature $\to$ 768-d projection. Disease-aware feature dropout $= 0.5$.
- Modifier-candidate & Gene-feature Fusion: concat with modifier-candidate embedding $\to \mathrm{Linear}(1{,}536 \to 768)$.
- Transformer: 10 layers, 12 heads, feed-forward network (FFN) 3,072, dropout 0.2.
- Latent head: $768 \to 32$-d latent z.

Loss hyperparameters. $\lambda_c = 0.1$, $\lambda_a = 0.1$, $\tau = 1.0$, $\nu = 0.1$. Warmup $= 2$ epochs (only $\mathcal{L}_{\mathrm{AE}}$ is applied during this period). After center initialization, coordinates with $|\mu_i| < \mathrm{center_eps}$ are clamped to $\pm\,\mathrm{center_eps}$.

Optimization. AdamW (Loshchilov & Hutter, 2019) with lr $= 1 \times 10^{-5}$, weight_decay $= 1 \times 10^{-6}$, cosine decay; gradient clip 1.0; 50 epochs; batch 2,048.

Score function: latent distance vs. reconstruction error. We use the latent distance ($\|z - \mu\|_2^2$) rather than the AE reconstruction error as the anomaly score for two reasons. (i) The latent distance is precisely the quantity that the cluster term in Section 2.3 actively minimizes during training. (ii) It does not pass through the decoder, so it avoids the AE over-generalization issue, in which the decoder reconstructs out-of-distribution inputs accurately and thereby misclassifies anomalies as normal.

H. ClinVar clinically validated variants curation

We curate the evaluation set from NIH/NLM ClinVar through a multi-step pipeline that keeps only clinically reported and validated variants.

1. **Gene panel definition.** *Positive panel (LDL pathway, 15 genes):* PCSK9, APOB, ANGPTL3, ANGPTL4, MTTP, APOC3, ABCG5, ABCG8, NPC1L1, LIPA, LDLRAP1, HMGCR, SORT1, CETP, LPA. These genes are directly involved in LDL cholesterol synthesis, absorption, transport, and receptor-mediated clearance. *Negative panel (30 genes, non-LDL):* ACTB, GAPDH, TP53, INS, ALB, HBB, COL1A1, MYH7, CFTR, BRCA1, RHO, OTOF, OR51E2, DRD2, TAS2R38, MAOA, CRYAA, USH1C, MC1R, GJB2, TGM1, PAH, SLC26A4, ZNF423, F8, FAM161A, TMEM204, SWSAP1, LRRC32, AQP11. These are single-organ disease or housekeeping genes unrelated to the LDL phenotype.
2. **Molecular consequence filter.** Keep only missense, nonsense, stop_gained, frameshift, splice donor/acceptor, and inframe (deletion/insertion).
3. **Coordinate normalization.** Keep only variants with cleanly normalized GRCh38 CPRA (chrom-pos-ref-alt), deduplicated by CPRA.
4. **Positive selection (disease-label based).** From LDL-pathway genes, accept variants whose ClinVar `disease_name` matches an LDL-lowering phenotype (*Familial hypobetalipoproteinemia*, *Hypocholesterolemia*, *Sitosterolemia 1/2*, *Wolman disease*, *Cholesteryl ester storage disease*, *Hypoalphalipoproteinemia*, "Protection against", *LDL cholesterol level QTL*). $\to n = 2{,}311$.
5. **Negative selection (pathogenicity based).** From the non-LDL panel, keep only variants whose ClinVar per-submission record (SCV) pathogenicity is judged Pathogenic or Likely Pathogenic, excluding VUS/benign noise. $\to n = 821$.
6. **Per-gene balancing.** Cap the negative pool at 50 variants per gene to prevent variant-rich genes such as TP53/BRCA1 from dominating the distribution.

Final composition. 2,311 positives + 821 negatives = 3,132 curated variants. Two large deletions (PAH 5,355 bp and BRCA1 2,594 bp, both negative-labeled) exceed the $\pm 1{,}000$ bp DLM input window and are excluded at scoring time, yielding $3{,}130 \times 3 = 9{,}390$ paired evaluations. The mean of variant scores within the same gene is used as the gene-level score.

I. Latent geometry analysis: visual verification of normal vs. anomaly

This appendix shows directly, with figures, what kind of latent geometry and score distribution produce the quantitative metrics (AUROC, Cohen's d, retrieval rank) reported in the main text. We use 12,000 patient pairs from the test split, 8,000 synthetic negatives, 162 synthetic positives (4-gene $\times$ 3-causal $\times$ all-modifier-CDS), and 9,390 clinically validated variant pairs.

I.1. Normal region of the latent space: synthetic positives lie outside the boundary

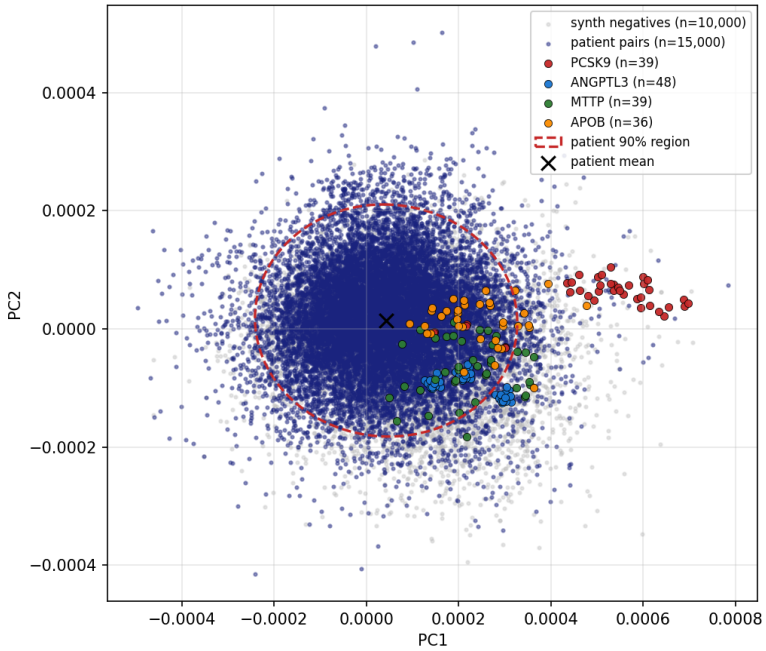

Figure 6. 2D PCA projection of the 32-d latent. Patient pairs (purple) form the normal region around the mean, while the 4-gene synthetic LoF positives (PCSK9 / ANGPTL3 / MTTP / APOB) lie on or outside the 90% Mahalanobis ellipse boundary.

I.2. Latent UMAP of patient + synthetic pairs

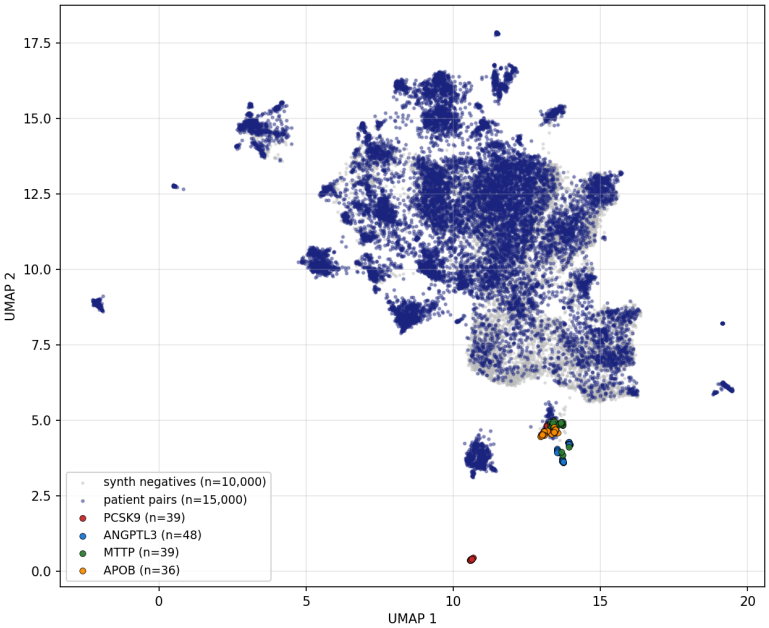

Figure 7. 32-d latent UMAP (McInnes et al., 2018) ($n_{\text{neighbors}} = 30$, $\text{min}_{\text{dist}} = 0.2$). Patient pairs (purple) plus synthetic negatives (gray) form the main cloud, and only the 4-gene synthetic positives occupy a separate region outside the main cloud.

I.3. Latent / disease group / candidate-gene panel distribution of clinically validated variants

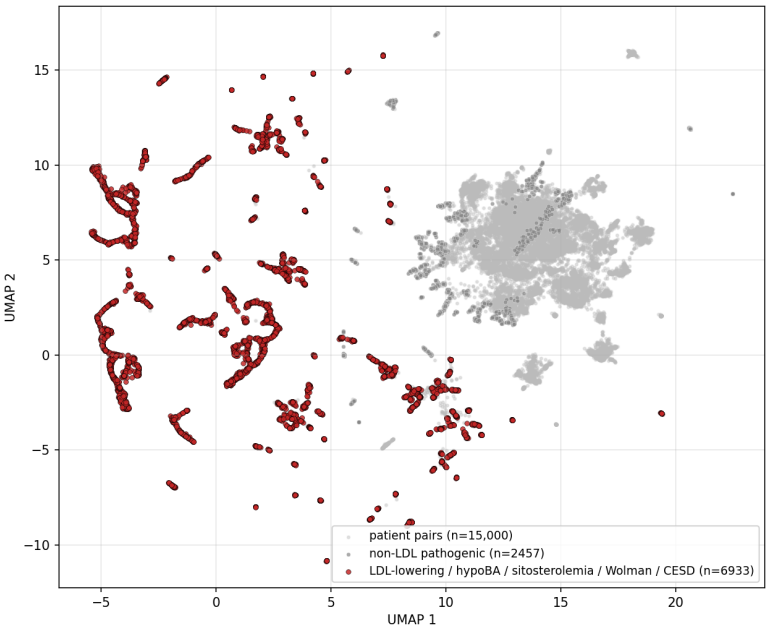

Figure 8. Latent UMAP of patient pairs plus ClinVar clinically validated variants. LDL-lowering positives form satellite clusters around the outer rim of the main patient cloud.

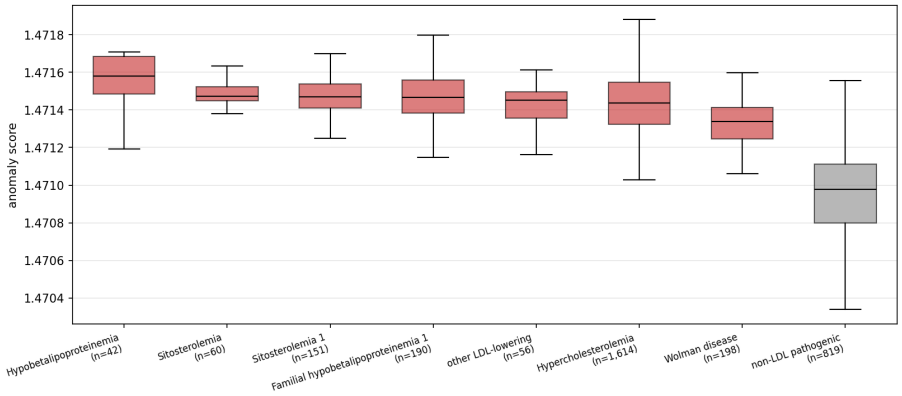

Figure 9. Box plot by disease group for the 3,130 ClinVar clinically validated variants evaluation set (variant-level mean score, averaged over 3 causals). The LDL-lowering / hypoBA / sitosterolemia / Wolman / CESD groups (red) consistently score higher than the non-LDL pathogenic negatives (gray).

I.4. Top-K cumulative recovery / Precision@K

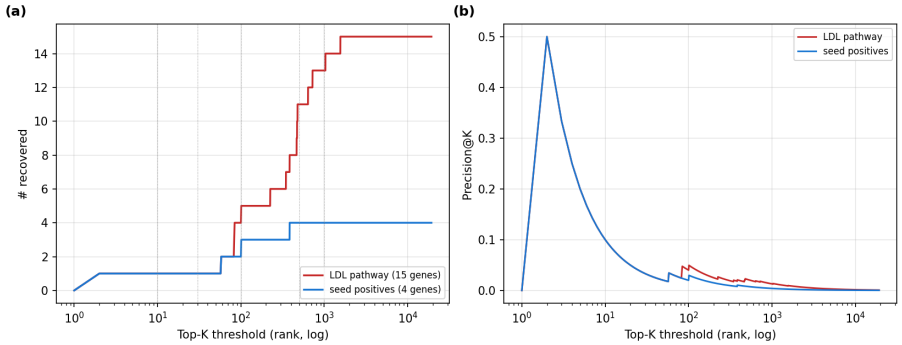

Figure 10. Cumulative recovery curves of the LDL-pathway 15-gene panel and the 4-gene seed positives along the synthetic LoF ranked gene list ($n = 19{,}292$). (a) cumulative number recovered, (b) Precision@K.

All four seed positives are recovered within the top-500, and 73% (11/15) of the LDL-pathway 15-gene panel members also enter the top-500, indicating pathway-level recovery rather than only single-gene retrieval.

J. Top-K candidate analysis and pathway enrichment

We analyze the top-30 of the synthetic LoF 19,292-gene ranking and its LDL-pathway content (Table 7).

Key observation. Ten LDL-pathway / lipid-metabolism / lipoprotein-related genes are retrieved within the top-30: PCSK9 (2), ANGPTL8 (5), RBP4 (6), TTPA (7), APOA1 (8), LRP2 (10), APOM (11), APOF (12), PNPLA3 (24), and RBP1 (27). The top-30 appearance of ANGPTL8, APOA1, APOM, and APOF, none of which are included in the 15-gene LDL-pathway panel (Appendix H), indicates that the model recovers the LDL-pathway signal from the data alone. (RNASEK-C17orf49 at rank 25 is a residual readthrough fusion symbol that may have escaped the Section 3 filter and is unrelated to the LDL-pathway interpretation.)

Notably, ANGPTL8 (rank 5) forms the ANGPTL3/8 complex with ANGPTL3, and the monoclonal antibody LY3475766 targeting this complex lowered LDL-C in a Phase 1 trial of mixed hyperlipidemia (Gaudet et al., 2025)—a related dyslipidemia indication—making ANGPTL8 an LDL-lowering drug target supported by direct clinical evidence, despite its absence from the 15-gene LDL-pathway panel (Appendix H).

Top-K Reactome / KEGG / GO BP / WikiPathways enrichment (g:Profiler g:SCS (Kolberg et al., 2023), custom background = 19,292 ranked genes):

Table 7. Top-30 of the best model (LDL-pathway / lipid-metabolism members in bold). The **Drug** column (✓) marks genes whose protein is the target of a drug with a demonstrated LDL-lowering effect, either FDA-approved for FH (PCSK9: evolocumab (Raal et al., 2015)) or in active clinical-stage development for a related dyslipidemia indication (ANGPTL8: LY3475766 ANGPTL3/8 mAb (Gaudet et al., 2025)).

rank	gene	LDL	Drug	rank	gene	LDL	Drug	rank	gene	LDL	Drug
1	MLDHR			**11**	**APOM**	★		21	UQCC3		
2	**PCSK9**	★	✓	**12**	**APOF**	★		22	REPS1		
3	SMIM10L3			13	UCP1			23	MARVELD1		
4	TAX1BP3			14	HOXC11			**24**	**PNPLA3**	★	
5	**ANGPTL8**	★	✓	15	SLC35D3			25	RNASEK-C17orf49		
6	**RBP4**	★		16	ADM2			26	ACOT4		
7	**TTPA**	★		17	ADH7			**27**	**RBP1**	★	
8	**APOA1**	★		18	HOXB8			28	AGMO		
9	SRD5A2			19	HNRNPA0			29	RSRP1		
10	**LRP2**	★		20	LBX2			30	PMAIP1		

K	source	term	p
200	KEGG	Cholesterol metabolism	5.0×10^{-18}
200	REAC	Plasma lipoprotein assembly/clearance	1.5×10^{-19}
200	REAC	Metabolism of fat-soluble vitamins	2.4×10^{-23}
200	GO:BP	lipid homeostasis	2.4×10^{-12}
200	WP	Metabolic pathway of LDL HDL TG	4.0×10^{-12}
500	KEGG	Cholesterol metabolism	$\mathbf{5.0 \times 10^{-26}}$
500	REAC	Plasma lipoprotein assembly/clearance	3.2×10^{-25}
500	GO:BP	lipid homeostasis	1.4×10^{-21}
500	GO:BP	cholesterol homeostasis	2.9×10^{-20}
500	GO:BP	sterol homeostasis	4.2×10^{-20}
500	WP	Cholesterol metabolism	7.5×10^{-20}
500	WP	Metabolic pathway of LDL HDL TG	7.1×10^{-16}

The top-500 candidate set is strongly enriched for KEGG cholesterol metabolism ($p = 5.0 \times 10^{-26}$), and every LDL/lipid/cholesterol-related term reaches statistical significance at $p < 10^{-15}$ (Figure 11), indicating that the model captures the LDL pathway broadly rather than only the seed positives.

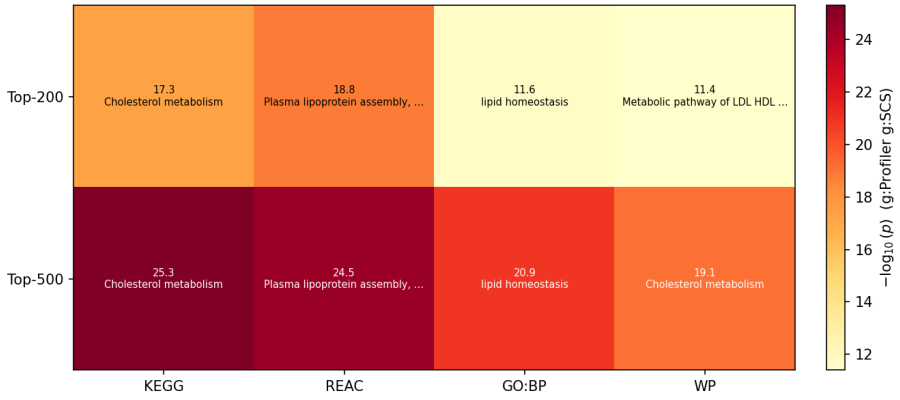

Figure 11. $-\log_{10}(p)$ heatmap of LDL/lipid/cholesterol pathways for the top-K candidate genes (g:Profiler g:SCS, custom background = 19,292 well-annotated genes).

K. Clinically validated variants: per-gene results

Three LDL-pathway positives with sufficient ClinVar coverage are all retrieved within the top 3% of 3,130 ranked variants (Table 8): PCSK9 hypoBA variants reach rank **1** (median **410** / 3,130, top 13%), APOB hypoBA variants reach rank **11**

Table 8. Per-gene rank statistics in the clinically validated variants benchmark. Ranks are over 3,130 variants. ANGPTL3 (0 qualifying variants) and MTTP (1 qualifying variant) are excluded from this table because their ClinVar coverage is too sparse to yield a statistically meaningful per-gene ranking.

Gene	n_{variants}	best variant rank	median rank
PCSK9	38	**1**	**410**
APOB	1,830	11	1,203
ABCG8	153	87	1,028

(median 1,203), and ABCG8 (sitosterolemia) variants reach rank **87** (median 1,028). The recovery of ABCG8, which is not a synthetic LoF benchmark seed, indicates that the model's signal extends beyond the four seeded positives to broader LDL-pathway biology.

L. Limitations

- **Single-disease validation.** FH (OMIM 143890) was chosen because it offers both a sizable patient cohort and an FDA-approved drug-target panel serving as ground truth; generalization to other monogenic diseases is future work.
- **AUROC is a conservative lower bound.** Only the four drug-validated targets are labeled positive, so genuine suppressor candidates retrieved at high ranks (e.g., ANGPTL8 at rank 5; Table 7) count as false positives; better predictions can paradoxically lower the metric.
- **Non-suppressor anomalies.** The one-class signal measures distance from the patient distribution, not suppressor identity, so high-ranking candidates may include non-suppressor biology (rare gain-of-function variants, comorbid pathogenic variants, sequence outliers). As future work, we plan to incorporate an unlabeled cohort of unaffected carriers (asymptomatic causal-variant carriers, enriched for true suppressors) as an additional training signal to sharpen this separation.

M. Code and Data Availability

Code and the trained model will be released as a public repository regardless of review outcome. Evaluation pipelines (synthetic LoF, ClinVar) are reproducible from public resources per Section 3 and Appendix H. The in-house patient corpus is subject to data-use restrictions; access may be coordinated with the corresponding author for collaboration.