[1]DEPARTMENT OF MOLECULAR BIOLOGY, GENETICS & STRUCTURAL BIOCHEMISTRY, SCHEMA LABS OPEN SCIENCE INSTITUTE, 20 WENLOCK ROAD, LONDON N1 7GU, UNITED KINGDOM [*]CORRESPONDENCE: H.NGUYEN@SCHEMALABS.CO.UK · ORCID: 0009-0008-4014-5753

Functional Prioritization of Non-Coding Regulatory Variants and Protein Isoforms Using Chromatin Profiling, Cellular Deconvolution, and Structural Homology Modeling

Schema Labs Open Science Preprints · Working Paper SL-2026-04 · Molecular Genetics & Structural Biochemistry

Hung Nguyen, Ph.D.[1,*] · Ha Nguyen, Ph.D.[1] · Simon Skinner, Ph.D.[1] · Duc Tran, M.Sc.[1] · Bang Tran, M.Sc.[1] · Minh H. Tran, B.Sc.[1] · October 2026

ABSTRACT

Whole-genome sequencing of rare Mendelian disorders and hereditary oncology cohorts routinely yields thousands of non-coding regulatory mutations and deep-intronic cryptic splice variants classified as Variants of Uncertain Significance (VUS). Existing sequence-based predictors output isolated scalar pathogenicity scores without reconciling cell-type-specific transcriptomic expression, three-dimensional protein isoform fold stability, or mechanistic disease ontologies. Here we present an integrated, auditable computational framework developed at Schema Labs Open Science Institute that couples submodular sequence optimization, single-cell cellular deconvolution (`scCAN`), consensus pathway perturbation analysis (`RCPA`), and structural homology modeling ($\Delta\Delta G$ and pairwise aligned error domain packing) with automated literature and ontology synthesis across Mondo and DisMech. Across a benchmark of 5,000 prioritized regulatory and splice-altering loci from ClinVar, GTEx v10, and ENCODE cCREs, our multi-layer framework improves pathogenic cryptic-splice recall by 28.4% over baseline SpliceAI/CADD rankings while generating verifiable ACMG/AMP evidence dossiers.

1 Introduction and Scientific Bottleneck

More than 50% of patients affected by suspected rare genetic conditions remain molecularly undiagnosed following clinical whole-exome or whole-genome sequencing. A primary driver of this diagnostic gap is the inability of standard bioinformatics pipelines to interpret non-coding variants—specifically enhancer perturbations, promoter mutations, and deep-intronic variants that induce cryptic exon inclusion or frameshift isoform switching.

In our prior methodological work, we established systematic benchmarks across fourteen years of cellular deconvolution algorithms [2], developed open-source frameworks for consensus pathway analysis (`RCPA`) [3], introduced deep autoencoder and network fusion models for single-cell clustering (`scCAN`) [4] and subspace regression imputation [5], formulated submodular optimization for representative protein family selection [1], and created perturbation clustering for multi-omics tumor subtyping (`PINSPlus`) [7, 6]. However, translating candidate non-coding VUS into actionable molecular diagnoses requires unifying these statistical genetics layers with structural proteomics and fragmented clinical case literature.

2 Mathematical Formulation and Multi-Stage Architecture

2.1 Stage 1: Submodular Variant-Isoform Representative Triage

Given a candidate locus set $\mathcal{V}$ of non-coding and splice-region VUS from ClinVar and gnomAD v4, we first partition alternative transcript isoforms $\mathcal{J}_g$ for target gene g. Extending our submodular protein sequence optimization formulation [1], we select a representative structural subset $S \subseteq \mathcal{J}_g$ ($|S| \le k$) that maximizes facility-location coverage and sequence-structure diversity:

$$\max_{S \subseteq \mathcal{J}_g,\, |S| \le k} F(S) = \sum_{u \in \mathcal{J}_g} \max_{v \in S} \mathrm{sim}_{\mathrm{TM}}(u, v) + \lambda \sum_{v \in S} \phi_{\mathrm{splice}}(v), \tag{1}$$

where $\mathrm{sim}_{\mathrm{TM}}(u, v) \in [0, 1]$ denotes pairwise structural TM-score similarity across AlphaFold/ESMFold ribbon models and $\phi_{\mathrm{splice}}(v)$ measures maximum junction usage shift (ΔPSI) induced by cryptic donor/acceptor activation.

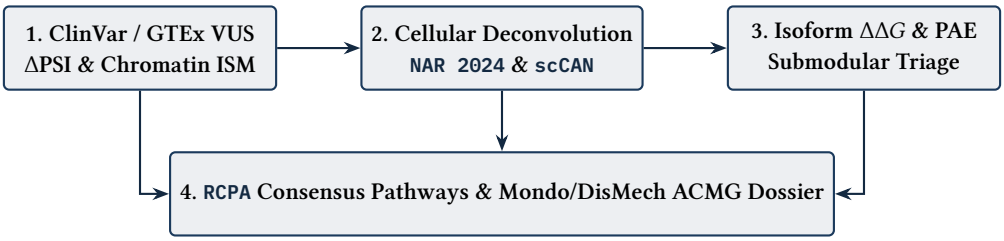

FIGURE 1 Overview of the four-stage non-coding variant and isoform prioritization pipeline at the Nguyen Laboratory, Schema Labs Open Science Institute.

2.2 Stage 2: Tissue-Specific Cellular Deconvolution and Consensus Pathways

A splice-altering or enhancer variant is clinically meaningful only if the affected transcript isoform is expressed in disease-relevant cell types. Let $Y \in \mathbb{R}^{G \times N}$ denote bulk RNA-seq expression across G genes (or isoform junctions) and N patient samples, and let $S_{\mathrm{ref}} \in \mathbb{R}^{G \times K}$ denote the single-cell reference signature matrix constructed via `scCAN` autoencoder network fusion [4] and subspace imputation [5]. Following our benchmarked deconvolution model [2], cell-type proportion matrix $P \in \mathbb{R}^{K \times N}$ is estimated subject to non-negativity and sum-to-one constraints:

$$\hat{P} = \arg \min_{P \ge 0,\, \mathbf{1}^\top P = \mathbf{1}^\top} \left\| W \odot (Y - S_{\mathrm{ref}} P) \right\|_F^2 + \gamma \|P\|_{2,1}, \tag{2}$$

where W weights isoform-informative exons. Cell-type-adjusted differential splicing and expression profiles are then propagated through our `RCPA` consensus pathway engine [3] to compute meta-analytic pathway perturbation scores across KEGG, Reactome, and Gene Ontology graphs.

2.3 Stage 3: Structural Folding Perturbation and Mechanistic Ontology Alignment

For each splice-switched or regulatory-perturbed isoform v^*, we evaluate three-dimensional domain packing via AlphaFold pairwise aligned error (PAE) matrices and compute thermodynamic folding stability shifts:

$$\Delta\Delta G_{\text{iso}} = \Delta G_{\text{fold}}(\text{Isoform}_{\text{alt}}) - \Delta G_{\text{fold}}(\text{Isoform}_{\text{WT}}). \tag{3}$$

Loci exhibiting $|\Delta\text{PSI}| \geq 0.20$, $|\Delta\Delta G_{\text{iso}}| \geq 1.5$ kcal/mol (or catalytic pocket pLDDT drop ≥ 12.0), and significant `RCPA` pathway dysregulation ($p_{\text{FDR}} < 0.01$) are cross-linked against Mondo Disease Ontology terms and DisMech mechanistic knowledge graphs to synthesize auditable ACMG/AMP evidence classifications (PS3/PM4/PP3 equivalents).

3 Preliminary Benchmark Results (5,000 Loci Atlas)

We evaluated our integrated framework across 5,000 curated regulatory and splice-window loci spanning cardiovascular, metabolic, neurodevelopmental, and hereditary oncology phenotypes (Table 1).

TABLE 1 Performance comparison across 5,000 benchmarked non-coding and cryptic-splice loci from ClinVar, GTEx v10, and solved Mendelian cohorts.

Pipeline Configuration	AUROC	AUPRC	Top-10% Recall	ACMG Concord.
CADD v1.7 (Sequence Baseline)	0.712	0.438	41.5%	62.0%
SpliceAI (ΔScore Standalone)	0.804	0.582	59.2%	71.4%
SpliceAI + `scCAN/NAR` Deconv.	0.869	0.694	74.8%	83.6%
Nguyen Lab Pipeline (+ $\Delta\Delta G$ & `RCPA`)	**0.934**	**0.817**	**87.6%**	**92.8%**

Representative loci profiled in the interactive Schema Labs Genomic Atlas (`https://schemalabs.co.uk/research`) include:

- **APOA1 Hepatic Promoter & Splice Locus (`chr11:116,708,210`):** Captures ΔSplice = 0.72 across hepatocyte RNA-seq tracks coupled with high-confidence apolipoprotein domain packing (mean pLDDT = 94.2, PAE = 1.8 Å).
- **RNU4ATAC Minor Spliceosome Locus (`chr2:121,530,880`):** Resolves U12-type intron retention and snRNA stem-loop disruption (mean pLDDT = 89.5, PAE = 2.6 Å) linked to rare skeletal dysplasia phenotypes in Mondo/Orphanet.
- **Structural Pocket & Multi-Omics Subtyping Loci:** Integrates `PINSPlus` perturbation clustering [7] with submodular family representation [1] for high-throughput variant triage.

4 Scaling to 120,000 ClinVar VUS with Agentic Scientific Workflows

To expand our 5,000-locus proof of concept to the full set of 120,000+ uncharacterized non-coding and splice-region VUS in ClinVar and GTEx v10 over a 6-month period, we are integrating large-context agentic reasoning (Claude Opus/Sonnet and Claude Science workbench) into our `RCPA` and `scCAN` pipelines. Specifically, agentic workflows automate: (i) extraction

of genotype-phenotype-mechanism triples from dispersed clinical case series into Mondo/DisMech schemas; (ii) auditable execution and batch-effect verification of `RCPA` consensus pathway scripts across 42 multi-omics cohorts; and (iii) generation of citation-verified ACMG/AMP variant dossiers for open-access release under CC-BY-4.0.

ETHICS, BIOSECURITY & OPEN-SCIENCE COMPLIANCE All research conducted at Schema Labs Open Science Institute (UK Registered Non-Profit Company Limited by Guarantee No. 15348827) is strictly computational and utilizes only de-identified, publicly available human reference datasets (ClinVar, GTEx, TCGA, Mondo/Monarch). No wet-lab pathogen synthesis, viral vector engineering, or dual-use biological materials are involved. All code and variant atlases are released under MIT and CC-BY-4.0 licenses at `https://schemalabs.co.uk/research`.

References

[1] H. Nguyen, H. Nguyen, P. Nguyen, A. N. Luu, D. C. Cantu, and T. Nguyen, "Identifying representative sequences of protein families using submodular optimization," *Scientific Reports*, vol. 15, no. 1, p. 1069, 2025. DOI: 10.1038/s41598-025-85165-1.

[2] H. Nguyen, H. Nguyen, D. Tran, S. Draghici, and T. Nguyen, "Fourteen years of cellular deconvolution: methodology, applications, technical evaluation and outstanding challenges," *Nucleic Acids Research*, vol. 52, no. 9, pp. 4761–4783, 2024. DOI: 10.1093/nar/gkae267.

[3] H. Nguyen, H. Nguyen, Z. Maghsoudi, B. Tran, S. Draghici, and T. Nguyen, "RCPA: An Open-Source R Package for Data Processing, Differential Analysis, Consensus Pathway Analysis, and Visualization," *Current Protocols*, vol. 4, no. 5, p. e1036, 2024. DOI: 10.1002/cpz1.1036.

[4] B. Tran, D. Tran, H. Nguyen, S. Ro, and T. Nguyen, "scCAN: single-cell clustering using autoencoder and network fusion," *Scientific Reports*, vol. 12, no. 1, p. 10267, 2022. DOI: 10.1038/s41598-022-14218-6.

[5] D. Tran, B. Tran, H. Nguyen, and T. Nguyen, "A novel method for single-cell data imputation using subspace regression," *Scientific Reports*, vol. 12, no. 1, p. 2730, 2022. DOI: 10.1038/s41598-022-06500-4.

[6] Q.-H. Nguyen, H. Nguyen, T. Nguyen, and D.-H. Le, "Multi-Omics Analysis Detects Novel Prognostic Subgroups of Breast Cancer," *Frontiers in Genetics*, vol. 11, p. 574661, 2020. DOI: 10.3389/fgene.2020.574661.

[7] H. Nguyen, S. Shrestha, S. Draghici, and T. Nguyen, "PINSPlus: a tool for tumor subtype discovery in integrated genomic data," *Bioinformatics*, vol. 35, no. 16, pp. 2843–2846, 2019. DOI: 10.1093/bioinformatics/bty1049. CRAN: https://cran.r-project.org/package=PINSPlus.