Literature-driven extraction and computational prediction of causal statements linking genetic variants to biological processes, pathways and phenotypes

Jici Jiang, Predrag Radivojac†, Benjamin M. Gyori†
Northeastern University
Boston, MA 02115, United States
†*E-mail: predrag@northeastern.edu; b.gyori@northeastern.edu*

Understanding the mechanistic basis of pathogenic genetic variants requires reconstructing the molecular pathways connecting the variant, via a chain of molecular intermediates, to a disease-causing biological process and phenotype. However, a literature-wide assembly of causal networks connecting variants, molecular pathways, biological processes and phenotypes has not been previously available. To create such a resource, we developed an automated pathway reconstruction approach building on the Integrated Network and Dynamical Reasoning Assembler (INDRA) system which extracts causal mechanistic statements (positive regulation, phosphorylation, complex formation, etc.) by combining structured databases and literature mining. We traversed INDRA statements extracted from publications to identify those describing a genetic variant resulting in a protein point mutation. We then reconstructed directed paths (consisting of one or more linked INDRA statements) connecting this variant to a term representing a biological process, phenotype or disease within the same publication. This resulted in a directed multigraph obtained from 25,862 paths for variants in 2,561 proteins. Each node in this graph corresponds to an ontology-grounded molecular or process term and each edge is explicitly linked to supporting literature evidence, enabling full auditability of inferred mechanisms. To leverage the assembled networks, we trained a classification model to predict likely downstream biological processes or specific disease associations for protein variants. As features to the model, we integrated molecular annotations (including protein sequence features, ClinVar pathogenicity labels, and UniProt domain mappings) in combination with representations from the ESM2 transformer-based protein language model. The performance achieved by this model shows promise for reconstructing causal mechanistic statements associated with function of genetic variants, a framing of the variant effect prediction task that goes significantly beyond simple assessment of pathogenicity. This integrative framework enables the mechanistic interpretation of known variants and prediction of functional relevance for variants lacking prior phenotypic annotation.

Keywords: text mining, machine learning, variant interpretation

1. Introduction

Advances in genomic sequencing have greatly increased the number of genetic variants observed in human populations; however, understanding how these variants affect biological processes and lead to disease remains an ongoing challenge.[1] Current computational approaches

for variant impact prediction are powerful.[2–4] They rely predominantly on sequence information, evolutionary conservation, and structural features, and are trained to predict specific functional alterations (e.g., impact on protein stability) or some form of variant viability, often using pathogenicity information from clinical databases.[5,6] While these methods have achieved widespread adoption in genomic pipelines,[7,8] they continue to exhibit fundamental limitations in deciphering the molecular mechanisms that lead to disease.[9] Specifically, these methods, along with recent studies that aim to establish genotype-phenotype causal relationships,[10] generally cannot reconstruct causal paths from individual variants, through specific molecular events, to downstream biological and disease processes. Moreover, current variant interpretation tools lack the ability to trace predictions back to supporting literature evidence, limiting their explainability and clinical applicability.

To address these challenges, we developed an integrated framework that combines automated path reconstruction with transformer-based machine learning for mechanistic variant interpretation. Building on the Integrated Network and Dynamical Reasoning Assembler (INDRA) system,[11,12] which extracts and assembles causal mechanistic statements from scientific literature and structured databases, we reconstructed mechanistic paths connecting protein variants to biological processes and disease phenotypes. This approach yielded a comprehensive network covering 6,268 variants and 25,862 causal paths across 2,561 proteins, where each causal path (chain of events) is explicitly linked to supporting literature evidence, enabling full auditability of predicted mechanisms. We then trained an attention-based model that integrates protein sequence features, domain annotations, pathogenicity scores, and pathway representations to predict chains of events that lead to both broad functional categories and specific biological process associations.

We proposed two prediction strategies for associating protein variants with biological processes and disease causal paths: a fine-grained model that directly scores all target labels, and a hierarchical model that incorporates intermediate broad functional categories to guide prediction. While both strategies achieve comparable overall performance, the hierarchical model provides a structured framework for organizing outputs and supports more interpretable classification. Incorporating mechanistic paths extracted from the literature substantially improves prediction quality, highlighting the importance of biological context in variant interpretation. Combined with domain and pathogenicity information, these features enable functionally explainable predictions and systematic exploration of variant impact across biological networks. We deployed an interactive web portal for exploring literature-derived mechanistic networks, which enables researchers to investigate the network-level propagation of variant effects. This approach advances mechanistic interpretation of variant effects and supports downstream applications in target prioritization, biomarker discovery, and precision medicine.

2. Methods

2.1. *Extracting causal paths from the literature*

To reconstruct causal paths connecting genetic variants to biological processes, we leveraged the INDRA Database (`https://db.indra.bio`). The INDRA Database contains mechanistic *Statements* extracted from PubMed abstracts and PubMed Central full text articles using

multiple natural language processing systems combined with the content of curated pathway databases, assembled using INDRA.[12] Each INDRA Statement represents an assertion about a causal relationship between biological *Agents* (genes, proteins, small molecules, biological processes, etc.), where an Agent may represent additional states such as mutations and post-translational modifications.[11] Statements can express, for instance that "CDK12 phosphorylates POLR2A on S1896" or "EGFR-G796D activates cell proliferation", with each statement linked to specific supporting evidence sentences and source publications.

From the INDRA Database we extracted all Statements in which an Agent representing the mutated form of a protein (corresponding to a genetic variant) was involved. Our path extraction process identified the shortest directed paths connecting genetic variants to biological process terms through chains of molecular intermediates within individual publications. The extracted paths form a directed graph connecting variants, via molecular intermediates, to disease-causing biological processes or phenotypes, with edges representing causal relationships. An example causal path for the A53T variant in *SNCA* is shown in Fig. 1. Each complete path was constrained to originate from a single publication (rather than drawing on statements from multiple publications) ensuring coherence of context in which the genetic variant and its effects are described.

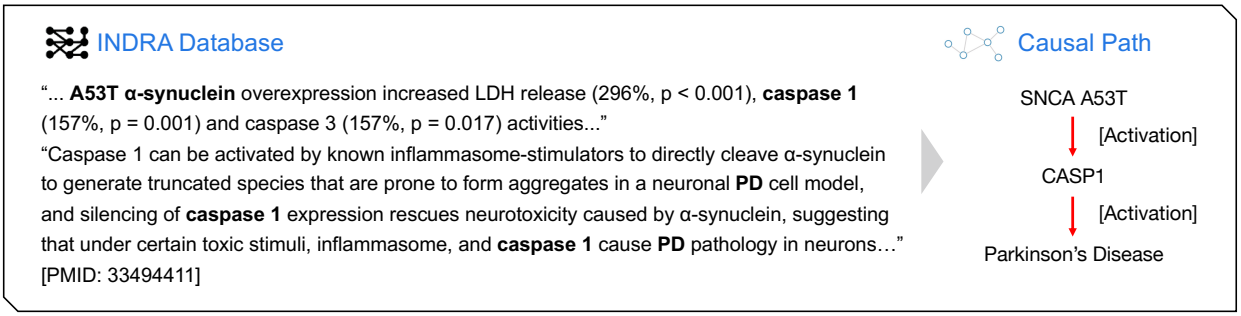

Fig. 1. An example of constructing a causal path for *SNCA* A53T based on literature evidence extracted from INDRA Database, with the cited publication supporting a mechanistic link from variant to Parkinson's disease via *CASP1*.

Through this extraction, we identified 19 types of relations capturing mechanistic interactions between molecular entities (Table 1). The "Identity" relation (ID 0) connects genetic variants to their corresponding proteins to facilitate graph neural network training. We treat the "Complex" (ID 3) relation as undirected, to describe mutual binding rather than directional causality. All other relations are directed, thus reflecting their inherent causal nature.

This approach yielded 25,862 causal records covering 2,561 proteins. The entities and the relations between them form a comprehensive network of causal relationships that serves as the foundation for our knowledge graph construction and subsequent analyses.

2.2. *Variant feature construction*

To ensure effective utilization of sequence information in model training, we validate position-to-amino acid correctness using UniProt Swiss-Prot canonical sequences and splice variant

Table 1. Mechanistic relation types between molecular entities in the knowledge graph

ID	Type	ID	Type	ID	Type	ID	Type
0	Identity	5	DecreaseAmount	10	Glycosylation	15	Palmitoylation
1	Acetylation	6	Deglycosylation	11	Hydroxylation	16	Phosphorylation
2	Activation	7	Demethylation	12	IncreaseAmount	17	Sumoylation
3	Complex	8	Dephosphorylation	13	Inhibition	18	Ubiquitination
4	Deacetylation	9	Deubiquitination	14	Methylation		

isoforms. The validation process filters for human proteins (OX = 9606) with canonical sequences taking precedence when isoform information was imprecise. This approach corrects potential text mining errors in variant position and amino acid assignment, generating protein sequence mappings for accurate variant feature construction, with 20,442 records successfully matching protein sequences across 1,989 proteins to yield 4,511 unique protein-variant pairs.

Next, we incorporate embedding from Evolutionary-Scale Modeling 2 (ESM2),[13] a pretrained protein language model developed for capturing evolutionary and structural information of proteins. Domain features provide binary encoding of the domain types where the variant is located. Amino acid features capture reference and alternative amino acid properties through concatenated embeddings. Position features represent the normalized position within the protein sequence. ClinVar annotations contribute two clinically curated features: pathogenicity scores and review star ratings. Categorical values are mapped to continuous scores as described in Table 2.

Table 2. Mapping of ClinVar pathogenicity and review star annotations to feature values

Value	Pathogenicity	Value	Review star rating
0.99	pathogenic	4	practice guideline
0.90	likely pathogenic	3	expert panel
0.50	uncertain significance/missing	2	multiple submitters, no conflicts
0.10	likely benign	1	single submitter
0.01	benign	0	no review/missing

The final variant feature vector x_v for variant v combines ESM2 embeddings, domain type features, ClinVar annotations, amino acid features, and position features as shown in Fig. 2.

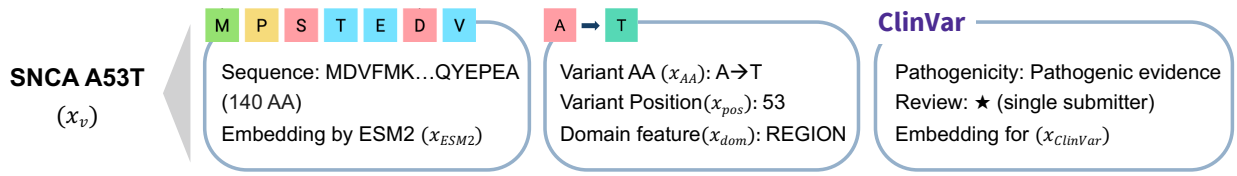

Fig. 2. Constructed feature representation for *SNCA* A53T, incorporating sequence embedding (via ESM2), variant annotations (substitution, domain type), ClinVar annotations.

2.3. *Dataset*

We constructed a hierarchical multi-label dataset to evaluate variant functional predictions across biological processes and diseases. The dataset (Fig. 3A) comprises 4,511 unique protein variants with experimentally validated functional annotations, organized into a two-level hierarchy: 1,085 fine-grained biological process and disease labels were manually mapped to 30 broader categories such as cancer and neoplasms, clinical symptoms, cellular transport, immune response, and cardiovascular diseases.

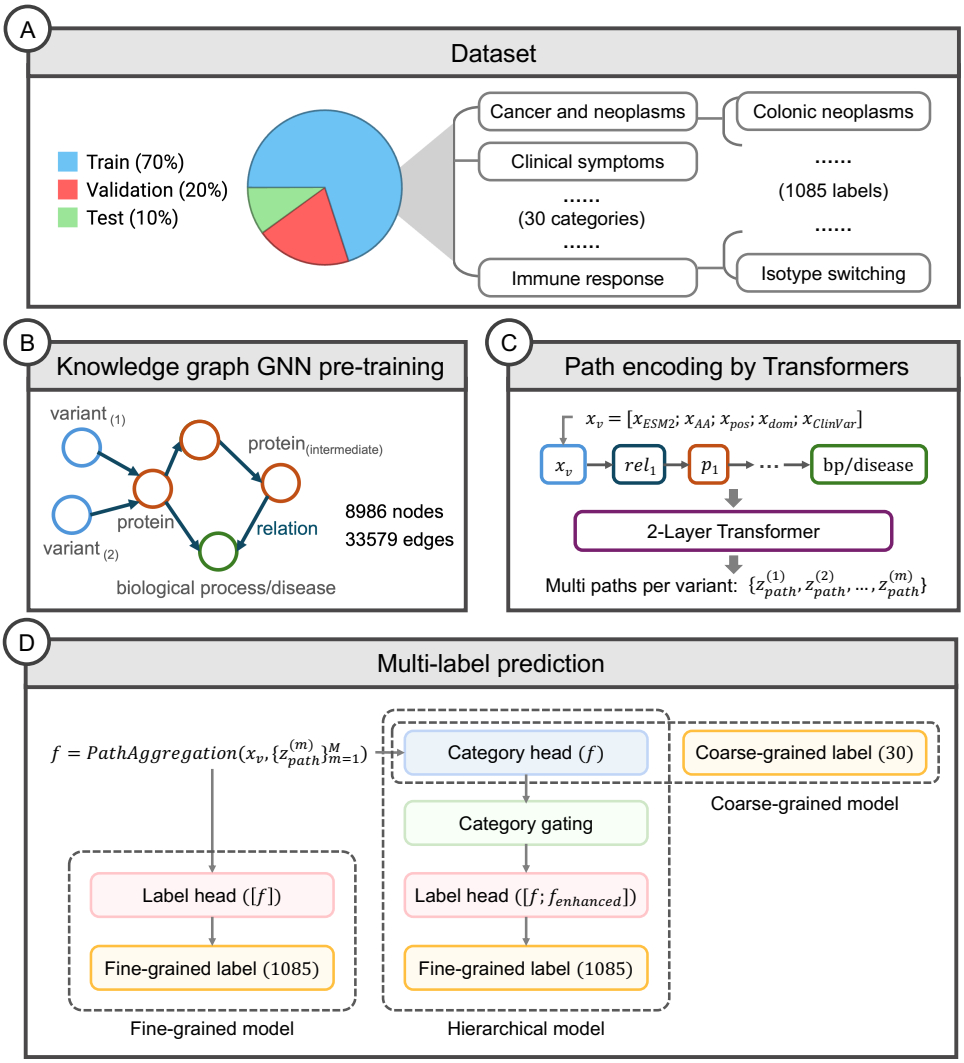

Fig. 3. Prediction framework. (A) Dataset construction with 1,085 fine-grained labels across 30 biological categories; variants are split into training, validation, and test sets. (B) Knowledge graph pre-training comprising 8,986 nodes (molecular entities) and 33,579 edges (mechanistic relations). (C) Causal paths are encoded using a 2-layer transformer. (D) Multi-label prediction: a fine-grained model (left) directly predicts all fine-grained labels; a hierarchical model (right) scores the categories first, then predicts fine-grained labels through the category-guided gating mechanism.

To prevent overfitting to specific proteins or categories, we adopted a stratified data splitting strategy repeated over 30 independent runs. In each run, variants were partitioned into training (70%), validation (20%), and test (10%) subsets. Across the 30 runs, the split sizes were highly consistent (mean ± std): 3,158.4 ± 59.8 for training, 889.5 ± 55.5 for validation, and 463.1 ± 29.3 for testing. To quantify the stability of variant allocation, we computed the Jensen–Shannon (JS) divergence between each variant's observed split distribution and the expected uniform distribution. The average JS divergence was 0.084 ± 0.046. All 30 categories were represented in every training and validation split on average, while the test set maintained 99.3% category coverage, with only occasional omissions in a few runs.

2.4. *Pre-training GNN model for knowledge graph construction*

Using validated sequences as described in Section 2.2, we constructed a knowledge graph $G = (V, E, R)$ comprising $|V| = 8986$ unique nodes representing variants, proteins, biological processes, and phenotypes. These nodes are connected by $|E| = 33579$ edges representing $|R| = 19$ distinct relation types, as shown in Table 1 and Fig. 3B.

We employed a 2-layer Relational Graph Convolutional Network (R-GCN)[14] to learn node embeddings, where the embedding dimensionality is chosen as 256. The initial node embeddings are randomly initialized from a standard normal distribution:

$$\mathbf{H}^{(0)} \in \mathbb{R}^{8986\times 256}, \quad \mathbf{h}_i^{(0)} \sim \mathcal{N}(0,1)$$

The R-GCN forward propagation is performed according to the following update rule:

$$\mathbf{h}_i^{(l+1)} = \sigma\left(\sum_{r\in\mathcal{R}}\sum_{j\in\mathcal{N}_i^r}\frac{1}{c_{i,r}}\mathbf{W}_r^{(l)}\mathbf{h}_j^{(l)}\right)$$

where $\mathbf{h}_i^{(l)}$ is the embedding of node i at layer l, $\mathcal{N}_i^r$ represents the set of neighbors of node i connected by relation r, $\mathbf{W}_r^{(l)}$ is the learnable weight matrix for relation r at layer l, $c_{i,r} = |\mathcal{N}_i^r|$ is the normalization constant, and σ denotes the ReLU activation function.

For knowledge graph completion, we employ the DistMult scoring function[15] $f(s,t,o)$ to evaluate the plausibility of triples:

$$f(s,t,o) = \langle \mathbf{h}_s, \mathbf{r}_t, \mathbf{h}_o\rangle = \sum_{d=1}^{256} h_{s,d}^{(2)} \cdot r_{t,d} \cdot h_{o,d}^{(2)}$$

where s is the subject node index, t is the relation type index, o is the object node index, and d denotes the dimension index of the embedding vectors.

The model parameters are optimized using a softplus margin ranking loss:

$$\mathcal{L}_{link} = \frac{1}{|E|}\sum_{(s,t,o)\in E}\left[\log(1+e^{-f(s,t,o)}) + \frac{1}{k}\sum_{i=1}^{k}\log(1+e^{f(s,t,o_i')})\right]$$

where $k = 2$ is the number of negative samples and o_i' denotes the i-th randomly sampled negative[16] object node.

Upon completion of training, the model produces node embeddings $\mathbb{E}_{node} \in \mathbb{R}^{8986\times 256}$ and relation embeddings $\mathbb{E}_{rel} \in \mathbb{R}^{19\times 256}$.

2.5. *Path encoding by transformers*

To encode causal paths from variants to biological processes or diseases, each path is represented as a sequential token sequence as illustrated in Fig. 3C. The path token $p^{(i)}$ is constructed as:

$$p^{(i)} = [x_v^{(proj)}, e_{r_1}, e_{n_1}, e_{r_2}, e_{n_2}, ..., e_{r_k}, e_{BP}]$$

where i denotes the path index, node embeddings $\mathbb{E}_{node}[n_i]$ are projected to e_{n_i} for each intermediate node n_i, relation embeddings $\mathbb{E}_{rel}[r_i]$ are projected to e_{r_i} for each relation r_i, and e_{BP} represents the final biological process or disease term. Each path $p^{(i)}$ is processed through a 2-layer transformer[17] to obtain the path encoding $z_{path}^{(i)}$.

2.6. *Category-guided multi-label prediction*

For a variant v with m encoded paths $\{z_{path}^{(1)}, z_{path}^{(2)}, ..., z_{path}^{(m)}\}$, an attention mechanism aggregates the path information by computing importance weights. The variant representation queries all paths to focus on the most relevant causal connections, giving the aggregated representation f_v. The prediction process follows a hierarchical approach. First, category-level predictions are generated through a classification head, producing probability distributions $p_{cat} \in [0,1]^{30}$ over 30 broader categories. Category information is then incorporated through a gated mechanism to enhance variant representations. The weighted category representation c_{weight} is computed using learned category embeddings weighted by predicted probabilities:

$$c_{weight} = p_{cat}^T C$$

where $C \in \mathbb{R}^{30\times512}$ contains learned embeddings for each category.

A variant-specific gate $g \in [0,1]^{512}$ determines which dimensions should incorporate category information,[18] which enhances the original variant representation f_v into $f_{enhanced}$ as:

$$f_{enhanced} = f_v + g \odot c_{weight}$$

where $\odot$ denotes element-wise multiplication.

Finally, fine-grained label predictions are generated using both original and enhanced representations to predict across 1,085 specific biological processes and diseases. To address the inherent class imbalance in multi-label biological process prediction, we employ asymmetric loss[19] which applies different focusing parameters for positive and negative samples, emphasizing learning from underrepresented positive classes. This hierarchical framework leverages category-level guidance for accurate fine-grained predictions while handling class imbalance through specialized loss functions.

3. Results

3.1. *Protein variant network*

We constructed a comprehensive protein variant network integrating 25,862 variant-to-disease paths across 2,561 proteins. Each causal edge is fully auditable with explicit links to supporting literature evidence, enabling researchers to trace any variant-disease association to its original experimental sources. We created an interactive web portal (http://variants.indra.bio) that

provides community access for browsing causal paths, visualizing network connections, and accessing supporting evidence. Table 3 summarizes the genes most frequently linked to causal paths in our network.

Table 3. Summary statistics for the top 20 genes ranked by the count of causal paths

Gene	Paths	Variants	BPs/diseases	PMIDs	Gene	Paths	Variants	BPs/diseases	PMIDs
TP53	1298	87	110	261	PIK3CA	180	9	47	55
BRAF	1218	30	130	562	EGFR	179	29	38	78
KRAS	1152	27	162	365	SNCA	178	10	45	52
TARDBP	909	35	54	30	NRAS	173	14	47	63
LRRK2	398	17	77	166	CTNNB1	156	20	36	59
SOD1	305	16	53	141	DNM1L	156	16	41	48
JAK2	231	16	50	125	RAC1	151	16	51	54
HRAS	225	14	56	89	MDM2	136	22	41	37
MAPT	187	19	77	55	GSK3B	128	15	39	40
RPS6KB1	183	4	58	100	PTEN	127	23	39	40

Figure 4 demonstrates a variant-centric mechanistic network derived for *SNCA*, a gene strongly associated with Parkinson's Disease. The visualization integrates multiple *SNCA* variants (e.g., A30P, E46K, G51D, A53E) and their downstream effects through molecular intermediates. Highlighted biological processes include unfolded protein response, autophagy, oxidative stress, and neuroinflammation. Key proteins such as *LRRK2*, *PRKN*, and *TNF* mediate diverse regulatory interactions including activation, inhibition, and post-translational modifications, ultimately converging on phenotypes such as cell death and Parkinson's Disease.

3.2. *Hierarchical model for prediction performance*

We compared three prediction strategies: (i) a fine-grained model that directly predicts all 1,085 biological process and disease labels; (ii) a coarse-grained model that predicts the 30 categories, which collectively cover all 1,085 fine-grained labels; (iii) a hierarchical model that first scores the categories and then uses those scores to guide fine-grained label predictions through a category-guided gating mechanism. Each strategy was trained and evaluated over 30 data splits using the same architecture and training protocol. Table 4 reports mean test performance over the 30 runs.

Table 4. Average test performance across 30 data splits for the three prediction strategies

Model	Micro-P	Micro-R	Micro-F_1	AUROC	AUPRC
fine-grained	**0.684**	0.622	**0.647**	0.913	0.286
coarse-grained	0.959	0.863	0.908	0.959	0.751
hierarchical	0.520	**0.644**	0.573	**0.919**	**0.307**

Bold values indicate the best performance between fine-grained and hierarchical models.

As expected, the coarse-grained model exhibited the strongest performance, achieving a

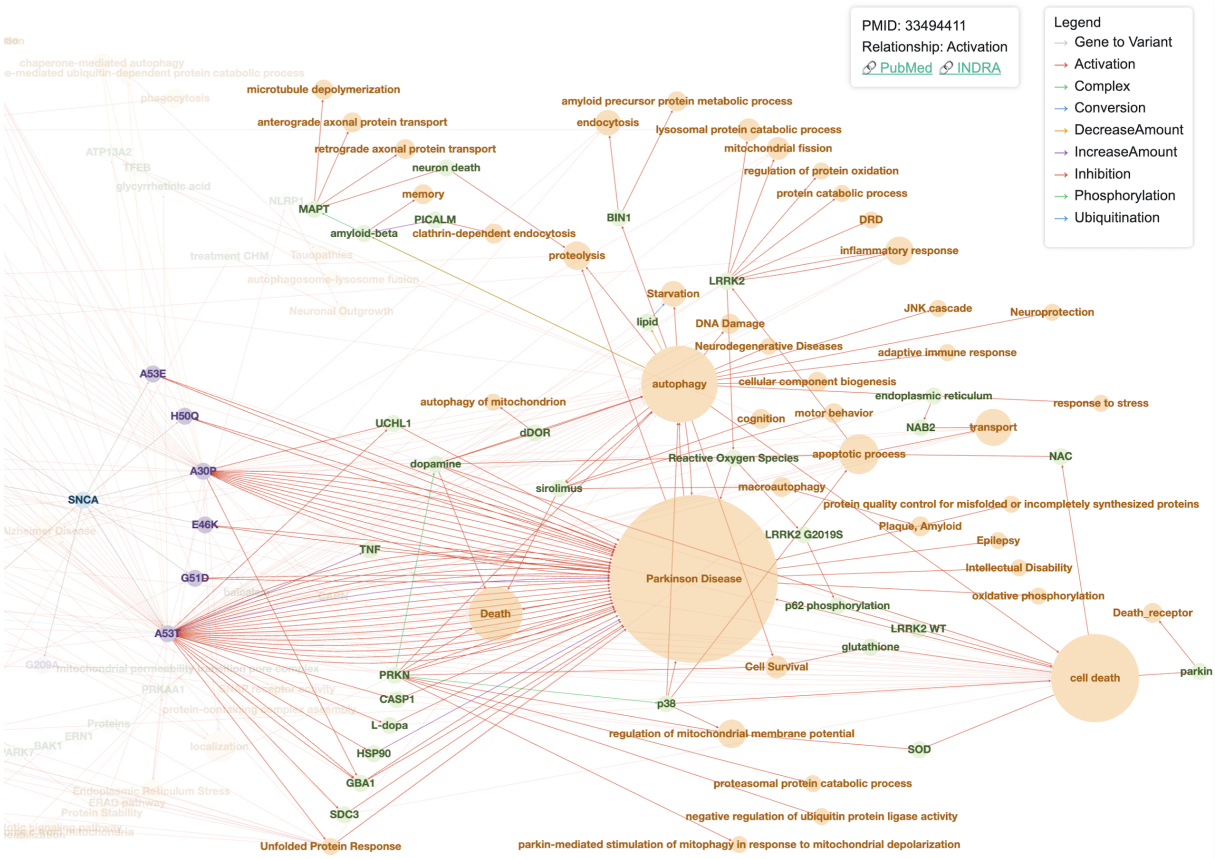

Fig. 4. Interactive visualization of *SNCA* variant network showing path to Parkinson's Disease. Blue nodes represent proteins, purple nodes indicate variants, green nodes denote intermediate processes, and orange nodes mark biological process or disease endpoints.

Micro-F_1 of 0.908. This result reflects the relative ease and reliability of coarse-grained classification, which benefits from stronger label signals and reduced sparsity. A more informative comparison concerns the two models that address the full 1,085-label space. The fine-grained model, which assigns probabilities to every label in a single step, achieves higher precision and an F_1 score, indicating stronger threshold-based classification on the frequent labels. By contrast, the hierarchical model achieves higher recall and slightly better AUROC and AUPRC. Because AUROC and especially AUPRC are less biased toward majority classes, these gains suggest that hierarchical supervision slightly helps find rare but relevant variant-label associations. Training all three models across the 30 random splits (100 epochs each) required approximately 12-14 hours using 4 H100 GPUs.

3.3. *Ablation study to assess feature contributions*

To evaluate the contribution of different features (Fig. 2), we conducted an ablation study in one of the 30 splits combining UniProt domain, ClinVar, and path-based inputs under three prediction strategies. The results are shown in Table 5. Among single-feature inputs, the path-only setting consistently outperformed domain or ClinVar in every prediction mode. With path

features alone, the coarse-grained model reaches an $F_1 = 0.922$, while hierarchical and fine-grained models attain 0.595 and 0.414, respectively. In contrast, using only domain or ClinVar information yields F_1 values below 0.3 across all modes, underscoring the comparatively weak signal carried by these annotations.

Table 5. Ablation results for different feature combinations and prediction strategies

Feature	Model	Micro-P	Micro-R	Micro-F_1	ACC	AUROC	AUPRC
Domain	fine-grained	0.201	0.196	0.199	0.016	0.564	0.056
	coarse-grained	0.462	0.117	0.187	0.059	0.596	0.160
	hierarchical	0.233	0.205	0.218	0.012	0.587	0.059
ClinVar	fine-grained	0.216	0.198	0.207	0.018	0.592	0.056
	coarse-grained	0.416	0.221	0.288	0.065	0.617	0.166
	hierarchical	0.219	0.253	0.235	0.008	0.574	0.059
Path	fine-grained	0.459	0.377	0.414	0.089	0.872	0.130
	coarse-grained	0.960	0.886	0.922	0.811	0.961	0.782
	hierarchical	0.527	0.683	0.595	0.175	0.925	0.279
Domain+ClinVar	fine-grained	0.263	0.205	0.230	0.008	0.594	0.061
	coarse-grained	0.522	0.124	0.200	0.075	0.611	0.172
	hierarchical	0.267	0.240	0.253	0.010	0.593	0.062
Domain+Path	fine-grained	0.501	0.482	0.491	0.142	0.890	0.177
	coarse-grained	0.969	0.895	0.931	0.815	0.974	0.791
	hierarchical	0.529	0.685	0.597	0.181	0.927	0.300
ClinVar+Path	fine-grained	0.563	0.545	0.554	0.258	0.900	0.182
	coarse-grained	0.961	0.911	0.936	0.823	0.964	0.779
	hierarchical	0.521	0.693	0.595	0.199	**0.929**	0.314
Domain+ClinVar+Path	fine-grained	**0.578**	0.562	0.570	**0.268**	0.902	0.200
	coarse-grained	0.968	0.908	0.937	0.827	0.969	0.783
	hierarchical	0.545	**0.695**	**0.611**	0.215	0.926	**0.317**

Bold values indicate the best performance between fine-grained and hierarchical models.

Adding path features to domain type or ClinVar inputs produces substantial gains, especially for fine-grained and hierarchical prediction. For example, adding path features to domain feature raises fine-grained F_1 from 0.198 to 0.491 and hierarchical F_1 from 0.218 to 0.597. A similar trend is observed for the ClinVar + path combination, where hierarchical F_1 goes to 0.595 and AUROC exceeds 0.928.

When all three feature types are combined, the hierarchical model achieves the best overall balance (Micro-$F_1 = 0.611$; AUPRC = 0.317), while the fine-grained model records its highest precision (0.578) and accuracy (0.268). These results confirm that path-derived context provides the dominant functional signal, while ClinVar pathogenicity annotation and UniProt domain type features add some complementary evidence that can further enhance precision and broaden label coverage when integrated with path information.

3.4. *De novo variant impact prediction*

To demonstrate the practical utility of our trained hierarchical model, we developed a prediction tool that annotates protein variants with biological process and disease associations. During inference, the model queries external databases to construct feature vectors, with missing features imputed using appropriate default values, as summarized in Fig. 2 and Table 2. The prediction tool outputs sigmoid-normalized probability scores (0-1) for each label and category, with predictions ranked by confidence to facilitate experimental prioritization.

Figure 5 illustrates a representative prediction for two variants, *VPS35* R524W and *PSEN1* L435F. *VPS35* encodes a retromer complex component involved in endosomal trafficking,[20] and the R524W substitution was not present in our training data. Our model predicts high-confidence associations with translation (62.9%), DNA damage processes (49.7%), and Parkinson's Disease (49.0%). These predictions align with *VPS35* known trafficking functions and established links between variants and neurodegeneration. Similarly, for *PSEN1* L435F, the model highlights associations with localization (74.1%), translation (62.5%), and apoptotic signaling (51.0%), consistent with *PSEN1* role in γ-secretase complex activity and its involvement in Alzheimer's Disease pathology. These examples demonstrate the predictor ability to capture both direct molecular consequences and downstream pathological processes.

Variant	VPS35_R524W		PSEN1_L435F	
Query	Sequence	MPTTQQSP...PIYEGLIL (796 AA)	Sequence	MTELPAPLSYF... QLAFHQFYI (467 AA)
	Domain	REGION	Domain	REGION, MOTIF
	ClinVar	Pathogenicity: not provided Review: 0	ClinVar	Pathogenicity: not provided Review: 0
Predict	biological process/disease	probability	biological process/disease	probability
	translation (protein synthesis degradation)	62.9%	localization (cellular transport)	74.1%
	DNA damage (DNA chromatin processes)	49.7%	translation (protein synthesis degradation)	62.5%
	Parkinson disease (neurological disorders)	49.0%	apoptotic process (cell death survival)	51.0%
	……		……	

Fig. 5. Predicted process and disease for novel variants *VPS35* R524W and *PSEN1* L435F; predictor queries external databases for sequences, domain types, and ClinVar annotations based on the input variant, outputs probability scores.

4. Discussion

The mechanistic paths reconstructed by our framework address a gap in variant interpretation in that it is able to generate specific, experimentally tractable hypotheses about variant function. Unlike pathogenicity scores that classify variants as "pathogenic" or "benign", our predictions identify particular molecular mechanisms that could facilitate targeted experimental validation and therapeutic design. This mechanistic understanding is valuable for interpreting variants of uncertain significance (VUS), where clinical care teams need biological context beyond pathogenicity scores to guide treatment decisions and experimental follow-up.

It is important to mention that this is not the first approach that attempts to predict molecular mechanisms of disease upon mutation. Prediction of disruption of protein stability

has been a long-standing problem in the community.[21,22] The MutPred[23–27] suite of tools has used unsupervised and supervised approaches to predict loss and gain of more than fifty different types of structural and functional disruption (e.g., loss of metal binding, gain of post-translational modification sites). Most recently, predictors capable of predicting disruption of specific protein-protein interactions have also emerged.[28] However, none of these approaches go beyond the effect of the variant on the protein's structure and function. They therefore do not identify a causal path from the variant to a broad biological process disrupted by the variant and ultimately a specific disease phenotype.

The network we constructed illustrates the power of automated literature synthesis for biological knowledge integration. It organizes mechanistic knowledge previously scattered across publications, with each path reconstructed from individual studies describing complete causal chains from variant to phenotype. Since they are constructed from the same publications, these paths provide internal consistency and minimize potential confounding effects from unrelated variants in the same protein. Such network-level analysis provides insights into complex diseases where multiple molecular pathways contribute to pathogenesis. Thus, our framework offers a distinct perspective on variant interpretation that prior pathogenicity predictors that focus on binary classification or local molecular effects do not capture, in particular, how these effects propagate through biological networks to influence disease outcomes.

The mechanistic insights provided by our framework have direct implications for precision medicine applications.[29] By identifying specific biological pathways disrupted by variants observed in patients, clinicians could select targeted therapies that address underlying molecular mechanisms rather than relying solely on pathogenicity scores. For instance, variants predicted to affect DNA repair pathways might guide the selection of PARP inhibitors, while those affecting metabolic processes could inform dietary or pharmacological interventions. The resolution of predictions can be adapted to specific use cases: 1085 fine-grained mechanistic labels support detailed experimental design for lab-based molecular studies, while broader 30 biological categories may suffice for initial clinical assessment and disease area identification.

While our literature-derived networks capture the latest findings and can identify novel mechanistic relationships not yet represented in curated databases, they are subject to inherent biases in the scientific literature. Well-studied proteins and diseases tend to have richer path coverage, while emerging areas of biology may be underrepresented until sufficient experimental evidence accumulates. The reliance on literature-curated pathways may limit prediction for novel variants lacking prior experimental characterization. Future developments should integrate multiple knowledge sources, including structural predictions and functional genomics data, to provide more complete mechanistic characterization of variant effects.

Availability

Code and data are available at `https://github.com/gyorilab/indra_variants`.

Acknowledgements

This work was funded by the DARPA ASKEM and ARPA-H BDF programs HR00112220036 (BMG), and the NIH award U01HG012022 (PR).

References

1. M. Arnaudi, M. Utichi, M. Tiberti and E. Papaleo, Predicting the structure-altering mechanisms of disease variants, *Curr Opin Struct Biol* **91**, 102994 (2025).
2. T. A. Peterson, E. Doughty and M. G. Kann, Towards precision medicine: advances in computational approaches for the analysis of human variants, *J Mol Biol* **425**, 4047 (2013).
3. P. Katsonis, K. Wilhelm, A. Williams and O. Lichtarge, Genome interpretation using in silico predictors of variant impact, *Hum Genet* **141**, 1549 (2022).
4. Y. J. Lin, A. S. Menon, Z. Hu and S. E. Brenner, Variant Impact Predictor database (VIPdb), version 2: trends from three decades of genetic variant impact predictors, *Hum Genomics* **18**, 90 (2024).
5. M. J. Landrum, J. M. Lee, M. Benson, G. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, J. Hoover, W. Jang, K. Katz, M. Ovetsky, G. Riley, A. Sethi, R. Tully, R. Villamarin-Salomon, W. Rubinstein and D. R. Maglott, ClinVar: public archive of interpretations of clinically relevant variants, *Nucleic Acids Res* **44**, D862 (2016).
6. P. D. Stenson, M. Mort, E. V. Ball, M. Chapman, K. Evans, L. Azevedo, M. Hayden, S. Heywood, D. S. Millar, A. D. Phillips and D. N. Cooper, The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting, *Hum Genet* **139**, 1197 (2020).
7. S. Richards, N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster, W. W. Grody, M. Hegde, E. Lyon, E. Spector, K. Voelkerding and H. L. Rehm, Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology, *Genet Med* **17**, 405 (2015).
8. V. Pejaver, A. B. Byrne, B. J. Feng, K. A. Pagel, S. D. Mooney, R. Karchin, A. O'Donnell-Luria, S. M. Harrison, S. V. Tavtigian, M. S. Greenblatt, L. G. Biesecker, P. Radivojac, S. E. Brenner and ClinGen Sequence Variant Interpretation Working Group, Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria, *Am J Hum Genet* **109**, 2163 (2022).
9. IGVF Consortium, Deciphering the impact of genomic variation on function, *Nature* **633**, 47 (2024).
10. K.-S. Benjamin, K. O. Buduka, P. W. Thomas and E. M. Adilson, Generative prediction of causal gene sets responsible for complex traits, *Proceedings of the National Academy of Sciences* **122**, p. e2415071122 (2025).
11. B. M. Gyori, J. A. Bachman, K. Subramanian, J. L. Muhlich, L. Galescu and P. K. Sorger, From word models to executable models of signaling networks using automated assembly, *Mol Syst Biol* **13**, 954 (2017).
12. J. A. Bachman, B. M. Gyori and P. K. Sorger, Automated assembly of molecular mechanisms at scale from text mining and curated databases, *Mol Syst Biol* **19**, e11325 (2023).
13. Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido and A. Rives, Evolutionary-scale prediction of atomic-level protein structure with a language model, *Science* **379**, 1123 (2023).
14. M. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov and M. Welling, Modeling relational data with graph convolutional networks, *arXiv preprint arXiv:1703.06103* (2017).
15. B. Yang, W. tau Yih, X. He, J. Gao and L. Deng, Embedding entities and relations for learning and inference in knowledge bases, *arXiv preprint arXiv:1412.6575* (2015).
16. A. Bordes, N. Usunier, A. Garcia-Durń, J. Weston and O. Yakhnenko, Translating embeddings for modeling multi-relational data, *Advances in Neural Information Processing Systems* **26**, p. 2787–2795 (2013).
17. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, Attention is all you need, *Advances in Neural Information Processing Systems* **30**, p.

6000–6010 (2017).
18. J. Hu, L. Shen and G. Sun, Squeeze-and-excitation networks, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.
19. E. Ben-Baruch, T. Ridnik, N. Zamir, A. Noy, I. Friedman, M. Protter and L. Zelnik-Manor, Asymmetric loss for multi-label classification, *arXiv preprint arXiv:2009.14119* (2020).
20. D. Hu, Z. Liu and X. Qi, Mitochondrial quality control strategies: Potential therapeutic targets for neurodegenerative diseases?, *Front Neurosci* **15**, 746873 (2021).
21. E. Capriotti, P. Fariselli and R. Casadio, I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure, *Nucleic Acids Res* **33**, W306 (2005).
22. J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau and L. Serrano, The FoldX web server: an online force field, *Nucleic Acids Res* **33**, W382 (2005).
23. B. Li, V. G. Krishnan, M. E. Mort, F. Xin, K. K. Kamati, D. N. Cooper, S. D. Mooney and P. Radivojac, Automated inference of molecular mechanisms of disease from amino acid substitutions, *Bioinformatics* **25**, 2744 (2009).
24. M. Mort, U. S. Evani, V. G. Krishnan, K. K. Kamati, P. H. Baenziger, A. Bagchi, B. J. Peters, R. Sathyesh, B. Li, Y. Sun, B. Xue, N. H. Shah, M. G. Kann, D. N. Cooper, P. Radivojac and S. D. Mooney, In silico functional profiling of human disease-associated and polymorphic amino acid substitutions, *Hum Mutat* **31**, 335 (2010).
25. K. A. Pagel, V. Pejaver, G. N. Lin, H. J. Nam, M. Mort, D. N. Cooper, J. Sebat, L. M. Iakoucheva, S. D. Mooney and P. Radivojac, When loss-of-function is loss of function: assessing mutational signatures and impact of loss-of-function genetic variants, *Bioinformatics* **33**, i389 (2017).
26. K. A. Pagel, D. Antaki, A. Lian, M. Mort, D. N. Cooper, J. Sebat, L. M. Iakoucheva, S. D. Mooney and P. Radivojac, Pathogenicity and functional impact of non-frameshifting insertion/deletion variation in the human genome, *PLoS Comput Biol* **15**, e1007112 (2019).
27. V. Pejaver, J. Urresti, J. Lugo-Martinez, K. A. Pagel, G. N. Lin, H.-J. Nam, M. Mort, D. N. Cooper, J. Sebat, L. M. Iakoucheva, S. D. Mooney and P. Radivojac, Inferring the molecular and phenotypic impact of amino acid variants with MutPred2, *Nat Commun* **11**, 5918 (2020).
28. J. C. Siwek, A. A. Omelchenko, P. Chhibbar, S. Arshad, A. Rosengart, I. Nazarali, A. Patel, K. Nazarali, J. Rahimikollu, J. S. Tilstra, M. J. Shlomchik, D. R. Koes, A. V. Joglekar and J. Das, Sliding window interaction grammar (SWING): a generalized interaction language model for peptide and protein interactions, *Nat Methods* **22**, 1707 (2025).
29. B. Rost, P. Radivojac and Y. Bromberg, Protein function in precision medicine: deep understanding with machine learning, *FEBS Lett* **590**, 2327 (2016).