GenePheno: Interpretable Gene Knockout-Induced Phenotype Abnormality Prediction from Gene Sequences

Jingquan Yan[1*], Yuwei Miao[1*], Lei Yu[2], Yuzhi Guo[1], Xue Xiao[2], Lin Xu[2], Junzhou Huang[1†]

[1]Department of Computer Science and Engineering, University of Texas at Arlington
[2]Quantitative Biomedical Research Center, School of Public Health, University of Texas Southwestern Medical Center
{jingquan.yan, yuwei.miao, yuzhi.guo, jzhuang}@uta.edu, {lei.yu, xue.xiao, lin.xu}@utsouthwestern.edu

Abstract

Exploring how genetic sequences shape phenotypes is a fundamental challenge in biology and a key step toward scalable, hypothesis-driven experimentation. The task is complicated by the large modality gap between sequences and phenotypes, as well as the pleiotropic nature of gene–phenotype relationships. Existing sequence-based efforts focus on the degree to which variants of specific genes alter a limited set of phenotypes, while general gene knockout-induced phenotype abnormality prediction methods heavily rely on curated genetic information as inputs, which limits scalability and generalizability. As a result, the task of broadly predicting the presence of multiple phenotype abnormalities under gene knockout directly from gene sequences remains underexplored. We introduce GenePheno, the first interpretable multi-label prediction framework that predicts knockout-induced phenotypic abnormalities from gene sequences. GenePheno employs a contrastive multi-label learning objective that captures inter-phenotype correlations, complemented by an exclusive regularization that enforces biological consistency. It further incorporates a gene function bottleneck layer, offering human-interpretable concepts that reflect functional mechanisms behind phenotype formation. To support progress in this area, we curate four datasets with canonical gene sequences as input and multi-label phenotypic abnormalities induced by gene knockouts as targets. Across these datasets, GenePheno achieves state-of-the-art gene-centric Fmax and phenotype-centric AUC, and case studies demonstrate its ability to reveal gene functional mechanisms.

Extended version — https://arxiv.org/abs/2511.09512

1 Introduction

Understanding the relationship between genetic information and phenotypic outcomes has long been a fundamental goal in biology. Genetic information, encoded in gene sequences, influences phenotypes through complex yet broadly consistent biological mechanisms across individuals (Feuermann et al. 2025). Predicting the phenotypic outcome from gene information is fundamental and essential to therapeutic discovery (Tang and Khvorova 2024), functional genomics (Zheng et al. 2024), and systems biology (Whiting et al. 2024; Islam et al. 2025). Existing approaches generally follow two directions. Variant effect prediction methods typically take gene sequences as inputs to estimate how specific genetic variants change the magnitude or degree of a limited set of phenotypes. Other methods for predicting large-scale phenotypic abnormalities at the gene or protein level rely heavily on labor-intensive curated information, such as protein–protein or gene–gene interaction networks, which limits their applicability to newly discovered or poorly annotated genes. Moreover, both types of methods offer limited insight into the intermediate functional mechanisms that connect genetic information to phenotypic outcomes. These limitations highlight a critical gap: the need for interpretable methods that can predict large-scale phenotypic abnormalities directly from gene sequences, thereby improving applicability to under-annotated genes and enhancing interpretability of the underlying biological mechanisms.

To address this gap, it is important to understand the biological flow of information from gene sequences to phenotypes. Research into the relationship between genetic information and phenotypic traits generally follows the central dogma of molecular biology. Figure 1 outlines the key biological modalities involved in passing genetic information from molecular-level gene sequences to organism-level phenotypes. At the molecular level, genetic information encoded in the linear sequence of DNA is transcribed into RNA and translated into proteins, whose amino acid composition and three-dimensional conformation determine their biochemical properties and functional roles within the cell (Crick 1970). The functions of genes and their products are systematically organized in a directed acyclic graph (DAG) known as the Gene Ontology (GO) (Consortium 2019). Predicting GO terms from protein sequence or structure has long been an active area of research (Kulmanov and Hoehndorf 2020a; Yuan et al. 2023; Liu, Zhang, and Freddolino 2024; Gligorijevic et al. 2019). Further attempts are made to predict large-scale knockout-induced phenotype abnormalities directly from the GO terms annotated to the corresponding protein (Kulmanov and Hoehndorf 2020b). Besides general GO functions, a key functional property of proteins is their capacity to interact with one another. Protein–protein interaction (PPI) networks, derived from assays such as yeast two-hybrid (Y2H), offer a curated modality of genetic information. Although PPI networks are widely

[*]These authors contributed equally.
[†]Corresponding author.

used in prediction of large-scale phenotype abnormalities, their utility is limited to proteins with experimentally validated interactions (Bi et al. 2023; Liu et al. 2022).

Phenotypic abnormalities are systematically organized in DAGs such as the Human Phenotype Ontology (HPO) and Mammalian Phenotype Ontology (MPO), where nodes represent phenotype abnormalities and edges encode subsumption relationships (Gargano et al. 2024; Smith and Eppig 2009). Due to pleiotropy and overlapped functional mechanisms, phenotypes often show strong dependencies and correlations (Mackay and Anholt 2024). Such dependencies are evident in genes whose functions influence multiple phenotypes, such as those involved in pigmentation, which are also associated with hearing or vision abnormalities (Reissmann and Ludwig 2013). However, most existing studies formulate the multi-label phenotype abnormality prediction as multiple binary classification tasks, neglecting correlations between phenotypes (Bi et al. 2023; Liu et al. 2022). These correlations include logical constraints, where certain phenotype abnormalities are mutually exclusive. For example, in HPO, abnormal muscle tone includes both hypotonia (reduced muscle tone) and hypertonia (increased muscle tone), which are semantically incompatible and should not co-occur. Existing prediction methods overlook these constraints, potentially producing logically inconsistent results.

To address these challenges, we propose GenePheno, the first interpretable multi-label prediction framework that predicts knockout-induced phenotypic abnormalities from gene sequences. GenePheno integrates GO functions at both fine and coarse granularity. Fine-grained GO terms, representing specific biological functions, are input alongside the gene sequence and fused via cross-attention. Coarse-grained GO categories serve as supervision targets at the bottleneck layer, providing human-interpretable concepts that reflect general mechanisms of phenotype formation. We further design a contrastive multi-label objective for phenotype abnormality prediction, capturing inter-phenotype correlations, and apply an exclusivity regularization to enforce biological logic consistency. By explicitly modeling the gene sequences, GO functions, and phenotype abnormalities, our work takes a significant step toward understanding how genetic information encodes large-scale observable traits.

Our key contributions are summarized as follows: (1) To the best of our knowledge, we are the first to formally formulate the deep learning task of predicting gene knockout-induced phenotype abnormality directly from gene sequences. (2) We curated four benchmark datasets with a stratified split to support future research in this area. (3) We propose GenePheno, the first interpretable end-to-end framework that maps gene sequences to multi-label phenotype abnormalities, guided by a biologically motivated objective that captures phenotype correlations and enforces phenotype exclusivity. (4) GenePheno offers an interpretable architecture with a functional bottleneck layer. We perform case studies showing that the resulting interpretations are consistent with the biological understanding of phenotypes. (5) We conduct comprehensive experiments showing that GenePheno achieves strong performance in both gene-centric $F_{\max}$ and phenotype-centric AUC, with ablation studies validating the contributions of each component.

2 Related Works

Phenotype Prediction from Curated Genetic Information Previous studies commonly formulate the multi-label phenotype prediction as multiple binary classification task and their input mostly rely on curated genetic information such as gene expression profiles, GO annotations, or PPI networks. For example, DeepPheno (Kulmanov and Hoehndorf 2020b) predicts HPO terms using protein GO functions and expression data from 53 tissues. HPOFiller (Liu, Mamitsuka, and Zhu 2021) constructs a bipartite graph connecting proteins and HPO terms, using GCN-based embeddings to infer associations. HPODNets (Liu, Mamitsuka, and Zhu 2022) processes three different PPI networks with parallel GCN pipelines and fuses the outputs for HPO prediction. GraphPheno (Liu et al. 2022) encodes protein sequences using Composition–Transition (CT) features as node inputs for a GCN over the PPI network. SSLPheno (Bi et al. 2023) integrates PPI and GO information to construct an attributed gene network and employs self-supervised contrastive learning to learn representations for predicting HPO terms via a downstream DNN classifier. Despite their contributions, these approaches face two fundamental limitations: they require pre-processed modalities, restricting their application to well-studied genes, and they fail to capture the biological interdependence and logical constraints among phenotypes by treating each as an isolated prediction task.

Sequence-based Genetic Property Prediction Recent efforts on sequence-based modeling have primarily focused on predicting GO functions from amino acid sequences. While many studies leverage curated genetic modalities such as PPI networks or 3D structural data to enhance performance, here we highlight approaches that rely primarily on protein sequence inputs. DeepGOPlus (Kulmanov and Hoehndorf 2020a) uses one-hot encoded protein sequences followed by convolution and max-pooling layers to predict GO terms. SPROF-GO (Yuan et al. 2023) encodes sequences with ProtT5 (Elnaggar et al. 2021), pools residue embeddings via gated attention, and refines predictions using homology-based diffusion from DIAMOND (Buchfink, Xie, and Huson 2015). InterLabelGO (Liu, Zhang, and Freddolino 2024) employs a rank-based loss and uses mean-pooled ESM-2 embeddings from the last three layers, followed by DIAMOND-based alignment and k-nearest-neighbor retrieval for GO term prediction.

Existing gene sequence-based methods such as DeepSEA (Zhou and Troyanskaya 2015), Basset (Kelley, Snoek, and Rinn 2016), ExPecto (Zhou et al. 2018), Enformer (Avsec et al. 2021), AlphaMissense (Cheng et al. 2023), EVE (Frazer et al. 2021), GENERator (Wu et al. 2025), and Evo2 (Brixi et al. 2025) emphasize on predicting the effect of local allelic variants by comparing short genomic windows under reference and alternate alleles. These methods differ fundamentally from our task, which involves modeling full-gene knockouts to predict the presence or absence of organism-level phenotypic abnormalities. Nonetheless, their advances in long-range DNA encoding provide valuable foundations for our encoder.

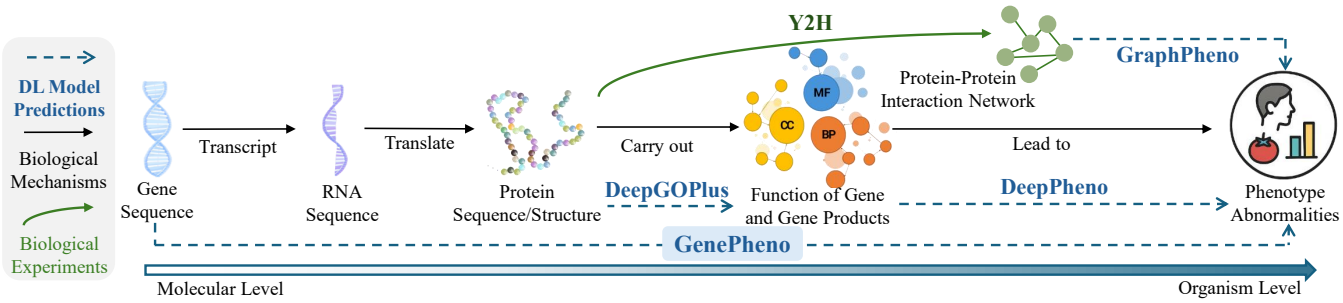

Figure 1: Overview of key biological modalities linking molecular-level DNA sequences to organism-level phenotypic traits and representative methods modeling different stages. Our method, GenePheno, bridges the modality gap in an end-to-end manner.

3 Problem Formulation

Let $\mathcal{G} = \{g_1, g_2, \ldots, g_N\}$ be the set of N genes, and let the training set be $\mathcal{D} = \{(X_i, \mathbf{y}_i)\}_{i=1}^N$, where each $X_i \in \mathcal{X}$ denotes the input features for gene g_i and $\mathbf{y}_i \in \{0,1\}^C$ is the binary label vector for C phenotypes. We aim to learn a mapping $f_\psi : \mathcal{X} \to [0,1]^C$ by solving $\min_\psi \frac{1}{N}\sum_{i=1}^N \mathcal{L}\big(\mathbf{y}_i, f_\psi(X_i)\big)$ where $\mathcal{L}(\cdot,\cdot)$ is an appropriate binary or multi-label classification loss function. Below, we detail the construction of input X_i for two representative methods.

PPI-based Method We encode protein-protein interaction as an undirected graph $G = (V, E)$ with adjacency matrix $A \in \{0,1\}^{|V|\times|V|}$, where $V = \{v_1, v_2, \ldots, v_N\}$ are the proteins corresponds to genes in $\mathcal{G}$. Let $H^{(0)} \in \mathbb{R}^{|V|\times d}$ denote the matrix of initial node features. A graph neural network computes node embeddings $H = \mathrm{GNN}_\phi(A, H^{(0)}) \in \mathbb{R}^{|V|\times d}$ and we write $\mathbf{x}_i = H_{(v_i,:)}$ for the embedding of node v_i. Then we predict multi-label outputs with $\hat{\mathbf{y}}_i = f_\psi(\mathbf{x}_i) \in [0,1]^C$.

Gene Sequence-based Method For each gene $g_i \in \mathcal{G}$, let seq_i denote its gene sequence. We first extract a token-level embedding via a pretrained sequence model f_θ to get the token-level embedding $\mathbf{e}_i = f_\theta(\mathrm{seq}_i) \in \mathbb{R}^{\ell\times d}$ where ℓ is the number of tokens. We then aggregate these token embeddings into a sequence-level embedding $\mathbf{x}_i = \mathrm{AGG}(\{\mathbf{e}_i\})$ and compute the multi-label output $\hat{\mathbf{y}}_i = f_\psi(\mathbf{x}_i) \in [0,1]^C$.

4 Method

In this section, we present the framework of our proposed GenePheno model. We systematically address the aforementioned challenges of label correlation, phenotype exclusivity, and interpretability requirements in the following subsections. Finally, we outline the comprehensive architecture of our learning framework and formulate its overall learning objective.

4.1 Multi-label Prediction

While phenotype prediction is typically formulated as a multi-label classification (MLC) task using binary cross-entropy (BCE) loss, this approach suffers from two limitations. First, BCE loss assumes label independence, failing to capture the inherent co-occurrence patterns in phenotypes that often exist in ontology hierarchies (e.g., "Astigmatism" is an "Abnormality of the eye"). Second, BCE yields higher-order exponential terms involving positive and negative classes (Su et al. 2022), where predominating negative classes will exacerbate the class imbalance. To address these limitations, we derive our supervised contrastive objective from InfoNCE (Oord, Li, and Vinyals 2018), traditionally used in unsupervised settings, to: (i) ties together all positive labels, which implicitly models label dependencies, and (ii) pools scores across positives and negatives globally.

To derive the contrastive MLC loss, we extend the InfoNCE loss to accommodate multiple positive and negative labels. Let $s_i(x)$ represent the classification logit for phenotype label i, Ω_+ the set of positive phenotype labels, and Ω_- the set of negative phenotype labels. Our contrastive MLC loss consists of two components:

$$\mathcal{L}^+_{\mathrm{NCE}} = \sum_{i\in\Omega_+} \log\Big(e^{\frac{s_i(x)}{\tau}} + \sum_{j\in\Omega_-} e^{-\frac{s_j(x)}{\tau}}\Big) - \frac{s_i(x)}{\tau} \quad (1)$$

$$\mathcal{L}^-_{\mathrm{NCE}} = -\sum_{j\in\Omega_-} \log\Big(e^{-\frac{s_j(x)}{\tau}} + \sum_{i\in\Omega_+} e^{\frac{s_i(x)}{\tau}}\Big) + \frac{s_j(x)}{\tau} \quad (2)$$

where τ is a temperature hyperparameter. This formulation encourages clustering of logits within the positive and negative label sets while maximizing the distance between two clusters. Specifically, $\mathcal{L}^+_{\mathrm{NCE}}$ pulls positive labels together while pushing them away from negatives, and $\mathcal{L}^-_{\mathrm{NCE}}$ ensures negative labels remain clustered and distinct from positives. Our final contrastive MLC learning objective is the sum of the positive and negative components:

$$\mathcal{L}_{\mathrm{MLC}} = \mathcal{L}^+_{\mathrm{NCE}} + \mathcal{L}^-_{\mathrm{NCE}} \quad (3)$$

The complete derivation from InfoNCE is detailed in Appendix B. Our contrastive loss effectively captures label dependencies through the explicit separation of positive and negative label clusters, while simultaneously avoiding high-order exponential terms that exacerbate class imbalance. Notably, the ZLPR loss (Su et al. 2022) used in InterLabelGO can be derived from our contrastive loss when $\tau = 1$ with the denominator shift $s(x)/\tau$ omitted. This equivalence confirms our loss inherits ZLPR's established properties of ca-

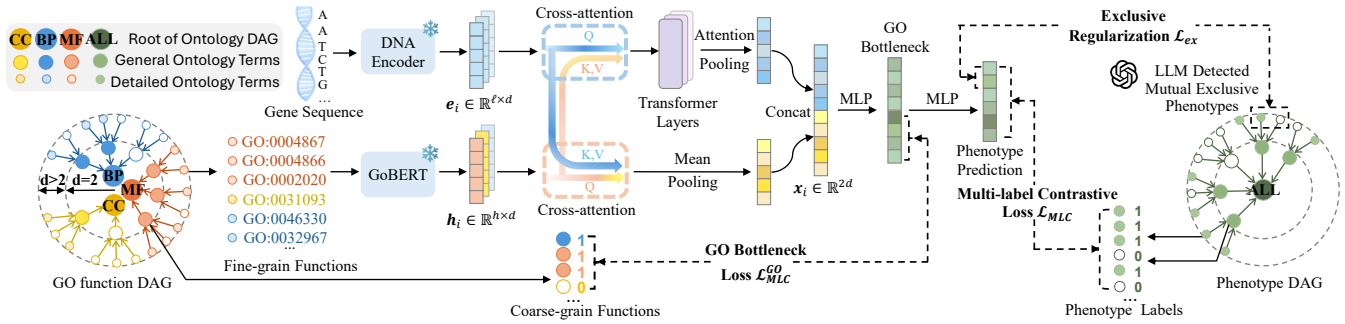

Figure 2: Overview of our learning framework. GO function DAG comprises three subgraphs, each with its own root, whereas the phenotype ontology DAG is single-rooted. Node depth d is the shortest path to the root, with deeper nodes indicating more specific functions or phenotypes. We use GO functions at dual granularity: fine-grain inputs ($d > 2$) for detailed information and coarse-grain bottleneck supervision ($d = 2$) for general mechanisms. Target phenotypes span general and specific categories.

pacity to capture label correlations and robustness to label imbalance (see Appendix C).

4.2 Phenotype Exclusivity Regularization

Despite effectively capturing label co-occurrence and mitigating class imbalance, the contrastive MLC loss fails to model semantic exclusivity in phenotype ontologies. Many abnormalities are inherently mutually exclusive—an organism cannot simultaneously exhibit both "hypotonia" and "hypertonia". These constraints encode critical biological knowledge, yet standard contrastive MLC losses lack mechanisms to impose such priors, often yielding biologically implausible predictions. The sparsity of phenotype annotations further impedes learning these constraints directly from data. To address this, we propose a complementary 2-step approach: (1) a large language model (LLM) pipeline for automated discovery of mutually exclusive phenotype pairs leveraging ontological structure, and (2) a soft exclusivity regularization term in the training objective that enforces these biological constraints.

Let $\mathcal{O} = (P, E)$ be a phenotype ontology where P represents the phenotype set and $E \subseteq P \times P$ denotes phenotype relationships. For phenotype $p_i \in P$, we define $D(p_i)$ as its set of direct descendants. Using an LLM with a structured query prompt (more details in Table 4), we identify $\mathcal{E}_i = \{(q_k, q_j) \in D(p_i) \times D(p_i) \mid p_k \text{ and } p_j \text{ are mutually exclusive}\}$, the set of mutually exclusive phenotype pairs among direct descendants of p_i. We then compute this set for all non-leaf phenotypes in the ontology graph and obtain the dataset-level exclusivity set $\mathcal{E} = \mathcal{E}_1 \bigcup \cdots \bigcup \mathcal{E}_n$.

For any pair of exclusive phenotype in $\mathcal{E}$, we impose a soft regularization on the exclusive pairs' prediction logits with Softplus. The exclusive regularization $\mathcal{L}_{\text{ex}}$ can be written as:

$$\mathcal{L}_{\text{ex}} = \frac{1}{N} \sum_{n=1}^{N} \sum_{(i,j) \in \mathcal{E}} \log\left(1 + e^{s_i(\mathbf{x}_n) + s_j(\mathbf{x}_n)}\right) \tag{4}$$

where $\mathcal{E}$ denotes a set of mutually-exclusive label pairs and $s_k \in \mathbb{R}$ is the logit for label k. We demonstrate the effectiveness of this loss using stationary analysis of gradient as follows.

Proposition 1 (Stationary Analysis of Exclusive Regularization)**.** *Let $\mathcal{L}$ be the contrastive MLC loss with our proposed soft exclusive regularization weighted by any $\lambda > 0$, namely $\mathcal{L} = \mathcal{L}_{\text{MLC}} + \lambda \mathcal{L}_{\text{ex}}$. For any pair $(i, j) \in \mathcal{E}$ and any input x, every first-order stationary point of $\mathcal{L}$ satisfies*

$$s_i(x) \leq 0 \quad \textit{or} \quad s_j(x) \leq 0 \tag{5}$$

The proof is deferred to Appendix D. Intuitively, Proposition 1 establishes that the regularization prevents any stationary solution where both logits in an exclusive pair are simultaneously positive, thereby enforcing the biological exclusive constraint. Furthermore, to demonstrate the necessity and advantages of our exclusive regularization, we show that the regularizer yields both a tighter generalization error bound and an explicit guarantee on conflict probability.

Theorem 1 (Generalization and Exclusivity Guarantee)**.** *Assume bounded inputs and linear logits as in Assumptions 1 and 2 (see Appendix A). Fix $\lambda > 0$ and let $\hat{R}_S(W)$ be the empirical risk on a sample $S = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^{N}$, drawn i.i.d. from $\mathcal{D}$, and let $W_\lambda^* := \arg\min_W \hat{R}_S(W)$. Define $C_\lambda := \hat{R}_S(W_\lambda^*)/\lambda$. Then, with probability at least $1-\delta$ over the draw of the training sample $\mathcal{S}$, we have the following* ***(a) generalization gap*** *and* ***(b) conflict probability****:*

$$\textbf{\textit{(a):}}\ R(W) \leq \widehat{R}_{\mathcal{S}}(W) + \frac{2C_\lambda}{\sqrt{N}} + 3\sqrt{\frac{\log(2/\delta)}{2N}} \tag{6}$$

$$\textbf{\textit{(b):}}\ \Pr_{(\mathbf{x},\mathbf{y}) \sim \mathcal{D}}\big(s_i(\mathbf{x}) > 0, s_j(\mathbf{x}) > 0\big) \leq \frac{R(W)}{\lambda \log 2}, \forall (i,j) \in \mathcal{E} \tag{7}$$

The detailed proof can be found in Appendix E. Part (a) of Theorem 1 indicates the exclusive regularization $\mathcal{L}_{\text{exc}}$ functions as an implicit norm regularizer, ensuring a generalization gap that is at least as tight as that of the unregularized loss $\mathcal{L}_{\text{MLC}}$. Part (b) further shows that the probability of predicting conflicting labels decays at least as fast as the generalization gap—a form of control not offered by plain loss $\mathcal{L}_{\text{MLC}}$. Together, these results provide a rigorous guarantee that our proposed objective can be broadly applied to ontology classification tasks with mutually exclusive label pairs.

Phenotype Frequency		11–30		31–100		101–300		≥301		All	
Dataset	**Method**	$F_{max}\uparrow$	AUC ↑	$F_{max}\uparrow$	AUC ↑	$F_{max}\uparrow$	AUC ↑	$F_{max}\uparrow$	AUC ↑	$F_{max}\uparrow$	AUC ↑
MPO	max-BLAST	8.14	50.61	10.96	50.52	13.92	50.55	21.78	50.79	16.65	50.59
	w-BLAST	11.78	52.74	13.59	53.16	17.23	53.71	28.02	53.46	22.73	53.10
	kmer2Vec+LR	9.82	55.63	11.46	55.36	16.78	55.57	33.74	54.91	28.75	55.45
	∗DeepGoPlus	1.42	52.58	8.56	52.52	11.79	52.74	33.26	52.53	28.51	52.58
	∗SPROF-GO	5.43	52.73	6.59	52.13	11.22	51.71	20.42	50.92	12.33	52.06
	∗InterLabelGO	11.90	59.27	13.19	57.36	17.03	56.03	34.15	54.29	29.66	57.61
	DeepPheno	8.45	58.09	13.98	64.05	18.03	61.94	33.84	60.59	28.80	59.89
	GenePheno	**13.86**	**68.02**	**16.12**	**69.73**	**19.49**	**68.09**	**36.06**	**61.43**	**31.14**	**67.86**
HPO	max-BLAST	14.71	51.25	17.72	51.83	21.51	52.27	35.74	52.43	25.77	51.74
	w-BLAST	16.19	53.65	19.76	54.40	24.12	54.81	40.67	54.82	31.03	54.22
	kmer2Vec+LR	15.08	54.02	20.83	54.97	26.32	54.62	48.58	54.52	40.11	54.47
	∗DeepGoPlus	5.86	51.33	16.94	51.02	25.25	51.03	48.71	51.88	40.18	51.24
	∗SPROF-GO	11.96	51.98	15.75	51.67	23.27	50.65	35.42	52.11	23.71	51.63
	∗InterLabelGO	16.04	57.52	19.13	58.40	26.09	56.96	49.17	58.39	40.45	57.84
	DeepPheno	17.48	55.68	18.73	58.32	27.08	60.59	49.02	63.03	40.37	58.17
	GenePheno	**22.02**	**72.79**	**26.34**	**73.09**	**31.65**	**69.03**	**51.95**	**65.82**	**43.17**	**71.44**
GWAS	max-BLAST	1.31	49.84	6.93	50.32	13.69	49.98	28.49	48.59	23.15	49.95
	w-BLAST	5.42	50.35	14.95	52.47	23.00	51.58	39.36	53.10	31.08	51.54
	kmer2Vec+LR	9.07	**57.64**	14.35	50.16	25.64	51.35	46.75	54.36	35.68	53.53
	∗DeepGoPlus	7.16	52.95	13.12	54.97	26.37	52.16	47.83	51.80	37.24	53.45
	∗SPROF-GO	6.70	51.01	9.40	52.24	26.69	53.36	38.35	53.01	31.08	52.03
	∗InterLabelGO	9.38	50.90	16.05	53.91	25.62	53.63	44.84	49.80	37.83	52.37
	DeepPheno	9.07	50.14	14.12	53.47	26.97	52.96	46.92	53.74	36.87	52.12
	GenePheno	**9.50**	56.64	**17.03**	**55.44**	**27.22**	**56.47**	**48.57**	**57.75**	**40.54**	**56.34**
CAFA2 wPPI	max-BLAST	14.35	50.16	16.97	49.93	20.92	49.65	31.39	49.77	20.21	49.94
	w-BLAST	16.50	51.37	21.10	50.98	26.73	51.30	39.23	51.79	26.37	51.24
	kmer2Vec+LR	17.28	54.61	24.29	52.98	30.14	51.58	45.41	52.02	36.07	53.13
	∗DeepGoPlus	12.95	51.84	16.58	54.59	28.01	53.49	46.69	52.59	37.18	53.32
	∗SPROF-GO	15.33	50.04	20.69	49.78	22.43	47.17	41.08	47.11	21.30	49.24
	∗InterLabelGO	16.07	54.37	19.55	51.75	26.69	52.73	45.02	55.46	35.47	53.31
	DeepPheno	15.20	54.62	20.47	54.63	30.93	57.44	47.37	56.37	37.55	55.35
	GraphPheno	18.01	56.35	24.68	59.04	31.88	59.47	46.60	57.13	36.02	58.11
	HPOFiller	14.37	49.57	22.58	48.44	28.84	49.14	43.18	49.10	23.32	48.87
	HPODNets	18.35	49.88	22.46	49.28	26.07	49.16	43.17	48.95	31.21	49.39
	SSLpheno	21.49	55.39	26.57	57.17	30.86	59.21	42.05	49.42	31.66	56.19
	GenePheno	**21.89**	**63.50**	**27.48**	**62.93**	**33.97**	**61.27**	**48.13**	**60.25**	**37.94**	**62.53**

Table 1: Phenotype prediction results on four datasets: MPO, HPO, GWAS, and CAFA2 (wPPI). Baselines marked with $*$ are adapted from protein sequence–to–GO function prediction models to gene sequence–to–phenotype prediction. Results are stratified by phenotype label frequency into four intervals: 11–30, 31–100, 101–300, and $\geq$301. "All" indicates performance on the full dataset. Percentage scores are reported for $F_{\max}$ and AUC. Bolded values denote the highest score in each column; underlined values indicate the second-best.

4.3 Mechanism-aware Learning with GO

Our regularized contrastive loss establishes the relationship between genetic information encoded in gene sequences and observable phenotypes. To further capture the biological mechanisms underlying phenotype formation, we incorporate GO functions as an additional modality through a dual-granularity approach.

Formally, for each gene g_i, we extract two complementary GO representations. First, we obtain fine-grained GO annotations with clear experimental evidence from UniEntrezDB (Miao et al. 2024) and process them through GoBERT to generate embeddings $\mathbf{h}_i$ that capture implicit functional correlations. These embeddings undergo cross-attention with gene sequence information to facilitate modality fusion. Second, we derive coarse-grained GO functions by propagating the fine-grained annotations along "is_a" and "part_of" relationships (Gaudet et al. 2017) in the GO DAG, resulting in a binary vector $\mathbf{g}_i \in \{0,1\}^n$ representing general biological mechanisms.

We designate the second-last layer of our network as a GO bottleneck, where outputs from n specific nodes serve as predictions $\widehat{\mathbf{g}}_i$ for the coarse-grained GO anno-

tations. This bottleneck is optimized using another contrastive loss $\mathcal{L}_{\mathrm{MLC}}^{\mathrm{GO}}(\widehat{\mathbf{g}}, \mathbf{g})$. During inference, the learned connection weights w_{ij} between GO node $\mathbf{g}_i$ and phenotype $\mathbf{y}_j$ quantify function-phenotype associations. The dual role of this bottleneck layer is significant: it not only introduces biologically-informed guidance throughout the end-to-end learning process, but also provides interpretable insights into the general biological mechanisms underlying phenotype formation during inference.

4.4 Overall Learning Framework

The architecture of our proposed model is illustrated in Figure 2. Our framework takes gene sequence embeddings $\mathbf{e}_i$ and GO embeddings $\mathbf{h}_i$ as input, facilitating information exchange between these modalities through cross-attention mechanism. These representations are subsequently integrated into a multi-modal embedding $\mathbf{x}_i$ via attention-pooling and concatenation. The final phenotype predictions are formulated as $\widehat{\mathbf{y}}_i = f_\psi(\mathbf{x_i}) = f_\psi(f_\theta(\mathbf{e}_i, \mathbf{h}_i))$ where $f_\theta(\cdot,\cdot)$ implements the cross-attention mechanism and $f_\psi(\cdot)$ includes MLP and GO bottleneck layer. Additionally, we derive exclusive label pairs set $\mathcal{E}$ using LLM and designate specific activations from intermediate layers as GO prediction outputs $\widehat{\mathbf{g}}_i$ along with corresponding ground-truth $\mathbf{g}_i$. Integrating these components, we formulate our comprehensive minimization objective as:

$$\mathcal{L} = \mathcal{L}_{\mathrm{MLC}}(\widehat{\mathbf{y}}, \mathbf{y}) + \lambda_1 \mathcal{L}_{\mathrm{ex}}(\widehat{\mathbf{y}}, \mathbf{y}, \mathcal{E}) + \lambda_2 \mathcal{L}_{\mathrm{MLC}}^{\mathrm{GO}}(\widehat{\mathbf{g}}, \mathbf{g}) \quad (8)$$

where λ_1 and λ_2 are hyperparameters controlling the contribution of exclusivity and bottleneck loss terms, respectively.

5 Experiments

We perform a comprehensive evaluation of GenePheno on four curated gebe sequence–to–phenotype datasets: HPO, MPO, GWAS, and CAFA2 wPPI. GenePheno is implemented with GENERator (Wu et al. 2025) as the gene sequence encoder and GoBERT (Miao et al. 2025) as the GO function encoder.

We compare GenePheno against several baselines, including max and weighted BLAST-based methods (Camacho et al. 2009), kmer2vec (Ren, Yin, and S.-T. Yau 2022) combined with logistic regression, and three adapted sequence-based gene function prediction models: DeepGOPlus (Kulmanov and Hoehndorf 2020a), SPROF-GO (Yuan et al. 2023), and InterLabelGO (Liu, Zhang, and Freddolino 2024).

As many genes lack experimentally validated PPI and GO annotations, we construct the CAFA2 wPPI dataset by intersecting the widely used CAFA2 challenge set (Jiang et al. 2016) with established PPI networks (Szklarczyk et al. 2015; Franz et al. 2018; Hwang et al. 2019). All genes in the resulting dataset have valid PPI information and GO annotations, enabling fair comparison with methods that rely on curated genetic modalities. On this dataset, we further evaluate curated-modality-based phenotype prediction methods, including DeepPheno (Kulmanov and Hoehndorf 2020b), GraphPheno (Liu et al. 2022), HPOFiller (Liu, Mamitsuka, and Zhu 2021), HPODNets (Liu, Mamitsuka, and Zhu 2022), and SSLPheno (Bi et al. 2023).

We follow CAFA-style preprocessing for all datasets and adopt evaluation protocols from prior work (Cho, Berger, and Peng 2016; Liu, Mamitsuka, and Zhu 2022; Bi et al. 2023), stratifying results by phenotype label frequency. Performance is reported using two widely accepted metrics: gene-centric $F_{\max}$ and phenotype-centric AUC (Jiang et al. 2016). More details on dataset construction, baseline adaptation, experimental setup and reproducibility details are provided in the Appendices G to J and L.

6 Results

We report results across phenotype frequency ranges using gene-centric $F_{\max}$ and phenotype-centric AUC to compare GenePheno against baseline methods. We perform an ablation study to quantify each module's contribution and present case studies showing that the GO bottleneck layer captures generalizable mechanisms of phenotype formation.

6.1 Benchmarking Results

Table 1 presents comprehensive experimental results across four phenotype-prediction datasets. GenePheno consistently outperforms all baseline methods, achieving substantial improvements in both $F_{\max}$ and AUC metrics. On the larger-scale datasets, GenePheno demonstrates particularly strong performance: for MPO, it surpasses the second-best method by +1.48 in $F_{\max}$ and +7.97 in AUC, while on HPO, the margins increase to +2.72 in $F_{\max}$ and +13.27 in AUC. These significant improvements can be attributed to the contrastive learning objective, which effectively captures inter-label relationships in high-dimensional phenotype spaces.

GenePheno maintains robust performance even on smaller datasets with limited genes and phenotypes. On GWAS, it improves performance from 37.83/53.53 ($F_{\max}$/AUC) achieved by the best baseline to 40.54/56.34, while consistently maintaining a 0.8–2.4-point advantage in high-frequency ($\geq$301) subsets. Similarly, on CAFA2, GenePheno demonstrates strong generalization capabilities under sparse supervision, particularly proved by its performance on low-frequency bins. This consistent performance across diverse data regimes validates the effectiveness of integrating DNA and GO encoders with cross-phenotype attention mechanisms for scalable phenotype prediction. The label-correlation aware contrastive loss further enhances the model's ability to generalize from limited supervision. Overall, GenePheno achieves state-of-the-art performance across all evaluation settings, establishing itself as a unified and effective approach for phenotype prediction tasks.

6.2 Ablation Studies

We conduct ablation studies to assess the impact of key components in GenePheno (Table 2). Removing the contrastive loss causes the largest drop in performance, highlighting its role in modeling phenotype correlations. GO inputs are also critical, as their removal notably reduces AUC. The mutual exclusivity and bottleneck losses consistently improve results, supporting model interpretability. Sequence input further adds complementary value, as its exclusion leads to a performance decline.

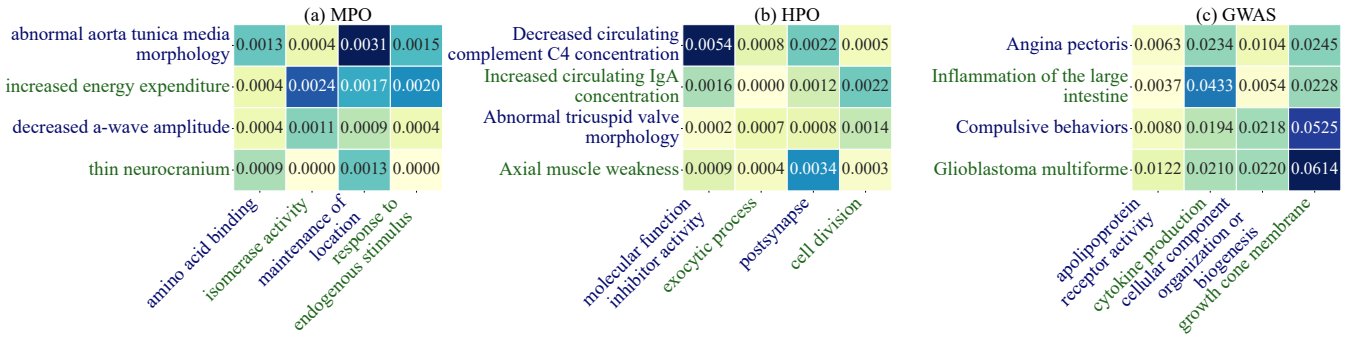

Figure 3: Sample bottleneck weight heatmap of MPO, HPO, and GWAS datasets. Darker colors indicate that the phenotype and GO function have a higher correlation.

	Interpretable	HPO		MPO		GWAS		CAFA2 (wPPI)	
Method	**Mechanism**	$F_{max}\uparrow$	AUC ↑	$F_{max}\uparrow$	AUC ↑	$F_{max}\uparrow$	AUC ↑	$F_{max}\uparrow$	AUC ↑
w/o GO Input	✓	39.96	56.60	28.21	54.19	36.82	52.59	35.65	53.83
w/o Sequence Input	✓	40.99	60.23	29.13	64.72	36.23	51.10	36.39	59.33
w/o Contrastive Loss	✓	41.83	65.75	29.99	64.64	36.05	52.25	37.06	50.86
w/o Mutual Exclusive Loss	✓	43.07	65.22	29.92	62.92	39.04	56.01	37.13	61.04
w/o Bottleneck Loss	✗	42.94	68.17	30.13	64.67	39.02	54.36	36.59	61.39
Complete Model	✓	**43.17**	**71.44**	**31.14**	**67.86**	**40.54**	**56.34**	**37.94**	**62.53**

Table 2: Ablation study results. Results are reported under the "All" setting for four datasets.

6.3 Case Studies

We analyze GenePheno's bottleneck weights to interpret phenotype formation. Representative links are shown in Figure 3, with detailed analysis in Appendix K.

MPO GenePheno identifies biologically plausible associations in this mouse gene knockout dataset. Two representative cases are shown in Figure 3(a). First, "maintenance of location" (GO:0051235) is linked to "abnormal aorta tunica media morphology" (MP:0009873), a feature of aortic aneurysms caused by disorganized cell and matrix structure (Tang et al. 2005). Second, "isomerase activity" (GO:0016853) is associated with "increased energy expenditure" (MP:0004889), aligning with findings that prolyl isomerases like Pin1 modulate skeletal muscle metabolism via SERCA activity (Nakatsu et al. 2024).

HPO The HPO dataset curates gene–phenotype associations from medical literature (Köhler et al. 2021). Figure 3(b) shows sample bottleneck weights from this dataset. GenePheno reveals a strong association between "molecular function inhibitor" (GO:0140678) and "decreased circulating complement C4 concentration" (HP:0045042), consistent with studies showing that complement inhibitors reduce C4 levels (Coss et al. 2023; Garred, Tenner, and Mollnes 2021). Another plausible link is found between "postsynapse" (GO:0098794) and "axial muscle weakness" (HP:0003327), as impaired signal transmission at the neuromuscular junction can lead to muscle weakness (Jones et al. 2017; Hirsch 2007).

GWAS The GWAS dataset links genes to phenotypes through SNP associations (Buniello et al. 2019; MacArthur et al. 2017), and focuses on a limited set of disease-related traits compared to the broader coverage of HPO and MPO. GenePheno interpretation weights are shown in Figure 3(c). One highlighted function, "growth cone membrane" (GO:0032580), is involved in axonal navigation (Vitriol and Zheng 2012; Igarashi 2019), and is strongly associated with "compulsive behaviors" (HP:0000722) and "glioblastoma multiforme" (HP:0012174). GenePheno also links "cytokine production" (GO:0001816) to "inflammation of the large intestine" (HP:0002037), consistent with the role of cytokines in coordinating immune responses (Zhang and An 2007; Sanchez-Muñoz, Dominguez-Lopez, and Yamamoto-Furusho 2008; Escalante et al. 2025).

7 Conclusion

We present GenePheno, the first end-to-end deep learning framework for directly mapping gene sequences to phenotypic traits. By integrating gene function information at both fine and coarse levels and modeling inter-phenotype dependencies and exclusivity, GenePheno bridges the modality gap between sequences and observable traits while providing interpretable, biologically grounded insights. To support this task, we curate four benchmark datasets and demonstrate strong, consistent performance through comprehensive experiments and case studies.

Acknowledgements

This work was partially supported by US National Science Foundation IIS-2412195, CCF-2400785, the Cancer Prevention and Research Institute of Texas (CPRIT) award (RP230363), the National Institutes of Health (NIH) R01 award (1R01AI190103-01) and Microsoft Accelerate Foundation Models Research (2024).

References

Avsec, Ž.; Agarwal, V.; Visentin, D.; Ledsam, J. R.; Grabska-Barwinska, A.; Taylor, K. R.; Assael, Y.; Jumper, J.; Kohli, P.; and Kelley, D. R. 2021. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10): 1196–1203.

Bi, X.; Liang, W.; Zhao, Q.; and Wang, J. 2023. Sslpheno: a self-supervised learning approach for gene–phenotype association prediction using protein–protein interactions and gene ontology data. *Bioinformatics*, 39(11): btad662.

Brixi, G.; Durrant, M. G.; Ku, J.; Poli, M.; Brockman, G.; Chang, D.; Gonzalez, G. A.; King, S. H.; Li, D. B.; Merchant, A. T.; et al. 2025. Genome modeling and design across all domains of life with Evo 2. *BioRxiv*, 2025–02.

Buchfink, B.; Reuter, K.; and Drost, H.-G. 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature methods*, 18(4): 366–368.

Buchfink, B.; Xie, C.; and Huson, D. H. 2015. Fast and sensitive protein alignment using DIAMOND. *Nature methods*, 12(1): 59–60.

Buniello, A.; MacArthur, J. A. L.; Cerezo, M.; Harris, L. W.; Hayhurst, J.; Malangone, C.; McMahon, A.; Morales, J.; Mountjoy, E.; Sollis, E.; et al. 2019. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, 47(D1): D1005–D1012.

Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; and Madden, T. L. 2009. BLAST+: architecture and applications. *BMC bioinformatics*, 10: 1–9.

Cheng, J.; Novati, G.; Pan, J.; Bycroft, C.; Žemgulytė, A.; Applebaum, T.; Pritzel, A.; Wong, L. H.; Zielinski, M.; Sargeant, T.; et al. 2023. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, 381(6664): eadg7492.

Cho, H.; Berger, B.; and Peng, J. 2016. Compact integration of multi-network topology for functional analysis of genes. *Cell systems*, 3(6): 540–548.

Consortium, G. O. 2019. The gene ontology resource: 20 years and still GOing strong. *Nucleic acids research*, 47(D1): D330–D338.

Coss, S. L.; Zhou, D.; Chua, G. T.; Aziz, R. A.; Hoffman, R. P.; Wu, Y. L.; Ardoin, S. P.; Atkinson, J. P.; and Yu, C.-Y. 2023. The complement system and human autoimmune diseases. *Journal of autoimmunity*, 137: 102979.

Crick, F. 1970. Central dogma of molecular biology. *Nature*, 227(5258): 561–563.

Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; et al. 2021. ProtTrans: towards cracking the language of life's code through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44: 7112–7127.

Escalante, J.; Artaiz, O.; Diwakarla, S.; and McQuade, R. M. 2025. Leaky gut in systemic inflammation: Exploring the link between gastrointestinal disorders and age-related diseases. *GeroScience*, 47(1): 1–22.

Feuermann, M.; Mi, H.; Gaudet, P.; Muruganujan, A.; Lewis, S. E.; Ebert, D.; Mushayahama, T.; dictyBase http://orcid. org/0000-0002-5638-3990 Chisholm Rex L. 10 http://orcid. org/0000 0002-4532-2703 Fey Petra 10; Evidence; http://orcid. org/0000-0001-7628-5565 Giglio Michelle 11 http://orcid. org/0000 0003-3643-281X Nadendla Suvarna 11, C. O.; et al. 2025. A compendium of human gene functions derived from evolutionary modelling. *Nature*, 1–9.

Franz, M.; Rodriguez, H.; Lopes, C.; Zuberi, K.; Montojo, J.; Bader, G. D.; and Morris, Q. 2018. GeneMANIA update 2018. *Nucleic acids research*, 46(W1): W60–W64.

Frazer, J.; Notin, P.; Dias, M.; Gomez, A.; Min, J. K.; Brock, K.; Gal, Y.; and Marks, D. S. 2021. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883): 91–95.

Gargano, M. A.; Matentzoglu, N.; Coleman, B.; Addo-Lartey, E. B.; Anagnostopoulos, A. V.; Anderton, J.; Avillach, P.; Bagley, A. M.; Bakštein, E.; Balhoff, J. P.; et al. 2024. The Human Phenotype Ontology in 2024: phenotypes around the world. *Nucleic acids research*, 52(D1): D1333–D1346.

Garred, P.; Tenner, A. J.; and Mollnes, T. E. 2021. Therapeutic targeting of the complement system: from rare diseases to pandemics. *Pharmacological reviews*, 73(2): 792–827.

Gaudet, P.; Škunca, N.; Hu, J. C.; and Dessimoz, C. 2017. Primer on the gene ontology. *The Gene Ontology Handbook*, 25–37.

Gligorijevic, V.; Renfrew, P. D.; Kosciolek, T.; Leman, J. K.; Cho, K.; Vatanen, T.; Berenberg, D.; Taylor, B.; Fisk, I. M.; Xavier, R. J.; Knight, R.; and Bonneau, R. 2019. Structure-Based Function Prediction using Graph Convolutional Networks. *bioRxiv*.

Hirsch, N. 2007. Neuromuscular junction in health and disease. *British journal of anaesthesia*, 99(1): 132–138.

Hwang, S.; Kim, C. Y.; Yang, S.; Kim, E.; Hart, T.; Marcotte, E. M.; and Lee, I. 2019. HumanNet v2: human gene networks for disease research. *Nucleic acids research*, 47(D1): D573–D580.

Igarashi, M. 2019. Molecular basis of the functions of the mammalian neuronal growth cone revealed using new methods. *Proceedings of the Japan Academy, Series B*, 95(7): 358–377.

Islam, U. I.; Campelo dos Santos, A. L.; Kanjilal, R.; and Assis, R. 2025. Learning genotype–phenotype associations from gaps in multi-species sequence alignments. *Briefings in Bioinformatics*, 26(1): bbaf022.

Jiang, Y.; Oron, T. R.; Clark, W. T.; Bankapur, A. R.; D'Andrea, D.; Lepore, R.; Funk, C. S.; Kahanda, I.; Verspoor, K. M.; Ben-Hur, A.; et al. 2016. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome biology*, 17: 1–19.

Jones, R. A.; Harrison, C.; Eaton, S. L.; Hurtado, M. L.; Graham, L. C.; Alkhammash, L.; Oladiran, O. A.; Gale, A.; Lamont, D. J.; Simpson, H.; et al. 2017. Cellular and molecular anatomy of the human neuromuscular junction. *Cell reports*, 21(9): 2348–2356.

Kelley, D. R.; Snoek, J.; and Rinn, J. L. 2016. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome research*, 26(7): 990–999.

Köhler, S.; Gargano, M.; Matentzoglu, N.; Carmody, L. C.; Lewis-Smith, D.; Vasilevsky, N. A.; Danis, D.; Balagura, G.; Baynam, G.; Brower, A. M.; et al. 2021. The human phenotype ontology in 2021. *Nucleic acids research*, 49(D1): D1207–D1217.

Kulmanov, M.; and Hoehndorf, R. 2020a. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics*, 36(2): 422–429.

Kulmanov, M.; and Hoehndorf, R. 2020b. DeepPheno: predicting single gene loss-of-function phenotypes using an ontology-aware hierarchical classifier. *PLoS computational biology*, 16(11): e1008453.

Liu, L.; Mamitsuka, H.; and Zhu, S. 2021. HPOFiller: identifying missing protein–phenotype associations by graph convolutional network. *Bioinformatics*, 37(19): 3328–3336.

Liu, L.; Mamitsuka, H.; and Zhu, S. 2022. HPODNets: deep graph convolutional networks for predicting human protein–phenotype associations. *Bioinformatics*, 38(3): 799–808.

Liu, Q.; Zhang, C.; and Freddolino, L. 2024. InterLabelGO+: unraveling label correlations in protein function prediction. *Bioinformatics*, 40(11): btae655.

Liu, Y.; He, R.; Qu, Y.; Zhu, Y.; Li, D.; Ling, X.; Xia, S.; Li, Z.; and Li, D. 2022. Integration of human protein sequence and protein-protein interaction data by graph autoencoder to identify novel protein-abnormal phenotype associations. *Cells*, 11(16): 2485.

MacArthur, J.; Bowler, E.; Cerezo, M.; Gil, L.; Hall, P.; Hastings, E.; Junkins, H.; McMahon, A.; Milano, A.; Morales, J.; et al. 2017. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic acids research*, 45(D1): D896–D901.

Mackay, T. F.; and Anholt, R. R. 2024. Pleiotropy, epistasis and the genetic architecture of quantitative traits. *Nature Reviews Genetics*, 25(9): 639–657.

Miao, Y.; Guo, Y.; Ma, H.; Yan, J.; Jiang, F.; An, W.; Gao, J.; and Huang, J. 2024. UniEntrezDB: Large-scale Gene Ontology Annotation Dataset and Evaluation Benchmarks with Unified Entrez Gene Identifiers. *arXiv preprint arXiv:2412.12688*.

Miao, Y.; Guo, Y.; Ma, H.; Yan, J.; Jiang, F.; Liao, R.; and Huang, J. 2025. GoBERT: Gene Ontology Graph Informed BERT for Universal Gene Function Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 622–630.

Nakatsu, Y.; Matsunaga, Y.; Nakanishi, M.; Yamamotoya, T.; Sano, T.; Kanematsu, T.; and Asano, T. 2024. Prolyl isomerase Pin1 in skeletal muscles contributes to systemic energy metabolism and exercise capacity through regulating SERCA activity. *Biochemical and Biophysical Research Communications*, 715: 150001.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Reissmann, M.; and Ludwig, A. 2013. Pleiotropic effects of coat colour-associated mutations in humans, mice and other mammals. In *Seminars in cell & developmental biology*, volume 24, 576–586. Elsevier.

Ren, R.; Yin, C.; and S.-T. Yau, S. 2022. kmer2vec: A novel method for comparing DNA sequences by word2vec embedding. *Journal of Computational Biology*, 29(9): 1001–1021.

Sanchez-Muñoz, F.; Dominguez-Lopez, A.; and Yamamoto-Furusho, J. K. 2008. Role of cytokines in inflammatory bowel disease. *World journal of gastroenterology: WJG*, 14(27): 4280.

Shalev-Shwartz, S.; and Ben-David, S. 2014. *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Smith, C. L.; and Eppig, J. T. 2009. The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 1(3): 390–399.

Su, J.; Zhu, M.; Murtadha, A.; Pan, S.; Wen, B.; and Liu, Y. 2022. Zlpr: A novel loss for multi-label classification. *arXiv preprint arXiv:2208.02955*.

Szklarczyk, D.; Franceschini, A.; Wyder, S.; Forslund, K.; Heller, D.; Huerta-Cepas, J.; Simonovic, M.; Roth, A.; Santos, A.; Tsafou, K. P.; et al. 2015. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, 43(D1): D447–D452.

Tang, P. C.; Coady, M. A.; Lovoulos, C.; Dardik, A.; Aslan, M.; Elefteriades, J. A.; and Tellides, G. 2005. Hyperplastic cellular remodeling of the media in ascending thoracic aortic aneurysms. *Circulation*, 112(8): 1098–1105.

Tang, Q.; and Khvorova, A. 2024. RNAi-based drug design: considerations and future directions. *Nature Reviews Drug Discovery*, 23(5): 341–364.

Vitriol, E. A.; and Zheng, J. Q. 2012. Growth cone travel in space and time: the cellular ensemble of cytoskeleton, adhesion, and membrane. *Neuron*, 73(6): 1068–1081.

Whiting, F. J.; Househam, J.; Baker, A.-M.; Sottoriva, A.; and Graham, T. A. 2024. Phenotypic noise and plasticity in cancer evolution. *Trends in Cell Biology*, 34(6): 451–464.

Wu, W.; Li, Q.; Li, M.; Fu, K.; Feng, F.; Ye, J.; Xiong, H.; and Wang, Z. 2025. GENERator: A Long-Context Generative Genomic Foundation Model. *arXiv preprint arXiv:2502.07272*.

Yuan, Q.; Xie, J.; Xie, J.; Zhao, H.; and Yang, Y. 2023. Fast and accurate protein function prediction from sequence through pretrained language model and homology-based label diffusion. *Briefings in bioinformatics*, 24(3): bbad117.

Zhang, J.-M.; and An, J. 2007. Cytokines, inflammation, and pain. *International anesthesiology clinics*, 45(2): 27–37.

Zheng, Z.; Liu, S.; Sidorenko, J.; Wang, Y.; Lin, T.; Yengo, L.; Turley, P.; Ani, A.; Wang, R.; Nolte, I. M.; et al. 2024. Leveraging functional genomic annotations and genome coverage to improve polygenic prediction of complex traits within and between ancestries. *Nature Genetics*, 56(5): 767–777.

Zhou, J.; Theesfeld, C. L.; Yao, K.; Chen, K. M.; Wong, A. K.; and Troyanskaya, O. G. 2018. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature genetics*, 50(8): 1171–1179.

Zhou, J.; and Troyanskaya, O. G. 2015. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, 12(10): 931–934.