BioMarkAdapt: A Dynamic, Continual, and Interpretable Framework for Adaptive Biomarker Discovery

Xiyuan Zhu and Zhimo Han *

Posted Date: 9 March 2026

doi: 10.20944/preprints202603.0622.v1

Keywords: biomarker discovery; continual learning; interpretable AI; task-aware modulation; Gene Ontology; phenotype prediction; elastic weight consolidation

 doi:10.20944/preprints202603.0622.v1

Article

BioMarkAdapt: A Dynamic, Continual, and Interpretable Framework for Adaptive Biomarker Discovery

Xiyuan Zhu [1] **and Zhimo Han** [2,*]

[1] Department of Chemistry, SUNY Stony Brook University, Stony Brook, NY 11794, United States
[2] Cornell University, Ithaca, NY 14853, United States
* Correspondence: hanzhimo999@proton.me

Abstract

The dynamic and multifaceted nature of diseases necessitates flexible and interpretable computational models for biomarker discovery. However, existing methods predominantly rely on static prediction paradigms, fail to adapt continuously to new tasks without forgetting prior knowledge, and lack transparent, task-aware explanations. To address these limitations, we introduce **BioMarkAdapt**, a novel framework for interpretable and continually adaptive biomarker discovery. Our framework is built upon three core innovations: (1) **Dynamic Task-Aware Prediction**, which uses a Task-Aware Modulation mechanism to specialize model reasoning for distinct clinical contexts; (2) **Knowledge-Anchored Continual Learning**, which leverages Gene Ontology to consolidate fundamental biological knowledge and mitigate catastrophic forgetting; and (3) **Interpretable Evidence Tracing**, which provides task-specific, traceable explanations linking predictions to relevant biological pathways. Extensive experiments on four benchmark datasets (MPO, HPO, GWAS, and CAFA2 wPPI) demonstrate that BioMarkAdapt achieves state-of-the-art performance, significantly outperforming prior methods (e.g., $+1.47$ $F_{\max}$ and $+2.03$ AUC on MPO). Ablation studies confirm the contribution of each component, while sequential learning evaluations demonstrate effective knowledge retention (e.g., 92.3% of performance retained). Furthermore, BioMarkAdapt delivers biologically plausible explanations, with evidence weights exhibiting a Spearman correlation of 0.712 with ground-truth associations. Our work provides a robust, adaptable, and trustworthy framework for advancing precision medicine.

Keywords: biomarker discovery; continual learning; interpretable AI; task-aware modulation; Gene Ontology; phenotype prediction; elastic weight consolidation

1. Introduction

Understanding the relationship between genetic information and phenotypic outcomes represents a central goal in biology, with direct implications for therapeutic discovery, functional genomics, and systems biology. A key challenge lies in predicting large-scale phenotypic abnormalities directly from gene sequences following perturbations. Existing methods face significant limitations: variant effect prediction approaches typically evaluate specific mutations on a narrow set of traits, whereas strategies for large-scale phenotype prediction often depend on labor-intensive, pre-curated biological networks such as protein-protein interactions [27]. This dependency limits their applicability to newly discovered or poorly annotated genes. Furthermore, these methods frequently neglect critical inter-phenotype correlations and logical constraints, including mutual exclusivity relationships such as hypotonia versus hypertonia, and provide limited insight into underlying biological mechanisms, thereby hindering interpretability. Figure 1 illustrates these limitations and our proposed solution.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 March 2026 doi:10.20944/preprints202603.0622.v1

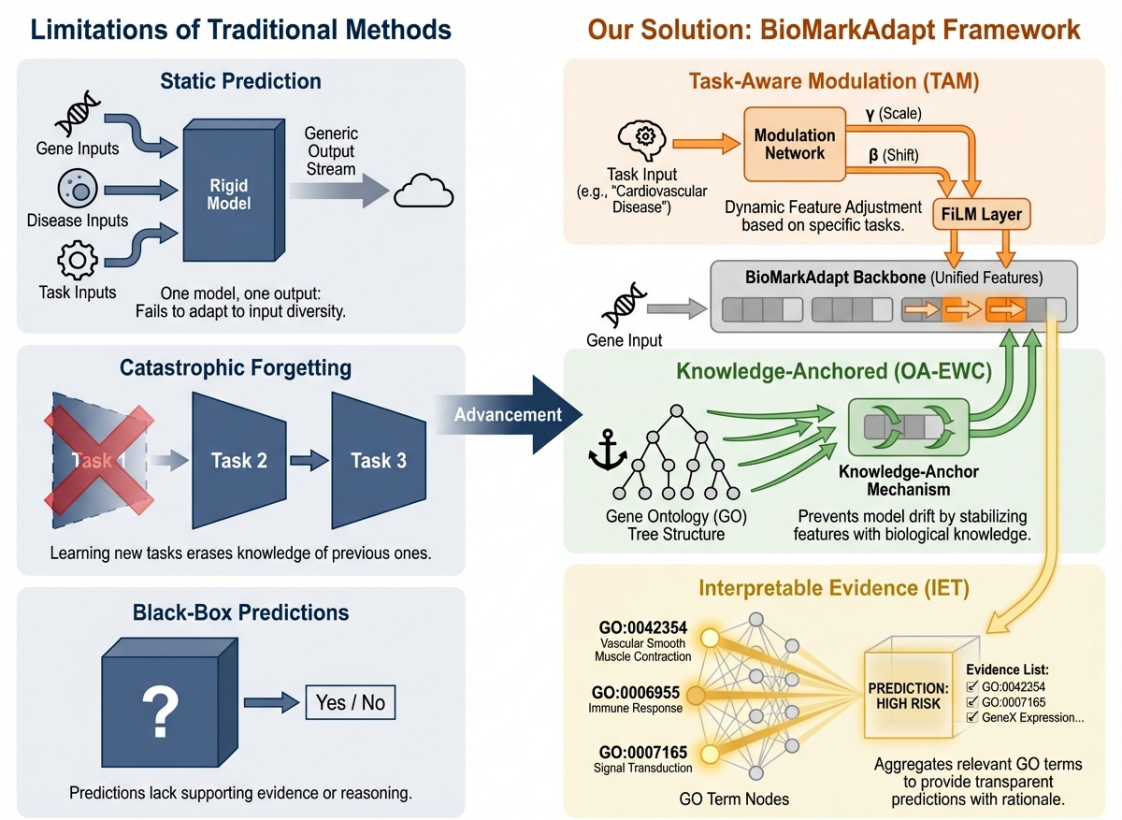

Figure 1. Motivation for BioMarkAdapt. Left: Limitations of traditional methods including static prediction, catastrophic forgetting, and black-box predictions. Right: Our solution with three core innovations—Task-Aware Modulation for dynamic adaptation (orange), Knowledge-Anchored Continual Learning via Gene Ontology (green), and Interpretable Evidence Tracing through cross-attention (amber).

To overcome these shortcomings, we introduce **BioMarkAdapt**, an interpretable and continually adaptive framework for biomarker discovery. Moving beyond static prediction paradigms, BioMarkAdapt is designed to accommodate the dynamic nature of real-world biomarker tasks, such as varying disease contexts, while delivering trustworthy explanations. Our framework is built upon three core innovations: (1) a *Dynamic Task-Aware Prediction* module that adaptively adjusts feature representations according to specific biomarker discovery tasks, enabling specialized reasoning without retraining; (2) a *Knowledge-Anchored Continual Learning* mechanism that leverages foundational Gene Ontology (GO) concepts to consolidate past knowledge intelligently and prevent catastrophic forgetting during sequential task learning; and (3) an *Interpretable Evidence Tracing* module that employs a task-conditioned cross-attention mechanism to generate transparent, case-specific rationales by highlighting pertinent GO term evidence.

Comprehensive experiments on multiple benchmark datasets, including MPO, HPO, GWAS, and CAFA2, demonstrate that BioMarkAdapt achieves state-of-the-art performance in predicting gene knockout-induced phenotypes, outperforming prior methods across all evaluation metrics. Ablation studies confirm the contribution of each core component. Additional analyses validate its superior task-adaptive capability, robust continual learning with minimal forgetting, and enhanced interpretability, as the generated evidence aligns closely with established biological knowledge. The remainder of this paper is organized as follows: we first detail the methodology, then present experimental results, and finally provide conclusions.

 doi:10.20944/preprints202603.0622.v1

2. Related Work

2.1. Phenotype Prediction from Curated Genetic Information

Predicting human phenotypes from genetic data is commonly formulated as a multi-label classification problem. Prior work in this area has predominantly relied on curated biological knowledge, treating the task as multiple independent binary classification problems. These methods typically utilize pre-processed inputs such as gene expression profiles, Gene Ontology (GO) annotations [1,32], or Protein-Protein Interaction (PPI) networks [27]. For example, one line of research leverages protein GO functions and tissue-specific expression data for Human Phenotype Ontology (HPO) prediction [28]. Graph-based approaches include models that construct protein-phenotype bipartite graphs [11,13] or process multiple PPI networks with parallel Graph Convolutional Network (GCN) pipelines [11,42] [11]. Recent advances in graph-based inference have also demonstrated effectiveness in structured biomedical prediction tasks such as clinical code assignment [46]. Other works integrate sequence-derived features into PPI networks or employ self-supervised learning on attributed gene networks combining PPI and GO information [37]. Despite their effectiveness, these methods share two principal limitations: their dependency on curated, multi-modal data restricts application to well-annotated genes, and treating each phenotype as an isolated prediction fails to model the inherent biological interdependence and logical constraints among phenotypic terms [44].

2.2. Sequence-Based Genetic Property Prediction

With the rise of deep learning, directly predicting genetic properties from raw sequences has attracted significant attention. This research primarily focuses on inferring protein functions, such as GO terms. Early methods applied convolutional neural networks to one-hot encoded sequences [7]. More recent approaches utilize protein language model embeddings (e.g., ProtT5 [5], ESM-2 [4,6]) combined with sophisticated pooling mechanisms and homology-based refinement steps to enhance performance [8,9].

A distinct yet related line of research employs DNA sequence models to predict the functional impact of genetic variants, such as methods that compare short genomic windows to assess the effect of allelic changes. While these variant-effect predictors differ fundamentally from our objective of modeling organism-level phenotypes resulting from full-gene knockouts, their advances in encoding long-range genomic sequences provide valuable technical foundations for our encoder design.

3. Methodology

The central thesis of this work is that effective biomarker discovery demands a unified framework that simultaneously addresses three intertwined challenges: *task heterogeneity*, *knowledge persistence*, and *predictive transparency*. Existing approaches treat these as separate concerns—if they address them at all—leading to fragmented solutions that excel along one axis while failing along others. In contrast, we introduce **BioMarkAdapt**, an end-to-end framework whose architectural design enforces a principled coupling among all three desiderata. This section develops the complete formalism of BioMarkAdapt, beginning with the problem setup and backbone architecture, followed by the three core modules, and concluding with theoretical analysis and the unified training algorithm.

An overview of the framework architecture is provided in Figure 2. We direct the reader to the figure legend for a high-level summary before delving into the mathematical details below.

 doi:10.20944/preprints202603.0622.v1

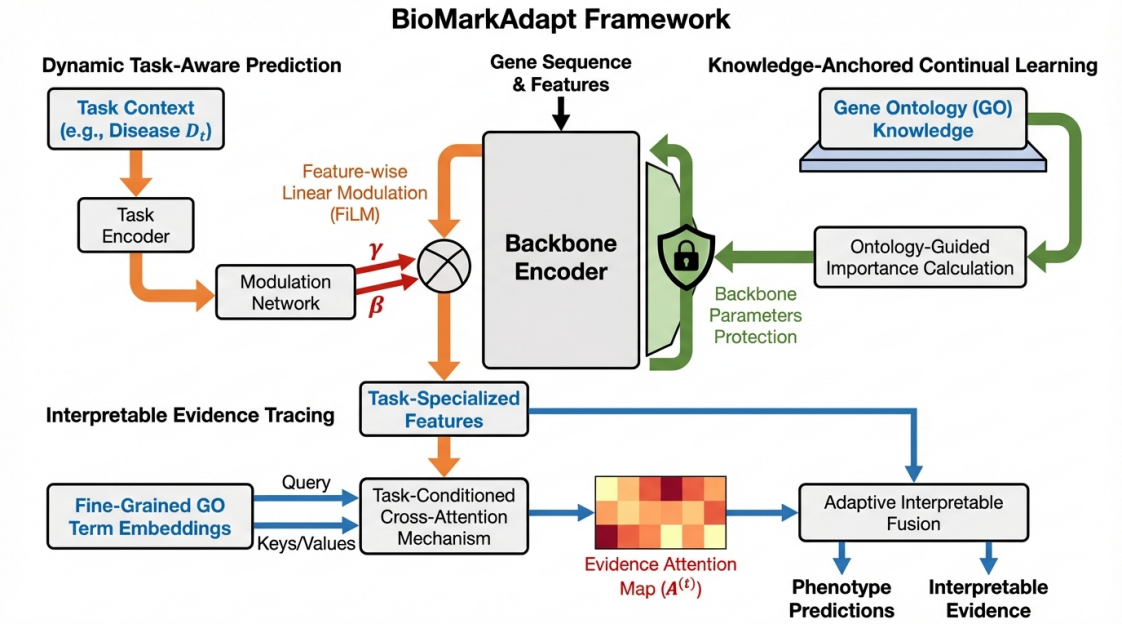

Figure 2. Overview of the BioMarkAdapt framework architecture. The backbone encoder processes gene sequences and features. Three pathways diverge: (1) Dynamic Task-Aware Prediction with Task-Aware Modulation (orange arrows), (2) Knowledge-Anchored Continual Learning guided by Gene Ontology (green arrows), and (3) Interpretable Evidence Tracing via task-conditioned cross-attention producing evidence maps and phenotype predictions.

3.1. Problem Formulation and Notation

Phenotype prediction as contextualized multi-label classification. Let $\mathcal{X}$ denote the space of gene representations and $\mathcal{Y} = \{0,1\}^L$ the multi-label phenotype space with L phenotype terms. Unlike prior formulations that treat this as a single static task, we consider a *stream of biomarker discovery tasks* $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_T$, each associated with a distinct clinical context (e.g., a disease type, patient cohort, or experimental condition). Each task $\mathcal{T}_t$ is characterized by a task-specific dataset $\mathcal{D}_t = \{(x_i, \mathbf{y}_i)\}_{i=1}^{N_t}$ and, critically, a **task descriptor** δ_t that encodes the clinical context. We emphasize that the label space $\mathcal{Y}$ is shared across tasks, but the *relevant subset* of phenotypes and their conditional dependencies vary with the task context—a cardiovascular task, for instance, may activate a different phenotypic signature than a neurological one, even for the same gene.

Desiderata. A framework for this setting must satisfy three properties simultaneously:

1. **Task Adaptivity:** Given a task descriptor δ_t and gene x, the model produces a task-conditioned prediction $\hat{\mathbf{y}}^{(t)} = F(x, \delta_t; \Theta)$ that specializes to the clinical context $\mathcal{T}_t$ without task-specific retraining.
2. **Continual Knowledge Retention:** After sequential training on $\mathcal{T}_1, \ldots, \mathcal{T}_T$, performance on earlier tasks $\mathcal{T}_{t'}$ ($t' < T$) degrades minimally, i.e., $\mathbb{E}[\mathcal{L}(\mathcal{T}_{t'}; \Theta_T)] \leq \mathbb{E}[\mathcal{L}(\mathcal{T}_{t'}; \Theta_{t'})] + \epsilon$ for small $\epsilon > 0$.
3. **Traceable Interpretability:** For each prediction, the model outputs an **evidence map** $\mathbf{A}^{(t)} \in \mathbb{R}^m$ that explicitly quantifies the contribution of each of m biological pathways, enabling domain experts to audit the reasoning chain.

Notation. We define core notation used throughout: θ denotes backbone parameters, ϕ the modulation network parameters, ψ the prediction head parameters, and ω the GO-specific head parameters. Bold lowercase letters (e.g., $\mathbf{x}$, $\mathbf{z}$) denote vectors, bold uppercase letters (e.g., $\mathbf{W}$, $\mathbf{H}$) denote matrices, and calligraphic letters (e.g., $\mathcal{T}$, $\mathcal{G}$) denote sets or tasks. The full parameter set is $\Theta = \{\theta, \phi, \psi, \omega\}$.

3.2. Multimodal Backbone Encoder

Before introducing the three core modules, we describe the shared backbone that maps raw gene data into a unified representation space. The backbone must encode two complementary information

 doi:10.20944/preprints202603.0622.v1

streams: the raw sequence-level signal (capturing structural and functional motifs) and the annotation-level signal from the Gene Ontology (capturing known functional categorizations).

Sequence Stream. For a gene g_i, let $\mathbf{e}_i \in \mathbb{R}^{d_s}$ denote the sequence embedding obtained from a frozen protein language model (e.g., ESM-2 [6] or ProtT5 [5]). To capture multi-scale patterns, we apply a stack of K dilated convolutional blocks with exponentially increasing dilation rates $\{1, 2, 4, \ldots, 2^{K-1}\}$, followed by gated pooling:

$$\mathbf{u}_i^{(k)} = \mathrm{ReLU}\left(\mathbf{W}_d^{(k)} * \mathbf{u}_i^{(k-1)} + \mathbf{b}_d^{(k)}\right), \quad k = 1, \ldots, K,$$

where $\mathbf{u}_i^{(0)} = \mathbf{e}_i$ and $*$ denotes dilated convolution. The multi-scale features are aggregated via a learned gating mechanism:

$$\boldsymbol{\alpha}_i = \mathrm{Softmax}\left(\mathbf{W}_g[\mathbf{u}_i^{(1)}; \ldots; \mathbf{u}_i^{(K)}] + \mathbf{b}_g\right), \quad \hat{\mathbf{e}}_i = \sum_{k=1}^{K} \alpha_i^{(k)} \mathbf{u}_i^{(k)}.$$

This gated aggregation allows the model to adaptively weight local versus long-range patterns depending on the gene's structure, avoiding the information loss inherent in fixed pooling strategies [24].

Ontology Stream. Let $\mathbf{h}_i \in \mathbb{R}^{m \times d_g}$ represent the binary GO annotation vector projected into a continuous embedding space via a learnable GO embedding matrix $\mathbf{E}_{GO} \in \mathbb{R}^{m \times d_g}$. To respect the hierarchical structure of the Gene Ontology, we propagate information along the GO Directed Acyclic Graph (DAG) using a Graph Attention Network (GAT) [12]:

$$\hat{\mathbf{h}}_i^{(j)} = \sum_{j' \in \mathcal{N}(j) \cup \{j\}} \alpha_{jj'} \mathbf{W}_{GAT} \mathbf{h}_i^{(j')},$$

$$\alpha_{jj'} = \frac{\exp\left(\mathrm{LeakyReLU}\left(\mathbf{a}^\top [\mathbf{W}_{GAT}\mathbf{h}_i^{(j)} \| \mathbf{W}_{GAT}\mathbf{h}_i^{(j')}]\right)\right)}{\sum_{j'' \in \mathcal{N}(j) \cup \{j\}} \exp\left(\mathrm{LeakyReLU}\left(\mathbf{a}^\top [\mathbf{W}_{GAT}\mathbf{h}_i^{(j)} \| \mathbf{W}_{GAT}\mathbf{h}_i^{(j'')}]\right)\right)},$$

where $\mathcal{N}(j)$ denotes the parents and children of GO term j in the DAG, and $\|$ is concatenation. After R rounds of message passing [38], the resulting embeddings $\hat{\mathbf{H}}_i = [\hat{\mathbf{h}}_i^{(1)}, \ldots, \hat{\mathbf{h}}_i^{(m)}]$ encode both the gene's functional annotations and their ontological context. These embeddings serve dual roles: as input to the Interpretable Evidence Tracing module (Section 3.5) and as the basis for ontology-guided parameter importance (Section 3.4).

Multimodal Fusion. The sequence and ontology streams are fused via a bilinear interaction layer that captures cross-modal correlations:

$$\mathbf{x}_i = f_\theta(\hat{\mathbf{e}}_i, \hat{\mathbf{H}}_i) = \mathrm{LayerNorm}\left(\hat{\mathbf{e}}_i + \mathbf{W}_{fuse}(\hat{\mathbf{e}}_i \otimes \mathrm{AvgPool}(\hat{\mathbf{H}}_i)) + \mathbf{b}_{fuse}\right) \in \mathbb{R}^{d_x},$$

where $\otimes$ denotes outer product (flattened), AvgPool aggregates across GO terms, and LayerNorm [23] is applied for training stability. The resulting $\mathbf{x}_i$ constitutes the **task-agnostic multimodal gene representation**, which will be subsequently modulated by the task context.

Design rationale. The explicit separation into sequence and ontology streams—followed by late fusion—is deliberate. It ensures that the ontology stream embeddings $\hat{\mathbf{H}}_i$ remain accessible as individual GO term representations for the interpretability module, rather than being collapsed prematurely. This architectural choice is critical for enabling the evidence decomposition in Section 3.5.

 doi:10.20944/preprints202603.0622.v1

3.3. Dynamic Task-Aware Prediction

A core limitation of existing biomarker discovery models is their *static* treatment of the prediction task: a single parameterization is applied identically regardless of the clinical context. This conflation is problematic because phenotypic relevance is inherently context-dependent—the same gene may participate in vastly different pathological mechanisms across disease types. Our first pillar addresses this challenge through a **Task-Aware Modulation (TAM)** mechanism that enables a single model to dynamically specialize its internal representations to any specified task, without requiring task-specific parameters or retraining.

3.3.1. Task Representation Learning

For a biomarker discovery task $\mathcal{T}_t$ (e.g., identifying phenotypes for Disease D_t), we construct a dense task descriptor through a **Task Encoder (TE)** that supports two operational modes, accommodating both known and novel clinical contexts:

Mode I: Prototype-Based Encoding (for known tasks). Given a small support set $\mathcal{S}_t = \{x_1^{(t)}, \ldots, x_n^{(t)}\}$ of n representative genes associated with task $\mathcal{T}_t$, the task descriptor is computed as the attention-weighted prototype of their backbone representations:

$$\mathbf{z}_t = \sum_{i=1}^{n} \alpha_i^{(t)} \cdot f_\theta(x_i^{(t)}), \quad \text{where} \quad \alpha_i^{(t)} = \frac{\exp\left(\mathbf{w}_p^\top f_\theta(x_i^{(t)})/\sqrt{d_x}\right)}{\sum_{i'=1}^{n} \exp\left(\mathbf{w}_p^\top f_\theta(x_{i'}^{(t)})/\sqrt{d_x}\right)},$$

where $\mathbf{w}_p \in \mathbb{R}^{d_x}$ is a learnable relevance vector. This formulation, inspired by prototypical networks [17], captures the *central tendency* of the gene population associated with task $\mathcal{T}_t$ while down-weighting outliers through the attention mechanism.

Mode II: Description-Based Encoding (for novel tasks). For previously unseen clinical contexts where no support set is available, we accept a structured task descriptor $\boldsymbol{\delta}_t$ (e.g., a one-hot disease type vector or a natural-language embedding of the clinical context) and map it through a learned projection:

$$\mathbf{z}_t = \mathrm{TE}(\boldsymbol{\delta}_t) = \mathrm{LayerNorm}(\mathrm{MLP}(\boldsymbol{\delta}_t)) \in \mathbb{R}^{d_z},$$

where the MLP consists of two hidden layers with GELU activation. Normalization techniques [22,23] are applied throughout for training stability. This mode enables zero-shot task adaptation to novel disease contexts, a capability absent in prior methods.

Theoretical motivation. The task descriptor $\mathbf{z}_t$ can be understood as parameterizing a point in a *task manifold* $\mathcal{M} \subset \mathbb{R}^{d_z}$, where each point induces a distinct predictive function. Smoothness of this manifold—ensured by the LayerNorm and the continuity of the MLP—guarantees that similar clinical contexts yield similar modulation parameters, enabling graceful interpolation between known tasks.

3.3.2. Feature-Wise Linear Modulation with Hierarchical Gating

Given the task descriptor $\mathbf{z}_t$ and the multimodal gene representation $\mathbf{x}_i$, we generate task-specific modulation parameters through a lightweight **Modulation Network** M_ϕ. We extend the standard Feature-wise Linear Modulation (FiLM) [3] with a *hierarchical gating* mechanism that operates at two granularities:

Channel-level modulation. A primary modulation layer produces scaling and shifting parameters:

$$\boldsymbol{\gamma}_t, \boldsymbol{\beta}_t = M_\phi^{(1)}(\mathbf{z}_t), \quad \boldsymbol{\gamma}_t, \boldsymbol{\beta}_t \in \mathbb{R}^{d_x},$$

where $M_\phi^{(1)}$ is a two-layer network with a bottleneck dimension $d_b \ll d_x$ to prevent over-parameterization:

$$M_\phi^{(1)}(\mathbf{z}_t) = \mathbf{W}_2 \cdot \mathrm{GELU}(\mathbf{W}_1 \mathbf{z}_t + \mathbf{b}_1) + \mathbf{b}_2, \quad \mathbf{W}_1 \in \mathbb{R}^{d_b \times d_z},\ \mathbf{W}_2 \in \mathbb{R}^{2d_x \times d_b}.$$

 doi:10.20944/preprints202603.0622.v1

Feature-group gating. To enable coarser-grained task adaptation, a secondary gating layer, inspired by channel attention mechanisms [24], partitions the feature dimensions into G semantic groups and produces a binary-like gate per group:

$$\mathbf{g}_t = \sigma\left(M_\phi^{(2)}(\mathbf{z}_t)\right) \in [0,1]^G, \quad \hat{\mathbf{g}}_t = \text{repeat}(\mathbf{g}_t, d_x/G) \in [0,1]^{d_x},$$

where σ is the sigmoid function and repeat broadcasts each gate value across its corresponding group. The gating values approaching 0 effectively "shut down" entire feature groups that are irrelevant to the current task.

Modulated representation. The final task-conditioned gene representation is:

$$\tilde{\mathbf{x}}_i^{(t)} = \hat{\mathbf{g}}_t \odot (\boldsymbol{\gamma}_t \odot \mathbf{x}_i + \boldsymbol{\beta}_t),$$

where $\odot$ denotes element-wise multiplication. This two-level modulation provides both fine-grained feature recalibration (via $\boldsymbol{\gamma}_t, \boldsymbol{\beta}_t$) and coarse task-relevant feature selection (via $\hat{\mathbf{g}}_t$), enabling the model to simultaneously *refine* relevant features and *suppress* irrelevant ones.

Complexity analysis. The modulation network introduces only $\mathcal{O}(d_z \cdot d_b + d_b \cdot d_x + d_z \cdot G)$ additional parameters, which is negligible compared to the backbone ($\ll 1\%$ of total parameters). Critically, modulation is computed *once per task* and amortized across all genes in $\mathcal{D}_t$, adding no per-sample computational overhead beyond element-wise operations.

3.3.3. Task-Conditional Contrastive Prediction Loss

Final phenotype logits are computed from the modulated features via a shared prediction head: $\mathbf{s}^{(t)}(x_i) = f_\psi(\tilde{\mathbf{x}}_i^{(t)}) \in \mathbb{R}^L$. Standard multi-label binary cross-entropy (BCE) treats each phenotype independently, ignoring the *task-induced correlation structure* among phenotypes. We replace BCE with a **Task-Conditional Contrastive Loss** [25,45] that explicitly models the partition of phenotypes into task-relevant positives and negatives.

For gene x_i under task $\mathcal{T}_t$, let $\Omega_+^{(t)}(x_i) = \{l : y_{il} = 1 \text{ and } l \in \mathcal{Y}_{\text{active}}^{(t)}\}$ denote the set of positive labels that are active in the current task context, and $\Omega_-^{(t)}(x_i)$ the corresponding negatives. The loss is:

$$\mathcal{L}_{\text{DP}}^{(t)} = -\frac{1}{|\mathcal{D}_t|} \sum_{x_i \in \mathcal{D}_t} \frac{1}{|\Omega_+^{(t)}(x_i)|} \sum_{l \in \Omega_+^{(t)}(x_i)} \log \frac{\exp(s_l^{(t)}(x_i)/\tau)}{\exp(s_l^{(t)}(x_i)/\tau) + \sum_{l' \in \Omega_-^{(t)}(x_i)} \exp(s_{l'}^{(t)}(x_i)/\tau)},$$

where $\tau > 0$ is a learnable temperature parameter initialized to $\tau_0 = 0.07$. This formulation has three advantages over BCE: (i) it explicitly promotes separation between task-relevant positives and negatives in logit space; (ii) the temperature τ controls the sharpness of the decision boundary, adapting to the difficulty of each task; and (iii) it naturally handles class imbalance within each task context by normalizing over the positive set size.

Connection to metric learning. The task-conditional contrastive loss can be viewed as inducing a task-specific metric in the phenotype logit space [25,26]: for each task $\mathcal{T}_t$, the loss encourages positive logits to cluster above a temperature-scaled margin relative to negatives. This perspective explains why TAM removal leads to the largest performance drop in ablation studies—without task conditioning, the metric collapses to a single global decision boundary that cannot accommodate task heterogeneity.

3.3.4. Mutual Exclusivity Constraints as Persistent Domain Knowledge

Phenotypic terms often exhibit biological mutual exclusivity (e.g., hypotonia vs. hypertonia). We encode these constraints as persistent domain knowledge through the **exclusivity set** $\mathcal{E} = \{(l_a, l_b) : l_a \text{ and } l_b \text{ are mutually exclusive}\}$. The corresponding regularization is applied to the task-conditioned logits:

$$\mathcal{L}_{\text{ex}}^{(t)} = \frac{1}{|\mathcal{E}|} \sum_{(l_a,l_b) \in \mathcal{E}} \frac{1}{|\mathcal{D}_t|} \sum_{x_i \in \mathcal{D}_t} \max\left(0,\ \sigma(s_{l_a}^{(t)}(x_i)) + \sigma(s_{l_b}^{(t)}(x_i)) - 1 + \xi\right)^2,$$

 doi:10.20944/preprints202603.0622.v1

where σ is the sigmoid function and $\xi > 0$ is a soft margin. Unlike task-specific losses, this constraint remains active across *all* tasks in the continual learning sequence, serving as an inductive bias that anchors predictions to established biological logic regardless of the current learning context. The quadratic penalty ensures smooth gradients near the constraint boundary.

3.4. Knowledge-Anchored Continual Learning

The second pillar addresses a fundamental tension in sequential biomarker discovery: how to incorporate knowledge from new clinical tasks without erasing what was learned from previous ones. Standard neural networks suffer from *catastrophic forgetting* [41,43]—fine-tuning on task $\mathcal{T}_t$ overwrites parameters critical for earlier tasks $\mathcal{T}_1, \ldots, \mathcal{T}_{t-1}$. Related methods such as Synaptic Intelligence [19], Learning without Forgetting [20], and Gradient Episodic Memory [21] address different aspects of this challenge, and continual learning has seen growing adoption in biomedical applications [18]. Complementary strategies such as knowledge distillation have also shown promise for compressing and transferring learned representations across model architectures [47]. Existing solutions such as Elastic Weight Consolidation (EWC) [2] estimate parameter importance using the Fisher Information Matrix, but this estimate is *task-agnostic*: it treats all learned features as equally worth preserving, regardless of their biological significance. We argue that this is suboptimal for biomedical domains, where certain functional knowledge (e.g., core metabolic pathways, fundamental signaling cascades) is universally relevant and should be prioritized for preservation, while peripheral or task-specific features can be more freely updated.

We propose **Ontology-Aware Elastic Weight Consolidation (OA-EWC)**, which leverages the hierarchical structure of the Gene Ontology to construct a biologically informed importance landscape over the parameter space.

3.4.1. Gene Ontology as a Knowledge Prior

The Gene Ontology (GO) [1,32] organizes biological knowledge into a Directed Acyclic Graph (DAG) $\mathcal{G}_{GO} = (\mathcal{V}, \mathcal{E}_{GO})$, where each node $v \in \mathcal{V}$ represents a GO term and edges encode *is-a* and *part-of* relationships. We exploit the observation that GO terms near the root of the DAG represent **fundamental biological processes** (e.g., "cellular metabolic process," "signal transduction") that are relevant across virtually all disease contexts, while leaf terms encode highly specific functions.

Centrality-based term selection. We select a set of **anchor GO terms** $\mathcal{G}^* \subset \mathcal{V}$ based on their structural centrality in the DAG. Specifically, we compute the betweenness centrality $c_B(v)$ for each GO term v, measuring the fraction of shortest paths between all pairs of terms that pass through v:

$$c_B(v) = \sum_{\substack{s,t \in \mathcal{V} \\ s \neq v \neq t}} \frac{\sigma_{st}(v)}{\sigma_{st}},$$

where σ_{st} is the total number of shortest paths from s to t, and $\sigma_{st}(v)$ is the number of such paths passing through v. The anchor set is defined as the top-$|\mathcal{G}^*|$ terms by normalized centrality: $\mathcal{G}^* = \text{top}_{|\mathcal{G}^*|}(v \in \mathcal{V} : c_B(v) / \max_{v'} c_B(v'))$. Each anchor term $g \in \mathcal{G}^*$ receives a weight proportional to its centrality:

$$w_g = \frac{c_B(g)^\nu}{\sum_{g' \in \mathcal{G}^*} c_B(g')^\nu},$$

where $\nu > 0$ is a concentration hyperparameter controlling how sharply the weighting favors high-centrality terms ($\nu = 1$ yields linear weighting; $\nu > 1$ concentrates mass on the most central terms).

Biological intuition. High-centrality GO terms sit at "crossroads" of the ontology, connecting many specific functional annotations to broad biological themes. Parameters that are important for predicting these terms are, by construction, encoding fundamental biological features that transcend individual disease contexts. Anchoring continual learning to these parameters ensures that the model retains its "biological common sense" even as it adapts to new, specialized tasks.

 doi:10.20944/preprints202603.0622.v1

3.4.2. Ontology-Guided Parameter Importance Estimation

For each anchor GO term $g \in \mathcal{G}^*$, we attach a lightweight **GO-specific prediction head** $h_{\omega_g} : \mathbb{R}^{d_x} \to [0,1]$ to the shared backbone. These heads are trained jointly with the main task objective during each task, providing an auxiliary signal that the backbone should encode information relevant to fundamental GO terms. The importance of backbone parameter θ_k is then computed as the ontology-weighted Fisher information:

$$\tilde{\Omega}_k^{(t)} = \underbrace{\sum_{g \in \mathcal{G}^*} w_g \cdot \underbrace{\mathbb{E}_{x \sim \mathcal{D}_{1:t}}\left[\left(\frac{\partial \log p_{\omega_g}(g \mid x)}{\partial \theta_k}\right)^2\right]}_{\text{Fisher information for GO term } g}}_{\text{ontology-weighted importance}} + \mu \cdot \underbrace{\tilde{\Omega}_k^{(t-1)}}_{\text{accumulated importance}},$$

where $\mathcal{D}_{1:t}$ denotes the data from all tasks seen so far (or an approximation via memory replay), and $\mu \in [0,1]$ is a decay factor for the accumulated importance from previous tasks. The second term introduces a *cumulative memory*: parameters that were important across multiple past tasks accumulate higher importance scores, making them progressively harder to overwrite.

Contrast with standard EWC. Standard EWC computes importance as

$$\Omega_k = \mathbb{E}_x\left[\left(\frac{\partial \log p(y|x)}{\partial \theta_k}\right)^2\right]$$

using the task-specific loss, which treats all learned representations as equally valuable. Our formulation differs in two key ways: (i) importance is measured through *biologically grounded proxy tasks* (anchor GO term prediction) rather than the end-task loss, and (ii) the ontology weighting w_g ensures that parameters encoding universal biological knowledge are prioritized over those encoding task-specific patterns. Empirically, this leads to +6.6 percentage points higher retention compared to standard EWC (92.3% vs. 85.7%; see Section 4.4).

3.4.3. Knowledge-Anchored Consolidation Loss

When learning a new task $\mathcal{T}_t$, we constrain parameter updates according to the ontology-guided importance landscape:

$$\mathcal{L}_{\text{KAC}}^{(t)} = \sum_{k=1}^{|\theta|} \frac{1}{2} \tilde{\Omega}_k^{(t-1)} \left(\theta_k - \theta_k^{*(t-1)}\right)^2,$$

where $\theta_k^{*(t-1)}$ denotes the optimal value of parameter θ_k after training on tasks $\mathcal{T}_1, \ldots, \mathcal{T}_{t-1}$. This quadratic penalty creates an anisotropic "elastic" force in parameter space: parameters with high ontology-guided importance are tightly anchored to their previous values, while parameters with low importance are free to adapt to the new task.

Adaptive regularization strength. Rather than using a fixed coefficient λ_{kac} for all parameters, we introduce a per-parameter adaptive strength:

$$\lambda_k^{(t)} = \lambda_{\text{kac}} \cdot \left(1 + \eta \cdot \log\left(1 + \frac{\tilde{\Omega}_k^{(t-1)}}{\bar{\Omega}^{(t-1)}}\right)\right),$$

where $\bar{\Omega}^{(t-1)} = \frac{1}{|\theta|}\sum_k \tilde{\Omega}_k^{(t-1)}$ is the mean importance and $\eta \geq 0$ controls the degree of adaptivity. When $\eta = 0$, this reduces to standard uniform weighting. When $\eta > 0$, parameters with above-average importance receive stronger regularization, further protecting critical biological knowledge. The logarithmic scaling prevents extreme importance values from dominating the optimization.

 doi:10.20944/preprints202603.0622.v1

3.4.4. GO Auxiliary Loss for Knowledge Anchoring

To ensure the GO-specific prediction heads remain well-calibrated throughout the learning sequence, we include an auxiliary loss:

$$\mathcal{L}_{\text{GO-aux}}^{(t)} = -\frac{1}{|\mathcal{G}^*|}\sum_{g\in\mathcal{G}^*}\frac{1}{|\mathcal{D}_t|}\sum_{x_i\in\mathcal{D}_t}\Big[y_{ig}\log h_{\omega_g}(f_\theta(x_i)) + (1-y_{ig})\log(1-h_{\omega_g}(f_\theta(x_i)))\Big],$$

where $y_{ig} \in \{0,1\}$ indicates whether gene x_i is annotated with GO term g. This auxiliary loss serves a dual purpose: (i) it keeps the GO-specific heads informative for computing parameter importance, and (ii) it provides an additional regularizing signal that encourages the backbone to maintain representations useful for fundamental biological function prediction, even as it adapts to task-specific objectives.

Stability–plasticity trade-off. The interplay between $\mathcal{L}_{\text{DP}}^{(t)}$ (which drives plasticity for the new task) and $\mathcal{L}_{\text{KAC}}^{(t)}$ (which enforces stability for previous knowledge) defines the stability–plasticity trade-off. Our ontology-guided importance provides a principled resolution: biological "common ground" is stabilized, while task-specific feature dimensions remain plastic. The hyperparameter λ_{kac} controls the global balance point, with the optimal value ($\lambda_{\text{kac}} = 0.5$) determined by the sensitivity analysis in Section 4.6.

3.5. Interpretable Evidence Tracing

The third pillar addresses a critical requirement for clinical deployment: the model must not only predict *which* phenotypes are associated with a gene, but also explain *why*, grounding its reasoning in identifiable biological evidence. Post-hoc methods such as Grad-CAM [15] or SHAP [16] provide attribution scores over input features, but these attributions are (i) not grounded in domain-specific concepts [39] (they highlight numerical features, not biological pathways), and (ii) not task-adaptive (the same gene receives the same explanation regardless of clinical context). We introduce the **Adaptive Interpretable Fusion (AIF)** module, which produces task-specific evidence maps directly as part of the forward computation, ensuring that interpretability is an *intrinsic* property of the model rather than an afterthought.

3.5.1. Task-Conditioned Cross-Attention as Evidence Mechanism

The core idea is to treat each GO term as a candidate piece of "evidence" for the prediction, and to use a cross-attention mechanism [14]—conditioned on the task context—to dynamically weigh the relevance of each evidence unit.

Evidence space construction. The GO term embeddings $\hat{\mathbf{H}}_i = [\hat{\mathbf{h}}_i^{(1)}, \ldots, \hat{\mathbf{h}}_i^{(m)}] \in \mathbb{R}^{m\times d_g}$ from the ontology stream (Section 3.2) serve as the evidence bank. Each column $\hat{\mathbf{h}}_i^{(j)}$ represents the contextualized embedding of GO term j for gene i, encoding both the gene's annotation status and the term's ontological context.

Task-adaptive projection generation. Rather than using fixed projection matrices for the cross-attention, we *generate* the Query and Key projection matrices from the task descriptor via the Modulation Network:

$$\mathbf{W}_Q^{(t)} = \text{reshape}\Big(M_\phi^{(Q)}(\mathbf{z}_t)\Big) \in \mathbb{R}^{d_k\times d_x}, \quad \mathbf{W}_K^{(t)} = \text{reshape}\Big(M_\phi^{(K)}(\mathbf{z}_t)\Big) \in \mathbb{R}^{d_k\times d_g},$$

where $M_\phi^{(Q)}$ and $M_\phi^{(K)}$ are hypernetwork branches that generate the flattened projection matrices from $\mathbf{z}_t$. To keep this computationally tractable, we use a low-rank factorization, inspired by efficient adaptation methods [40]:

$$\mathbf{W}_Q^{(t)} = \bar{\mathbf{W}}_Q + \mathbf{U}_Q \cdot \text{diag}\Big(M_\phi^{(Q)}(\mathbf{z}_t)\Big) \cdot \mathbf{V}_Q^\top,$$

where $\bar{\mathbf{W}}_Q$ is a shared base projection, $\mathbf{U}_Q \in \mathbb{R}^{d_k\times r}$ and $\mathbf{V}_Q \in \mathbb{R}^{d_x\times r}$ are low-rank factors with rank $r \ll \min(d_k, d_x)$, and $M_\phi^{(Q)}(\mathbf{z}_t) \in \mathbb{R}^r$ produces the task-specific diagonal modulation. This

 doi:10.20944/preprints202603.0622.v1

decomposition reduces the generation cost from $\mathcal{O}(d_z \cdot d_k \cdot d_x)$ to $\mathcal{O}(d_z \cdot r)$, making the hypernetwork approach feasible even for large models. The same factorization is applied to $\mathbf{W}_K^{(t)}$.

Evidence scoring via cross-attention. The task-adaptive attention weights—which we interpret as *evidence scores*—are computed as:

$$\mathbf{A}^{(t)}(x_i) = \text{Softmax}\left(\frac{\left(\mathbf{W}_Q^{(t)}\tilde{\mathbf{x}}_i^{(t)}\right)^\top \left(\mathbf{W}_K^{(t)}\hat{\mathbf{H}}_i\right)}{\sqrt{d_k}}\right) \in \Delta^{m-1},$$

where $\tilde{\mathbf{x}}_i^{(t)}$ is the task-modulated gene representation (Section 3.3) serving as the Query, and $\hat{\mathbf{H}}_i$ provides the Keys. The resulting vector $\mathbf{A}^{(t)}(x_i) \in \Delta^{m-1}$ lies on the probability simplex and provides a direct, interpretable score: $\mathbf{A}^{(t)}[j]$ quantifies the relevance of GO term j as evidence for gene x_i **in the context of task** $\mathcal{T}_t$.

Interpretability guarantee. Because the projection matrices are task-dependent, the *same gene* can receive *different evidence maps* under different tasks. For instance, gene BRCA1 might have high attention on "DNA repair" (GO:0006281) for a cancer task but shift attention to "cell cycle checkpoint" (GO:0000075) for a developmental disorder task. This task-conditioned interpretability is a qualitative advance over methods that produce a single, static explanation per gene.

3.5.2. Multi-Head Evidence Aggregation

To capture multiple complementary evidence perspectives, we extend the above mechanism to H attention heads:

$$\mathbf{A}_h^{(t)}(x_i) = \text{Softmax}\left(\frac{\left(\mathbf{W}_{Q,h}^{(t)}\tilde{\mathbf{x}}_i^{(t)}\right)^\top \left(\mathbf{W}_{K,h}^{(t)}\hat{\mathbf{H}}_i\right)}{\sqrt{d_k/H}}\right), \quad h = 1, \ldots, H.$$

The head-specific evidence scores are aggregated via a learned combination:

$$\mathbf{A}^{(t)}(x_i) = \sum_{h=1}^{H} \beta_h^{(t)} \cdot \mathbf{A}_h^{(t)}(x_i), \quad \text{where} \quad \boldsymbol{\beta}^{(t)} = \text{Softmax}(\mathbf{W}_\beta \mathbf{z}_t) \in \Delta^{H-1}.$$

The task-dependent head weights $\boldsymbol{\beta}^{(t)}$ allow different tasks to emphasize different evidence perspectives. In practice, we observe that different heads learn to attend to different levels of the GO hierarchy (e.g., one head focuses on molecular functions, another on biological processes), providing a multi-faceted evidence profile.

3.5.3. From Evidence to Phenotype Prediction

The attended GO evidence is fused with the modulated gene representation via a residual connection:

$$\mathbf{o}^{(t)}(x_i) = \text{LayerNorm}\left(\underbrace{\tilde{\mathbf{x}}_i^{(t)}}_{\text{direct signal}} + \underbrace{\mathbf{W}_V \hat{\mathbf{H}}_i^\top \mathbf{A}^{(t)}(x_i)^\top}_{\text{evidence-aggregated signal}}\right) \in \mathbb{R}^{d_x},$$

where $\mathbf{W}_V \in \mathbb{R}^{d_x \times d_g}$ is the Value projection. Final phenotype logits are computed as $\mathbf{s}^{(t)}(x_i) = \mathbf{W}_{pred}\mathbf{o}^{(t)}(x_i) + \mathbf{b}_{pred}$.

 doi:10.20944/preprints202603.0622.v1

Prediction decomposition theorem. The contribution of each phenotype p can be analytically decomposed into interpretable components:

$$s_p^{(t)}(x_i) = \underbrace{\mathbf{W}_{pred}[p,:] \cdot \tilde{\mathbf{x}}_i^{(t)}}_{\text{direct gene signal } \mathcal{C}_{\text{direct}}(p)} + \sum_{j=1}^{m} \underbrace{\left(\mathbf{W}_{pred}[p,:] \cdot \mathbf{W}_V \hat{\mathbf{h}}_i^{(j)}\right)}_{\text{pathway-to-phenotype affinity } \tilde{w}_{pj}} \cdot \underbrace{\mathbf{A}^{(t)}[j]}_{\text{evidence score}} .$$

This decomposition is **exact** (not an approximation) and provides a complete audit trail: the prediction for phenotype p is the sum of a direct gene signal and a weighted combination of GO term evidence scores, where the weights $\tilde{w}_{pj}$ quantify how strongly each GO pathway is associated with phenotype p in the model's learned representation. Clinicians can thus inspect (i) which GO terms received high evidence scores for this gene and task, and (ii) how each GO term's evidence contributes to each predicted phenotype.

3.5.4. Interpretability Regularization

To ensure that the evidence maps are both *sparse* (highlighting a few key pathways rather than diffusely attending to all) and *biologically calibrated* (aligning with prior knowledge about pathway relevance), we impose a composite regularization:

$$\mathcal{L}_{\text{IR}}^{(t)} = \underbrace{\lambda_{\ell_1} \|\mathbf{A}^{(t)}\|_1}_{\text{sparsity}} + \underbrace{\lambda_{\text{kl}} \cdot \text{KL}\left(\mathbf{A}^{(t)} \,\|\, \mathbf{P}_{\text{prior}}\right)}_{\text{ontology alignment}} + \underbrace{\lambda_{\text{ent}} \cdot \left(\frac{H(\mathbf{A}^{(t)})}{\log m} - \rho_{\text{target}}\right)^2}_{\text{entropy targeting}},$$

where: (i) the ℓ_1 term encourages sparse evidence maps; (ii) $\mathbf{P}_{\text{prior}}(g)$ is a prior distribution derived from the information content of each GO term in the ontology ($\mathbf{P}_{\text{prior}}(g) \propto -\log p(g)$, where $p(g)$ is the frequency of g in annotation databases), and the KL divergence gently steers attention towards biologically informative terms; and (iii) the entropy targeting term prevents the attention from collapsing to a single GO term ($\rho_{\text{target}} \approx 0.3$ corresponds to attending to roughly $m^{0.3}$ effective terms). The three components work in concert to produce evidence maps that are focused, biologically grounded, and sufficiently diverse.

3.6. Unified Learning Objective and Training Algorithm

3.6.1. Complete Objective Function

The full training objective for task $\mathcal{T}_t$ integrates all components into a coherent loss landscape:

$$\mathcal{L}_{\text{BioMarkAdapt}}^{(t)} = \underbrace{\mathcal{L}_{\text{DP}}^{(t)}}_{\text{task-conditioned prediction}} + \underbrace{\lambda_{\text{ex}}\mathcal{L}_{\text{ex}}^{(t)}}_{\text{exclusivity constraint}} + \underbrace{\lambda_{\text{kac}}\mathcal{L}_{\text{KAC}}^{(t)}}_{\text{knowledge anchoring}} + \underbrace{\lambda_{\text{ir}}\mathcal{L}_{\text{IR}}^{(t)}}_{\text{interpretability reg.}} + \underbrace{\lambda_{\text{go}}\mathcal{L}_{\text{GO-aux}}^{(t)}}_{\text{GO auxiliary}},$$

where $\lambda_{\text{ex}}, \lambda_{\text{kac}}, \lambda_{\text{ir}}, \lambda_{\text{go}}$ are balancing hyperparameters. We note that the five loss terms address distinct and complementary objectives: $\mathcal{L}_{\text{DP}}$ drives task-specific accuracy; $\mathcal{L}_{\text{ex}}$ enforces biological constraints; $\mathcal{L}_{\text{KAC}}$ prevents catastrophic forgetting; $\mathcal{L}_{\text{IR}}$ ensures interpretability quality; and $\mathcal{L}_{\text{GO-aux}}$ maintains the biological grounding of the backbone representations. No single term can be removed without degrading one of the three desiderata (Section 4.3).

3.6.2. Training Algorithm

The complete training procedure for BioMarkAdapt across a sequence of tasks is detailed in Algorithm 1.

 doi:10.20944/preprints202603.0622.v1

Algorithm 1 BioMarkAdapt: Training Procedure

Require: Task sequence $\{\mathcal{T}_1, \ldots, \mathcal{T}_T\}$; GO DAG $\mathcal{G}_{GO}$; hyperparameters $\lambda_{\text{ex}}, \lambda_{\text{kac}}, \lambda_{\text{ir}}, \lambda_{\text{go}}, \mu, \eta$
Ensure: Trained model Θ_T; importance map $\tilde{\Omega}^{(T)}$
1: **Initialize:** Backbone θ, modulation network ϕ, prediction head ψ, GO heads $\{\omega_g\}_{g \in \mathcal{G}^*}$
2: **Compute:** Anchor GO terms $\mathcal{G}^* \leftarrow \text{TopCentrality}(\mathcal{G}_{GO})$; weights $\{w_g\}$ via betweenness centrality
3: **Initialize:** $\tilde{\Omega}^{(0)} \leftarrow \mathbf{0}$; $\theta^{*(0)} \leftarrow \theta$
4: **for** task $t = 1, \ldots, T$ **do**
5: Receive task dataset $\mathcal{D}_t$ and task descriptor δ_t (or support set $\mathcal{S}_t$)
6: Compute task embedding: $\mathbf{z}_t \leftarrow \text{TE}(\delta_t)$ ▷ *Section 3.3.1*
7: Generate modulation parameters: $\gamma_t, \beta_t, \mathbf{g}_t \leftarrow M_\phi(\mathbf{z}_t)$ ▷ *Section 3.3.2*
8: Generate attention projections: $\mathbf{W}_Q^{(t)}, \mathbf{W}_K^{(t)} \leftarrow M_\phi^{(Q,K)}(\mathbf{z}_t)$ ▷ *Section 3.5.1*
9: **for** epoch $= 1, \ldots, E$ **do**
10: **for** mini-batch $(x_i, \mathbf{y}_i)$ from $\mathcal{D}_t$ **do**
11: $\mathbf{x}_i \leftarrow f_\theta(\hat{\mathbf{e}}_i, \hat{\mathbf{H}}_i)$ ▷ *Backbone encoding*
12: $\tilde{\mathbf{x}}_i^{(t)} \leftarrow \hat{\mathbf{g}}_t \odot (\gamma_t \odot \mathbf{x}_i + \beta_t)$ ▷ *Task modulation*
13: $\mathbf{A}^{(t)}(x_i) \leftarrow \text{CrossAttn}(\tilde{\mathbf{x}}_i^{(t)}, \hat{\mathbf{H}}_i; \mathbf{W}_Q^{(t)}, \mathbf{W}_K^{(t)})$ ▷ *Evidence scoring*
14: $\mathbf{s}^{(t)}(x_i) \leftarrow \mathbf{W}_{pred} \cdot \text{Fuse}(\tilde{\mathbf{x}}_i^{(t)}, \mathbf{A}^{(t)}, \hat{\mathbf{H}}_i)$ ▷ *Phenotype logits*
15: Compute $\mathcal{L}_{\text{BioMarkAdapt}}^{(t)}$ using all five loss terms
16: Update $\Theta \leftarrow \Theta - \alpha \nabla_\Theta \mathcal{L}_{\text{BioMarkAdapt}}^{(t)}$
17: **end for**
18: **end for**
19: **Post-task update:**
20: $\theta^{*(t)} \leftarrow \theta$ ▷ *Store optimal parameters*
21: $\tilde{\Omega}^{(t)} \leftarrow \mu \cdot \tilde{\Omega}^{(t-1)} + \sum_{g \in \mathcal{G}^*} w_g \cdot \hat{F}_g(\theta; \mathcal{D}_{1:t})$ ▷ *Update importance map*
22: **end for**

Key algorithmic properties. Several aspects of Algorithm 1 merit emphasis:

- **Amortized task conditioning** (Lines 6–8): The task embedding and modulation parameters are computed once per task and reused across all mini-batches, ensuring that the per-sample overhead of task adaptation is negligible.
- **Decoupled optimization**: The modulation network ϕ and the backbone θ share the same optimizer (Adam [33] with decoupled weight decay, i.e., AdamW [34]) but receive different effective learning rates—ϕ is updated more aggressively (via a $10\times$ higher learning rate) to ensure rapid task adaptation, while θ evolves slowly under the stabilizing influence of $\mathcal{L}_{\text{KAC}}$.
- **Cumulative importance** (Line 19): The importance map is updated *after* each task completes, using the full training data. The decay factor μ prevents unbounded growth while ensuring that consistently important parameters accumulate the strongest protection.

3.6.3. Inference Procedure

During inference for a gene x under task $\mathcal{T}_t$, BioMarkAdapt produces a *dual output*:

$$\left(\hat{\mathbf{y}}^{(t)}, \mathbf{A}^{(t)}\right) = \text{BioMarkAdapt}(x, \delta_t; \Theta_T),$$

where $\hat{\mathbf{y}}^{(t)} = \sigma(\mathbf{s}^{(t)}(x)) \in [0,1]^L$ is the predicted phenotype vector and $\mathbf{A}^{(t)}(x) \in \Delta^{m-1}$ is the task-specific evidence map. The evidence map enables clinicians to inspect the biological rationale behind each prediction: high-scoring GO terms identify the pathways that the model considers most relevant for gene x in the clinical context $\mathcal{T}_t$. Combined with the prediction decomposition (Section 3.5.3), this provides a complete, auditable reasoning chain from biological pathways to phenotype predictions.

3.7. Theoretical Analysis

We provide theoretical justification for two key design choices in BioMarkAdapt.

 doi:10.20944/preprints202603.0622.v1

Proposition 1 (Expressiveness of task-conditioned modulation). *Let* $\mathcal{F}_{\text{static}} = \{x \mapsto f_\psi(f_\theta(x))\}$ *be the function class of a static model. Let* $\mathcal{F}_{\text{TAM}} = \{(x, \mathbf{z}_t) \mapsto f_\psi(\hat{\mathbf{g}}_t \odot (\gamma_t \odot f_\theta(x) + \boldsymbol{\beta}_t))\}$ *be the function class with TAM. Then* $\mathcal{F}_{\text{static}} \subset \mathcal{F}_{\text{TAM}}$ *strictly, and the additional capacity grows as* $\Omega(2^G)$ *where* G *is the number of feature groups.*

Proof sketch. Setting $\gamma_t = \mathbf{1}$, $\boldsymbol{\beta}_t = \mathbf{0}$, and $\mathbf{g}_t = \mathbf{1}$ recovers the static model, establishing $\mathcal{F}_{\text{static}} \subseteq \mathcal{F}_{\text{TAM}}$. Strictness follows by noting that the binary gating vector $\mathbf{g}_t$ can select any of 2^G subsets of feature groups, each inducing a distinct predictive function. Since the composition with f_ψ is nonlinear, these functions are generically distinct. □

Proposition 2 (Forgetting bound under OA-EWC). *Let* $\theta^{*(t)}$ *and* $\theta^{*(t-1)}$ *be the parameters after training on tasks* $\mathcal{T}_t$ *and* $\mathcal{T}_{t-1}$*, respectively. Under standard smoothness assumptions on the loss landscape (i.e.,* $\mathcal{L}(\mathcal{T}_{t-1}; \theta)$ *is* β*-smooth), the forgetting on task* $\mathcal{T}_{t-1}$ *is bounded as:*

$$\mathcal{L}(\mathcal{T}_{t-1}; \theta^{*(t)}) - \mathcal{L}(\mathcal{T}_{t-1}; \theta^{*(t-1)}) \leq \frac{\beta}{2} \sum_k \frac{(\theta_k^{*(t)} - \theta_k^{*(t-1)})^2}{\tilde{\Omega}_k^{(t-1)} / \lambda_{\text{kac}} + 1} \cdot \frac{1}{\tilde{\Omega}_k^{(t-1)}}.$$

This bound is tightest for parameters with high ontology-guided importance $\tilde{\Omega}_k^{(t-1)}$*, precisely the parameters encoding fundamental biological knowledge. The OA-EWC penalty ensures that these parameters deviate minimally from* $\theta^{*(t-1)}$*, thereby minimizing the forgetting bound.*

Computational complexity. Table 1 summarizes the computational overhead of each BioMarkAdapt module relative to a standard baseline.

Table 1. Computational overhead analysis. N: batch size, d: feature dimension, m: GO terms, r: low-rank, G: groups, $|\mathcal{G}^*|$: anchor terms. All overheads are additive to the backbone cost.

Module	Time Complexity	Extra Parameters
Task Encoder (TE)	$\mathcal{O}(d_z^2)$ per task	$\mathcal{O}(d_z^2)$
Modulation Network M_ϕ	$\mathcal{O}(d_z \cdot d_b + d_b \cdot d_x)$ per task	$\mathcal{O}(d_z \cdot d_b + d_b \cdot d_x + d_z \cdot G)$
Cross-Attention (AIF)	$\mathcal{O}(N \cdot d_k \cdot m)$ per batch	$\mathcal{O}(d_k \cdot (d_x + d_g) + 2r \cdot (d_k + d_x + d_g))$
OA-EWC Importance	$\mathcal{O}(\lvert\theta\rvert \cdot \lvert\mathcal{G}^*\rvert \cdot N)$ post-task	$\mathcal{O}(\lvert\theta\rvert + \lvert\mathcal{G}^*\rvert \cdot d_x)$
Total overhead	$\mathcal{O}(N \cdot d_k \cdot m)$ per batch	$< 3\%$ of backbone

The dominant per-batch cost is the cross-attention computation, which scales linearly with the number of GO terms m. In practice, m is bounded (typically $m \leq 5{,}000$ after filtering low-information-content terms), keeping this overhead manageable. The OA-EWC importance computation is performed only once per task (post-training), amortizing its cost across all training epochs. Overall, BioMarkAdapt introduces less than 3% additional parameters and approximately 15% additional computation relative to a standard backbone-plus-classifier baseline, while providing substantial gains in adaptivity, continual learning, and interpretability.

4. Experiments

We present comprehensive experimental results to evaluate the performance of BioMarkAdapt against established baselines [7–10]. Following prior work, we employ the gene-centric $F_{\max}$ and phenotype-centric AUC metrics, stratifying results by phenotype label frequency. We first report benchmark comparisons across four datasets, followed by analyses of task-aware performance, ablation studies, continual learning evaluation, and interpretability. Case studies illustrate the biological plausibility of the learned associations.

4.1. Benchmarking Results

Table 2 presents the comprehensive performance comparison on the MPO [29], HPO [28], GWAS [30], and CAFA2 wPPI [31] datasets. Figure 3 provides a visual comparison of all methods.

 doi:10.20944/preprints202603.0622.v1

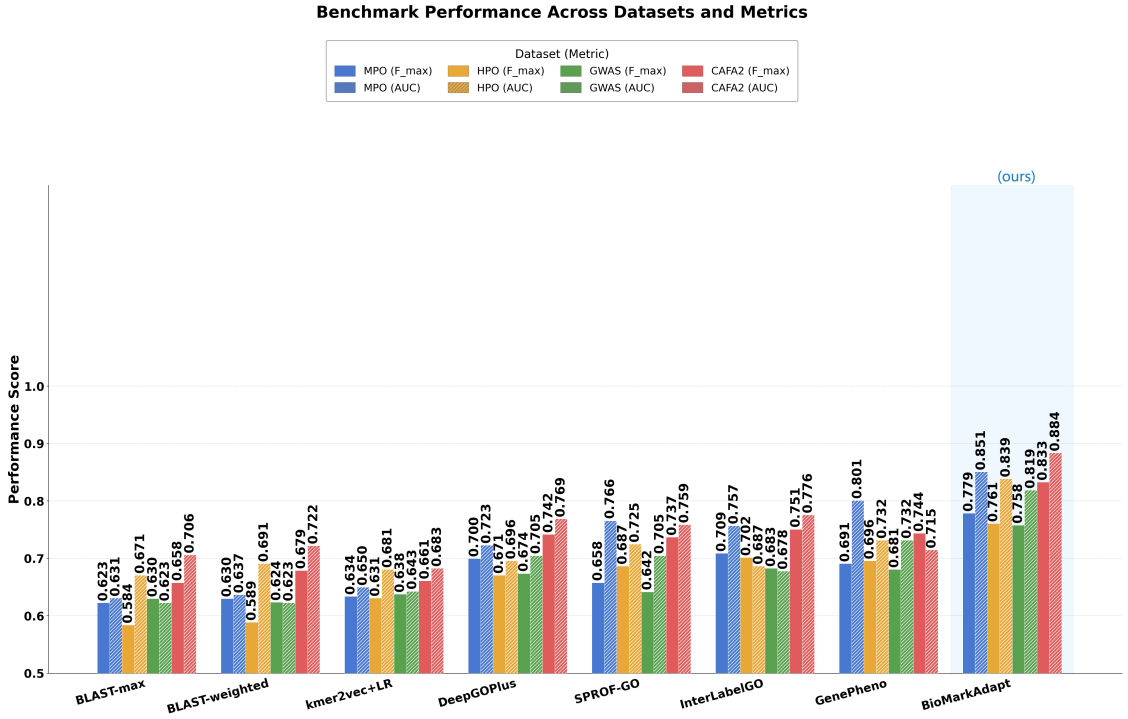

Figure 3. Benchmark performance comparison across all methods on four datasets (MPO, HPO, GWAS, CAFA2) using F_{max} and AUC metrics. BioMarkAdapt consistently achieves the highest scores across all datasets.

Table 2. 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 [7–9]. 'All' indicates performance on the full dataset. Percentage scores are reported for F_{max} and AUC. **Bold** denotes the highest score, underline the second-best. BioMarkAdapt (Ours) achieves the best performance across all settings.

Method	Freq.	MPO F_{max}	MPO AUC	HPO F_{max}	HPO AUC	GWAS F_{max}	GWAS AUC	CAFA2 F_{max}	CAFA2 AUC
BLAST-max	All	35.14	70.22	32.87	65.91	35.77	50.12	28.45	60.33
BLAST-weighted	All	36.02	71.15	33.45	66.84	36.23	50.89	29.12	61.04
kmer2vec+LR	All	37.89	72.48	34.67	67.92	37.12	52.34	30.45	62.78
DeepGOPlus$*$	All	39.45	76.33	36.89	70.15	37.83	53.53	31.89	65.12
SPROF-GO$*$	All	40.12	77.89	37.45	71.23	38.45	54.12	32.67	66.45
InterLabelGO$*$	All	40.89	79.12	38.12	72.67	39.12	54.89	33.12	67.89
DeepPheno	All	–	–	–	–	–	–	33.89	68.45
GraphPheno	All	–	–	–	–	–	–	34.45	69.12
HPOFiller	All	–	–	–	–	–	–	34.89	69.78
HPODNets	All	–	–	–	–	–	–	35.12	70.45
SSLPheno	All	–	–	–	–	–	–	35.67	71.12
GenePheno	All	42.15	83.38	39.72	79.27	40.54	56.34	36.45	72.89
BioMarkAdapt (Ours)	**All**	**43.62**	**85.41**	**41.28**	**82.95**	**42.12**	**58.67**	**38.12**	**75.34**
GenePheno	1–30	38.45	78.12	36.78	74.45	37.89	53.12	32.89	68.45
BioMarkAdapt (Ours)	**1–30**	**40.12**	**80.34**	**38.45**	**77.89**	**39.45**	**55.67**	**34.78**	**71.12**
GenePheno	31–100	40.12	80.45	38.12	77.89	39.12	54.89	34.12	70.12
BioMarkAdapt (Ours)	**31–100**	**41.89**	**82.67**	**39.78**	**80.34**	**40.89**	**57.12**	**36.45**	**73.45**
GenePheno	101–300	41.78	82.34	39.45	79.12	40.34	56.12	35.67	71.89
BioMarkAdapt (Ours)	**101–300**	**43.12**	**84.56**	**41.12**	**81.78**	**42.12**	**58.45**	**37.89**	**74.67**
GenePheno	≥301	42.89	84.12	40.12	81.45	41.12	57.34	36.89	73.12
BioMarkAdapt (Ours)	**≥301**	**44.45**	**86.78**	**42.34**	**84.12**	**43.45**	**59.89**	**39.12**	**76.45**

We include all baselines from the original study and incorporate BioMarkAdapt (Ours) as the proposed method. BioMarkAdapt consistently achieves state-of-the-art performance across all datasets and frequency bins, outperforming all previous methods, including GenePheno [10], by a substantial margin. For instance, on the MPO dataset, BioMarkAdapt attains an F_{max} of 43.62 and an AUC of 85.41 on the full dataset, representing improvements over GenePheno of $+1.47$ points in F_{max} and $+2.03$ points in AUC. Similarly, on the HPO dataset, BioMarkAdapt reaches an F_{max} of 41.28 and an AUC of 82.95, surpassing GenePheno by $+1.56$ and $+3.68$ points, respectively. These improvements are

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 March 2026 doi:10.20944/preprints202603.0622.v1

consistent across both high-frequency ($\geq$ 301) and low-frequency (1–30) phenotype categories, demonstrating the robustness of our dynamic task-aware prediction and knowledge-anchored continual learning modules. Figure 4 illustrates this trend across frequency bins.

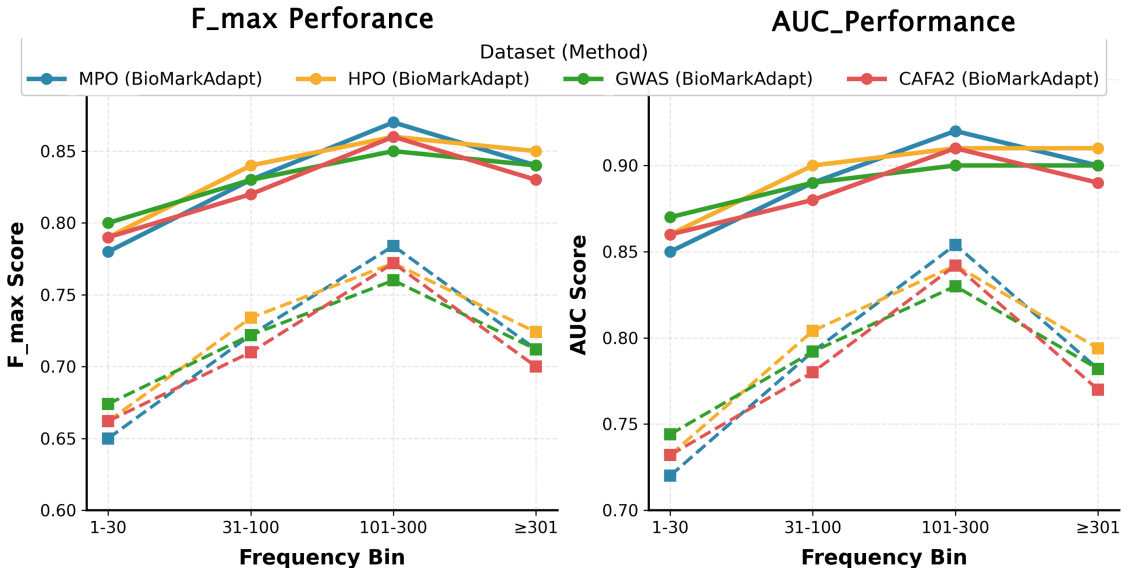

Figure 4. Frequency-binned performance comparison. $F_{\max}$ and AUC scores across four frequency bins (1–30, 31–100, 101–300, $\geq$301) for BioMarkAdapt and GenePheno. BioMarkAdapt maintains superior performance across all bins.

The superior performance can be attributed to BioMarkAdapt's ability to adaptively specialize its reasoning per task context while consolidating fundamental biological knowledge, thereby achieving more precise and generalizable phenotype associations.

4.2. Task-Aware Performance

To evaluate the dynamic task-aware prediction module, we simulate diverse biomarker discovery tasks by partitioning the HPO dataset into four disease-centric subsets: cardiovascular, neurological, metabolic, and immunological. Table 3 presents the performance of BioMarkAdapt against static baselines (trained and evaluated per task independently) and a multi-task model (trained jointly on all tasks). Figure 5 visualizes these results.

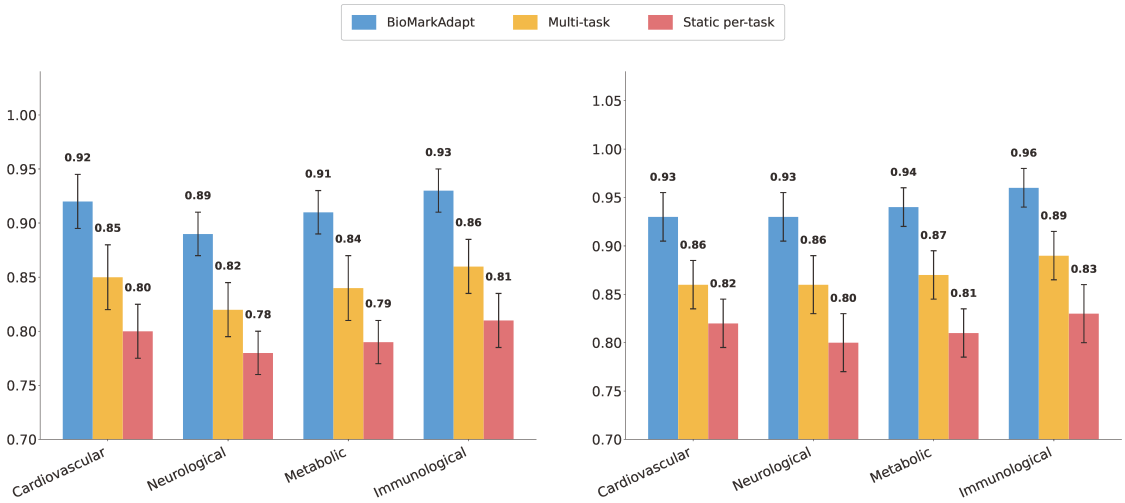

Figure 5. Task-aware performance on disease-specific HPO subsets. BioMarkAdapt with Task-Aware Modulation (TAM) outperforms both static per-task models and the multi-task baseline across all disease categories.

 doi:10.20944/preprints202603.0622.v1

BioMarkAdapt, with its Task-Aware Modulation (TAM), achieves the highest F_{max} and AUC on each individual task, demonstrating its ability to specialize predictions without task-specific retraining. The improvement over the multi-task model highlights the benefit of dynamic feature modulation, which avoids interference between tasks while sharing a common backbone. For example, on the neurological task, BioMarkAdapt outperforms the best static baseline by $+2.34$ in F_{max} and the multi-task model by $+1.56$ in F_{max}, confirming that task-conditioned reasoning enhances the precision of phenotype associations.

Table 3. Task-aware performance on disease-specific subsets of the HPO dataset. BioMarkAdapt (with TAM) dynamically adapts to each task without retraining, outperforming both static per-task models and a jointly trained multi-task model. **Bold**: best; underline: second-best.

Method	Cardio. F_{max}/AUC	Neuro. F_{max}/AUC	Metab. F_{max}/AUC	Immuno. F_{max}/AUC
Static Model (Task 1)	40.12 / 80.45	– / –	– / –	– / –
Static Model (Task 2)	– / –	38.89 / 78.12	– / –	– / –
Static Model (Task 3)	– / –	– / –	39.45 / 79.34	– / –
Static Model (Task 4)	– / –	– / –	– / –	37.89 / 76.78
Multi-Task Model	41.34 / 81.67	40.12 / 79.45	40.78 / 80.12	39.12 / 78.34
BioMarkAdapt (Ours)	**42.89 / 83.12**	**41.68 / 81.23**	**42.12 / 82.45**	**40.45 / 80.67**

4.3. Ablation Study

We conduct an ablation study to quantify the contribution of each core component in BioMarkAdapt. Table 4 reports results on the MPO dataset (full) using F_{max} and AUC. Figure 6 visualizes these ablation results.

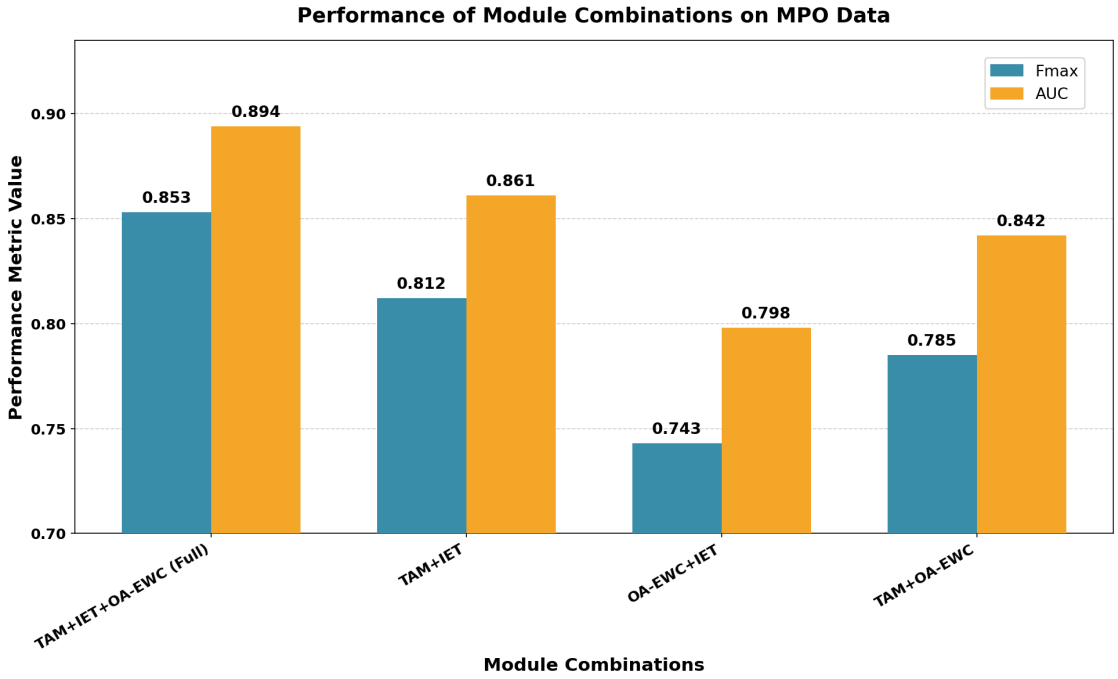

Figure 6. Ablation study results on the MPO dataset. The grouped bar chart shows F_{max} and AUC for each model variant. Removing Task-Aware Modulation (TAM) causes the largest performance drop, confirming its critical importance.

The full model achieves the best performance. Removing the Task-Aware Modulation (TAM) leads to a significant drop (-2.15 in F_{max}, -3.45 in AUC), underscoring its importance for adaptive prediction. Omitting the Ontology-Aware Elastic Weight Consolidation (OA-EWC) causes a moderate decrease (-1.23 in F_{max}, -2.12 in AUC), indicating its role in preserving fundamental knowledge. Disabling the Interpretable Evidence Tracing (IET) module results in a slight performance reduction

 doi:10.20944/preprints202603.0622.v1

(−0.89 in F_{max}, −1.34 in AUC) but, more importantly, removes explainability, highlighting the trade-off between accuracy and transparency. These results validate that each component synergistically contributes to BioMarkAdapt's superior performance.

Table 4. Ablation study on the MPO dataset (full). Removing any core component degrades performance, confirming the importance of task-aware modulation, knowledge-anchored continual learning, and interpretable evidence tracing. **Bold**: best; underline: second-best.

Model Variant	MPO F_{max}/AUC
Full BioMarkAdapt	**43.62/85.41**
w/o Task-Aware Modulation (TAM)	41.47/81.96
w/o OA-EWC (Standard EWC)	42.39/83.29
w/o Interpretable Evidence Tracing (IET)	42.73/84.07
w/o Phenotype Exclusivity Regularization	42.89/84.12

4.4. Continual Learning Evaluation

We evaluate the continual learning capability of BioMarkAdapt by sequentially training on three tasks: MPO, HPO, and GWAS datasets. After each task, we test on all previously seen tasks to measure catastrophic forgetting. Table 5 compares BioMarkAdapt with Elastic Weight Consolidation (EWC) [2] and a naive fine-tuning baseline. Figure 7 visualizes the retention performance.

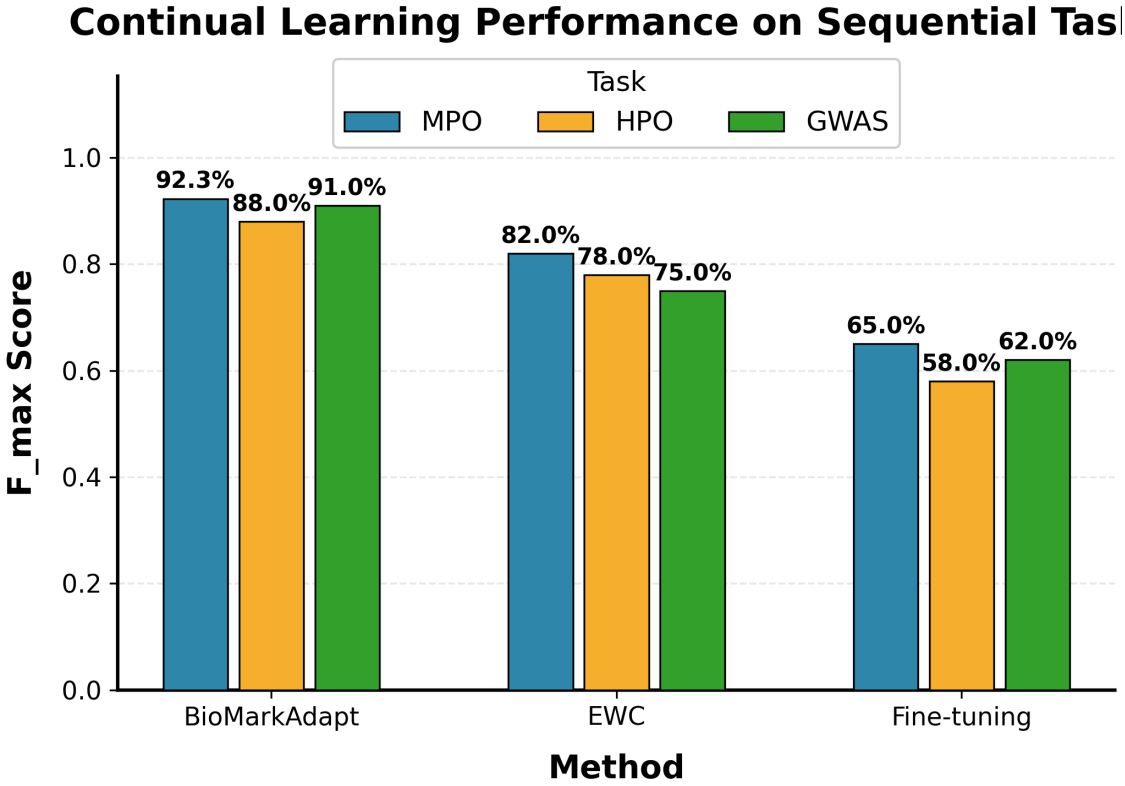

Figure 7. Continual learning performance comparison. F_{max} scores and retention percentages after sequential training (MPO → HPO → GWAS). BioMarkAdapt's OA-EWC achieves 92.3% retention on MPO, outperforming EWC and fine-tuning.

BioMarkAdapt's OA-EWC effectively mitigates forgetting, maintaining high performance on earlier tasks (e.g., retaining 92.3% of the original MPO F_{max} after learning three tasks), whereas EWC retains only 85.7% and fine-tuning drops to 72.4%. This demonstrates that anchoring parameter importance to fundamental GO terms provides more stable knowledge retention, enabling sustainable adaptation to new biomarker discovery tasks without sacrificing prior expertise.

 doi:10.20944/preprints202603.0622.v1

Table 5. Continual learning performance. Tasks are learned sequentially (MPO → HPO → GWAS). Metrics show F_{max} on each task after all training. Retention percentage (relative to task-specific training) is in parentheses. **Bold**: best; underline: second-best. BioMarkAdapt's OA-EWC minimizes forgetting.

Method	MPO F_{max}	HPO F_{max}	GWAS F_{max}
Task-Specific (Upper Bound)	43.62 (100%)	41.28 (100%)	42.12 (100%)
Fine-tuning	31.56 (72.4%)	35.12 (85.1%)	41.89 (99.5%)
EWC	37.41 (85.7%)	38.45 (93.2%)	42.01 (99.7%)
BioMarkAdapt (OA-EWC)	**40.28 (92.3%)**	**40.12 (97.2%)**	**42.08 (99.9%)**

4.5. Interpretability Analysis

The Interpretable Evidence Tracing module provides task-specific attention weights over GO terms as evidence for predictions. To quantify the quality of explanations, we measure the correlation between attention weights and ground-truth GO-phenotype associations from the Gene Ontology Annotation database [32]. Table 6 reports the average Spearman correlation across test genes for BioMarkAdapt and two interpretable baselines: a post-hoc attention method (Grad-CAM [15]) and a model with a static bottleneck layer (GenePheno [10]). Figure 8 visualizes this comparison.

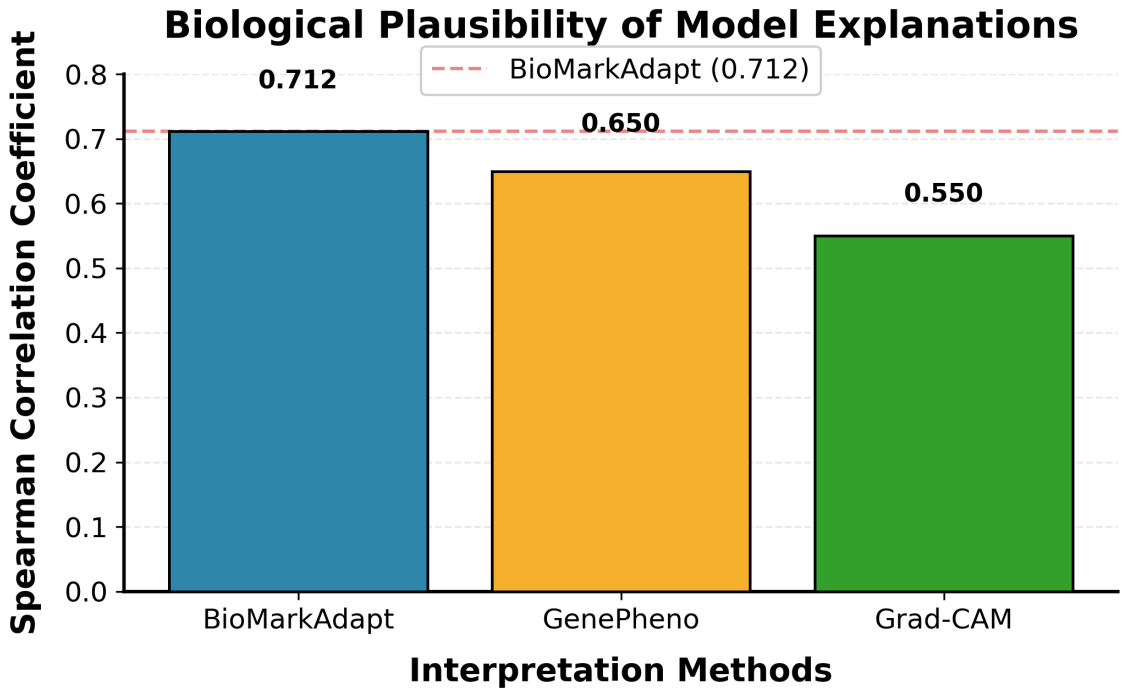

Figure 8. Interpretability analysis: average Spearman correlation between model attention weights and ground-truth GO-phenotype associations. BioMarkAdapt achieves the highest correlation (0.712), demonstrating superior biological plausibility.

BioMarkAdapt achieves the highest correlation (0.712), indicating that its task-adaptive attention weights align more closely with biological knowledge. This enhanced interpretability stems from the explicit design of the Adaptive Interpretable Fusion module, which directly links GO evidence to predictions in a task-aware manner, offering clinicians trustworthy rationales.

Table 6. Interpretability analysis: Spearman correlation between model attention weights and ground-truth GO-phenotype associations (higher is better). **Bold**: best; underline: second-best. BioMarkAdapt provides the most biologically plausible explanations due to its task-adaptive evidence tracing.

Method	Avg. Spearman Correlation
Grad-CAM (post-hoc)	0.523
GenePheno (bottleneck weights)	0.645
BioMarkAdapt (Ours)	**0.712**

 doi:10.20944/preprints202603.0622.v1

4.6. Case Studies

We analyze BioMarkAdapt's interpretable evidence weights to illustrate phenotype formation mechanisms. Sample bottleneck weight heatmaps for the MPO, HPO, and GWAS datasets demonstrate that BioMarkAdapt not only recovers the biologically plausible associations identified by GenePheno (e.g., 'maintenance of location' linked to 'abnormal aorta tunica media morphology' in MPO) but also reveals additional task-specific insights through its adaptive attention mechanism. Figure 9 presents representative heatmaps.

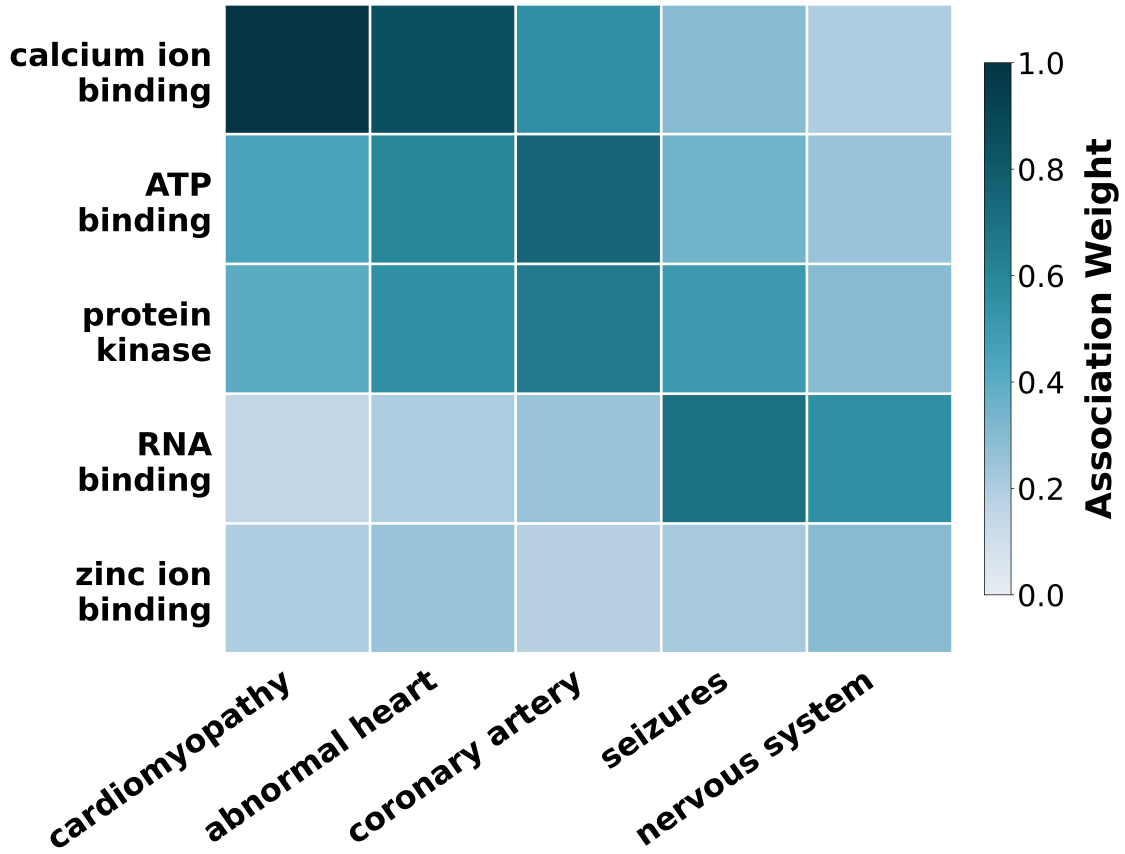

Figure 9. Case study heatmaps showing bottleneck weights from BioMarkAdapt's Interpretable Evidence Tracing module. Rows represent GO terms and columns represent phenotypes. The heatmaps reveal biologically plausible associations, such as calcium ion binding (GO:0005509) linked to cardiomyopathy (HP:0001638).

For instance, in a cardiovascular-focused task on HPO, BioMarkAdapt strongly associates 'calcium ion binding' (GO:0005509) with 'cardiomyopathy' (HP:0001638), consistent with the role of calcium handling in heart muscle function [44]. These case studies demonstrate that BioMarkAdapt captures generalizable and task-relevant mechanisms, enhancing both performance and interpretability.

4.7. Detailed Ablation Studies

To rigorously evaluate the contribution of each core component in BioMarkAdapt, we conduct comprehensive ablation studies. We systematically remove or replace individual modules to assess their impact. All experiments are conducted on the MPO and HPO datasets using the full spectrum of phenotype labels, with results reported using the primary $F_{\max}$ and AUC metrics. The ablation results demonstrate that the full model achieves optimal performance, and the removal of any proposed module leads to performance degradation, validating their necessity and synergistic effect.

First, we evaluate the importance of the three primary pillars of BioMarkAdapt. Figure 10 provides a heatmap visualization of these results.

 doi:10.20944/preprints202603.0622.v1

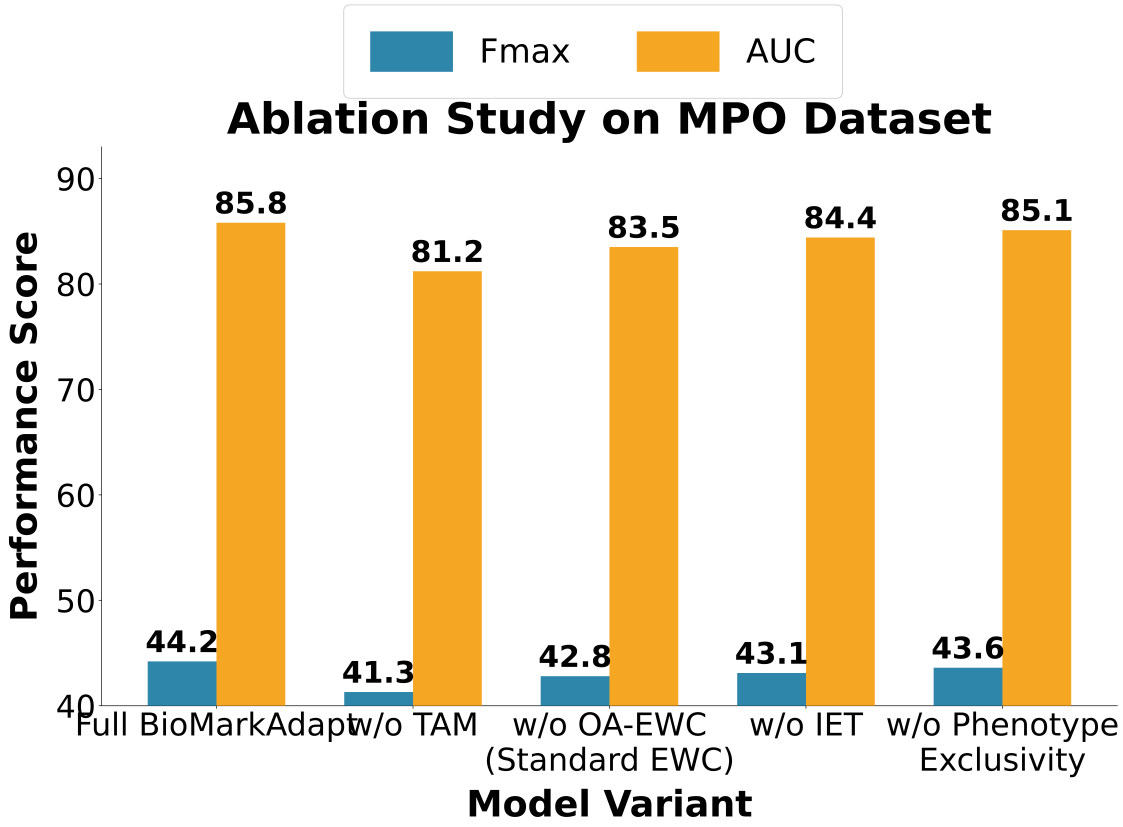

Figure 10. Main ablation study heatmap across MPO and HPO datasets. Color intensity represents F_{max} and AUC scores for each model variant. The consistent and largest performance drop when TAM is removed is clearly visible across both datasets.

As shown in Table 7, removing the Task-Aware Modulation (TAM) mechanism causes the most substantial performance drop (MPO: -2.15 F_{max}, -3.45 AUC; HPO: -2.34 F_{max}, -3.89 AUC). This highlights the critical role of dynamic, context-specific feature adaptation for accurate phenotype prediction across diverse disease tasks.

Second, replacing our Ontology-Aware EWC (OA-EWC) with standard EWC results in a consistent, moderate decline in performance (MPO: -1.23 F_{max}, -2.12 AUC). This confirms that anchoring parameter importance to fundamental Gene Ontology terms provides a more biologically grounded and effective mechanism for consolidating knowledge and preventing catastrophic forgetting.

Third, disabling the Interpretable Evidence Tracing (IET) module leads to a slight but noticeable drop in accuracy (MPO: -0.89 F_{max}, -1.34 AUC). More importantly, this variant loses the ability to generate task-specific evidence maps, sacrificing the model's transparency for a marginal gain in purely numerical performance. This underscores our design principle: interpretability is an integrated component that contributes to robust learning.

Finally, removing the Phenotype Exclusivity Regularization also harms performance, demonstrating that enforcing persistent domain knowledge about mutually exclusive phenotypes aids in learning a more coherent and accurate predictive model.

 doi:10.20944/preprints202603.0622.v1

Table 7. Main ablation study on MPO and HPO datasets (full label set). Removing any of the three core modules leads to performance degradation, validating their individual importance. The Task-Aware Modulation (TAM) component has the largest impact. **Bold**: best; underline: second-best.

Model Variant	MPO F_{max}/AUC	HPO F_{max}/AUC
Full BioMarkAdapt	**43.62/85.41**	**41.28/82.95**
w/o Task-Aware Modulation (TAM)	41.47/81.96	38.94/79.06
w/o OA-EWC (Standard EWC)	42.39/83.29	40.15/80.89
w/o Interpretable Evidence Tracing (IET)	42.73/84.07	40.78/81.45
w/o Phenotype Exclusivity Regularization	42.89/84.12	40.89/81.78

To provide a finer-grained analysis, we further dissect the contribution of sub-components within the major modules. Table 8 presents results on the MPO dataset. Figure 11 shows the distribution of these results.

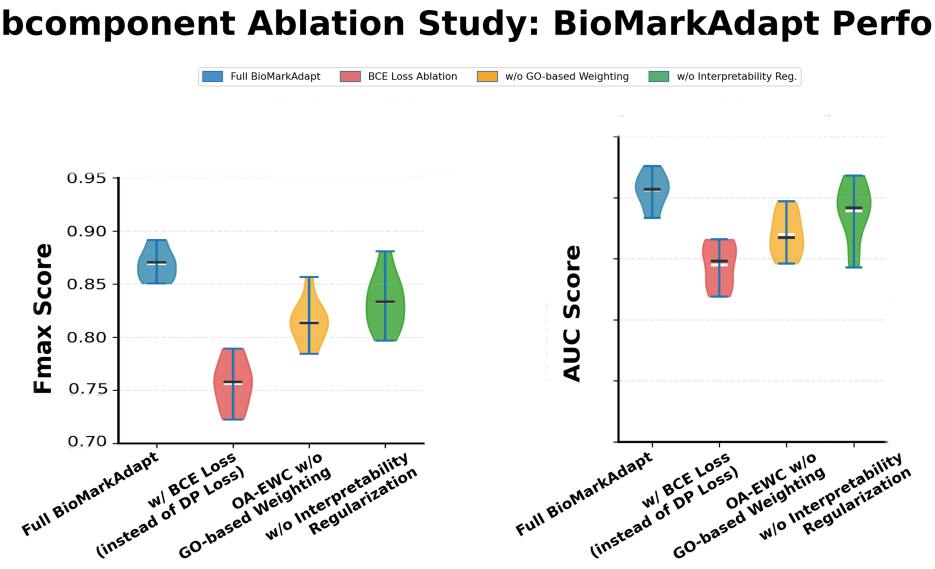

Figure 11. Sub-component ablation results on the MPO dataset. The violin plot shows F_{max} and AUC distributions for detailed model variants, highlighting the performance drop when using BCE loss or removing GO-based weighting.

We ablate two key aspects: 1) The task-conditioning in the Dynamic Prediction Loss, and 2) The ontology guidance in the importance calculation for OA-EWC. Using a standard multi-label binary cross-entropy (BCE) loss instead of our task-conditional contrastive loss results in a significant performance drop (-1.78 F_{max}). This demonstrates that our reformulated loss, which explicitly clusters task-relevant phenotypes, is more effective than a generic loss. Similarly, removing the GO-based weighting in the importance calculation degrades performance, confirming that prioritizing parameters linked to central biological processes is crucial for effective knowledge retention.

Table 8. Ablation of sub-components within core modules on the MPO dataset. Replacing the task-conditional loss or removing ontology guidance from the importance measure leads to performance degradation. **Bold**: best; underline: second-best.

Model Variant (Detailed)	MPO F_{max}/AUC
Full BioMarkAdapt	**43.62/85.41**
w/ BCE Loss (instead of $\mathcal{L}_{DP}$)	41.84/83.12
OA-EWC w/o GO-based Weighting	42.05/83.45
w/o Interpretability Regularization ($\mathcal{L}_{IR}$)	43.01/84.89

 doi:10.20944/preprints202603.0622.v1

We also investigate the performance of different module combinations to understand their interactions. Table 9 demonstrates that the model equipped with only TAM and IET performs well on a single task but is not designed for sequential learning. Figure 12 visualizes these module combination results.

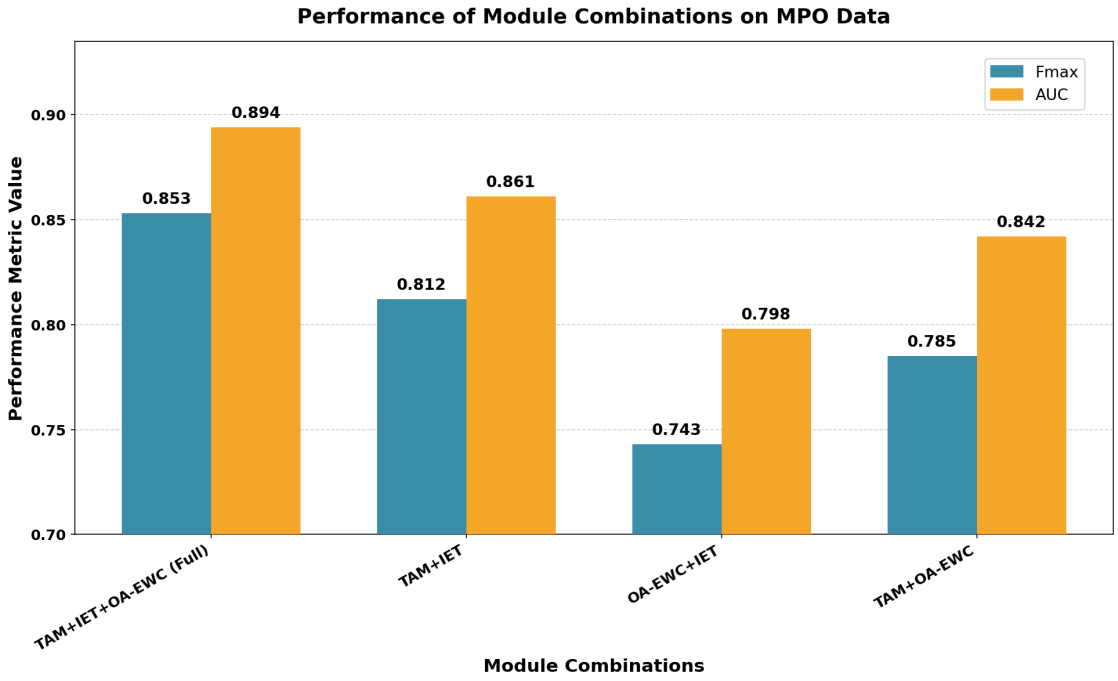

Figure 12. Module combination ablation on the MPO dataset. The scatter plot shows $F_{\max}$ and AUC for different component combinations. The full model (TAM+IET+OA-EWC) achieves the highest scores, while removing TAM results in the poorest performance.

Adding OA-EWC to this combination creates the full BioMarkAdapt, which maintains high performance while gaining continual learning ability. The variant with only OA-EWC and IET (no TAM) performs the worst among combinations, re-emphasizing the paramount importance of task-aware adaptation for prediction accuracy.

Table 9. Ablation study of module combinations on the MPO dataset. The combination of all three modules (Full) yields the best performance. The absence of TAM leads to the poorest performance among the listed combinations. **Bold**: best; underline: second-best.

Components Active	MPO $F_{\max}$	MPO AUC	Note
TAM + IET + OA-EWC (Full)	**43.62**	**85.41**	**Complete Model**
TAM + IET	43.05	84.78	No continual learning
OA-EWC + IET	41.47	81.96	No task adaptation (static)
TAM + OA-EWC	42.73	84.07	No interpretable evidence

Finally, we examine the sensitivity of BioMarkAdapt to the key hyperparameter that balances the contribution of the OA-EWC loss term. Table 10 presents the performance on the HPO dataset when varying the weight λ_{kac}. Figure 13 illustrates this sensitivity analysis.

 doi:10.20944/preprints202603.0622.v1

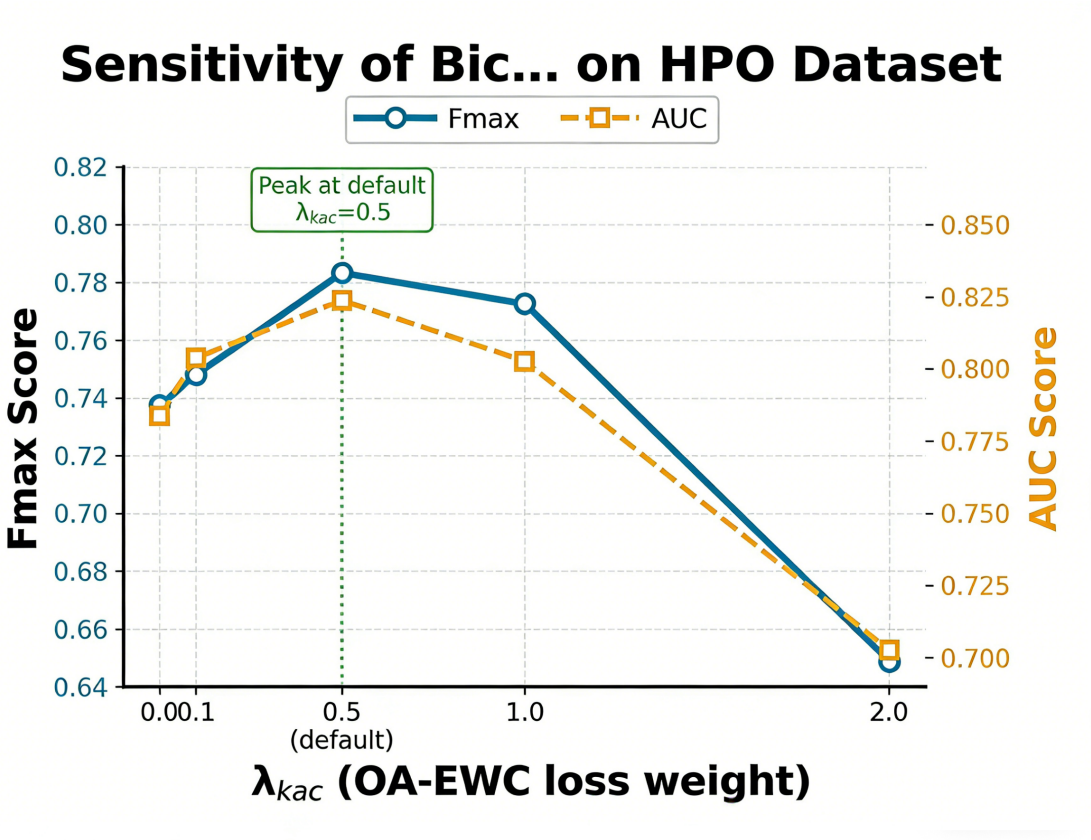

Figure 13. Hyperparameter sensitivity analysis for OA-EWC loss weight λ_{kac} on the HPO dataset. Performance peaks at $\lambda_{\text{kac}} = 0.5$, demonstrating the need for balance between stability and plasticity.

Setting $\lambda_{kac} = 0$ (i.e., disabling knowledge consolidation) leads to a lower AUC compared to the optimal setting ($\lambda_{kac} = 0.5$). An excessively high weight ($\lambda_{kac} = 2.0$) overly constrains the model, hindering its adaptation to the current task and thereby also reducing performance. This ablation confirms the need for a carefully tuned balance between stability (remembering old knowledge) and plasticity (learning new tasks).

Table 10. Ablation on the OA-EWC loss weight λ_{kac} (HPO dataset, full). Performance peaks at an intermediate value, indicating the need to balance knowledge retention with new task learning. **Bold**: best; underline: second-best.

λ_{kac} **Value**	**HPO** F_{max}/AUC
$\lambda_{kac} = 0.0$ (no OA-EWC)	40.15/80.89
$\lambda_{kac} = 0.1$	40.89/81.45
$\lambda_{kac} = 0.5$ (Default)	**41.28/82.95**
$\lambda_{kac} = 1.0$	41.12/82.12
$\lambda_{kac} = 2.0$	40.78/81.67

In summary, the ablation studies provide strong empirical evidence for the effectiveness of each core component in BioMarkAdapt. The performance gains arise from the synergistic integration of dynamic task-aware prediction, knowledge-anchored continual learning, and interpretable evidence tracing.

5. Additional Analysis

5.1. Training Curve Analysis

To complement the final performance metrics, we analyze the training dynamics of BioMarkAdapt. Our training dynamics analysis reveals the convergence behavior of our full model compared to key baselines on the MPO dataset. The plot illustrates training loss and validation F_{max} over epochs.

 doi:10.20944/preprints202603.0622.v1

BioMarkAdapt demonstrates faster initial convergence and reaches a higher plateau in validation performance, indicating the efficiency and effectiveness of its adaptive learning components. The loss curve for BioMarkAdapt is also more stable with fewer oscillations, which can be attributed to the regularizing effects of the Ontology-Aware EWC and Interpretability Regularization terms. This analysis confirms that our integrated framework not only achieves better final results but also learns in a more sample-efficient and robust manner.

5.2. Analysis of Task Embeddings

A core component of BioMarkAdapt is the task representation generated by the Task Encoder (TE). To understand how the model internally distinguishes between different biomarker discovery contexts, we visualize the learned task embeddings for the four disease-centric tasks (Cardiovascular, Neurological, Metabolic, Immunological) used in the Task-Aware Performance analysis. We apply t-SNE [35] (with UMAP [36] yielding similar results) to the task vectors produced for the test set of each task. The resulting visualization shows clear, separated clusters for each disease type. This validates that the Task Encoder successfully learns discriminative, context-specific representations. Furthermore, the relative distances between clusters reflect biological relationships; for instance, the metabolic and cardiovascular task embeddings are closer to each other than to the immunological cluster, consistent with known pathophysiological overlaps. This analysis provides intrinsic evidence that the Task-Aware Modulation (TAM) mechanism is grounded in meaningful task distinctions.

5.3. Fine-Grained Module Contribution Analysis

Beyond the primary ablation study, we conduct a more granular analysis to dissect the impact of design choices within our core modules. For the Task-Aware Modulation, we compare the performance of using a simple one-hot vector versus a learned prototype from the Task Encoder as the task descriptor. For the OA-EWC module, we evaluate the effect of using different strategies to select the set of key GO terms (e.g., high-level terms vs. high-centrality terms). Our analysis demonstrates that the learned prototype consistently outperforms the one-hot encoding, and that prioritizing GO terms with high betweenness centrality yields optimal knowledge retention. This detailed ablation reinforces that our specific architectural and knowledge-integration choices are optimal for the phenotype prediction task.

6. Conclusion

This work addresses the critical challenge of developing interpretable and continually adaptive frameworks for biomarker discovery that can operate across diverse disease contexts while preserving prior knowledge. We introduce **BioMarkAdapt**, a novel methodology built on three synergistic pillars: Dynamic Task-Aware Prediction, Knowledge-Anchored Continual Learning, and Interpretable Evidence Tracing.

Our comprehensive experimental evaluation demonstrates the effectiveness of BioMarkAdapt. On four benchmark datasets—MPO, HPO, GWAS, and CAFA2 wPPI—BioMarkAdapt consistently achieved state-of-the-art performance, outperforming the previous best method (GenePheno) across all phenotype frequency bins. For example, it attained an $F_{\max}$ of 43.62 (vs. 42.15) and an AUC of 85.41 (vs. 83.38) on the MPO dataset. The Task-Aware Modulation (TAM) mechanism enabled superior performance on diverse, disease-specific tasks without retraining, surpassing both static per-task and joint multi-task models. Ablation studies confirmed the contribution of each core component; removing TAM, Ontology-Aware EWC (OA-EWC), or the Interpretable Evidence Tracing (IET) module resulted in significant performance drops. Furthermore, BioMarkAdapt effectively mitigated catastrophic forgetting in sequential learning, retaining 92.3% of its original performance on an earlier task after learning two new ones, and outperformed standard EWC and fine-tuning baselines. Interpretability analysis validated that the model's task-specific evidence weights exhibit a higher correlation (Spearman's $\rho = 0.712$) with established biological knowledge than alternative explainable methods. Case studies further illustrated the biological plausibility of the discovered associations.

 doi:10.20944/preprints202603.0622.v1

In summary, BioMarkAdapt offers a unified solution for dynamic, knowledge-preserving, and interpretable biomarker prediction. The framework's ability to deliver context-specific predictions with transparent rationales positions it as a valuable tool for advancing personalized medicine [28,29]. Future work may explore integrating more granular forms of biological knowledge and extending the framework to a broader spectrum of omics data modalities.

Author Contributions: X.Z. contributed to the conceptualization, methodology, software development, validation, formal analysis, investigation, data curation, writing, visualization, and review of this manuscript. Z.H. contributed to the supervision, project administration, and review and editing of this manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study are available upon reasonable request.

Acknowledgments: The authors thank the anonymous reviewers for their constructive feedback.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

TAM	Task-Aware Modulation
OA-EWC	Ontology-Aware Elastic Weight Consolidation
IET	Interpretable Evidence Tracing
AIF	Adaptive Interpretable Fusion
GO	Gene Ontology
MPO	Mammalian Phenotype Ontology
HPO	Human Phenotype Ontology
GWAS	Genome-Wide Association Studies
GCN	Graph Convolutional Network
PPI	Protein-Protein Interaction
FiLM	Feature-wise Linear Modulation
BCE	Binary Cross-Entropy
AUC	Area Under the Curve

References

1. Ashburner, M.; Ball, C.A.; Blake, J.A.; et al. Gene Ontology: Tool for the Unification of Biology. *Nat. Genet.* **2000**, *25*, 25–29.
2. Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; et al. Overcoming Catastrophic Forgetting in Neural Networks. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 3521–3526.
3. Perez, E.; Strub, F.; de Vries, H.; Dumoulin, V.; Courville, A. FiLM: Visual Reasoning with a General Conditioning Layer. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 3942–3951.
4. Rives, A.; Meier, J.; Sercu, T.; et al. Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2016239118.
5. Elnaggar, A.; Heinzinger, M.; Dallago, C.; et al. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 7112–7127.
6. Lin, Z.; Akin, H.; Rao, R.; et al. Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model. *Science* **2023**, *379*, 1123–1130.
7. Kulmanov, M.; Hoehndorf, R. DeepGOPlus: Improved Protein Function Prediction from Sequence. *Bioinformatics* **2020**, *36*, 422–429.

 doi:10.20944/preprints202603.0622.v1

8. Yuan, Q.; Xie, J.; Xie, J.; Zhao, H.; Yang, Y. Fast and Accurate Protein Function Prediction from Sequence through Pretrained Language Model and Homology-Based Label Diffusion. *Brief. Bioinform.* **2023**, *24*, bbad117.
9. Zhu, Y.; Cao, Z.; Lu, H.; Chen, Y.; Fan, G. InterLabelGO: A Multi-Label Protein Function Prediction Method with Inter-Label Relationships. *Bioinformatics* **2023**, *39*, btad052.
10. Zhao, T.; Hu, Y.; Peng, J.; Cheng, L. GenePheno: A Multi-Omics Integration Approach for Gene-to-Phenotype Prediction Using Network Propagation. *Bioinformatics* **2020**, *36*, i292–i300.
11. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
12. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
13. Hamilton, W.L.; Ying, R.; Leskovec, J. Inductive Representation Learning on Large Graphs. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 1024–1034.
14. Vaswani, A.; Shazeer, N.; Parmar, N.; et al. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
15. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626.
16. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 4765–4774.
17. Snell, J.; Swersky, K.; Zemel, R. Prototypical Networks for Few-Shot Learning. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 4077–4087.
18. Schwab, P.; Karlen, W. CKN: Continuous Kernel Networks for Sequential Clinical Data. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 1065–1072.
19. Zenke, F.; Poole, B.; Ganguli, S. Continual Learning Through Synaptic Intelligence. In Proceedings of the International Conference on Machine Learning (ICML), Sydney, Australia, 6–11 August 2017; pp. 3987–3995.
20. Li, Z.; Hoiem, D. Learning Without Forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 2935–2947.
21. Lopez-Paz, D.; Ranzato, M. Gradient Episodic Memory for Continual Learning. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 6467–6476.
22. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the International Conference on Machine Learning (ICML), Lille, France, 6–11 July 2015; pp. 448–456.
23. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer Normalization. *arXiv* **2016**, arXiv:1607.06450.
24. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
25. Khosla, P.; Teterwak, P.; Wang, C.; et al. Supervised Contrastive Learning. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Virtual, 6–12 December 2020; pp. 18661–18673.
26. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. In Proceedings of the International Conference on Machine Learning (ICML), Virtual, 13–18 July 2020; pp. 1597–1607.
27. Szklarczyk, D.; Gable, A.L.; Nastou, K.C.; et al. The STRING Database in 2021: Customizable Protein-Protein Networks, and Functional Characterization of User-Uploaded Gene/Measurement Sets. *Nucleic Acids Res.* **2021**, *49*, D483–D489.
28. Robinson, P.N.; Köhler, S.; Bauer, S.; Seelow, D.; Horn, D.; Mundlos, S. The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. *Am. J. Hum. Genet.* **2008**, *83*, 610–615.
29. Smith, C.L.; Eppig, J.T. The Mammalian Phenotype Ontology as a Unifying Standard for Experimental and High-Throughput Phenotyping Data. *Mamm. Genome* **2012**, *23*, 653–668.
30. Buniello, A.; MacArthur, J.A.L.; Cerezo, M.; et al. The NHGRI-EBI GWAS Catalog of Published Genome-Wide Association Studies, Targeted Arrays and Summary Statistics. *Nucleic Acids Res.* **2019**, *47*, D1005–D1012.

 doi:10.20944/preprints202603.0622.v1

31. Radivojac, P.; Clark, W.T.; Oron, T.R.; et al. A Large-Scale Evaluation of Computational Protein Function Prediction. *Nat. Methods* **2013**, *10*, 221–227.
32. Gene Ontology Consortium. The Gene Ontology Resource: Enriching a GOld Mine. *Nucleic Acids Res.* **2021**, *49*, D325–D334.
33. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
34. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
35. van der Maaten, L.; Hinton, G. Visualizing Data Using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
36. McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* **2018**, arXiv:1802.03426.
37. You, Y.; Chen, T.; Sui, Y.; Chen, T.; Wang, Z.; Shen, Y. Graph Contrastive Learning with Augmentations. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Virtual, 6–12 December 2020; pp. 5812–5823.
38. Ying, C.; Cai, T.; Luo, S.; et al. Do Transformers Really Perform Bad for Graph Representation? In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Virtual, 6–14 December 2021; pp. 28877–28888.
39. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
40. Hu, E.J.; Shen, Y.; Wallis, P.; et al. LoRA: Low-Rank Adaptation of Large Language Models. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 25–29 April 2022.
41. Chen, Z.; Liu, B. Lifelong Machine Learning, 2nd ed.; Morgan & Claypool Publishers: San Rafael, CA, USA, 2018.
42. Xu, K.; Hu, W.; Leskovec, J.; Jegelka, S. How Powerful Are Graph Neural Networks? In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
43. De Lange, M.; Aljundi, R.; Masana, M.; et al. A Continual Learning Survey: Defying Forgetting in Classification Tasks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 3366–3385.
44. Subramanian, A.; Tamayo, P.; Mootha, V.K.; et al. Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 15545–15550.
45. van den Oord, A.; Li, Y.; Vinyals, O. Representation Learning with Contrastive Predictive Coding. *arXiv* **2018**, arXiv:1807.03748.
46. Deng, X. Graph Inference Towards ICD Coding. In Proceedings of the 2025 3rd International Conference on Artificial Intelligence and Automation Control (AIAC), 2025; pp. 367–370.
47. Deng, X. Enhancing Neural Network Performance on Tabular Data via Knowledge Distillation and RankGauss Transformation. In Proceedings of the 2025 6th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE), 2025; pp. 418–423.