GRASP: Gene-Relation Adaptive Soft Prompt for Universal Gene Network Inference

Yuqiang Feng[1], Kaiwen Deng[2,*], Yuanfang Guan[2,3,*]

1. Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, USA
2. Department of Computational Medicine and Bioinformatics, Michigan Medicine, University of Michigan, Ann Arbor, USA
3. Department of Internal Medicine, University of Michigan, Ann Arbor, USA

* Co-corresponding authors: Yuanfang Guan (gyuanfan@umich.edu) and Kaiwen Deng (dengkw@umich.edu)

Abstract

Gene networks (GNs) represent the complex molecular interactions that regulate cellular states. Existing sequence and graph based deep learning approaches have advanced GN inference but remain constrained by their reliance on specialized data. Large language models (LLMs) offer a promising alternative by reframing structured biological problems as natural language inference tasks. However, a major bottleneck lies in prompt design, where fixed prompts or task-specific soft prompts cannot adapt to the heterogeneity of gene-gene relationships. Here we introduce GRASP (Gene-Relation Adaptive Soft Prompt), a framework that performs universal GN inference directly from gene symbols. GRASP generates pair-specific adaptive tokens by encoding gene-level representations and relational contrasts into latent vectors, enabling context-aware inference while preserving parameter efficiency. Across large-scale protein-protein interaction (PPI) benchmarks, GRASP consistently outperforms other baselines, achieving ROC-AUC values up to 0.937. GRASP further demonstrates strong cross-species transferability, robust performance in phosphorylation network inference, and discovery potential by recovering previously mislabeled gene interactions.

Introduction

Cells are governed by complex molecular interactions that collectively dictate cellular states. These interactions can be effectively abstracted and represented as gene networks (GNs), which are graph-based models linking genes that coordinately regulate cellular functions[1]. Understanding how genes interact in networks is fundamental to deciphering complex biological processes and disease mechanisms[2–5]. GNs include different forms, such as protein-protein interaction (PPI) networks[6], which encodes the physical contacts among proteins and underpins the structural and functional connectome of the cell. Other critical subtypes include phosphorylation networks[7], wherein edges represent the enzymatic modification of substrate proteins by kinases.

Numerous computational methods have been developed to infer GNs across diverse biological data types. Some deep learning approaches infer interaction propensities directly from amino-acid sequences[8–11], while others exploit network topology and graph learning to capture higher-order relational patterns[12–15]. For example, the recently developed TUnA[16] model leverages ESM-2 protein language model embeddings and a Transformer-Gaussian Process hybrid to predict PPIs from sequence data. And graph-based approaches like KINference[17] integrate phosphoproteomics data with network filters such as co-phosphorylation correlations to infer condition-specific kinase-substrate subnetworks.

Recently, converting knowledge graph data into natural language prompts and leveraging large language models (LLMs) for link prediction has outperformed traditional graph neural network (GNN) approaches[18,19]. Within the biological field, methods such as ProLLM[20] replicates the biological mechanism of signaling pathways as natural language descriptions for PPI prediction. Despite these

advances, natural language based GN inference remains underexplored relative to sequence and graph based methods.

A major bottleneck lies in prompt design. Fixed prompts constructed from gene symbols or augmented with short gene summaries are simple but static, limiting adaptability to diverse relational contexts. Soft prompt tuning provides a more flexible alternative by introducing learnable virtual tokens that guide inference during fine-tuning. But the conventional task-specific soft prompt assumes a single shared embedding for all gene pairs. While effective in many NLP tasks, this approach struggles with the heterogeneity of gene pair semantics.

To address this limitation, we introduce GRASP (Gene-Relation Adaptive Soft Prompt), a framework that generates pair-specific prompts by encoding each gene and their relational contrast into latent vectors. Through a factorized process combining gene-specific signals with shared prototypes, GRASP produces three adaptive tokens per pair, which are appended to the input prompt to enable context-aware inference directly from gene symbols. Across large-scale protein-protein interaction benchmarks, cross-species transfer settings, phosphorylation networks, and gene interaction discovery, GRASP consistently outperforms other methods. Remarkably, GRASP even surpasses fixed prompts augmented with real biological summaries, demonstrating its strong ability to capture relational semantics. These results establish GRASP as a general and scalable framework for natural language based GN inference.

Results

Domain adaptation and prompting methods for gene network inference

For downstream analysis, we selected two parameter-efficient large language models (LLMs), Gemma-3-4B-IT[21] and Llama-3.1-8B-Instruct[22], as base models. Each model was continually pretrained[23] on a domain-specific corpus consisting of titles and abstracts from 6.3 million gene-related PubMed articles (Fig. 1A). This adaptation phase was designed to enhance the LLMs' capacity to capture gene relationships and relevant biological context. As a result, we obtained both the original and domain-adapted versions of each LLM.

We first established two baseline methods for gene network inference: few-shot inference and fixed prompt (Fig. 1B). Few-shot inference serves as a minimal-effort comparison, leveraging the in-context learning ability of LLMs with a few examples but without any fine-tuning or architectural changes. In contrast, fixed-prompt methods adopted from a fine-tuning framework by introducing a lightweight classification head atop the frozen LLM backbone, outputting interaction probabilities. Each input query consists of a gene pair and its binary label. Two prompt styles were tested: a symbol-only template and a biocontext-informed template incorporating short database-derived gene descriptions, allowing us to assess the benefit of external biological context.

The static nature of fixed prompts restricts the model's adaptability across diverse contexts. Thus we investigated soft prompt-based strategies that optimize prompts during fine-tuning (Fig. 1C). Unlike discrete textual prompts, soft prompts consist of learnable virtual tokens concatenated with the input embeddings. These tokens are updated during fine-tuning and act as continuous context representations that guide model behavior. The task-specific soft prompt follows a conventional design in which a single shared embedding is learned for each task[24]. In gene network inference, this embedding is applied uniformly across all gene pairs. While this approach has proven effective in many natural language processing tasks, a single, shared soft prompt may be insufficient to capture the diverse relational patterns across different gene pairs.

To address this limitation, we propose Gene-Relation Adaptive Soft Prompt (GRASP), a dynamic prompting framework that conditions the soft prompt on the input gene pair. Inspired by chain-of-thought reasoning[25], GRASP first guides the LLM to address subquestions by generating summary information

for each gene. Rather than appending textual summaries to the input prompt, we encode them into latent gene representations. For each gene pair, we compute two gene vectors (one per gene) and an additional difference vector that captures their relational contrast. These three vectors form the basis for generating adaptive soft prompts. To construct the final soft prompt embeddings, we factorize the prompt generation process into two components: a gene coefficient matrix and a gene prototype matrix. The gene coefficient matrix captures the private, gene-specific information and is computed directly from the gene vectors via a linear projection. The gene prototype matrix serves as a shared information base. It is constructed as a weighted combination of a set of learnable matrices. This design promotes parameter sharing while preserving gene-specific information. Each of the three vectors is used to generate a distinct type of soft prompt embedding. In practice, we set the token length of each embedding to one, resulting in three soft prompt tokens per instance, which are appended as a suffix to the input prompt.

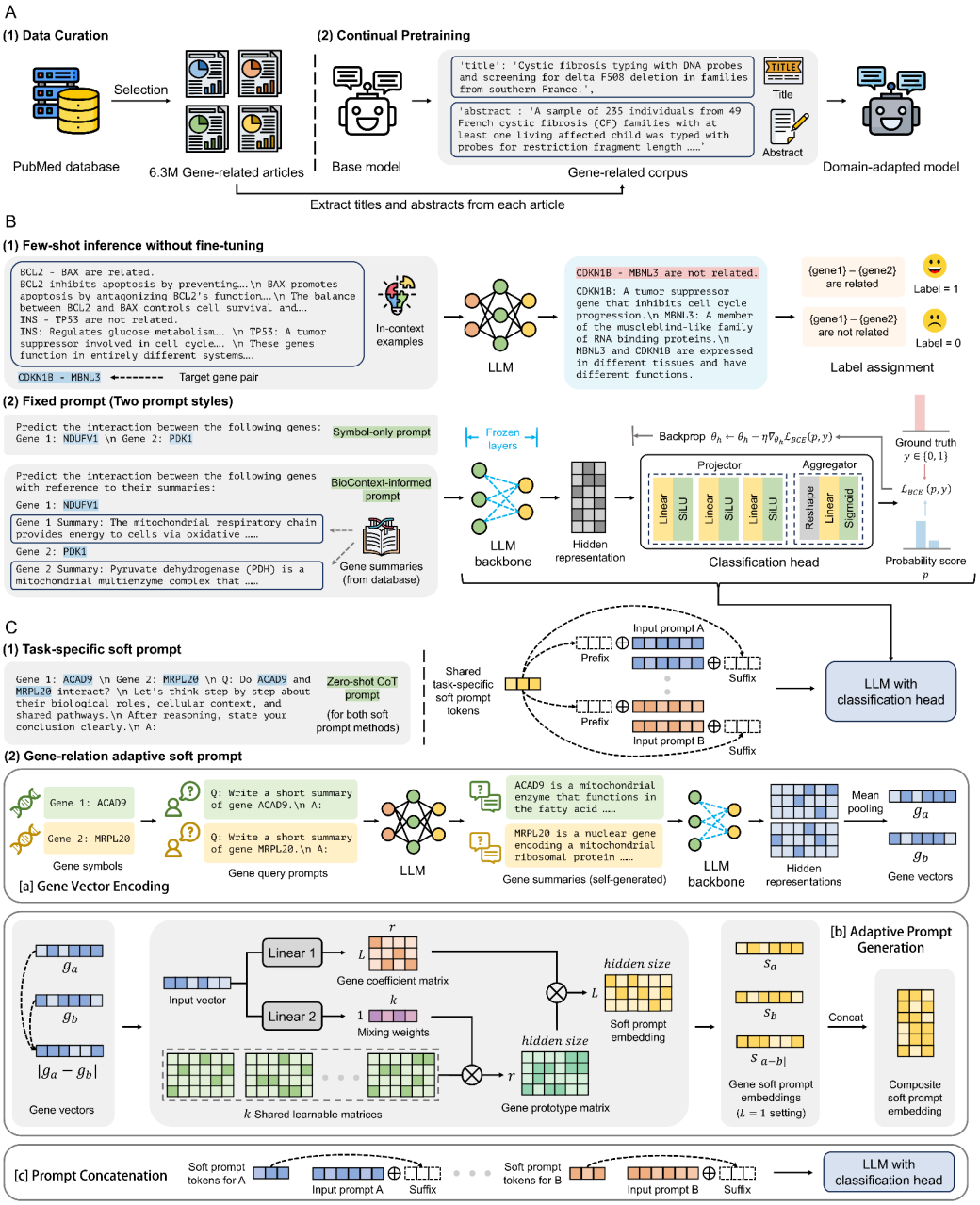

Figure 1. Framework overview. (A) Domain-specific continual pretraining pipeline. We collected approximately 6.3 million gene-related articles from the PubMed database. Titles and abstracts were extracted from each article to create a domain-specific corpus for continual pretraining. **(B)** Baseline methods. Few-shot inference uses in-context examples without model fine-tuning. Fixed prompt applies either symbol-only or biocontext-informed templates, followed by a classification head that predicts gene interactions from frozen LLM representations. **(C)** Soft prompt-based methods. Task-specific soft prompt learns a fixed set of virtual tokens shared across gene pairs. GRASP generates context-aware prompts: gene summaries are first produced to build latent vectors, which are then transformed into adaptive soft prompt embeddings via coefficient-prototype factorization. The resulting tokens are appended to the input before classification.

GRASP improves large-scale human PPI inference

We first evaluated GRASP on a large-scale benchmark of human protein–protein interactions (PPIs) comprising 2,127,124 gene pairs, balanced between experimentally validated physical interactions and randomly sampled non-interacting pairs. Physical PPIs were obtained from curated resources including BioGRID[26], MIPS[27], IntAct[28], and STRING[29], representing the most reliable subset of gene interactions supported by experimentally confirmed protein–protein contacts. We fine-tuned models on this dataset using a standard train-validation-test split, comprising 1,361,359 training pairs (64.0%), 340,340 validation pairs (16.0%), and 425,425 test pairs (20.0%). Each experiment was repeated five times with different random splits, and mean ± SD values were reported. Performance was assessed on 50,000 pairs randomly selected from the test set using precision, recall, accuracy, and ROC-AUC to capture both overall accuracy and sensitivity–specificity trade-offs.

Across all settings, GRASP consistently achieved the best precision–recall balance (Fig. 2A). Fixed prompts modestly outperformed few-shot inference, while task-specific soft prompts slightly improved performance but showed higher variance. GRASP consistently occupied the upper-right region of the precision–recall space, with domain adaptation further enhancing its advantage. ROC-AUC analyses (Fig. 2B) corroborated these findings: GRASP achieved 0.923 ± 0.005 and 0.931 ± 0.004 with Gemma 4B, and 0.925 ± 0.003 and 0.937 ± 0.002 with Llama 8B before and after domain adaptation, surpassing all other methods. Accuracy results (Supplementary Fig. 1a) further confirmed GRASP's robustness and scalability in human PPI inference. These results surpassed all other prompting methods, demonstrating both robustness and scalability of relation-adaptive soft prompting

We next examined per-gene error rate distributions in the test set to assess whether GRASP's improvements were uniform across the network or concentrated in specific regions. The analysis showed that GRASP achieved the lowest error rates for high-degree genes and exhibited even greater relative improvements among moderately connected genes (Fig. 2C, Supplementary Fig. 2). Representative cases such as COG4, SLC22A9, ZIC1, and CDCA7L illustrate GRASP's ability to correct systematic misclassifications observed with other prompting methods. In the COG4 subnetwork, GRASP recovered known retrograde transport partners (NPY2R, BEX2) while reducing false positives; in the ZIC1 subnetwork, it retained transcriptional regulators (RNASE7, OXER1) missed by task-specific soft prompts (Fig. 4D). Comparative subnetwork analyses across all prompting methods (Supplementary Fig. 3) reinforced these findings, demonstrating that GRASP not only improved global metrics but also reconstructed biologically coherent local structures.

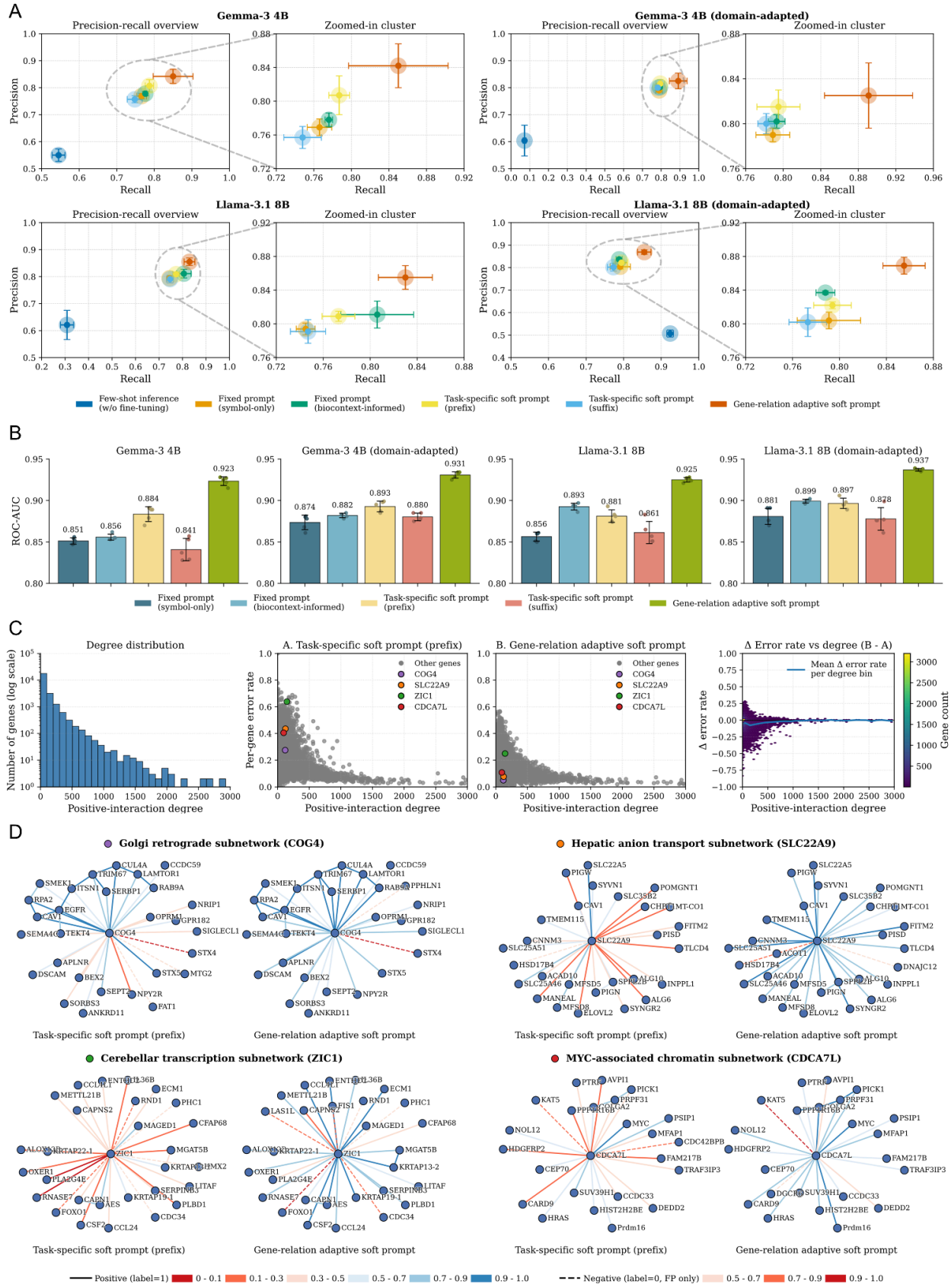
A
Gemma-3 4B
Precision-recall overview
Zoomed-in cluster
Gemma-3 4B (domain-adapted)
Llama-3.1 8B
Llama-3.1 8B (domain-adapted)
Precision
Recall
Few-shot inference (w/o fine-tuning)
Fixed prompt (symbol-only)
Fixed prompt (biocontext-informed)
Task-specific soft prompt (prefix)
Task-specific soft prompt (suffix)
Gene-relation adaptive soft prompt
B
ROC-AUC
0.851
0.856
0.884
0.841
0.923
0.874
0.882
0.893
0.880
0.931
0.856
0.893
0.881
0.861
0.925
0.881
0.899
0.897
0.878
0.937
C
Degree distribution
Number of genes (log scale)
Positive-interaction degree
A. Task-specific soft prompt (prefix)
B. Gene-relation adaptive soft prompt
Other genes
COG4
SLC22A9
ZIC1
CDCA7L
Per-gene error rate
Δ Error rate vs degree (B - A)
Mean Δ error rate per degree bin
Δ error rate
Gene count
D
Golgi retrograde subnetwork (COG4)
Hepatic anion transport subnetwork (SLC22A9)
Cerebellar transcription subnetwork (ZIC1)
MYC-associated chromatin subnetwork (CDCA7L)
Task-specific soft prompt (prefix)
Gene-relation adaptive soft prompt
Positive (label=1)
0 - 0.1
0.1 - 0.3
0.3 - 0.5
0.5 - 0.7
0.7 - 0.9
0.9 - 1.0
Negative (label=0, FP only)

Figure 2. Comparative performance of prompting methods on the human PPI dataset. (A) Precision-recall analyses for Gemma-3 4B and Llama-3.1 8B models, with and without domain adaptation. Each panel shows an overview plot (right) and a zoomed-in view of high-performing clusters (right). **(B)** ROC-AUC scores for each prompting method. **(C)** Per-gene error rate analyses. Left: degree distribution of positive interactions (log scale). Middle: scatter plots of per-gene error rates versus positive-interaction degree under task-specific soft prompt (prefix) and GRASP. Highlighted genes illustrate differences in error patterns. Right: Δ error rates between GRASP and task-specific soft prompt (prefix), plotted against positive-interaction degree, with negative values indicating GRASP improvements. The blue line shows the mean Δ error rate per bin. **(D)** Subnetwork visualizations of representative genes. Networks are centered on COG4, SLC22A9, ZIC1, and CDCA7L, under task-specific soft prompting (prefix) versus GRASP. Edge colors represent prediction confidence: solid blue, true positives (TPs); solid red, false negatives (FNs); dashed red, false positives (FPs).

GRASP enhances cross-species transferability of PPI inference

To assess model transferability across species, we directly applied models fine-tuned on the human PPI dataset to chicken, cow, and dog PPI datasets without additional adaptation. This cross-species setting provides a stringent test of how well prompting strategies generalize beyond the training domain. The chicken, cow and dog datasets were derived from experimentally validated PPIs in the Integrated Interactions Database (IID)[30], with all pairs overlapping with the human dataset removed. Each dataset was balanced by adding an equal number of randomly sampled non-interacting pairs, resulting in 153,252 pairs for chicken, 165,574 for cow, and 149,960 for dog. For each species, 50,000 pairs were randomly sampled for evaluation.

In the precision–recall analysis (Fig. 3A), transfer across species proved challenging, with data points tightly clustered. To emphasize small performance differences, we overlaid F1-score contours at 0.02 intervals, increasing from the lower left to the upper right; methods positioned further toward the upper-right region therefore achieve higher F1 values. GRASP generally attained higher precision but slightly lower recall compared with other methods. On the dog PPI dataset using Gemma models, GRASP achieved the best overall balance, while in other cases its performance was marginally lower but remained within a single 0.02 F1 contour.

ROC-AUC analysis (Fig. 3B), which provides a threshold-independent measure of performance, revealed clearer separation. GRASP achieved the highest AUC in most conditions, with only two exceptions, chicken with Gemma-3 4B (domain-adapted) and cow with Llama-3.1-8B (domain-adapted), where its performance was slightly lower but still comparable. Overall, these results demonstrate that GRASP confers strong cross-species generalization, with near-equivalent F1 scores and consistently superior ROC-AUC in the majority of transfer scenarios.

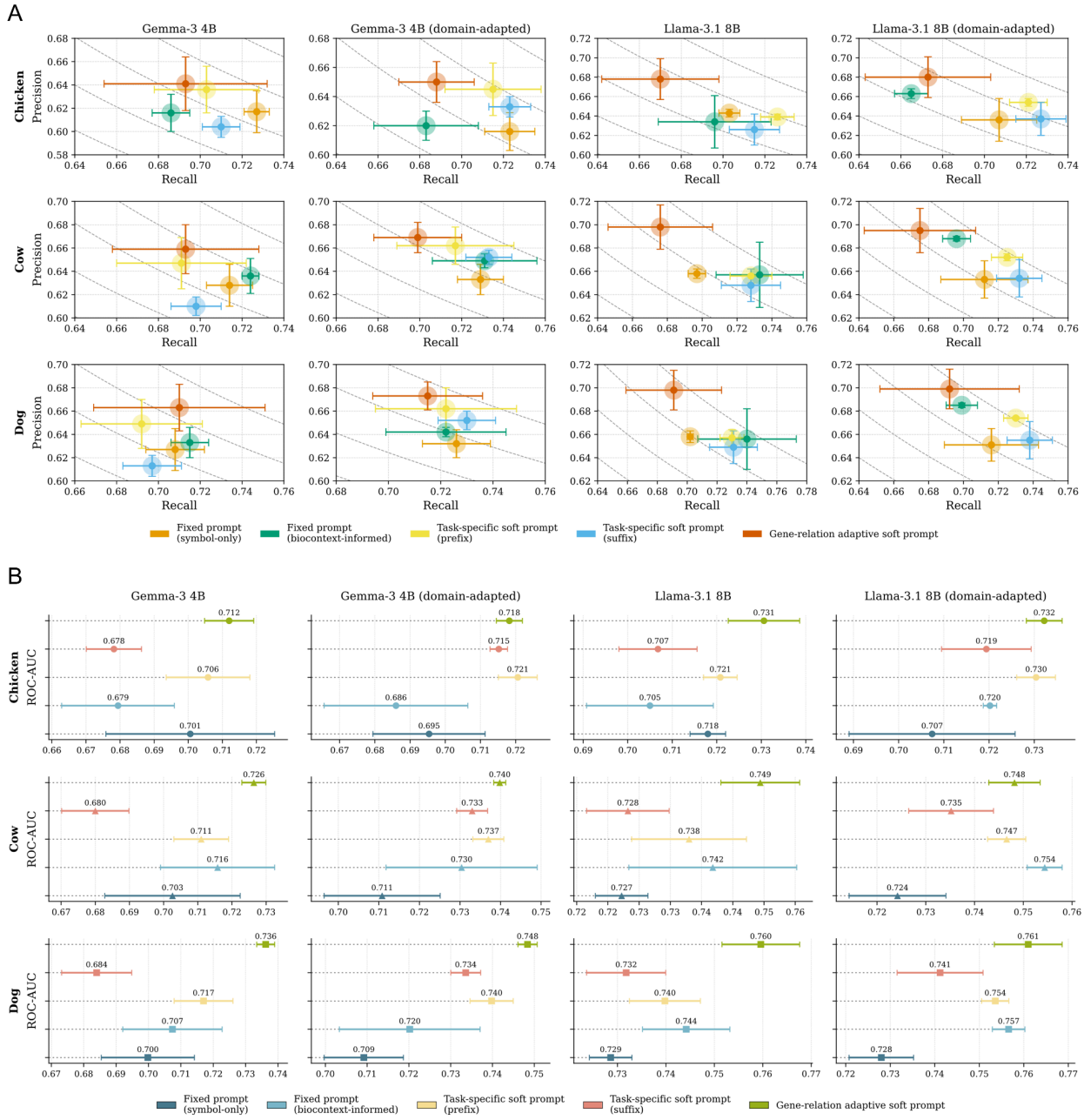

Figure 3. Cross-species transfer performance of PPI network inference. (A) Precision-recall analysis for chicken, cow and dog PPI networks using models fine-tuned on human PPIs. Contours indicate F1 score differences of 0.02 to highlight separation among closely clustered data points. **(B)** ROC-AUC results for the same transfer setting.

GRASP achieves top performance in phosphorylation network inference

Building on the evaluation of protein-protein interactions (PPIs), we next assessed whether GRASP can generalize to other types of gene interactions. To this end, we applied the framework to phosphorylation networks, focusing on kinase-substrate relationships. We used the PhosphoNetworks[31] database, which provides a high-resolution map of phosphorylation signaling pathways, and constructed a balanced dataset of 8,750 kinase-substrate pairs. The dataset was partitioned into training (5,600 samples), validation (1,400 samples) and test (1,750 samples) subsets. The training settings and evaluation metrics were kept consistent with those used in the PPI experiments.

Consistent with the PPI findings, GRASP generalized to kinase-substrate prediction and dominated the high-precision, high-recall regime (Fig. 4A). In the overview precision-recall plots, few-shot inference clustered well below the high-performing region, whereas all fine-tuned methods concentrated near the upper-right corner. In the zoom-in panels, GRASP occupied the frontier for all settings, while fixed prompts and task-specific soft-prompts trailed by several points of precision at comparable recall. Domain adaptation shifted all methods up and to the right, with the largest gain visible for GRASP.

ROC-AUC analyses confirmed these trends (Fig. 4B). Across models, GRASP achieved the highest AUC, with domain adaptation delivering additional improvements (Gemma 4B: 0.953(±0.009) - 0.963(±0.002); Llama 8B: 0.960(±0.005) - 0.965(±0.003)). Accuracy results (Supplementary Fig. 1C) further supported these findings, showing that GRASP consistently maintained the highest accuracy across all prompting strategies. Interestingly, enriching fixed prompts with biological summaries even lowered performance, suggesting that additional textual context can introduce noise rather than aid generalization. Moreover, task-specific soft prompts performed better in the prefix than the suffix position, highlighting the sensitivity of prompt tuning to design choices.

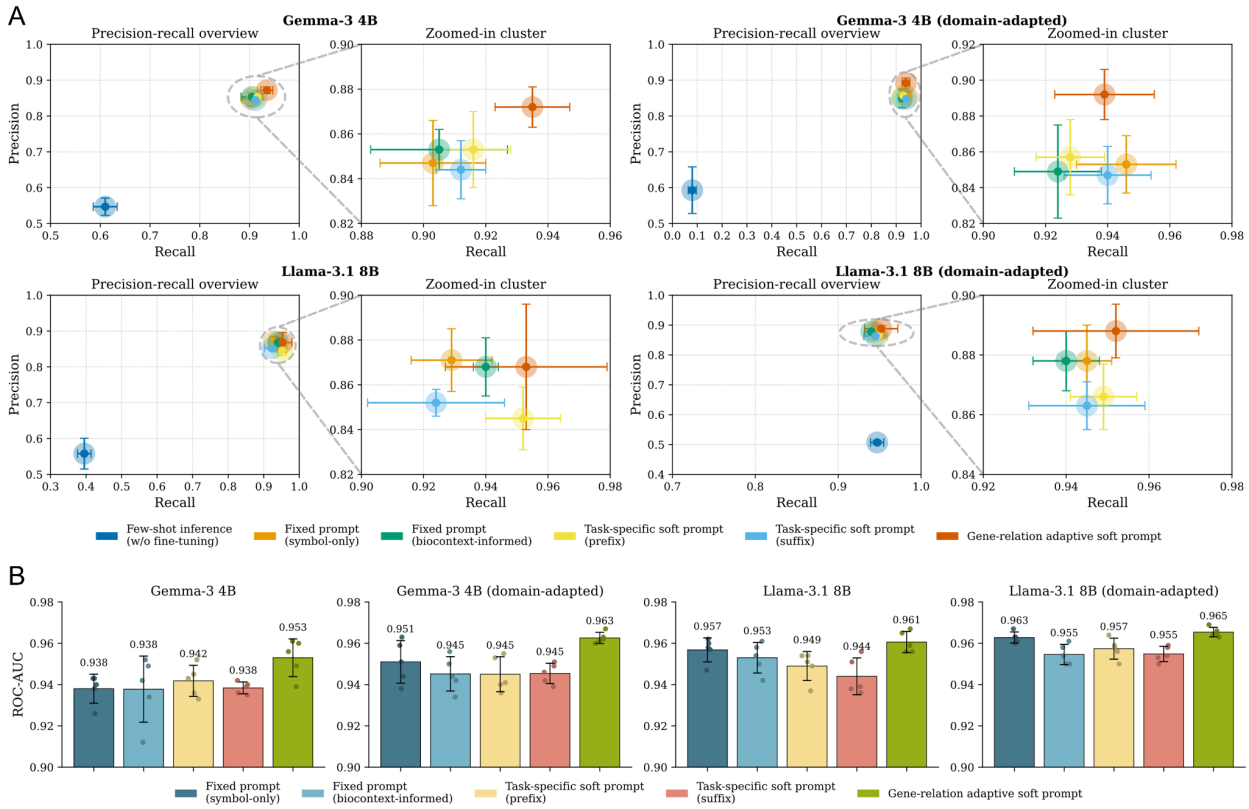

Figure 4. Comparative performance of prompting strategies for kinase-substrate relationship inference. (A) Precision-recall analyses for Gemma-3 4B and Llama-3.1 8B models before and after domain adaptation. Each panel shows an overview plot (left) and a zoomed-in view of high-performing clusters (right). **(B)** ROC-AUC for the same prompting strategies.

GRASP facilitates discovery of previously mislabeled gene interactions

Although our human PPI benchmark incorporated over one million experimentally validated pairs from BioGRID[26], MIPS[27], IntAct[28], and STRING[29], the negative set was constructed synthetically through random sampling. Inevitably, some of these synthetic negatives correspond to real but unrecorded interactions, introducing mislabeled positives. A robust inference method should therefore exhibit the capacity to recover such hidden interactions, reflecting discovery potential beyond the training corpus.

To assess this property, we incorporated an external source from the Integrated Interactions Database (IID)[30], which contains 561,029 experimentally verified human PPIs across diverse studies. We cross-referenced IID against our synthetic negative set of 1,063,562 pairs and identified 164 pairs that had been mislabeled as negatives. These mislabeled positives were not restricted to the test partition, as their presence in the training and validation sets also provides an opportunity to evaluate whether models can generalize to interactions not annotated in the benchmark databases. We then compared model predictions on the 164 mislabeled positives against a reference cohort of 10,000 true negative test pairs, ensuring no overlap with the mislabeled set. To quantify discovery ability, we analyzed the score distributions produced by each prompting method using both the Wasserstein distance, which captures global distributional separation, and Cohen's d, an effect size metric reflecting standardized mean differences. Larger values for either measure indicate greater separation between mislabeled positives and true negatives, and thus stronger ability to recover novel interactions.

Across all settings, GRASP achieved the largest separation between mislabeled positives and true negatives, demonstrating a clear advantage in uncovering biologically relevant interactions absent from the benchmark databases (Fig. 5A). With both Gemma-3 4B and Llama-3.1 8B models, domain adaptation further amplified these gains, yielding a Wasserstein distance of 0.463(±0.024) and a Cohen's d of 1.880(±0.051). In contrast, fixed prompts and task-specific soft prompts produced substantially smaller separations, underscoring their limited ability to distinguish hidden positives from synthetic negatives.

To illustrate how such predictions may be interpreted in practice, we further queried the domain-adapted Llama-3.1 8B model with a zero-shot chain-of-thought style prompt (Fig. 5B). Given the candidate interaction between INSR and PTPRF, the model generated a reasoning-based explanation emphasizing their shared role in insulin signaling. Specifically, it identified INSR as the receptor for insulin and PTPRF as a phosphatase that modulates INSR activity and downstream signaling. This hypothesis is consistent with previous reports showing that PTPRF can dephosphorylate the insulin receptor and suppress insulin signaling in cell and animal models[32,33]. But we should also notice that while many generated hypotheses align with established biological knowledge, some aspects remain unverified and will require experimental validation. Additional representative examples of LLM-generated hypotheses for gene pairs are provided in Supplementary Table 1.

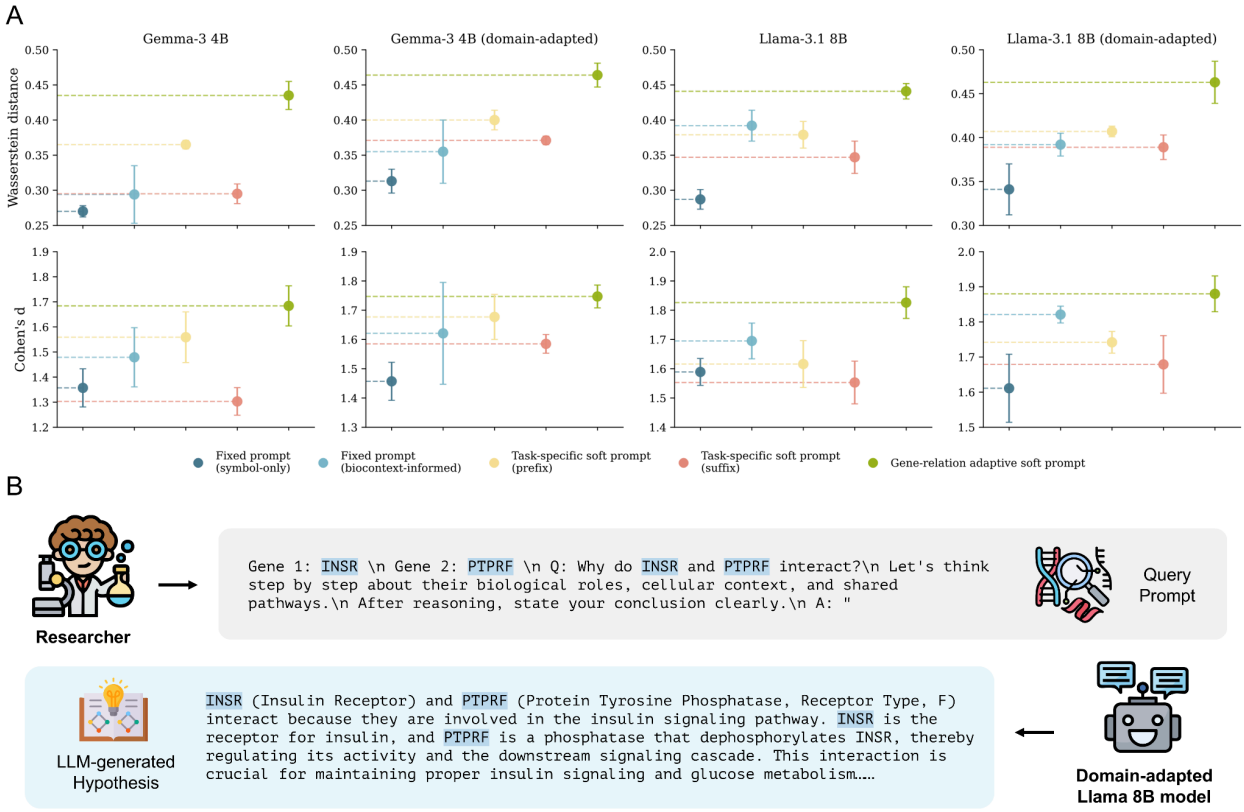

Figure 5. Discovery of previously mislabeled gene interactions. (A) Separation between mislabeled positives and true negatives measured by Wasserstein distance and Cohen's d. **(B)** Example of hypothesis generation for a previously mislabeled interaction using the domain-adapted Llama-3.1-8B model. Given a gene pair (INSR and PTPRF), the model receives a zero-shot chain-of-thought style query prompt and produces a reasoning-based explanation.

Discussion

In this work, we present GRASP, a gene-relation adaptive soft prompting framework that enables universal gene network inference directly from gene symbols. By generating pair-specific adaptive tokens conditioned on gene-level and relational representations, GRASP overcomes the rigidity of fixed prompts and the limitations of task-specific soft prompts. Across large-scale human protein-protein interaction (PPI) datasets, cross-species transfer tasks, and phosphorylation networks, GRASP consistently outperformed existing baselines. These results establish GRASP as a general and scalable approach for natural language based gene network inference.

A key advantage of GRASP is its ability to adaptively represent the heterogeneity of gene-gene relationships. Task-specific soft prompts rely on a single shared embedding across all input pairs, which often fails to capture diverse biological contexts. GRASP decomposes pairwise context into gene-specific and relational components, producing compact adaptive tokens that modulate inference in a structured manner. This design improved predictive accuracy and enhanced interpretability, as shown by reduced per-gene error rates and more biologically coherent subnetworks. GRASP also maintains parameter efficiency by freezing the pretrained backbone, demonstrating that relation-aware prompting can extract latent biological knowledge from large language models without requiring full fine-tuning.

Our evaluation shows that GRASP is robust and transferable. The framework generalized effectively across species despite differences in interaction landscapes, highlighting its ability to capture conserved relational semantics. GRASP also extended beyond PPIs to phosphorylation networks, outperforming fix prompts that incorporated external biological summaries. This suggests that dynamic relational encoding is more effective than static context for gene network inference. In addition, GRASP demonstrated discovery potential by recovering mislabeled positives from true negatives, illustrating its potential as a discovery tool that can complement experimental studies.

There are still limitations to address. Our analysis was based on curated binary interaction datasets, which do not reflect dynamic regulation over time or condition-specific dependencies of gene interactions. Extending GRASP to integrate tissue, disease, or temporal context could enable finer-grained network reconstruction. Domain adaptation improved results, but the continual pretraining corpus was restricted to abstracts and titles. Incorporating full texts, experimental protocols, or multi-omic metadata could further strengthen the biological grounding of adaptive prompts.

Future work could expand GRASP along several directions. The factorized design of adaptive prompts can integrate embeddings from other molecular modalities, such as gene expression profiles or structural annotations. Beyond PPIs and phosphorylation networks, GRASP could be applied to other regulatory, metabolic, and signaling networks, providing a unified approach to molecular interaction inference. By enabling pair-specific reasoning in a parameter-efficient manner, GRASP represents a step toward bridging natural language modeling with the complexity of cellular systems.

Methods

Human protein-protein interaction network dataset construction

We collected experimentally validated physical protein-protein interaction (PPI) pairs for human and mouse from four curated databases: BioGRID[26], MIPS[27], IntAct[28], and STRING[29]. In total, we obtained 1,063,562 human gene pairs. These confirmed interaction pairs are denoted as positive pairs. To focus on symbol-level inference, we retained only the gene symbols from the original records and discarded all other associated data.

To construct a balanced dataset, we generate synthetic negative gene interaction pairs that are not present in any of the curated physical PPI sources. Our task formulation assumes undirected interactions, where the existence of a relationship is independent of gene order. We first extract all unique gene symbols from the positive pairs and generate all possible undirected gene-gene combinations among them. Negative pairs are randomly sampled from these combinations, excluding any that overlap with the positive set. This sampling process continues until the number of negative samples matches the number of positive ones. Following this procedure, we retrieved a total of 2,127,124 human gene pairs for downstream model fine-tuning and evaluation, comprising the same number of positive and negative examples.

Cross-species protein-protein interaction network dataset construction

We constructed cross-species datasets for chicken, cow and dog by collecting physical PPI pairs from the Integrated Interactions Database (IID)[30]. IID integrates multiple types of evidence, including experimental detection, orthology, and machine learning predictions. To ensure high-quality data, we retained only experimentally verified PPIs and discarded all pairs supported exclusively by orthology or computational predictions. We further removed any pairs overlapping with the human PPI dataset used in earlier experiments. This filtering resulted in 76,626 positive pairs for chicken, 82,787 for cow and 74,980 for dog. To create balanced datasets, we generated an equal number of negative pairs by randomly sampling gene-gene combinations not present in the positive set. To create balanced datasets, we generated an equal number of negative pairs by randomly sampling gene-gene combinations not present

in the positive set, yielding final datasets of 153,252 pairs for chicken, 165,574 for cow and 149,960 for dog, consistent with the procedure used for the human PPI dataset.

Phosphorylation network dataset construction

The phosphorylation relationships among the genes were extracted from a kinase-substrate relationship (KSR) study[31]. We treated the gene pairs in the combined KSR dataset (comKSR) as the positive interaction samples, including 3,656 refined KSRs (refKSR) and 744 literature-curated KSRs. An equal number of negative pairs (4,375) was generated using the same sampling procedure described above for the human PPI dataset. To further ensure the validity of the negative set, any gene pairs that appeared in the raw KSRs (24,046 KSRs identified through protein microarray assays) were excluded. As a result, the final kinase-substrate interaction dataset comprises 8,750 gene pairs, equally divided between positive and negative examples

Curation of gene summaries for BioContext-informed prompt

To construct the BioContext-informed prompts, we retrieved gene summaries from MyGene.info (https://mygene.info)[34]. We queried the gene symbols in bulk using the "querymany" function from the Mygene Python package, with the scopes parameter set to "symbol" and the fields parameter set to "summary". The API returned biological summaries describing gene functions, and we extracted only the first sentence of each gene's summary. We constructed BioContext-informed natural language prompts that included gene symbols and their corresponding descriptions.

Gene-relation adaptive soft prompt

GRASP is a gene-relation adaptive soft prompt tuning method that appends a few learnable virtual tokens conditioned on a gene pair. Let the backbone have hidden size H. For each gene g we build a fixed summary vector $s_g \in \mathbb{R}^H$. To obtain s_g, we first prompt the model to write a short synopsis of gene g. Then we feed this synopsis back into the same backbone, collect the last- layer hidden states for all tokens, and compute the mean across the token dimension. The result is a single H-dimensional vector that captures gene context. All gene summary vectors are cached in a look-up table S and kept fixed during training.

The core idea is to synthesize a soft prompt embedding from any context vector $z \in \mathbb{R}^H$ through a two-matrix factorization. GRASP maps z to a length-L prompt embedding $P(z) \in \mathbb{R}^{L \times H}$ as:

$$P(z) = C(z) \times B(z) \tag{1}$$

where $C(z) \in \mathbb{R}^{L \times r}$ is a gene-specific coefficient matrix and $B(z) \in \mathbb{R}^{r \times H}$ is a shared gene prototype matrix selected by routing. The coefficient matrix is produced by a lightweight linear projection:

$$C(z) = Reshape((W_c z + b_c), (L, r)) \tag{2}$$

and the prototype matrix is a convex combination of K global bases $\{B^{(k)} \in \mathbb{R}^{r \times H}\}_{k=1}^{K}$:

$$B(z) = \sum_{k=1}^{K} \alpha_k(z) \times B^{(k)},\ \alpha(z) = Softmax(W_b z + b_b) \tag{3}$$

where $\alpha_k(z) \geq 0$ and $\sum_{k=1}^{K} \alpha_k(z) = 1$. This factorization yields virtual tokens $P(z)$ whose content varies with z through $C(z)$ while sharing structure through the global bases in $B(z)$. The internal rank r controls capacity and the number of bases K controls diversity.

same parameterization (W_c, b_c, W_b, b_b) with inputs $z = s_a$ and $z = s_b$. The third segment captures the relation between the two genes and uses its own projection parameters $(W_c^{rel}, b_c^{rel}, W_b^{rel}, b_b^{rel})$ while sharing the same global basis bank $\{B^{(k)} \in \mathbb{R}^{r \times H}\}_{k=1}^{K}$. The relation context is a single vector:For a gene pair (a, b), we instantiate three prompt segments. Two segments are gene-specific and use the

$$u_{a,b} = \varphi(s_a, s_b) \tag{4}$$

with $\varphi(s_a, s_b) = |s_a - s_b|$ to emphasize contrastive biological features. Its prompt is:

$$P_{rel} = C^{rel}(u_{a,b}) \times B^{rel}(u_{a,b}) \tag{5}$$

The three segments are concatenated as:

$$P_{a,b} = Concat(P(s_a), P(s_b), P_{rel}) \in \mathbb{R}^{3L \times H} \tag{6}$$

and appended as a suffix to the query embeddings. With $L = 1$ the method contributes exactly three adaptive tokens per example.

The two-matrix view clarifies how GRASP balances specialization and sharing. The coefficient matrix $C(z)$ acts as a compact hypernetwork that selects r directions within the prototype subspace for each instance. The prototype matrix $B(z)$ restricts those directions to mixtures of a small bank of orthogonalized bases, which stabilizes training and promotes generalization across semantically related genes. This design turns pair-specific reasoning into structured modulation of a shared prompt space, enabling dynamic behavior through a compact factorized representation. Even with $L = 1$, a compact setting adding only three adaptive tokens per pair, the approach remains expressive by deriving each token from a full $r \times H$ prototype matrix and maintains efficiency through projection layers.

Classification head and training objective

To map token-level representations to a pairwise interaction probability, we attach a classification head to the frozen LLM backbone (replacing the language modeling head). Let $H \in \mathbb{R}^{T \times D}$ denote the last-layer hidden states for an input of T tokens and hidden size D. A position-wise three-layer MLP, with SiLU[35] activation after each linear layer, transforms each token embedding into a scalar local logit:

$$Z = \phi(\phi(\phi(HW_1 + b_1)W_2 + b_2)W_3 + b_3) \in \mathbb{R}^{T \times 1} \tag{7}$$

where b_1, b_2, b_3 are learnable biases and ϕ is the SiLU function.

These token-level logits are then aggregated into a single sequence logit through a learnable linear projection across positions:

$$s = a^T Z + b_{agg},\ a \in \mathbb{R}^T \tag{8}$$

where a contains the aggregation weights and b_{agg} is a learnable scalar bias. When soft prompt tokens are present, they are included in the aggregation and can influence the decision; when absent, the aggregation is over query tokens only.

The predicted probability is $\hat{y} = \sigma(s)$, where $\sigma(s) = 1/(1 + e^{-s})$ is the sigmoid function mapping logits to $(0, 1)$. Given a binary label $y \in \{0, 1\}$, the model is trained with the batch-averaged binary cross-entropy (BCE) loss:

$$\mathcal{L}_{BCE} = -\frac{1}{B}\sum_{i=1}^{B}[y_i log\, \hat{y}_i + (1 - y_i) log(1 - \hat{y}_i)] \tag{9}$$

During fine-tuning, optimization is restricted to the classification head and the soft prompt parameters, while all backbone parameters remain frozen to preserve the pre-trained knowledge and maintain training efficiency.

Hyperparameter settings

Across all methods we evaluated both Gemma-3-4B-IT[21] and Llama-3.1-8B-Instruct[22] together with their domain-adapted variants pretrained on biomedical corpora. Training was run for 15 epochs in every setting. We adopted BF16 precision and AdamW defaults from Hugginface Transformers with weight decay 0.01, learning rate 1e-4, warmup ratio 0.05. The batch size was set to 128 for the PPI network datasets and 64 for the phosphorylation network dataset.

Few-shot inference: The model was not fine-tuned. We prompted with six examples (three positive and three negative gene-pair relations) and decoded the answer for each query pair. Decoding used a maximum sequence length of 512 tokens when generating responses, and predictions were parsed from the text.

Fix prompt: We attached a small MLP classification head to the frozen backbone and fine-tuned only that head. We used two query styles: (i) Symbol-only prompts with sequence length 64, and (ii) BioContext-informed prompts that incorporate brief gene summaries with sequence length 256.

Task-specific soft prompt: We learned a short prompt of length 3 tokens and concatenated it either as a prefix or a suffix to the input while keeping the backbone frozen. We tokenized 80 query tokens (left padding with EOS as pad) so the classification head received 83 tokens.

Gene-relation adaptive soft prompt: GRASP generated 3 adaptive soft-prompt tokens per pair (two gene-specific and one relation-specific) via a matrix factorization. We kept the query length at 80 and set the soft-prompt length to 1 per segment. The internal rank r was set to 1024 for PPI network and 512 for phosphorylation network. In both cases we used K=4 routed base. Gene vectors were built once by mean-pooling the last hidden layer over a generated one-paragraph summary (maximum 128 new tokens per summary) and cached in an embedding table that remains fixed during training.

Reference

1. Badia-i-Mompel, P. *et al.* Gene regulatory network inference in the era of single-cell multi-omics. *Nat. Rev. Genet.* **24**, 739–754 (2023).

2. Xia, S., Chen, J., Arsala, D., Emerson, J. J. & Long, M. Functional innovation through new genes as a general evolutionary process. *Nat. Genet.* 1–15 (2025).

3. Bonev, B. *et al.* Opportunities and challenges of single-cell and spatially resolved genomics methods for neuroscience discovery. *Nat. Neurosci.* **27**, 2292–2309 (2024).

4. Zhu, Q. M. *et al.* Protein interaction networks in the vasculature prioritize genes and pathways underlying coronary artery disease. *Commun. Biol.* **7**, 87 (2024).

5. Kim, S. & Wysocka, J. Deciphering the multi-scale, quantitative cis-regulatory code. *Mol. Cell* **83**, 373–392 (2023).

6. Sevimoglu, T. & Arga, K. Y. The role of protein interaction networks in systems biomedicine. *Comput. Struct. Biotechnol. J.* **11**, 22–27 (2014).

7. De Oliveira, P. S. L. *et al.* Revisiting protein kinase–substrate interactions: Toward therapeutic development. *Sci. Signal.* **9**, (2016).

8. Sledzieski, S., Singh, R., Cowen, L. & Berger, B. D-SCRIPT translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions. *Cell Syst.* **12**, 969-982.e6 (2021).

9. Schmirler, R., Heinzinger, M. & Rost, B. Fine-tuning protein language models boosts predictions across diverse tasks. *Nat. Commun.* **15**, 7407 (2024).

10. Deznabi, I., Arabaci, B., Koyutürk, M. & Tastan, O. DeepKinZero: zero-shot learning for predicting kinase–phosphosite associations involving understudied kinases. *Bioinformatics* **36**, 3652–3661 (2020).

11. Zhou, Z. *et al.* Phosformer: an explainable transformer model for protein kinase-specific phosphorylation predictions. *Bioinformatics* **39**, btad046 (2023).

12. Jha, K., Saha, S. & Singh, H. Prediction of protein–protein interaction using graph neural networks. *Sci. Rep.* **12**, 8360 (2022).

13. Gao, Z. *et al.* Hierarchical graph learning for protein–protein interaction. *Nat. Commun.* **14**, 1093 (2023).

14. Zhang, F., Chang, S., Wang, B. & Zhang, X. DSSGNN-PPI: A Protein–Protein Interactions prediction model based on Double Structure and Sequence graph neural networks. *Comput. Biol. Med.* **177**, 108669 (2024).

15. Nováček, V. *et al.* Accurate prediction of kinase-substrate networks using knowledge graphs. *PLoS Comput. Biol.* **16**, e1007578 (2020).

16. Ko, Y. S., Parkinson, J., Liu, C. & Wang, W. TUnA: an uncertainty-aware transformer model for sequence-based protein–protein interaction prediction. *Brief. Bioinform.* **25**, bbae359 (2024).

17. Meyerhöfer, N., Krogan, N. J., Polacco, B. J. & Blumenthal, D. B. Inference of differential kinase interaction networks with KINference. *Bioinformatics* **41**, btaf349 (2025).

18. Shu, D. *et al.* Knowledge Graph Large Language Model (KG-LLM) for Link Prediction. Preprint at https://doi.org/10.48550/arXiv.2403.07311 (2025).

19. Ye, R., Zhang, C., Wang, R., Xu, S. & Zhang, Y. Language is All a Graph Needs. in *Findings of the Association for Computational Linguistics: EACL 2024* (eds. Graham, Y. & Purver, M.) 1955–1973 (Association for Computational Linguistics, St. Julian's, Malta, 2024).

20. Jin, M. *et al.* ProLLM: Protein Chain-of-Thoughts Enhanced LLM for Protein-Protein Interaction Prediction. Preprint at https://doi.org/10.48550/arXiv.2405.06649 (2024).

21. Team, G. *et al.* Gemma 3 Technical Report. Preprint at https://doi.org/10.48550/arXiv.2503.19786 (2025).

22. Grattafiori, A. *et al.* The llama 3 herd of models. *ArXiv Prepr. ArXiv240721783* (2024).

23. Gururangan, S. *et al.* Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. Preprint at https://doi.org/10.48550/arXiv.2004.10964 (2020).

24. Lester, B., Al-Rfou, R. & Constant, N. The Power of Scale for Parameter-Efficient Prompt Tuning. Preprint at https://doi.org/10.48550/arXiv.2104.08691 (2021).

25. Wei, J. *et al.* Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **35**, 24824–24837 (2022).

26. Oughtred, R. *et al.* The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci.* **30**, 187–200 (2021).

27. Mewes, H.-W. *et al.* MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **30**, 31–34 (2002).

28. Del Toro, N. *et al.* The IntAct database: efficient access to fine-grained molecular interaction data. *Nucleic Acids Res.* **50**, D648–D653 (2022).

29. Szklarczyk, D. *et al.* The STRING database in 2025: protein networks with directionality of regulation. *Nucleic Acids Res.* **53**, D730–D737 (2025).

30. Kotlyar, M. *et al.* IID 2021: towards context-specific protein interaction analyses by increased coverage, enhanced annotation and enrichment analysis. *Nucleic Acids Res.* **50**, D640–D647 (2022).

31. Hu, J. *et al.* PhosphoNetworks: a database for human phosphorylation networks. *Bioinformatics* **30**, 141–142 (2014).

32. Zabolotny, J. M. *et al.* Overexpression of the LAR (leukocyte antigen-related) protein-tyrosine phosphatase in muscle causes insulin resistance. *Proc. Natl. Acad. Sci.* **98**, 5187–5192 (2001).

33. Sevillano, J., Sánchez-Alonso, M. G., Pizarro-Delgado, J. & Ramos-Álvarez, M. del P. Role of receptor protein tyrosine phosphatases (RPTPs) in insulin signaling and secretion. *Int. J. Mol. Sci.* **22**, 5812 (2021).

34. Wu, C., MacLeod, I. & Su, A. I. BioGPS and MyGene. info: organizing online, gene-centric information. *Nucleic Acids Res.* **41**, D561–D565 (2013).

35. Elfwing, S., Uchibe, E. & Doya, K. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Netw.* **107**, 3–11 (2018).