A DATA-DRIVEN RECOMMENDATION FRAMEWORK FOR GENOMIC DISCOVERY

Ying Yang[*1], **Zhaoying Pan**[*1,2], **Jinge Ma**[1,2], **Daniel J. Klionsky**[1]

[1] University of Michigan
[2] Purdue University

ABSTRACT

Data-driven approaches to genomic discovery have been accelerated by emerging efforts in machine learning. However, due to the inherent complexity of genomic data, it can be challenging to model or utilize the data and their intricate relationships. In this work, we propose a framework for genomic prediction utilizing information from various genomic databases. We use a knowledge graph following existing work to extract gene representations and either use XGBoost or construct a graph to rank feature importance. By filtering key features and computing relevancy scores with genes that are known to be associated or unassociated with a specified area, we recommend unlabeled gene candidates with a high likelihood of association for further genomic research. We demonstrate how this framework works by applying it to autophagy genomics, illustrating its potential as a powerful recommendation system for genomic discovery.

1 INTRODUCTION

The explosion of various biological data types has opened a data-centric paradigm for the discovery of genes with novel functions. Within this paradigm, machine learning (ML) data pipelines deployed on extensive biological datasets (Libbrecht & Noble, 2015; Zitnik et al., 2019; Greener et al., 2022; Yang et al., 2024) can perform computational analysis to screen potential genes linked to particular pathways and diseases, revolutionizing the discovery of novel genes and the understanding of their functions. The challenges mainly lie in how to identify a universal and task-agnostic representation that can facilitate efficient search among the extensive and mostly unlabeled genomics space to recommend desirable gene candidates.

Traditionally, determining gene functions largely relies on functional genomics mining methods, such as CRISPR/Cas9 (Zhou et al., 2014), RNA interference (Vanhecke & Janitz, 2005), or gene knockouts (Skarnes et al., 2011). These methods alter gene expression, followed by phenotypic screening to observe changes that the genetic modifications incur. Novel functionalities often emerge as distinct phenotypic features or changes in cellular processes. However, this conventional method is often slow and demands significant resources. Efforts have been made to facilitate the gene identification process by utilizing computational network-based methods. For example, Erten et al. (2011) showed that the effectiveness of global prioritization techniques, including random walk and network propagation, is significantly influenced by the degree of candidate genes. Oprea et al. (2019) proposed a MetaPath framework that can be applied in conjunction with various classification algorithms, which has been shown to identify novel genes associated with autophagy successfully. Most of the existing graph representation learning methods in biological networks are based on extending the idea of random walks, which learn a continuous low-dimensional vector representation of nodes in a graph.

In this paper, we present a recommendation system that is inspired by the existing work (Oprea et al., 2019) while enhancing the integration of gene interactions. Following Oprea et al. (2019), we utilized the information collected from various genomic databases and generated the representation vectors for both labeled and unlabeled genes. Our framework prioritizes feature importance

[*] Equal contribution.

to select a subset of features using two methods: a graph-based approach integrating edge information for gene representations, and an approach relying solely on gene representations based on XGBoost (Chen & Guestrin, 2016).

For the graph-based method, we constructed graph convolutional networks (GCNs) (Kipf & Welling, 2016) with the representation vectors as node embedding and protein-protein interactions as edges. We then apply GNNExplainer (Ying et al., 2019) to calculate the average contributions for nodes with positive labels and select the top features based on the contribution ranking. In contrast, XGBoost is trained on the representation vectors and ranks features by their importance within the trained model. Our genomics recommendation framework uses the selected features to compute the relevancy score between unlabeled genes and genes with positive labels and recommends candidate genes that are most likely to be associated with the target biological term.

To summarize, our contribution is a flexible framework for gene recommendation based on feature selection using either XGBoost or graph convolutional networks. Given a biological term for a specific species, our framework identifies the most important features and recommends candidate genes that, while unlabeled, are likely to be associated with the term. We demonstrate the effectiveness of this approach by applying it to recommend autophagy-related genes for the model species yeast, showcasing its practical workflow.

2 Related Work

2.1 Deep learning for Genomic Application

Deep learning has been applied in many fields to assist research, due to its advantages of learning from large-scale data and capturing complex or subtle patterns. There has been a line of deep learning approaches applied in genomics (Libbrecht & Noble, 2015; Eraslan et al., 2019; Quazi, 2022), with popular deep learning methods having been widely adopted. One key task is to predict the gene expression from DNA sequences, where methods using CNNs (Zhou et al., 2018; Kelley et al., 2018; Kelley, 2020; Agarwal & Shendure, 2020) or Transformers (Avsec et al., 2021) showed strong performance. Another important task is the identification of genomic sequence elements, such as promoters, enhancers, or other regulatory elements. Methods (Alipanahi et al., 2015; Zhou & Troyanskaya, 2015; Kelley et al., 2016; Zeng et al., 2016) utilized CNNs to achieve this goal successfully, and popular methods include DeepBind (Alipanahi et al., 2015), DeepSEA (Zhou & Troyanskaya, 2015), and Basset (Kelley et al., 2016). The protein sequences can be utilized to predict protein functions with various methods based on CNNs (Kulmanov et al., 2018; Kulmanov & Hoehndorf, 2020), RNNs (Liu, 2017; Cao et al., 2017), and GNNs (Gligorijević et al., 2021). Unlike protein function prediction, which relies on sequence data, gene function prediction uses graph-based methods built upon protein-protein interaction (PPI) to represent gene interactions. Methods have been proposed to identify specified gene functions, such as functional genes (Oprea et al., 2019; Peng et al., 2021) or disease-related genes (Binder et al., 2022; Peng et al., 2022).

2.2 Feature Selection for Genomic Application

Feature selection refers to identifying and ranking the importance of features to select a subset of the most relevant ones, which can be used to select core features or remove redundant features, or gain biological insights for better explainability (Libbrecht & Noble, 2015; Eraslan et al., 2019). There are a number of methods in machine learning and deep learning that allow for ranking feature importance (Yang et al., 2024), such as Random Forest (Ho, 1995), LASSO Cox regression (Tibshirani, 1997), XGBoost (Chen & Guestrin, 2016), and GNNs with the GNNExplainer (Ying et al., 2019). Additionally, techniques including Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016) and SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017) provide explanations for general deep learning models. Given the reliance of PPI, which is generally modeled as a graph, we adopt the GNNs and GNNExplainer to rank the feature importance.

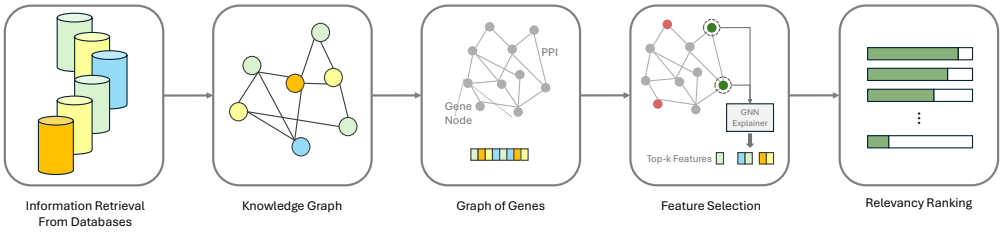

Figure 1: **The Example Workflow of the Graph-based Recommendation System**. The workflow begins with information retrieval from various databases to collect genomic and protein data to construct a knowledge graph integrating labeled and unlabeled genes. A graph of genes is created with protein-protein interactions as edges and representation vectors as node embeddings. Then, feature selection is performed using GNNExplainer, and relevancy ranking computes scores to recommend gene candidates most likely associated with the target biological term.

3 METHOD

3.1 OVERVIEW

This section discusses the workflow for the application in genomic discovery. Using a specified keyword of genetic function and its synonyms, we first retrieve relevant genomic data for a given species from multiple databases. To curate a set of negative examples, we collect data with negative labels based on keywords unassociated with autophagy, identified by domain experts. Following prior work (Sun et al., 2011; 2012; Oprea et al., 2019), we construct a knowledge graph to capture gene interactions specific to the keyword pathway. Each gene is then represented by a vector embedding calculated through MetaPath (Oprea et al., 2019), a method that encodes complex relationships within the knowledge graph. On top of these representations, we implement either an XGBoost-based or graph-based method to rank feature importance, identifying high-impact features for further analysis. Finally, relevancy scores are computed between each unlabeled gene and all genes that are known to be associated with the given term. The framework uses the relevancy scores to recommend the most relevant unlabeled genes as potential candidates for future research. The example of the graph-based workflow is illustrated in Figure 1.

This section discusses the general workflow of our framework, and Section 4 applies our approach to this well-studied model organism *saccharomyces cerevisiae* (yeast, brewer's yeast, or baker's yeast), and validate the framework's ability to uncover meaningful genomic insights within a defined biological term "autophagy". This approach serves as a powerful recommendation tool, guiding targeted genomic discovery for genomics practitioners in autophagy and beyond.

3.2 DATA COLLECTION

Given the keyword of a biological context or genomic function for one specified organism, we first narrow down our query to retrieve relevant information from the databases. Generally, we need the information of a species name, the biological keyword for the genomics discovery, and other biological keywords that are distinct from the given keyword to serve as negative labels. Due to the inconsistencies across different databases, we suggest using not only the given keyword and species name, but also considering their synonyms.

To embed the representations of genes, we need to collect information from various databases, such as databases for genomic and sequences (NCBI (Brown et al., 2015), Ensembl (Martin et al., 2023)), proteins (UniProt (Apweiler et al., 2004), InterPro (Paysan-Lafosse et al., 2023)), metabolic pathways (KEGG (Kanehisa & Goto, 2000), Reactome (Fabregat et al., 2018)), functional annotation (GO (Ashburner et al., 2000), UniProt), protein-protein interactions (STRING (Szklarczyk et al., 2023)), and homology/evolutionary relationships (NCBI, Ensembl). The comprehensive integration of these databases ensures a rich and unified representation of gene and protein attributes. Following Oprea et al. (2019), we collect the protein-specified information, including SwissProt accessions, symbols, names, and taxonomic identifiers from UniProt, Ensembl, and NCBI to facilitate

cross-referencing across gene and protein identifiers. The protein domains, families, and conserved sites are categorized using InterPro. Functional annotations, including Gene Ontology (GO) terms, are drawn from UniProt and GO, and metabolic pathways are integrated from KEGG and Reactome. STRING provides the protein-protein interactions (PPI), and homology and evolutionary relationships come from NCBI and Ensembl, linking homologous genes, taxonomic information, and protein identifiers.

3.3 GENE REPRESENTATION

To obtain the gene representation that fully utilizes the information from the above databases, we build a knowledge graph (KG) and obtain the embeddings of the KG with MetaPath, informed by existing work (Oprea et al., 2019). We integrate multiple biological data types that characterize genes (proteins), their interactions, and functional roles. The nodes in the KG include proteins identified by unique UniProt accessions, along with nodes representing pathways (from KEGG and Reactome), Gene Ontology (GO) terms for biological processes, molecular functions, and cellular components, and InterPro entries for conserved protein families and domains. The edges in the KG capture various relationships: protein-protein interactions (PPI) represent known interactions between proteins, pathway memberships link proteins to KEGG and Reactome pathways, GO annotations connect proteins to biological functions, and InterPro associations indicate structural or functional features shared by specific protein families. Note that despite Oprea *et al.*constructed the KG mainly for proteins, genes always share the same names with proteins in our case, therefore, we can construct the KG to consider genes and their associations as well.

To generate gene representations from the knowledge graph, MetaPath (Oprea et al., 2019) identifies meaningful pathways that connect target nodes to related genes through a sequence of intermediary nodes. Starting from a target node, the associated genes are gathered by examining connections within the graph, distinguishing between positively associated paths (direct relationships) and negatively associated paths (indirect or absent relationships). Each MetaPath captures distinct relationships, such as protein-protein interactions, pathway memberships, or functional annotations, allowing us to analyze genes based on their proximity and type of linkage to target nodes. For each relevant feature in the graph, the MetaPath values are computed by examining its connections and aggregating the topological information from each unique path type. This approach produces a feature matrix where each gene is represented by MetaPath-based scores that reflect its network relationships and associations within the graph.

3.4 FEATURE RANKING

Among all features generated from MetaPath on the knowledge graph, we aim to further identify the most influential features in the predictions for genes of our interest. We apply two methods for feature importance ranking: an XGBoost-based approach and a graph-based approach. XGBoost is a powerful ensemble learning algorithm that is efficient and effective in various machine learning scenarios. XGBoost iteratively constructs decision trees to minimize the predictive error. In addition to its superior performance, XGBoost has desirable properties to evaluate the contribution of each feature by measuring the improvement in accuracy when the feature is used to split a node in a decision tree. Fitting the gene representations and their labels, XGBoost can provide insights into the most influential features in the prediction.

In contrast, our graph-based approach uses not only gene representations but also gene interactions via protein-protein interactions. We use a simple Graph Convolutional Network (GCN), which builds on a graph with MetaPath representations as node embedding, and PPI from STRING as edges. To evaluate feature importance, we apply GNNExplainer for each gene labeled as relevant to the given term and compute the contribution for features averaged from all relevant gene nodes. The average contribution allows us to rank features effectively based on their roles in the predictions with the graph. Both approaches enable the revealing of the most important features in prediction based on known genes. They not only facilitate the gene recommendation for the next step but also promote transparency of prediction. Besides, the effectiveness of the framework can be validated by examining the relation between the top features and the given term. Moreover, the top features can serve as inspiration for wet experiments as well.

3.5 CANDIDATE GENE RECOMMENDATION

We utilize the masked gene representations that contain only the top features for new gene recommendations. We apply the Targeted Attribute Prediction Distance (TAPD) method to calculate relevancy scores and rank the scores for top candidate genes. TAPD computes the distance between each unlabeled candidate gene and each gene labeled as relevant in the feature space:

$$TAPD = \sum_{k=1}^{K} \frac{|h_k^c - h_k^p|}{h_k^p}$$

where TAPD is the total absolute percent difference, K is the total number of selected features, h^c and h^p are the features of unlabeled candidate and positively labeled data, respectively. TAPD measures the composite deviation of candidate features from positively labeled data features, and a lower score is preferred. We define *relevancy score* R as the reciprocal of the TAPD averaged on all selected features and positively-labeled data:

$$R = \frac{NK}{\sum_{n=1}^{N} \sum_{k=1}^{K} \frac{|h_k^c - h_k^p|}{h_k^p}}$$

where N is the total number of genes labeled to be associated with the specified term and K is the total number of features selected to be most influential.

The relevancy scores enable us to rank the unlabeled genes with the likelihood of association with genes that are known to be associated with the given biological term for gene functions. We filter the top candidate genes with the highest relevancy scores as the recommendation results, aiming to facilitate the selection process for wet experiments to identify gene functions. Moreover, the framework can be flexible and adapt to more annotations for more accurate and reliable recommendations in the future.

4 EXPERIMENTS

4.1 "AUTOPHAGY" FOR YEAST: AN EXAMPLE

Although our framework aims to serve for general genomics discovery, we use one example, an autophagy-related gene (ATG) recommendation for the model organism yeast, to exhibit the workflow of our framework. Autophagy is a fundamental cellular process that maintains cell health by breaking down and recycling damaged parts of cells (Yang et al., 2024). Therefore, discovering genes involved in autophagy is crucial, as it can lead to insights into disease mechanisms and potential therapeutic targets like cancer, neurodegeneration, and infectious diseases (Yang et al., 2024). As we discussed in Section 3.1, we start with the specification of biological terminologies. Given the keyword "autophagy" and the organism "saccharomyces cerevisiae" (we refer to it as *yeast* for simplicity), the following information is listed per the suggestions from domain experts:

Organism name Saccharomyces cerevisiae (NCBI taxonomy Id 4932)

Alternative organism names brewer's yeast, budding yeasts, ATCC 18824, Candida robusta, NRRL Y-12632, S. cerevisiae, Saccharomyces capensis, Saccharomyces italicus, Saccharomyces oviformis, Saccharomyces uvarum var. melibiosus, lager beer yeast, yeast.

Keyword for Genomics Discovery Autophagy (KEGG pathway map04138, titled "Autophagy - yeast")

Synonyms of the Keyword Mitophagy, autophagic pathway, CVT pathway.

Keywords for Negative Labels TCA (tricarboxylic acid cycle), cell cycle, meiosis.

With the information above, we collect data from various genomics databases, as discussed in Section 3.2. Specifically, we query relevant genes, pathways, and functional annotations linked to autophagy from sources like KEGG (particularly map04138, the "Autophagy - yeast" pathway), Reactome, and Gene Ontology (GO). This query gathers both positive and negative samples based on the specified keywords and synonyms, where genes associated with "autophagy" and its related

terms are labeled as positive, while those linked to unrelated processes, such as "TCA", "cell cycle" and "meiosis" are labeled as negative.

Using this curated information, we construct the knowledge graph by integrating protein-protein interactions, pathway associations, and functional annotations. Each gene is embedded in the graph using MetaPath-based representations to capture the relational context between ATG and non-ATG. In the end, we used 868 labeled genes in total and generated 1060 features for each gene.

The edges for the GCN are derived from protein-protein interaction (PPI) data sourced from the STRING database. This data includes connections between proteins based on evidence of physical or functional interactions. To ensure relevance to our study, we filtered the interactions to retain only those between proteins that match genes in our labeled data. We also removed redundant edges, standardizing each protein pair to maintain an undirected edge structure. This process results in a refined set of edges that accurately represent meaningful interactions within our specific set of genes, serving as the foundation for the GCN to learn from the network structure and relationships among proteins.

4.2 Recommendation Results

As we discussed in Section 3.4, we use XGBoost or GCN to rank the importance of features. By applying both methods, we offer flexibility on whether to use gene representations solely and validate the prediction across methods. Based on how the XGBoost or GCN works, the results from the graph-based method utilize the protein-protein interactions as the edge information, while the results from the XGBoost-based method might focus on the genes that have connections with known ATG or pathways.

Table 1: **Top-5 Most Important Features Ranked by Graph-based Method.** This table presents the top five features identified as most significant by the graph-based method, highlighting their potential importance in the underlying biological or structural relationships.

Feature	Description	Gain Value
GCN4	General control transcription factor GCN4 (Amino acid biosynthesis regulatory protein) (General control protein GCN4)	0.162
PSD1	Phosphatidylserine decarboxylase proenzyme 1, mitochondrial (EC 4.1.1.65) [Cleaved into: Phosphatidylserine decarboxylase 1 beta chain; Phosphatidylserine decarboxylase 1 alpha chain]	0.157
SEC13	Protein transport protein SEC13	0.155
RPL15B	Large ribosomal subunit protein eL15B (60S ribosomal protein L15-B) (L13) (RP15R) (YL10) (YP18)	0.151
SNF4	5'-AMP-activated protein kinase subunit gamma (AMPK gamma) (AMPK subunit gamma) (Regulatory protein CAT3) (Sucrose non-fermenting protein 4)	0.149

For XGBoost-based and graph-based methods of feature ranking, we report the top-10 most important features calculated by them, respectively, along with the feature descriptions, in Table 2 and 1. Compared with the graph-based results, XGBoost-based results favor more direct connections with autophagy-related genes, such as ATG11 and ATG5, which aligns well with our expectation for XGBoost methods. In contrast, the graph-based method provides a list of more general features that do not relate to autophagy explicitly. We provide further evidence from recent research in Section 4.3 regarding the rationality of the top features. In addition, we identified 293 common features out of the top 500 features from both methods, underscoring the effectiveness of identifying features for both methods. More information on the full top 500 features is attached in the supplementary materials.

With the features selected by either XGBoost or GCN, we compute the relevancy scores between unlabeled genes and known autophagy-related genes to recommend genes that are most likely to be associated with autophagy. The recommended genes from both methods are displayed in Table 5

Table 2: **Top-5 Most Important Features Ranked by XGBoost-based Method.** Similarly, this table presents the top five features identified as most significant by the XGBoost-based method.

Feature	Description	Gain Value
VPS21	Vacuolar protein sorting-associated protein 21 (GTP-binding protein YPT51) (Vacuolar protein-targeting protein 12)	0.114
ATG11	Autophagy-related protein 11 (Cytoplasm to vacuole targeting protein 9)	0.083
ATG5	Autophagy protein 5	0.036
CDC14	Tyrosine-protein phosphatase CDC14 (EC 3.1.3.48)	0.035
PEP5	E3 ubiquitin-protein ligase PEP5 (EC 2.3.2.27) (Carboxypeptidase Y-deficient protein 5) (Histone E3 ligase PEP5) (RING-type E3 ubiquitin transferase PEP5) (Vacuolar biogenesis protein END1) (Vacuolar morphogenesis protein 1) (Vacuolar protein sorting-associated protein 11) (Vacuolar protein-targeting protein 11)	0.034

Table 3: **Top-10 Genes Ranked by Both Methods.** X-Rank and G-Rank provide the rank of the gene in the lists from XGBoost-based recommendation and graph-based recommendation. In addition, the information of each gene from databases is provided.

X-Rank	G-Rank	Symbol	Gene Name	Swissprot	UniProt	NCBI ID	Description
1	35	UTP30	YKR060W	RL1D1_YEAST	P36144	853934	Ribosome biogenesis protein UTP30 (U3 snoRNP-associated protein UTP30)
2	38	MET13	YGL125W	MTHR2_YEAST	P53128	852752	Methylenetetrahydrofolate reductase 2 (EC 1.5.1.53) (YmL45)
20	23	RPL27A	YHR010W	RL27A_YEAST	P0C2H6	856401	Large ribosomal subunit protein eL27A (60S ribosomal protein L27-A)
3	43	LSO1	YJR005C-A	LSO1_YEAST	Q3E827	1466469	Protein LSO1 (Late-annotated small open reading frame 1)
48	2	GNA1	YFL017C	GNA1_YEAST	P43577	850529	Glucosamine 6-phosphate N-acetyltransferase (EC 2.3.1.4) (Phosphoglucosamine acetylase) (Phosphoglucosamine transacetylase)
5	48	MNN4	YKL201C	MNN4_YEAST	P36044	853634	Protein MNN4
53	1	UIP4	YPL186C	UIP4_YEAST	Q08926	855916	ULP1-interacting protein 4
37	17	SDS24	YBR214W	SDS24_YEAST	P38314	852515	Protein SDS24
7	50	WRS1	YOL097C	SYWC_YEAST	Q12109	854056	Tryptophan–tRNA ligase, cytoplasmic (EC 6.1.1.2) (Tryptophanyl-tRNA synthetase) (TrpRS)
8	52	YKE4	YIL023C	YKE4_YEAST	P40544	854789	Zinc transporter YKE4

and 4 for XGBoost-based and graph-based recommendations, respectively. Additionally, we provide the systematic gene names (denoted as Gene Name), SwissProt identifiers (denoted as SwissProt), UniProt accession numbers (denoted as UniProt), NCBI IDs, and their descriptions for more convenient usage across databases. We compared the top 500 recommended genes from both methods and identified 359 common recommendations among them, which provides a promising list of possible ATGs for future research. We show the ranks in both methods for the overlapping 359 genes in Figure 2, and a consistent trend in ranking is observed.

4.3 Validation from Recent Research

Since our framework aims to identify new genes that are likely to be associated with the given keyword for a biological process, it is challenging to validate the recommended genes without relevant genomics experiments. In this section, we compiled information from several autophagy studies to substantiate the relationship between our top features and the given keyword autophagy in terms of their functionality. For example, GCN4, which is the most important feature selected by the graph-based method, is a primary transcriptional activator involved in the induction of specific ATG genes in response to amino acid starvation, playing a crucial role in autophagy regulation Natarajan

et al. (2001); Prigent et al. (2024). The second most important feature, PSD1, is a yeast enzyme for autophagy regulation and its overexpression can increase autophagy Rockenfeller et al. (2015). SEC13, ranked third, was reported that its silence can result in a defect of autophagy.

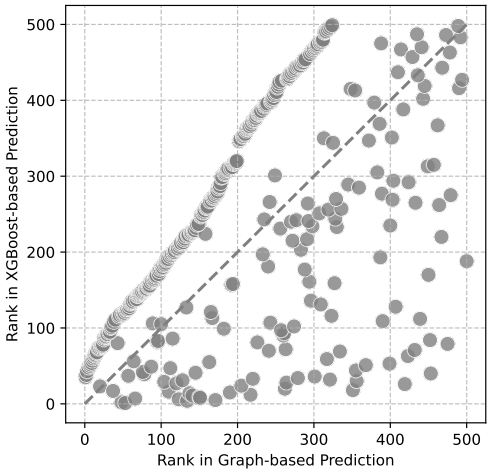

Figure 2: **The Ranks of Genes Recommended by Both Methods**. There are 359 overlapping predictions among the top 500 genes recommended by both methods. This figure shows the ranks in each method for the 359 predictions, where we can observe the majority of them achieved similar ranks in both methods.

Compared to the graph-based selected features, XGBoost uses features that are associated with autophagy more explicitly, such as ATG11 and ATG5, ranked as the second and the third respectively, which are key genes in autophagy. Besides, VPS21, as the first feature, regulates autophagy, and its deletion of the module results in autophagy defects and accumulation of autophagosomal clusters Chen et al. (2014).

Since our recommendation system aims to guide the revealing of gene functions for new genes, there is little research working on the relationship between these recommended genes and autophagy.

We provide the information of recommended genes with high ranks in both methods in Table 3, which might offer insights for domain experts in experiment designs to explore possible relations. For example, the UTP30 recommended by the graph-based method, plays a crucial role in ribosome biogenesis according to its description. Despite not relating to autophagy directly, the disruptions in UTP30 function might relate to autophagy through ribophagy, thus leading to autophagy for clearing out incomplete or faulty ribosomes. On the other hand, UIP4, which was primarily recommended by XGBoost, has several functions primarily related to nuclear. The cell might activate nucleophagy, a form of autophagy for nucleus parts if UIP4 misfunctions in maintaining the stability of the nucleus.

5 Discussion

In this work, we present a recommender framework for genomic discovery utilizing varied databases. By integrating data from multiple sources, such as KEGG, STRING, and Gene Ontology, we built a graph that represents genes as nodes with rich contextual embeddings computed from a knowledge graph, while protein-protein interactions form the edges. This multi-relational graph provides a structured approach to capturing complex genomic relationships that would be challenging to model with traditional methods. Either using XGBoost with the sole gene representations or using the graph explanations, we filter key features and generate relevancy scores for recommending unlabeled gene candidates. We demonstrated the workflow of our framework with the example of recommending autophagy-related genes for yeast.

Broader Impact. Our framework serves for genomic discovery for general purposes, given an organism and the term for genomic context. It has the potential to advance genomic research by suggesting key features and highly related genes for the term. Besides, the framework might accelerate insights into the biological processes and pathways, in terms of their relations to the given term. In addition, the scalability and flexibility of this framework enable it to be broadly applicable across organisms and gene functions, and allow researchers to prioritize candidate genes in experimental setups.

Limitation. The functionality of the framework heavily depends on the data from the genomic databases and the way to utilize their interactions. The possible bias or gaps in the databases, such as under-representation in the study of certain pathways, might lead to biased or incorrect predictions.

Additionally, the specification of the negative labels can introduce bias as well. To obtain the top important features, we use data with positive (related to the given term) and negative labels to train the XGBoost or GCN. Although we suggest that the choice of negative labels should be proceeded with caution under the advice from domain experts, it can lead to incorrect features or predictions if they are chosen improperly. Lastly, our framework relies on the feature ranking from XGBoost or GCN. Despite their powerful feature selection mechanism, it is possible that they may not be able to capture all nuances of biological relationships, thus introducing bias in the results.

REFERENCES

Vikram Agarwal and Jay Shendure. Predicting mrna abundance directly from genomic sequence using deep convolutional neural networks. *Cell reports*, 31(7), 2020.

Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 33 (8):831–838, 2015.

Rolf Apweiler, Amos Bairoch, Cathy H Wu, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, et al. Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 32(suppl_1):D115–D119, 2004.

Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18 (10):1196–1203, 2021.

Jessica Binder, Oleg Ursu, Cristian Bologa, Shanya Jiang, Nicole Maphis, Somayeh Dadras, Devon Chisholm, Jason Weick, Orrin Myers, Praveen Kumar, et al. Machine learning prediction and tau-based screening identifies potential alzheimer's disease genes relevant to immunity. *Communications Biology*, 5(1):125, 2022.

Garth R Brown, Vichet Hem, Kenneth S Katz, Michael Ovetsky, Craig Wallin, Olga Ermolaeva, Igor Tolstoy, Tatiana Tatusova, Kim D Pruitt, Donna R Maglott, et al. Gene: a gene-centered information resource at ncbi. *Nucleic acids research*, 43(D1):D36–D42, 2015.

Renzhi Cao, Colton Freitas, Leong Chan, Miao Sun, Haiqing Jiang, and Zhangxin Chen. Prolango: protein function prediction using neural machine translation based on a recurrent neural network. *Molecules*, 22(10):1732, 2017.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Yong Chen, Fan Zhou, Shenshen Zou, Sidney Yu, Shaoshan Li, Dan Li, Jingzhen Song, Hui Li, Zhiyi He, Bing Hu, et al. A vps21 endocytic module regulates autophagy. *Molecular biology of the cell*, 25(20):3166–3177, 2014.

Gökcen Eraslan, Žiga Avsec, Julien Gagneur, and Fabian J Theis. Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, 20(7):389–403, 2019.

Sinan Erten, Gurkan Bebek, Rob M Ewing, and Mehmet Koyutürk. Dada: degree-aware algorithms for network-based disease gene prioritization. *BioData mining*, 4:1–20, 2011.

Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, et al. The reactome pathway knowledgebase. *Nucleic acids research*, 46(D1):D649–D655, 2018.

Vladimir Gligorijević, P Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C Taylor, Ian M Fisk, Hera Vlamakis, et al. Structure-based protein function prediction using graph convolutional networks. *Nature communications*, 12(1):3168, 2021.

Joe G Greener, Shaun M Kandathil, Lewis Moffat, and David T Jones. A guide to machine learning for biologists. *Nature reviews Molecular cell biology*, 23(1):40–55, 2022.

Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pp. 278–282. IEEE, 1995.

Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

David R Kelley. Cross-species regulatory sequence activity prediction. *PLoS computational biology*, 16(7):e1008050, 2020.

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

David R Kelley, Yakir A Reshef, Maxwell Bileschi, David Belanger, Cory Y McLean, and Jasper Snoek. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome research*, 28(5):739–750, 2018.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Maxat Kulmanov and Robert Hoehndorf. Deepgoplus: improved protein function prediction from sequence. *Bioinformatics*, 36(2):422–429, 2020.

Maxat Kulmanov, Mohammed Asif Khan, and Robert Hoehndorf. Deepgo: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics*, 34 (4):660–668, 2018.

Maxwell W Libbrecht and William Stafford Noble. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6):321–332, 2015.

Xueliang Liu. Deep recurrent neural network for protein function prediction from sequence. *arXiv preprint arXiv:1701.08318*, 2017.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017. URL `https://arxiv.org/abs/1705.07874`.

Fergal J Martin, M Ridwan Amode, Alisha Aneja, Olanrewaju Austine-Orimoloye, Andrey G Azov, If Barnes, Arne Becker, Ruth Bennett, Andrew Berry, Jyothish Bhai, et al. Ensembl 2023. *Nucleic acids research*, 51(D1):D933–D941, 2023.

Krishnamurthy Natarajan, Michael R Meyer, Belinda M Jackson, David Slade, Christopher Roberts, Alan G Hinnebusch, and Matthew J Marton. Transcriptional profiling shows that gcn4p is a master regulator of gene expression during amino acid starvation in yeast. *Molecular and cellular biology*, 2001.

Tudor I Oprea, Jeremy J Yang, Daniel R Byrd, and Vojo Deretic. Autophagy dark genes: Can we find them with machine learning? *bioRxiv*, pp. 715037, 2019.

Typhaine Paysan-Lafosse, Matthias Blum, Sara Chuguransky, Tiago Grego, Beatriz Lázaro Pinto, Gustavo A Salazar, Maxwell L Bileschi, Peer Bork, Alan Bridge, Lucy Colwell, et al. Interpro in 2022. *Nucleic acids research*, 51(D1):D418–D427, 2023.

Jiajie Peng, Hansheng Xue, Zhongyu Wei, Idil Tuncali, Jianye Hao, and Xuequn Shang. Integrating multi-network topology for gene function prediction using deep neural networks. *Briefings in bioinformatics*, 22(2):2096–2105, 2021.

Wei Peng, Qi Tang, Wei Dai, and Tielin Chen. Improving cancer driver gene identification using multi-task learning on graph convolutional network. *Briefings in Bioinformatics*, 23(1):bbab432, 2022.

Magali Prigent, Hélène Jean-Jacques, Delphine Naquin, Stéphane Chédin, Marie-Hélène Cuif, Renaud Legouis, and Laurent Kuras. Sulfur starvation-induced autophagy in saccharomyces cerevisiae involves sam-dependent signaling and transcription activator met4. *Nature Communications*, 15(1):6927, 2024.

Sameer Quazi. Artificial intelligence and machine learning in precision and genomic medicine. *Medical Oncology*, 39(8):120, 2022.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Patrick Rockenfeller, M Koska, F Pietrocola, N Minois, O Knittelfelder, V Sica, J Franz, D Carmona-Gutierrez, G Kroemer, and F Madeo. Phosphatidylethanolamine positively regulates autophagy and longevity. *Cell Death & Differentiation*, 22(3):499–508, 2015.

William C Skarnes, Barry Rosen, Anthony P West, Manousos Koutsourakis, Wendy Bushell, Vivek Iyer, Alejandro O Mujica, Mark Thomas, Jennifer Harrow, Tony Cox, et al. A conditional knockout resource for the genome-wide study of mouse gene function. *Nature*, 474(7351):337–342, 2011.

Yizhou Sun, Rick Barber, Manish Gupta, Charu C Aggarwal, and Jiawei Han. Co-author relationship prediction in heterogeneous bibliographic networks. In *2011 International Conference on Advances in Social Networks Analysis and Mining*, pp. 121–128. IEEE, 2011.

Yizhou Sun, Jiawei Han, Charu C Aggarwal, and Nitesh V Chawla. When will it happen? relationship prediction in heterogeneous information networks. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pp. 663–672, 2012.

Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, et al. The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research*, 51(D1):D638–D646, 2023.

Robert Tibshirani. The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997.

Dominique Vanhecke and Michal Janitz. Functional genomics using high-throughput rna interference. *Drug discovery today*, 10(3):205–212, 2005.

Ying Yang, Zhaoying Pan, Jianhui Sun, Joshua Welch, and Daniel J Klionsky. Autophagy and machine learning: Unanswered questions. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, pp. 167263, 2024.

Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, 32, 2019.

Haoyang Zeng, Matthew D Edwards, Ge Liu, and David K Gifford. Convolutional neural network architectures for predicting dna–protein binding. *Bioinformatics*, 32(12):i121–i127, 2016.

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

Jian Zhou, Chandra L Theesfeld, Kevin Yao, Kathleen M Chen, Aaron K Wong, and Olga G Troyanskaya. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature genetics*, 50(8):1171–1179, 2018.

Yuexin Zhou, Shiyou Zhu, Changzu Cai, Pengfei Yuan, Chunmei Li, Yanyi Huang, and Wensheng Wei. High-throughput screening of a crispr/cas9 library for functional genomics in human cells. *Nature*, 509(7501):487–491, 2014.

Marinka Zitnik, Francis Nguyen, Bo Wang, Jure Leskovec, Anna Goldenberg, and Michael M Hoffman. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion*, 50:71–91, 2019.

A TOP RECOMMENDATIONS FROM EACH METHOD

We provide the top 10 genes that rank high in the recommendation from both methods in Table 3. In addition, we display the top 10 genes recommended by each method for further reference in Table 4 and Table 5.

Table 4: **Top-10 Genes Recommended by Graph-based Method.**

Symbol	Gene Name	Swissprot	UniProt	NCBI ID	Description
UTP30	YKR060W	RL1D1_YEAST	P36144	853934	Ribosome biogenesis protein UTP30 (U3 snoRNP-associated protein UTP30)
MET13	YGL125W	MTHR2_YEAST	P53128	852752	Methylenetetrahydrofolate reductase 2 (EC 1.5.1.53) (YmL45)
LSO1	YJR005C-A	LSO1_YEAST	Q3E827	1466469	Protein LSO1 (Late-annotated small open reading frame 1)
YMD8	YML038C	YMD8_YEAST	Q03697	854970	Putative nucleotide-sugar transporter YMD8
MNN4	YKL201C	MNN4_YEAST	P36044	853634	Protein MNN4
RFU1	YLR073C	RFU1_YEAST	Q08003	850762	Regulator of free ubiquitin chains 1
WRS1	YOL097C	SYWC_YEAST	Q12109	854056	Tryptophan–tRNA ligase, cytoplasmic (EC 6.1.1.2) (Tryptophanyl-tRNA synthetase) (TrpRS)
YKE4	YIL023C	YKE4_YEAST	P40544	854789	Zinc transporter YKE4
COX26	YDR119W-A	COX26_YEAST	Q2V2P9	3799970	Cytochrome c oxidase subunit 26, mitochondrial
THS1	YIL078W	SYTC_YEAST	P04801	854732	Threonine–tRNA ligase, cytoplasmic (EC 6.1.1.3) (Threonyl-tRNA synthetase) (ThrRS)

Table 5: **Top-10 Genes Recommended by XGBoost-based Method.**

Symbol	Gene Name	Swissprot	UniProt	NCBI ID	Description
UIP4	YPL186C	UIP4_YEAST	Q08926	855916	ULP1-interacting protein 4
GNA1	YFL017C	GNA1_YEAST	P43577	850529	Glucosamine 6-phosphate N-acetyltransferase (EC 2.3.1.4) (Phosphoglucosamine acetylase) (Phosphoglucosamine transacetylase)
TDA10	YGR205W	TDA10_YEAST	P42938	853119	Probable ATP-dependent kinase TDA10 (EC 2.7.-.-) (Topoisomerase I damage affected protein 10)
ERR3	YMR323W	ERR3_YEAST	P42222	855373	Enolase-related protein 3 (EC 4.2.1.11) (2-phospho-D-glycerate hydro-lyase) (2-phosphoglycerate dehydratase)
SAP1	YER047C	SAP1_YEAST	P39955	856771	Protein SAP1 (SIN1-associated protein)
MDJ2	YNL328C	MDJ2_YEAST	P42834	855388	Mitochondrial DnaJ homolog 2
ASN1	YPR145W	ASNS1_YEAST	P49089	856268	Asparagine synthetase [glutamine-hydrolyzing] 1 (EC 6.3.5.4) (Glutamine-dependent asparagine synthetase 1)
SRT1	YMR101C	SRT1_YEAST	Q03175	855127	Dehydrodolichyl diphosphate synthase complex subunit SRT1 (EC 2.5.1.87) (Ditrans,polycis-polyprenyl diphosphate synthase ((2E,6E)-farnesyl diphosphate specific))
MRK1	YDL079C	MRK1_YEAST	P50873	851480	Serine/threonine-protein kinase MRK1 (EC 2.7.11.1)
MSA1	YOR066W	MSA1_YEAST	Q08471	854232	G1-specific transcription factors activator MSA1 (MBF and SBF-associated protein 1)

B Database Information and Usage

Following our discussion in Section 4.2, we retrieved information from a range of databases for autophagy research on yeast (Saccharomyces cerevisiae) for the knowledge graph (Oprea et al., 2019). From the Gene Ontology (GO) database, we obtained the core ontology structure (go.obo) and species-specific annotations for Saccharomyces cerevisiae from the GO Association File (sgd.gaf). Protein-protein interaction data were retrieved from the STRING database v11.0, specifically using the protein links file (protein.links) and protein information file (protein.info) for S. cerevisiae with a taxonomy ID of 4932. For pathway analysis, we incorporated Reactome data, utilizing both the pathway definitions (ReactomePathways.txt) and protein-pathway associations (UniProt2Reactome_All_Levels.txt). InterPro data, particularly from protein2ipr.dat.gz and entry.list, is utilized to associate proteins with known protein families and domains. To ensure consistent identifier mapping across databases, we used the Ensembl database (release 109) UniProt mapping files which provide us with gene identifiers linked to protein accession numbers. In addition, the label information is retrieved from GO, KEGG, and UniProt according to the specified biological terms.