Cross-Species Graph Neural Network for Translating Animal Disease Resistance to Human Drug Targets

Deek Guruge
Fauna Bio
deek@faunabio.com

Michelle M. Li*
Harvard Medical School
michelleli@g.harvard.edu

Natalie DeForest
Fauna Bio
natalie@faunabio.com

Phil McNamara
Fauna Bio
phil@faunabio.com

Manashvi Borad
Fauna Bio
manashvi@faunabio.com

Linda Goodman*
Fauna Bio
linda@faunabio.com

Abstract

Therapeutic target discovery has traditionally relied on data from human patients and mouse models, which primarily capture disease states rather than mechanisms of resistance or repair. Animals with evolved disease-protective adaptations offer complementary insights that, when properly combined with human data, can drive more discoveries. Here, we introduce CENTAURGNN, a cross-species graph neural network for novel target discovery. It leverages species-specific message passing to integrate animal data with large-scale human data. In particular, CENTAURGNN incorporates data from animal disease resistance (ADR) adaptations to conditions like obesity, heart disease, kidney disease, and neurodegeneration, providing complementary biological signals to human-only datasets. We demonstrate that CENTAURGNN has up to 2 times stronger generalizability (via ROC) to unseen gene-indication associations than state-of-the-art baselines. We also find that integrating animal data to a model trained only on human data improves predictive performance by 9% (via ROC). Additionally, feature attribution analysis confirms the significant contributions of animal data toward CENTAURGNN's predictions, with animal-derived features having at least an order of magnitude higher average importance than human-derived features. Finally, we validate our findings using gene-based burden tests in the UK Biobank ($\sim$400,000 exomes). Our analysis shows that CENTAURGNN, when enhanced with ADR data, nominates an average of 448 more genetically-validated novel targets than our human-only baseline, indicating its translational potential for drug discovery.

1 Introduction

Identifying novel, effective drug targets is a critical challenge in pharmaceutical research, characterized by high costs, long timelines, and a low probability of success [11, 53]. A key reason for this is a fundamental reliance on data that capture disease states, rather than innate mechanisms of health and resilience [57]. Preclinical research has historically centered on inducing disease symptoms in a narrow range of model organisms, offering a limited view of biology that often fails to translate to human clinical outcomes [18, 21, 34]. This approach overlooks a complementary source of insight: certain animal species have *naturally* evolved robust mechanisms to resist pressures like heart disease, kidney disease, metabolic dysfunction, and neurodegeneration [52, 48, 5, 10, 12, 51, 23].

*Co-corresponding authors

The NeurIPS 2025 Workshop on AI Virtual Cells and Instruments: A New Era in Drug Discovery and Development (AI4D3 2025), San Diego, California, USA, 2025.

For example, hibernating ground squirrels develop tau hyperphosphorylation (similar to Alzheimer's patients) as they enter hibernation, but are able to reverse this pathology upon arousal [3].

Harnessing these unique biological signals from animal studies requires integrating them with large-scale human data, a task that machine learning is well-suited for due to its ability to uncover complex patterns from vast, heterogeneous datasets [60, 1]. To this end, we introduce CENTAURGNN, a novel cross-species graph neural network designed to enable genome-wide *in silico* screens by explicitly bridging this conceptual gap. **Our core contributions** are:

1. The construction of a large-scale multi-species knowledge graph, CENTAUR, that integrates human genomic data with high-resolution data from species with animal disease resistance.
2. A species-specific message passing mechanism in CENTAURGNN that captures cross-species biological principles to enable animal-to-human transfer learning.

We empirically demonstrate that CENTAURGNN's cross-species integration provides substantial performance gains in predicting gene-indication associations. Concretely, we investigate the benefit of leveraging animal disease resistance data to enhance target discovery, and quantify the contribution of animal data (i.e., animal disease resistance information) towards model predictions.

2 Related Work

2.1 Graph machine learning for biomedicine

Graph machine learning is a predominant approach for encoding complex biomedical systems into low-dimensional vector embeddings to accelerate discovery [31]. Shallow embedding algorithms (e.g., DeepWalk [44], node2vec [15]) and message-passing neural networks (e.g., graph convolutional network [27], GraphSAGE [16], graph attention network [56]) have been widely adapted for biomedical tasks [31, 24]. Contextual graph learning is advancing our understanding of single-cell protein biology by representing context-specific topology [32]. Few-shot learning on graphs can accelerate rare disease diagnosis by enriching patients subgraph representations with biomedical knowledge [2]. Heterogeneous graph neural networks (GNNs) have been developed for predicting adverse drug reactions [13], and graph foundation models are enabling clinician-centered drug repurposing [20]. **Our contribution.** We design a novel species-specific message passing mechanism to translate the complex biological signals of ADR across a multi-species knowledge graph.

2.2 Machine learning for target identification

Applying machine learning to target identification has evolved from early feature-based methods to sophisticated deep learning approaches [58]. Due to the interconnectedness of biology, GNNs have become a dominant paradigm [9, 20, 32]. Pioneering models like Decagon [60] and DTINet [35] have demonstrated the ability of GNNs to predict drug-target and drug-drug interactions by learning on experimentally-derived associations between drugs and proteins. More recently, GNNs that leverage biological knowledge graphs to learn contextualized embeddings of genes, proteins, and diseases have facilitated cell type specific target identification and large-scale explainable drug repositioning [32, 20]. In addition to architectural innovation, integrating diverse datasets and contexts is critical for models to better capture the complexities of human disease [24, 20, 14]. **Our contribution.** We introduce a novel class of complementary data based on preclinical animal studies on ADR. Our knowledge graph and model architecture leverage the underutilized biological insights from evolved disease resistance to accelerate target discovery [57, 43].

2.3 Cross-species transfer learning for target discovery

Machine learning has been widely used to enrich human-centric datasets, bridging the translational gap between preclinical animal research and human outcomes [40]. In particular, transfer learning enables a model that is trained on animal-derived data (e.g., predicting gene-phenotype relationships in mice with induced disease) to be adapted for inferring insights in humans [41, 17]. These approaches often rely on conserved features, such as orthologous gene expression [41, 37], or unified knowledge graphs that connect animal-derived nodes to human disease nodes [47, 7]. While valuable, this paradigm is still fundamentally focused on using animals to mimic human disease pathology. **Our**

contribution. In contrast, we learn from protective mechanisms that have naturally evolved in certain species [48, 5]. Our model formalizes this approach via a cross-species GNN, co-training on ADR and human data to map resilience phenotypes to their corresponding human disease indications.

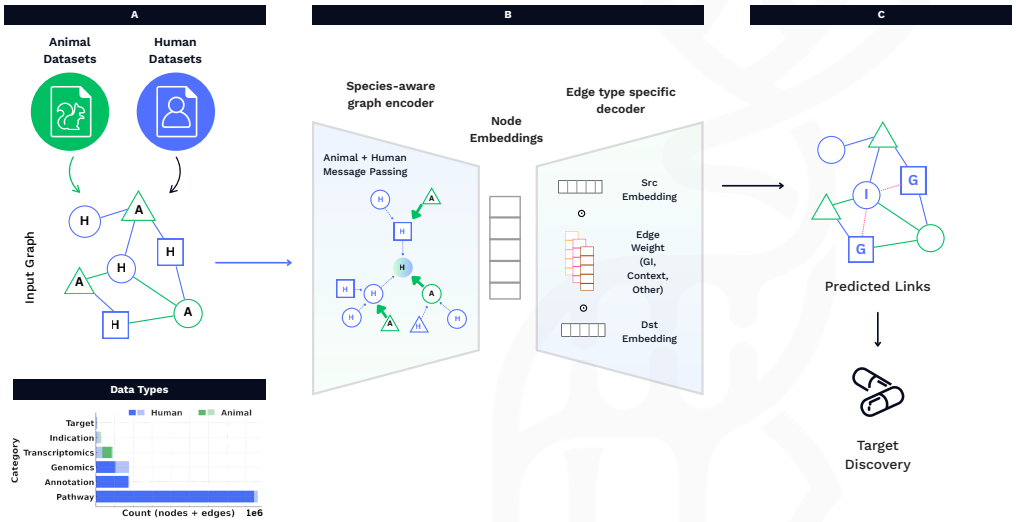

Figure 1: **Overview of CENTAURGNN.** (A) We construct the CENTAUR knowledge graph that captures both ADR data and large-scale human data. (B) The architecture of CENTAURGNN consists of a species-aware encoder that integrates cross-species knowledge and an edge type specific decoder for heterogeneous link prediction. (C) Learning on the CENTAUR knowledge graph, CENTAURGNN enables genome-wide *in silico* screening of candidate therapeutic targets.

3 Methods

We introduce CENTAURGNN, a cross-species graph neural network (GNN) that leverages species-specific message passing to integrate animal disease resistance (ADR) data and human genomic data for novel target discovery. In this section, we define CENTAURGNN's architecture, including the novel species-specific message passing mechanism and objective function, that captures cross-species biological mechanisms to enable ADR-to-human transfer learning (Figure 1).

3.1 Species-aware graph encoder

CENTAURGNN learns a species-aware graph encoder to generate embeddings that capture ADR- and human-derived data. It leverages graph attention neural networks (GAT) [56]. Given a graph $\mathcal{G} = (V, H{+}A)$, CENTAURGNN learns an embedding $\mathbf{h}_v^l$ at each layer l for node v by aggregating messages from its neighboring nodes $u \in \mathcal{N}_v$ via edges derived from cross-species $s \in \{H, A\}$ data:

$$\mathbf{h}_v^{l+1} = \sigma\Big(\sum_{s \in \{H,A\}} \mathbf{m}_{vu}^{(s,l)} \Big) \tag{1}$$

where messages $\mathbf{m}^{(s,l)}$ are propagated via attention weights $\alpha_{vu}^{(i,l)}$ computed for each neighbor u connected to v by species-specific relation $i \in r_s$. For notation, we omit l from $\mathbf{W}$ unless needed.

$$\mathbf{m}_{vu}^{(s,l)} = \sum_{i \in r_s} \sum_{u \in \mathcal{N}_v^{(i)}} \Big(\alpha_{vu}^{(i,l)} \mathbf{W}_i^{tgt} \mathbf{h}_u^l \Big) \tag{2}$$

$$\alpha_{vu}^{(i,l)} = \underset{u \in \mathcal{N}_v^{(i)}}{\text{softmax}}\ e_{vu}^{(i,l)} \tag{3}$$

$$e_{vu}^{(i,l)} = \mathbf{a}_i^{l\top}\ \text{ACT}\Big(\big[\mathbf{W}_i^{src} \mathbf{h}_v^l \| \mathbf{W}_i^{tgt} \mathbf{h}_u^l \| \mathbf{W}_i^{edge} \mathbf{e}_{vu} \big] \Big) \tag{4}$$

where $\mathbf{W}^{src}$, $\mathbf{W}^{tgt}$, and $\mathbf{W}^{edge}$ refer to weight matrices for source nodes, target nodes, and edges, respectively. $\mathbf{W}$ and $\mathbf{a}_i$ are learned. ACT refers to the activation function. Features of edges

$\mathbf{e}_{vu}$ and nodes $\mathbf{h}_v^0$ are extracted from the original databases (Section 4.1). The edge type specific message propagation is aggregated via summation to generate final node embeddings $\mathbf{z}_v$ that capture cross-species biomedical knowledge. The number of GNN layers L and attention heads K are hyperparameters; here, $L = 2$ and $K = 8$. Following the first GNN layer is a BatchNorm layer [22] for normalization, a LeakyReLU activation, and a Dropout layer to mitigate overfitting; only a BatchNorm layer follows the second GNN layer.

3.2 Heterogeneous edge type specific decoder

For self-supervised learning via heterogeneous link prediction, CENTAURGNN adapts an edge type specific DistMult [59] decoder:

$$\mathbf{z}_{(v,r,u)} = \mathbf{z}_v^T \mathbf{Q}_r \mathbf{z}_u \tag{5}$$

where $\mathbf{z}_v$ is the embedding of the source node v, $\mathbf{z}_u$ is the embedding of the target node u, and $\mathbf{Q}_r$ is a trainable edge type specific weight matrix for relation r. The resulting embedding $\mathbf{z}_{(v,r,u)}$ is fed into a two-layer multilayer perceptron (MLP), with ReLU activation and dropout layers in between:

$$\hat{y} = \sigma\Big(\mathbf{U}_2(\mathbf{U}_1\mathbf{z}_{(v,r,u)} + \mathbf{b}_1) + \mathbf{b}_2\Big) \tag{6}$$

where $\mathbf{U}_1, \mathbf{U}_2, \mathbf{b}_1, \mathbf{b}_2$ are learned. The output is a scalar logit $\hat{y}$ indicating whether the edge exists. The number of layers in the MLP is a hyperparameter. To transfer knowledge between different sources of gene-indication associations (e.g., based on ADR data or human population-level genome-wide association studies), one learnable weight matrix for gene-indication associations is shared.

3.3 Objective function

CENTAURGNN is trained by optimizing a composite objective function that simultaneously learns to predict gene-indication associations $\mathcal{L}_{\text{GI}}$ as well as the underlying topology of the knowledge graph ($\mathcal{L}_{\text{CONTEXT}}$ and $\mathcal{L}_{\text{OTHER}}$). $\mathcal{L}_{\text{CONTEXT}}$ focuses on the biological knowledge that provides relevant contextual information to perform target discovery (e.g., gene-pathway and gene-phenotype relationships), and $\mathcal{L}_{\text{OTHER}}$ encompasses the remaining biological knowledge that indirectly contributes to target discovery (e.g., disease ontology). It adapts binary cross entropy loss for $\mathcal{L}$:

$$\mathcal{L}_{\text{TOTAL}} = \lambda_1\big(\mathcal{L}_{\text{GI}}^A + \mathcal{L}_{\text{GI}}^H\big) + \lambda_2\big(\mathcal{L}_{\text{CONTEXT}}^A + \mathcal{L}_{\text{CONTEXT}}^H\big) + (1 - \lambda_1 - \lambda_2)\,\mathcal{L}_{\text{OTHER}}^H \tag{7}$$

where λ_1 and λ_2 are hyperparameters. For CENTAURGNN, $\lambda_1 = 0.79$ and $\lambda_2 = 0.2$. The ratio of positive and negative examples for self-supervised link prediction is 1:1.

4 Experimental Setup

4.1 Datasets

CENTAUR is a large-scale multi-species knowledge graph that combines human genomic and clinical data with high-resolution animal disease resistance (ADR) data (Figure 1a). It is constructed by integrating data from a wide array of sources, which can be broadly categorized into two groups: (1) publicly available databases of human pathway, genomics, and disease transcriptomics data, and (2) proprietary transcriptomic datasets derived from ADR (primarily the thirteen-lined ground squirrel, *Ictidomys tridecemlineatus*). Refer to Appendix 7.1 for data statistics on CENTAUR.

We perform the following **data processing** steps to harmonize gene identifiers between the proprietary animal data and the external human disease datasets. Animal genes are mapped to human orthologs, and human Ensembl gene stable identifiers are used as the index for any subsequent graph operations. Disease indications are mapped via a combination of International Classification of Diseases (ICD-10) codes, MONDO terms, and Unified Medical Language System (UMLS) disease identifiers. Genetic variants are indexed based on rs identifiers (or reference SNP cluster IDs).

Public human datasets on human genomics and disease form the backbone of CENTAUR. DisGeNET [46], ClinVar [28], GeneBass [25], and Cortellis [8] provide edges that capture curated and literature-derived associations between genes, genetic variants, and human diseases and traits. Genome- and phenome-wide association studies and PheWAS [54] connect genetic variants to phenotypes and clinical outcomes. To capture functional context, we include protein–protein interactions and pathway information. MONDO [55] is used to standardize disease and indication terms.

Proprietary ADR datasets help bridge the translational gap by allowing the model to learn from biological signals that reflect resistance and repair to pathologies similar to human clinical phenotypes. ADR phenotypes are nodes in CENTAUR and are linked to human indications. We incorporate differential gene expression data from internal studies (ADR DE Gene I) and gene co-expression modules from Weighted Correlation Network Analysis (WGCNA) [29] performed on various animal tissues; these datasets link genes to specific ADR phenotypes and biological states. CENTAUR is enriched with high confidence gene-indication links via internal curation (ADR Gene-Indication).

4.2 Baseline, ablations, and implementation details

To evaluate the performance of CENTAURGNN and understand the contribution of its components, we compare several model configurations. We perform ablation studies by training CENTAURGNN on different subsets of the CENTAUR knowledge graph: Animal+Human (i.e., curated subset of non-human genomics animal and human data sources), HumanOnly (i.e., curated subset of human non-genomics data), and AnimalOnly (i.e., curated subset of ADR data). We additionally evaluate CENTAURGNN against a transformer-based model that leverages only CENTAUR's node features (i.e., without the relational structure) to measure the performance gained by explicitly modeling graph topology. We also compare against a domain specific model, Exomiser [50] (Appendix 8.4).

Refer to Appendix 7.2-7.4 for implementation details (e.g., model architecture, hyperparameters).

5 Results

We evaluate CENTAURGNN on four key questions. **R1:** To what extent does incorporating ADR or large-scale human data affect predictive performance? **R2:** What components of CENTAUR contribute the most to nominating therapeutic targets? **R3:** What is the benefit of ADR data for predicting novel drug targets? **R4:** What is the clinical translatability potential and generalizability across multiple disease indications of CENTAURGNN's nominated novel targets?

5.1 R1: Benchmarking models that leverage animal disease resistance data

To quantify the effects of incorporating animal disease resistance (ADR) data for predicting therapeutic targets, we evaluate CENTAURGNN trained on different subsets of data: human data only, ADR data only, and combined ADR data with human data. These data subsets are subgraphs of CENTAUR that contain biological knowledge that provides relevant contextual information to perform target discovery. They are curated to isolate the contributions of ADR data versus human data. CENTAURGNN models that leverage ADR data, namely CENTAURGNN (AnimalOnly) and CENTAURGNN (Animal+Human), achieve the strongest performance across all metrics (Table 1). We find that **integrating both ADR data and large-scale human data leads to improvements in all performance metrics compared to modeling human data alone** (Table 1).

We also isolate the contribution of the species-specific message passing by comparing CENTAURGNN models against a transformer architecture. Compared to Transformer (Animal+Human), CENTAURGNN (Animal+Human) achieves 4.2% higher ROC, indicating the significant benefit of our cross-species message propagation mechanism (Table 1). Further, we compare the predictive performance of CENTAURGNN and the transformer architecture trained on the full CENTAUR knowledge graph. We find that CENTAURGNN achieves stronger performance than the transformer architecture in all metrics (Table 2). Notably, CENTAURGNN achieves a 5.0% higher ROC than the transformer baseline (Table 2). These **performance gaps underscore the benefit of leveraging ADR data in addition to large-scale human data**, as well as the impact of CENTAURGNN's **species-specific message propagation and objective function that enable the cross-species data integration**.

5.2 R2: Attribution analysis on the contributions of animal disease resistance data

To understand which components of CENTAUR are most influential to CENTAURGNN's therapeutic target predictions, we perform attribution analysis using Captum to calculate feature importance scores. We focus on the attributions for the following indications, chosen due to their relevance to the physiological adaptations of our primary animal model, the 13-lined ground squirrel (*Ictidomys tridecemlineatus, 13LGS*): kidney disease, myocardial reperfusion, obesity, ischemia, and Alzheimer's

Table 1: **Benchmarking** performance in predicting gene-indication associations after being trained on different subsets of the CENTAUR knowledge graph. These data subsets are subgraphs of CENTAUR, curated to isolate the contributions of ADR data versus human data. Models are trained on 3 seeds.

Model	ROC	Accuracy	F1	Precision
CENTAURGNN (HumanOnly)	0.9226 ± 0.0003	0.8453 ± 0.0003	0.8543 ± 0.0005	0.9103 ± 0.0003
CENTAURGNN (AnimalOnly)	0.9904 ± 0.0000	0.9530 ± 0.0001	0.9534 ± 0.0002	0.9891 ± 0.0000
CENTAURGNN (Animal+Human)	0.9935 ± 0.0001	0.9708 ± 0.0001	0.9711 ± 0.0000	0.9909 ± 0.0001
Transformer (Animal+Human)	0.9517 ± 0.0001	0.8935 ± 0.0001	0.8962 ± 0.0002	0.9397 ± 0.0002

Table 2: **Benchmarking** performance on predicting gene-indication associations after being trained on the full CENTAUR knowledge graph. Models are trained on 3 seeds.

Model	ROC	Accuracy	F1	Precision
CENTAURGNN	0.9512 ± 0.0001	0.8696 ± 0.0001	0.8692 ± 0.0001	0.9476 ± 0.0001
Transformer	0.9010 ± 0.0001	0.8221 ± 0.0002	0.8139 ± 0.0002	0.9061 ± 0.0001

disease. We compare the attribution scores of animal versus human data in CENTAURGNN, trained on the full CENTAUR knowledge graph, and CENTAURGNN (Animal+Human) (Appendix 8.5).

The **mean attribution scores of ADR data are consistently at least an order of magnitude higher than those of human data** (Table 3; sum-aggregated scores are in Appendix 8.5). In CENTAURGNN, the average animal-derived node is approximately 35 times more influential than the average human-derived node; similarly, the average animal-derived edge has 25 times higher influence (Appendix Table 17). Further, the average animal-derived edge and node are about 119 times and 23 times more influential, respectively, in CENTAURGNN (Animal+Human) (Appendix Table 18).

Table 3: Mean **attribution** scores of CENTAURGNN (trained on the full CENTAUR knowledge graph) and CENTAURGNN (Animal+Human) for animal- or human-derived graph components.

Data Origin	Component Type	CENTAURGNN Mean Attribution	CENTAURGNN (A+H) Mean Attribution
Animal	Edges	8.765×10^{-2}	7.648×10^{-2}
Human	Edges	3.546×10^{-3}	6.448×10^{-4}
Animal	Nodes	5.806×10^{-5}	2.371×10^{-5}
Human	Nodes	1.680×10^{-6}	1.036×10^{-6}

Notably, **ADR data indicate high feature importance despite only making up 0.35% of nodes and 5.90% of edges** in CENTAUR (Figure 2). Of the nodes and edges that contribute to the CENTAURGNN model's predictions of therapeutic targets, 13.7% and 25.3% are animal-derived nodes and edges, respectively. In the CENTAURGNN (Animal+Human) model, **edges derived from animal studies have higher feature importance (57.7%) than those of humans** despite only making up 1.6% of the edges in the Animal+Human subset of CENTAUR (Appendix Figure 3b). Our attribution analyses demonstrate the substantial benefits of leveraging ADR data for target discovery.

5.3 R3: Contribution of animal disease resistance data to model generalizability

To approximately quantify the translational value of ADR data, we evaluate CENTAURGNN models' ability to generalize to unseen, out-of-distribution gene-indications. Concretely, all gene-indication edges from the GeneBass dataset, which summarizes the exome-based genetic analyses on ∼400,000 participants in the UK Biobank, are held out from training. We compare the performance of

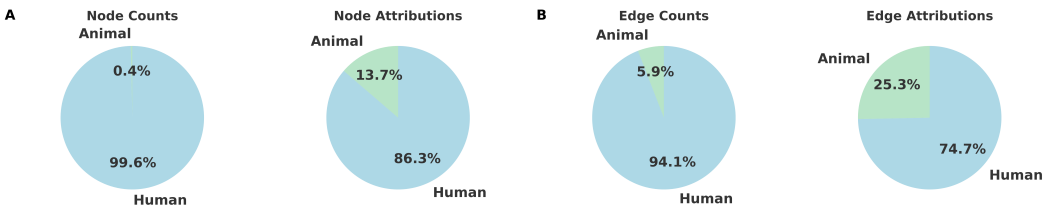

Figure 2: Counts of (A) nodes or (B) edges compared to their corresponding overall contribution (or **attribution**) to predictions of the CENTAURGNN model. Refer to Appendix Figure 3 for this attribution analysis on the CENTAURGNN (Animal+Human) model.

CENTAURGNN and baselines trained to predict gene-indication associations from other sources excluding human genomics data on zero-shot prediction of unseen associations from GeneBass.

Our results demonstrate the benefits of incorporating ADR data for improving model generalizability (Table 4). CENTAURGNN (Animal+Human) achieves the highest performance across nearly all metrics, with an average ROC of 0.8636. Notably, it outperforms CENTAURGNN (HumanOnly) by almost 9% in ROC. Interestingly, CENTAURGNN (AnimalOnly), a model that is trained only on ADR data, performs better than CENTAURGNN (HumanOnly) by nearly 6% in ROC. This suggests that **ADR data is informative for inferring therapeutic targets that may be effective in humans**. The performance gains are likely due to the transcriptomic and gene-phenotype associations in the ADR data, which provide **genomic evidence that support generalizability to unseen gene-indication associations derived from human genomics**. Such an advantage is particularly evidenced by the improvement over CENTAURGNN (HumanOnly), which is trained on human non-genomics data. The additional performance lift by CENTAURGNN models trained on both ADR data and human data underscores their synergistic effects. Further, ablating the species-specific message passing mechanism, as shown by the transformer baseline, leads to a substantial decrease in all performance metrics. For instance, the Transformer(Animal+Human) baseline yields a 52% decrease in ROC compared to CENTAURGNN (Animal+Human). This suggests that **cross-species data integration via CENTAURGNN's species-specific message passing improves model generalizability**.

Table 4: Model **generalizability on predicting out-of-distribution targets** (i.e., gene–indication associations). The threshold (Youden's J Statistic) for CENTAURGNN (HumanOnly), CENTAURGNN (AnimalOnly), CENTAURGNN (Animal+Human), and Transformer (Animal+Human) are 0.7109, 0.6982, 0.3391, and 0.4739, respectively. Models are trained on 3 seeds.

Model	ROC	Accuracy	F1	Precision
CENTAURGNN (HumanOnly)	0.7739 ± 0.0008	0.7186 ± 0.0007	0.6924 ± 0.0006	0.7412 ± 0.0009
CENTAURGNN (AnimalOnly)	0.8311 ± 0.0005	0.7664 ± 0.0013	0.7877 ± 0.0021	0.7756 ± 0.0017
CENTAURGNN (Animal+Human)	0.8636 ± 0.0002	0.8245 ± 0.0001	0.8433 ± 0.0005	0.7756 ± 0.0006
Transformer (Animal+Human)	0.4161 ± 0.0001	0.5001 ± 0.0000	0.6667 ± 0.0000	0.4314 ± 0.0002

5.4 R4: Evaluating the clinical translatability potential of nominated novel predictions using human genetic evidence

Drug targets supported by human genetic evidence are predicted to be 2.6 times more likely to succeed in clinical development compared to those without such support [38]. We evaluate the contribution of ADR data on the human relevance and clinical translatability of predicted novel targets. Concretely, we compare the novel predictions generated by CENTAURGNN (Animal+Human) and CENTAURGNN (HumanOnly) in the GeneBass exome-based genetic analyses of nearly 400,000

participants from the UK Biobank [26, 6]. Novel target predictions are defined as those surpassing the Youden's J Statistic thresholds (Table 4), which are computed on the validation set of each model.

Our analysis focuses on five indications: kidney disease, myocardial reperfusion, obesity, ischemia, and Alzheimer's disease. For each indication, we aggregate the corresponding UK Biobank phenocodes to define testable human disease-relevant endpoints (Appendix Table 19). In brief, **for each indication-specific novel prediction, we evaluate gene-based association tests in 394,841 exome-sequenced UK Biobank participants** [26], considering both the predicted loss-of-function (pLOF) and missense variant masks. A prediction is "validated" to have human genetic support if it shows a significant association ($p < 0.05$) in any of the gene-based burden, SKAT-O, or SKAT tests for at least one UK Biobank phenocode mapped to the relevant indication.

For all five indications, CENTAURGNN (Animal+Human) produces a greater number of validated novel predictions compared to CENTAURGNN (HumanOnly) (Table 5). On average across the five indications, 448 validated predictions are unique to CENTAURGNN (Animal+Human) (Table 5). These results demonstrate that **integrating ADR data produces novel targets that would not have been prioritized** by CENTAURGNN (HumanOnly), and even **prioritizes a larger pool of novel targets with clinical translatability potential across multiple disease contexts**.

Table 5: Validation of **novel therapeutic target predictions** for five indications by CENTAURGNN (Animal+Human) and CENTAURGNN (HumanOnly) using human genetics evidence from nearly 400,000 UK Biobank exome sequenced participants. "Total" indicates the total number of novel targets predicted, a subset of which are validated by human evidence ("Validated").

Indication	**Animal+Human**		**HumanOnly**		**Animal+Human & not in HumanOnly**
	Total	**Validated**	**Total**	**Validated**	**Validated**
Kidney disease	176	167	40	38	**160**
Myocardial reperfusion	224	211	0	0	**211**
Obesity	679	628	7	7	**624**
Ischemia	628	590	0	0	**590**
Alzheimer's disease	914	660	7	6	**657**

6 Conclusion

We have addressed a fundamental limitation in drug discovery of over-relying on data from disease states. We introduce CENTAURGNN, a cross-species GNN that learns from mechanisms of ADR. By integrating these unique biological signals with large-scale human data, CENTAURGNN **effectively translates insights from naturally protected animals into human therapeutic contexts**.

Our comprehensive experiments support two central hypotheses: (1) ADR data provides complementary signals for target discovery, as confirmed through benchmarking and feature attribution analysis; and (2) a specialized graph-based architecture is essential for integrating these disparate data sources. We show that CENTAURGNN's species-specific message passing outperforms transformer-based baselines, highlighting the importance of explicitly modeling biological relationships. Further, we demonstrate the impact of translating insights from animal to human contexts through CENTAURGNN's ability to **nominate on average 448 more novel drug targets, validated with independent human genetics evidence, than models relying on human data alone**.

Broadly, our work establishes a **framework for a new class of target discovery models**, demonstrating that specialized cross-species graph neural networks enable the systematic mining of the evolved disease resistance and repair capabilities of the animal kingdom. Ultimately, our findings indicate that by shifting the focus from modeling disease to learning from resilience, we can unlock a new source of therapeutic targets with strong translational potential.

References

[1] J. N. Acosta, G. J. Falcone, P. Rajpurkar, and E. J. Topol. Multimodal biomedical AI. *Nature Medicine*, 28:1773–1784, 2022.

[2] E. Alsentzer, M. M. Li, S. N. Kobren, A. Noori, the Undiagnosed Diseases Network, I. S. Kohane, and M. Zitnik. Few shot learning for phenotype-driven diagnosis of patients with rare genetic diseases. *npj Digital Medicine*, 8(1):Article 380, 2025.

[3] Thomas Arendt, Jens Stieler, Arjen M. Strijkstra, Roelof A. Hut, Jan Rüdiger, Eddy A. Van der Zee, Tibor Harkany, Max Holzer, and Wolfgang Härtig. Reversible paired helical filament-like phosphorylation of tau is an adaptive process associated with neuronal plasticity in hibernating animals. *Journal of Neuroscience*, 23(18):6972–6981, 2003.

[4] Meet Barot, Vladimir Gligorijević, Kyunghyun Cho, and Richard Bonneau. Netquilt: deep multispecies network-based protein function prediction using homology-informed network similarity. *Bioinformatics*, 37(16):2414–2423, 2021.

[5] A. Bonis, L. Anderson, G. Talhouarne, et al. Cardiovascular resistance to thrombosis in 13-lined ground squirrels. *Journal of Comparative Physiology B*, 189:167–177, 2019.

[6] Clare Bycroft, Colin Freeman, Desislava Petkova, and et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[7] P. Cacheiro, D. Pava Mejia, H. Parkinson, et al. Computational identification of disease models through cross-species phenotype comparison. *Disease Models & Mechanisms*, 17(6):dmm050604, 2024.

[8] Clarivate Analytics. Cortellis drug discovery intelligence. `https://clarivate.com/life-sciences-healthcare/research-development/discovery-development/cortellis-pre-clinical-intelligence/`, 2024. Accessed January 31, 2024.

[9] Hejie Cui, Jiaying Lu, Ran Xu, Shiyu Wang, Wenjing Ma, Yue Yu, Shaojun Yu, Xuan Kan, Chen Ling, Liang Zhao, Zhaohui S. Qin, Joyce C. Ho, Tianfan Fu, Jing Ma, Mengdi Huai, Fei Wang, and Carl Yang. A review on knowledge graphs for healthcare: Resources, applications, and promises. *Journal of Biomedical Informatics*, 169:104861, 2025.

[10] C. de Veij Mestdagh, T. Bulavintseva, L. Zeef, et al. Torpor induces reversible tau hyperphosphorylation and accumulation in mice expressing human tau. *Acta Neuropathologica Communications*, 12:112, 2024.

[11] K. D. S. Fernald, P. C. Förster, E. Claassen, and L. H. M. van de Burgwal. The pharmaceutical productivity gap—incremental decline in R&D efficiency despite transient improvements. *Drug Discovery Today*, 29(11):104160, 2024.

[12] Gregory L. Florant and Jessica E. Healy. The regulation of food intake in mammalian hibernators: a review. *Journal of Comparative Physiology B*, 182(4):451–467, 2012.

[13] Yang Gao, Xiang Zhang, Zhongquan Sun, Payal Chandak, Jiajun Bu, and Haishuai Wang. Precision adverse drug reactions prediction with heterogeneous graph neural network. *Advanced Science*, 12(4):e2404671, 2024.

[14] T. Gaudelet, B. Day, A. R. Jamasb, et al. Utilizing graph machine learning within drug discovery and development. *Briefings in Bioinformatics*, 22(6):bbab159, 2021.

[15] A. Grover and J. Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 855–864, 2016.

[16] W. Hamilton, Z. Ying, and J. Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1024–1034, 2017.

[17] J. Han, H. Zhang, and K. Ning. Techniques for learning and transferring knowledge for microbiome-based classification and prediction: review and assessment. *Briefings in Bioinformatics*, 26(1):bbaf015, 2025.

[18] T. Hartung. The (misleading) role of animal models in drug development. *Frontiers in Drug Discovery*, 2024. Article 1355044.

[19] Robert Hoehndorf, Paul N. Schofield, and Georgios V. Gkoutos. Phenomenet: a whole-phenome approach to disease gene discovery. *Nucleic Acids Research*, 39(18):e119, 2011.

[20] Kexin Huang, Payal Chandak, Sarah Gady, and et al. A foundation model for clinician-centered drug repurposing. *Nature Medicine*, 30(12):3601–3613, 2024.

[21] B. V. Ineichen, E. Furrer, S. L. Grüninger, W. E. Zürrer, and M. R. Macleod. Analysis of animal-to-human translation shows that only 5% of animal-tested therapeutic interventions obtain regulatory approval for human applications. *PLOS Biology*, 22(6):e3002667, 2024.

[22] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, pages 448–456. PMLR, 2015.

[23] Swati Jain and Alkesh Jani. Renal tubular cells from hibernating squirrels are protected against cisplatin induced apoptosis. *International Journal of Nephrology*, 2020:6313749, 2020.

[24] Ruth Johnson, Michelle M. Li, Ayush Noori, Owen Queen, and Marinka Zitnik. Graph AI in medicine. *Annual Review of Biomedical Data Science*, 7:345–368, 2024.

[25] K. J. Karczewski, M. Solomonson, K. R. Chao, J. K. Goodrich, G. Tiao, et al. Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK biobank exomes. *Cell Genomics*, 2(9):100168, 2022.

[26] Konrad J. Karczewski, Miles Solomonson, Keith R. Chao, and et al. Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 uk biobank exomes. *Cell Genomics*, 2(9):100168, 2022.

[27] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[28] M. J. Landrum et al. Clinvar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 44:D862–D868, 2016.

[29] P. Langfelder and S. Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC Bioinformatics*, 9:559, 2008.

[30] Lechuan Li, Ruth Dannenfelser, Yu Zhu, Nathaniel Hejduk, Santiago Segarra, and Vicky Yao. Joint embedding of biological networks for cross-species functional alignment. *Bioinformatics*, 39(9):btad529, 2023.

[31] M. M. Li, K. Huang, and M. Zitnik. Graph representation learning in biomedicine and healthcare. *Nature Biomedical Engineering*, 6:1353–1372, 2022.

[32] Michelle M. Li, Yadi Huang, Manuja Sumathipala, Ming Q. Liang, Alberto Valdeolivas, Ashwin N. Ananthakrishnan, Katherine Liao, Daniel Marbach, and Marinka Zitnik. Contextual AI models for single-cell protein biology. *Nature Methods*, 21(8):1546–1557, 2024.

[33] Mengdi Liu, Zhangyang Gao, Hong Chang, Stan Z. Li, Shiguang Shan, and Xilin Chen. G2pdiffusion: Genotype-to-phenotype prediction with diffusion models, 2025.

[34] A. Loewa, J. J. Feng, and S. Hedtrich. Human disease models in drug development. *Nature Reviews Bioengineering*, 1:545–559, 2023.

[35] Y. Luo et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nature Communications*, 8:573, 2017.

[36] Nil Mamano and Wayne B. Hayes. Sana: simulated annealing far outperforms many other search algorithms for biological network alignment. *Bioinformatics*, 33(14):2156–2164, 2017.

[37] C. A. Mancuso, K. A. Johnson, R. Liu, and A. Krishnan. Joint representation of molecular networks from multiple species improves gene classification. *PLOS Computational Biology*, 20(1):e1011773, 2024.

[38] Eric Vallabh Minikel, Jeffery L. Painter, Coco Chengliang Dong, Matthew R. Nelson, Frank R. Day, Adam S. Butterworth, Robert M. Plenge, and et al. Refining the impact of genetic evidence on clinical success. *Nature*, 629(8012):624–629, 2024.

[39] Muhammad Muneeb, Shu Feng, and Andreas Henschel. Transfer learning for genotype–phenotype prediction using deep learning models. *BMC Bioinformatics*, 23:444, 2022.

[40] Rachelly Normand, Wenfei Du, Mayan Briller, Renaud Gaujoux, Elina Starosvetsky, Amit Ziv-Kenet, Gali Shalev-Malul, Robert J. Tibshirani, and Shai S. Shen-Orr. Found in translation: a machine learning model for mouse-to-human inference. *Nature Methods*, 15(12):1067–1073, 2018.

[41] Y. Park, N. P. Muttray, and A.-C. Hauschild. Species-agnostic transfer learning for cross-species transcriptomics data integration without gene orthology. *Briefings in Bioinformatics*, 25(2):bbae004, 2024.

[42] Yeqi Park, Niklas P. Muttray, and Anne-Christin Hauschild. Species-agnostic transfer learning for cross-species transcriptomics data integration without gene orthology. *Briefings in Bioinformatics*, 25(2):bbae004, 2024.

[43] G. Pei, A. Balkema-Buschmann, and A. Dorhoi. Disease tolerance as immune defense strategy in bats: One size fits all? *PLoS Pathogens*, 20(9):e1012471, 2024.

[44] B. Perozzi, R. Al-Rfou, and S. Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 701–710, 2014.

[45] Rodrigo Petegrosso, Jisoo Park, and TaeHyun Hwang. Transfer learning across ontologies for phenome–genome association prediction. *Bioinformatics*, 33(4):529–536, 2017.

[46] J. Piñero et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, 48:D845–D855, 2020.

[47] T. E. Putman, K. Schaper, N. Matentzoglu, et al. The monarch initiative in 2024: an analytic platform integrating phenotypes, genes and diseases across species. *Nucleic Acids Research*, 52(D1):D938–D949, 2024.

[48] M. W. Saxton, B. W. Perry, V. B. Fedorov, et al. Serum plays an important role in reprogramming the seasonal transcriptional profile of brown bear adipocytes (reversible insulin resistance during hibernation). *iScience*, 25(10):105084, 2022.

[49] Wenxiang Shi, Xixia Zhao, Jin Jin, Yu Liu, Lei Song, Xiaoping Yu, and Zhaoxiang Yang. Dapbch: a disease association prediction model based on cross-species heterogeneous networks. *Frontiers in Genetics*, 14:1222346, 2023.

[50] Damian Smedley, Julius O. B. Jacobsen, Markus Jäger, Sebastian Köhler, Manuel Holtgrewe, Max Schubach, Enrico Siragusa, Tomasz Zemojtel, Orion J. Buske, Nicole L. Washington, Melissa A. Haendel, and Peter N. Robinson. Next-generation diagnostics and disease-gene discovery with the exomiser. *Nature Protocols*, 10(12):2004–2015, 2015.

[51] Bing Su, Xiaojun Wang, Kelly L. Drew, George Perry, Mark A. Smith, and Xiongwei Zhu. Physiological regulation of tau phosphorylation during hibernation. *Journal of Neurochemistry*, 105(6):2098–2108, 2008.

[52] P. F. Sullivan, J. R. S. Meadows, S. Gazal, Zoonomia Consortium, et al. Leveraging base-pair mammalian constraint to understand genetic variation and human disease. *Science*, 380(6643):eabn2937, 2023.

[53] D. Sun, W. Gao, H. Hu, and S. Zhou. Why 90% of clinical drug development fails and how to improve it? *Acta Pharmaceutica Sinica B*, 12(7):3049–3062, 2022.

[54] Vanderbilt University Medical Center. Phewas catalog (rrid:scr_003562). https://phewascatalog.org, 2022. Accessed December 16, 2022.

[55] N. A. Vasilevsky et al. Mondo: unifying diseases for the world, by the world. *medRxiv*, 2022. doi:10.1101/2022.10.12.22280969.

[56] P. Veličković et al. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018.

[57] H. Weavers. Biological resilience in health and disease. *Disease Models & Mechanisms*, 17(7):dmm050799, 2024.

[58] Aaron Wenteler, Claudia P. Cabrera, Wei Wei, Victor Neduva, and Michael R. Barnes. Ai approaches for the discovery and validation of drug targets. *Cambridge Prisms: Precision Medicine*, 2:e7, 2024.

[59] B. Yang et al. Embedding entities and relations for learning and inference in knowledge bases. In *International Conference on Learning Representations (ICLR)*, 2015.

[60] M. Zitnik, M. Agrawal, and J. Leskovec. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34:i457–i466, 2018.

7 Appendix: Data and Methods

7.1 CENTAUR Knowledge Graph

We evaluate CENTAURGNN on the CENTAUR knowledge graph, including subsets of the data for various ablation analyses (Tables 6-8).

Table 6: Node types and counts of the CENTAUR knowledge graph. “Source" indicates whether the node is derived from animal or human data.

Node type	Counts	Source	Category
Variant	211,992	Human	Genomics
Metabolite Pathway	48,655	Human	Pathway
Gene Pathway	30,481	Human	Pathway
Indication	29,900	Human	Indication
Protein	19,909	Human	Target
Gene	19,822	Human	Target
Publication	4578	Human	Annotation
Indication Pathway	2489	Human	Pathway
Trait	2107	Human	Indication
ADR WGCNA	1229	Animal	Transcriptomics
Genomic Phenotype	1186	Human	Genomics
ADR Phenotype	76	Animal	Indication

Table 7: Edge types and counts of the CENTAUR knowledge graph. “Source" indicates whether the node is derived from animal or human data.

Edge type	Counts	Species	Category
Gene Pathway	3,793,726	Human	Pathway
Protein Interactions I	710,561	Human	Annotation
Protein Pathway	303,045	Human	Pathway
DisGeNet Variant	259,706	Human	Genomics
Gene in ADR WGCNA Module	253,790	Animal	Transcriptomics
Genomic Publication	148,945	Human	Genomics
Trait Association	145,230	Human	Genomics
Indication Pathway	116,927	Human	Pathway
Disease Transcriptomics	88,629	Human	Transcriptomics
DisGeNet Gene	76,022	Human	Annotation
ADR Gene-Indication	72,362	Animal	Indication
Protein Interactions II	61,719	Human	Annotation
Genomic Association	60,692	Human	Genomics
GeneBass	53,881	Human	Genomics
Cortellis	31,553	Human	Indication
ADR DE Gene I	25,240	Animal	Transcriptomics
Gene Annotation	19,859	Human	Annotation
ADR DE Gene II	12,687	Animal	Transcriptomics
Indication Association	11,334	Human	Indication
MONDO	10,113	Human	Pathway
ADR GRN	9519	Animal	Pathway
Clinvar	9136	Human	Annotation
Indication Interactions	7875	Human	Annotation
Variant Association	3246	Human	Genomics
DisGeNET enrichment for ADR WGCNA Module	2189	Animal	Pathway
Variant Phenotype	1118	Human	Genomics
Variant	1096	Human	Genomics
Publication	751	Human	Annotation
Indication Mapping for ADR Phenotype	49	Animal	Indication

Table 8: Node and edge counts of the input graph for each model configuration (i.e., full CENTAUR knowledge graph, Animal+Human, AnimalOnly, or HumanOnly). Animal- or human-derived edges are subsets of the total edges in CENTAUR

Configuration	Total nodes	Total edges	Animal-derived edges	Human-derived edges
Full	372,424	6,346,903	364,240	5,982,663
Animal+Human	132,510	5,417,983	119,544	5,298,439
AnimalOnly	18,942	364,284	364,284	0
HumanOnly	119,487	5,031,714	0	5,031,714

7.2 Model Architectures and Parameters

Table 9 details the architectural specifications, including the total number of trainable parameters, for CENTAURGNN, baselines, and ablations.

Table 9: Architectural specifications and total parameter counts for all evaluated models. Hidden channels per layer are separated by a comma.

Model	Encoder layers	Decoder layers	Hidden channels	Output channels	Decode channels	Heads	Total Parameters
CENTAURGNN	2	2	64	32	16	8	628842273
CENTAURGNN (Animal+Human)	3	1	96, 64	32	16	4	347302785
CENTAURGNN (AnimalOnly)	3	1	96, 64	32	16	4	103084033
CENTAURGNN (HumanOnly)	3	1	96, 64	32	16	4	237303169
Transformer	4	–	256	–	–	8	85977345
Transformer (Animal+Human)	4	–	256	–	–	4	63337217

7.3 Hyperparameters

All models are optimized via the Adam optimizer. We perform a hyperparameter sweep for each model configuration to select its best hyperparameters (Table 10).

Table 10: Training hyperparameters used for each model configuration. “Neighborhood" refers to the number of neighbors at each layer. “Epochs" refers to the epoch of the best model checkpoint.

Model	Learning rate	Weight decay	Dropout	Neighborhood	λ_1, λ_2	Epochs
CENTAURGNN (HumanOnly)	1.46×10^{-4}	2.32×10^{-5}	0.608	13, 5, 2	0.79, 0.20	72
CENTAURGNN (AnimalOnly)	1.46×10^{-4}	2.32×10^{-5}	0.608	13, 5, 2	0.79, 0.21	149
Transformer (Animal+Human)	2.80×10^{-4}	2.35×10^{-4}	0.367	12, 3, 3	N/A	21
CENTAURGNN (Animal+Human)	1.46×10^{-4}	2.32×10^{-5}	0.608	13, 5, 2	0.79, 0.20	44
Transformer	1.52×10^{-4}	7.70×10^{-5}	0.638	10, 5, 1	N/A	52
CENTAURGNN	2.72×10^{-4}	4.05×10^{-5}	0.713	13, 3	0.79, 0.20	21

7.4 Implementation Details

We construct and analyze the CENTAUR knowledge graph using Neo4j and the py2neo toolkit. CENTAURGNN, baselines, and ablations are developed using Pytorch and Pytorch Geometric.

Models are trained on NVIDIA H100 80 GB (SXM, Hopper architecture). Standard deviation is computed by training models on three seeds: 32, 42, and 52.

8 Appendix: Results

8.1 Benchmarking ADR Models

We apply the one-sided student t-test to show that both **AnimalOnly** and **Animal+Human** models significantly outperform the **HumanOnly** model across all reported metrics (Table 11). The gains indicate that ADR data contributes complementary signal beyond human data alone and that integrating ADR data yields robust improvements in generalization.

Table 11: One-sided Student's t-test: leveraging ADR data significantly improves performance.

Comparison	**Metric**	**p-value**
AnimalOnly $>$ HumanOnly	ROC	2.38×10^{-6}
	Accuracy	9.27×10^{-9}
	F1	7.78×10^{-8}
	Precision	1.84×10^{-6}
Animal+Human $>$ HumanOnly	ROC	3.64×10^{-7}
	Accuracy	9.36×10^{-8}
	F1	2.57×10^{-6}
	Precision	5.86×10^{-7}

8.2 Data size and Model Performance

We compare dataset size (number of nodes and edges) with model performance to assess whether a relationship exists (Table 12). We find that data size does not correlate with performance. For instance, CENTAURGNN (AnimalOnly) is trained on the smallest subset of data, but its performance is ranked between that of CENTAURGNN (Animal+Human) and CENTAURGNN (HumanOnly).

Table 12: Comparing the performance of CENTAURGNN models and the numbers of nodes and edges in the underlying knowledge graph.

Model	**# nodes**	**# edges**	**ROC**	**Accuracy**	**F1**	**Precision**
CENTAURGNN (Animal+Human)	132,510	5,417,983	0.9935 ± 0.0001	0.9708 ± 0.0001	0.9711 ± 0.0000	0.9909 ± 0.0001
CENTAURGNN (AnimalOnly)	20,042	366,314	0.9904 ± 0.0000	0.9531 ± 0.0006	0.9534 ± 0.0002	0.9891 ± 0.0000
CENTAURGNN (HumanOnly)	129,887	5,138,004	0.9226 ± 0.0003	0.8453 ± 0.0003	0.8543 ± 0.0005	0.9103 ± 0.0003

8.3 Loss Function Ablation

We ablate different components of the loss function ($\mathcal{L}_{gi}, \mathcal{L}_{context}, \mathcal{L}_{rest}$) to evaluate the contribution of different subsets of our data (Table 13). We evaluate three ablations: **Ablation I** drops $\mathcal{L}_{context}$ and $\mathcal{L}_{rest}$; **Ablation II** drops $\mathcal{L}_{rest}$; and **Ablation III** drops $\mathcal{L}_{context}$. We find that the full model, CENTAURGNN, outperforms all ablations.

Table 13: Comparing the performance of CENTAURGNN and loss ablations.

Model	ROC	Accuracy	F1	Precision
CENTAURGNN	0.9512 ± 0.0001	0.8696 ± 0.0001	0.8692 ± 0.0001	0.9476 ± 0.0001
Ablation I	0.9374 ± 0.0001	0.8574 ± 0.0002	0.8524 ± 0.0001	0.9329 ± 0.0002
Ablation II	0.9407 ± 0.0001	0.8628 ± 0.0002	0.8578 ± 0.0004	0.9352 ± 0.0001
Ablation III	0.9378 ± 0.0001	0.8584 ± 0.0002	0.8537 ± 0.0002	0.9332 ± 0.0002

8.4 Evaluation Against Domain-Specific Models

We evaluate CENTAURGNN against domain-specific models. After surveying cross-species and related machine learning literature, we find that Exomiser [50] is the only domain-specific model that is appropriate to compare against our approach. It uses model organism phenotype annotations (e.g., mouse, zebrafish) to score the similarity of the input Human Phenotype Ontology (HPO) terms and candidate causal genes (i.e., gene that harbors the variant that causes the disease). While there exist many efforts to model cross-species relationships for biological discovery, they are not comparable to our approach because of at least one of the following:

- **Data mismatch** (e.g., need ortholog/grouped multi-species networks, additional annotations, or IsoRank alignments): GenePlexusZoo [37], ETNA [30], NetQuilt [4].
- **Task mismatch** (e.g., optimize alignment/semantic similarity, not binary gene–indication link prediction): PhenoMeNET [19], SemSimian [47].
- **Interface mismatch** (e.g., produce embeddings/alignments requiring custom link-prediction heads and nonstandard evaluation): DapBCH [49], SATL [42], SANA [36].
- **Modality/domain mismatch** (e.g., single-cell transfer, genotype–image, or intra-human transfer): G2PDiffusion [33], CEU–YRI [39], GO–HPO [45].

To evaluate Exomiser on the indications in our test set, we map the indication IDs (ICD10/UMLS/MONDO) to HPO Phenotypes. Among the 4,713 (out of 9,369) indications that can be mapped to HPO terms, we perform a fair comparison by evaluating both CENTAURGNN and Exomiser on the exact same set of gene-indication pairs. This ensures that differences in performance reflect model capability rather than differences in evaluation data. We construct a test set that spans 4,362 indications and 18,580 genes. **We find that CENTAURGNN substantially outperforms Exomiser across all metrics** (Table 14).

Table 14: Benchmarking CENTAURGNN against domain-specific model, Exomiser.

Model	ROC	Accuracy	F1	Precision
Exomiser	0.6542	0.6133	0.6401	0.6537
CENTAURGNN	0.9550	0.8834	0.8925	0.9026

8.5 Model Attributions

To quantify the overall contribution of each data type, we compute the mean of the summed absolute attributions (Tables 15-16). This metric represents the average total importance per prediction for each feature set (Node/Edge, Human/Animal). We additionally calculate the mean of absolute attribution scores (Tables 17-18), or the average impact of an individual feature normalized by the number of features in each category. We break down these scores by disease indication, reporting the mean across all indications and the fold-difference between human and animal data.

Table 15: Sum attribution scores per indication for predicting gene-indication associations using CENTAURGNN models.

Indication	Abs Node (Animal)	Abs Node (Human)	Abs Edge (Animal)	Abs Edge (Human)
Alzheimer disease	18.66	125.86	45.01	193.14
Ischemia	27.01	117.12	95.23	132.18
Kidney diseases	5.28	38.83	13.04	73.99
Myocardial reperfusion injury	3.41	11.82	6.27	17.53
Obesity	15.68	147.75	46.67	191.76
Mean	14.01	88.28	41.24	121.72
		6x		**3x**

Table 16: Sum attribution scores per indication for predicting gene-indication associations using CENTAURGNN (Animal+Human) models.

Indication	Abs Node (Animal)	Abs Node (Human)	Abs Edge (Animal)	Abs Edge (Human)
Alzheimer's disease	22.59	52.53	65.36	49.62
Ischemia	10.02	26.28	50.31	31.12
Kidney diseases	6.74	5.60	14.54	7.16
Myocardial reperfusion injury	6.48	14.16	40.41	24.58
Obesity	10.22	18.53	32.25	36.50
Mean	11.21	23.42	40.57	29.80
		2x		**0.73x**

Table 17: Mean attribution scores per indication for predicting gene-indication associations using CENTAURGNN models.

Indication	Abs Node (Animal)	Abs Node (Human)	Abs Edge (Animal)	Abs Edge (Human)
Alzheimer's disease	5.64×10^{-5}	1.17×10^{-6}	1.20×10^{-1}	5.62×10^{-3}
Ischemia	1.67×10^{-4}	2.41×10^{-6}	1.98×10^{-1}	3.71×10^{-3}
Kidney diseases	1.40×10^{-5}	2.50×10^{-6}	3.64×10^{-3}	2.24×10^{-3}
Myocardial reperfusion injury	1.49×10^{-5}	4.31×10^{-7}	1.57×10^{-3}	4.13×10^{-4}
Obesity	3.79×10^{-5}	1.89×10^{-6}	1.14×10^{-1}	5.74×10^{-3}
Mean	5.81×10^{-5}	1.68×10^{-6}	8.76×10^{-2}	3.55×10^{-3}
		35x		**25x**

Table 18: Mean attribution scores per indication for predicting gene-indication associations using CENTAURGNN (Animal+Human) models.

Indication	Abs Node (Animal)	Abs Node (Human)	Abs Edge (Animal)	Abs Edge (Human)
Alzheimer's disease	3.56×10^{-5}	1.09×10^{-6}	1.81×10^{-1}	1.46×10^{-3}
Ischemia	1.20×10^{-5}	7.70×10^{-7}	9.32×10^{-2}	4.68×10^{-4}
Kidney diseases	1.54×10^{-5}	1.33×10^{-6}	5.86×10^{-3}	1.80×10^{-4}
Myocardial reperfusion injury	2.38×10^{-5}	1.41×10^{-6}	2.66×10^{-3}	2.23×10^{-4}
Obesity	3.18×10^{-5}	5.81×10^{-7}	9.93×10^{-2}	8.93×10^{-4}
Mean	2.37×10^{-5}	1.04×10^{-6}	7.65×10^{-2}	6.45×10^{-4}
		23x		**119x**

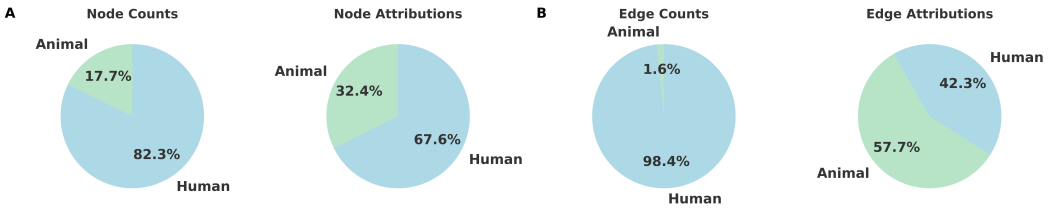

Figure 3: Comparing the counts of (A) nodes or (B) edges to their corresponding overall contribution (or attribution) to predictions of CENTAURGNN (Animal+Human). Preclinical animal data is more scarce than human data, yet they account for a much larger share (relative to their size) of the information that CENTAURGNN (Animal+Human) relies on to predict gene-indication associations.

8.6 UK Biobank Phenocodes

We provide a list of phenocodes used for each phenotype analyzed in this study (Table 19), defined in Karczewski et al. [25]

Table 19: Corresponding UK Biobank phenocodes to selected indications.

Indication	**UK Biobank phenocodes**
Alzheimer's disease	Alzheimers_custom1, Alzheimers_custom2, 130836, 131036, FH_Alzheimer_disease_dementia_custom, 130840, 130842, Frontotemporal_dementia_custom, 20018, 20002
Heart disease	131300, 131298, 20002, 6150, 131304, 131306, CAD_custom, 131288, 20002, 6150, 20002, 131288, 131294, 131286, 131354, heart_failure_custom, 20002, 6150, 6150, 20004, 131310, FH_Heart_disease_custom, 131356, 132474, 20004, 20002, 20002, 20004, 20002, 131346, 131352, 131338, 20002
Ischemia	131304, 20002, 131306, 131056, 20002, 131346, 131300, 131298, 20002, 6150, CAD_custom, 131288, 20002, 6150, 20002, 131288, 131294, 131286, 6150, 6150, 20004, 20002, 131346, 131352
Kidney disease	20002, 132034, 131290, 132014, 132032, 132050, 20002, 132044, 132532, C64, 132530, 41210, 132036, 132042, 20002, 132046, 20002, 20004, 20002, 20004, 20001, 20002, 132030
Liver disease	NAFLD_custom, 131668, 20002, 131670, 131666, 132244, 20004, C22, all_cause_liver_custom, 131658, 41210
Myocardial	41200, 131304, 20002, 131306, 131056, 20002, 131346, 131300, 131298, 20002, 6150, CAD_custom, 131288, 20002, 6150, 20002, 131288, 131294, 131286, 6150, 6150, 20004, 20002, 131346, 131352
Obesity	130792, 21001, BMI_custom, BMI_male_custom, 23104, BMI_female_custom, waist_circ_male_custom, WCadjBMI_female_custom, WCadjBMI_custom, 48, waist_circ_custom, waist_circ_female_custom, WCadjBMI_male_custom, WHR_custom, WHR_female_custom, WHRadjBMI_custom, WHRadjBMI_female_custom, WHR_male_custom, WHRadjBMI_male_custom, weight_male_custom, weightadjheight_custom, weightadjheight_male_custom, 21002, weight_custom, 23098, weightadjheight_female_custom, weight_female_custom, 23099, BFPadjBMI_custom, BFPadjBMI_female_custom, body_fat_percent_custom, body_fat_percent_female_custom, BFPadjBMI_male_custom, body_fat_percent_male_custom, 23100