DRIVE-KG: Enhancing variant-phenotype association discovery in understudied complex diseases using heterogeneous knowledge graphs

Ananya Rajagopalan[1], Tram Anh Nguyen[1], Lindsay A. Guare[1], Andre Luis Garao Rico[2], Rasika Venkatesh[1], Lannawill Caruth[3], Regeneron Genetics Center[8], Penn Medicine BioBank[4], Anurag Verma[5], Marylyn D. Ritchie[6,7], Molly A. Hall[2],
Joseph D. Romano[7†*], Shefali Setia-Verma[3‡*]

[1]*Genomics and Computational Biology Graduate Program,* [2]*Department of Genetics,* [3]*Department of Pathology and Laboratory Medicine,* [4]*Penn Medicine Biobank,* [5]*Department of Medicine,* [6]*Department of Genetics, Division of Informatics,* [7]*Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA, USA*
[8]*777 Old Saw Mill River Rd, Tarrytown, NY 10591, USA*
E-mails: †*joseph.romano@pennmedicine.upenn.edu,* ‡*shefali.setiaverma@pennmedicine.upenn.edu*

**co-corresponding authors*

Multi-omics data are instrumental in obtaining a comprehensive picture of complex biological systems. This is particularly useful for women's health conditions such as endometriosis, which has been historically understudied despite having a high prevalence (around 10% of women of reproductive age). Subsequently, endometriosis has limited genetic characterization: current genome-wide association studies explain only 11% of its 47% total estimated heritability, underscoring the need for integrative approaches. Graph representations provide an intuitive and meaningful way to harmonize biological data, using nodes to represent biological concepts (e.g., genes, single nucleotide polymorphisms, proteins, and phenotypes) and edges to represent their relationships. We present DRIVE-KG (Disease Risk Inference and Variant Exploration Knowledge Graph), which uses a heterogeneous graph representation to integrate data from diverse multi-omics datasets. We trained two distinct models using DRIVE-KG: a link prediction model to suggest associations between SNPs and two pilot phenotypes (endometriosis and obesity), and a graph convolutional network (GCN) for patient-level classification of endometriosis/adenomyosis as a combined phenotype. We conducted patient-level classification using data from 1,441 Penn Medicine BioBank participants with gold standard chart-reviewed endometriosis/adenomyosis status. The link prediction model uncovered 66 high-confidence (model score ≥ 0.95) candidate SNP-endometriosis associations, representing largely distinct genetic signals ($R^2 < 0.1$). These variants were enriched for obesity/body mass index traits (24.2%), lipid metabolism (6%), and depressive disorders (4.5%), showing agreement with emerging hypotheses about endometriosis etiology. In contrast, of the high-confidence, candidate SNP–obesity associations that could be evaluated using LDlink, 38.22% were in high linkage disequilibrium ($R^2 \geq 0.8$) with known obesity or comorbidity associations. The GCN to classify patient endometriosis/adenomyosis status had an F1 score of 0.752 compared to 0.698 for a genetic risk score. Despite this moderate improvement, we found that the GCN learned meaningful stratification of underlying adenomyosis signal and severe endometriosis grades. Together, these results demonstrate that heterogeneous integration of multi-omics data is valuable for diverse downstream tasks—including discovery and clinical prediction—particularly for understudied diseases where traditional genomic approaches are insufficient.

Keywords: Multi-Omics; Knowledge Graph; Women's Health; Endometriosis; Complex Diseases; Variant Discovery

1. Introduction

Multi-omics data integration involves combining diverse biological data types (including genomic, proteomic, epigenomic, and metabolomic data) to provide comprehensive insights into the mechanisms of diseases.[1] This approach is particularly useful for uncovering complex molecular interactions and biological pathways that remain hidden when analyzing individual data types in isolation, and leads to a more complete understanding of disease etiology and progression.[2] Despite this potential, significant computational challenges exist with multi-omics data integration. The high-dimensional, heterogeneous, and sparse nature of multi-omics data creates substantial barriers in developing analysis methods that are both biologically interpretable and computationally efficient.[3] These integration challenges highlight the need for analytical frameworks that can effectively represent the inherently complex relationships in multi-omics data.

Knowledge graphs have emerged as a promising solution to address these multi-omics integration challenges. By representing biological entities as nodes and the known relationships that exist between them as edges, knowledge graphs provide a framework that can capture the complex, interconnected nature of biological systems across data modalities.[3] Graph-based approaches are particularly well-suited for large-scale, multi-omics data because they can naturally accommodate heterogeneous data types while preserving their semantic relationships and biological context. For multi-omics integration, a heterogeneous knowledge graph representation offers particular advantages.[4] Unlike homogeneous graphs where all nodes and edges represent the same type of entity or relationship, heterogeneous knowledge graphs can simultaneously represent multiple entity types (e.g., genes, proteins, metabolites) and diverse relationship types (e.g., regulatory interactions, protein-protein interactions, metabolic pathways) within a single structure.[3] Another advantage of knowledge graph-based approaches for multi-omics integration is that they do not require data from the same individuals across different -omics layers. Instead, these methods can leverage population-level patterns and biological knowledge to create comprehensive representations that integrate information from disparate studies and cohorts. Existing biological knowledge encoded in the graph structure can then be leveraged via network analyses for a variety of purposes: making use of existing relationships in the graph to predict new connections, or for downstream tasks such as patient-level classification.

For many well-characterized complex diseases like obesity, extensive genetic studies have successfully explained a substantial proportion of disease heritability. Furthermore, polygenic risk scores (PRS) demonstrate strong predictive performance; for obesity, typical AUC values for PRS are around 0.8.[5] These genetic studies are enabled by large-scale biobanks (e.g., UK Biobank,[6] Penn Medicine BioBank,[7] and All of Us[8]) which are collections of de-identified genetic, lifestyle, clinical and biological data. In contrast, endometriosis—one of many complex gynecological conditions—affects approximately 10% of reproductive age women yet remains severely understudied and underfunded.[9] Endometriosis is characterized by the presence of endometrial-like tissue outside the uterus and can cause chronic pain symptoms along with comorbidities including infertility.[10] Due to its underdiagnosis, traditional genomic approaches fall short as there is a much lower prevalence observed in biobanks. Despite a broad sense heritability estimation from a twin study of 47%,[11] variants identified through GWAS studies of endometriosis to-date ($N = 928,413$; $n_{cases} = 44,125$)[12] have only been able to explain roughly 11% of this heritability, and the predictive accuracy of genetic risk scores have been inconsistent as a result.[13] The ability of knowledge graph methods to integrate information across disparate studies and populations, without requiring matched individual-level data, makes them especially suited to address these gaps. While previous studies have applied graph representations to integrate multi-omics data for disease discovery[14] and prediction,[15] we are not aware of effective applications of these methods to complex and understudied conditions such as those in women's health. By leveraging existing multi-omics data from independent studies and incorporating established biological relationships contributing to disease, these approaches can potentially uncover disease mechanisms and improve predictive models for conditions where traditional large-scale genomic studies have been insufficient.

This study presents an innovative framework for integrating multi-omics data about disease associations through a heterogeneous knowledge graph representation. We hypothesize that by systematically incorporating knowledge from diverse -omics data sources within a unified graph structure, we can identify previously hidden candidate relationships that are both biologically meaningful and mechanistically relevant to disease pathogenesis. Combining large-scale biological data in this way will improve our ability to discover meaningful and interpretable insights about disease pathogenesis. This is particularly impactful for historically understudied women's health conditions where traditional genomic methods have not fully captured the complexity of disease etiology.

2. Methods

2.1. *Knowledge Graph Data Sources*

To construct a heterogeneous knowledge graph (KG), we integrated multiple biological data sources (node and edge types summarized in Table 1). The graph is comprised of four primary node types: single nucleotide polymorphisms (SNPs), genes, proteins, and phenotypes.

2.1.1. *Node Data Sources*

We derived SNP nodes from the Single Nucleotide Polymorphism Database (dbSNP).[16] To avoid node type imbalance caused by including tens of millions of variants, we included only SNPs with minor allele frequency (MAF) ≥ 0.01 in the Penn Medicine BioBank (PMBB). We obtained phenotype nodes from the Human Phenotype Ontology (HPO),[17] which contains standardized disease and trait terminology. We sourced gene information from the National Center for Biotechnology Information (NCBI) gene database,[18] and derived protein nodes from the Universal Protein Resource (UniProt).[19]

2.1.2. *Edge Data Sources*

We incorporated five distinct edge types connecting the four biological entities (SNP, gene, phenotype, and protein) into the KG. We established SNP-to-gene associations using expression quantitative trait loci (eQTL) data from version 8 of the Genotype-Tissue Expression (GTEx) project.[20] We derived SNP-to-phenotype relationships from Open Targets,[21] which aggregates evidence from genome-wide association studies (GWAS) and other sources. We obtained gene-to-phenotype associations from PheWAS (phenome-wide association study) summary statistics, published from OmicsPred.[22] We established SNP-to-protein connections through OmicsPred, and sourced gene-to-protein relationships from UniProt annotations.[19]

Table 1. Knowledge Graph Data Sources

Graph Concept	Data Source	Features
Node: SNP	dbSNP[16]	Bulk RefSNP JSON files
Node: Phenotype	Human Phenotype Ontology[17]	HP Ontology
Node: Gene	NCBI Gene[18]	Homo Sapiens, filtered to 9606 Taxa ID
Node: Protein	UniProtKB[19]	Swiss-Prot Entries (XML)
Edge: SNP to Gene	GTEx[20]	version 8, eQTL
Edge: SNP to Phenotype	Open Targets[21]	GWAS Summary Statistics
Edge: Gene to Phenotype	OmicsPred[22]	PheWAS Summary Statistics
Edge: SNP to Protein	OmicsPred[22]	Olink, Somalogic (Proteomics)
Edge: Gene to Protein	UniProtKB[19]	Swiss-Prot Entries (XML)

2.2. *Disease Risk Inference and Variant Exploration Knowledge Graph (DRIVE-KG) Construction*

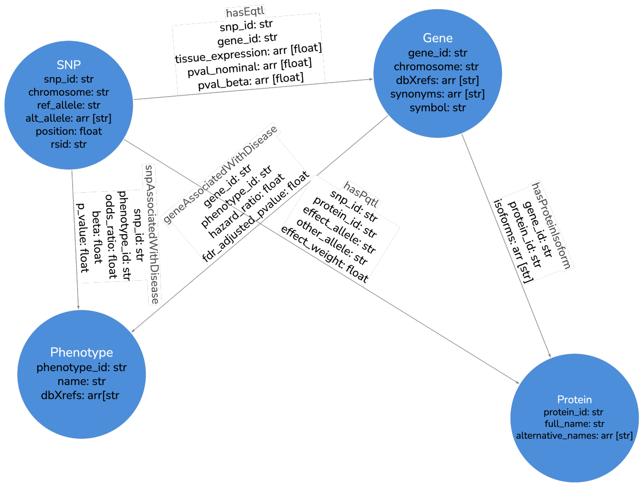

Figure 1. Knowledge Graph Schema.

When designing the DRIVE-KG schema (Figure 1), we prioritized having dense node properties and fewer edge types to avoid sparsity issues when performing machine learning analyses on the graph. We maintained cross references of node type identifiers across databases (‘dbXrefs’) as properties of the gene and phenotype node for possible future querying by other researchers. The ‘tissue_expression’ property of the ‘hasEqtl’ edge is an array of length 53 (number of tissues) that contains all tissue-specific beta values for eGenes in this dataset. Since eGenes (a gene whose expression is influenced by eQTLs) are tissue-specific in the Omics Pred dataset, we aggregated all tissue-specific effects for a particular SNP–Gene pair into this list. If a particular SNP–Gene pair did not have a tissue-specific association, we set the beta value for that entry to 0.

Omics Pred pQTl data pertain to the same 53 tissue types as published by the GTEx consortium, with the exception of Bladder, Fallopian tube, and Cervix (Endocervix & Ectocervix), resulting in a total of 49 tissue types. We mapped the tissue expression array in the same alphabetical order as used by the GTEx consortium,[20] so the index of a given beta value in this array can be used to infer from which tissue it came from.

The DRIVE-KG construction process involved several key steps: first, we standardized identifiers across all data sources (e.g., harmonized to the same Entrez identifier for all gene-related data and ensured that hg38 was used to represent all SNP-related information). Finally, we ingested these files into Memgraph for subsequent visualization and analysis.

For the purposes of model training, we removed all unconnected nodes (node with degree = 0) in the graph (removing 94.2% of the original 13,806,079 nodes), yielding a total of 801,112 nodes and 1,390,440 edges in the graph. The removed nodes consisted mainly of highly specific genetic variants and phenotypes without documented associations in our data sources. This pre-processing step preserved the core connectivity structure necessary for effective graph-based learning and link prediction analyses.

2.3. *Link Prediction*

To identify previously unreported, biologically relevant relationships within DRIVE-KG, we employed link prediction methods that leverage its multi-modal data structure. The core premise of our link prediction framework is that missing edges can be inferred from graph connectivity patterns, because biologically meaningful associations often exhibit characteristic topologies. For instance, if a SNP is associated with multiple genes that are collectively linked to a specific phenotype, this suggests a potential direct SNP-phenotype relationship that may not be explicitly captured in current databases. Similarly, proteins sharing common genetic variants and phenotypic associations may indicate previously uncharacterized protein-protein interactions or shared functional pathways.

Given that DRIVE-KG is created with disease-agnostic biological data, it can be queried for links to any phenotype (using its HPO identifier). We were interested in predicting whether there should be a link between a given SNP node and two pilot phenotypes: endometriosis and obesity, both complex traits with contrasting levels of genetic characterization (67 versus 112 existing genome-wide significant variants, respectively).

We generated 64-dimension node embeddings using Memgraph's node2vec[23] module and then used these embeddings as input for a pairwise multilayer perceptron (MLP) model developed using the PyTorch Lightning[24] module. The three-layer feed-forward MLP used rectified linear unit (ReLU) activations and dropout, projecting 64-dim embeddings to 128, then sequentially reducing to 64 and finally 1 dimension. The output scalar represents the probability of a particular SNP-phenotype edge.

We trained the link prediction module on all *snpAssociatedWithDisease* (SNP-Phenotype) edges with Memgraph's native link prediction module, applying a 70-30 training-validation split. To evaluate specific phenotypes, we queried the trained model with their HPO identifiers (HP_0030127 for endometriosis and HP_0001513 for obesity). The model outputs a score in the range [0,1] (from the final sigmoid activation), representing its confidence in a potential association. To assess the sensibility of these predictions, we focused on high-confidence edges (model score $\geq$ 0.95) and cross-referenced them with published GWAS results.[12] We also explored the neighborhood of these candidate associations within DRIVE-KG.

2.4. *Endometriosis Patient Classification*

In order to assess the utility of DRIVE-KG for disease risk stratification, we trained a graph convolutional network (separate from the link prediction procedure). In particular, we explored the potential of this graph to accurately stratify disease risk at the patient level when applied to endometriosis/adenomyosis. We obtained chart reviews for a cohort of 1,441 patients from the PMBB[7] with their status of endometriosis or adenomyosis (case or control) labeled, as these conditions are often grouped together during phenotyping despite likely differences in their underlying pathogenesis.[25] For endometriosis cases, the chart review also documented surgical disease stage according to established clinical classification systems, enabling both binary case-control analyses and stage-specific investigations. To calculate genetically inferred ancestry (GIA) for each participant, principal components were calculated from PMBB genotype data and then projected onto the reference HapMap3 dataset.[26] GIA for each participant was assigned by comparing their projected principal component scores to the HapMap3 reference populations and classifying them using a kernel density estimation approach.[27] Table 2 contains additional cohort demographics and GIA; a supplemental version with relevant co-morbidities is available in the project's Github repository.

Table 2. Demographics of PMBB Chart-Reviewed Cohort

Characteristic	Category	Overall (n=1441)	Control (n=732)	Case (n=709)
Sequenced gender	Female	1441 (100.0%)	732 (100.0%)	709 (100.0%)
GIA group	AFR	613 (42.5%)	256 (35.0%)	357 (50.4%)
	AMR	30 (2.1%)	20 (2.7%)	10 (1.4%)
	EAS	20 (1.4%)	13 (1.8%)	7 (1.0%)
	EUR	746 (51.8%)	422 (57.7%)	324 (45.7%)
	SAS	30 (2.1%)	19 (2.6%)	11 (1.6%)
	UNKNOWN	2 (0.1%)	2 (0.3%)	—

To curate the graph classification dataset, we generated per-patient copies of DRIVE-KG from the chart-reviewed cohort. We did this by first exporting the base KG (with existing network edges) and graph connectivity information from Memgraph into Python (PyTorch Geometric library[28]). Then, we assigned the value of each SNP node, for each patient's graph, as the imputed genotype dosage (where the value indicates the dosage of the alternate allele at a given locus) from imputed genotyping data in the 2024 PMBB data release 3.0, which was processed using PLINK2.0.[29] Here, the prediction task is whether the entire graph represents endometriosis/adenomyosis (case) or not (control). For all non-SNP node types, we represented them with 8-dimension embeddings learned using the node2vec module. We also added per-patient covariates (age and the first 5 of overall ancestry principal components (PCs)) to the classifier.

We leveraged PyTorch Lightning to load the patient graph objects and train the graph classifier. The graph classifier architecture contains two convolutional layers for all edge types, with SAGEConv layers wrapping each of the edges. We chose SAGEConv due to its inductive capabilities, and ability to efficiently scale to large graphs. We used 64-dimension projections of the 8-dimension embeddings learned on all nodes as model input. We used ReLU activation functions for the two convolutional layers, and the 'sum' aggregation function (which generates 64-dimension tensors for each node type). We then generated pooled GNN embeddings over the SNP node type, which creates 64-dimension tensors for the entire graph. We pooled over the SNP node type alone since only the SNP node values differ across patient graphs; therefore, it is likely to contain biological signal that can be used to distinguish patients. We concatenated the pooled, 64-dimension tensors with the per-patient covariates data (6-dimension: age and 5 PCs) to get 70-dimension tensors. We fed these tensors into the first linear layer, which transformed them into 64-dimension tensors with the ReLU activation function. The second linear layer contained a sigmoid activation function that generated a binary classification value (single probability per graph/patient) from the 64-dimension tensors. For each epoch, we specified a batch size of 16, a learning rate of 0.001, and used the Adam optimization function.

We compared the performance of disease risk prediction using DRIVE-KG to a genetic risk score (GRS)[12] developed and evaluated on the same patient cohort (with the same training-validation split of patients between DRIVE-KG and the GRS). We calculate the GRS by linearly combining the patient's genetic information, age and 5 PCs (as covariates) to ensure a fair comparison.

3. Results

3.1. *Link Prediction*

We investigated link predictions for two phenotypes of interest: endometriosis (HP_0030127) and obesity (HP_0001513). For each phenotype, we obtained recommendations from the model of SNP-phenotype links, ranked by their binary classification score and annotated with whether they were existing or candidate associations.

3.1.1. *Endometriosis*

Our link prediction analysis identified 66 high confidence, candidate SNP-endometriosis associations. Notably, many of these predicted associations involved SNPs with well-established links to other complex traits and conditions yet had not been previously associated with endometriosis in the literature (except for rs10828249 located on *MLLT10*[30]). Specifically, 24.2% of the top-ranked predicted SNP-endometriosis links involved variants previously associated with body mass index (BMI)/obesity-related traits, 6% with lipid metabolism (triglycerides and HDL), and 4.5% with depressive disorders. Additionally, these top-ranked predicted SNPs are enriched in previous endometriosis GWAS for p-values with an average magnitude of 0.473. The majority of these associations (60/66) are also not in LD with each other ($R^2 \leq 0.1$), with the exception of 3 variant pairs with R^2 values ranging from 0.4-0.65. We also estimated narrow sense heritability within the PMBB cohort, using previous GWAS associations versus these associations in addition to the 66 SNP-endometriosis candidates. The variant set from previous GWAS (136 SNPs)[12] explained an average variance of 16.73% (pseudo$-R^2$ 90% CI: 13.03–20.39%). Adding the 66 DRIVE-KG variants increased this to 24.64% (pseudo$-R^2$ 90% CI: 20.30–28.82%), an improvement of 7.91%. We observed a higher confidence for candidate associations compared to existing association scores. This could be because the model is undertrained on endometriosis associations compared to more well-studied diseases (i.e., those with a greater number of genome-wide significant associations) and biased towards positive predictions as a result. Candidate SNP-endometriosis associations had a mean confidence of 0.757 ± 0.179, whereas the 67 known SNP-endometriosis edges had a mean confidence of 0.634 ± 0.163. Figure 2 compares known versus candidate SNP$-$endometriosis associations in DRIVE-KG with $-log_2$ link prediction score on the y$-$axis. We also visualize the neighborhood of candidate signals, which are not explicitly connected to endometriosis as a phenotype, but rather relevant pathways and associated conditions.

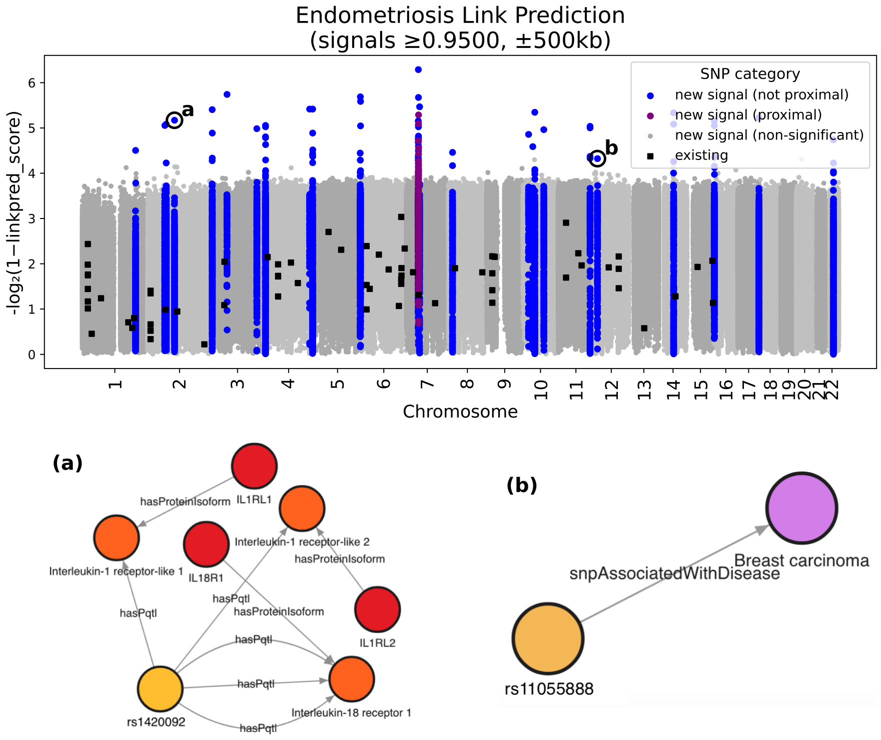

Figure 2. Endometriosis Link Prediction Results. (a) Subgraph view of rs1420092, with 100% of edges from pQTLs in the *IL−1* and *IL−8* family, which are key mediators of immune and inflammatory pathways. (b) Subgraph view of rs1105588 linked to breast cancer (GWAS), which is associated with endometriosis.

3.1.2. *Obesity*

In contrast to the endometriosis results, link prediction for obesity yielded markedly different patterns, reflecting extensive existing knowledge for these well-characterized phenotypes. The link prediction model ranked the 112 existing SNP-obesity associations with a mean confidence of 0.937 ± 0.111 (much higher than that of endometriosis), whereas the candidate SNP-obesity associations were ranked substantially lower with a mean confidence of 0.255 ± 0.178. Of the 1,575 high-confidence, candidate SNP–obesity associations, 348 could be evaluated using LDlink, as many variants lacked corresponding entries in the GWAS Catalog and thus could not be assessed. Among these, 38.22% were in high LD ($R^2 \geq 0.8$) with previously established loci (*NEGR1*[31] and *DNAJC27-AS1*[32]) for obesity or related comorbidities, rather than unreported associations. This suggests that the predicted associations largely recapitulated existing knowledge of metabolic pathways and obesity-related biological mechanisms already well-documented in the literature. The network visualizations for the high-confidence, candidate SNP-obesity associations in Figure 3 (rs74432706 and rs329642) demonstrate that these predictions are primarily driven by neighboring connections to obesity phenotypes, as opposed to distinct biological signals.

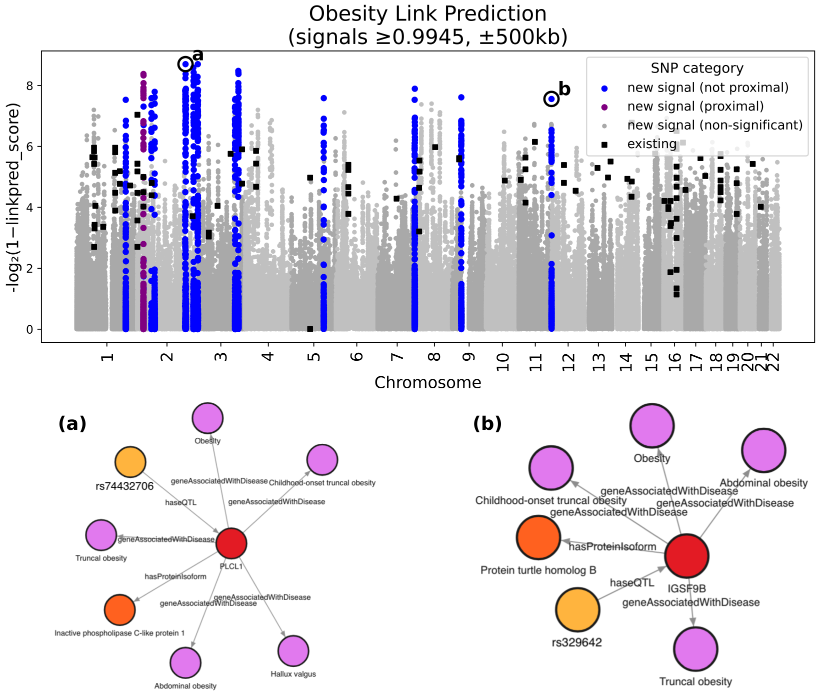

Figure 3. Obesity Link Prediction Results. (a) Subgraph view of rs74432706, with an eQTL edge (driven by the Brain Putamen basal ganglia tissue type) to *PLC1*, which modulates intracellular signaling pathways and has 4 PheWAS associations with obesity phenotypes. (b) Subgraph view of rs329642 with an eQTL edge (driven by Whole Blood and Sigmoid Colon tissue types) to *IGSF9B*, a cell adhesion molecule with 4 PheWAS associations with obesity phenotypes.

3.2. *Endometriosis Patient Classification*

Our cohort consisted of 1,411 PMBB participants assigned female at birth with imputed genotyping data and a chart reviewed label (0 or 1) for having endometriosis/adenomyosis or not. With a 70-30 training-validation split, the graph classifier had an F1 score of 0.752. We compared this to a GRS developed and evaluated on the same patient cohort, which had an F1 score of 0.698 (additional metrics displayed in Table 3). Since this is a binary classification task, the model provides a score (value between 0 and 1, inclusive) for its predicted likelihood of a given patient having either an endometriosis or adenomyosis diagnosis. Figure 4 provides model score distributions stratified both by disease subtypes from chart reviews (Figure 4a) and by surgically confirmed endometriosis stages (1-4) (Figure 4b).

Table 3. Patient Classification Evaluation (Mean ± Standard Deviation)

Model Type	AUC-ROC	AUPRC	F_1	Accuracy
DRIVE-KG	0.571 ± 0.032	0.739 ± 0.033	0.752 ± 0.0193	0.642 ± 0.0225
GRS	0.564 ± 0.032	0.738 ± 0.030	0.698 ± 0.022	0.598 ± 0.024

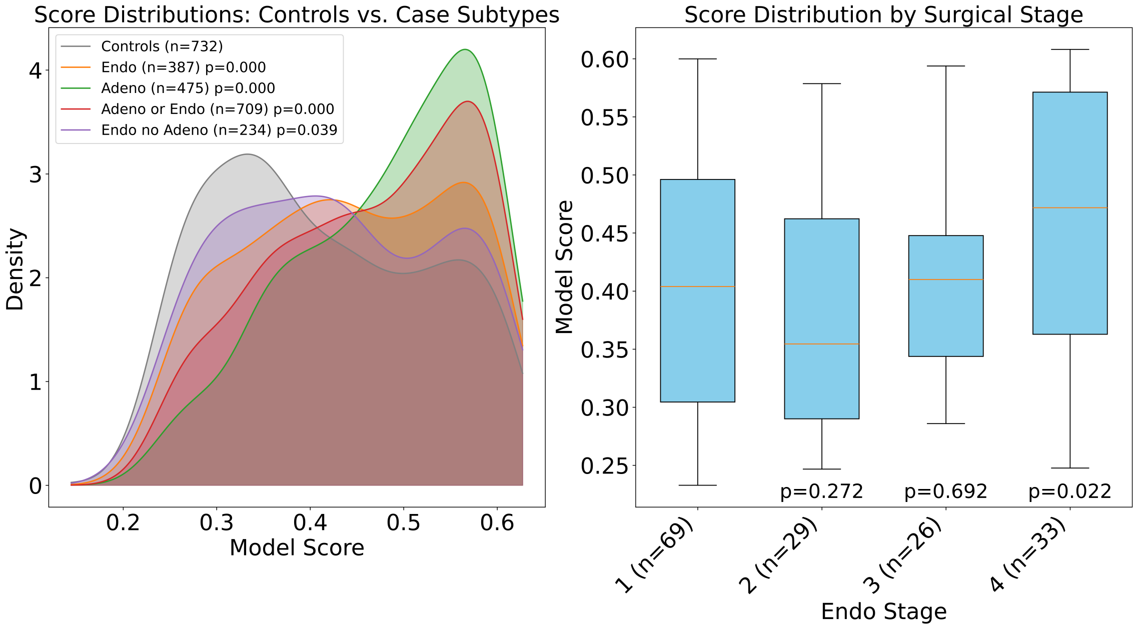

Figure 4. (a) Model score distribution for chart-reviewed disease subtypes. Density distribution indicates that high model scores for patient classification are primarily driven by adenomyosis signal. (b) Model score distribution by chart-reviewed, surgically confirmed endometriosis stages. Boxplots indicate that the model demonstrates particular strength in identifying the most severe endometriosis stage (stage 4).

4. Discussion

This study demonstrates the potential of heterogeneous KG approaches to advance our understanding of understudied diseases where traditional genomic methods have plateaued. Our findings provide several key insights into the utility of integrating multi-omics data through graph-based representations for both discovery and clinical prediction applications. Link prediction revealed a striking contrast between well-studied and understudied phenotypes, validating our central hypothesis. For endometriosis, we identified high-confidence candidate SNPs largely independent of one another ($R^2 < 0.1$) and enriched for variants linked to BMI, lipid metabolism, and depressive disorders—traits frequently co-morbid with endometriosis.[33–35] Network visualization further connected these variants to underlying inflammatory and cell proliferation pathways, providing additional insight into shared biological mechanisms. This pattern suggests shared biological pathways between endometriosis and these co-morbid conditions, which is consistent with emerging mechanistic hypotheses[12,30,36] and our PMBB chart reviewed cohort, which shows elevated risks of hyperlipidemia (RR 1.85), obesity (2.53), CVD (2.03), hypertension (1.79), and depression (1.95). These findings help address the "missing heritability" gap between traditional polygenic risk scores (explaining 11%) and the broad sense heritability estimate from a twin study (47%).[12] Conversely, link predictions for obesity—a trait with extensive genomic characterization—yielded lower confidence scores and predominantly identified variants in LD with known associations. This validates that DRIVE-KG appropriately finds fewer novel discoveries for well-mapped biological landscapes while successfully uncovering genuine candidates for understudied conditions.

Our patient classification for endometriosis demonstrates modest but meaningful improvements over traditional genetic approaches. The DRIVE-KG-based model achieved an F1 score of 0.752 compared to 0.698 for the GRS. These moderate effect sizes show potential clinical relevance for a condition where existing genomic tools perform poorly: the model showed enhanced prediction for adenomyosis and demonstrated particular strength in identifying patients with severe endometriosis (stage 4). The improved performance likely stems from DRIVE-KG's ability to capture complex multi-omics relationships that single-modality approaches miss, even when the only patient-specific input is genetic data. Rather than relying solely on individual genetic variants, our framework leverages broader biological knowledge encoded in protein associations, gene expression patterns, and phenotypic associations. This integrated approach may be particularly valuable for conditions like endometriosis where the genetic architecture is complex and poorly understood.

These findings have broader implications for addressing healthcare disparities in women's health research. Endometriosis exemplifies the challenges facing many gynecological conditions: high prevalence (affecting around 10% of women of reproductive-age[9]), significant clinical impact, yet persistent research neglect leading to limited therapeutic options. Our results suggest that graph-based approaches can help bridge these knowledge gaps by extracting additional insights from existing multi-omics data without requiring massive new cohort studies that may be infeasible for understudied populations. The identification of shared pathways with well-studied traits (obesity,[33] depression[35]) also suggests opportunities for re-purposing existing therapeutic targets and biomarkers. If endometriosis shares mechanistic pathways with metabolic or psychiatric conditions, this could accelerate drug development and improve patient care through a better understanding of comorbidity patterns.

Our study has several limitations. First, the construction of DRIVE-KG relied on existing databases that may contain biases toward well-studied biological pathways, potentially limiting discovery of entirely new mechanisms. Second, while we harmonized variants across alleles when creating DRIVE-KG, we do not fully account for differences in the effect allele when processing the PMBB genotype data to make patient-specific graphs and simply drop variants that are in PMBB data but not in DRIVE-KG. Changing the SNP node representation to be per alternate allele would enable us to more completely utilize genotype data during patient classification. Third, we filtered the SNP nodes at a MAF ≥ 0.01 to accommodate memory constraints from Memgraph, but this restricted our analysis to common variants. Our future work will incorporate rare variants to explore associations with higher effect sizes. For patient classification, our modest AUC improvements will require further validation in independent cohorts, and generalizability across diverse, multi-institution populations remains to be established. Lastly, we cannot yet automatically incorporate data from source databases of DRIVE-KG. Ensuring the continual refreshment of its underlying data is critical for scalable biological discovery, and we plan to address this in subsequent development.

Our future work will focus on incorporating additional data modalities (e.g., transcriptomics, metabolomics, protein-protein interaction and existing genetic pathways) to further enhance the comprehensiveness of DRIVE-KG. Additionally, experimental validation of our predicted SNP-endometriosis associations will be crucial for translating these computational findings into biological insights.

In conclusion, our study demonstrates that DRIVE-KG offers a promising approach to advance precision medicine in understudied diseases. By integrating multi-omics data in a unified framework, we identified previously unreported biological candidates for endometriosis (that underscore emerging hypotheses about disease mechanisms) and improved patient classification beyond traditional genomic approaches. These findings highlight the value of graph-based methods for conditions where conventional genomic tools have reached their current limits, offering a path forward to address knowledge gaps in understudied populations.

5. Acknowledgments

We gratefully acknowledge the PMBB for providing data and thank the patient-participants of Penn Medicine who consented to participate in this research program. The PMBB is approved under IRB protocol #813913 and #852155 and supported by Perelman School of Medicine at University of Pennsylvania, a gift from the Smilow family, and the National Center for Advancing Translational Sciences of the National Institutes of Health under CTSA award number UL1TR001878. The research reported in this publication was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health under award number R01HD110567-01A0, and by the US National Library of Medicine under award number R00LM013646.

Code Availability: `https://github.com/Setia-Verma-Lab/multi_omics_kg`.

References

1. Yehudit Hasin, Marcus Seldin, and Aldons Lusis. Multi-omics approaches to disease. *Genome Biol*, 18(1), December 2017. Publisher: Springer Science and Business Media LLC.
2. Srinivasan Mani, Seema R. Lalani, and Mohan Pammi. Genomics and multiomics in the age of precision medicine. *Pediatr Res*, 97(4):1399–1410, March 2025. Publisher: Springer Science and Business Media LLC.
3. Rachit Kumar, Joseph D. Romano, and Marylyn D. Ritchie. Network-based analyses of multi-omics data in biomedicine. *BioData Mining*, 18(1), May 2025. Publisher: Springer Science and Business Media LLC.
4. Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. Knowledge Graphs. *ACM Comput. Surv.*, 54(4):1–37, May 2022. Publisher: Association for Computing Machinery (ACM).
5. Roelof A. J. Smit, Kaitlin H. Wade, Qin Hui, Joshua D. Arias, Xianyong Yin, Malene R. Christiansen, Loic Yengo, Michael H. Preuss, Mariam Nakabuye, Ghislain Rocheleau, Sarah E. Graham, Victoria L. Buchanan, Geetha Chittoor, Marielisa Graff, Marta Guindo-Martínez, Yingchang Lu, Eirini Marouli, Saori Sakaue, Cassandra N. Spracklen, Sailaja Vedantam, Emma P. Wilson, Shyh-Huei Chen, Teresa Ferreira, Yingjie Ji, Tugce Karaderi, Kreete Lüll, Moara Machado, Deborah E. Malden, Carolina Medina-Gomez, Amy Moore, Sina Rüeger, Masato Akiyama, Matthew A. Allison, Marcus Alvarez, Mette K. Andersen, Vivek Appadurai, Liubov Arbeeva, Eric Bartell, Seema Bhaskar, Lawrence F. Bielak, Joshua C. Bis, Sailalitha Bollepalli, Jette Bork-Jensen, Jonathan P. Bradfield, Yuki Bradford, Caroline Brandl, Peter S. Braund, Jennifer A. Brody, Ulrich Broeckel, Kristoffer S. Burgdorf, Brian E. Cade, Qiuyin Cai, Silvia Camarda, Archie Campbell, Marisa Cañadas-Garre, Jin-Fang Chai, Alessandra Chesi, Seung Hoan Choi, Paraskevi Christofidou, Christian Couture, Gabriel Cuellar-Partida, Rebecca Danning, Frauke Degenhardt, Graciela E. Delgado, Alessandro Delitala, Ayşe Demirkan, Xuan Deng, Alexander Dietl, Maria Dimitriou, Latchezar Dimitrov, Rajkumar Dorajoo, Fabian Eichelmann, Anders U. Eliasen, Jorgen E. Engmann, Michael R. Erdos, Zammy Fairhurst-Hunter, Aliki-Eleni Farmaki, Jessica D. Faul, Juan-Carlos Fernandez-Lopez, Lukas Forer, Mirjam Frank, Sandra Freitag-Wolf, Lars G. Fritsche, Christian Fuchsberger, Tessel E. Galesloot, Yan Gao, Frank Geller, Olga Giannakopoulou, Franco Giulianini, Anette P. Gjesing, Anuj Goel, Scott D. Gordon, Mathias Gorski, Jakob Grove, Xiuqing Guo, Stefan Gustafsson, Jeffrey Haessler, Thomas F. Hansen, Aki S. Havulinna, Simon J. Haworth, Nancy Heard-Costa, Daiane Hemerich, Heather M. Highland, George Hindy, Yuk-Lam Ho, Edith Hofer, Elizabeth Holliday, Katrin Horn, Whitney E. Hornsby, Jouke-Jan Hottenga, Hongyan Huang, Jie Huang, Alicia Huerta-Chagoya, Shaofeng Huo, Mi Yeong Hwang, Chii-Min Hwu, Hiroyuki Iha, Daisuke D. Ikeda, Masato Isono, Anne U. Jackson, Iris E. Jansen, Yunxuan Jiang, Ingegerd Johansson, Anna Jonsson, Torben Jørgensen, Ioanna P. Kalafati, Masahiro Kanai, Stavroula Kanoni, Line L. Kårhus, Anuradhani Kasturiratne, Tomohiro Katsuya, Takahisa Kawaguchi, Rachel L. Kember, Katherine A. Kentistou, Daeeun Kim, Han-Na Kim, Young Jin Kim, Marcus E. Kleber, Maria J. Knol, Azra Kurbasic, Marie Lauzon, Phuong Le, Rodney Lea, Jong-Young Lee, Wen-Jane Lee, Hampton L. Leonard, Hengtong Li, Shengchao A. Li, Xiaohui Li, Xiaoyin Li, Jingjing Liang, Honghuang Lin, Kuang Lin, Jun Liu, Xueping Liu, Ken Sin Lo, Jirong Long, Laura Lores-Motta, Jian'an Luan, Valeriya Lyssenko, Leo-Pekka Lyytikäinen, Anubha Mahajan, Md Zubbair Malik, Vasiliki Mamakou, Massimo Mangino, Ani Manichaikul, Jonathan Marten, Manuel Mattheisen, Aaron F. McDaid, Quanshun Mei, Heike Meiselbach, Tori L. Melendez, Yuri Milaneschi, Jason E. Miller, Iona Y. Millwood, Pashupati P. Mishra, Ruth E. Mitchell, Line T. Møllehave, Nina Mononen,

Sören Mucha, Matthias Munz, Juha Mykkänen, Masahiro Nakatochi, Giuseppe Giovanni Nardone, Christopher P. Nelson, Maria Nethander, Chu Won Nho, Aneta A. Nielsen, Ilja M. Nolte, Suraj S. Nongmaithem, Raymond Noordam, Ioanna Ntalla, Teresa Nutile, Anita Pandit, Marc Pauper, Eva R. B. Petersen, Liselotte V. Petersen, Francesco Piluso, Ozren Polašek, Alaitz Poveda, Saiju Pyarajan, Laura M. Raffield, Hiromi Rakugi, Julia Ramirez, Asif Rasheed, Dennis Raven, Nigel W. Rayner, Carlos Riveros, Rebecca Rohde, Daniela Ruggiero, Sanni E. Ruotsalainen, Kathleen A. Ryan, Maria Sabater-Lleal, Aurora Santin, Richa Saxena, Markus Scholz, Botong Shen, Jingchunzi Shi, Jae Hun Shin, Carlo Sidore, Julia Sidorenko, Xueling Sim, Roderick C. Slieker, Albert V. Smith, Jennifer A. Smith, Laura J. Smyth, Lorraine Southam, Valgerdur Steinthorsdottir, Liang Sun, Fumihiko Takeuchi, Kent D. Taylor, Bamidele O. Tayo, Catherine Tcheandjieu, Natalie Terzikhan, Paola Tesolin, Alexander Teumer, Elizabeth Theusch, Deborah J. Thompson, Gudmar Thorleifsson, Paul R. H. J. Timmers, Stella Trompet, Constance Turman, Simona Vaccargiu, Sander W. Van Der Laan, Peter J. Van Der Most, Jan B. Van Klinken, Jessica Van Setten, Shefali S. Verma, Niek Verweij, Yogasudha Veturi, Carol A. Wang, Chaolong Wang, Jun-Sing Wang, Lihua Wang, Ya Xing Wang, Zhe Wang, Helen R. Warren, Wen Bin Wei, Wanqing Wen, William A. Wheeler, Ananda R. Wickremasinghe, Matthias Wielscher, Bendik S. Winsvold, Andrew Wong, Matthias Wuttke, Rui Xia, Ken Yamamoto, Jingyun Yang, Jie Yao, Hannah Young, Noha A. Yousri, Lei Yu, Lingyao Zeng, Weihua Zhang, Xinyuan Zhang, Jing-Hua Zhao, Wei Zhao, Wei Zhou, Martina E. Zimmermann, Magdalena Zoledziewska, Leen M. 'T Hart, Linda S. Adair, Hieab H. H. Adams, Carlos A. Aguilar-Salinas, Fahd Al-Mulla, Donna K. Arnett, Folkert W. Asselbergs, Bjørn Olav Åsvold, John Attia, Bernhard Banas, Stefania Bandinelli, Lawrence J. Beilin, David A. Bennett, Tobias Bergler, Dwaipayan Bharadwaj, Ginevra Biino, Eric Boerwinkle, Carsten A. Böger, Judith B. Borja, Claude Bouchard, Donald W. Bowden, Ivan Brandslund, Ben Brumpton, Julie E. Buring, Mark J. Caulfield, John C. Chambers, Giriraj R. Chandak, Stephen J. Chanock, Nish Chaturvedi, Yii-Der Ida Chen, Zhengming Chen, Ching-Yu Cheng, Yoon Shin Cho, Kaare Christensen, Ingrid E. Christophersen, Marina Ciullo, John W. Cole, Francis S. Collins, Maria Pina Concas, Richard S. Cooper, Miguel Cruz, Francesco Cucca, Michael J. Cutler, Scott M. Damrauer, Thomas M. Dantoft, Gert J. De Borst, Eco J. C. De Geus, Lisette C. P. G. M. De Groot, Philip L. De Jager, Dominique P. V. De Kleijn, H. Janaka De Silva, George V. Dedoussis, Anneke I. Den Hollander, Shufa Du, Douglas F. Easton, Kai-Uwe Eckardt, Petra J. M. Elders, A. Heather Eliassen, Patrick T. Ellinor, Sölve Elmståhl, Jeanette Erdmann, Michele K. Evans, Diane Fatkin, Bjarke Feenstra, Mary F. Feitosa, Luigi Ferrucci, Jose C. Florez, Ian Ford, Myriam Fornage, Andre Franke, Paul W. Franks, Barry I. Freedman, Christian Gieger, Giorgia Girotto, Yvonne M. Golightly, Clicerio Gonzalez-Villalpando, Penny Gordon-Larsen, Harald Grallert, Struan F. A. Grant, Niels Grarup, Lyn Griffiths, Vilmundur Gudnason, Christopher Haiman, Hakon Hakonarson, Torben Hansen, Catharina A. Hartman, Andrew T. Hattersley, Caroline Hayward, Iris M. Heid, Chew-Kiat Heng, Christian Hengstenberg, Karl-Heinz Herzig, Alex W. Hewitt, Haretsugu Hishigaki, David M. Hougaard, Carel B. Hoyng, Paul L. Huang, Wei Huang, Wen-Yi Huang, Jennifer E. Huffman, Steven C. Hunt, Nina Hutri, Kristian Hveem, Elina Hyppönen, William G. Iacono, Sahoko Ichihara, M. Arfan Ikram, Carmen R. Isasi, Marjo-Riitta Jarvelin, Zi-Bing Jin, Karl-Heinz Jöckel, Jost B. Jonas, Peter K. Joshi, Pekka Jousilahti, J. Wouter Jukema, Mika Kähönen, Yoichiro Kamatani, Kui Dong Kang, Jaakko Kaprio, Sharon L. R. Kardia, Fredrik Karpe, Norihiro Kato, Maryam Kavousi, Frank Kee, Thorsten Kessler, Amit V. Khera, Chiea Chuen Khor, Lambertus A. L. M. Kiemeney, Bong-Jo Kim, Eung Kweon Kim, Hyung-Lae Kim, Paulus Kirchhof, Mika Kivimaki, Woon-Puay Koh, Heikki A. Koistinen, Alexander Kokkinos, Jaspal S. Kooner, Charles Kooperberg, Peter Kovacs, Adriaan Kraaijeveld, Peter Kraft, Ronald M. Krauss, Meena Kumari, Zoltan Kutalik, Markku Laakso, Leslie A. Lange, Claudia Langenberg, Lenore J. Launer, Hyejin Lee, Nanette R. Lee, Terho Lehtimäki, Rozenn N. Lemaitre, Huaixing Li, Liming Li, Wolfgang Lieb,

Xu Lin, Lars Lind, Allan Linneberg, Ching-Ti Liu, Jianjun Liu, Markus Loeffler, Barry London, Fan Lu, Steven A. Lubitz, David A. Mackey, Patrik K. E. Magnusson, JoAnn E. Manson, Gregory M. Marcus, Pedro Marques Vidal, Nicholas G. Martin, Winfried März, Fumihiko Matsuda, Mark I. McCarthy, Robert W. McGarrah, Matt McGue, Amy Jayne McKnight, Sarah E. Medland, Dan Mellström, Andres Metspalu, Braxton D. Mitchell, Paul Mitchell, Dennis O. Mook-Kanamori, Trevor A. Mori, Andrew D. Morris, Lorelei A. Mucci, Patricia B. Munroe, Mike A. Nalls, Saman Nazarian, Amanda E. Nelson, Matt J. Neville, Christopher Newton-Cheh, Christopher S. Nielsen, Harri Niinikoski, Kjell Nikus, Markus M. Nöthen, Adesola Ogunniyi, Claes Ohlsson, Albertine J. Oldehinkel, Lorena Orozco, Katja Pahkala, Päivi Pajukanta, Colin N. A. Palmer, Esteban J. Parra, Cristian Pattaro, Oluf Pedersen, Craig E. Pennell, Brenda W. J. H. Penninx, Louis Perusse, Annette Peters, Patricia A. Peyser, David J. Porteous, Danielle Posthuma, Chris Power, Peter P. Pramstaller, Michael A. Province, Bruce M. Psaty, Qibin Qi, Jia Qu, Daniel J. Rader, Olli T. Raitakari, Loukianos S. Rallidis, Dabeeru C. Rao, Susan Redline, Dermot F. Reilly, Alexander P. Reiner, Sang Youl Rhee, Paul M. Ridker, Michiel Rienstra, Samuli Ripatti, Marylyn D. Ritchie, Fernando Rivadeneira, Dan M. Roden, Frits R. Rosendaal, Jerome I. Rotter, Igor Rudan, Femke Rutters, Seungho Ryu, Charumathi Sabanayagam, Babatunde Salako, Danish Saleheen, Veikko Salomaa, Nilesh J. Samani, Dharambir K. Sanghera, Naveed Sattar, Börge Schmidt, Helena Schmidt, Reinhold Schmidt, Matthias B. Schulze, Heribert Schunkert, Laura J. Scott, Rodney J. Scott, Peter Sever, Wayne H. H. Sheu, M. Benjamin Shoemaker, Xiao-Ou Shu, Eleanor M. Simonsick, Mario Sims, Andrew B. Singleton, Moritz F. Sinner, J. Gustav Smith, Harold Snieder, Tim D. Spector, Beatrice Spedicati, Meir J. Stampfer, Klaus J. Stark, David P. Strachan, Yasuharu Tabara, E. Shyong Tai, Hua Tang, Jean-Claude Tardif, Thangavel A. Thanaraj, Anke Tönjes, Tiinamaija Tuomi, Jaakko Tuomilehto, Maria-Teresa Tusié-Luna, Rob M. Van Dam, Pim Van Der Harst, Nathalie Van Der Velde, Cornelia M. Van Duijn, Natasja M. Van Schoor, Veronique Vitart, Marie-Claude Vohl, Uwe Völker, Peter Vollenweider, Henry Völzke, Scott Vrieze, Niels H. Wacher-Rodarte, Mark Walker, Gurpreet S. Wander, Nicholas J. Wareham, Richard M. Watanabe, Hugh Watkins, David R. Weir, Thomas M. Werge, Elisabeth Widen, Gonneke Willemsen, Walter C. Willett, James F. Wilson, Peter W. F. Wilson, Tien Y. Wong, Jeong-Taek Woo, Alan F. Wright, Huichun Xu, Chittaranjan S. Yajnik, Jian Yang, Mitsuhiro Yokota, Jian-Min Yuan, Eleftheria Zeggini, Babette S. Zemel, Wei Zheng, Xiaofeng Zhu, M. Carola Zillikens, Alan B. Zonderman, John-Anker Zwart, 23andMe Research Team, DiscovEHR (DiscovEHR and MyCode Community Health Initiative), eMERGE (Electronic Medical Records and Genomics Network), GPC-UGR, The PRACTICAL Consortium, Understanding Society Scientific Group, VA Million Veteran Program, Goncalo R. Abecasis, Themistocles L. Assimes, Adam Auton, Michael Boehnke, Daniel I. Chasman, Tõnu Esko, Kari Stefansson, Guillaume Lettre, Cecilia M. Lindgren, Maggie C. Y. Ng, Christopher J. O'Donnell, Unnur Thorsteinsdottir, Peter M. Visscher, Robin G. Walters, Thomas W. Winkler, Andrew R. Wood, Panos Deloukas, Timothy M. Frayling, Anne E. Justice, Tuomas O. Kilpeläinen, Adam E. Locke, Karen L. Mohlke, Kari E. North, Yukinori Okada, Cristen J. Willer, Kristin L. Young, Segun Fatumo, Jeanne M. McCaffery, Nicholas J. Timpson, Joel N. Hirschhorn, Yan V. Sun, Sonja I. Berndt, and Ruth J. F. Loos. Polygenic prediction of body mass index and obesity through the life course and across ancestries. *Nat Med*, July 2025. Publisher: Springer Science and Business Media LLC.

6. Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med*, 12(3):e1001779, March 2015. Publisher: Public Library of Science (PLoS).

7. Anurag Verma, Scott M. Damrauer, Nawar Naseer, JoEllen Weaver, Colleen M. Kripke, Lindsay Guare, Giorgio Sirugo, Rachel L. Kember, Theodore G. Drivas, Scott M. Dudek, Yuki Bradford, Anastasia Lucas, Renae Judy, Shefali S. Verma, Emma Meagher, Katherine L. Nathanson, Michael Feldman, Marylyn D. Ritchie, Daniel J. Rader, and For The Penn Medicine BioBank. The Penn Medicine BioBank: Towards a Genomics-Enabled Learning Healthcare System to Accelerate Precision Medicine in a Diverse Population. *JPM*, 12(12):1974, November 2022. Publisher: MDPI AG.
8. The All of Us Research Program Investigators. The "All of Us" Research Program. *N Engl J Med*, 381(7):668–676, August 2019. Publisher: Massachusetts Medical Society.
9. Krina T. Zondervan, Christian M. Becker, Kaori Koga, Stacey A. Missmer, Robert N. Taylor, and Paola Viganò. Endometriosis. *Nat Rev Dis Primers*, 4(1), July 2018. Publisher: Springer Science and Business Media LLC.
10. J. Prescott, L.V. Farland, D.K. Tobias, A.J. Gaskins, D. Spiegelman, J.E. Chavarro, J.W. Rich-Edwards, R.L. Barbieri, and S.A. Missmer. A prospective cohort study of endometriosis and subsequent risk of infertility. *Hum. Reprod.*, 31(7):1475–1482, July 2016.
11. Rama Saha, Hans Järnbert Pettersson, Pia Svedberg, Matts Olovsson, Agneta Bergqvist, Lena Marions, Per Tornvall, and Ralf Kuja-Halkola. Heritability of endometriosis. *Fertility and Sterility*, 104(4):947–952, October 2015. Publisher: Elsevier BV.
12. Lindsay A Guare, Jagyashila Das, Lannawill Caruth, Ananya Rajagopalan, Alexis T. Akerele, Ben M Brumpton, Tzu-Ting Chen, Leah Kottyan, Yen-Feng Lin, Elisa Moreno, Ashley J Mulford, Vita Rovite, Alan R Sanders, Marija Simona Dombrovska, Noemie Elhadad, Andrew Hill, Gail Jarvik, James Jaworski, Yuan Luo, Shinichi Namba, Yukinori Okada, Yue Shi, Yuya Shirai, Jonathan Shortt, Wei-Qi Wei, Chunhua Weng, Yuji Yamamoto, Sinead Chapman, Wei Zhou, Digna R. Velez Edwards, and Shefali Setia-Verma. Expanding the genetic landscape of endometriosis: Integrative -omics analyses uncover key pathways from a multi-ancestry study of over 900,000 women, November 2024.
13. Kirstine Kloeve-Mogensen, Palle Duun Rohde, Simone Twisttmann, Marianne Nygaard, Kristina Magaard Koldby, Rudi Steffensen, Christian Møller Dahl, Dorte Rytter, Michael Toft Overgaard, Axel Forman, Lene Christiansen, and Mette Nyegaard. Polygenic Risk Score Prediction for Endometriosis. *Front. Reprod. Health*, 3, December 2021. Publisher: Frontiers Media SA.
14. Antoine Bodein, Marie-Pier Scott-Boyer, Olivier Perin, Kim-Anh Lê Cao, and Arnaud Droit. Interpretation of network-based integration from multi-omics longitudinal data. *Nucleic Acids Research*, 50(5):e27–e27, March 2022.
15. Tongxin Wang, Wei Shao, Zhi Huang, Haixu Tang, Jie Zhang, Zhengming Ding, and Kun Huang. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nat Commun*, 12(1):3445, June 2021.
16. S. T. Sherry. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311, January 2001. Publisher: Oxford University Press (OUP).
17. Peter N. Robinson, Sebastian Köhler, Sebastian Bauer, Dominik Seelow, Denise Horn, and Stefan Mundlos. The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. *The American Journal of Human Genetics*, 83(5):610–615, November 2008. Publisher: Elsevier BV.
18. Garth R. Brown, Vichet Hem, Kenneth S. Katz, Michael Ovetsky, Craig Wallin, Olga Ermolaeva, Igor Tolstoy, Tatiana Tatusova, Kim D. Pruitt, Donna R. Maglott, and Terence D. Murphy. Gene: a gene-centered information resource at NCBI. *Nucleic Acids Research*, 43(D1):D36–D42, January 2015. Publisher: Oxford University Press (OUP).
19. A. Bairoch. The Universal Protein Resource (UniProt). *Nucleic Acids Research*, 33(Database issue):D154–D159, December 2004. Publisher: Oxford University Press (OUP).

20. John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, Barbara Foster, Mike Moser, Ellen Karasik, Bryan Gillard, Kimberley Ramsey, Susan Sullivan, Jason Bridge, Harold Magazine, John Syron, Johnelle Fleming, Laura Siminoff, Heather Traino, Maghboeba Mosavel, Laura Barker, Scott Jewell, Dan Rohrer, Dan Maxim, Dana Filkins, Philip Harbach, Eddie Cortadillo, Bree Berghuis, Lisa Turner, Eric Hudson, Kristin Feenstra, Leslie Sobin, James Robb, Phillip Branton, Greg Korzeniewski, Charles Shive, David Tabor, Liqun Qi, Kevin Groch, Sreenath Nampally, Steve Buia, Angela Zimmerman, Anna Smith, Robin Burges, Karna Robinson, Kim Valentino, Deborah Bradbury, Mark Cosentino, Norma Diaz-Mayoral, Mary Kennedy, Theresa Engel, Penelope Williams, Kenyon Erickson, Kristin Ardlie, Wendy Winckler, Gad Getz, David DeLuca, Daniel MacArthur, Manolis Kellis, Alexander Thomson, Taylor Young, Ellen Gelfand, Molly Donovan, Yan Meng, George Grant, Deborah Mash, Yvonne Marcus, Margaret Basile, Jun Liu, Jun Zhu, Zhidong Tu, Nancy J Cox, Dan L Nicolae, Eric R Gamazon, Hae Kyung Im, Anuar Konkashbaev, Jonathan Pritchard, Matthew Stevens, Timothèe Flutre, Xiaoquan Wen, Emmanouil T Dermitzakis, Tuuli Lappalainen, Roderic Guigo, Jean Monlong, Michael Sammeth, Daphne Koller, Alexis Battle, Sara Mostafavi, Mark McCarthy, Manual Rivas, Julian Maller, Ivan Rusyn, Andrew Nobel, Fred Wright, Andrey Shabalin, Mike Feolo, Nataliya Sharopova, Anne Sturcke, Justin Paschal, James M Anderson, Elizabeth L Wilder, Leslie K Derr, Eric D Green, Jeffery P Struewing, Gary Temple, Simona Volpi, Joy T Boyer, Elizabeth J Thomson, Mark S Guyer, Cathy Ng, Assya Abdallah, Deborah Colantuoni, Thomas R Insel, Susan E Koester, A Roger Little, Patrick K Bender, Thomas Lehner, Yin Yao, Carolyn C Compton, Jimmie B Vaught, Sherilyn Sawyer, Nicole C Lockhart, Joanne Demchok, and Helen F Moore. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*, 45(6):580–585, June 2013. Publisher: Springer Science and Business Media LLC.
21. Gautier Koscielny, Peter An, Denise Carvalho-Silva, Jennifer A. Cham, Luca Fumis, Rippa Gasparyan, Samiul Hasan, Nikiforos Karamanis, Michael Maguire, Eliseo Papa, Andrea Pierleoni, Miguel Pignatelli, Theo Platt, Francis Rowland, Priyanka Wankar, A. Patrícia Bento, Tony Burdett, Antonio Fabregat, Simon Forbes, Anna Gaulton, Cristina Yenyxe Gonzalez, Henning Hermjakob, Anne Hersey, Steven Jupe, Şenay Kafkas, Maria Keays, Catherine Leroy, Francisco-Javier Lopez, Maria Paula Magarinos, James Malone, Johanna McEntyre, Alfonso Munoz-Pomer Fuentes, Claire O'Donovan, Irene Papatheodorou, Helen Parkinson, Barbara Palka, Justin Paschall, Robert Petryszak, Naruemon Pratanwanich, Sirarat Sarntivijal, Gary Saunders, Konstantinos Sidiropoulos, Thomas Smith, Zbyslaw Sondka, Oliver Stegle, Y. Amy Tang, Edward Turner, Brendan Vaughan, Olga Vrousgou, Xavier Watkins, Maria-Jesus Martin, Philippe Sanseau, Jessica Vamathevan, Ewan Birney, Jeffrey Barrett, and Ian Dunham. Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res*, 45(D1):D985–D994, January 2017. Publisher: Oxford University Press (OUP).
22. Yu Xu, Scott C. Ritchie, Yujian Liang, Paul R. H. J. Timmers, Maik Pietzner, Loïc Lannelongue, Samuel A. Lambert, Usman A. Tahir, Sebastian May-Wilson, Carles Foguet, Åsa Johansson, Praveen Surendran, Artika P. Nath, Elodie Persyn, James E. Peters, Clare Oliver-Williams, Shuliang Deng, Bram Prins, Jian'an Luan, Lorenzo Bomba, Nicole Soranzo, Emanuele Di Angelantonio, Nicola Pirastu, E. Shyong Tai, Rob M. Van Dam, Helen Parkinson, Emma E. Davenport, Dirk S. Paul, Christopher Yau, Robert E. Gerszten, Anders Mälarstig, John Danesh, Xueling Sim, Claudia Langenberg, James F. Wilson, Adam S. Butterworth, and Michael Inouye. An atlas of genetic scores to predict multi-omic traits. *Nature*, 616(7955):123–131, April 2023. Publisher: Springer Science and Business Media LLC.
23. Aditya Grover and Jure Leskovec. node2vec: Scalable Feature Learning for Networks, 2016. Version Number: 1.
24. William Falcon, Jirka Borovec, Adrian Wälchli, Nic Eggert, Justus Schock, Jeremy Jordan,

Nicki Skafte, Ir1dXD, Vadim Bereznyuk, Ethan Harris, Tullie Murrell, Peter Yu, Sebastian Præsius, Travis Addair, Jacob Zhong, Dmitry Lipin, So Uchida, Shreyas Bapat, Hendrik Schröter, Boris Dayma, Alexey Karnachev, Akshay Kulkarni, Shunta Komatsu, Martin.B, Jean-Baptiste SCHIRATTI, Hadrien Mary, Donal Byrne, Cristobal Eyzaguirre, Cinjon, and Anton Bakhtin. PyTorchLightning/pytorch-lightning: 0.7.6 release, May 2020.

25. Andrew W Horne and Stacey A Missmer. Pathophysiology, diagnosis, and management of endometriosis. *BMJ*, page e070750, November 2022.
26. The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58, September 2010.
27. Yen-Chi Chen. A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology*, 1(1):161–187, January 2017.
28. Matthias Fey and Jan Eric Lenssen. Fast Graph Representation Learning with PyTorch Geometric, May 2019.
29. Christopher C Chang, Carson C Chow, Laurent Cam Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, 4(1), December 2015. Publisher: Oxford University Press (OUP).
30. Nilufer Rahmioglu, Sally Mortlock, Marzieh Ghiasi, Peter L. Møller, Lilja Stefansdottir, Geneviève Galarneau, Constance Turman, Rebecca Danning, Matthew H. Law, Yadav Sapkota, Paraskevi Christofidou, Sini Skarp, Ayush Giri, Karina Banasik, Michal Krassowski, Maarja Lepamets, Błażej Marciniak, Margit Nõukas, Danielle Perro, Eeva Sliz, Marta Sobalska-Kwapis, Gudmar Thorleifsson, Nura F. Topbas-Selcuki, Allison Vitonis, David Westergaard, Ragnheidur Arnadottir, Kristoffer S. Burgdorf, Archie Campbell, Cecilia S. K. Cheuk, Caterina Clementi, James Cook, Immaculata De Vivo, Amy DiVasta, O. Dorien, Jacqueline F. Donoghue, Todd Edwards, Pierre Fontanillas, Jenny N. Fung, Reynir T. Geirsson, Jane E. Girling, Paivi Harkki, Holly R. Harris, Martin Healey, Oskari Heikinheimo, Sarah Holdsworth-Carson, Isabel C. Hostettler, Henry Houlden, Sahar Houshdaran, Juan C. Irwin, Marjo-Riitta Jarvelin, Yoichiro Kamatani, Stephen H. Kennedy, Ewa Kepka, Johannes Kettunen, Michiaki Kubo, Bartosz Kulig, Venla Kurra, Hannele Laivuori, Marc R. Laufer, Cecilia M. Lindgren, Stuart MacGregor, Massimo Mangino, Nicholas G. Martin, Charoula Matalliotaki, Michail Matalliotakis, Alison D. Murray, Anne Ndungu, Camran Nezhat, Catherine M. Olsen, Jessica Opoku-Anane, Sandosh Padmanabhan, Manish Paranjpe, Maire Peters, Grzegorz Polak, David J. Porteous, Joseph Rabban, Kathyrn M. Rexrode, Hanna Romanowicz, Merli Saare, Liisu Saavalainen, Andrew J. Schork, Sushmita Sen, Amy L. Shafrir, Anna Siewierska-Górska, Marcin Słomka, Blair H. Smith, Beata Smolarz, Tomasz Szaflik, Krzysztof Szyłło, Atsushi Takahashi, Kathryn L. Terry, Carla Tomassetti, Susan A. Treloar, Arne Vanhie, Katy Vincent, Kim C. Vo, David J. Werring, Eleftheria Zeggini, Maria I. Zervou, DBDS Genomic Consortium, Kari Stefansson, Mette Nyegaard, FinnGen Study, Paivi Harkki, Oskari Heikinheimo, Johannes Kettunen, Venla Kurra, Hannele Laivuori, Outi Uimari, FinnGen Endometriosis Taskforce, The Celmatix Research Team, Geneviève Galarneau, Caterina Clementi, Piraye Yurttas-Beim, The 23andMe Research Team, Pierre Fontanillas, Joyce Y. Tung, Sosuke Adachi, Julie E. Buring, Paul M. Ridker, Thomas D'Hooghe, George N. Goulielmos, Dharani K. Hapangama, Caroline Hayward, Andrew W. Horne, Siew-Kee Low, Hannu Martikainen, Daniel I. Chasman, Peter A. W. Rogers, Philippa T. Saunders, Marina Sirota, Tim Spector, Dominik Strapagiel, Joyce Y. Tung, David C. Whiteman, Linda C. Giudice, Digna R. Velez-Edwards, Outi Uimari, Peter Kraft, Andres Salumets, Dale R. Nyholt, Reedik Mägi, Kari Stefansson, Christian M. Becker, Piraye Yurttas-Beim, Valgerdur Steinthorsdottir, Mette Nyegaard, Stacey A. Missmer, Grant W. Montgomery, Andrew P. Morris, and Krina T. Zondervan. The genetic basis of endometriosis and comorbidity with other pain and inflammatory conditions. *Nat Genet*, 55(3):423–436, March 2023.
31. Anonymous Reviewer and Stephen Cj Parker. eLife Assessment: 3D genomic features across >50

diverse cell types reveal insights into the genomic architecture of childhood obesity.

32. Arjen J. Boender, Margriet A. Van Gestel, Keith M. Garner, Mieneke C. M. Luijendijk, and Roger A. H. Adan. The obesity-associated gene *Negr1* regulates aspects of energy balance in rat hypothalamic areas. *Physiol Rep*, 2(7):e12083, July 2014.
33. Ingrid J. Rowlands, Richard Hockey, Jason A. Abbott, Grant W. Montgomery, and Gita D. Mishra. Body mass index and the diagnosis of endometriosis: Findings from a national data linkage cohort study. *Obesity Research & Clinical Practice*, 16(3):235–241, May 2022. Publisher: Elsevier BV.
34. Rong Zheng, Xin Du, and Yan Lei. Correlations between endometriosis, lipid profile, and estrogen levels. *Medicine*, 102(29):e34348, July 2023. Publisher: Ovid Technologies (Wolters Kluwer Health).
35. Dora Koller, Gita A. Pathak, Frank R. Wendt, Daniel S. Tylee, Daniel F. Levey, Cassie Overstreet, Joel Gelernter, Hugh S. Taylor, and Renato Polimanti. Epidemiologic and Genetic Associations of Endometriosis With Depression, Anxiety, and Eating Disorders. *JAMA Netw Open*, 6(1):e2251214, January 2023. Publisher: American Medical Association (AMA).
36. Nina Shigesi, Holly R. Harris, Hai Fang, Anne Ndungu, Matthew R. Lincoln, The International Endometriosis Genome Consortium, The 23andMe Research Team, Chris Cotsapas, Julian Knight, Stacey A. Missmer, Andrew P. Morris, Christian M. Becker, Nilufer Rahmioglu, and Krina T. Zondervan. The Association between Endometriosis and Immunological diseases, July 2024.