Discovery of Disease Relationships via Transcriptomic Signature Analysis Powered by Agentic AI*

Ke Chen

School of Information Sciences,
University of Illinois Urbana-Champaign,
Urbana, IL, USA
E-mail: kec10@illinois.edu

Haohan Wang

School of Information Sciences,
University of Illinois Urbana-Champaign,
Urbana, IL, USA
E-mail: haohanw@illinois.edu

Modern disease classification often overlooks molecular commonalities hidden beneath divergent clinical presentations. This study introduces a transcriptomics-driven framework for discovering disease relationships by analyzing over 1,300 disease–condition pairs using GenoMAS, a fully automated agentic AI system. Beyond identifying robust gene-level overlaps, we develop a novel pathway-based similarity framework that integrates multi-database enrichment analysis to quantify functional convergence across diseases. The resulting disease similarity network reveals both known comorbidities and previously undocumented cross-category links. By examining shared biological pathways, we explore potential molecular mechanisms underlying these connections—offering functional hypotheses that go beyond symptom-based taxonomies. We further show how background conditions such as obesity and hypertension modulate transcriptomic similarity, and identify therapeutic repurposing opportunities for rare diseases like autism spectrum disorder based on their molecular proximity to better-characterized conditions. In addition, this work demonstrates how biologically grounded agentic AI can scale transcriptomic analysis while enabling mechanistic interpretation across complex disease landscapes. All results are publicly accessible at `github.com/KeeeeChen/Pathway_Similarity_Network`.

Keywords: Disease similarity network; Transcriptomic associations; AI-driven discovery

*This research was supported by the National AI Research Resource (NAIRR) under grant number 240283.

1. Introduction

Modern disease classification is predominantly grounded in clinical symptoms, anatomical locations, and observable phenotypes.[1–3] While practical for diagnosis and treatment, this symptom-centric taxonomy often obscures deeper biological relationships between diseases—especially those with divergent clinical manifestations but shared molecular origins.[4,5] In contrast, transcriptomic signatures[6] capture gene expression patterns directly reflective of underlying cellular mechanisms, offering a biologically principled lens to reexamine disease relationships.

Recent studies have shown that transcriptomic profiling not only reveals disease-specific pathways related to susceptibility,[7,8] progression,[9–11] and resilience,[12,13] but also uncovers shared molecular programs[6,14,15] across phenotypically distinct diseases. These shared patterns, often invisible to clinical observation,[16,17] have profound implications for disease reclassification,[18,19] biomarker discover,[20,21] and therapeutic repurposing.[22,23] However, realizing these benefits at scale remains challenging:[24,25] each transcriptomic dataset requires extensive preprocessing, normalization, and analysis—an effort that is labor-intensive and difficult to replicate consistently across diverse biological and demographic contexts. Existing frameworks such as Hetionet[26] and DisGeNET[27] have provided valuable resources for studying disease–disease associations, yet they are largely knowledge-based: relying on curated gene–disease links or heterogeneous databases. As a result, they are limited in capturing transcriptomic mechanisms directly from large-scale molecular data or in accounting for condition-specific variation across diseases.

To address this, we leveraged GenoMAS,[28] a fully automated, agentic AI system that performs large-scale transcriptomic analyses across 1,384 disease–condition pairs drawn from the GenoTEX[29] benchmark dataset. Each *pair* represents a disease under a specific biological or demographic condition (e.g., age, sex, obesity, comorbidity), enabling nuanced profiling across 132 diseases and 911 cohorts. Powered by a team of specialized LLM agents, the agentic system performs end-to-end processing, from data cleaning to statistical inference, to identify the genes associated with the disease status under the conditions. In other words, the agentic system identified the *transcriptomic signatures* for each *pair* of disease and condition.

Building on these results, we construct a disease relation network through transcriptomic signatures, identifying statistically significant transcriptomic overlaps between thousands of disease–condition pairs. We validate this network against ICD-10-CM categories and observe both strong within-category clustering and biologically plausible cross-category links—highlighting hidden disease relationships overlooked by traditional taxonomy.

To further interpret the functional basis of these relationships, we extend our analysis to the pathway level. By conducting multi-database enrichment and introducing a novel pathway-based similarity scoring framework, we identify over 1,000 disease combinations that converge on shared molecular pathways. These shared pathways reveal fundamental biological mechanisms that transcend clinical presentation and reflect the cellular logic underlying diverse disease states.

Our analysis recovers well-established comorbidities (e.g., epilepsy and Canavan disease), confirms mechanistically plausible cross-category relationships (e.g., ankylosing spondylitis

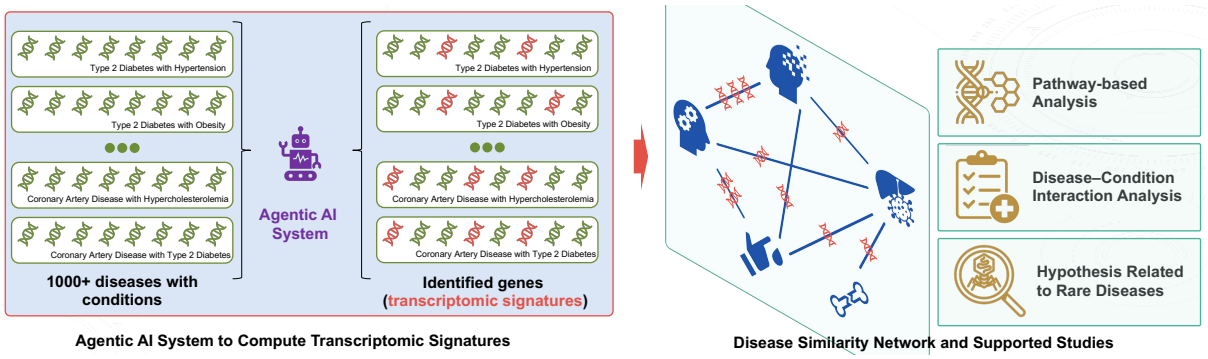

Fig. 1: Agenetic AI analysis of transcriptomic data for transcriptomic signatures and the network of diseases constructed from the signatures.

and osteoporosis), and—most notably—uncovers novel disease links that have not been previously reported in the literature. For these unexpected pairs, we hypothesize potential biological mechanisms supported by shared pathways and gene functions, providing initial interpretive insights to be explored in future studies—for instance, immune and glycosylation-related convergence between Gaucher disease and kidney cancer, or shared metabolic signaling and oxidative stress patterns observed in neurodegenerative diseases and `Ocular Melanoma`.

Finally, we explore how background conditions like obesity and hypertension modulate disease–disease transcriptomic similarity and highlight rare disease cases, such as autism spectrum disorder, where shared molecular signatures with more common diseases may inform drug repurposing opportunities.

To encourage broader exploration of hidden disease relationships, we have made our full results publicly available at `github.com/KeeeeChen/Pathway_Similarity_Network`. Additionally, an initial biological plausibility assessment was conducted using GPT-4o to highlight approximately 200 disease-combinations that exhibit interpretable functional convergence. We hope this resource can inspire new hypotheses and offer alternative perspectives for understanding disease mechanisms beyond established taxonomies.

In summary, the contributions of this paper are illustrated in Figure 1 and as follows:

- We perform large-scale transcriptomic signature analysis across 1,384 disease-condition pairs using an agentic AI system (GenoMAS).
- We construct a gene-level transcriptomic similarity network based on transcriptomic signatures, revealing both strong within-category and cross-category connections.
- We introduce a pathway-level similarity framework based on multi-database enrichment and joint pathway scoring, identifying over 1,000 disease-condition combinations that converge on interpretable molecular mechanisms.
- We highlight examples of transcriptomic convergence in both well-established and unexpected disease pairings, including several cases with no previously documented clinical or molecular connection.
- We study how background conditions such as obesity and hypertension modulate transcriptomic similarity between diseases, and identify rare diseases whose molecular profiles suggest potential therapeutic strategies based on cross-disease alignment.

2. Results

Before presenting our main findings, we first clarify several key terms used throughout the analysis. Our study involves multiple levels of comparison across diseases, biological conditions, and their combinations. Table 1 summarizes the terminology used.

Table 1: Terminology used in this study

Term	Definition / Example
Disease	A clinical diagnosis or condition label. *e.g., Liver Cancer*
Condition	A biological or demographic modifier that contextualizes the disease. *e.g., Obesity, Sex, Age, Hypertension*
Pair	A disease combined with a specific condition. *e.g., Liver Cancer–Obesity*
Combination	A pairwise comparison between two disease–condition pairs. *e.g., (Liver Cancer–Obesity) vs. (Schizophrenia–Gender)*

2.1. *Gene-Based Similarity Network*

To investigate inter-disease relationships at the transcriptomic level, we preliminarily analyzed the statistical significance of overlap of transcriptomic signatures between every *combination* of the 1,384 disease–condition pairs (hereafter, *"combinations"*).

Based on these shared gene relationships, we constructed a graph in which each node represents a disease–condition pair, and each edge connects two pairs that significantly share a set of genes (see Section 3.2 for details). To validate the biological plausibility of the resulting network, we compared our result with the ICD-10-CM classification system.[1] Specifically, we prompted GPT-4o to assign an ICD category to each disease, and constructed a heatmap of average pairwise gene similarity scores for both pairs within the same ICD category and cross-category pairs. (Figure 2)

As expected, many chapters show the strongest similarity within their own category—e.g., Chapter 6, 13, and 9 all display elevated diagonal values. However, the heatmap also reveals that several chapters exhibit their highest similarity scores with other categories rather than their own. For instance, certain subtypes within Chapter 2 and 3 share stronger transcriptomic profiles with Chapter 13 than within their own chapters, suggesting biologically meaningful cross-category overlap. While some of these connections may arise from annotation bias or shared tissue origin, others may reflect previously overlooked biological commonalities.

Together, these findings suggest that while disease taxonomy based on anatomy or symptoms often aligns with molecular signatures, gene-level similarity can also uncover latent biological relationships that transcend clinical classifications. This motivated our subsequent pathway-level analysis to probe deeper into shared mechanisms.

2.2. *Pathway-Based Disease Similarity*

While these combinations significantly shared some genes, their biological relevance remained unclear without understanding what molecular processes these genes are involved in. To better

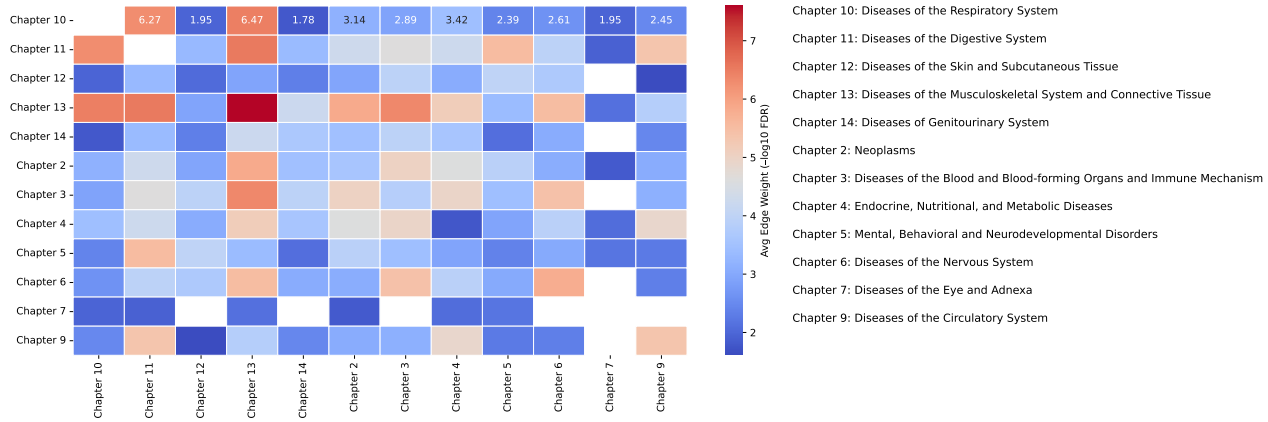

Fig. 2: Heatmap of average gene-based similarity between ICD-10-CM chapters. Diagonal blocks (e.g., Chapters 6, 9, 13) show strong within-category similarity, while several off-diagonal blocks indicate cross-category transcriptomic convergence. Notably, chapter 2, 3, 10, and 11 show higher similarity to other categories than within their own, suggesting latent biological overlap.

interpret the functional basis of disease similarity, we examined pathway-level overlap among the 1,293 significant disease–condition combinations identified by our gene-based analysis (see Section 3.3 for details). Among these combinations, 1,060 were found to share at least one enriched pathway. To visualize these relationships, we constructed a weighted undirected graph (see Figure 3) to provide a systems-level view of transcriptomic convergence across diseases.

This network reveals a clear tendency for nodes of the same ICD-10-CM category to cluster together, which suggests that our pathway-based analysis, while agnostic to clinical labels, nonetheless recapitulates key elements of traditional disease taxonomy. At the same time, many edges span across categories, hinting at molecular commonalities that transcend existing clinical boundaries.

Subsequent analyses in this study are grounded in this network representation. Specifically, we focus on interpretable subgraphs extracted from the full network—such as highly connected modules, cross-category clusters, and rare disease neighborhoods—to uncover novel patterns of comorbidity, shared vulnerability, and potential therapeutic convergence. This pathway network thus serves as the functional scaffold for the biological insights that follow.

2.2.1. *Transcriptomic Similarity Reflects Symptom-Based Taxonomy*

Many high-scoring combinations correspond closely to well-established disease relationships. For example, `Canavan Disease` and `Epilepsy`—both neurological disorders—significantly shared pathways such as **detection of chemical stimulus**, **sensory perception**, and **G protein-coupled receptor signaling pathway**. These pathways are central to neuronal communication and signal transduction, especially in sensory and stimulus-related neural activity. This is consistent with clinical consensus.

There are also other top-scoring combinations aligned with known biological and clinical groupings, including: - `Stomach Cancer` and `Peptic Ulcer Disease`, both involving the gas-

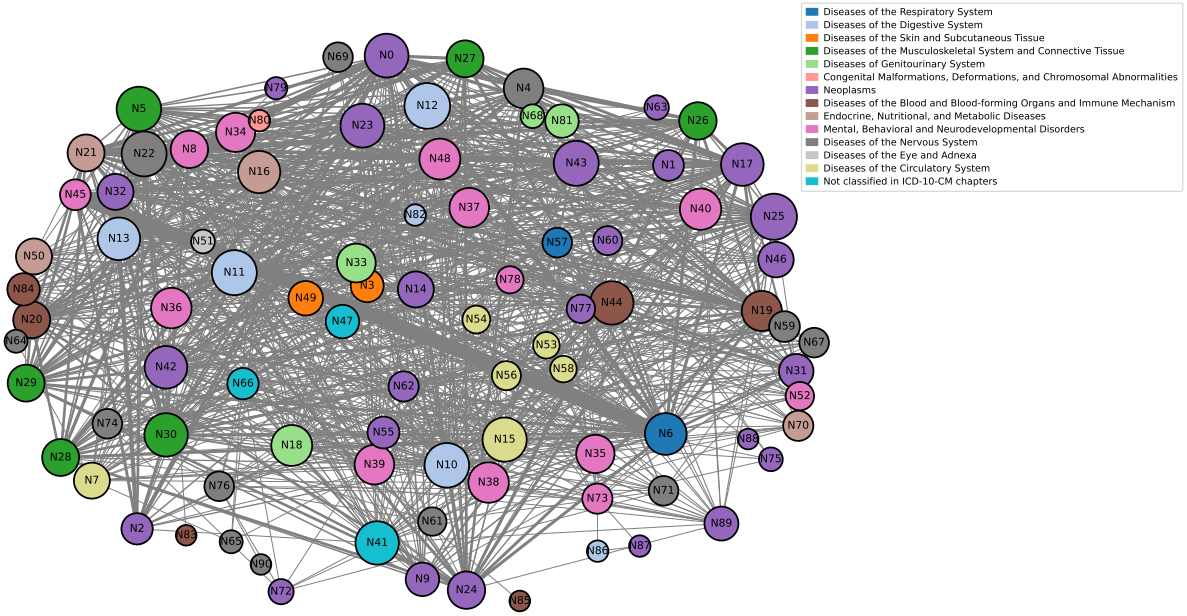

Fig. 3: Pathway-level similarity network. Each node represents a disease–condition pair, colored by ICD-10-CM category. Edges indicate statistically significant overlap in enriched pathways. Both the thickness and the length of each edge reflect the strength of similarity—stronger pathway-level similarity results in shorter and thicker edges. Node size reflects degree centrality. While many nodes are connected, this visualization is designed to emphasize the strength of similarity rather than the presence of connection.

Table 2: P-values of pathological-related shared Pathways in Canavan Disease and Epilepsy

Pathway	**Canavan_Disease–None**	**Epilepsy–None**
detection of chemical stimulus	2.04×10^{-22}	3.72×10^{-26}
sensory perception	6.01×10^{-16}	1.50×10^{-22}
G protein-coupled receptor signaling pathway	1.17×10^{-13}	2.43×10^{-17}

trointestinal system; - `Depression` and `Schizophrenia`, both major psychiatric disorders; - `Bladder Cancer` and `Endometrioid Cancer`, which share hormonal and tissue-level commonalities.

2.2.2. *Cross-Category Transcriptomic Similarity with Empirical Support*

Beyond well-established within-category associations, our pathway-based analysis also revealed biologically meaningful links across some phenotypically unrelated disease categories.

One example is `Ankylosing Spondylitis (AS)` and `Osteoporosis`, two conditions traditionally categorized under musculoskeletal and metabolic disorders, respectively. They significantly share genes such as *AAMDC*, *ABCB1*, and *ABCA5*, along with enriched pathways

related to lipid metabolism, cholesterol regulation, ABC transporters, steroid biosynthesis, and xenobiotic response.

These functions jointly regulate inflammation, immune activity, and bone remodeling—suggesting a shared biological axis linking chronic inflammation, lipid dysregulation, and bone loss. This supports the hypothesis that inflammatory mechanisms in AS may drive osteoporosis risk through disrupted metabolic signaling. Our findings are consistent with recent empirical studies confirming an elevated osteoporosis risk in AS patients,[30,31] and with transcriptomic evidence highlighting immune-driven bone density reduction.[32] Our results further clarify potential shared molecular mechanisms underlying this comorbidity.

We also observed high pathway-based similarity between `Hemochromatosis` and `Liver Cancer`, supported by significantly shared genes such as *AADAT*, *A1BG*, *A4GNT*, and *AARS2*. These genes participate in pathways related to amino acid metabolism, mitochondrial function, immune regulation, and glycoprotein processing.

These shared pathways converge on several key processes: iron overload in Hemochromatosis promotes oxidative stress and chronic inflammation in the liver—an organ central to both conditions. Pathways such as *Tryptophan metabolism*, *immune response signaling*, and *protein glycosylation* highlight a potential mechanistic chain involving metabolic disruption, immune imbalance, and epithelial cell proliferation—all of which may facilitate hepatocarcinogenesis.

These findings align with prior epidemiological studies confirming elevated liver cancer risk in patients with HFE-related Hemochromatosis,[33] and extend beyond prior expression analyses by identifying a broader set of molecular mediators.[34]

These examples illustrate how pathway-level similarity can provide complementary context to gene-level overlap, offering candidate functional processes that may help interpret co-occurrence patterns between diseases.

2.2.3. *Transcriptomic Similarities That Are Potentially Unexpected from Conventional View*

One interesting outcome of our transcriptomic similarity analysis is the resemblance observed between several phenotypically and clinically unrelated conditions. One example is `Gaucher Disease` and `Kidney Chromophobe`. Although `Gaucher Disease` is a lysosomal storage disorder and `Kidney Chromophobe` is a renal carcinoma subtype, they share significant expression of genes such as *A1BG*, *A4GNT*, and *A2M*, alongside co-enrichment in pathways involving immune signaling, extracellular matrix (ECM) remodeling, and protein glycosylation.

These overlapping genes suggest a common functional landscape shaped by immune regulation, protein processing, and inflammation. *A1BG* has been linked to tumor-associated immune modulation;[35,36] *A4GNT* influences glycosylation—a process central to immune escape and cellular signaling;[37] and *A2M* is involved in ECM maintenance and inflammatory control.[38] Pathway-level analysis further reveals enrichment in immune response, ECM organization, glycoprotein biosynthesis, and cellular stress adaptation. Together, these findings point to a shared cellular environment marked by chronic inflammation and metabolic stress—hallmarks of both lysosomal disorders and tumorigenesis. While no direct clinical relationship has been reported between `Gaucher Disease` and `Kidney Cancer`, our results suggest

a potentially overlooked biological intersection that warrants further investigation.

A second example involves an unexpected transcriptomic connection between `Alzheimer's Disease & Parkinson's Disease`, and `Ocular Melanoma`. These conditions share significant expression of *AADAT* and *AASDH*, genes involved in lysine[39,40] and tryptophan metabolism,[39] which regulate NAD^+ biosynthesis, glutamate balance, oxidative stress response, and immune modulation.[41–43] Though these processes diverge in pathological outcomes, they are central to both neurodegeneration and cancer.

As shown in Figure 4, we observed shared enrichment in pathways related to amino acid catabolism, β-oxidation, and cellular response to oxidative stress. In neurodegenerative diseases, these pathways are often impaired,[44] leading to energy failure and excitotoxicity. In contrast, `Ocular Melanoma` exhibits enhanced β-oxidation,[45] supporting tumor proliferation and immune evasion. This inverse utilization of the same metabolic axis may reflect a mechanistic fork, shaped by the shared neural crest origin of retinal and neural tissues.[46] AADAT's dual role in neural excitotoxicity and tumor immune regulation further supports this.

Although no clinical relationship has been established between neurodegenerative disorders and `Ocular Melanoma`, the observed transcriptomic similarities may reflect a shared developmental or metabolic context. These findings raise the possibility of underlying molecular features that span traditionally unrelated disease categories, which may merit further investigation through functional or mechanistic studies.

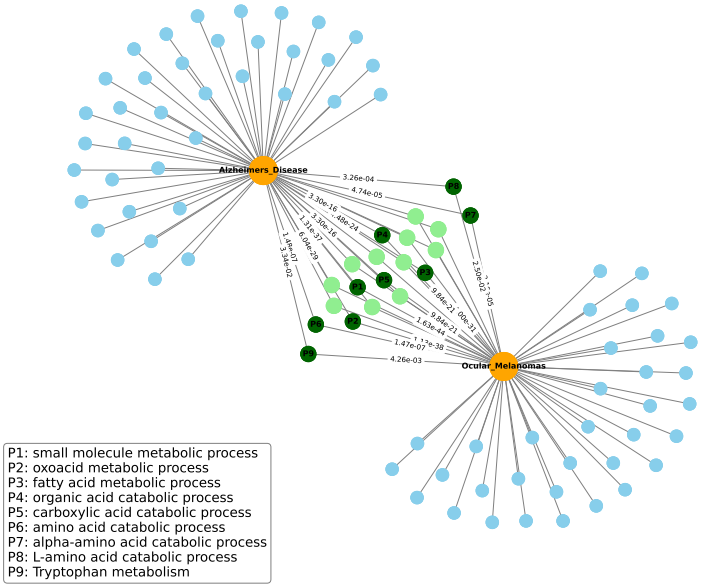

Fig. 4: Shared transcriptomic pathways between Alzheimer's Disease and Ocular Melanoma. The graph displays the top 50 most significant enriched pathways in each disease. Blue nodes represent highly enriched but not shared pathways. Green nodes indicate pathways shared by both diseases, with darker green highlighting the pathways discussed in 2.2.3, which are potentially relevant to the comorbidity. Edge labels reflect pathway significance (p-values), and edge lengths scale with significance.

2.3. *Disease–Condition Interaction Analysis: Triggers of Comorbidity*

While disease–disease similarities often reflect shared genetic programs or pathological mechanisms, the presence of specific physiological or environmental conditions can further modulate the expression of such relationships. In our analysis, we explored disease–condition pairs to understand how background factors—such as obesity or hypertension—may shape transcriptomic overlaps and increase the risk of co-occurrence.

One interesting case involves the co-occurrence of `Celiac Disease` and `Uterine Carcinosarcoma` in obese individuals. Though one is an autoimmune enteropathy and the other a rare uterine malignancy, they share dysregulation of genes such as *A1CF*, *AACS*, and *ABCB1*, which point to altered mRNA editing, amino acid metabolism, and xenobiotic transport. These genes are enriched in pathways that are particularly sensitive to metabolic dysregulation in obesity, including glycosphingolipid biosynthesis, bile secretion, and branched-chain amino acid degradation. In obese individuals, chronic inflammation, disrupted metabolic homeostasis, and impaired detoxification mechanisms may jointly promote both autoimmune activation and tumorigenesis, thus creating a fertile biological landscape for comorbidity.

Another example is the comorbidity of `Acute Myeloid Leukemia` and `Osteoarthritis` in individuals with hypertension. These diseases converge on genes such as *A2ML1* and *A2M*, which regulate extracellular matrix homeostasis and inflammation, as well as *A1CF*, which modulates immune signaling through RNA editing. The two diseases also share enrichment in immune and complement pathways, ECM degradation, and glucocorticoid response—features that are frequently exacerbated in hypertensive individuals. Hypertension, by promoting systemic inflammation, endothelial dysfunction, and hormonal imbalance, may amplify shared transcriptomic vulnerabilities in both hematologic and joint tissues.

This is not an isolated observation. In fact, 11 of the top 20 highest-scoring disease–condition pairs in our transcriptomic similarity analysis involve hypertension, including connections with `Acute Myeloid Leukemia, Adrenocortical Cancer, Gaucher Disease`, and `Osteoarthritis`. These findings underscore the wide-reaching systemic impact of hypertension—not just as a cardiovascular risk factor, but as a molecular amplifier of disease vulnerability across diverse biological systems. Given its high prevalence and silent progression, we emphasize the importance of early detection and integrative management of hypertension to mitigate its far-reaching comorbidity burden.

2.4. *Hypotheses Related to Rare Diseases*

In addition to mapping disease–disease similarity, we also examined whether transcriptomic overlaps with well-characterized conditions could highlight underexplored disorders that worth further investigation. For example, we extracted the subgraph centered on `Autism Spectrum Disorder (ASD)`. As shown in Figure 5, this local network reveals close transcriptomic and pathway-level similarity between ASD and other conditions, including `Osteoporosis` and `Type 1 Diabetes`. While these two diseases are typically studied in distinct clinical domains, their established therapeutic pipelines and shared molecular features with ASD raise the possibility of identifying underexamined connections or therapeutic hypotheses, particularly in individuals with overlapping metabolic or immune phenotypes.

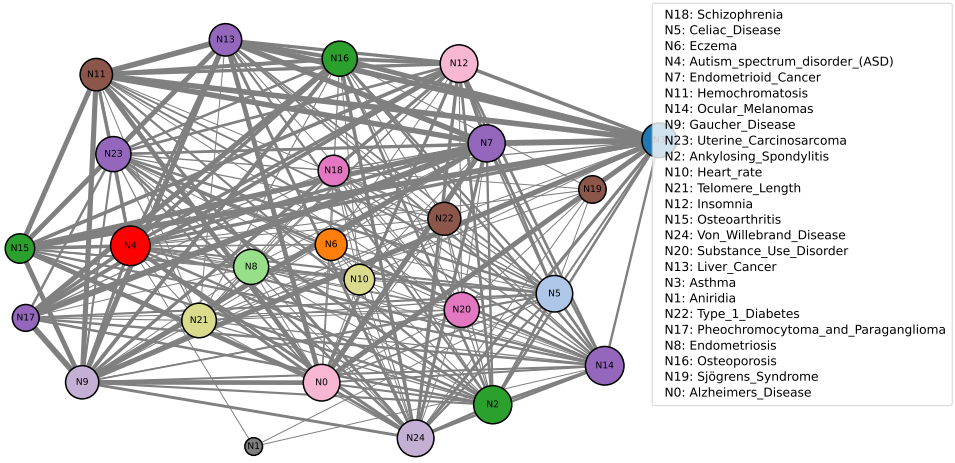

Fig. 5: Subnetwork of diseases showing significant pathway-level similarity to ASD (the red node with index N4). Osteoporosis and Type 1 Diabetes (T1D) emerge as strongly connected conditions, both with established pharmacological pipelines.

In the case of Osteoporosis, ASD shares genes such as *AADAC*, *ABCF3*, and *ABCA7*, which participate in lipid metabolism and ABC transporter pathways. While these genes have not been directly targeted in ASD, several lipid-modulating agents—such as statins, bisphosphonates, and ANGPTL3/APOC3 inhibitors—have demonstrated activity along the same pathways. Their mechanistic action on lipid regulation and inflammatory balance raises the possibility that they could be repurposed for ASD, particularly in individuals with lipid signaling or neuroinflammation phenotypes.

A similar pattern emerges with T1D: shared genes like *AADAT*, *ABCD1*, and *AATF* point to convergence in fatty acid oxidation, mitochondrial stress, and immune dysregulation. Corresponding therapeutic approaches—ranging from anti-inflammatory agents (e.g., α1-antitrypsin, TYK2 inhibitors) to metabolic modulators (e.g., *ABCD1* gene therapy, PPARγ agonists like Leriglitazone)—may offer a foundation for exploratory ASD interventions aimed at metabolic or immune correction.

3. Methodology

3.1. *Large-Scale Gene Analysis via Transcriptomic Agentic AI System*

To explore disease relationships from a transcriptomic perspective, we leveraged the GenoTEX dataset,[29] a large-scale, biologically curated benchmark for automated gene expression analysis. GenoTEX comprises 1,384 gene–disease association problems, spanning 132 distinct human diseases, each analyzed under varying biological or demographic conditions (e.g., age, sex, obesity, hypertension, or comorbidities). For clarity, we refer to the combination of a disease and a condition as a "disease–condition pair", or simply a "pair", throughout this study. The dataset encompasses 911 unique cohorts, totaling over 150,000 biological samples, with each cohort containing more than 18,000 normalized gene features on average.

To process this data at scale, we employed GenoMAS,[28] a multi-agent agentic AI framework

built for code-level automation in genomic analysis. In practice, GenoMAS automates the full workflow of gene expression analysis—from dataset retrieval and preprocessing to regression-based gene–trait association modeling. During preprocessing, the system standardizes heterogeneous data formats (e.g., RNA-seq and microarray), performs gene identifier normalization, and applies corrections for batch effects and population stratification. For statistical inference, it primarily relies on regularized regression (e.g., Lasso) to select trait-associated genes under high-dimensional settings, while incorporating confounder adjustment and imputation strategies to ensure robust signal detection. Using this system, we performed end-to-end gene significance analysis for all 1,384 disease–condition pairs. The results include gene-level effect sizes (regression coefficients), lists of significant genes.

3.2. *Gene-Based Similarity Network*

To quantify transcriptomic similarity, we assessed gene overlap significance for each of the pairwaise combinations of 1384 pairs. For each pair, we retained genes with $|\beta| > 0.05$ from Lasso regression, filtering out weak associations. We then computed shared genes between each combination and performed a bidirectional hypergeometric test, testing enrichment of set A within set B and vice versa, so that significance was not driven solely by large gene sets, while accounting for gene set sizes and the full gene universe (18,000+ genes). Benjamini–Hochberg correction was applied to adjust for multiple testing.

We retained only combinations with false discovery rate (FDR) ≤ 0.05, yielding approximately 65,000 significant pairwise links out of nearly 1 million tested combinations.

To avoid redundancy, we further filtered out overlapping combinations involving generalized condition entries labeled as `None` (i.e., entries not conditioned on any specific biological factor). For example, if a significant link was already identified between `Disease1--None` and `Disease2--Obesity`, then additional links such as `Disease1--Sex` or `Disease1--Age` with `Disease2--Obesity` were considered redundant, as the shared signal likely reflects disease-level rather than condition-specific effects. This de-duplication step prevents generalized associations from inflating or obscuring condition-specific links, reducing the network from over 65,000 initially significant combinations to 1,293 unique links after filtering.

We constructed a gene-level similarity network using `NetworkX`. In this network, each node corresponds to a disease–condition pair, and an edge is drawn between two nodes if their overlap of significant genes passes the bidirectional hypergeometric test after FDR correction. To quantify edge strength, we defined the weight as $-\log_{10}(\text{FDR})$, such that more significant overlaps are assigned larger weights. The resulting weighted graph thus captures both the presence and relative strength of transcriptomic similarity among disease–condition pairs.

For visualization, we employed the `nx.spring_layout()` algorithm to arrange nodes in 2D space. To account for edge strength, we defined the layout weight as the inverse of the edge similarity score ($w_{uv} = 1/(\text{weight}_{uv} + 10^{-5})$), so that stronger similarities correspond to shorter distances. We set the repulsive force parameter to $k = 0.2$ to generate a more compact layout. These settings allow the visualization to reflect both the local clustering and the global structure of the similarity network in a consistent manner.

To annotate the biological identity of each node, we assigned an ICD-10-CM category to

every disease using GPT-4o. These categories were also used as node colors in the visualization for pathway-based similarity network (Figure 3).

3.3. *Pathway-Based Disease Similarity*

To investigate functional overlap among disease–condition pairs, we performed pathway enrichment analysis and similarity scoring for all 1,293 significant combinations identified in the gene-level analysis.

Pathway Enrichment per Pair. For each disease–condition pair, we mapped significantly expressed genes (with $|\beta| > 0.05$) to pathways using six complementary annotation databases: GO:Biological Process (GO:BP),[47] Reactome (REAC), KEGG,[48] transcription factor targets (TF), miRNA targets (MIRNA), and Human Phenotype Ontology (HP).[49] This ensures both biological breadth and low redundancy. We adopted the g:Profiler[50] framework to retrieve enriched terms. g:Profiler selects higher-level, abstracted pathway terms to mitigate semantic variability across databases and maximize interpretability.

Identification of Shared Pathways. For each disease–condition pairwise combination (hereafter, "combination"), we focused only on the genes that were shared between the two pairs. For each such shared gene, we retrieved its pathway annotations in both pairs. A pathway was considered "shared" if it was enriched for the same gene in both pairs.

Similarity Scoring. To quantify the overall strength of pathway-level similarity between two disease–condition pairs, we computed a cumulative score across all shared pathways using the following formula:

$$\text{Similarity Score} = \sum_{k=1}^{N} [\log(1 - p_{1k}) + \log(1 - p_{2k})]$$

Here, N is the number of shared pathways between the two pairs, and p_{1k}, p_{2k} are the enrichment p-values of the k-th pathway in the first and second pair, respectively. For each pathway, this score reduces to $\log((1 - p_{1k}) \times (1 - p_{2k}))$, which reflects the joint probability that both enrichments are non-random. That is, the score becomes more positive when both p_1 and p_2 are small, indicating that the pathway is likely involved in both disease contexts. Summing across all such shared pathways allows us to capture not just the presence of overlap, but the joint confidence in their functional relevance.

Filtering. We retained only combinations where at least one shared pathway had a positive similarity score, indicating non-random co-enrichment. This yielded 1,060 pathway-supported combinations out of the original 1,293 gene-sharing ones.

Pathway-level Graph Construction To visualize cross-disease functional similarity, we constructed an undirected weighted network in which each node represents a disease-condition pair. An edge is drawn between two nodes if the two diseases share at least one enriched pathway. The edge weight corresponds to the pathway-level similarity score. Node size reflects its

degree (i.e., the number of connected neighbors), and node color encodes ICD-10-CM categories. Similar to the gene-level network, we applied a spring-force layout in which edge lengths are inversely proportional to similarity weights. To better illustrate the strength of functional similarity, edge thickness is scaled proportionally to the similarity score—stronger similarities are rendered as thicker connections. In subsequent analyses, we also constructed disease-level subgraphs by collapsing the condition dimension (e.g., Figure 5). In these subgraphs, node size reflects the average degree of each disease across all associated conditions, while edge width and length correspond to the average pathway similarity between the connected diseases.

4. Limitation and Discussion

This work has several limitations. Most importantly, our results are hypothesis-generating rather than experimentally validated. While we provide mechanistic hypotheses for selected examples, these remain speculative until confirmed by biological experiments. For other findings, we performed an initial interpretability assessment using GPT-4o, but these outputs still require expert review and validation to ensure robustness.

We envision several ways in which our resources may support future research. The similarity network may help generate hypotheses on disease etiology, comorbidity, or molecular mimicry. For instance, rare-disease investigators may identify molecular patterns shared with more prevalent conditions, suggesting avenues for drug repurposing or biomarker development. Gene- and pathway-level results may also guide validation studies, especially in evaluating the functional relevance of shared molecular programs. Finally, by providing structured and interpretable outputs from AI-generated analyses, we aim to lower the barrier for translational researchers to engage with complex transcriptomic datasets.

More broadly, we advocate closer integration of AI systems with biomedical research—not just automation, but biologically interpretable and clinically useful output. Agentic AI, when grounded in biological context, can help bridge the gap between large-scale computation and meaningful biological interpretation.

5. Conclusion

This study introduces a transcriptomics-driven framework for rethinking disease relationships beyond traditional clinical boundaries. Leveraging the GenoMAS system, we analyzed over 1,300 disease–condition pairs to construct both gene- and pathway-level similarity networks.

Our multi-layered approach reveals both well-established comorbidities and novel cross-category links, proposing molecular connections across diseases through shared pathways in metabolism, immune response, and cellular stress. We further show that systemic conditions such as obesity and hypertension modulate transcriptomic similarity, while rare diseases like autism spectrum disorder may benefit from therapeutic hypotheses derived from better-characterized conditions.

By publicly sharing our results and network resources, we aim to support hypothesis generation and translational research. More broadly, this work demonstrates how biologically grounded agentic AI can scale transcriptomic analysis while enabling mechanistic interpretation across complex disease landscapes.

References

1. C. for Medicare & Medicaid Services (U.S.) and N. C. for Health Statistics (U.S.), *ICD-10-CM Official Guidelines for Coding and Reporting FY 2024 – UPDATED October 1, 2023 (October 1, 2023 - September 30, 2024)*, tech. rep., Centers for Medicare & Medicaid Services (CMS) (July 2023), Published July 24, 2023.
2. A. S. Fauci, E. Braunwald, D. L. Kasper *et al.*, *Harrison's Principles of Internal Medicine* (McGraw-Hill, 2008).
3. K. I. Goh, M. E. Cusick, D. Valle *et al.*, The human disease network, *Proceedings of the National Academy of Sciences* **104**, 8685 (2007).
4. L. P. Santamaría, E. P. G. del Valle, M. Zanin *et al.*, Classifying diseases by using biological features to identify potential nosological models, *Scientific Reports* **11**, p. 21096 (2021).
5. M. Zanin, J. M. Tuñas and E. Menasalvas, Understanding diseases as increased heterogeneity: A complex network computational framework, *Journal of The Royal Society Interface* **15**, p. 20180405 (2018).
6. N. M. Ferraro *et al.*, Transcriptomic signatures across human tissues identify functional rare genetic variation, *Science* **369**, p. eaaz5900 (2020).
7. F. Q. Wang *et al.*, Unraveling transcriptomic signatures and dysregulated pathways in systemic lupus erythematosus across disease states, *Arthritis Research & Therapy* **26**, p. 99 (2024).
8. L. Shao *et al.*, T-cell transcriptomic signatures in adults with primary untreated immune thrombocytopenia segregate by age, revealing distinct druggable pathways, *Blood* **144**, p. 124 (2024).
9. Z. Poon *et al.*, Transcriptomic signature and functional abnormalities of bone marrow mesenchymal stromal cells mediate disease progression of myelodysplastic/myeloproliferative neoplasms, *Blood* **142**, p. 5618 (2023).
10. Q. Wang *et al.*, Deep learning-based brain transcriptomic signatures associated with the neuropathological and clinical severity of alzheimer's disease, *Brain Communications* **4**, p. fcab293 (2022).
11. S. C. Mendelsohn *et al.*, Transcriptomic signatures of progression to tuberculosis disease among close contacts in brazil, *The Journal of Infectious Diseases* , p. jiae237 (2024).
12. S. Chaudhuri *et al.*, Cell-specific transcriptional signatures of vascular cells in alzheimer's disease: Perspectives, pathways, and therapeutic directions, *Molecular Neurodegeneration* **20**, p. 12 (2025).
13. T. M. Pelaia, M. Shojaei and A. S. McLean, The role of transcriptomics in redefining critical illness, in *Annual Update in Intensive Care and Emergency Medicine 2023*, 2023 pp. 3–14.
14. D. B. Antcliffe *et al.*, Transcriptomic signatures in sepsis and a differential response to steroids: From the vanish randomized trial, *American Journal of Respiratory and Critical Care Medicine* **199**, 980 (2019).
15. J. Bigot *et al.*, Transcriptomic signature of the cd, *American Journal of Transplantation* **16**, 3430 (2016).
16. R. Q. Figueiredo *et al.*, Towards a global investigation of transcriptomic signatures through co-expression networks and pathway knowledge for the identification of disease mechanisms, *Nucleic Acids Research* **49**, 7939 (2021).
17. A. Jha *et al.*, Identifying common transcriptome signatures of cancer by interpreting deep learning models, *Genome Biology* **23**, p. 117 (2022).
18. J. A. Harrill *et al.*, Signature analysis of high-throughput transcriptomics screening data for mechanistic inference and chemical grouping, *Toxicological Sciences* **202**, 103 (2024).
19. M. Kim *et al.*, Refining diagnosis of renal cell carcinoma subtypes through single-cell resolution transcriptomic signatures, *Cancer Research* **84**, 3500 (2024).
20. S. Namba, M. Iwata and Y. Yamanishi, From drug repositioning to target repositioning: Predic-

tion of therapeutic targets using genetically perturbed transcriptomic signatures, *Bioinformatics* **38**, i68 (2022).
21. Z. Zhai *et al.*, Disignatlas: An atlas of human and mouse disease signatures based on bulk and single-cell transcriptomics, *Nucleic Acids Research* **52**, D1236 (2024).
22. S. Lessard *et al.*, Leveraging large-scale multi-omics evidences to identify therapeutic targets from genome-wide association studies, *BMC Genomics* **25**, p. 1111 (2024).
23. F. Wang and C. A. Barrero, Multi-omics analysis identified drug repurposing targets for chronic obstructive pulmonary disease, *International Journal of Molecular Sciences* **25**, p. 11106 (2024).
24. Anaconda, *The State of Data Science 2020: Moving from Hype Toward Maturity*, tech. rep., Anaconda, Inc. (2020).
25. R. BPC, *Navigating the Intersection of Biostatistics, Bioinformatics, and Machine Learning*, tech. rep. (2023), Publication details not fully specified.
26. D. S. Himmelstein, A. Lizee, C. Hessler, L. Brueggeman, S. L. Chen, D. Hadley, A. Green, P. Khankhanian and S. E. Baranzini, Systematic integration of biomedical knowledge prioritizes drugs for repurposing, *eLife* **6**, p. e26726 (2017).
27. J. Piñero, J. M. Ramírez-Anguita, J. Saüch-Pitarch, F. Ronzano, E. Centeno, F. Sanz and L. I. Furlong, The disgenet knowledge platform for disease genomics: 2019 update, *Nucleic Acids Research* **48**, D845 (2020).
28. H. Liu, Y. Li and H. Wang, Genomas: A multi-agent framework for scientific discovery via code-driven gene expression analysis, *arXiv preprint arXiv:2507.21035* (2025).
29. H. Liu, S. Chen, Y. Zhang *et al.*, Genotex: An llm agent benchmark for automated gene expression data analysis, *arXiv preprint arXiv:2406.15341* (2024).
30. K. Sharif, A. M. Tsur, N. Ben-Shabat *et al.*, The risk of osteoporosis in patients with ankylosing spondylitis—a large retrospective matched cohort study, *Medicina Clínica* **160**, 373 (2023).
31. J. Mei, H. Hu, H. Ding *et al.*, Investigating the causal relationship between ankylosing spondylitis and osteoporosis in the european population: A bidirectional mendelian randomization study, *Frontiers in Immunology* **14**, p. 1163258 (2023).
32. D. Zhang, J. Liu, B. Gao *et al.*, Immune mechanism of low bone mineral density caused by ankylosing spondylitis based on bioinformatics and machine learning, *Frontiers in Genetics* **13**, p. 1054035 (2022).
33. J. K. Olynyk and G. A. Ramm, Risk of liver cancer in hfe-hemochromatosis (2021).
34. A. Jayachandran, R. Shrestha, K. R. Bridle *et al.*, Association between hereditary hemochromatosis and hepatocellular carcinoma: A comprehensive review, *Hepatoma Research* **6**, N/A (2020).
35. M. Tian, Y.-Z. Cui, G.-H. Song *et al.*, Proteomic analysis identifies mmp-9, dj-1 and a1bg as overexpressed proteins in pancreatic juice from pancreatic ductal adenocarcinoma patients, *BMC Cancer* **8**, p. 241 (2008).
36. N. Piyaphanee, Q. Ma, O. Kremen *et al.*, Discovery and initial validation of α1-b glycoprotein fragmentation as a differential urinary biomarker in pediatric steroid-resistant nephrotic syndrome, *Proteomics Clinical Applications* (2011).
37. C. Fujii, S. Harumiya, Y. Sato, M. Kawakubo, H. Matoba and J. Nakayama, α1,4-linked n-acetylglucosamine suppresses gastric cancer development by inhibiting mucin-1-mediated signaling, *Cancer Science* **113**, 3852 (2022).
38. C. Sun, C. Cao, T. Zhao *et al.*, A2m inhibits inflammatory mediators of chondrocytes by blocking il-1β/nf-κb pathway, *Journal of Orthopaedic Research* **41**, 241 (2023).
39. E. Okuno, M. Tsujimoto, M. Nakamura and R. Kido, 2-aminoadipate-2-oxoglutarate aminotransferase isoenzymes in human liver: A plausible physiological role in lysine and tryptophan metabolism, *Enzyme and Protein* **47**, 136 (1993).
40. A. Hallen, J. F. Jamie and A. J. L. Cooper, Lysine metabolism in mammalian brain: An update

on the importance of recent discoveries, *Amino Acids* **45**, 1249 (2013).
41. M. M. Essa, S. Subash, N. Braidy *et al.*, Role of nad$^+$, oxidative stress, and tryptophan metabolism in autism spectrum disorders, *International Journal of Tryptophan Research* **6**, p. IJTR.S11355 (2013).
42. L. Yang, Z. Chu, M. Liu *et al.*, Amino acid metabolism in immune cells: Essential regulators of the effector functions, and promising opportunities to enhance cancer immunotherapy, *Journal of Hematology & Oncology* **16**, p. 59 (2023).
43. C. Michaudel, C. Danne, A. Agus *et al.*, Rewiring the altered tryptophan metabolism as a novel therapeutic strategy in inflammatory bowel diseases, *Gut* **72**, 1296 (2023).
44. E. O. Olufunmilayo, M. B. Gerke-Duncan and R. M. D. Holsinger, Oxidative stress and antioxidants in neurodegenerative disorders, *Antioxidants* **12**, p. 517 (2023).
45. D. Lumaquin-Yin, E. Montal, E. Johns, A. Baggiolini, T. H. Huang, Y. Ma, C. LaPlante, S. Suresh, L. Studer and R. M. White, Lipid droplets are a metabolic vulnerability in melanoma, *Nature Communications* **14**, p. 3192 (2023).
46. E. Castro-Pérez, M. Singh, S. Sadangi *et al.*, Connecting the dots: Melanoma cell of origin, tumor cell plasticity, trans-differentiation, and drug resistance, *Pigment Cell & Melanoma Research* **36**, 330 (2023).
47. P. Gaudet, N. Škunca, J. C. Hu *et al.*, Primer on the gene ontology, *Methods in Molecular Biology* **1446**, 25 (2017).
48. M. Kanehisa, The kegg database, in *In Silico Simulation of Biological Processes: Novartis Foundation Symposium 247*, (John Wiley & Sons, Ltd, Chichester, UK, 2002).
49. S. Köhler, L. Carmody, N. Vasilevsky *et al.*, Expansion of the human phenotype ontology (hpo) knowledge base and resources, *Nucleic Acids Research* **47**, D1018 (2019).
50. U. Raudvere, L. Kolberg, I. Kuzmin *et al.*, g:profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update), *Nucleic Acids Research* **47**, W191 (2019).