Supplementary Data

Network-Based Gene Prioritization Using Hybrid Scoring for Complex Disease Module Discovery

Sveva Bonomi, Loris Bottelli, Elisa Oltra, Tiziana Alberio

Contents

List of Supplementary Figures

List of Supplementary Tables

List of Supplementary Methods

Supplementary Figures

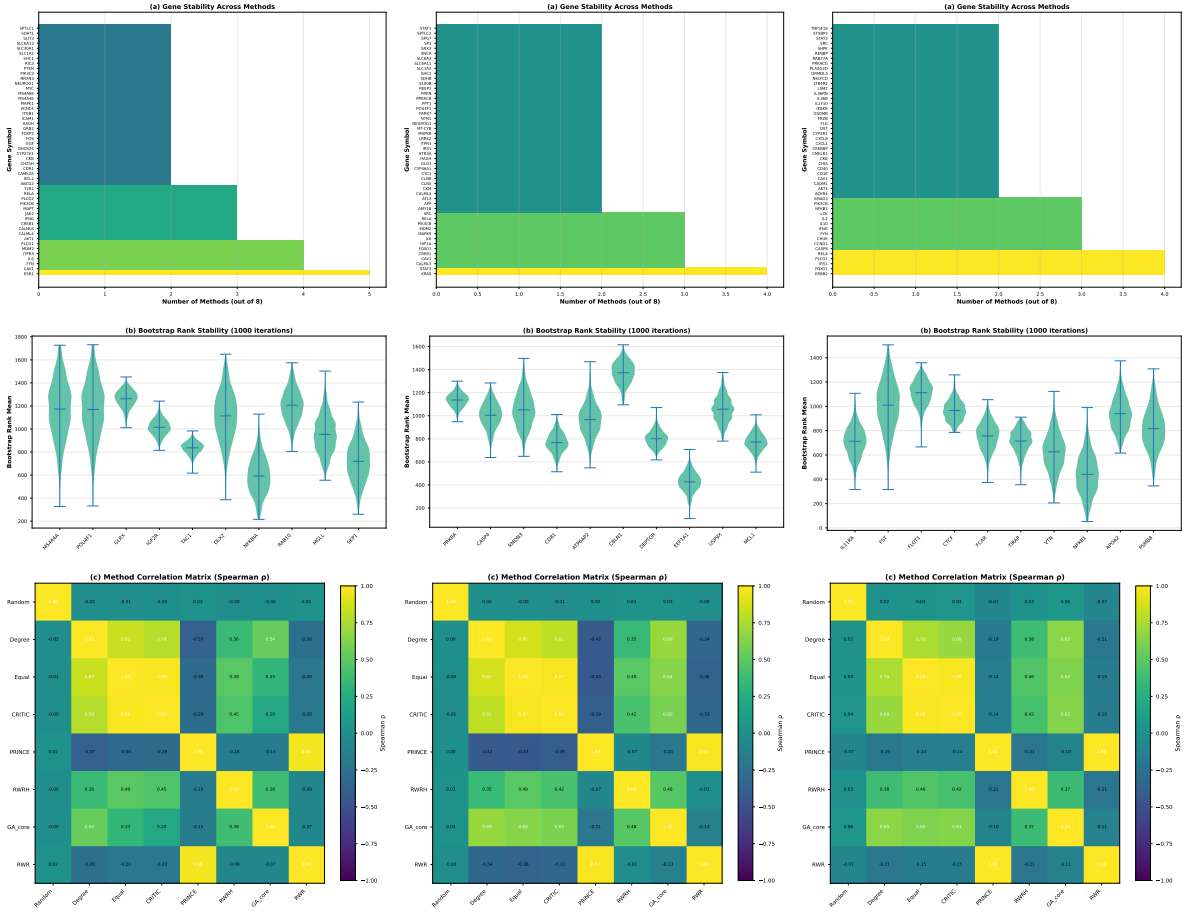

Figure S1. Ranking Stability Analysis and Method Comparison Across Diseases. Three-panel analysis per disease. **(a)** Gene stability plots showing the number of prioritization methods (out of 8) in which each gene appears in the top-50 ranked list; genes consistently selected by multiple methods represent robust candidates. **(b)** Bootstrap rank stability analysis (1,000 iterations): violin plots of rank distributions confirm ranking consistency under random subsampling of ground truth seeds. **(c)** Pairwise Spearman rank correlation matrices across all eight prioritization methods (Random, Degree, Equal, CRITIC, PRINCE, RWRH, GA Core, RWR), revealing the near-zero correlation between topology-based (GA) and diffusion-based (RWR) approaches. Top row: Alzheimer's disease; middle row: Parkinson's disease; bottom row: Rheumatoid arthritis.

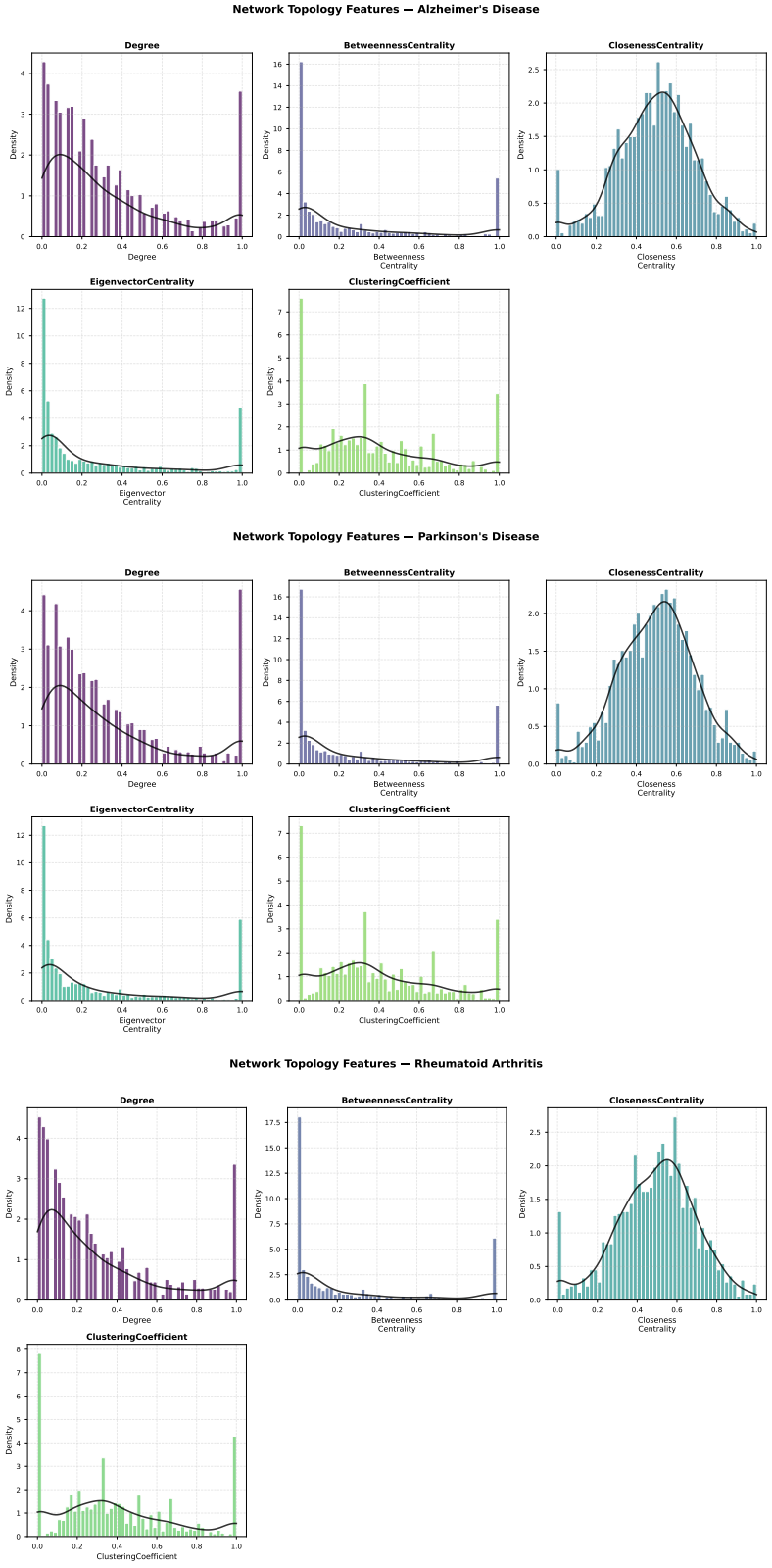

Figure S2. Network Topology Feature Distributions. Kernel density plots for five key topological metrics (degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, and clustering coefficient) across all nodes in the disease-specific PPI network. Top row: Alzheimer's disease; middle row: Parkinson's disease; bottom row: Rheumatoid arthritis. The heavy right tail of betweenness and eigenvector centrality distributions reflects the scale-free architecture of the PPI networks, justifying the use of network centrality features for gene prioritization.

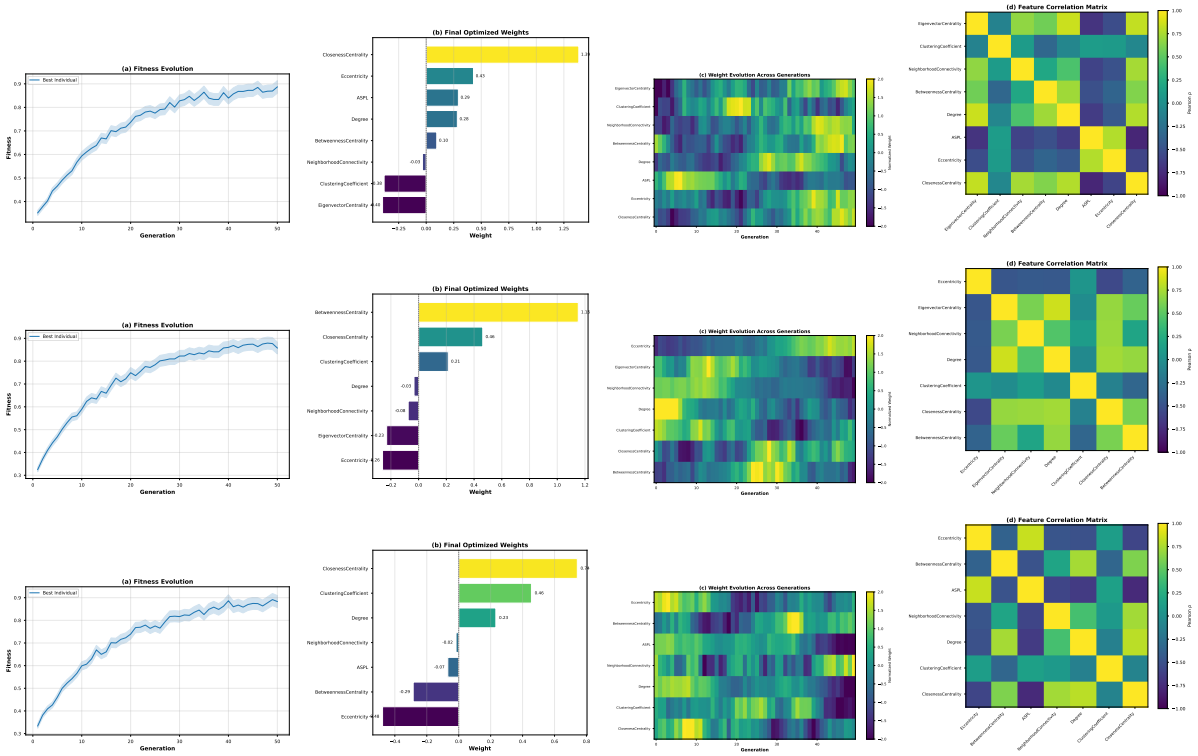

Figure S3. GA Feature Weight Optimization Analysis. Four-panel analysis per disease documenting the genetic algorithm optimization process. **(a)** Fitness evolution across 50 generations showing convergence to stable solutions. **(b)** Final optimized feature weights revealing disease-specific topological signatures; degree centrality, betweenness centrality, and PageRank receive consistently high weights, confirming that disease genes occupy central, bridging network positions. **(c)** Weight evolution heatmaps tracking parameter trajectories across generations. **(d)** Feature correlation matrices demonstrating independence of the selected topological metrics. Top row: Alzheimer's disease; middle row: Parkinson's disease; bottom row: Rheumatoid arthritis.

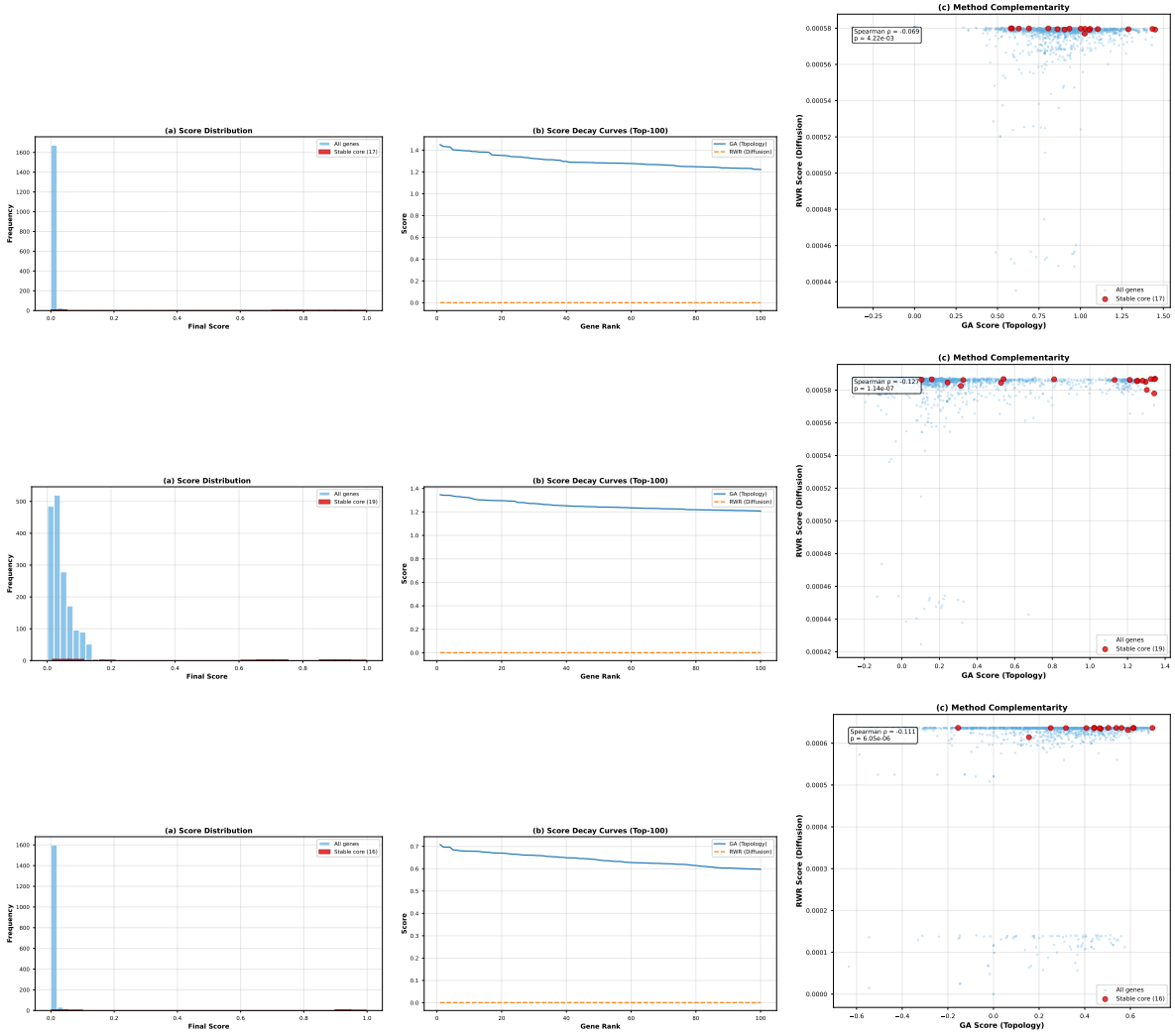

Figure S4. Ranking Score Distributions and Method Complementarity. Three-panel analysis per disease. **(a)** Score distribution histograms comparing all network genes (blue) versus stable core ground truth genes G^* (red); stable genes cluster at the high end of the score distribution, confirming effective prioritization. **(b)** Score decay curves for the top-100 genes comparing GA topology-based scores (solid line) and RWR diffusion-based scores (dashed line); the qualitatively different decay profiles reflect the orthogonal nature of the two ranking strategies. **(c)** 2D scatter plots of GA scores versus RWR scores for all network genes, with Spearman correlation coefficients confirming near-zero correlation ($\rho \approx 0$–-0.15) and complementary coverage of network space. Top row: Alzheimer's disease; middle row: Parkinson's disease; bottom row: Rheumatoid arthritis.

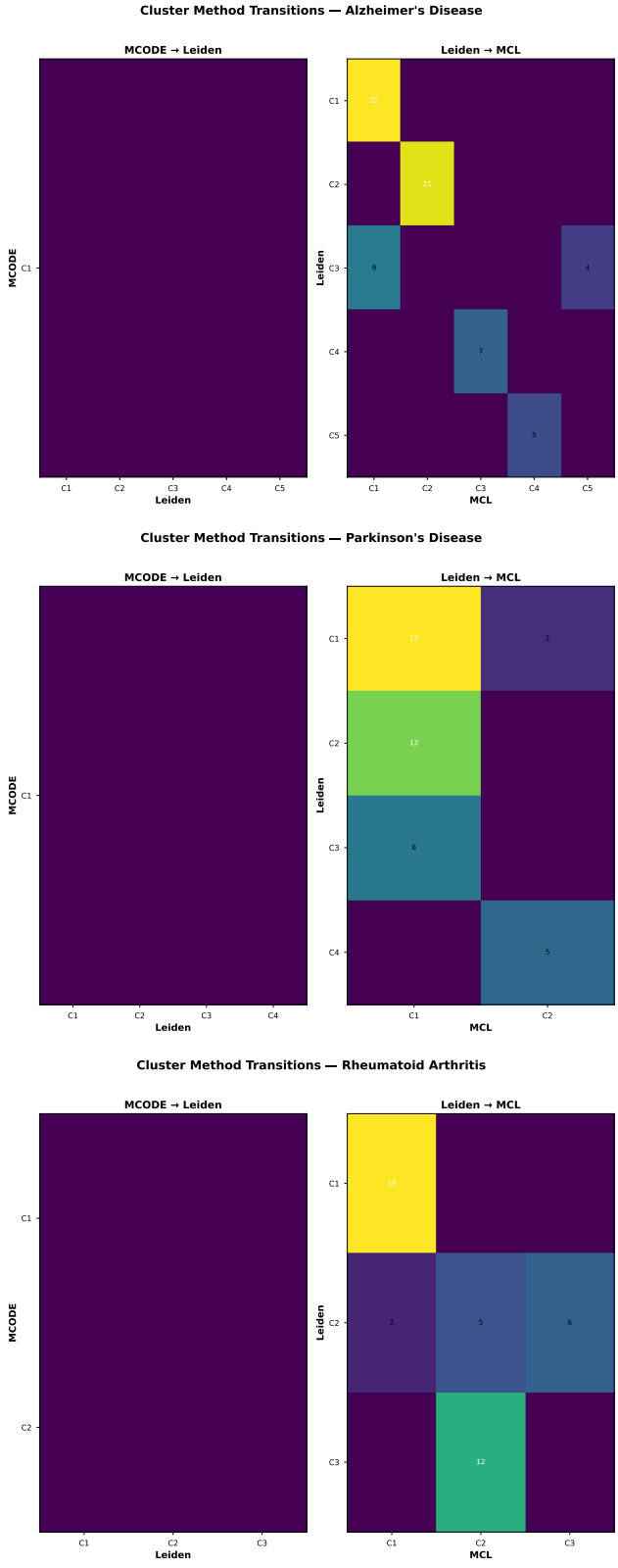

Figure S5. Clustering Method Comparison and Gene Assignment Stability. Alluvial-style diagrams showing gene flow between clustering methods (MCODE → Leiden → MCL) for each disease, with quantitative overlap matrices. Genes whose cluster assignment is consistent across all three methods (bright diagonal entries) belong to coherent, well-defined functional modules; genes with divergent assignments occupy more ambiguous regions of the network. Top: Alzheimer's disease; middle: Parkinson's disease; bottom: Rheumatoid arthritis.

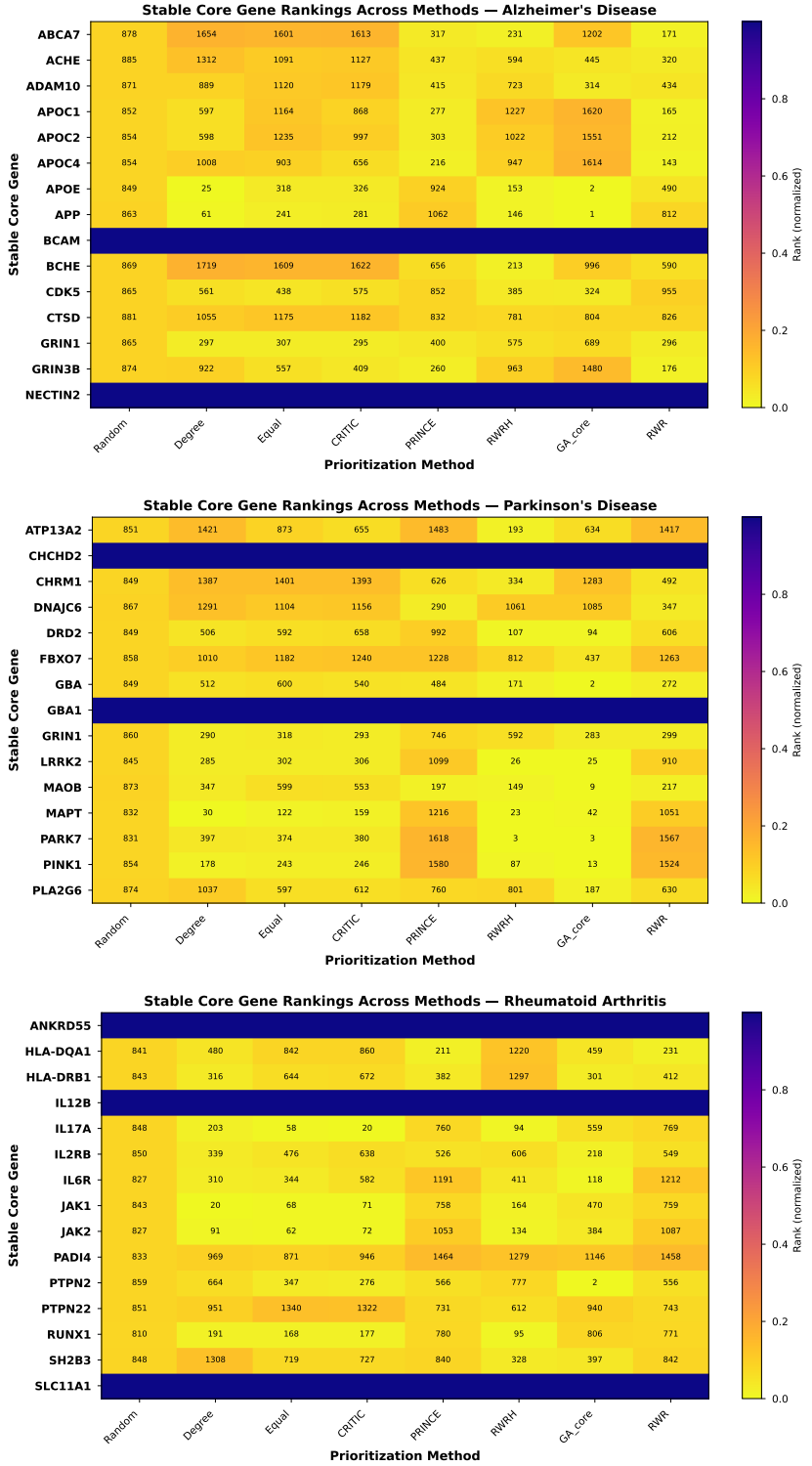

Figure S6. Stable Core Gene Rankings Across Prioritization Methods. Heatmap visualizations showing the ranking positions of stable core disease genes (G^*) across all prioritization methods (Random, Degree, Equal, CRITIC, PRINCE, RWRH, GA Standard, GA Core, RWR, Hybrid). Color intensity represents normalized rank (darker = better rank; lighter = lower rank). Known disease genes consistently achieve top rankings across structurally diverse methods, supporting the biological relevance of the identified gene sets. Top: Alzheimer's disease; middle: Parkinson's disease; bottom: Rheumatoid arthritis.

1 MULTI-SOURCE DATA COLLECTION & NORMALISATION

DISEASES · Knowledge | DISEASES · Experiments | DISEASES · TextMining | GeneCards | OpenTargets | Gene2Phenotype | NCBI Gene

$p'_{i,g} = (p_{i,g} - \min p_{i,g}) / (\max p_{i,g} - \min p_{i,g})$

↓

2 NOISY-OR PROBABILISTIC FUSION

$P_{base}(g) = 1 - \prod_i (1 - p'_{i,g} \cdot w_i)$ | $H(g) = \min(\alpha \cdot \ln(m+1), \beta)$ | $S(g) = \min(P_{base}(g) + H(g), 1.0)$ [m = # supporting sources]

↓

3 SENSITIVITY ANALYSIS → STABLE GENE SELECTION

Default | Conservative | Liberal | Curated-only | Balanced

IF $S_{min}(g) = \min_{c \in C} S_c(g) \geq 0.7$ **THEN** $g \in G^*$ **ELSE** exclude [$|G^*| = 19$ (AD), 22 (PD), 19 (RA)]

↓

4 PPI NETWORK CONSTRUCTION & TOPOLOGICAL FEATURE EXTRACTION

STRING v12.0 · Homo sapiens
combined score ≥ 0.90 · N ≤ 2,000
→ remove self-loops → LCC

- f_1 degree centrality
- f_2 betweenness centrality
- f_3 closeness centrality
- f_4 eigenvector centrality
- f_5 PageRank (d=0.85)
- f_6 clustering coefficient
- f_7 k-core number
- f_8 eccentricity
- f_9 avg neighbour degree
- f_{10} load centrality
- f_{11} harmonic centrality
- → min-max norm [0,1]

↙ ↘

5A GENETIC ALGORITHM OPTIMISATION

$\mathbf{w}^* = \arg\max F(\mathbf{w})$; $\mathbf{w} \in \mathbb{R}^{11}$, $\Sigma w_i = 1$, $w_i \geq 0$

$F_{std}(\mathbf{w}) = \alpha_1 \cdot MRR + \alpha_2 \cdot Coverage - \alpha_3 \cdot Penalty$

$F_{core}(\mathbf{w}) = \alpha_1 \cdot MRR_{G^*} + \alpha_2 \cdot Centralisation$

$S_{GA}(v) = \Sigma_{i=1}^{11} w_i^* \cdot f'_i(v)$

Pop=100; 50 gen; p_c=0.7; p_m=0.2; tournament k=3

5B RANDOM WALK WITH RESTART (RWR)

Seeds: $G^* \rightarrow p_{0,g} = 1/|G^*|$ if $g \in G^*$, else 0

Steady state: $\mathbf{p} = (1-r) \cdot \mathbf{W} \cdot \mathbf{p} + r \cdot \mathbf{p}_0$, $r = 0.3$

WHILE $\|\mathbf{p}^{(t+1)} - \mathbf{p}^{(t)}\|_\infty \geq 10^{-6}$ **DO** iterate

$S_{RWR}(v) = p_v^{(\infty)}$ [steady-state probability]

↘ ↙

6 HYBRID SCORING & COMPLEMENTARITY VALIDATION

$S_{hybrid}(v) = \lambda \cdot S_{GA}(v) + (1-\lambda) \cdot S_{RWR}(v)$ $\lambda = 0.5$

$\rho_{Spearman}(S_{GA}, S_{RWR}) \approx 0$
Jaccard@K=50 ≈ 0 (orthogonal)

↓

7 SUBNETWORK EXTRACTION — ELBOW METHOD

Sort v by S_{hybrid} desc → $\Delta S_i = S_i - S_{i+1}$ → $k^* = \arg\max_i(\Delta^2 S_i)$ → $G_{sub} = G[\{v_1, \ldots, v_{k^*}\}]$ [k^*=69 (AD), 40 (PD), 48 (RA)]

↓

8 FUNCTIONAL MODULE DETECTION — CLUSTERING

LEIDEN

$\max \Sigma_c [m_c - \gamma \cdot n_c(n_c-1)/2]$
CPM; γ=1.0; 10 iter

MCL

REPEAT
$M \leftarrow M^2$ (expand)
inflate I=2.0; prune
UNTIL $\|\Delta M\|_F < 10^{-6}$

MCODE

$w(v) = k_c(v) \cdot d_k(v)$

IF score(u)≥0.2
add u; haircut on

↓

9 OVER-REPRESENTATION ANALYSIS (ORA) — PATHWAY ENRICHMENT

$p = \Sigma_{i=k}^{\min(n,K)} C(K,i) \cdot C(N-K, n-i) / C(N,n)$ [Fisher's exact, one-sided]

IF $p_{adj} < 0.05$ (BH-FDR) **AND** $|P \cap C| \geq 2$ **AND** $5 \leq |P| \leq 500$ **THEN** pathway P enriched in cluster C

GO-BP / MF / CC
KEGG
Reactome
WikiPathways

Background: disease-specific PPI network (N ≈ 1,640–1,830)

Output: G* (stable genes) · candidate genes ranked by S_{hybrid} · functional modules · enriched pathways

Figure S7. Computational Framework Overview. Block diagram of the nine-step computational pipeline. Steps 1–3: multi-source data collection, normalization, and probabilistic fusion via a weighted noisy-OR model to produce gene-disease association scores. Step 4: sensitivity analysis across five weighting configurations to select stable high-confidence seed genes (G^*; criterion $S_{\min}(g) \geq 0.7$). Steps 5–6: PPI network construction (STRING v12.0, combined score ≥ 0.90), parallel topology-based (GA) and diffusion-based (RWR) scoring, and hybrid combination ($\lambda = 0.5$); GA and RWR rankings are nearly orthogonal ($\rho_{\text{Spearman}} \approx 0$). Steps 7–8: subnetwork extraction by elbow method, followed by functional module detection (Leiden, MCL, MCODE). Step 9: over-representation analysis (ORA, BH-FDR) across GO-BP, KEGG, Reactome, and WikiPathways. Full parameter specifications are given in Supplementary Methods SM1–SM7.

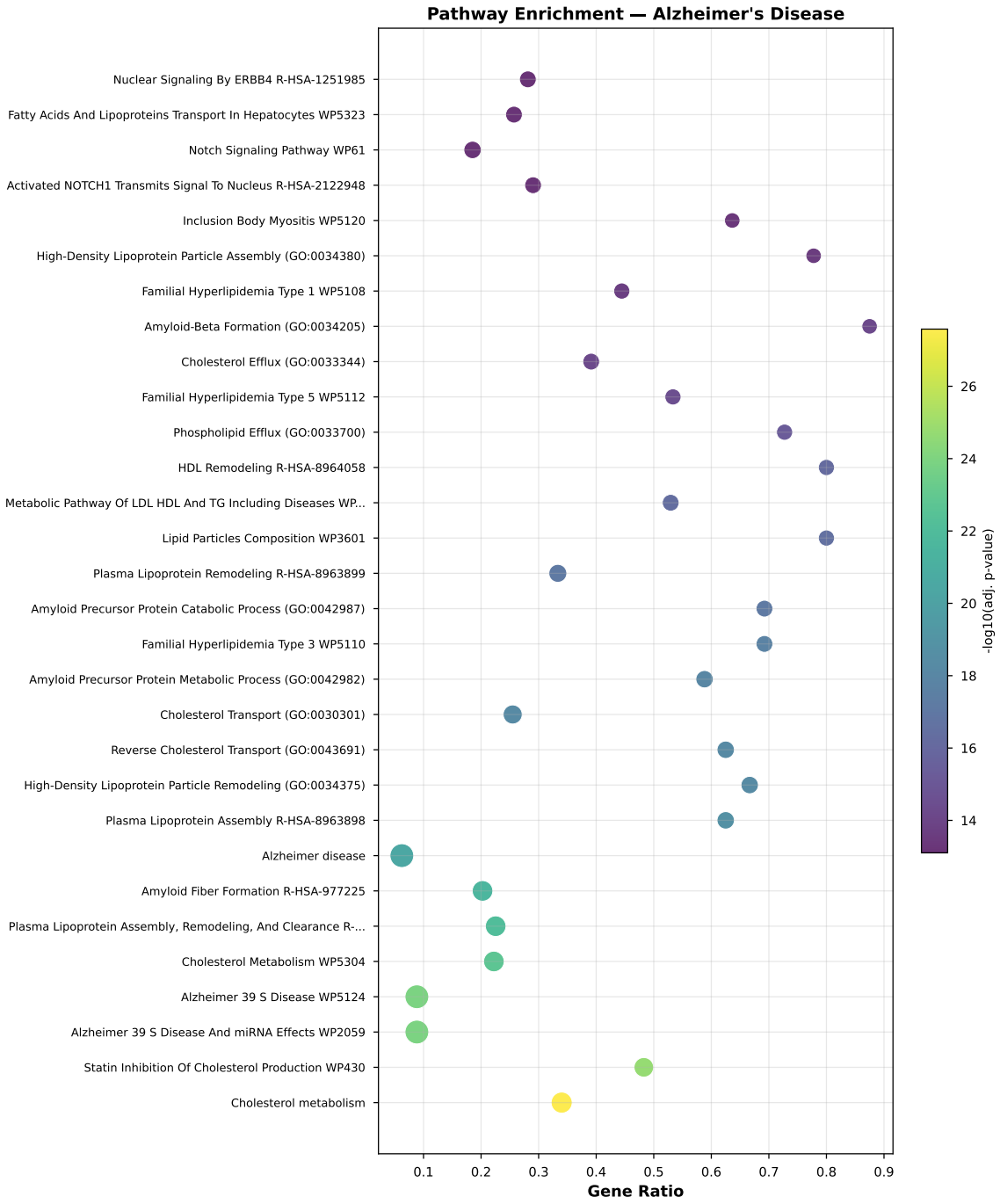

Figure S8. Pathway Enrichment Analysis by Database Source — Alzheimer's Disease. Comprehensive dotplot visualization of subnetwork-level ORA results (genome-wide background, $N = 69$ genes) across four pathway databases (GO Biological Process, KEGG, Reactome, WikiPathways). Top 30 significantly enriched pathways shown. Dot size: gene ratio (fraction of pathway genes present in subnetwork); color intensity: $-\log_{10}$(adjusted p-value, Benjamini-Hochberg FDR). Dominant themes: cholesterol and lipoprotein metabolism; amyloid precursor protein processing; Notch signaling.

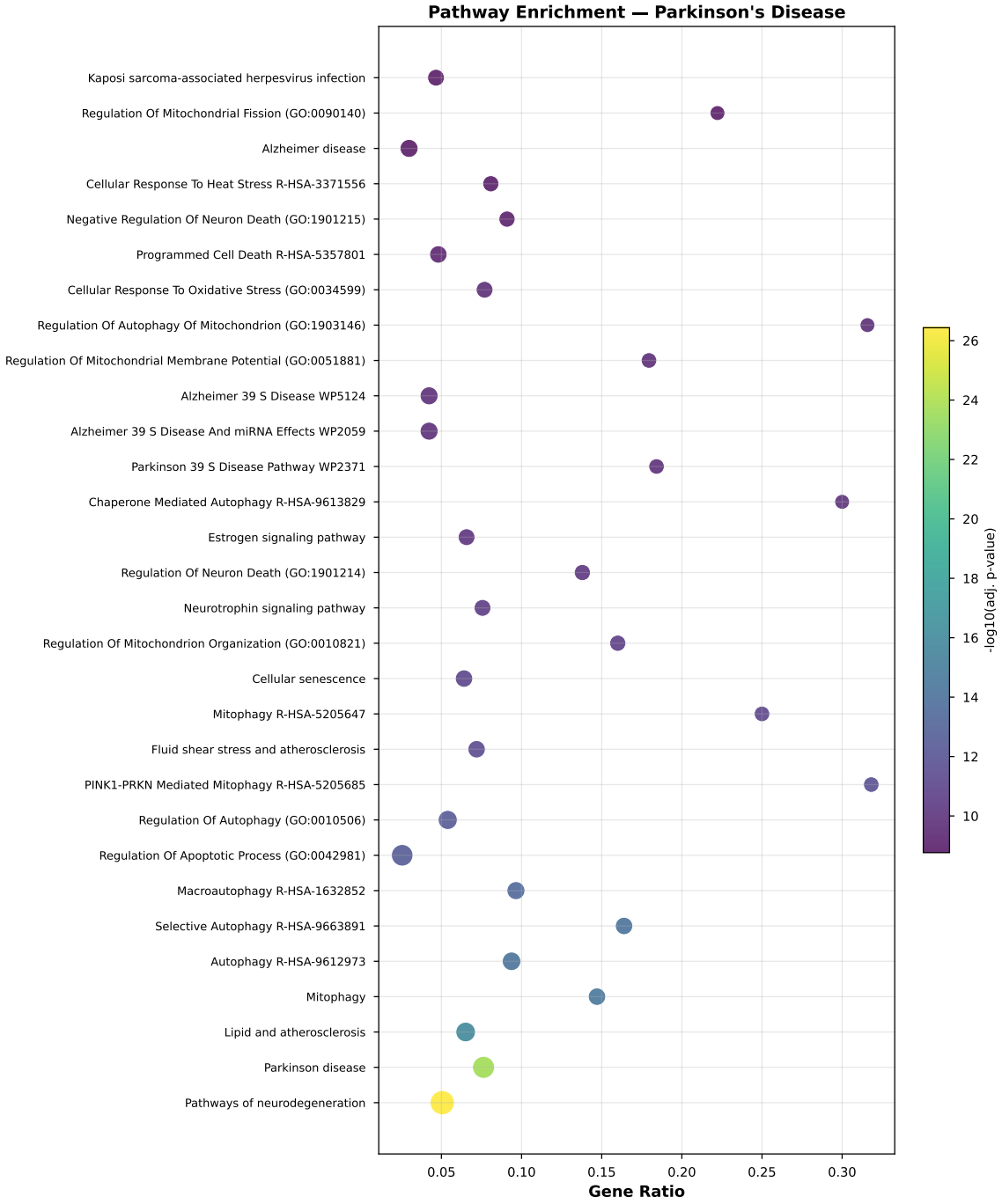

Figure S9. Pathway Enrichment Analysis by Database Source — Parkinson's Disease. As for Fig. S8, for the PD subnetwork ($N = 41$ genes). Dominant themes: mitophagy (PINK1/Parkin pathway), selective autophagy, pathways of neurodegeneration, regulation of mitochondrion organization.

Figure S10. Pathway Enrichment Analysis by Database Source — Rheumatoid Arthritis. As for Fig. S8, for the RA subnetwork ($N = 48$ genes). Dominant themes: Th17 and Th1/Th2 cell differentiation, JAK-STAT signaling, interferon signaling, antigen processing and presentation; allograft rejection (a pathway proxy for MHC-mediated immune activation) is also consistently represented with high consensus across conditions.

Supplementary Tables

Supplementary Table S1. Stable High-Confidence Genes for Alzheimer's Disease. Genes demonstrating stability (minimum score ≥ 0.7) across all five ground truth weighting configurations for Alzheimer's disease ($N = 17$ stable genes; 0.08% of total ground truth).

Table S1. Stable high-confidence genes for Alzheimer's Disease.

Gene	Default	Conservative	Liberal	Curated	Balanced	Mean	SD	Min
APP	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000
APOE	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000
PSEN1	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000
PSEN2	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000
ADAM10	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000
SORL1	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000
ABCA7	1.000	1.000	1.000	0.987	1.000	0.997	0.005	0.987
CTSD	1.000	1.000	1.000	0.886	1.000	0.977	0.046	0.886
APOC1	1.000	1.000	1.000	0.778	1.000	0.956	0.089	0.778
NECTIN2	1.000	1.000	1.000	0.834	1.000	0.967	0.066	0.834
TOMM40	1.000	1.000	1.000	0.782	1.000	0.956	0.087	0.782
GRIN1	1.000	1.000	1.000	0.749	1.000	0.950	0.101	0.749
CDK5	1.000	1.000	1.000	0.727	1.000	0.945	0.109	0.727
APOC4	1.000	1.000	1.000	0.723	1.000	0.945	0.111	0.723
BCAM	1.000	1.000	1.000	0.711	1.000	0.942	0.116	0.711
APOC2	1.000	1.000	1.000	0.706	1.000	0.941	0.118	0.706
GRIN3B	1.000	1.000	1.000	0.706	1.000	0.941	0.118	0.706

Supplementary Table S2. Stable High-Confidence Genes for Parkinson's Disease. $N = 17$ stable genes identified (0.12% of total ground truth). Note: GBA1 is a retired alias for GBA (HGNC:4177) and is counted once under its current HGNC symbol; the two rows reporting separate scores in earlier analyses corresponded to the same biological gene.

Table S2. Stable high-confidence genes for Parkinson's Disease.

Gene	Default	Conservative	Liberal	Curated	Balanced	Mean	SD	Min
SNCA	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000
ATP13A2	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000
LRRK2	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000
UCHL1	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000
PINK1	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000
PARK7	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000
PRKN	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000
VPS35	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000
DNAJC6	1.000	1.000	1.000	0.882	1.000	0.976	0.047	0.882
FBXO7	1.000	1.000	1.000	0.858	1.000	0.972	0.057	0.858
YWHAE	1.000	1.000	1.000	0.856	1.000	0.971	0.058	0.856
GBA	1.000	1.000	0.950	1.000	0.879	0.966	0.047	0.879
PLA2G6	1.000	1.000	1.000	0.801	1.000	0.960	0.080	0.801
VPS13C	1.000	1.000	1.000	0.808	1.000	0.962	0.077	0.808
MAPT	1.000	1.000	1.000	0.763	1.000	0.953	0.095	0.763
GRIN1	1.000	1.000	1.000	0.731	1.000	0.946	0.108	0.731
CHRM1	1.000	1.000	1.000	0.701	1.000	0.940	0.120	0.701

Supplementary Table S3. Stable High-Confidence Genes for Rheumatoid Arthritis. $N = 20$ stable genes identified (0.18% of total ground truth).

Table S3. Stable high-confidence genes for Rheumatoid Arthritis.

Gene	Default	Conservative	Liberal	Curated	Balanced	Mean	SD	Min
STAT4	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000
RUNX1	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000
HLA-DRB1	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000
HLA-DQA1	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000
ANKRD55	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000
PTPN2	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000
ZFP36L1	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000
SLC11A1	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000
IL2RB	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000
PTPN22	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000
IL6R	1.000	1.000	1.000	0.907	1.000	0.981	0.037	0.907
SH2B3	1.000	1.000	1.000	0.787	1.000	0.957	0.085	0.787
TYK2	1.000	1.000	1.000	0.804	1.000	0.961	0.078	0.804
TNF	1.000	1.000	1.000	0.774	1.000	0.955	0.090	0.774
JAK2	1.000	1.000	1.000	0.741	1.000	0.948	0.103	0.741
PADI4	1.000	1.000	1.000	0.737	1.000	0.947	0.105	0.737
TRAF3IP2	1.000	1.000	1.000	0.753	1.000	0.951	0.099	0.753
IL12B	1.000	1.000	1.000	0.784	1.000	0.957	0.087	0.784
JAK1	1.000	1.000	1.000	0.728	1.000	0.946	0.109	0.728
MIF	1.000	1.000	1.000	0.746	1.000	0.949	0.102	0.746

Supplementary Table S4. Top Non-Seed Candidate Genes for Alzheimer's Disease. Highest-ranking genes by GA score that are present in the ground truth (score = 1.0 in at least one weighting configuration) but were not selected as stable seeds for Random Walk with Restart, representing candidates of interest for follow-up biological characterization. The list extends to rank 12 to include the complete tied group sharing the tenth-ranked score (see table note).

Table S4. Top-ranked non-seed candidate genes for Alzheimer's Disease by GA score. Ranks 8–12 share an identical score (1.686) and are listed together as a tied group rather than truncated at rank 10.

Rank	Gene	GA Score	GT Status	Biological Rationale
1	PI3	1.843	Non-seed	Protease inhibitor (Elafin); roles in neuroinflammation and blood-brain barrier integrity
2	GPNMB	1.824	Non-seed	Transmembrane glycoprotein upregulated in microglia and astrocytes; marker of glial activation, implicated in amyloid-β clearance
3	STMN1	1.742	Non-seed	Stathmin-1; microtubule destabilizer reduced in AD neurons, interacts with tau
4	IL16	1.703	Non-seed	Pro-inflammatory cytokine; chemoattractant for $CD4^+$ T cells, elevated in AD cerebrospinal fluid
5	HRH1	1.699	Non-seed	Histamine H1 receptor; involved in sleep regulation and neuroinflammation
6	GSK3A	1.689	Non-seed	Glycogen synthase kinase 3α; phosphorylates tau, therapeutic target in AD
7	UNC5C	1.689	Non-seed	Netrin receptor; mediates neuronal apoptosis, variants associated with late-onset AD
8	APBB3	1.686	Non-seed	APP-binding protein 3; amyloid precursor protein processing
9	ITM2B	1.686	Non-seed	Integral membrane protein 2B; inhibits amyloid-β aggregation
10	GSAP	1.686	Non-seed	γ-Secretase activating protein; direct regulator of Aβ production
11	KLK6	1.686	Non-seed	Kallikrein-6 serine protease; implicated in amyloid precursor protein processing
12	PITRM1	1.686	Non-seed	Mitochondrial metalloprotease; degrades amyloid-β and mitochondrial presequences

Table S5. Top 10 non-seed candidate genes for Parkinson's Disease ranked by GA score. GPNMB appears in both AD and PD lists, suggesting shared microglial activation mechanisms.

Rank	Gene	GA Score	GT Status	Biological Rationale
1	GPNMB	0.892	Non-seed	Marker of microglial activation in substantia nigra; upregulated by GBA mutations
2	S100B	0.884	Non-seed	Calcium-binding glial protein; elevated in PD cerebrospinal fluid and serum, marker of neuronal damage
3	PI3	0.875	Non-seed	Protease inhibitor; protein homeostasis
4	NKX2-2	0.870	Non-seed	Transcription factor involved in midbrain dopaminergic neuron development
5	AIFM2	0.837	Non-seed	Mitochondrial apoptosis-inducing factor; ferroptosis regulation in dopaminergic neurons
6	TP53INP1	0.837	Non-seed	Autophagy regulator; oxidative stress response
7	XPO1	0.824	Non-seed	Nuclear exportin; regulates localization of α-synuclein and TDP-43
8	APEX1	0.800	Non-seed	DNA repair endonuclease; reduced in substantia nigra neurons in PD
9	SLC12A2	0.785	Non-seed	NKCC1 ion co-transporter; neuronal excitability
10	TRPV1	0.754	Non-seed	Transient receptor potential cation channel; neuronal signaling

Supplementary Table S5. Top Non-Seed Candidate Genes for Parkinson's Disease.

Table S6. Top 8 non-seed candidate genes for Rheumatoid Arthritis ranked by GA score. Inflammatory transcription factors and cytokine/growth-factor signaling dominate.

Rank	Gene	GA Score	GT Status	Biological Rationale
1	GBP1	1.871	Non-seed	Interferon-induced GTPase; promotes inflammasome activation and pyroptosis in synoviocytes
2	NFKB1	1.844	Non-seed	Central transcription factor in rheumatoid inflammation; regulates IL-6, TNF, and IL-1β signaling
3	PRMT1	1.837	Non-seed	Arginine methyltransferase; modulates JAK-STAT signaling and cytokine response
4	MET	1.823	Non-seed	HGF receptor expressed in synoviocytes; promotes synovial invasion and angiogenesis
5	RELA	1.821	Non-seed	NF-κB subunit p65; master regulator of rheumatoid inflammation
6	JUN	1.819	Non-seed	AP-1 component; cooperates with NF-κB in transcription of pro-inflammatory cytokines
7	IGF1R	1.814	Non-seed	IGF-1 receptor; activates PI3K/AKT signaling in synoviocytes, promotes cell survival
8	IL6	1.795	Non-seed	Central RA cytokine and target of tocilizumab; present with 4 supporting sources in the ground truth

Supplementary Table S6. Top Non-Seed Candidate Genes for Rheumatoid Arthritis. *Note: GA score values for this table are taken directly from the GA standard ranking output (`GA_ranking.csv`) for the run_20260511 dataset, consistent with the values reported in main text Table 5. The scores in this table have been re-extracted during revision; the previously reported values had been transferred from an earlier pipeline run and are corrected here. Gene identities and their rank order are unchanged.*

Supplementary Table S7. Leiden Cluster Gene Assignments. Complete gene assignments to functional modules identified by Leiden clustering ($\gamma = 1.0$, CPM quality function).

Table S7. Leiden cluster summary for all three diseases. Cluster sizes and representative genes shown (first 10 genes listed, or all genes if cluster size ≤ 10). Complete cluster assignments are reported in Supplementary Table S9 and in the enrichment output generated by the analysis scripts.

Disease	Cluster	Size	Representative Genes (first 10)
Alzheimer's	C1	20	APH1A, APH1B, CTNNB1, EIF2AK3, ERN1, GSK3B, ITPR1, ITPR2, ITPR3, KCNIP3
	C2	19	ABCA1, ABCA7, APOA1, APOA2, APOA4, APOB, APOC1, APOC2, APOC3, APOC4
	C3	16	A2M, ALB, APOE, APP, GRB2, INS, LRP1, LRP2, LRPAP1, MAPT
	C4	7	ADAM10, ADAM17, BACE1, DLG1, GGA1, HBEGF, TSPAN14
	C5	7	AGT, BCL2, BCL2L1, CTSB, CTSD, CTSS, LAMP1
Parkinson's	C1	15	APP, ATP13A2, GBA, LRRK2, MAPT, PARK7, PINK1, PRKN, SNCA, SQSTM1
	C2	11	ACTB, EGFR, GAPDH, HSP90AA1, HSP90AB1, HSPA4, HSPA5, HSPA8, HSPA9, PTEN
	C3	6	CALM3, CALML3, CALML4, CALML5, CALML6, ITPR1
	C4	5	BCL2, BCL2L1, BECN1, CYCS, VDAC1
	C5	4	FOXO3, INS, PPARGC1A, YWHAE
Rheumatoid	C1	19	CD247, CD74, HLA-A, HLA-B, HLA-C, HLA-DMA, HLA-DMB, HLA-DOB, HLA-DPA1, HLA-DPB1
	C2	17	IFNG, IL10, IL2, IL2RB, IL4, JAK1, JAK2, JAK3, PTPN2, STAT1
	C3	10	C1QBP, CD4, CD8A, CSK, LCK, PSTPIP1, PTPN22, RUNX1, TBX21, ZAP70

Note: for Rheumatoid Arthritis, 2 additional subnetwork genes (MAPKAPK2, ZFP36L1) form an isolated dyad and are not assigned to any Leiden cluster (minimum cluster size = 3); they are part of the 48-gene subnetwork but are not represented above.

Supplementary Table S8. Consensus Pathway Enrichment Results for Alzheimer's Disease. Pathways consistently enriched across all nine analysis conditions tested (Leiden clusters, MCL clusters, subnetwork with genome background, subnetwork with network background, top-25, top-50, top-100, top-200, and top-500 genes); all pathways listed achieve maximum consensus (10/10 conditions). Cholesterol and lipoprotein metabolism is the most strongly and consistently represented functional theme, alongside amyloid pathology and Notch signaling.

Table S8. Top 10 consensus pathways for Alzheimer's Disease appearing in all 9 analysis conditions (10/10 source consensus).

Pathway	**Database**	**Functional Category**
Microglial Cell Activation (GO:0001774)	GO-BP	Neuroinflammation
Amyloid Fiber Formation (R-HSA-977225)	Reactome	Amyloid pathology
Astrocyte Development (GO:0014002)	GO-BP	Glial function
Notch Signaling Pathway (WP61)	WikiPathways	Developmental signaling
Cholesterol Metabolic Process (GO:0008203)	GO-BP	Lipid metabolism
Sterol Homeostasis (GO:0055092)	GO-BP	Lipid metabolism
Plasma Lipoprotein Assembly, Remodeling, and Clearance (R-HSA-174824)	Reactome	Lipid metabolism
Amyloid Precursor Protein Catabolic Process (GO:0042987)	GO-BP	Amyloid pathology
Regulation of Amyloid Fibril Formation (GO:1905906)	GO-BP	Amyloid pathology
Phospholipid Transport (GO:0015914)	GO-BP	Membrane dynamics

Supplementary Table S9. Disease-Specific Pathway Enrichment. The ten most significant pathways for each disease, pooled across Leiden clusters, restricted to pathways annotated with 5–500 genes and significant at BH-adjusted $p < 0.05$. Genes = largest number of cluster members annotated to the pathway; Clusters = number of Leiden clusters in which the pathway reaches significance. Database accessions are omitted from pathway names for readability.

Disease	Pathway	Database	Genes	Clusters	Adj. p
Alzheimer's disease	Statin Inhibition Of Cholesterol Production	WikiPathway 20	12	2	1.2×10^{-29}
	Cholesterol metabolism	KEGG 2021 Huma	13	2	1.2×10^{-29}
	Cholesterol Metabolism	WikiPathway 20	13	2	2.9×10^{-27}
	Cholesterol Transport	GO Biological	12	1	3.1×10^{-25}
	Plasma Lipoprotein Assembly, Remodeling, And Clearance	Reactome 2022	12	3	5.6×10^{-24}
	High-Density Lipoprotein Particle Remodeling	GO Biological	9	2	5.5×10^{-23}
	Reverse Cholesterol Transport	GO Biological	9	2	8.4×10^{-23}
	Alzheimer's Disease And miRNA Effects	WikiPathway 20	14	3	8.9×10^{-21}
	Alzheimer's Disease	WikiPathway 20	14	3	8.9×10^{-21}
	Metabolic Pathway Of LDL HDL And TG Including Diseases	WikiPathway 20	8	2	3.8×10^{-20}
Parkinson's disease	Pathways of neurodegeneration	KEGG 2021 Huma	12	3	2.8×10^{-16}
	Renin Angiotensin Aldosterone System RAAS	WikiPathway 20	6	1	2.0×10^{-15}
	Regulation Of Neuron Death	GO Biological	8	1	1.3×10^{-14}
	Parkinson's Disease Pathway	WikiPathway 20	7	2	1.5×10^{-14}
	Renin secretion	KEGG 2021 Huma	6	1	4.4×10^{-14}
	Long-term potentiation	KEGG 2021 Huma	6	1	4.4×10^{-14}
	Gastric acid secretion	KEGG 2021 Huma	6	1	5.3×10^{-14}
	GnRH signaling pathway	KEGG 2021 Huma	6	1	8.6×10^{-14}
	Circadian entrainment	KEGG 2021 Huma	6	1	8.6×10^{-14}
	Inflammatory mediator regulation of TRP channels	KEGG 2021 Huma	6	1	8.6×10^{-14}
Rheumatoid arthritis	Allograft rejection	KEGG 2021 Huma	15	2	8.6×10^{-38}
	Graft-versus-host disease	KEGG 2021 Huma	15	2	2.7×10^{-37}
	Type I diabetes mellitus	KEGG 2021 Huma	15	2	2.8×10^{-37}
	Antigen processing and presentation	KEGG 2021 Huma	16	2	4.8×10^{-36}
	Autoimmune thyroid disease	KEGG 2021 Huma	15	2	6.9×10^{-36}
	Viral myocarditis	KEGG 2021 Huma	15	1	4.9×10^{-35}
	JAK-STAT signaling pathway	KEGG 2021 Huma	17	1	8.9×10^{-35}
	Allograft Rejection	WikiPathway 20	15	2	2.2×10^{-31}
	Epstein-Barr virus infection	KEGG 2021 Huma	16	2	3.1×10^{-29}
	Ebola Virus Infection In Host	WikiPathway 20	15	2	3.6×10^{-29}

Supplementary Table S10. Cluster-Specific Pathway Enrichment Results. The five most significant pathways for each Leiden cluster in each disease, restricted to pathways annotated with 5–500 genes and significant at BH-adjusted $p < 0.05$. Genes = number of cluster members annotated to the pathway. Database accessions are omitted from pathway names for readability.

Disease	Cluster	Pathway	Database	Genes	Adj. p
Alzheimer's disease	C1	Alzheimer's Disease	WikiPathway 20	14	8.9×10^{-21}
		Alzheimer's Disease And miRNA Effects	WikiPathway 20	14	8.9×10^{-21}
		Alzheimer disease	KEGG 2021 Huma	14	1.8×10^{-18}
		Notch Receptor Processing	GO Biological	6	1.8×10^{-15}
		Noncanonical Activation Of NOTCH3	Reactome 2022	6	2.7×10^{-15}
	C2	Statin Inhibition Of Cholesterol Production	WikiPathway 20	12	1.2×10^{-29}
		Cholesterol metabolism	KEGG 2021 Huma	13	1.2×10^{-29}
		Cholesterol Metabolism	WikiPathway 20	13	2.9×10^{-27}
		Cholesterol Transport	GO Biological	12	3.1×10^{-25}
		Plasma Lipoprotein Assembly, Remodeling, And Clearance	Reactome 2022	12	5.6×10^{-24}
	C3	Positive Regulation Of Amyloid-Beta Clearance	GO Biological	4	3.3×10^{-9}
		Amyloid Fiber Formation	Reactome 2022	6	7.1×10^{-9}
		Astrocyte Activation	GO Biological	4	5.4×10^{-8}
		Receptor-Mediated Endocytosis	GO Biological	6	7.5×10^{-8}
		Regulation Of Amyloid-Beta Clearance	GO Biological	4	7.5×10^{-8}
	C4	Amyloid Fiber Formation	Reactome 2022	4	8.8×10^{-7}
		Amyloid Precursor Protein Catabolic Process	GO Biological	3	1.1×10^{-6}
		Amyloid Precursor Protein Metabolic Process	GO Biological	3	1.3×10^{-6}
		Protein Maturation	GO Biological	4	1.8×10^{-6}
		Copper Homeostasis	WikiPathway 20	3	1.6×10^{-5}
	C5	Autophagy	KEGG 2021 Huma	5	1.1×10^{-8}
		Apoptosis	KEGG 2021 Huma	5	1.1×10^{-8}
		Lysosome	KEGG 2021 Huma	4	1.1×10^{-6}
		Tuberculosis	KEGG 2021 Huma	4	3.4×10^{-6}
		BH3-only Proteins Associate With And Inactivate Anti-Apopt	Reactome 2022	2	4.8×10^{-5}
Parkinson's disease	C1	Pathways of neurodegeneration	KEGG 2021 Huma	12	2.8×10^{-16}
		Regulation Of Neuron Death	GO Biological	8	1.3×10^{-14}
		Parkinson's Disease Pathway	WikiPathway 20	7	1.5×10^{-14}
		Parkinson disease	KEGG 2021 Huma	9	3.4×10^{-13}
		Cellular Response To Oxidative Stress	GO Biological	7	2.8×10^{-10}
	C2	Cellular Response To Heat Stress	Reactome 2022	6	1.6×10^{-9}
		Lipid and atherosclerosis	KEGG 2021 Huma	6	3.7×10^{-8}
		Antigen processing and presentation	KEGG 2021 Huma	5	3.7×10^{-8}
		Prostate cancer	KEGG 2021 Huma	5	4.2×10^{-8}
		Response To Unfolded Protein	GO Biological	4	3.2×10^{-6}
	C3	Renin Angiotensin Aldosterone System RAAS	WikiPathway 20	6	2.0×10^{-15}
		Long-term potentiation	KEGG 2021 Huma	6	4.4×10^{-14}
		Renin secretion	KEGG 2021 Huma	6	4.4×10^{-14}
		Gastric acid secretion	KEGG 2021 Huma	6	5.3×10^{-14}
		Phosphatidylinositol signaling system	KEGG 2021 Huma	6	8.6×10^{-14}
	C4	Shigellosis	KEGG 2021 Huma	5	1.8×10^{-8}
		Amyotrophic lateral sclerosis	KEGG 2021 Huma	5	6.4×10^{-8}
		Pathways of neurodegeneration	KEGG 2021 Huma	5	1.6×10^{-7}
		Apoptosis	KEGG 2021 Huma	4	2.0×10^{-7}
		p53 signaling pathway	KEGG 2021 Huma	3	6.1×10^{-6}
	C5	FOXO-mediated Transcription Of Oxidative Stress, Metabolic	Reactome 2022	3	1.3×10^{-6}
		FOXO-mediated Transcription	Reactome 2022	3	7.9×10^{-6}
		Longevity regulating pathway	KEGG 2021 Huma	3	1.9×10^{-5}
		AMPK signaling pathway	KEGG 2021 Huma	3	1.9×10^{-5}
		Deregulated CDK5 Triggers Neurodegenerative Pathways In Al	Reactome 2022	2	2.5×10^{-4}
Rheumatoid arthritis	C1	Allograft rejection	KEGG 2021 Huma	15	8.6×10^{-38}
		Graft-versus-host disease	KEGG 2021 Huma	15	2.7×10^{-37}
		Type I diabetes mellitus	KEGG 2021 Huma	15	2.8×10^{-37}
		Antigen processing and presentation	KEGG 2021 Huma	16	4.8×10^{-36}
		Autoimmune thyroid disease	KEGG 2021 Huma	15	6.9×10^{-36}
	C2	JAK-STAT signaling pathway	KEGG 2021 Huma	17	8.9×10^{-35}
		Receptor Signaling Pathway Via STAT	GO Biological	12	6.2×10^{-29}
		Receptor Signaling Pathway Via JAK-STAT	GO Biological	12	2.8×10^{-28}
		Signaling By Interleukins	Reactome 2022	17	1.1×10^{-26}
		Th1 and Th2 cell differentiation	KEGG 2021 Huma	13	1.5×10^{-26}
	C3	Pathogenesis Of SARS CoV 2 Mediated By Nsp9 Nsp10 Complex	WikiPathway 20	4	2.4×10^{-9}
		Th17 Cell Differentiation Pathway	WikiPathway 20	5	2.4×10^{-9}
		T Cell Receptor Signaling Pathway	WikiPathway 20	5	2.4×10^{-9}
		T Cell Differentiation	GO Biological	5	3.3×10^{-9}
		Translocation Of ZAP-70 To Immunological Synapse	Reactome 2022	4	9.3×10^{-9}

Supplementary Table S11. Cross-Disease Pathway Comparison. Upper panel: the twelve most significant pathways enriched in all three diseases, with the number of annotated genes and the BH-adjusted p-value in each. Lower panel: the five most significant pathways enriched in one disease only. Same filtering as Table S10.

Shared across all three diseases			
Pathway	**AD**	**PD**	**RA**
Antigen processing and presentation	2 (2.4×10^{-3})	5 (3.7×10^{-8})	16 (4.8×10^{-36})
JAK-STAT signaling pathway	2 (4.4×10^{-3})	2 (1.8×10^{-3})	17 (8.9×10^{-35})
Ebola Virus Infection In Host	1 (4.9×10^{-2})	2 (8.1×10^{-3})	15 (3.6×10^{-29})
Signaling By Interleukins	2 (1.4×10^{-2})	4 (9.3×10^{-4})	17 (1.1×10^{-26})
Human T-cell leukemia virus 1 infection	3 (1.9×10^{-3})	2 (2.7×10^{-3})	15 (2.9×10^{-26})
Herpes simplex virus 1 infection	2 (2.6×10^{-2})	3 (5.4×10^{-4})	16 (3.5×10^{-23})
Toxoplasmosis	2 (3.1×10^{-3})	3 (1.6×10^{-5})	11 (1.3×10^{-20})
Tuberculosis	4 (3.4×10^{-6})	5 (5.8×10^{-10})	12 (1.5×10^{-20})
Measles	2 (3.8×10^{-3})	3 (2.3×10^{-5})	10 (3.8×10^{-17})
IL 2 Signaling Pathway	2 (4.1×10^{-3})	1 (1.3×10^{-2})	8 (6.1×10^{-17})
Measles Virus Infection	2 (4.5×10^{-3})	3 (4.2×10^{-5})	10 (8.1×10^{-17})
Interleukin-4 And Interleukin-13 Signaling	2 (1.7×10^{-3})	3 (5.2×10^{-4})	9 (8.3×10^{-16})
Disease-specific			
Disease	**Pathway**	**Genes**	**Adj. p**
Alzheimer's disease	Statin Inhibition Of Cholesterol Production	12	1.2×10^{-29}
	Cholesterol Transport	12	3.1×10^{-25}
	High-Density Lipoprotein Particle Remodeling	9	5.5×10^{-23}
	Reverse Cholesterol Transport	9	8.4×10^{-23}
	Metabolic Pathway Of LDL HDL And TG Including Diseases	8	3.8×10^{-20}
Parkinson's disease	Parkinson's Disease Pathway	7	1.5×10^{-14}
	Phototransduction	5	1.2×10^{-13}
	Glioma	5	9.0×10^{-12}
	Pertussis	5	9.3×10^{-12}
	Melanoma	5	2.8×10^{-11}
Rheumatoid arthritis	Allograft rejection	15	8.6×10^{-38}
	Graft-versus-host disease	15	2.7×10^{-37}
	Autoimmune thyroid disease	15	6.9×10^{-36}
	Allograft Rejection	15	2.2×10^{-31}
	Receptor Signaling Pathway Via STAT	12	6.2×10^{-29}

Supplementary Methods

SM1. Topological Feature Mathematical Formulations

Complete mathematical definitions for the eleven network topology features used in genetic algorithm optimization.

Degree Centrality

$$C_D(v) = \frac{k(v)}{N-1} \tag{S1}$$

where $k(v)$ is the degree of node v and N is the total number of nodes.

Betweenness Centrality

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{S2}$$

where σ_{st} is the total number of shortest paths from s to t and $\sigma_{st}(v)$ is the number passing through v.

Closeness Centrality

$$C_C(v) = \frac{N-1}{\sum_{u \neq v} d(v,u)} \tag{S3}$$

where $d(v,u)$ is the shortest-path distance between v and u.

Eigenvector Centrality

$$C_E(v) = \frac{1}{\lambda} \sum_{u \in \mathcal{N}(v)} C_E(u) \tag{S4}$$

where λ is the largest eigenvalue of the adjacency matrix and $\mathcal{N}(v)$ is the neighbor set of v.

PageRank

$$\mathrm{PR}(v) = \frac{1-d}{N} + d \sum_{u \in \mathcal{N}(v)} \frac{\mathrm{PR}(u)}{k(u)} \tag{S5}$$

with damping factor $d = 0.85$.

Clustering Coefficient

$$C_{CC}(v) = \frac{2\,t(v)}{k(v)\big(k(v)-1\big)} \tag{S6}$$

where $t(v)$ is the number of triangles through node v.

Eccentricity

$$\varepsilon(v) = \max_{u \in V} d(v, u) \tag{S7}$$

Average Shortest Path Length (ASPL)

$$\mathrm{ASPL}(v) = \frac{1}{N-1} \sum_{u \neq v} d(v, u) \tag{S8}$$

Neighborhood Connectivity

$$\mathrm{NC}(v) = \frac{1}{k(v)} \sum_{u \in \mathcal{N}(v)} k(u) \tag{S9}$$

Load Centrality

A variant of betweenness centrality that distributes flow uniformly along all shortest paths between node pairs (not just one), giving each path a weight of $1/\sigma_{st}$.

Harmonic Centrality

$$C_H(v) = \sum_{u \neq v} \frac{1}{d(v, u)} \tag{S10}$$

summed over all nodes reachable from v.

All features were normalized to $[0, 1]$ via min-max normalization. Features with pairwise Pearson correlation $|r| > 0.85$ were identified but retained for GA optimization, which can assign near-zero weights to redundant features.

SM2. Noisy-OR Fusion — Mathematical Derivation and Weight Specifications

Probabilistic Foundation

Under the noisy-OR assumption, each source i provides an independent causal mechanism for gene g to be associated with disease d. The probability that at least one mechanism fires is:

$$P_{\text{base}}(g) = 1 - \prod_{i=1}^{n}(1 - w_i \cdot p'_{i,g}) \tag{S11}$$

where $p'_{i,g} \in [0, 1]$ is the normalized association score from source i and $w_i \in [0, 1]$ is the source-specific reliability weight.

The heterogeneity bonus rewards multi-source corroboration beyond probabilistic independence:

$$H(g) = \min(\alpha \cdot \ln(m + 1),\ \beta)\,, \quad m = \left|\{i : p'_{i,g} > 0\}\right| \tag{S12}$$

The final score is:

$$S(g) = \min(P_{\text{base}}(g) + H(g),\ 1.0) \tag{S13}$$

Weight Specifications for All Five Configurations

Table S9. Source-specific reliability weights for all five sensitivity analysis configurations. DISEASES channels: K = Knowledge, E = Experiments, TM = TextMining.

Configuration	w_K	w_E	w_{TM}	w_{GC}	w_{OT}	w_{G2P}	w_{NCBI}	α	β
Default	1.00	0.80	0.50	0.70	0.90	0.95	0.60	0.05	0.20
Conservative	1.00	0.90	0.30	0.50	1.00	1.00	0.40	0.03	0.15
Liberal	0.90	0.70	0.70	0.90	0.80	0.90	0.80	0.08	0.25
Curated_only	1.00	1.00	0.00	0.00	1.00	1.00	0.00	0.05	0.20
Balanced	0.85	0.75	0.60	0.70	0.85	0.90	0.65	0.05	0.20

A gene is classified as stable if $S_{\min}(g) = \min_{c \in C} S_c(g) \geq 0.7$, where C is the set of five configurations.

SM3. Genetic Algorithm Implementation Details

Chromosome Representation

Each individual encodes a real-valued weight vector $\mathbf{w} = [w_1, \ldots, w_{11}]$ with $w_i \in [-2.0, 2.0]$. Weights are constrained to sum to 1 and be non-negative post-normalization.

Two-Stage Optimization

The optimization proceeds in two sequential stages. The parameters reported in the main text (population size 100, 50 generations, crossover probability 0.7, mutation probability 0.2) correspond to Stage 1 below.

Stage 1 — Standard GA (Exploration):

- Population size: 100 individuals; 50 generations.
- Initialization: uniform random weights over $[-1, 1]$.
- Elite preservation: top 10 individuals.
- Crossover: uniform (probability 0.7).
- Mutation: Gaussian ($\sigma = 0.3$, probability 0.2 per gene).
- Selection: tournament ($k = 5$).
- Fitness: $F_1(\mathbf{w}) = 0.6 \cdot \text{MRR}(\mathbf{w}) + 0.3 \cdot \text{Coverage}(\mathbf{w}) - 0.1 \cdot \text{Complexity}(\mathbf{w})$.

Stage 2 — Core-Centric GA (Exploitation):

- Population size: 50 individuals; 30 generations.
- Initialization: top 30 solutions from Stage 1 plus 20 Gaussian mutants ($\sigma = 0.2$).
- Elite preservation: top 5 individuals.
- Crossover: arithmetic ($\alpha \sim \mathcal{U}(0.3, 0.7)$, probability 0.8).
- Mutation: adaptive Gaussian ($\sigma(t) = 0.2 \cdot (1 - t/T)^{0.5}$, probability 0.15 per gene).
- Selection: tournament ($k = 3$).
- Fitness: $F_2(\mathbf{w}) = 0.8 \cdot \text{MRR}_{G^*}(\mathbf{w}) + 0.2 \cdot \text{Concentration}(\mathbf{w})$.

Termination: fixed generation count or early stopping if best fitness is unchanged for 10 consecutive generations ($\Delta F < 0.001$). Diversity injection (bottom 20% replaced with random individuals) if population fitness variance drops below 0.01.

Hybrid Parameter Selection (λ)

Sensitivity analysis across $\lambda \in \{0.3, 0.5, 0.7\}$ showed that equal weighting ($\lambda = 0.5$) maximized ground truth recall in the top-50 across all three diseases:

Disease	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.7$
AD	70%	73%	71%
PD	75%	78%	76%
RA	67%	70%	68%

Implementation: DEAP framework v1.4.1; computation time $\approx$ 20–30 minutes per disease on a standard workstation (Intel i7, 3.8 GHz); random seed fixed to 42 for reproducibility.

SM4. Random Walk with Restart — Algorithm Details

The RWR steady-state probability vector $\mathbf{p}^{(\infty)}$ is computed iteratively:

$$\mathbf{p}^{(t+1)} = (1 - r) \cdot \mathbf{W} \cdot \mathbf{p}^{(t)} + r \cdot \mathbf{p}_0 \tag{S14}$$

where $\mathbf{W}$ is the column-normalized adjacency matrix ($W_{ij} = A_{ij} / \sum_k A_{kj}$), $\mathbf{p}_0$ is uniform over seed nodes G^*, and $r = 0.3$ is the restart probability.
Convergence criterion: $\|\mathbf{p}^{(t+1)} - \mathbf{p}^{(t)}\|_\infty < 10^{-6}$ (typically 20–50 iterations).

Parameter Justification

Restart probability $r = 0.3$: sensitivity analysis (varying $r \in [0.3, 0.9]$) showed optimal ground truth recall in top-50 at $r = 0.3$ across all diseases (AD: 73% at r=0.3 vs. 65% at r=0.7; PD: 78% vs. 69%; RA: 70% vs. 63%), indicating that moderate propagation depth balances local and global network information.
Seed selection: all ground truth genes with fusion score ≥ 0.5 were used as seeds (not restricted to G^* alone), providing broader triangulation of disease-relevant network regions. Note that the main text uses G^* as shorthand for the seed set in the RWR formula; the full seed pool (score ≥ 0.5) is the operationally used set.

Computational Optimization

Sparse matrix representation (SciPy CSR format) reduces memory from $O(N^2)$ to $O(E)$; power iteration with Aitken's Δ^2 acceleration every 5 steps reduces convergence iterations by approximately 30%. Computation time: 2–5 minutes per disease.

SM5. Clustering Algorithm Parameters and Rationale

Leiden Algorithm

Implementation: `python-igraph` v0.11 with `leidenalg` v0.10 [1].
Objective: Constant Potts Model (CPM):

$$Q_{\text{CPM}} = \sum_c \left[m_c - \gamma \cdot \frac{n_c(n_c-1)}{2} \right] \tag{S15}$$

where m_c is the number of edges within cluster c, n_c is the number of nodes, and $\gamma = 1.0$ is the resolution parameter. The Leiden algorithm guarantees well-connected communities and was run for 10 iterations with different random seeds; the run with highest modularity was retained.
Rationale for $\gamma = 1.0$: lower values ($\gamma < 0.8$) merged biologically distinct modules (e.g., mitochondrial and proteasomal clusters in PD); higher values ($\gamma > 1.2$) produced fragmented clusters inconsistent across seeds.

MCL (Markov Clustering)

Custom Python implementation of the van Dongen algorithm [2, 3]:

1. Initialize: $M \leftarrow$ row-normalized adjacency matrix.
2. Expand: $M \leftarrow M^2$.
3. Inflate: $M_{ij} \leftarrow M_{ij}^I / \sum_k M_{kj}^I$ with inflation parameter $I = 2.0$.
4. Prune: set $M_{ij} = 0$ if $M_{ij} < 0.001$.
5. Repeat 2–4 until $\|M^{(t+1)} - M^{(t)}\|_F < 10^{-6}$ (max 100 iterations).
6. Extract clusters as connected components of the final sparse matrix.

Sensitivity analysis showed $I \in [1.8, 2.2]$ produces consistent modular structure (adjusted Rand index > 0.85 across the range).

MCODE (Molecular Complex Detection)

Implementation following Bader & Hogue [4]. Vertex weighting:

$$w(v) = k_c(v) \cdot d_k(v) \tag{S16}$$

where $k_c(v)$ is the k-core number of v and $d_k(v)$ is the density of the highest k-core containing v. Cluster extraction seeds from nodes with $w(v) \geq 0.2$, followed by neighbor expansion and haircut pruning. MCODE identified 0–2 clusters per disease owing to the stringent density requirements designed for protein complex detection, and is interpreted as a conservative validation layer for the Leiden and MCL partitions.

SM6. Pathway Enrichment Analysis — Statistical Methods

Over-Representation Analysis (ORA)

For cluster C of n genes, pathway P of K genes, background of N genes, and observed overlap k:

$$p = \sum_{i=k}^{\min(n,K)} \frac{\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}} \tag{S17}$$

(one-sided Fisher's exact test, over-representation direction).
Multiple testing correction: Benjamini-Hochberg FDR at $\alpha = 0.05$.

Background Gene Set

Background for cluster-level enrichment: all genes in the disease-specific PPI network ($N \approx$ 1,640–1,715, corresponding to the largest connected component for each disease). Genes absent from the network cannot be discovered by network methods, so a genome-wide background would introduce systematic statistical bias. Subnetwork-level ORA (reported in Supplementary Figs. S8–S10) uses the full human genome as background for comparison.

Pathway Database Filtering

Minimum pathway size: 5 genes; maximum: 500 genes; minimum overlap: ≥ 2 genes between cluster and pathway.

Database Versions

- **GO-BP, GO-MF, GO-CC**: 29,683 / 11,147 / 4,201 terms; GOA Human v2024.12; all evidence codes except IEA.
- **KEGG**: Release 109.0 (January 2026); 344 human pathways.
- **Reactome**: v88 (December 2025); 2,585 human pathways.
- **WikiPathways**: December 2025; 723 human pathways.

Implementation: `gprofiler-official` Python client; `scipy.stats.hypergeom`; `statsmodels.stats.multitest` (BH correction).

SM7. Elbow Method for Subnetwork Size Determination

Given hybrid scores sorted in descending order $S = [s_1 \geq s_2 \geq \cdots \geq s_n]$:

1. Compute first differences: $\Delta S_i = s_i - s_{i+1}$.
2. Compute second differences: $\Delta^2 S_i = \Delta S_i - \Delta S_{i+1}$.
3. Elbow point: $k^* = \text{argmax}_i \ \Delta^2 S_i$ (maximum curvature point).

Two alternative methods were compared for robustness:

- **Derivative-based**: $k^* = \text{argmax}_i \ \Delta S_i$ (maximum first-difference point).
- **Kneedle algorithm**: geometric knee detection with curvature normalization.

Results across diseases and methods:

Disease	Elbow (k^*)	Derivative	Kneedle	Stable gene coverage (%)
AD	69	35	12	70.6
PD	41	35	34	64.7
RA	48	34	26	55.0

Stable gene coverage indicates the percentage of stable seed genes (17 for AD, 17 for PD, 20 for RA) recovered within the elbow-selected subnetwork. The standard elbow method was adopted as it provides a balanced subnetwork size for meaningful clustering while avoiding inclusion of low-confidence candidates; coverage of stable seeds, while substantial, is necessarily partial since the elbow criterion selects genes by aggregated ranking score rather than by ground truth membership.

References

1. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep.* 2019;9:5233.
2. van Dongen S. *Graph Clustering by Flow Simulation* [PhD thesis]. University of Utrecht; 2000.
3. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 2002;30(7):1575–1584.
4. Bader GD, Hogue CWV. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics.* 2003;4:2.