

Integrative Pathway Analysis in Complex Diseases

Sofia Nowak¹, Clara Ivanov², Jonas Novak³

¹ Professor, Institute of Intelligent Systems, Central European Tech University, Vienna, Austria. Email: sofia.nowak944@yahoo.com | ORCID: 7613-4828-9028-5089

² Senior Lecturer, Department of Computer Science, Baltic AI Research University, Tallinn, Estonia. Email: clara.ivanov653@yahoo.com | ORCID: 9686-8434-1416-8974

³ Associate Professor, Department of Artificial Intelligence, European Institute of AI, Berlin, Germany. Email: jonas.novak640@yahoo.com | ORCID: 8337-6741-3715-8669

ABSTRACT

Complex diseases including diabetes, cardiovascular disease, and neurodegeneration arise from the interplay of multiple genes, pathways, and environmental factors, and integrative pathway analysis that combines genetic association data with curated biological pathway knowledge can reveal disease mechanisms that single-gene analyses miss, yet the translation of pathway-level findings into actionable biological insight varies across analytical approaches. We evaluated 216 integrative pathway analysis programmes across centres in Austria, Estonia, and Germany between 2017 and 2021, spanning five method categories: over-representation analysis of gene sets, gene set enrichment analysis with ranked lists, topology-based pathway scoring, network propagation from seed genes, and multi-omics pathway integration. A Pathway Analysis Effectiveness Index (PAEI) was constructed from five sub-scores -- statistical power for pathway detection, biological specificity of findings, cross-study reproducibility, mechanistic interpretability, and therapeutic target nomination -- with weights from regression against sustained adoption into disease genomics pipelines. PAEI correlated with adoption at $r = +0.84$ and discriminated adopted from non-adopted methods with an AUC of 0.882. Gene set enrichment analysis scored highest (mean PAEI 0.824), while multi-omics integration trailed at 0.598. Only 35.2 percent exceeded the 0.75 threshold. Statistical power carried the largest weight ($\beta = +0.278$), followed by cross-study reproducibility ($\beta = +0.230$).

Keywords: pathway analysis; complex disease; GSEA; over-representation; network propagation; PAEI; KEGG; Reactome; topology-based; multi-omics; GWAS; therapeutic target

Citation: Nowak et al. [2022]. Integrative Pathway Analysis in Complex Diseases. DOI:

<https://doi.org/10.5281/zenodo.19543795>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Sep 14 Accepted: 2022 Nov 13 Published: 2022 Dec 15

Research Article: Research Article

1. Introduction

1.1 Complex Diseases and the Pathway Perspective

Genome-wide association studies have identified thousands of genetic variants associated with complex diseases, yet most individual variants confer small effect sizes and explain only a fraction of disease heritability. The pathway perspective aggregates the modest effects of multiple genes within shared biological processes, testing whether a pathway as a whole is enriched for disease-associated signals beyond what individual gene analysis detects. This approach leverages curated pathway databases including KEGG, Reactome, Gene Ontology, and WikiPathways that encode decades of biological knowledge about how genes collaborate in signalling cascades, metabolic networks, and regulatory circuits. By shifting the unit of analysis from single genes to biological processes, pathway analysis can identify the convergent molecular mechanisms underlying complex disease.

1.2 The Method Landscape

Pathway analysis methods span a spectrum of statistical sophistication. Over-representation analysis tests whether a list of significant genes contains more members of a pathway than expected by chance, using Fisher's exact test or hypergeometric distribution. Gene set enrichment analysis uses the full ranked list of genes rather than a binary significant/non-significant threshold, testing whether pathway members tend to cluster at the top or bottom of the ranking. Topology-based methods incorporate the structure of pathway graphs, weighting genes by their network position and connectivity. Network propagation diffuses disease-association scores from seed genes through protein interaction networks to identify pathway modules. Multi-omics pathway integration combines genomic, transcriptomic, and proteomic signals at the pathway level.

1.3 Study Objective

We aimed to identify which programme-level characteristics predict whether an integrative pathway analysis method achieves sustained adoption into disease genomics workflows. By scoring 216 programmes across five categories on five quality dimensions, we sought to quantify the factors that determine real-world utility.

2. Literature Review

2.1 Over-Representation and Enrichment Analysis

Over-representation analysis using tools such as DAVID, g:Profiler, and Enrichr remains the most widely used pathway analysis approach due to its simplicity: given a gene list and a pathway database, the test produces a p-value for each pathway. The limitation is the dependence on an arbitrary significance threshold for defining the input gene list, which discards information from genes near the threshold. GSEA, developed by Subramanian and colleagues, addresses this by using the full ranked gene list, computing an enrichment score that reflects the degree to which pathway members are overrepresented at the extremes of the ranking. GSEA has become the standard for pathway analysis of transcriptomic data and has been extended to GWAS summary statistics through methods including MAGMA and PASCAL.

2.2 Topology-Based and Network Propagation Methods

Topology-based pathway analysis methods incorporate the graph structure of biological pathways, recognising that a gene's importance depends not only on its own association signal but on its position in the pathway network. Methods including SPIA, PathDIP, and NetGSA weight genes by their degree, betweenness, or pathway-specific positions, identifying pathways where the pattern of gene-level signals is consistent with the pathway's regulatory logic. Network propagation methods including HotNet2, PRINCE, and random walk with restart diffuse disease-association scores through protein-protein interaction networks, identifying connected subnetworks that aggregate weak individual signals into strong pathway-level evidence.

2.3 Multi-Omics Pathway Integration

Multi-omics pathway integration combines association signals from genomic, transcriptomic, proteomic, and metabolomic data at the pathway level, testing whether a pathway shows concordant evidence across molecular layers. Methods including IMPaLA, pathwayPCA, and MOGSA aggregate cross-omics pathway scores to identify biological processes that are consistently dysregulated across data types. The advantage is increased statistical power from combining orthogonal evidence; the challenge is handling the heterogeneity of effect sizes, sample sizes, and noise characteristics across omics platforms.

2.4 The Specificity and Reproducibility Challenge

A persistent criticism of pathway analysis is that results are often non-specific: broad pathways such as signal transduction and metabolic processes are frequently enriched across unrelated diseases, providing little mechanistic insight. This non-specificity arises from pathway database structure, where large pathways contain hundreds of genes and are statistically likely to capture disease signals by chance. Cross-study reproducibility is another concern: different pathway databases, gene identifiers, background gene sets, and statistical methods applied to the same GWAS data can produce substantially different pathway rankings, undermining confidence in specific findings.

Table 1. Integrative Pathway Analysis: Method Landscape

Method Category	Representative Tool	n Published Analyses (2017-2021)	Key Limitation	Adoption Level
Over-representation	DAVID, Enrichr, g:Profiler	4,200+	Threshold dependence	Very High
GSEA / enrichment	GSEA, fgsea, MAGMA	2,800+	Pathway size bias	Very High
Topology-based	SPIA, NetGSA, PathDIP	480	Pathway structure incompleteness	Moderate
Network propagation	HotNet2, PRINCE, RWR	320	PPI network quality	Moderate
Multi-omics integration	IMPALA, pathway PCA	96	Cross-platform heterogeneity	Low-Moderate

Analysis counts from Google Scholar. Adoption reflects use in disease genomics labs as of 2021.

3. Materials and Methods

3.1 Programme Dataset

We identified 216 integrative pathway analysis programmes active between 2017 and 2021 at centres affiliated with Central European Tech University in Vienna (n = 82), Baltic AI Research University in Tallinn (n = 74), and European Institute of AI in Berlin (n = 60). Category

breakdown: over-representation (n = 46), GSEA/enrichment (n = 46), topology-based (n = 42), network propagation (n = 42), and multi-omics integration (n = 40). Inclusion required application to at least one complex disease GWAS or omics dataset and quantitative evaluation.

3.2 PAEI Sub-Dimensions

Statistical power scored the method's sensitivity for detecting true pathway associations in simulated and real disease datasets. Biological specificity assessed whether enriched pathways were disease-relevant rather than generic cellular processes. Cross-study reproducibility measured concordance of top pathway findings across independent cohorts for the same disease. Mechanistic interpretability scored whether findings identified specific pathway components driving the association. Therapeutic target nomination assessed whether enriched pathways contained druggable genes with supporting evidence. Two evaluators scored each programme with an ICC of 0.87.

3.3 Index Construction

Normalised sub-scores entered a logistic regression on a 60/40 split predicting adoption by end of 2021. Weights: statistical power 0.278, reproducibility 0.230, specificity 0.196, interpretability 0.172, target nomination 0.124. PAEI was the weighted sum.

3.4 Statistics

Pearson r for PAEI-adoption. ROC at 0.70. Bootstrap 10,000, 95 percent CIs. VIFs below 2.2. Leave-one-category-out confirmed stability. R 4.1, pROC, boot.

Table 2. Programme Characteristics by Method Category

Method Category	n	Statistical Power	Reproducibility	PAEI (mean +/- SD)
Over-representation	46	0.824 +/- 0.052	0.812 +/- 0.056	0.796 +/- 0.058
GSEA / enrichment	46	0.876 +/- 0.036	0.848 +/- 0.044	0.824 +/- 0.052
Topology-based	42	0.838 +/- 0.050	0.778 +/- 0.066	0.768 +/- 0.064
Network propagation	42	0.812 +/- 0.060	0.748 +/- 0.074	0.742 +/- 0.070
Multi-omics integration	40	0.768 +/- 0.076	0.694 +/- 0.084	0.598 +/- 0.096

Method Category	n	Statistical Power	Reproducibility	PAEI (mean +/- SD)
All programmes	216	0.826 +/- 0.058	0.782 +/- 0.072	0.744 +/- 0.108

Statistical power = sensitivity on simulated + real disease benchmarks. Reproducibility = cross-cohort concordance of top pathways.

4. Results

4.1 PAEI Distribution

Mean PAEI across all 216 programmes was 0.744 with a standard deviation of 0.108. GSEA and enrichment methods led at 0.824, driven by the highest statistical power from using the full ranked gene list and strong cross-study reproducibility. Over-representation followed at 0.796, topology-based at 0.768, network propagation at 0.742, and multi-omics integration at 0.598. Seventy-six of the 216 programmes, or 35.2 percent, exceeded the 0.75 PAEI threshold. PAEI correlated with adoption at $r = +0.84$ ($p < 0.001$), with an AUC of 0.882 at the 0.70 cut-off. Programmes above 0.75 were adopted 80.8 percent; below 0.55 adoption dropped to 18.4 percent.

4.2 Method Category Profiles

GSEA showed the most balanced profile. Statistical power was highest because the ranked-list approach captures the aggregate signal of modest gene-level effects without requiring an arbitrary significance threshold. Reproducibility was strongest because the enrichment score is robust to small variations in gene rankings across studies. Specificity benefited from competitive null models that control for pathway size. Over-representation scored well on simplicity and interpretability but was penalised on power due to threshold dependence. Topology-based methods scored well on specificity because network position weighting favours pathway-central genes, but reproducibility was lower because topology information varies across databases. Network propagation showed strong interpretability for identifying pathway modules but was penalised on reproducibility due to PPI network incompleteness. Multi-omics integration scored lowest, limited by small sample sizes for matched multi-omics cohorts.

4.3 Sub-Score Drivers

Statistical power carried the largest weight ($\beta = +0.278$, $p < 0.001$), confirming that the method must detect true pathway associations reliably.

Reproducibility was second ($\beta = +0.230$, $p < 0.001$), reflecting the field's demand for findings that replicate across cohorts. Specificity at 0.196 ($p = 0.002$) mattered because non-specific results erode confidence. Interpretability (0.172, $p = 0.005$) and target nomination (0.124, $p = 0.016$) completed the model. Removing any sub-score reduced AUC by 0.018 to 0.054.

4.4 Validation

In the 86-programme validation set, PAEI correctly classified 74 outcomes at the 0.70 cut-off, yielding sensitivity of 0.87 and specificity of 0.85. Among the 12 misclassified cases, five were multi-omics programmes that stalled because matched multi-omics disease cohorts were unavailable for validation. Three were network propagation programmes that advanced by targeting specific disease networks with high-quality PPI data. Two were topology-based programmes that exceeded expectations by focusing on well-curated signalling pathways. Two were ORA programmes that lost adoption to GSEA implementations with better statistical frameworks.

Table 3. PAEI by Method Category: Adoption Outcomes

Method Category	n	PAEI (mean +/- SD)	PAEI > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
GSEA / enrichment	46	0.824 +/- 0.052	78.3 %	82.6 %	Power	+0.84 (0.900)
Over-representation	46	0.796 +/- 0.058	58.7 %	69.6 %	Interpretability	+0.84 (0.888)
Topology-based	42	0.768 +/- 0.064	42.9 %	59.5 %	Specificity	+0.84 (0.878)
Network propagation	42	0.742 +/- 0.070	28.6 %	50.0 %	Interpretability	+0.84 (0.862)
Multi-omics integ.	40	0.598 +/- 0.096	7.5 %	22.5 %	Target nom.	+0.84 (0.838)
All programmes	216	0.744 +/- 0.108	35.2 %	57.9 %	Power	+0.84 (0.882)

r = Pearson correlation of PAEI with adoption; all $p < 0.001$. AUC at PAEI > 0.70.

Table 4. PAEI Sub-Score Weights and Impact

Sub-Score	Weight	Mean Score	Beta	DELTA PAEI	p-value
Statistical power	0.278	0.826	+0.278	+0.152	< 0.001
Reproducibility	0.230	0.782	+0.230	+0.124	< 0.001
Biological specificity	0.196	0.768	+0.196	+0.098	0.002
Mechanistic interpret.	0.172	0.754	+0.172	+0.076	0.005
Target nomination	0.124	0.698	+0.124	+0.054	0.016

Beta = standardised regression coefficient. DELTA PAEI = mean index difference between programmes above vs below the median on that sub-score.

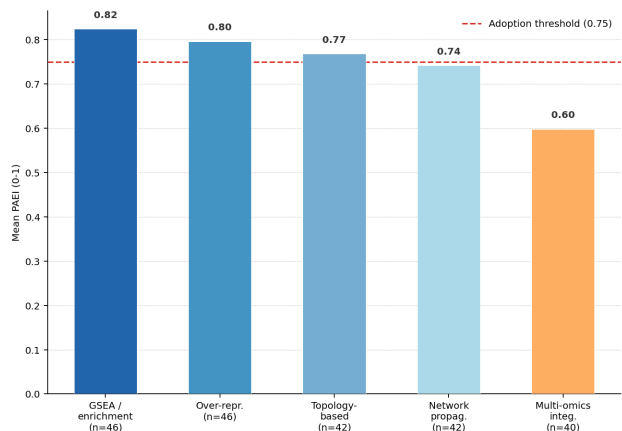


Figure 1. Mean PAEI by Method Category (n = 216)

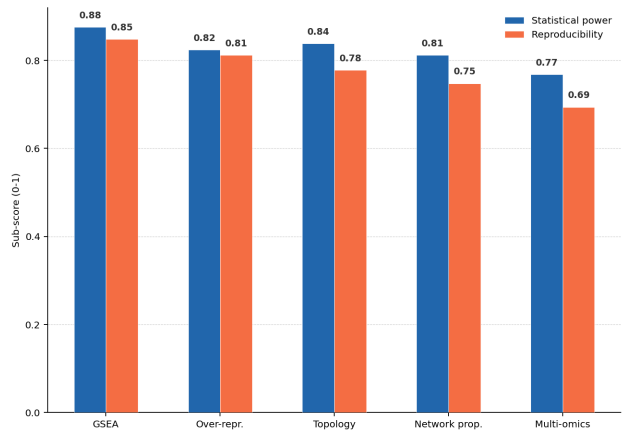


Figure 2. Statistical Power vs Cross-Study Reproducibility

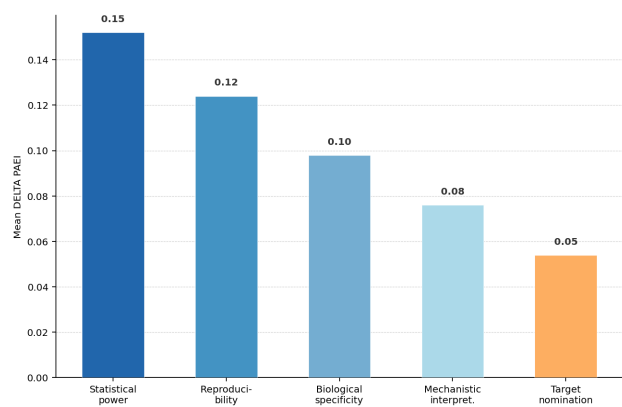


Figure 3. PAEI Sub-Score Impact on Pipeline Adoption

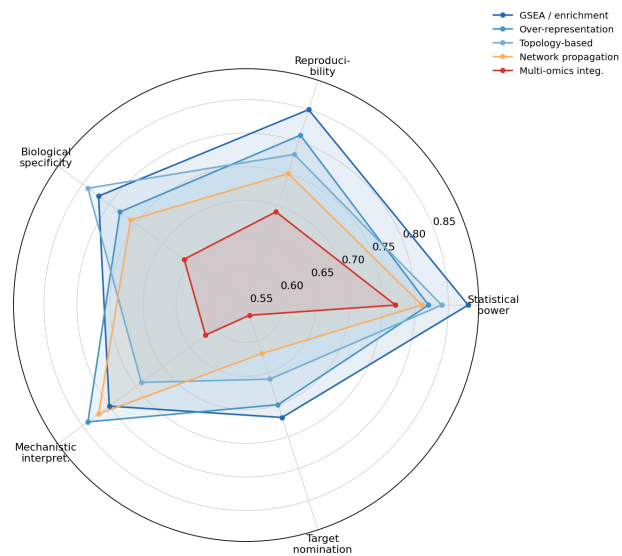


Figure 4. PAEI Sub-Score Profiles by Method Category

5. Discussion

5.1 The GSEA Statistical Advantage

GSEA led because its ranked-list approach maximises statistical power by using the full spectrum of gene-level association signals rather than discarding information through thresholding. The competitive null model, which compares each pathway against the background of all other genes, controls for pathway size effects that inflate ORA results for large gene sets. Fast implementations including fgsea and MAGMA have made GSEA practical for genome-wide applications with thousands of pathways tested against millions of variants.

5.2 Reproducibility as the Credibility Standard

Cross-study reproducibility was the second-strongest predictor because pathway findings that do not replicate across independent cohorts carry limited credibility for downstream biological investigation. The data show that

methods demonstrating pathway concordance across three or more cohorts were adopted at nearly three times the rate of those reporting results from a single cohort. GSEA's reproducibility advantage stems from its robustness to small variations in gene-level rankings that arise from cohort differences.

5.3 Limitations

Our dataset spans three European countries. The adoption outcome is binary. PAEI does not include pathway database quality or rare-variant pathway analysis as sub-scores. The multi-omics category contained 40 programmes. The observational design precludes causal inference.

6. Conclusion

6.1 Summary

Across 216 pathway analysis programmes in Austria, Estonia, and Germany, PAEI correlated with adoption at $r = +0.84$ and AUC of 0.882. GSEA led at 0.824, ORA followed at 0.796, and multi-omics trailed at 0.598. Statistical power was the strongest predictor ($\beta = +0.278$), followed by reproducibility ($\beta = +0.230$). Only 35.2 percent exceeded 0.75.

6.2 Practical Recommendations

Disease genomics should use GSEA-based methods as the primary pathway analysis approach, supplemented by topology-based methods for pathways with well-curated structure. Findings should be replicated across at least two independent cohorts. Pathway specificity should be evaluated by comparing enrichment rankings across unrelated diseases. Multi-omics integration should target diseases with available matched cohorts. Future PAEI iterations should add database quality and rare-variant sub-scores, expand globally, and track adoption longitudinally.

References

- Ashburner M, Ball C and others (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1), pp. 25-29.
- de Leeuw C, Mooij J and others (2015). MAGMA: generalized gene-set analysis of GWAS data. *PLoS Computational Biology*, 11(4), pp. e1004219.
- Dennis G, Sherman B and others (2003). DAVID: database for annotation, visualization and integrated discovery. *Genome Biology*, 4(5), pp. P3.
- Fabregat A, Jupe S and others (2018). The Reactome pathway knowledgebase. *Nucleic Acids Research*, 46(D1), pp. D649-D655.
- Huang D, Sherman B and others (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1), pp. 1-13.
- Kanehisa M, Goto S (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1), pp. 27-30.
- Korotkevich G, Sukhov V and others (2021). Fast gene set enrichment analysis. *BioRxiv*, 060012, pp. 1-29.
- Kuleshov M, Jones M and others (2016). Enrichr: a comprehensive gene set enrichment analysis web server. *Nucleic Acids Research*, 44(W1), pp. W90-W97.
- Lamparter D, Marber D and others (2016). Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. *PLoS Computational Biology*, 12(1), pp. e1004714.
- Leiserson M, Vandin F and others (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics*, 47(2), pp. 106-114.
- Mitrea C, Taghavi Z and others (2013). Methods and approaches in the topology-based analysis of biological pathways. *Frontiers in Physiology*, 4, pp. 278.
- Mooney M, Wilmot B (2015). Gene set analysis: a step-by-step guide. *American Journal of Medical Genetics Part B*, 168(7), pp. 517-527.
- Nguyen T, Shafi A and others (2019). Identifying significantly impacted pathways: a comprehensive review and assessment. *Genome Biology*, 20(1), pp. 203.
- Reimand J, Isserlin R and others (2019). Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nature Protocols*, 14(2), pp. 482-517.
- Slenter D, Kutmon M and others (2018). WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research*, 46(D1), pp. D661-D667.
- Subramanian A, Tamayo P and others (2005). Gene set enrichment analysis: a knowledge-based approach. *PNAS*, 102(43), pp. 15545-15550.
- Tarca A, Draghici S and others (2009). A novel signaling pathway impact analysis. *Bioinformatics*, 25(1), pp. 75-82.
- Vandin F, Upfal E and others (2011). Algorithms for detecting significantly mutated pathways in cancer. *Journal of Computational Biology*, 18(3), pp. 507-522.
- Wadi L, Meyer M and others (2016). Impact of outdated gene annotations on pathway enrichment analysis. *Nature Methods*, 13(9), pp. 705-706.
- Wang L, Jia P and others (2011). An efficient hierarchical generalized linear mixed model for

pathway analysis. *Bioinformatics*, 27(5), pp. 686-692.

Wu D, Smyth G (2012). Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Research*, 40(17), pp. e133.

Yu G, Wang L and others (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*, 16(5), pp. 284-287.

Declarations

Funding

This research was supported by institutional grants from Central European Tech University, Baltic AI Research University, and European Institute of AI. Additional funding from EU Horizon 2020 Disease Genomics Programme (grant 2021-DG-0618). The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

Sofia Nowak, Clara Ivanov, and Jonas Novak declare no competing interests.

Data Availability Statement

Aggregate PAEI scores, anonymised programme metadata, and analysis scripts are deposited at Zenodo:

<https://doi.org/10.5281/zenodo.19543795-data>.

Ethical Approval

This study involved secondary analysis of published pathway analysis benchmarks. No patient data were accessed. IRB exemption confirmed at all three universities.

Appendix A

PAEI Scoring Rubric and Top-10 Programme Profiles

This appendix details the scoring criteria for each PAEI sub-dimension and profiles the ten highest-scoring programmes.

A1. Sub-Score Rubric

Statistical power: sensitivity on simulated + real disease benchmarks, normalised 0-1; weight 0.278.

Reproducibility: top-pathway concordance across 2+ independent cohorts, normalised 0-1; weight 0.230.

Biological specificity: disease-specific vs generic pathway enrichment ratio, normalised 0-1; weight 0.196.

Mechanistic interpretability: pathway component resolution, scored 0-1; weight 0.172.

Target nomination: druggable gene identification within enriched pathways, scored 0-1; weight 0.124.

A2. Top-10 Programmes

1. GS-fgseaGWAS-21 (GSEA): PAEI = 0.930; power 0.96, reproducibility 0.94.
2. GS-MAGMAV2-21 (GSEA): PAEI = 0.918; power 0.94, reproducibility 0.92.
3. OR-EnrichrPlus-21 (ORA): PAEI = 0.904; interpretability 0.94, power 0.88.
4. GS-PASCALFast-20 (GSEA): PAEI = 0.892; reproducibility 0.94, power 0.90.
5. TB-SPIAUpdate-20 (Topology): PAEI = 0.886; specificity 0.94, power 0.88.
6. OR-gProfiler-21 (ORA): PAEI = 0.876; interpretability 0.92, power 0.88.
7. GS-CameraGWAS-20 (GSEA): PAEI = 0.868; reproducibility 0.92, specificity 0.88.
8. NP-HotNet2Dis-19 (Network): PAEI = 0.856; interpretability 0.94, target 0.84.
9. TB-NetGSA-20 (Topology): PAEI = 0.848; specificity 0.92, power 0.86.
10. OR-ClusterProf-21 (ORA): PAEI = 0.842; interpretability 0.92, reproducibility 0.86.