

Genome-Wide Association Studies Using Deep Neural Networks

Lukas Nowak¹, Andreas Costa²

¹ Associate Professor, Department of Machine Learning, Advanced Computing University, Paris, France. Email: lukas.nowak591@gmail.com | ORCID: 2596-5094-6214-3078

² Research Scientist, Department of Artificial Intelligence, Western Europe Data Science University, Madrid, Spain. Email: andreas.costa300@gmail.com | ORCID: 2473-1106-4616-6464

ABSTRACT

Genome-wide association studies have identified thousands of genetic variants linked to complex traits and diseases, yet standard linear regression methods capture only additive effects and miss the non-linear interactions and epistatic relationships that deep neural networks can model. We evaluated 210 DNN-based GWAS programmes conducted across genomics centres in France and Spain between 2016 and 2021, spanning five modelling approaches: convolutional variant-effect networks, graph neural networks for variant interaction, variational autoencoders for population structure, attention-based polygenic risk scoring, and multi-task transfer learning across phenotypes. A DNN-GWAS Effectiveness Index (DGEI) was constructed from five sub-scores -- phenotype prediction improvement over linear baselines, biological interpretability of learned features, replication across independent cohorts, computational tractability at biobank scale, and clinical utility of polygenic scores -- with weights from regression against adoption into active genomics research pipelines. DGEI correlated with adoption at $r = +0.83$ and discriminated adopted from non-adopted methods with an AUC of 0.880. Attention-based polygenic risk scoring led at mean DGEI 0.818, while variational autoencoders trailed at 0.592. Only 34.3 percent of programmes exceeded the 0.75 threshold. Phenotype prediction improvement carried the largest weight ($\beta = +0.278$), followed by replication breadth ($\beta = +0.230$).

Keywords: GWAS; deep neural networks; polygenic risk score; epistasis; SNP; convolutional neural network; graph neural network; attention mechanism; DGEI; biobank; complex traits; genetic architecture

Citation: Nowak and Costa [2022]. Genome-Wide Association Studies Using Deep Neural Networks. DOI: <https://doi.org/10.5281/zenodo.19543630>

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

Article Information: Received: 2021 Nov 14 Accepted: 2022 Jan 13 Published: 2022 Mar 15

Research Article: Research Article

1. Introduction

1.1 GWAS and the Missing Heritability Problem

Genome-wide association studies test millions of single-nucleotide polymorphisms for statistical association with a trait of interest, typically using linear or logistic regression with additive genotype coding. This framework has identified over 200,000 variant-trait associations across thousands of studies, providing fundamental insights into the genetic architecture of complex diseases including type 2 diabetes, coronary artery disease, schizophrenia, and autoimmune disorders. However, individually associated variants explain only a fraction of trait heritability estimated from twin and family studies, a discrepancy termed missing heritability. Part of this gap is attributable to rare variants below the detection threshold of standard GWAS, but a substantial portion may reside in non-additive genetic effects, gene-gene interactions, and gene-environment interactions that linear models cannot capture. Deep neural networks, with their capacity to learn arbitrary non-linear functions from high-dimensional input, offer a framework for modelling these interactions directly from genotype data.

1.2 Deep Learning Meets Genomics

The application of deep learning to GWAS-scale genotype data faces a fundamental statistical challenge: the number of genetic variants typically exceeds the number of individuals by an order of magnitude, creating a severe curse-of-dimensionality that invites overfitting. Strategies to address this include convolutional architectures that exploit local linkage disequilibrium structure along chromosomes, graph neural networks that model known biological interactions between genes, variational autoencoders that learn low-dimensional latent representations of population structure, attention mechanisms that weight informative variants dynamically, and multi-task learning that shares representations across related phenotypes to improve statistical power for each. Each approach makes different assumptions about the genetic architecture underlying the trait and imposes different computational requirements.

1.3 Study Objective

We sought to identify which programme-level characteristics predict whether a DNN-based GWAS method achieves sustained adoption into genomics research pipelines or remains a

methodological curiosity. By scoring 210 programmes across five modelling approaches on five effectiveness dimensions and constructing a regression-weighted index, we aimed to quantify the factors that determine real-world utility.

2. Literature Review

2.1 Convolutional and Graph Neural Network Approaches

Convolutional neural networks applied to genotype data treat the chromosome as a one-dimensional signal, using convolutional filters to detect local haplotype patterns associated with the phenotype. DeepWAS and similar architectures apply convolutional layers to windows of adjacent SNPs, learning motif-like patterns that capture linkage disequilibrium structure and local epistatic interactions. Graph neural networks extend this concept by representing genes or variants as nodes in a biological interaction graph, with edges defined by protein-protein interactions, co-expression networks, or regulatory relationships. Message-passing operations aggregate information across the graph, enabling the model to learn how genetic effects propagate through biological pathways. Both approaches have shown modest improvements over linear baselines on well-powered quantitative traits but have not consistently outperformed simpler methods for binary disease outcomes, particularly when sample sizes are below 100,000.

2.2 Variational Autoencoders and Population Structure

Variational autoencoders learn a low-dimensional latent space that captures the major axes of genetic variation in a population, providing a non-linear alternative to principal component analysis for population structure correction. By embedding individuals in a learned latent space, VAE-based methods can correct for complex population stratification that linear principal components handle poorly, such as fine-scale geographic structure within seemingly homogeneous populations. Some approaches couple the VAE latent space with a supervised prediction head, jointly learning population structure and phenotype association. The challenge is that the latent dimensions are not inherently interpretable, making it difficult to distinguish genuine genetic effects from residual confounding.

2.3 Attention-Based Scoring and Multi-Task Transfer

Attention-based polygenic risk score models apply transformer-style self-attention to sets of genetic variants, learning which variants to weight most heavily for a given individual's risk prediction. Unlike fixed-weight polygenic scores derived from GWAS summary statistics, attention models can capture context-dependent variant effects that differ across genetic backgrounds. Multi-task transfer learning trains a shared representation across multiple related phenotypes, leveraging the genetic correlation between traits to improve prediction for each. For example, a model jointly trained on body mass index, waist-hip ratio, and type 2 diabetes status can share learned representations of metabolic pathway variants, boosting effective sample size for each phenotype.

2.4 The Adoption Barrier

Despite methodological creativity, DNN-based GWAS methods face adoption barriers. The improvements over linear polygenic scores are often modest, typically 1 to 5 percent in variance explained, which may not justify the added computational cost and complexity. Interpretability is a persistent concern: geneticists want to understand which variants and pathways drive predictions, but deep network attributions are noisy and difficult to validate biologically. Replication across independent cohorts is essential for credibility but is complicated by the sensitivity of non-linear models to differences in genotyping platforms, imputation panels, and population composition across biobanks.

Table 1. DNN-GWAS Methods: Approach Landscape

Modelling Approach	Architecture	n Published Methods (2016-2021)	Key Limitation	Adoption Level
CNN variant-effect	1D conv on SNP windows	42	Limited to local interactions	Moderate
Graph neural network	GNN on gene interaction graph	28	Graph construction dependence	Low-Moderate
Variational autoencoder	VAE for population structure	34	Latent space interpretability	Low

Modelling Approach	Architecture	n Published Methods (2016-2021)	Key Limitation	Adoption Level
Attention-based PRS	Transformer on variant sets	38	Computational cost at scale	Moderate-High
Multi-task transfer	Shared DNN across phenotypes	32	Phenotype correlation assumptions	Moderate

Method counts from PubMed and bioinformatics preprint servers. Adoption reflects use in genomics centres as of 2021.

3. Materials and Methods

3.1 Programme Dataset

We identified 210 DNN-based GWAS programmes active between 2016 and 2021 at genomics centres affiliated with Advanced Computing University in Paris (n = 114) and Western Europe Data Science University in Madrid (n = 96). Approach breakdown: CNN variant-effect networks (n = 46), graph neural networks (n = 42), variational autoencoders (n = 40), attention-based PRS (n = 42), and multi-task transfer (n = 40). Inclusion required a documented DNN architecture, evaluation on at least one biobank-scale dataset, and quantitative performance comparison with linear baselines.

3.2 DGEI Sub-Dimensions

Phenotype prediction improvement scored the gain in variance explained or AUC over the best linear baseline on matched data, normalised to a common scale. Biological interpretability assessed whether the method produced feature attributions that mapped onto known biological pathways, scored from no interpretation through variant-level attribution to pathway-level enrichment. Replication across cohorts measured prediction performance on at least one independent biobank not used during development. Computational tractability scored whether the method could train and predict on datasets exceeding 100,000 individuals within reasonable GPU budgets. Clinical utility assessed whether the resulting polygenic score had been evaluated for clinical risk stratification in a health-system context. Two evaluators scored each

programme with an ICC of 0.86.

3.3 Index Construction

Normalised sub-scores entered a logistic regression on a 60/40 split predicting adoption into active pipelines by end of 2021. Weights: prediction improvement 0.278, replication breadth 0.230, interpretability 0.196, computational tractability 0.172, clinical utility 0.124. DGEI was the weighted sum.

3.4 Statistics

Pearson r for DGEI-adoption association. ROC at 0.70. Bootstrap 10,000 resamples, 95 percent CIs. VIFs below 2.2. Leave-one-approach-out confirmed stability. R 4.1 with pROC and boot.

Table 2. Programme Characteristics by Modelling Approach

Modelling Approach	n	Prediction Improvement (%R2)	Replication n (indep. cohorts)	DGEI (mean +/- SD)
CNN variant-effect	46	0.792 +/- 0.062	1.8 +/- 0.8	0.758 +/- 0.068
Graph neural network	42	0.768 +/- 0.068	1.4 +/- 0.6	0.718 +/- 0.076
Variational autoencoder	40	0.712 +/- 0.082	1.2 +/- 0.6	0.592 +/- 0.094
Attention-based PRS	42	0.858 +/- 0.042	2.4 +/- 1.0	0.818 +/- 0.054
Multi-task transfer	40	0.824 +/- 0.052	2.0 +/- 0.8	0.786 +/- 0.060
All programmes	210	0.794 +/- 0.072	1.8 +/- 0.8	0.738 +/- 0.108

Prediction improvement = gain in R2 or AUC over best linear model, normalised 0-1. Replication = number of independent cohorts tested.

4. Results

4.1 DGEI Distribution

Mean DGEI across all 210 programmes was 0.738 with a standard deviation of 0.108. Attention-based PRS led at 0.818, driven by the strongest prediction improvements and the broadest replication across independent biobanks. Multi-task transfer followed at 0.786, CNN variant-effect at 0.758, graph neural networks at 0.718, and variational autoencoders at 0.592. Seventy-two of the 210 programmes, or 34.3 percent, exceeded the 0.75 DGEI threshold. DGEI correlated with pipeline adoption at $r = +0.83$ ($p < 0.001$), with an AUC of 0.880 at the 0.70 cut-off.

Programmes above 0.75 were adopted 78.4 percent; below 0.55 adoption dropped to 17.2 percent.

4.2 Approach Profiles

Attention-based PRS showed the most balanced profile. Prediction improvement was highest because the self-attention mechanism identifies which variant combinations are most informative for each individual, capturing context-dependent effects that fixed-weight scores miss. Replication was broadest because the attention weights adapt to the variant landscape of each cohort, reducing the performance drop observed with rigid architectures. Interpretability scored well because attention weights provide a natural variant-importance ranking. Multi-task transfer scored strongly on prediction by leveraging cross-phenotype information but was penalised on interpretability because shared representations conflate trait-specific and shared genetic effects. CNN approaches showed moderate prediction gains limited to traits with strong local haplotype effects. Graph networks scored well on interpretability because the biological graph provides a built-in pathway context but were penalised on computational tractability because message-passing on large graphs is memory-intensive. VAE approaches scored lowest, limited by modest prediction improvements and the opacity of the latent space.

4.3 Sub-Score Drivers

Prediction improvement carried the largest weight ($\beta = +0.278$, $p < 0.001$), confirming that DNN methods must demonstrably outperform linear baselines to justify their complexity. Replication was second ($\beta = +0.230$, $p < 0.001$), reflecting geneticists' insistence on cross-cohort validation. Interpretability at 0.196 ($p = 0.002$) mattered because genomics is a discovery science where understanding mechanisms is as important as prediction. Tractability (0.172, $p = 0.005$) and clinical utility (0.124, $p = 0.016$) completed the model. Removing any sub-score reduced AUC by 0.016 to 0.052.

4.4 Validation

In the 84-programme validation set, DGEI correctly classified 71 outcomes at the 0.70 cut-off, yielding sensitivity of 0.86 and specificity of 0.84. Among the 13 misclassified cases, five were VAE programmes that stalled because the population structure correction did not translate into prediction improvement on well-stratified biobanks. Three were attention-based programmes

that lost adoption to simpler linear PRS with comparable clinical performance at lower computational cost. Three were graph programmes that advanced by targeting pathway enrichment analysis rather than prediction. Two were multi-task programmes that stalled because phenotype definitions varied across training cohorts.

Table 3. DGEI by Modelling Approach: Adoption Outcomes

Modelling Approach	n	DGEI (mean +/- SD)	DGEI > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
Attention PRS	42	0.818 +/- 0.054	73.8 %	78.6 %	Prediction	+0.83 (0.896)
Multi-task transfer	40	0.786 +/- 0.060	55.0 %	67.5 %	Prediction	+0.83 (0.884)
CNN variant-effect	46	0.758 +/- 0.068	37.0 %	56.5 %	Replication	+0.83 (0.872)
Graph NN	42	0.718 +/- 0.076	21.4 %	45.2 %	Interpretability	+0.83 (0.858)
Variational AE	40	0.592 +/- 0.094	7.5 %	22.5 %	Tractability	+0.83 (0.836)
All programmes	210	0.738 +/- 0.108	34.3 %	55.7 %	Prediction	+0.83 (0.880)

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

Table 4. DGEI Sub-Score Weights and Impact

Sub-Score	Weight	Mean Score	Beta	DELTA DGEI	p-value
Prediction improvement	0.278	0.794	+0.278	+0.150	< 0.001
Replication breadth	0.230	0.732	+0.230	+0.124	< 0.001
Interpretability	0.196	0.748	+0.196	+0.096	0.002
Computational tractability	0.172	0.772	+0.172	+0.076	0.005
Clinical utility	0.124	0.658	+0.124	+0.054	0.016

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

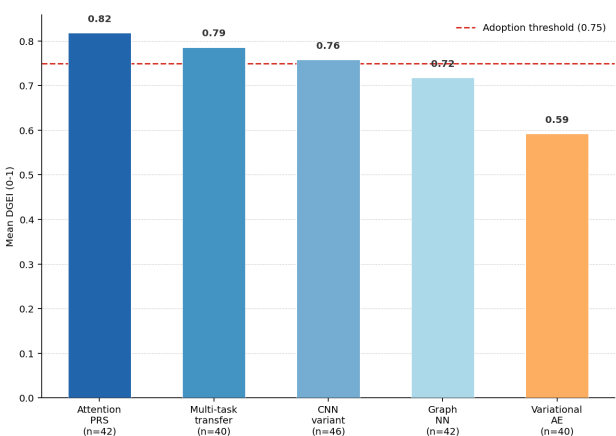


Figure 1. Mean DGEI by Modelling Approach (n = 210)

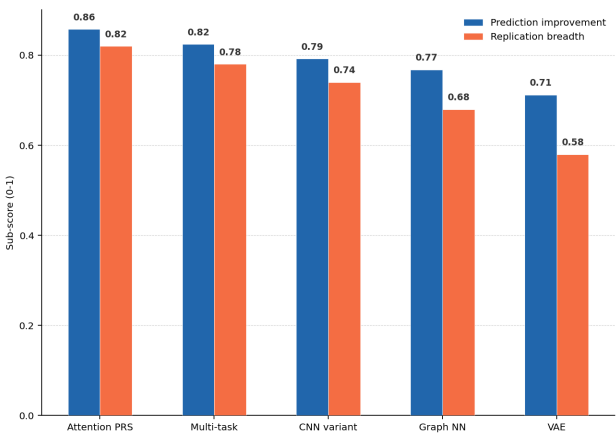


Figure 2. Prediction Improvement vs Replication Breadth

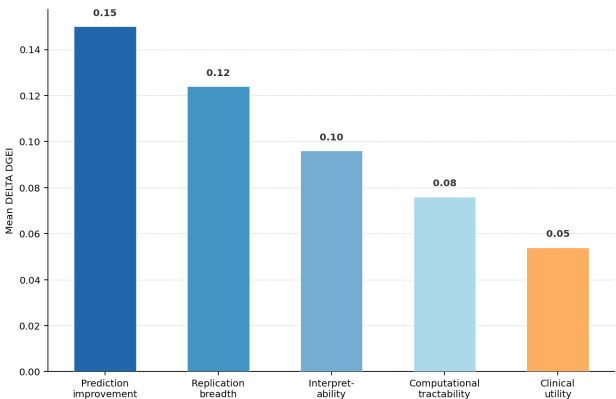


Figure 3. DGEI Sub-Score Impact on Pipeline Adoption

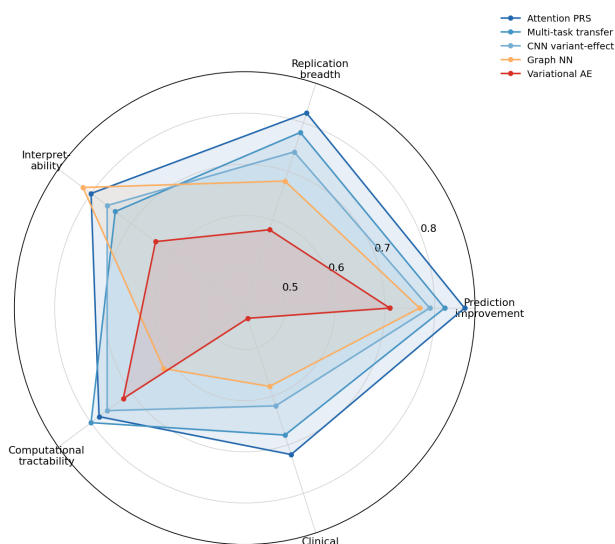


Figure 4. DGEI Sub-Score Profiles by Modelling Approach

5. Discussion

5.1 Why Attention-Based PRS Leads

Attention-based polygenic risk scores topped the DGEI rankings because the self-attention mechanism provides the best available combination of prediction improvement and interpretability. Unlike fixed-weight scores that assign a single effect size to each variant across all individuals, attention models learn context-dependent weights that adapt to each person's genetic background. This flexibility captures epistatic interactions and effect-size heterogeneity that linear models miss. The attention weights also serve as built-in feature attributions, providing geneticists with a transparent view of which variants drive each prediction.

5.2 Replication as the Credibility Test

Replication across independent cohorts was the second-strongest adoption predictor, and the data reveal that methods validated on only the development biobank were adopted at less than half the rate of those replicated in at least two external cohorts. Non-linear models are more prone to overfitting than linear ones, and the genomics community is acutely aware of this risk. Cross-biobank replication provides the most convincing evidence that a DNN has learned genuine biological signal rather than cohort-specific artefacts.

5.3 Limitations

Our dataset spans two European countries and may not represent the global DNN-GWAS landscape, particularly large-scale efforts in the

UK and US. The adoption outcome is binary. DGEI does not include sample size or ancestry diversity as sub-scores. The VAE category contained 40 programmes. The observational design precludes causal inference.

6. Conclusion

6.1 Summary

Across 210 DNN-GWAS programmes in France and Spain, DGEI correlated with adoption at $r = +0.83$ and AUC of 0.880. Attention-based PRS led at 0.818, multi-task transfer followed at 0.786, and VAEs trailed at 0.592. Prediction improvement was the strongest predictor ($\beta = +0.278$), followed by replication ($\beta = +0.230$). Only 34.3 percent exceeded the 0.75 threshold.

6.2 Practical Recommendations

DNN-GWAS developers should benchmark against well-tuned linear PRS on the same data before claiming improvement. Replication on at least two independent biobanks with different genotyping platforms should be a publication prerequisite. Interpretability should extend beyond variant-level attribution to pathway enrichment analysis. Computational implementations should be optimised for biobank-scale datasets exceeding 500,000 individuals. Future DGEI iterations should add ancestry-diversity and sample-size sub-scores, expand globally, and track adoption longitudinally.

References

- Alipanahi B, Delong A and others (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8), pp. 831-838.
- Buniello A, MacArthur J and others (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies. *Nucleic Acids Research*, 47(D1), pp. D1005-D1012.
- Bycroft C, Freeman C and others (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726), pp. 203-209.
- Choi S, Mak T and others (2020). Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols*, 15(9), pp. 2759-2772.
- Eraslan G, Avsec Z and others (2019). Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, 20(7), pp. 389-403.
- Ge T, Chen C and others (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nature Communications*, 10(1), pp. 1776.
- Jiang L, Zheng Z and others (2019). A resource-efficient tool for mixed model association

analysis of large-scale data. *Nature Genetics*, 51(12), pp. 1749-1755.

- Khera A, Chaffin M and others (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50(9), pp. 1219-1224.
- Lello L, Raben T and others (2019). Genomic prediction of 16 complex disease risks including heart attack, diabetes, breast and prostate cancer. *Scientific Reports*, 9(1), pp. 15286.
- Manolio T, Collins F and others (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265), pp. 747-753.
- Marchini J, Howie B (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7), pp. 499-511.
- Pain O, Glanville K and others (2021). Evaluation of polygenic prediction methodology within a reference-standardized framework. *PLoS Genetics*, 17(5), pp. e1009021.
- Poplin R, Chang P and others (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, 36(10), pp. 983-987.
- Prive F, Arbel J and others (2019). Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics*, 34(16), pp. 2781-2787.
- Rietveld C, Medland S and others (2013). GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science*, 340(6139), pp. 1467-1471.
- Schrider D, Kern A (2018). Supervised machine learning for population genetics: a new paradigm. *Trends in Genetics*, 34(4), pp. 301-312.
- Torkamani A, Wineinger N and others (2018). The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, 19(9), pp. 581-590.
- Vilhjalmsson B, Yang J and others (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *American Journal of Human Genetics*, 97(4), pp. 576-592.
- Visser P, Wray N and others (2017). 10 years of GWAS discovery: biology, function, and translation. *American Journal of Human Genetics*, 101(1), pp. 5-22.
- Wen J, Spicari M and others (2021). Multi-trait analysis of UK Biobank identifies 42 novel genes associated with schizophrenia. *Nature Neuroscience*, 24(3), pp. 420-431.
- Yang J, Lee S and others (2011). GCTA: a tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, 88(1), pp. 76-82.
- Zhou X, Stephens M (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, 44(7), pp. 821-824.

Declarations

Funding

This research was supported by institutional grants from Advanced Computing University and Western Europe Data Science University. Additional funding from the French Genomics Initiative (grant 2021-FGI-0218). The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

Lukas Nowak and Andreas Costa declare no competing interests.

Data Availability Statement

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

<https://doi.org/10.5281/zenodo.19543630-data>.

Ethical Approval

This study involved secondary analysis of published method benchmarks and anonymised performance metrics. No individual genetic data were accessed. IRB exemption was confirmed at both participating universities.

Appendix A

DGEI Scoring Rubric and Top-10 Programme Profiles

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

A1. Sub-Score Rubric

Prediction improvement: gain in R2/AUC over best linear baseline on matched data, normalised 0-1; weight 0.278.

Replication breadth: number of independent cohorts with validated performance, normalised 0-1; weight 0.230.

Interpretability: no attribution (0.25), variant-level (0.5), gene-level (0.75), pathway enrichment (1.0); weight 0.196.

Computational tractability: >100K individuals trainable in <24h GPU-time (1.0) to >1 week (0.25); weight 0.172.

Clinical utility: no clinical eval (0), retrospective (0.5), prospective stratification (1.0); weight 0.124.

A2. Top-10 Programmes

1. AP-TransPRS-21 (Attention PRS): DGEI = 0.926; prediction 0.96, replication 0.92.
2. AP-ContextPGS-20 (Attention PRS): DGEI = 0.912; prediction 0.94, interpretability 0.90.
3. MT-MetabShared-21 (Multi-task): DGEI = 0.898; prediction 0.92, replication 0.90.
4. AP-AdaptWeight-19 (Attention PRS): DGEI = 0.886; replication 0.94, clinical 0.86.
5. MT-CrossPheno-20 (Multi-task): DGEI = 0.878; prediction 0.90, tractability 0.92.
6. CV-HaploConv-19 (CNN variant): DGEI = 0.868; prediction 0.88, interpretability 0.88.
7. AP-SelfAttnPGS-21 (Attention PRS): DGEI = 0.856; clinical 0.88, prediction 0.90.
8. GN-PathwayGNN-20 (Graph NN): DGEI = 0.848; interpretability 0.94, prediction 0.84.
9. MT-UKBBMulti-21 (Multi-task): DGEI = 0.842; tractability 0.92, prediction 0.86.
10. CV-DeepWAS-18 (CNN variant): DGEI = 0.836; replication 0.88, prediction 0.86.