METHODOLOGY | Article in Press

G2PT: a genotype–phenotype transformer to assess and explain polygenic risk

Ingoo Lee[1] · Zachary Wallace[1,2] · Yuqi Wang[1] · Sungjoon Park[1] · Hojung Nam[3,4,5] · Amit R. Majithia[1] · Trey Ideker[1,3,6]

Received: 20 January 2026 / Accepted: 21 October 2026
© The Author(s) 2026

Abstract

Genome-wide association studies have linked millions of genetic variants to biomedical phenotypes, but their utility has been limited by lack of mechanistic understanding and widespread epistatic interactions. Here, we introduce the Genotype-to-Phenotype Transformer (G2PT), a framework for modeling hierarchical information flow among variants, genes, multigenic systems, and phenotypes. We train G2PT to model the genetics of metabolic traits including insulin resistance, LDL and type − 2 diabetes. G2PT predicts these traits with accuracy comparable to best-in-class methods, demonstrates comparable transferability across cohort, and captures epistatic effects. This work positions hierarchical graph transformers as a next-generation approach to polygenic risk.

Background

Common diseases such as type 2 diabetes (T2D) [1], cardiovascular disease [2], and fatty liver [3] are highly polygenic and physiologically heterogeneous, involving complex networks of interactions within and among multigenic molecular systems [4]. In these complex traits, individual examination of single nucleotide polymorphisms (SNPs) and other genetic variants has had limited utility for risk prediction and stratification in most individuals. Rather, progress has been made by systematically scanning for associated SNPs through genome-wide association studies (GWAS) [5], and then combining these many SNP-phenotype associations using methodologies collectively termed Polygenic Risk Scores (PRS) [6–8]. PRS methods have been applied to predict risk for a wide range of common multigenic diseases [1, 9–11] but have two major limitations. First, they model loci additively and thus miss functional dependencies among variants, including genetic epistasis. Second, they assign a summary risk score for the whole collection of SNPs comprising an individual's genotype, without explaining how that risk is attributed to perturbations to molecular and physiological pathways, or recommending what action the individual should take. Addressing these aspects could markedly improve risk prediction, biological interpretation and ability to guide treatment [12].

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.



https://doi.org/10.1186/s13059-026-04327-0

Published online: 23 November 2026

Recent developments in deep learning [13–15] offer significant opportunities to advance the PRS framework [16–19], as they have the capacity to model both complex epistatic interactions and knowledge of molecular mechanisms. Among these, the Transformer has shown strong potential to address longstanding problems in the biomedical sciences, including prediction of 3D protein structures [20–22], biomedical image analysis [23], inference of gene expression from genome sequence [24–26], and mapping a sequence of SNPs to a predicted phenotype [27, 28]. The Transformer architecture is known for its central use of an "attention" mechanism [29], an operation that dynamically computes the importance of each input element relative to others, enabling the model to focus on the most relevant features [29–31]. While this mechanism can aid in interpretation, the vast number of input features typical of biomedical data can greatly complicate interpretability and transparency. To address this challenge, incorporating biological knowledge into Transformer attention – such as that provided by gene function databases like the Gene Ontology [32] – has the potential to improve model explainability, transparency, and fairness, all of which are crucial for clinical applications [33–38].

Here we describe the Genotype-to-Phenotype Transformer (G2PT), a graph transformer architecture for general genotype-to-phenotype translation and interpretation (Fig. 1a). The G2PT model analyzes the complex set of genetic variants in a genotype by directing attention to their effect on embedded representations of genes and a hierarchy of multigenic systems. As proof-of-concept we use G2PT to study metabolic traits including insulin resistance (as measured by the triglyceride to high-density lipoprotein cholesterol ratio: TG/HDL), serum cholesterol (low-density lipoprotein cholesterol: LDL) and T2D [39, 40], the most common metabolic disease. Hundreds of loci have been mapped for each of these (TG/HDL [41–43], LDL [44], T2D [45, 46]) but the genetic circuits that regulate their status and molecular physiology are still poorly understood. In what follows, application of G2PT yields a predictive genomic model that explains metabolic traits via a constellation of genetic factors, biological systems, and nonlinear epistatic interactions.

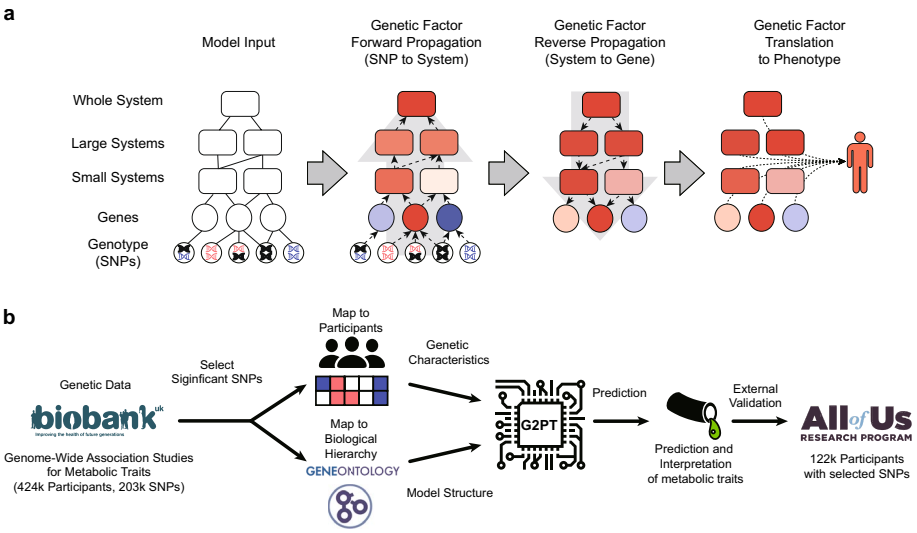


Fig. 1 G2PT workflow. **a** Inputs of the Transformer model include genotypic data (SNPs, bottom layer), a mapping of SNPs to genes (second layer), and a mapping of genes into a hierarchy of multi-genic molecular systems (top layers). The presence of SNP minor alleles modifies the embedding states of downstream genes and multigenic systems (forward propagation). Conversely, state changes in systems influence the states of sub-systems and genes they contain (reverse propagation). Finally, all gene and system states are integrated to predict phenotype. **b** Proof of concept via prediction of metabolic traits such as triglyceride/HDL cholesterol ratio (TG/HDL), low-density lipoprotein cholesterol (LDL), and type – 2 diabetes (T2D). Genotypic data and corresponding metabolic traits are extracted for participants from the UK Biobank. SNPs are selected based on their independent association with metabolic traits and mapped to the closest genes, which in turn map to multigenic systems defined by the Gene Ontology. This information is used by G2PT to predict metabolic phenotypes

Results

G2PT model overview

The G2PT framework models the states of biological entities, including variants, genes, multigenic systems, and phenotypes, as coordinates within a machine learning embedding. An embedding is a simplified low-dimensional representation of a high-dimensional dataset, optimized so that similar entities are assigned similar embedding coordinates [47]. Positions in the embedding (i.e. the states of each entity) are governed by a hierarchical graph transformer, a deep neural network that models flow of information across a hierarchy of connected entities. Such information flow includes the effects of variants on the states of genes (SNP-gene mapping), the effects of altered genes on multigenic systems and their supersystems (gene-system and system-system mapping), and the reciprocal influences of systems on the states of their component systems and genes (Fig. 1a, Methods). Based on the collection of variants comprising an individual's genotype, the model uses a multi-head attention mechanism to propagate these effects to select biological entities in the hierarchy, resulting in updates to their embedding coordinates. Finally, the entire collection of embedding states for genes and systems is used to predict phenotype.

Using G2PT to model metabolic traits and disease

As proof-of-concept, we used G2PT to study human metabolism, focusing on insulin resistance (TG/HDL), LDL cholesterol, and T2D (Fig. 1b). Serum TG, HDL, LDL, and T2D case:control status were obtained from 423,888 participants profiled in the UK Biobank with accompanying genetic data (covering 203,126 SNPs; Methods, Additional file 1: Fig. S1) [41, 48]. SNPs were mapped to genes using any of three lines of evidence: expression quantitative trait loci (eQTL) [49], curated mappings from the cS2G mapping tool [50], or the closest gene in genomic coordinates (hg37 reference genome, Methods). Genes were mapped to a hierarchy of multigenic systems condensed from Biological Process terms recorded in the Gene Ontology (Fig. 1b) [32].

Using this information, G2PT models were trained to translate an individual's complement of SNP alleles into a prediction of each of their three metabolic trait values (separate models for TG/HDL, LDL, T2D). Training and evaluation were carried out in a robust framework of five-fold nested cross-validation [47] (Methods). Input features for G2PT were defined as SNPs that have an independent marginal association with metabolic phenotypes, where significance of association was defined across a series of p-value thresholds of decreasing stringency (models built on SNP sets ranging from $p < 10^{-8}$ to $p < 10^{-4}$, Methods, Additional file 2: Table S1). The entire training procedure across all models required approximately 10 h using 4 NVIDIA A30 Graphics Processing Units (GPUs, Methods).

Following model training, we assessed the performance of G2PT in predicting the three metabolic traits for held-out UK Biobank individuals. Performance was benchmarked against a panel of widely used PRS and machine learning approaches. This panel included PRS methods based on SNP thresholding or thresholding and clumping (PRS T, PRS C + T, PRScs with threshold, SBayesRC with threshold) [51]; regularized linear and nonlinear regression models (ElasticNet, XGBoost) [52, 53]; and polygenic scoring methods based on the entire genome–wide collection of SNPs (Lassosum, LDpred2, PRScs, SBayesRC [54–57]). Across all SNP p-value thresholds, G2PT achieved an explained variance (R^2) that was significantly higher than either PRS or regularized regression approaches (Fig. 2a–c). Notably, this performance advantage persisted even in comparison to advanced models utilizing genome-wide summary statistics (LDPRED2, Lassosum2, PRScs, and SBayesRC). G2PT outperformed these dense models despite relying on only thousands of prioritized SNPs, in contrast to the hundreds of thousands of genome-wide approaches (Fig. 2a–c, right panels).

In addition to cross-validation studies within the UK Biobank, we also evaluated the degree to which predictions of the G2PT model were transferable to an independent cohort not seen during training. For this purpose we obtained genotype–phenotype data from self-reported White participants in the All of Us cohort [58], comprising 88,252 individuals of whom 21,260 had a prior diagnosis of T2D (covering 706 SNPs, Fig. 1b). When trained in

https://doi.org/10.1186/s13059-026-04327-0

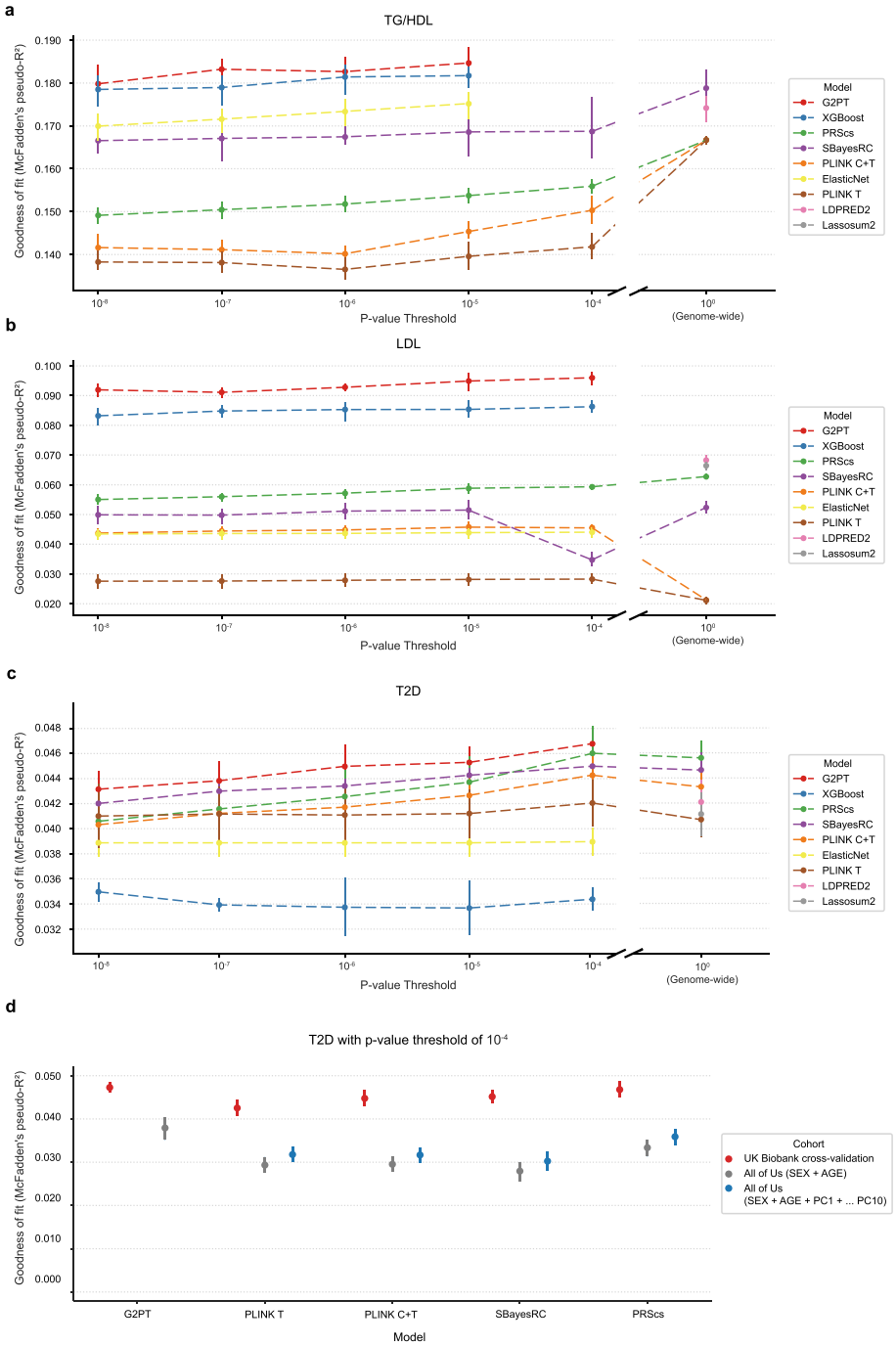


Fig. 2 Predictive performance of G2PT compared with PRS models. **a** Explained variance (R^2) of prediction of TG/HDL phenotype. The left section of each panel compares G2PT against ElasticNet, XGBoost, PRS clumping + thresholding (PLINK C + T), PRS thresholding (PLINK T), PRScs, and SBayesRC, across varying marginal p-value thresholds for SNP selection and corresponding number of SNPs. The right section (after the axis break) compares performance against various genome-wide PRS methods: PLINK T, PLINK C + T, LDPRED2, Lassosum2, PRScs, and SBayesRC. **b**, **c** As for panel (**a**) but for LDL and T2D phenotypes, respectively; T2D performance is measured using McFadden's pseudo-R^2. **d**, Validation of T2D prediction performance using a p-value threshold of 10^{-4}. Models are trained on the UK Biobank and performance is evaluated either in a held-out partition of UK Biobank (red) or on the All of Us cohort (blue) to assess transferability. Points represent mean performance over nested five folds of cross-validation, with error bars showing 95% confidence intervals

the UK Biobank and evaluated on this All of Us cohort, G2PT retained significant predictive power ($R^2 = 0.025$) albeit lower than seen in cross-validation (Fig. 2d). Notably, all competing methods also showed substantial reductions in predictive performance when transferred from UK Biobank to All of Us for T2D case–control prediction, with the degree of performance degradation varying across methods (Fig. 2d).

To investigate which aspects of the G2PT model were most responsible for the increased performance, we repeated our analysis over a series of ablation studies in which key G2PT architectural modules were removed. These studies included [1] removing knowledge of multi-gene systems, thereby training the Transformer with SNP inputs and SNP-to-gene mappings only; [2] removing knowledge of both systems and genes, leaving only

SNPs; and [3] removing hierarchical information flow between systems and supersystems, treating each system as an independent entity. The full G2PT model outperformed all of these ablations either modestly (most comparisons) or substantially (removal of systems and genes), indicating that incorporating hierarchical biological knowledge does not limit predictive capability compared to simpler black-box approaches (Additional file 1: Fig. S2). We also examined the effects of simplifying our SNP-to-gene mapping policy. We found that removing the more advanced eQTL and cS2G mapping methods, leaving a model based on the closest gene only, led to a significant albeit modest decrease in prediction performance (Additional file 1: Fig. S2). In addition, we quantified performance as a function of training set size (number of individuals) by subsampling the cohort and retraining G2PT with identical architecture and SNPs (Additional file 1: Fig. S3). We found that stable performance emerges by ~ 100 k samples for the continuous traits (TG/HDL, LDL) but continues to improve for the binary trait (T2D), consistent with lower per-sample information in case–control settings.

Transformer attention reveals genes and mechanisms underlying phenotype

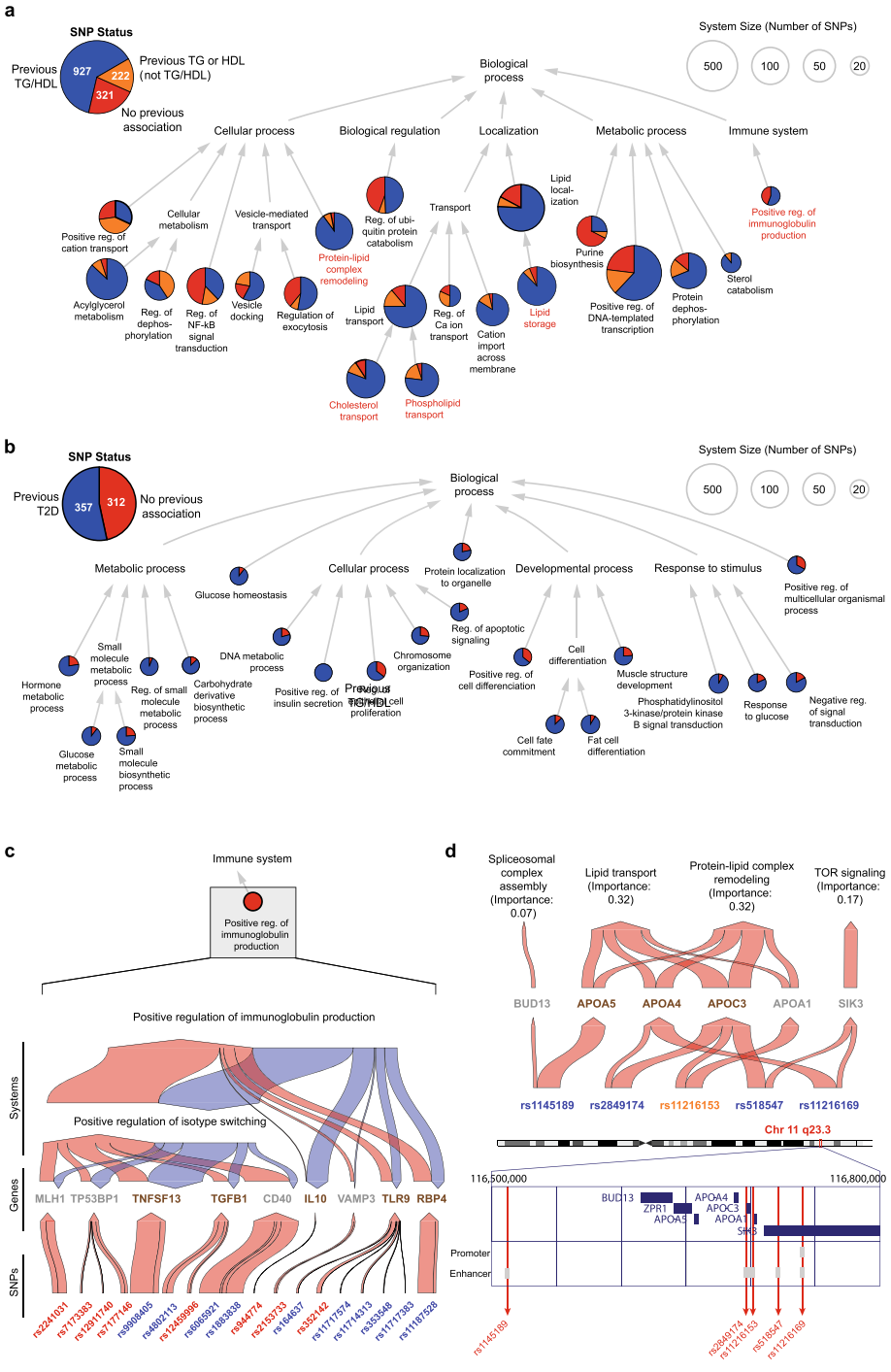
We next turned to mechanistic interpretation – studying the model's transformer attention mechanism to reveal the key genes and systems it had used for prediction, focusing on TG/HDL (Methods). As a positive control, we examined the lipoprotein lipase (*LPL*) gene, a known insulin resistance gene [59] which we expected should be given high attention by the G2PT framework. We indeed found participants for which G2PT gave high attention to variants in *LPL*, and that these individuals tended to have high predicted TG/HDL (Additional file 1: Fig. S4a). A similar relationship was observed for the molecular systems in which this gene is involved (e.g., lipid localization, Additional file 1: Fig. S4b).

Across all TG/HDL predictions for the 420 K + individuals, we identified significant model attention on 20 multigenic systems incorporating a total of 1,395 SNPs linked to 253 genes (Fig. 3a, Methods). Of these genes, 172 had been associated with TG/HDL ratio in our most recent GWAS [41], whereas the remaining 81 had not (Fig. 3a). In general, model attention on systems was higher than on genes (Additional file 1: Fig. S4c) and not significantly correlated with system size (number of genes), depth in the systems hierarchy, or gene-level importance (Additional file 1: Fig. S4d-f). Key systems included cholesterol and phospholipid transport, lipid storage, vesicle docking, and transcriptional regulation – corresponding to known metabolic pathways for serum lipids and their accompanying regulatory factors (Fig. 3a, Extended Data 1). We compared these results with MAGMA gene set analysis [60], finding moderate agreement between the Gene Ontology systems identified by the two approaches ($\rho = 0.26$, $p < 0.01$, Additional file 2: Table S2).

We also performed a parallel importance analysis for T2D using DIAMANTE consortium GWAS summary statistics [61] as reference. The top 20 important systems incorporated a total of 350 SNPs linked to 87 genes, and were generally smaller in gene content (10–30 genes per system) compared to those identified for TG/HDL, consistent with the more focused genetic architecture of T2D. Of the 350 SNPs, 131 had no prior association reported in DIAMANTE [61], suggesting that G2PT's hierarchical aggregation captures genetic signals not detected by single-variant analyses. The majority of identified systems reflected canonical T2D biology, including glucose homeostasis, carbohydrate derivative biosynthesis, regulation of insulin secretion, and fat cell differentiation. However, several systems warrant further investigation, including chromosome organization [62, 63], regulation of apoptotic signaling [64, 65], and muscle structure development [66, 67] — processes not prominently highlighted in previous T2D GWAS analyses but with plausible mechanistic connections to beta-cell dysfunction and peripheral insulin resistance. Comparison with MAGMA gene set analysis showed moderate concordance for canonical systems, while these discordant systems received high G2PT importance despite minimal MAGMA support (Additional file 2: Table S3), further demonstrating G2PT's ability to surface biologically relevant signals beyond the reach of standard single-variant approaches.

G2PT highlighted several systems not emphasized in previous TG/HDL gene mapping studies or in the MAGMA results, suggesting hypotheses for further investigation. In particular, immunoglobulin production, the

https://doi.org/10.1186/s13059-026-04327-0



fundamental process of adaptive immunity which involves the creation of antibodies by B cells in response to antigens, was broadly supported by model attention to 18 SNPs with effects distributed over 9 genes (Fig. 3c). Supporting this observation, clinical investigations of individuals with immunoglobulin-related deficiencies report alterations in serum lipids and insulin resistance [68]. Further investigation revealed that ten of the SNPs impacting this system had in fact been associated with TG/HDL previously [41], but independently from one another without any indication of a common biological process.

◂**Fig. 3** Multigenic systems determining prediction of TG/HDL and T2D. **a** Hierarchy of important multigenic systems. Top 20 systems are represented as circular nodes, with size proportional to the number of genes annotated to the system. Arrows represent involvement in broader systems (Gene Ontology "is_a" relation) or containment of one system by another ("part_of" relation). Pie charts show the proportion of genetic variants assigned to each system in three categories: SNPs previously associated with TG/HDL ratio (blue); SNPs associated with TG or HDL individually (orange); SNPs not previously reported for any of these phenotypes (red). The pie chart in the legend (upper left) shows the total number of variant counts across the systems hierarchy. Red labels denote systems highlighted in subsequent figures. **b** Hierarchy of important multigenic systems for T2D, displayed as in panel (a). Pie charts show the proportion of variants previously associated with T2D (blue) or with no previous association (red). **c** Genetic information flow within Positive regulation of immunoglobulin production (GO:0002639), a top-20 informative system. Red arrows represent information flow from SNPs to genes to systems (ascending layers). Blue arrows indicate flow from systems back to genes (reverse propagation). Arrow width is proportional to the amount of attention given by the model. SNPs previously associated with TG/HDL ratio are highlighted in blue; SNPs not previously associated with TG, HDL, or TG/HDL phenotypes are highlighted in red. Genes implicated due to a 'closest gene' mapping policy are colored in gray, whereas genes implicated due to other SNP-to-gene mappings (Methods) are dark brown. **d** Genetic information flow at chromosomal locus 11q23.3. Top: genetic information flow visualized similarly to panel c, focusing on forward propagation of information only. SNPs previously associated with TG/HDL ratio are highlighted in blue; SNP previously associated with HDL is highlighted in orange. Shown beneath each gene is the sum of attention scores from its potential SNP-to-gene mappings. Bottom: Positions of SNPs with respect to the chromosomal locus. The whole chromosome 11 ideogram is shown, with coordinates 116.5–116.8 Mb expanded to detail gene open reading frames (dark blue rectangles) and regulatory regions (promoters or enhancers; light gray)

For SNPs with several potential SNP-to-gene mappings (Methods), we found that G2PT often prioritized one of these genes in particular due to its membership in a high-attention system. For example, the chr11q23.3 locus contains multiple genes including the *APOA1/C3/A4/A5* gene cluster [69] (Fig. 3d) which is well-known to govern lipid transport, an important system for G2PT predictions (Fig. 3a). Due to high linkage disequilibrium in the region, all of its associated SNPs had multiple alternative gene mappings available. For example, SNP rs1145189 mapped not only to *APOA5* but to the more proximal *BUD13*, a gene functioning in spliceosomal assembly [70] (a system receiving substantially lower G2PT attention). Here, the relevant information flow learned by G2PT was from rs1145189 to *APOA5* to lipid transport and protein-lipid complex remodeling (Fig. 3d; and conversely, deprioritizing *BUD13* as an effector gene for TG/HDL). We found that this particular genetic flow was corroborated by exome sequencing [71, 72], which implicates *APOA5* but not *BUD13* in regulation of TG/HDL, using data that were not available to G2PT. Similarly, two other SNPs at this locus – rs518547 and rs11216169 – had potential mappings to their closest gene *SIK3*, where they reside within an intron, but also to regulatory elements for the more distant lipid transport genes *APOC3* and *APOA4*. Here, G2PT preferentially weighted the mappings to *APOC3* and *APOA4* rather than to *SIK3* (Fig. 3d). Altogether, of the 1,395 SNPs important to G2PT predictions, 252 were involved in multiple potential SNP-to-gene mappings where G2PT had prioritized one of these genes substantially above the others (> 2-fold difference in attention). These findings demonstrate how modeling genetic information flow through a hierarchy of SNPs, genes and multigenic systems provides a means to identify the major molecular pathways underlying a phenotype.

Model predictions invoke epistatic relationships

The favorable performance of G2PT in comparison to linear models (Fig. 2) suggested that its predictions may leverage nonlinear (epistatic) interactions among SNPs. To investigate, we initially focused on SNPs impacting phospholipid transport, a system that was highly important to G2PT predictions of TG/HDL (Fig. 3a) predominantly through the phospholipid efflux subsystem. SNPs in this system were first selected based on high model attention (Fig. 4a), from which pairs of SNPs spanning distinct loci were tested for pairwise interactions using a standard statistical model of epistasis (Methods). This screen yielded six SNPs involved in seven epistatic interactions (Fig. 4b). For example, we identified an interaction between SNPs rs11216169 and rs7499892 located at distinct chromosomal loci encoding the genes apolipoprotein A-IV (*APOA4*, 11q23.3) and cholesteryl ester

https://doi.org/10.1186/s13059-026-04327-0

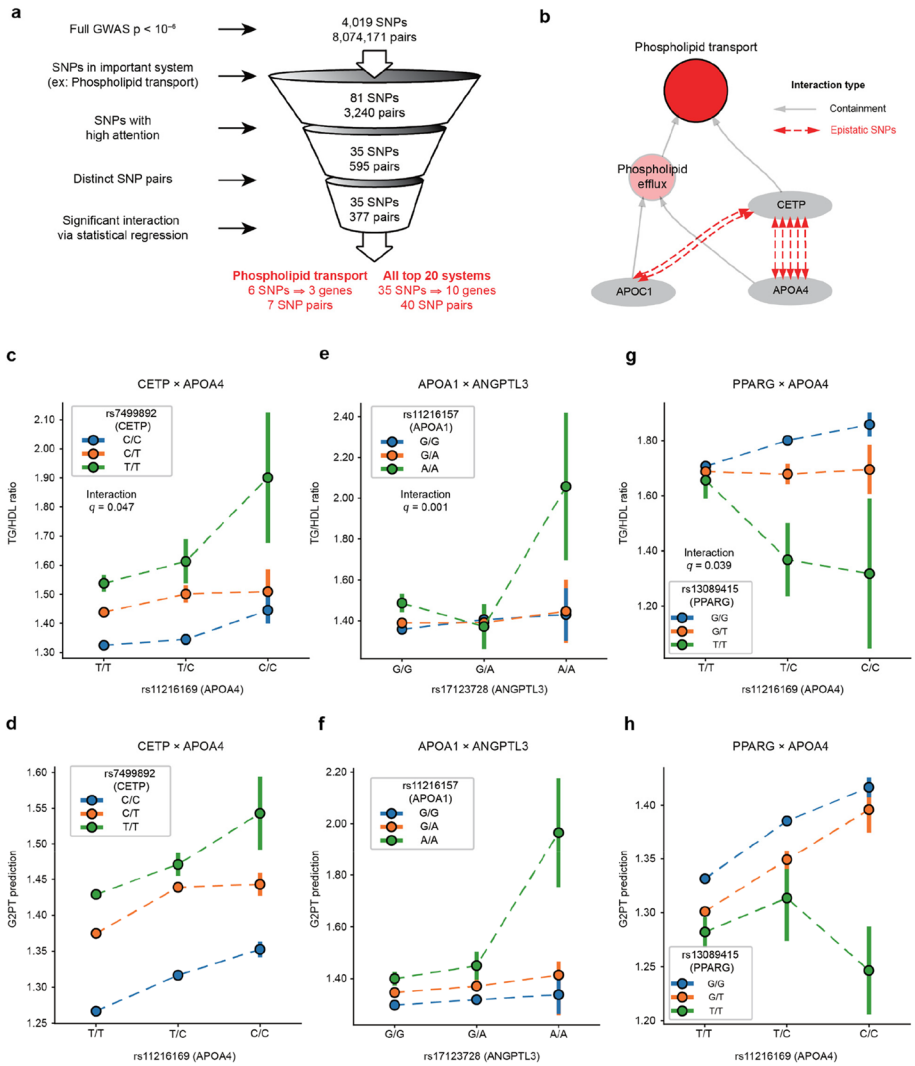


Fig. 4 Epistatic interactions identified by attention-based epistasis search. **a** Pipeline for detecting epistasis. Interacting SNP pairs are identified through a progressive multi-step screening process based on the system under study (here, Phospholipid transport), model attention and validation by statistical regression. **b** Hierarchy of subsystems and individual gene products within phospholipid transport (red circle), with gray arrows representing containment of a gene product (gray ovals) or subsystem (pink circles) within a larger system. Red dashed arrows indicate significant epistatic interactions. **c** Epistatic interaction of rs7499892 (CETP) with rs11216169 (APOA4) on measured TG/HDL ratio. Points and error bars show medians with standard error of TG/HDL ratio for subsets of individuals stratified by genotype. Colors denote rs7499892 genotype: C/C (blue, homozygous major allele), C/T (orange, heterozygous), and T/T (green, homozygous minor allele). X-axis denotes rs11216169 genotype: T/T (homozygous major allele), T/C (heterozygous), C/C (homozygous minor allele). **d** Similar to panel (c), but showing the TG/HDL values predicted by the G2PT model. **e**, **f** Epistatic interaction of rs17123728 (ANGPTL3) with rs11216157 (APOA1) on (**e**) measured or (**f**) predicted TG/HDL ratio. Display as per panel (c). **g**, **h** Epistatic interaction of rs13089415 (PPARG) with rs11216169 (APOA4) on (**g**) measured or (**h**) predicted TG/HDL ratio. Display as per panel (c)

transfer protein (*CETP*, 16q13), respectively. For both SNPs, the minor alleles were linked to an increase in TG/HDL (Fig. 4c). However, the impact of rs11216169 was significantly amplified in the rs7499892 T/T homozygous minor genotype, an effect which was not only seen in the observed TG/HDL ratio (Fig. 4c) but captured in the model prediction (Fig. 4d). Beyond this particular interaction, the *CETP* locus harbored a total of five SNPs exhibiting epistasis with *APOA4* (Fig. 4b, Additional file 2: Table S4). Notably, the *APOA4* and *CETP*

gene products are secreted into blood (from intestine and liver respectively), in which they associate with HDL particles in regulation of cholesterol ester transport [73, 74], supporting a functional biochemical interaction.

Beyond phospholipid transport, we broadly applied this same epistasis screening procedure across all 20 systems most important to TG/HDL predictions (Fig. 4a). This broader screen identified 40 SNP-SNP pairs with significant epistasis (Additional file 2: Table S4). One significant interaction was identified in the lipid storage system, between variants at the apolipoprotein A-I (*APOA1*, 11q23.3) and angiopoietin-like 3 (*ANGPTL3*, 1p31.3) loci; here phenotypic impact was seen predominantly in the combination of homozygous minor states for both variants (Fig. 4e, f). This interaction was also supported by the All of Us cohort (Additional file 1: Fig. S5). As another example, screening of the cholesterol transport system again highlighted the *APOA4* gene, in which the rs11216169 minor allele exposed a strong suppressive effect of the rs13089415 variant upstream of peroxisome proliferator-activated receptor gamma (*PPARG* at locus 3p25.2, Fig. 4g, h).

Validation of epistasis through simulation studies

Having identified epistatic interactions in the study of metabolic phenotypes, we sought to characterize more generally G2PT's ability to recover such non-additive effects under controlled conditions. For this purpose we developed a simulation–retrieval framework mirroring G2PT's hierarchical architecture (Additional file 1: Fig. S6a). In each simulation a random systems hierarchy was constructed, after which SNPs were assigned phenotypic effects of varying strengths as well as, for select SNP pairs, epistatic interactions. The choice of epistatic SNP pairs was governed by an ontology coherence parameter, ranging from $\rho = 1$ (every epistatic interaction occurs between genes in a common leaf system) to $\rho = 0$ (every epistatic interaction occurs between genes in two unrelated systems). Analysis of simulation results confirmed that epistasis detection accuracy increases consistently with ontology coherence (Additional file 1: Fig. S7). High ρ values improved recovery by concentrating epistasis within coherent systems, whereas low ρ values reduced epistasis detection to random chance. While absolute accuracy depended on other simulation parameters such as the sample size and number of epistatic interactions, G2PT maintained substantial improvement over random chance and over baseline methods such as PLINK [75]. At parameter settings consistent with current GWAS sample sizes and understanding of heritability [76, 77] (Additional file 1: Fig. S6b), we observed a simulated AUROC of 0.694 (Additional file 1: Fig. S6c), corresponding to recovery of approximately 10% of all epistatic relationships with $> 50\%$ precision (Additional file 1: Fig. S6d). Collectively, these simulations demonstrate that G2PT identifies epistatic interactions across diverse genetic architectures and clarify how detection strength scales with ontology coherence, sample number, and effect size.

Discussion

Here we have introduced an approach to genotype–phenotype translation using a graph-based transformer (Fig. 1), a significant neural network architecture arising from recent machine learning research [78, 79]. When applied to traits inducing insulin resistance (TG/HDL), serum LDL cholesterol, and T2D case status, this architecture outperforms genome-wide PRS methods (Fig. 2) despite being constrained to a few thousand loci. When trained in UK Biobank participants of European genetic ancestry and evaluated in self-reported White participants from the All of Us cohort, G2PT retained measurable predictive power for T2D, whereas competing PRS-based methods exhibited substantial performance attenuation under the same UK Biobank-to- All of Us transfer setting. Importantly, as all methods were evaluated under identical transfer conditions, the relative performance advantage of G2PT remains informative (Fig. 2d). This divergence is consistent with a well-recognized limitation of PRS portability, which can be influenced by differences in linkage disequilibrium, allele frequencies, and residual population structure [80], and it suggests that G2PT's hierarchical aggregation of variant effects through genes and biological systems yields representations that are more robust to cross-cohort heterogeneity.

Detailed examination of model attention identified a core set of biological systems driving phenotypic predictions, which integrate signals from hundreds of genes including expected and unexpected factors (Fig. 3). Among these was an unexpected system, immunoglobulin production, based on genotypic aggregation of 18 genetic variants impacting 9 genes (Fig. 3a, b). While some genes in this system were newly linked to metabolic phenotypes, others such as RBP4 are established biomarkers associated with insulin resistance in humans [81, 82] (but through distinct mechanisms from immunoglobulin production). Although few studies have directly linked insulin resistance to immunoglobulins, research in mouse models of diet-induced obesity suggests that B-cell infiltration into adipose tissue — producing pathogenic IgG — can drive inflammation and disrupt insulin signaling, ultimately contributing to systemic insulin resistance [83]. These observations suggest plausible mechanisms for further investigation in humans.

This ability of the G2PT model to identify key SNPs and systems was greatly aided by the formal mechanisms of attention inherent to Transformer architectures. During the 'genetic factor propagation phase' of modeling (Fig. 1a), we used multi-head attention to transmit the effects of SNPs across the knowledge hierarchy of genes and systems, allowing us to inspect and trace the cascade of molecular entities impacted by a genotype (Fig. 3). During subsequent 'genetic factor translation', single-head attention was used to quantify the impacts of the altered genes and systems on an individual's phenotype. In this way, genetic factor propagation facilitates interpretation of the model by revealing complex information flows, whereas genetic factor translation integrates across these mechanisms to produce a single unified attention value used for interpreting the mechanisms underlying phenotypic risk.

A second notable aspect of the G2PT architecture is the bidirectional flow of information among genes and multigenic systems. Variant effects are first transmitted upwards in scale to impact genes and their collective functions, after which this flow is reversed to enable the functional states of systems to influence how specific variants impacting that system are interpreted (Fig. 1a). This reverse propagation step captures the biological context in which genes and variants operate (e.g., Fig. 3b), and it promotes cross-talk among multiple genetic variants that may have conditional interrelationships. These aspects enabled G2PT to learn gene–gene epistatic interactions across a variety of biochemical mechanisms. For example, in addition to interactions among apolipoproteins and their potential interactors (e.g. *APOA4* and *CETP*, Fig. 4c), G2PT identified novel epistasis between *ANGPTL3*, the major secreted inhibitor of lipoprotein lipase targeted by triglyceride-reducing therapies [84], and *APOA1*, the major protein component of HDL particles (Fig. 4e). While these two gene products have not been reported to interact physically, recent data suggests they are transcriptionally regulated by the same hepatic nuclear factor *HNF1A* [85], providing a potential mechanism for the observed genetic epistasis. Similarly, the identified epistatic interaction (Fig. 4g) between *APOA4*, an intestinally secreted triglyceride-rich lipoprotein factor, and *PPARG*, a nuclear hormone receptor and transcription factor, has a strong biological basis. The promoter of *APOA4* contains *PPAR* response elements, and *APOA4* expression is upregulated by the closely associated nuclear hormone receptor *PPARA* [86]. G2PT also found a non-canonical epistatic interaction between *TGFB1* and *FKBP1A* related to protein dephosphorylation (Additional file 2: Table S4). In particular, variants at the *TGFB1* locus can serve to increase TGF-β ligand–driven receptor/SMAD phosphorylation [87], while variants at *FKBP1A* can weaken receptor activity [88]. Where both risk genotypes co-occur, this could plausibly lead to disproportionate SMAD activation, exaggerating transcriptional programs that raise TG and lower HDL, producing an epistatic increase in the TG/HDL ratio.

Despite the promise of Transformers for genotype–phenotype modeling, computational cost remains a principal bottleneck. Our use of the xFormer memory-efficient attention mechanism substantially reduced the memory footprint and accelerated training (Additional file 1: Fig. S8). Regardless, the current pipeline was capped at ~5,000 SNP inputs before memory and interconnect overheads were found to dominate. We expect that parallelized, sharded execution can lift this limit in the near future, enabling larger SNP sets, faster convergence, and deeper hyperparameter exploration. We also experienced challenges arising from our use of Gene Ontology as a prior. This knowledgebase includes many redundant groups of systems with nearly identical sets of genes, and it has a

natural bias towards well-studied systems [32]. An important step moving forward will be to explore alternative knowledge structures, such as Reactome [89], GO Causal Activity Models (GO-CAMS) [90] or maps of biological structures and systems resolved directly from 'omics data [91–93]. Finally, our interpretability results relied on model attention, which we treat as hypothesis-generating rather than causal evidence. Correlations between attention and predictions do not establish mechanisms, and while we included stress tests (e.g., masking/ablation and permutation), such tests provide model-level, not biological, validation. Regardless, G2PT has immediate application to the genetic analysis of diverse phenotypes of interest, including those related to multigenic diseases such as autism, aging, or cancer. This work also has implications outside of the life sciences, insofar as it presents a general template for constructing interpretable Transformer architectures across deep learning challenges.

Conclusions

We have presented G2PT, a hierarchical graph transformer that translates an individual's genotype into phenotype predictions while simultaneously revealing the biological systems and epistatic interactions underlying those predictions. Applied to metabolic traits including insulin resistance, LDL, and T2D, G2PT consistently outperforms state-of-the-art polygenic risk score methods and demonstrates stronger transferability than competing methods in an independent external cohort restricted to self-reported White participants. Beyond prediction, G2PT identifies biologically meaningful systems and epistatic interactions — including novel relationships among apolipoproteins, lipid transport regulators, and immune factors — that generate testable hypotheses for follow-up investigation. Together, these results establish hierarchical graph transformers as a powerful and interpretable framework for genotype-to-phenotype modeling. As biobank cohorts continue to grow and biological knowledge graphs become more comprehensive, we anticipate that architectures of this kind may play an increasingly central role in precision medicine, enabling not only risk stratification but mechanistic understanding of complex polygenic disease.

Methods

UK biobank data processing

Human genotype and phenotype data were obtained from the UK Biobank [48], in which participants of European ancestry had been genotyped with SNP arrays and characterized for TG and HDL levels in millimoles per liter (mmol/L). Type 2 diabetes designation was based on a combination of self-report or ICD-10 codes of E11.X in hospitalization or primary care records or having HbA1C level $\geq$ 48 mmol/mol as detailed in previous studies [11, 94] (Additional file 2: Table S5). Genotypes were represented by vectors of SNPs, where each SNP was encoded as 0 to represent the homozygous major allele (reference), 1 the heterozygous major/minor allele, and 2 the homozygous minor allele, yielding a participant-by-SNP genotype matrix. Variants were restricted to 203,126 SNPs (Additional file 2: Table S6) retained after intersection and allele harmonization among the LDpred2 HapMap3/HapMap3 + candidate set, the imputed UK Biobank genotypes, and the GWAS summary statistics after allele matching and QC. We performed a Bayesian regression analysis using BOLT-LMM [95] to identify SNPs that statistically associate with $\log_2$(TG/HDL), normalized LDL level, or T2D status, while including sex, age, and the top 10 principal components as covariates. Significantly associated SNPs for each trait were then retained based on various p-value thresholds from 10^{-4} to 10^{-8} (Fig. 2, from 10^{-5} for TG/HDL due to computational limitations). These SNPs were provided as inputs to the G2PT model, described below, without further pruning or clumping. This simple input policy was by design, since a major advantage of the transformer architecture is to direct attention to individual features (here SNPs) within a set of correlated input signals (here a locus).

All of us data processing

Genotype data from the All of Us Research Program (v8) were accessed through the Researcher Workbench using whole-genome sequencing data stored in Variant Dataset format. For external evaluation, we restricted the All of Us cohort to self-reported White participants and analyzed only the SNP subset used as G2PT input features in UK Biobank, thereby constructing a matched variant set across cohorts. Because UK Biobank variant coordinates were defined in GRCh37, target loci were lifted over to GRCh38 using Hail's built-in liftover function with the UCSC chain file, and loci that failed liftover were excluded. The All of Us WGS VDS was then subset to these loci using Hail (v0.2). The local genotype field (LGT) was reassigned as the standard genotype (GT) field to ensure export compatibility. All processing was performed within the All of Us Researcher Workbench using Hail, Apache Spark, and Google Cloud Platform. For phenotype and covariate processing, sex was defined using the recorded sex-at-birth field, and age was calculated as the difference between the measurement date and the participant's date of birth. For the PC-adjusted evaluation of non-G2PT PRS methods, we used the first ten genetic principal components (PC1–PC10) provided in the All of Us Controlled Tier CDR v9 ancestry resource. These components represent projections of All of Us genotypes into a PCA space defined by Human Genome Diversity Project and 1000 Genomes reference samples [96, 97]. These PCs were not added to the transferred G2PT model. To harmonize phenotype definitions across cohorts, T2D status in All of Us was assigned using a definition analogous to that used in UK Biobank, based on diabetes-related diagnosis records and/or HbA1c $\geq$ 6.5% (48 mmol/mol) (Additional file 2: Table S5).

SNP-to-gene mapping

Significantly associated SNPs were mapped to (potentially multiple) protein-coding genes using the union of (1) associations from cS2G [50], (2) eQTL associations from GTEx v7 [49], or (3) the nearest gene in genomic coordinates (Genome Reference Consortium hg19 assembly [98]). For the eQTL SNP-to-gene mappings, we used adipose subcutaneous, adipose visceral omentum, liver, pancreas, adrenal gland, muscle skeletal, and uterus tissue types [41, 43].

Gene ontology knowledge hierarchy

The knowledge graph used for G2PT was based on the Gene Ontology [32] Biological Process database (version 2023–07–27). This GO version was pruned using the DDOT package [99] to include only those terms (systems) relevant to the genes with significant SNP mapping (see 'SNP-to-Gene Mapping' above). In particular, systems with < 5 mapped genes were excluded, after which we further excluded all systems for which the annotated genes exactly matched to those of a subsystem. The final directed acyclic graph contained three types of nodes, representing SNPs, genes and systems, and three types of directed edges, representing SNP $\rightarrow$ gene mappings, gene $\rightarrow$ system annotations, and subsystem $\rightarrow$ supersystem (is_a and part_of) relations from GO.

G2PT model

G2PT uses a hierarchical graph transformer to integrate and distribute genotypic information across different levels in a knowledge hierarchy (here Gene Ontology as detailed above). The model uses self-attention for residual message-passing between SNPs and other biological entities in the hierarchy (systems or genes), thus propagating the effects of genetic variations on higher order biological states and translating these altered states to predict phenotypes (Fig. 1). Specific details are divided into the following subsections: Model Initialization and Graph Construction, Hierarchical Graph Transformer, Genetic Factor Propagation Phase, Genetic Factor Translation Phase, Model Training and Comparative Evaluation, Model Interpretation by Scoring Importance of Genes and Systems.

Model Initialization and graph construction

All node embeddings—corresponding to SNPs, genes, and ontology-defined systems—were initialized randomly using uniform Xavier initialization, consistent with the model's hidden dimension size. The attention mechanism utilizes the pre-defined connectivity between successive ontology levels to guide attention flow. During training, these hierarchical masks constrain message passing and self-attention to occur only along valid biological connections, allowing variant-level information to propagate upward through genes and systems while enabling feedback across connected subsystems.

Hierarchical graph transformer

The hierarchical graph transformer (HiGT) is a modified version of a Graph Attention Transformer [29, 100] that leverages a hierarchical knowledge graph encoded into three sparse adjacency matrices (see above sections 'Gene Ontology Knowledge Hierarchy' and 'Model Initialization and Graph Construction'). For each node i with embedding state E_i, the HiGT function transforms this embedding by incorporating effects from graph neighbors j (Eq. 1). This transformation is computed from the neighbor embeddings via multiple attention heads Attn_h, which are provided as input to a Feed Forward Neural Network (FFNN) with Layer Normalization (LN). Each head is used to perform a linear projection of the neighbor embeddings (Eq. 2) scaled by an attention weight α (Eq. 3). In particular:

$$\text{HiGT}\left(E_i; \left\{E_j | j \in \text{neighbors}_i\right\}\right) = \text{FFNN}\left(\text{LN}\left(\sum_j \text{MultiHeadAttn}\left(E_i, E_j\right) W^O\right)\right) \tag{1}$$

$$\text{MultiHeadAttn}\left(E_i, E_j\right) = \text{Concat}\left(\text{Attn}_1\left(E_i, E_j\right), \cdots, \text{Attn}_h\left(E_i, E_j\right), \cdots, \text{Attn}_n\left(E_i, E_j\right)\right) \tag{2}$$

$$\text{Attn}_h\left(E_i, E_j\right) = \alpha_{ij,h} E_j W_h^v \tag{3}$$

$$\alpha_{ij,h} = \text{softmax}\left(\frac{\left(E_i W_h^q\right)\left(E_j W_h^k\right)^T}{\sqrt{d}}\right) \tag{4}$$

where W^q, W^k, W^v are learnable weight matrices of dimension $(d/h) \times d$ encoding the central concepts of query, key, and value used in computing self-attention [29]. W^O is a learnable weight matrix of dimension $d \times d$. The dimension d encoding the size of each node embedding was set to $d = 64$, the largest dimensionality that was practical given the available compute infrastructure (4 A30 GPUs). To balance complexity in a multi-headed attention model, the number of heads was set to $n = 4$. The feed-forward neural networks follow the original Transformer architecture—a two-layer position-wise network with a hidden dimension four times the node embedding size (4×64) and a GeLU activation function. This HiGT function was used as the central mechanism to update the embeddings for all nodes in sequence, as described in 'Genetic Factor Propagation' below.

Genetic factor propagation phase

G2PT models the effects of genetic alterations via the sequential update of node states by forward propagation from SNPs to genes to systems, followed by reverse propagation from systems to genes, in a single pass. All of these updates use the HiGT formulation (Eq. 1) selecting from specific types of (target → source) edges in the knowledge graph. First, the embedding of each gene is updated given its incoming SNPs:

$$E_i = E_i + \sum_z \mathrm{HiGT}_z\left(E_i, \left\{E_j | j \in \mathrm{source}_i\right\}\right) \forall i \in \mathrm{genes}, \forall j \in \mathrm{SNPs}_z \tag{5}$$

Different HiGT modules are applied based on the zygosity of the SNP alteration, $z \in \{\text{hetero}, \text{homo}_{\text{minor}}\}$. Homozygous major SNPs represent the reference state and thus are not considered a genetic alteration. Second, the updated genes propagate their state changes upward to the systems containing these genes:

$$E_i = E_i + \mathrm{HiGT}\left(E_i, \left\{E_j | j \in \mathrm{source}_i\right\}\right) \forall i \in \mathrm{systems}, \forall j \in \mathrm{genes} \tag{6}$$

Third, state changes of systems j are propagated to their supersystems i:

$$E_i = E_i + \mathrm{HiGT}\left(E_i, \left\{E_j | j \in \mathrm{source}_i\right\}\right) \forall i \in \mathrm{systems}, \forall j \in \mathrm{systems} \tag{7}$$

Following forward propagation, reverse propagation steps distribute the state changes downward to subsystems and genes. These reverse steps model how the particular state of a molecular system can either protect against, or expose vulnerabilities to, alterations in its component subsystems or genes. Here, state changes of systems j are reverse propagated to subsystems i:

$$E_i = E_i + \mathrm{HiGT}\left(E_i, \left\{E_j | j \in \mathrm{target}_i\right\}\right) \forall i \in \mathrm{systems}, \forall j \in \mathrm{systems} \tag{8}$$

after which state changes of systems j are reverse propagated to genes i:

$$E_i = E_i + \mathrm{HiGT}\left(E_i, \left\{E_j | j \in \mathrm{target}_i\right\}\right) \forall i \in \mathrm{genes}, \forall j \in \mathrm{systems} \tag{9}$$

In summary, the effect of the above forward and reverse procedure is to:

(1) Forward: update all genes based on component SNPs
(2) Forward: update all systems based on component genes
(3) Forward: update all systems based on component subsystems
(4) Reverse: update all systems based on containing supersystems
(5) Reverse: update all genes based on containing systems

Overall, the HiGT functions used in each of these steps form distinct layers of the model with separately learned weights.

Genetic factor translation phase

For each participant, G2PT projects a feature vector of covariates $P^{cov} = (\text{sex}, \text{age}, \text{PC1}, \ldots, \text{PC10})$ onto an embedding of size d using a Multi-Layer Perceptron, yielding a participant embedding P^{embed}. For the T2D model used in the UK Biobank-to-All of Us transfer analysis, principal components were excluded from the covariate inputs because the UK Biobank and All of Us PC coordinates were not derived from the same PCA reference and were therefore not directly comparable across cohorts; accordingly, $\mathrm{P}^{\mathrm{cov}} = (\text{sex}, \text{age})$.

$$P^{\mathrm{embed}} = \mathrm{MLP}\left(P^{\mathrm{cov}}\right), where \mathrm{MLP} : \mathbb{R}^m \rightarrow \mathbb{R}^d \tag{10}$$

where $m = 12$ for the UK Biobank models that include ten PCs and $\mathrm{m} = 2$ for the transfer model; m denotes the number of covariates. G2PT then uses the final embedding states of genes and systems (see above section 'Genetic Factor Propagation') to update P^{embed}. This update occurs via the HiGT function (Eq. 1) by representing the participant as node i with embedding state $E_i = P^{embed}$, and all genes and systems as graph neighbors j with states E_j. Due to the large number of genes and systems updating a single P^{embed}, the attention values in these operations, $\alpha_{ij,h}$, are computed using Differential Attention [101] with a single attention head $n = 1$ to facilitate interpretation

(see below section 'Model Interpretation'). The updated embeddings from these modules are concatenated and projected through a final layer to yield $\widehat{Y}\widehat{Y}$, a prediction of phenotype YY, the $\log_2$(TG/HDL) phenotype. Altogether, this layer can be formulated as:

$$\widehat{Y} = W^{\text{pred}}\left(\text{concat}\left(\text{HiGT}\left(P^{\text{embed}}, \{E_j|\forall j \in \text{systems}\}\right), \text{HiGT}\left(P^{\text{embed}}, \{E_j|\forall j \in \text{genes}\}\right)\right)\right) \tag{11}$$

where W^{pred} represents the learned prediction weights.

Model training and comparative evaluation

Separate models were trained for the three traits under study (TG/HDL, LDL, T2D), using nested cross-validation to train and robustly evaluate model performance in genotype–phenotype prediction. In each of five cross-validation folds, data were split into training, validation, and test sets in 3:1:1 proportion. The training set was used for selecting significant SNPs and fitting model parameters, the validation set was used for tuning model hyperparameters, and the test set comprised held-out samples for independent evaluation of performance. The model was trained to minimize the mean squared error between the true and predicted trait values, optimized with decoupled weight decay regularization (AdamW and L2 regularization) [102]. Hyperparameters were optimized through a grid search with p-value threshold set to $p = 10^{-8}$ (Additional file 1: Fig. S9). Starting with the optimal baseline configuration, we further tuned the G2PT p-value threshold and training epochs, and we modified the architecture by reducing the hidden dimension and increasing the number of attention heads to manage computational costs. For T2D, the G2PT checkpoint used for performance evaluation and subsequent transfer to the All of Us cohort was selected based on the highest McFadden's pseudo-R^2 observed in the UK Biobank validation set. For non-G2PT T2D PRS evaluation in UK Biobank, we fitted logistic regression models with T2D status as the outcome and PRS, age, sex, and PC1–PC10 as predictors, including an intercept. Performance of the fitted full model was quantified using McFadden's pseudo-R^2 as follows:

$$R^2_{\text{McFadden}} = 1 - \frac{\log L_{\text{full}}}{\log L_{\text{null}}} \tag{12}$$

where L_{full} denotes the likelihood of the full logistic regression model, including the PRS and the specified covariates, and L_{null} denotes the likelihood of the intercept-only model, evaluated on the same individuals. This metric reflects the full model's fit relative to an intercept-only baseline, not the incremental contribution of the PRS beyond covariates. We also optimized parameters for other models: for XGBoost, the number of estimators, maximum depth, subsampling rate, and learning rate; and for ElasticNet, the alpha (overall strength of regularization), L1 ratio, and tolerance. For LDPred2 and Lassosum models, effect sizes of SNPs were adjusted by LD information from the training population. For performance evaluation in the All of Us cohort, McFadden's pseudo-R^2 was calculated for each non-G2PT PRS using two logistic regression specifications: PRS + age + sex, and PRS + age + sex + the first ten All of Us-provided genetic PCs (PC1–PC10).

Model interpretation by scoring importance of genes and systems

During the genetic factor translation phase (see above), G2PT assigns two sets of attention values, $\{\alpha_{ik}\}$ and $\{\alpha_{jk}\}$, to weight the effects of $g_i \in \{genes\}$ and $s_j \in \{systems\}$ towards the phenotype predictions of $p_k \in \{participants\}$. To assess the importance of these genes and systems across the population, we used the fully trained G2PT model and calculated Pearson correlations between the individual attention values and the predicted phenotype values, separately for males and females. We then averaged the absolute values of these correlations, which served as the importance score assigned to each system and gene.

Detection of epistatic interactions in important systems

We implemented a search for epistatic interactions (Fig. 4a) for each system falling in the top 20 by importance (see 'Model Interpretation by Scoring Importance' above). For each of these systems, we first selected all SNPs mapping to genes (see 'SNP-to-Gene Mapping' above) with GO annotations to that system or its children. A chi-square test was used to select SNPs that exhibited significant differences in frequency between individuals with high versus low attention to the system in question (top or bottom 10% in the whole population of individuals ranked by system attention). Retained SNPs passing multiple-testing control within the system advanced to pairwise testing. SNP pairs (*SNP1*, *SNP2*) were tested for a statistical interaction in prediction of the phenotype according to a combinatorial linear model [103, 104]:

$$y = \alpha_1 \cdot \mathrm{SNP1} + \alpha_2 \cdot \mathrm{SNP2} + \alpha_{12} \cdot \mathrm{SNP1} \times \mathrm{SNP2} + \beta \cdot \mathrm{sex} + \gamma \cdot \mathrm{age} \tag{13}$$

where y is the continuous phenotype value (TG/HDL or LDL). Each α represents the effect size of the corresponding SNP or SNP interaction, and β and γ represent the effect sizes of the sex and age covariates, respectively, with all α, β and γ values estimated from the population data. The significance of the interaction term ($\alpha_{12} \not\approx 0$) was evaluated using Wald test corrected through the Benjamini–Hochberg procedure with adjusted p-value threshold of 0.05. SNP pairs located within 1 Mb of each other were excluded in the above search, to avoid the confounding effects of linkage disequilibrium. For T2D, a binary trait, Eq. (13) above implemented logistic rather than linear regression.

Epistasis simulation

To evaluate G2PT's ability to recover non-additive interactions, we simulated genotype–phenotype datasets following a sequential three-step process. First, we used a previous method [105, 106] to generate a random genotype matrix Z of dimensions $p \times n$ containing allele states across p individuals and n SNPs. We randomly designated a subset of these SNPs, A, to carry additive effects, and a distinct subset of SNPs to carry epistatic interactions via SNP pairs E (set sizes $|A|$ and $|E|$ were determined by hyperparameters). A vector of phenotypes y over all individuals was then generated by an additive-epistatic model:

$$y = \sum_{i \in \mathrm{A}} \beta_i z_i + \sum_{(j,k) \in \mathrm{E}} \gamma_{j,k} \left(z_j \odot z_k \right) + \varepsilon \tag{14}$$

where the vector $\boldsymbol{z_i}$ holds values of SNP i drawn from $\boldsymbol{Z}$ across all individuals, the scalar β_i is the SNP i additive effect size, the scalar $\gamma_{j,k}$ is the interaction effect size between SNPs j and k, $\odot$ denotes the element-wise product, and ε represents standard normal error. Effect sizes for each additive SNP and epistatic SNP pair were determined using the equations:

$$\beta = \beta^{\mathrm{raw}} \sqrt{\frac{h^2_{\mathrm{add}}}{\mathrm{Var}(Z_{\mathrm{add}} \beta^{\mathrm{raw}})}}, \gamma = \gamma^{\mathrm{raw}} \sqrt{\frac{h^2_{\mathrm{epi}}}{\mathrm{Var}\left(Z_{\mathrm{epi}} \gamma^{\mathrm{raw}}\right)}} \tag{15}$$

where Z_{add} is the $p \times |A|$ sub-matrix of $\mathbf{Z}$ restricted to the additive SNP set A, $\mathbf{Z_{epi}}$ is the $p \times |E|$ matrix of interaction terms formed by the element-wise products ($z_j \odot z_k$) for all epistatic pairs in E, and base effect size vectors ($\boldsymbol{\beta^{raw}}$, $\boldsymbol{\gamma^{raw}}$) were drawn from a standard normal distribution and scaled by heritability hyperparameters $h^2_{\mathrm{add}} h^2_{\mathrm{add}}$ and $h^2_{\mathrm{epi}} h^2_{\mathrm{epi}}$ to yield the final effect size vectors ($\boldsymbol{\beta}$, $\boldsymbol{\gamma}$) [107, 108].

Second, a systems ontology was constructed in a hierarchical bottom-up manner by recursively aggregating a set of initial leaf systems S to yield a balanced binary tree, in which the number of systems decreased by a factor of two at each successive level until a single aggregate system remained. A universe of genes G was created and mapped to leaf system chosen uniformly at random (set sizes $|S|$ and $|G|$ were determined by hyperparameters).

Third, SNPs were assigned to the ontology as follows. SNPs in the additive SNP set A were assigned to randomly chosen genes. SNP pairs in the epistasis set E were mapped 1:1 to pairs of leaf systems (u,v) selected as follows:

$$\left\{(u,v)|\ u,v \in\ S, \mathrm{Depth}(\mathrm{LCA}(u,v)) \approx \lfloor \rho\ \cdot D_{\max} \rfloor \right\} \tag{16}$$

where LCA(u, v) denotes the Lowest Common Ancestor of the leaf systems u and v, and D_{max} represents the maximum depth of the ontology (distance from the root to the leaves). The "ontology coherence" hyperparameter, ρ, ranges from 0 to 1, shifting (u,v) from distant systems (connected only by the ontology root) to identical leaf systems ($u = v$).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1186/s13059-026-04327-0.

Acknowledgements We thank the participants of the UK Biobank and the All of Us Research Program for their contributions to this research. The All of Us Research Program is supported by the National Institutes of Health, Office of the Director. UK Biobank data were accessed under application numbers 51436. We thank the members of the Ideker and Majithia laboratories for helpful discussions.

Peer review information This article was first peer reviewed at Review Commons and the original paper, reviewer reports and point-by-point response are available online113. The peer-review history at Genome Biology is available in the online version of this article. Claudia Feng and Leonardo Bottolo were the primary editors of this article at Genome Biology and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions I.L. conceived the study, developed the G2PT framework, performed the computational analyses, and wrote the original draft of the manuscript. Y.W. performed data processing. S.P. and H.N. provided guidance on the study design and methodology. Z.W. contributed to the development of ideas and revised the manuscript. A.R.M. and T.I. supervised the project, interpreted the results, and critically reviewed the manuscript for important intellectual content. All authors read and approved the final manuscript.

Funding We are grateful to the National Institutes of Health for funding for this project through the following awards: Bridge2AI Common Fund (T.I.; OD032742); National Resource for Network Biology (T.I.; GM103504); National Institute of Diabetes and Digestive and Kidney Diseases (A.R.M.; DK123422) and National Heart Lung and Blood Institute (A.R.M.; HL159760). This research was also supported by the US Department of Veterans Affairs (VA, A.R.M.; I01BX006293).

Data availability Individual-level genomic and phenotypic data from the UK Biobank are available to researchers through an application process at https://ukbiobank.ac.uk (application numbers 51436). Functional genomic annotations used for SNP-to-gene mapping were obtained from the cS2G resource. eQTL data were sourced from GTEx v7 (dbGaP accession phs000424.v7.p2), accessible at https://www.gtexportal.org/home/downloads/adult-gtex/qtl. GO data were downloaded from the 2023–07-27 release. The source code for G2PT is available on GitHub (https://github.com/idekerlab/G2PT) under the MIT License. A versioned release of the source code used in this manuscript is archived on Zenodo at https://doi.org/10.5281/zenodo.19713271.

Declarations

Ethics approval and consent to participate The UK Biobank study was approved by the North West Multi-Centre Research Ethics Committee (REC reference: 11/NW/0382). Analysis of UK Biobank data was conducted under application numbers 51436. All UK Biobank participants provided written informed consent at the time of enrollment. The All of Us Research Program was approved by the Multi-Regional Clinical Trials Center of Brigham and Women's Hospital and Partners HealthCare Institutional Review Board (Protocol #: IRB16-0360). All All of Us participants provided written informed consent. All procedures were conducted in accordance with the ethical standards of the responsible institutional and national committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2013.

Consent for publication Not applicable.

Competing interests T.I. is a co-founder, advisor, and holder of equity for Data4Cure and Serinus Biosciences, and he is an advisor and shareholder for Ideaya BioSciences and Eikon Therapeutics. A.R.M is an advisor to Terns Pharmaceuticals.

The terms of these arrangements have been reviewed and approved by UC San Diego in accordance with its conflict of interest policies.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Suzuki K, et al. Genetic drivers of heterogeneity in type 2 diabetes pathophysiology. Nature. 2024;627:347–57.
2. Aragam KG, et al. Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants. Nat Genet. 2022;54:1803–15.
3. Chen Y, et al. Genome-wide association meta-analysis identifies 17 loci associated with nonalcoholic fatty liver disease. Nat Genet. 2023;55:1640–50.
4. Schughart K, Williams RW, editors. Systems genetics: methods and protocols. 2018.
5. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. Science. 2008;322:881–8.
6. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. Genome Med. 2020;12:44.
7. Choi SW, Mak TSH, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. Nat Protoc. 2020;15:2759–72.
8. Wray NR, et al. From basic science to clinical application of polygenic risk scores: a primer. JAMA Psychiatry. 2021;78:101–9.
9. Hahn S-J, Kim S, Choi YS, Lee J, Kang J. Prediction of type 2 diabetes using genome-wide polygenic risk score and metabolic profiles: a machine learning analysis of population-based 10-year prospective cohort study. EBioMedicine. 2022;86:104383.
10. International Schizophrenia Consortium et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009;460:748–752.
11. Khera AV, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet. 2018. https://doi.org/10.1038/s41588-018-0183-z.
12. Visscher PM, et al. 10 years of GWAS discovery: biology, function, and translation. Am J Hum Genet. 2017;101:5–22.
13. Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw. 2014. https://doi.org/10.1016/j.neunet.2014.09.003.
14. Sarker IH. Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions. SN Comput Sci. 2021;2:420.
15 Mienye ID, Swart TG. A comprehensive review of deep learning: architectures, recent advances, and applications. Information (Basel). 2024;15:755.
16. van Hilten A, et al. GenNet framework: interpretable deep learning for predicting phenotypes from genetic data. Commun Biol. 2021;4:1094.
17. Li Y, et al. DeepGWAS: enhance GWAS signals for neuropsychiatric disorders via deep neural network. Res Sq. 2023. https://doi.org/10.21203/rs.3.rs-2399024/v1.
18. Sigala RE, et al. Machine learning to advance human genome-wide association studies. Genes. 2023;15:34.
19. Miao J, et al. Valid inference for machine learning-assisted genome-wide association studies. Nat Genet. 2024;56:2361–9.
20. Jumper J, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–9.
21. Lin Z, et al. Evolutionary-scale prediction of atomic level protein structure with a language model. Synthetic Biol. 2022;379(6637):1123–30.
22. Abramson J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024;630:493–500.
23. Kim C, et al. Transparent medical image AI via an image-text foundation model grounded in medical literature. Nat Med. 2024;30:1154–65.
24. Avsec Ž, et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods. 2021;18:1196–203.
25. Cui H, et al. ScGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nat Methods. 2024;21:1470–80.

26. Dalla-Torre H, et al. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. Nat Methods. 2025;22:287–97.
27. Elmes K, et al. SNVformer: An Attention-based Deep Neural Network for GWAS Data. Bioinformatics (2022).
28. Wu C, et al. A transformer-based genomic prediction method fused with knowledge-guided module. Brief Bioinform. 2023;25:bbad438.
29. Vaswani A, et al. Attention Is All You Need. arXiv e-prints arXiv:1706.03762 (2017).
30. Bahdanau, D., Cho, K. & Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv [cs.CL] (2014).
31. Kim Y, Denton C, Hoang L, Rush AM. Structured Attention Networks. arXiv [cs.CL] (2017).
32. Gene Ontology Consortium, et al. The Gene Ontology knowledgebase in 2023. Genetics. 2023;224:iyad031.
33. Azodi CB, Tang J, Shiu S-H. Opening the black box: interpretable machine learning for geneticists. Trends Genet. 2020;36:442–55.
34. Kamal MS, Dey N, Chowdhury L, Hasan SI, Santosh KC. Explainable AI for glaucoma prediction analysis to understand risk factors in treatment planning. IEEE Trans Instrum Meas. 2022;71:1–9.
35. Watson DS. Interpretable machine learning for genomics. Hum Genet. 2022;141:1499–513.
36. Qiu W, et al. Interpretable machine learning prediction of all-cause mortality. Commun Med. 2022;2:125.
37. Susnjak T, Griffin E. Towards clinical prediction with transparency: An explainable AI approach to survival modelling in residential aged care. bioRxiv (2024) https://doi.org/10.1101/2024.01.14.24301299.
38. Lundberg SM, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng. 2018;2:749–60.
39. Reaven G. Insulin resistance and coronary heart disease in nondiabetic individuals. Arterioscler Thromb Vasc Biol. 2012;32:1754–9.
40. Orgel E, Mittelman SD. The links between insulin resistance, diabetes, and cancer. Curr Diab Rep. 2013;13:213–22.
41. DeForest N, et al. Genome-wide discovery and integrative genomic characterization of insulin resistance loci using serum triglycerides to HDL-cholesterol ratio as a proxy. Nat Commun. 2024;15:8068.
42. Hindy G, et al. FNIP1 variants are associated with favourable metabolism in 1 million humans. Nature 1–10 (2026).
43. Oliveri A, et al. Comprehensive genetic study of the insulin resistance marker TG:HDL-C in the UK Biobank. Nat Genet. 2024;56:212–21.
44. Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of Bio-Medical Research, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science. 2007;316:1331–6.
45. Voight BF, et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat Genet. 2010;42:579–89.
46. Morris AP, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet. 2012;44:981–90.
47. Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press; 2021.
48. Sudlow C, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015;12:e1001779.
49. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369:1318–30.
50. Gazal S, et al. Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. Nat Genet. 2022;54:827–36.
51. Privé F, Vilhjálmsson BJ, Aschard H, Blum MGB. Making the most of clumping and thresholding for polygenic scores. Am J Hum Genet. 2019;105:1213–21.
52. Zou H, Hastie T. Addendum: regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol. 2005;67:768–768.
53. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. arXiv [cs.LG] (2016).
54. Mak TSH, Porsch RM, Choi SW, Zhou X, Sham PC. Polygenic scores via penalized regression on summary statistics. Genet Epidemiol. 2017;41:469–80.
55. Privé F, Arbel J, Vilhjálmsson BJ. LDpred2: better, faster, stronger. Bioinformatics. 2021;36:5424–31.
56. Ge T, Chen CY, Ni Y, Feng YCA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat Commun. 2019;10:1776.
57. Zheng Z, et al. Leveraging functional genomic annotations and genome coverage to improve polygenic prediction of complex traits within and between ancestries. Nat Genet. 2024;56:767–77.
58. All of Us Research Program Investigators et al. The 'All of Us' Research Program. N. Engl. J. Med. 381, 668–676 (2019).
59. Munshi A, et al. Association of LPL gene variant and LDL, HDL, VLDL cholesterol and triglyceride levels with ischemic stroke and its subtypes. J Neurol Sci. 2012;318:51–4.

60. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol. 2015;11:e1004219.
61. Mahajan A, et al. Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. Nat Genet. 2022;54:560–72.
62. Kim H, Kulkarni RN. Epigenetics in β-cell adaptation and type 2 diabetes. Curr Opin Pharmacol. 2020;55:125–31.
63. Su C, et al. 3D chromatin maps of the human pancreas reveal lineage-specific regulatory architecture of T2D risk. Cell Metab. 2022;34:1394-1409.e4.
64. Ardestani A, et al. MST1 is a key regulator of beta cell apoptosis and dysfunction in diabetes. Nat Med. 2014;20:385–97.
65. Costes S, Bertrand G, Ravier MA. Mechanisms of beta-cell apoptosis in type 2 diabetes-prone situations and potential protection by GLP-1-based therapies. Int J Mol Sci. 2021;22:5303.
66. Scott LJ, et al. The genetic regulatory signature of type 2 diabetes in human skeletal muscle. Nat Commun. 2016;7:11764.
67. Parker S, Varshney A, Fusion tissue study, group AMP-T2D skeletal muscle working. 242-OR: Single nuclei skeletal muscle chromatin and gene expression profiling across 287 individuals reveals cell type-specific genetic regulatory architectures at T2D GWAS loci. Diabetes. 2021;70:242-OR.
68. Gonzalez-Quintela A, et al. Serum levels of immunoglobulins (IgG, IgA, IgM) in a general adult population and their relationship with alcohol consumption, smoking and common metabolic abnormalities: serum immunoglobulin levels in adults. Clin Exp Immunol. 2008;151:42–50.
69. Wang F, et al. Apolipoprotein A-IV: a protein intimately involved in metabolism. J Lipid Res. 2015;56:1403–18.
70. Dziembowski A, et al. Proteomic analysis identifies a new complex required for nuclear pre-mRNA retention and splicing. EMBO J. 2004;23:4847–56.
71. Karczewski KJ, et al. Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. Cell Genom. 2022;2:100168.
72. Dornbos P, et al. Evaluating human genetic support for hypothesized metabolic disease genes. Cell Metab. 2022;34:661–6.
73 Andraski AB, et al. The distinct metabolism between large and small HDL indicates unique origins of human apolipoprotein A4. JCI Insight. 2023;8:e162481.
74. de Grooth GJ, et al. A review of CETP and its relation to atherosclerosis. J Lipid Res. 2004;45:1967–74.
75. Purcell S, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.
76. Bloom JS, et al. Genetic interactions contribute less than additive effects to quantitative trait variation in yeast. Nat Commun. 2015;6:8712.
77. Speed D, Evans DM. Estimating disease heritability from complex pedigrees allowing for ascertainment and covariates. Am J Hum Genet. 2024;111:680–90.
78. Yun S, Jeong M, Kim R, Kang J, Kim HJ. Graph transformer networks. Adv Neural Inf Process. Syst. 2019;32.
79. Shehzad A, et al. Graph transformers: a survey. IEEE Trans Neural Netw Learn Syst. 2024. https://doi.org/10.1109/tnnls.2025.3646122.
80. Moreno-Grau S, et al. Polygenic risk score portability for common diseases across genetically diverse populations. Hum Genomics. 2024;18:93.
81. Yang Q, et al. Serum retinol binding protein 4 contributes to insulin resistance in obesity and type 2 diabetes. Nature. 2005;436:356–62.
82. Graham TE, et al. Retinol-binding protein 4 and insulin resistance in lean, obese, and diabetic subjects. N Engl J Med. 2006;354:2552–63.
83. Winer DA, et al. B cells promote insulin resistance through modulation of T cells and production of pathogenic IgG antibodies. Nat Med. 2011;17:610–7.
84. Graham MJ, et al. Cardiovascular and metabolic effects of ANGPTL3 antisense oligonucleotides. N Engl J Med. 2017;377:222–32.
85. DeForest N, et al. Human gain-of-function variants in HNF1A confer protection from diabetes but independently increase hepatic secretion of atherogenic lipoproteins. Cell Genom. 2023;3:100339.
86. Nagasawa M, et al. Identification of a functional peroxisome proliferator-activated receptor (PPAR) response element (PPRE) in the human apolipoprotein A-IV gene. Biochem Pharmacol. 2009;78:523–30.
87. Yang L, et al. Transforming growth factor beta signaling in hepatocytes participates in steatohepatitis through regulation of cell death and lipid metabolism in mice: Yang et al. Hepatology. 2014;59:483–95.
88. Chen Y-G, Liu F, Massagué J. Mechanism of TGFβ receptor inhibition by FKBP12. EMBO J. 1997;16:3866–76.
89. Gillespie M, et al. The reactome pathway knowledgebase 2022. Nucleic Acids Res. 2022;50:D687-92.
90. Thomas PD, et al. Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems. Nat Genet. 2019;51:1429–33.
91. Qin Y, et al. A multi-scale map of cell structure fusing protein images and interactions. Nature. 2021;600:536–42.
92 Zheng F, et al. Interpretation of cancer mutations using a multiscale map of protein systems. Science. 2021;374:eabf3067.
93 Schaffer LV, et al. Multimodal cell maps as a foundation for structural and functional genomics. Nature. 2025;642:1–10.

94. Tan X, Benedict C. Increased risk of myocardial infarction among patients with type 2 diabetes who carry the common rs10830963 variant in the MTNR1B gene. Diabetes Care. 2020;43:2289–92.
95. Loh P-R, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet. 2015;47:284–90.
96. The All of Us Research Program Genomics Investigators. Genomic data in the all of us research program. Nature. 2024;627:340–6. https://doi.org/10.1038/s41586-023-06957-x.
97. Lambert SA, et al. Enhancing the polygenic score catalog with tools for score calculation and ancestry normalization. Nat Genet. 2024;56:1989–94. https://doi.org/10.1038/s41588-024-01937-x.
98. Homo sapiens genome assembly GRCh37. NCBI https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.13/.
99. Yu MK, et al. DDOT: a Swiss army knife for investigating data-driven biological ontologies. Cell Syst. 2019;8:267-273.e3.
100. Veličković, P. et al. Graph Attention Networks. arXiv [stat.ML] (2017).
101. Ye, T. et al. Differential Transformer. arXiv [cs.CL] (2024).
102. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. arXiv [cs.LG] (2017) https://doi.org/10.48550/ARXIV.1711.05101.
103. Phillips PC. Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet. 2008;9:855–67.
104. Niel C, Sinoquet C, Dina C, Rocheleau G. A survey about methods dedicated to epistasis detection. Front Genet. 2015;6:285.
105. Wu MC, et al. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011;89:82–93.
106. Behrouzi P, Arends D, Wit EC. Netgwas: an R package for network-based genome-wide association studies. R J. 2017. https://doi.org/10.32614/rj-2023-011.
107. Meyer HV, Birney E. PhenotypeSimulator: A comprehensive framework for simulating multi-trait, multi-locus genotype to phenotype relationships. Bioinformatics. 2018;34:2951–6.
108. Ning C, et al. A rapid epistatic mixed-model association analysis by linear retransformations of genomic estimated values. Bioinformatics. 2018;34:1817–25.
109. Gazal S. Combining SNP-to-gene linking strategies to pinpoint disease genes and assess disease omnigenicity. Preprint at https://doi.org/10.5281/ZENODO.7754032 (2021).
110. Carbon S, Mungall C. Gene Ontology Data Archive. Zenodo https://doi.org/10.5281/ZENODO.8200914 (2023).
111. Lee I, Wallace Z, Wang Y, Park S, Nam H, Majithia RA, Ideker T. G2PT: A Genotype-Phenotype Transformer to Assess and Explain Polygenic Risk. (Github).
112. Lee I, Wallace Z, Wang Y, Park S, Nam H, Majithia RA, Ideker T. G2PT: A Genotype-Phenotype Transformer to Assess and Explain Polygenic Risk. (Zenodo, 2026). 10.5281/ZENODO.19713271.
113. Lee I, Wallace Z, Wang Y, Park S, Nam H, Majithia RA, et al. Peer review of 'G2PT: a genotype-phenotype transformer to assess and explain polygenic risk.' Europ Mol Biol Organization. 2025. https://doi.org/10.1101/2024.10.23.619940.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Ingoo Lee[1] · Zachary Wallace[1,2] · Yuqi Wang[1] · Sungjoon Park[1] · Hojung Nam[3,4,5] · Amit R. Majithia[1] · Trey Ideker[1,3,6]

✉ Amit R. Majithia
amajithia@ucsd.edu

✉ Trey Ideker
tideker@health.ucsd.edu

[1] Department of Medicine, University of California San Diego, La Jolla, CA 92093, USA

[2] Bioinformatics and Systems Biology Program, University of California San Diego, La Jolla, CA 92093-0419, USA

[3] Department of Pharmacy, College of Pharmacy, Yonsei University, 85 Songdogwahak-Ro, Yeonsu-Gu, Incheon 21983, Republic of Korea

[4] Department of Integrative Biotechnology and Translational Medicine, Yonsei University, 85 Songdogwahak-Ro, Yeonsu-Gu, Incheon 21983, Republic of Korea

[5] Yonsei Institute of Pharmaceutical Sciences, and Yonsei-SL Bigen Research Institute, College of Pharmacy, Yonsei University, 85 Songdogwahak-Ro, Yeonsu-Gu, Incheon 21983, Republic of Korea

[6] Big Data Institute, University of Oxford, Oxford OX3 7LF, UK