A genotype-phenotype transformer to assess and explain polygenic risk

Ingoo Lee[1,2], Zachary S. Wallace[1,3], Yuqi Wang[6], Sungjoon Park[1], Hojung Nam[2,4,5], Amit R. Majithia[6,7†], Trey Ideker[1,3,8,9,†]

1. Division of Human Genomics and Precision Medicine, Department of Medicine, University of California at San Diego, La Jolla CA 92093
2. School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, 123 Cheomdangwagi-ro, Buk-gu, Gwangju 61005, Republic of Korea
3. Bioinformatics and Systems Biology Program, University of California at San Diego, La Jolla CA 92039
4. AI Graduate School, Gwangju Institute of Science and Technology, 123 Cheomdangwagi-ro, Buk-gu, Gwangju 61005, Republic of Korea
5. Center for AI-Applied High Efficiency Drug Discovery (AHEDD), Gwangju Institute of Science and Technology, 123 Cheomdangwagi-ro, Buk-gu, Gwangju 61005, Republic of Korea
6. Division of Endocrinology, Department of Medicine, University of California at San Diego, La Jolla CA 92093
7. Division of Biomedical Informatics, Department of Medicine, University of California at San Diego, La Jolla CA 92093
8. Department of Computer Science and Engineering, University of California at San Diego, La Jolla CA 92093
9. Department of Bioengineering, University of California at San Diego, La Jolla CA 92093

† Correspondence to: amajithia@ucsd.edu, tideker@health.ucsd.edu

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. Recently, Transformer models have emerged as a powerful machine learning architecture with potential to address these and other challenges. Accordingly, here we introduce the Genotype-to-Phenotype Transformer (G2PT), a framework for modeling hierarchical information flow among variants, genes, multigenic systems, and phenotypes. As proof-of-concept, we use G2PT to model the genetics of TG/HDL (triglycerides to high-density lipoprotein cholesterol), an indicator of metabolic health. G2PT predicts this trait via attention to 1,395 variants underlying at least 20 systems, including immune response and cholesterol transport, with accuracy exceeding state-of-the-art. It implicates 40 epistatic interactions, including epistasis between *APOA4* and *CETP* in phospholipid transfer, a target pathway for cholesterol modification. This work positions hierarchical graph transformers as a next-generation approach to polygenic risk.

Introduction

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].

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 prior 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 the triglyceride to high-density lipoprotein cholesterol ratio (TG/HDL), a primary marker of insulin resistance and associated risk factor for T2D, cardiovascular disease, fatty liver and certain cancers[39,40]. Hundreds of loci have been recently mapped for TG/HDL, encompassing genes operating in adipose, liver and muscle tissues[41,42], but the genetic circuits that regulate this trait and its molecular physiology are

still poorly understood. In what follows, application of G2PT yields a predictive genomic model that explains TG/HDL via a constellation of genetic factors, biological systems, and nonlinear epistatic interactions.

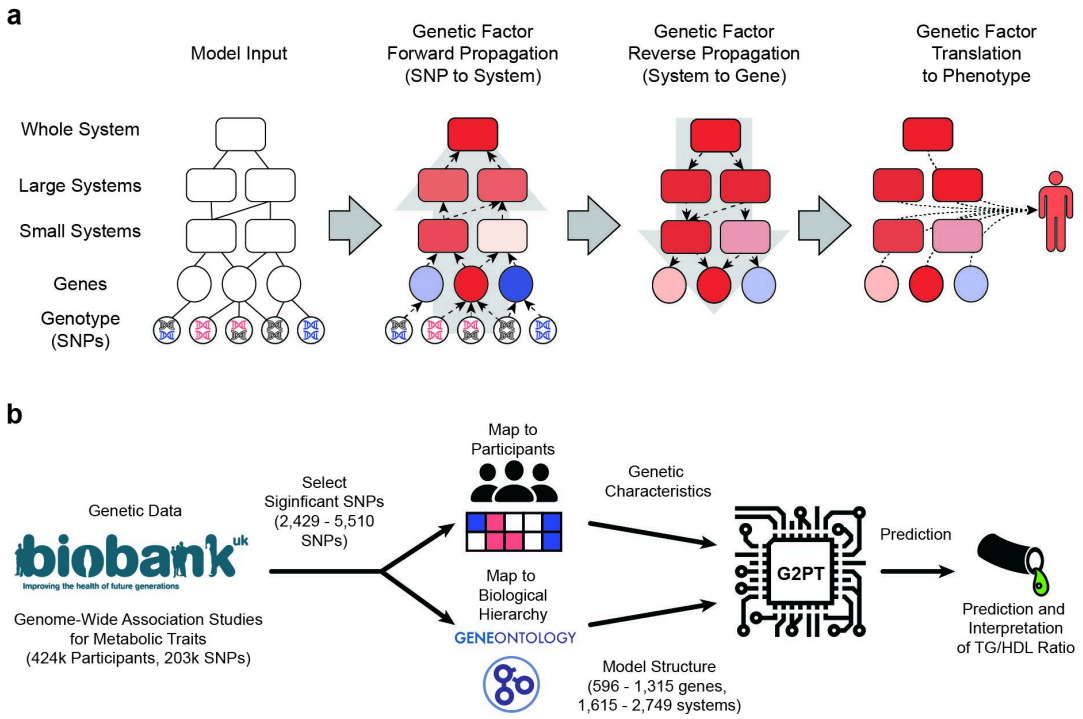

Fig. 1: G2PT Workflow and Application to Serum TG/HDL Ratio. a, Inputs of the Transformer model include genotypic data (SNPs, bottom layer), a mapping of SNPs to genes (second layer), and a mapping of genes into a hierarchy of multi-genic molecular systems (top layers). The presence of SNP minor alleles 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 triglyceride/HDL cholesterol ratio (TG/HDL). Genotypic data and corresponding metabolic traits are extracted for participants from the UK Biobank. SNPs are selected based on their independent association with TG/HDL ratio 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 the TG/HDL phenotype.

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[43]. 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 Insulin Resistance

As proof-of-concept, we used G2PT to study human metabolism, focusing on TG/HDL, a biomarker of insulin resistance (**Fig. 1b**). TG/HDL values were obtained from 423,888 participants profiled in the UK Biobank with accompanying genetic data (covering 203,126 SNPs; **Supplementary Fig. 1**)[42,44]. SNPs were mapped to genes using any of three lines of evidence: expression quantitative trait loci (eQTL)[45], curated mappings from the cS2G mapping tool[46], 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 TG/HDL. Training and evaluation were carried out in a robust framework of five-fold nested cross-validation[43] (**Methods**). Input features for G2PT were defined as SNPs that have an independent marginal association with TG/HDL phenotype, where significance of association was defined across a series of p-value thresholds of decreasing stringency (**Methods**). The resulting scope ranged from 2,429 SNPs, identified at a strict threshold of $p = 10^{-8}$, to 5,510 SNPs, identified at a permissive threshold of $p = 10^{-5}$ (**Fig. 2a**). Separate G2PT models were trained on each of these inputs. The entire training procedure required approximately 168 hours using 4 NVIDIA A30 Graphics Processing Units (GPUs, **Methods**).

Following model training, we assessed the performance of G2PT in predicting TG/HDL levels for held-out 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)[47]; regularized linear and nonlinear regression models (ElasticNet, XGBoost)[48,49], and polygenic scoring methods based on the entire genome–wide collection of SNPs (Lassosum, LDpred2)[50,51]. Across all ranges of SNP p-value thresholds (10^{-8} to 10^{-5}), G2PT achieved an explained variance (R^2) that was significantly higher than either PRS or regularized regression approaches (**Fig. 2a**). Its performance was also significantly higher than LDPRED2 or Lassosum, a notable result given that G2PT predictions were based on only thousands of SNPs versus the hundreds of thousands used by these competing methods (**Fig. 2b**).

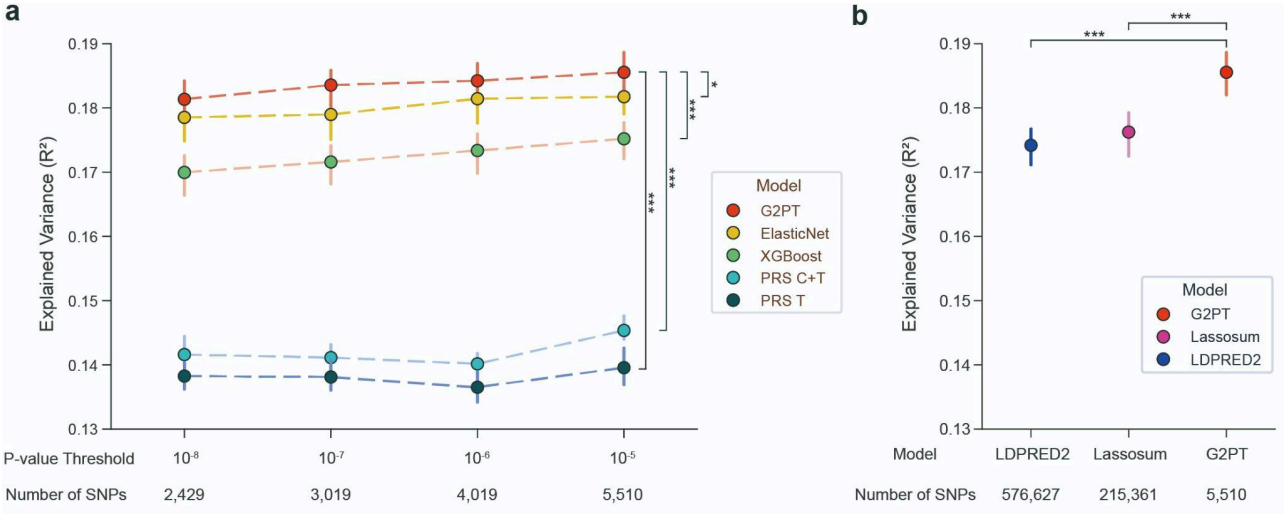

Fig. 2: Predictive Performance of G2PT compared with PRS Models. a, Explained variance (R^2) of G2PT compared against ElasticNet, XGBoost, PRS clumping + thresholding (PRS C+T), and PRS thresholding (PRS T), across varying marginal p-value thresholds for SNP selection and corresponding number of SNPs. **b**, Performance of G2PT relative to state-of-the-art Linkage Disequilibrium (LD)-aware PRS methods, Lassosum and LDPRED2. In both panels, points represent mean performance over nested five folds of cross-validation, with error bars showing 95% confidence intervals. *, significant differences in mean R^2 with $p < 0.05$, *** $p < 0.001$.

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 (**Supplementary Fig. 2**). Lastly, 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 (**Supplementary Fig. 2**).

Transformer Attention Reveals Genes and Mechanisms Underlying Phenotype

We next turned from phenotypic prediction to mechanistic interpretation – studying the model's transformer attention mechanism to reveal the key genes and systems it had used to predict TG/HDL (**Methods**). As a positive control, we examined the lipoprotein lipase (*LPL*) gene, a known insulin resistance gene associated with TG/HDL ratio[52] 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 (**Supplementary**

Fig. 3a). A similar relationship was observed for the molecular systems in which this gene is involved (e.g., lipid localization, **Supplementary Fig. 3b**).

Across all predictions for the 420K+ 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[42], whereas the remaining 81 had not (**Fig. 3a**). In general, however, model attention on systems was higher than on genes (**Supplementary Fig. 3c**) and not significantly correlated with system size (number of genes), depth in the systems hierarchy, or gene-level importance (**Supplementary Fig. 3d-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**).

G2PT also placed attention on unexpected systems, suggesting novel mechanisms regulating TG/HDL. In particular, immunoglobulin production, the 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. 3b**). While this pathway had not been implicated in prior GWAS of TG/HDL or related traits, it is concordant with clinical observations for individuals with genetic deficiencies in immunoglobulin-related genes, who collectively show significant alterations in serum lipids and insulin resistance[53]. Further investigation revealed that ten of the SNPs impacting this system had in fact been associated with TG/HDL previously[42], but independently from one another without any indication of a common biological process.

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[54] (**Fig. 3c**) 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[55] (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. 3c**; and conversely, deprioritizing *BUD13* as an effector gene for TG/HDL). We found that this particular genetic flow was corroborated by exome sequencing[56,57], 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. 3c**).

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.

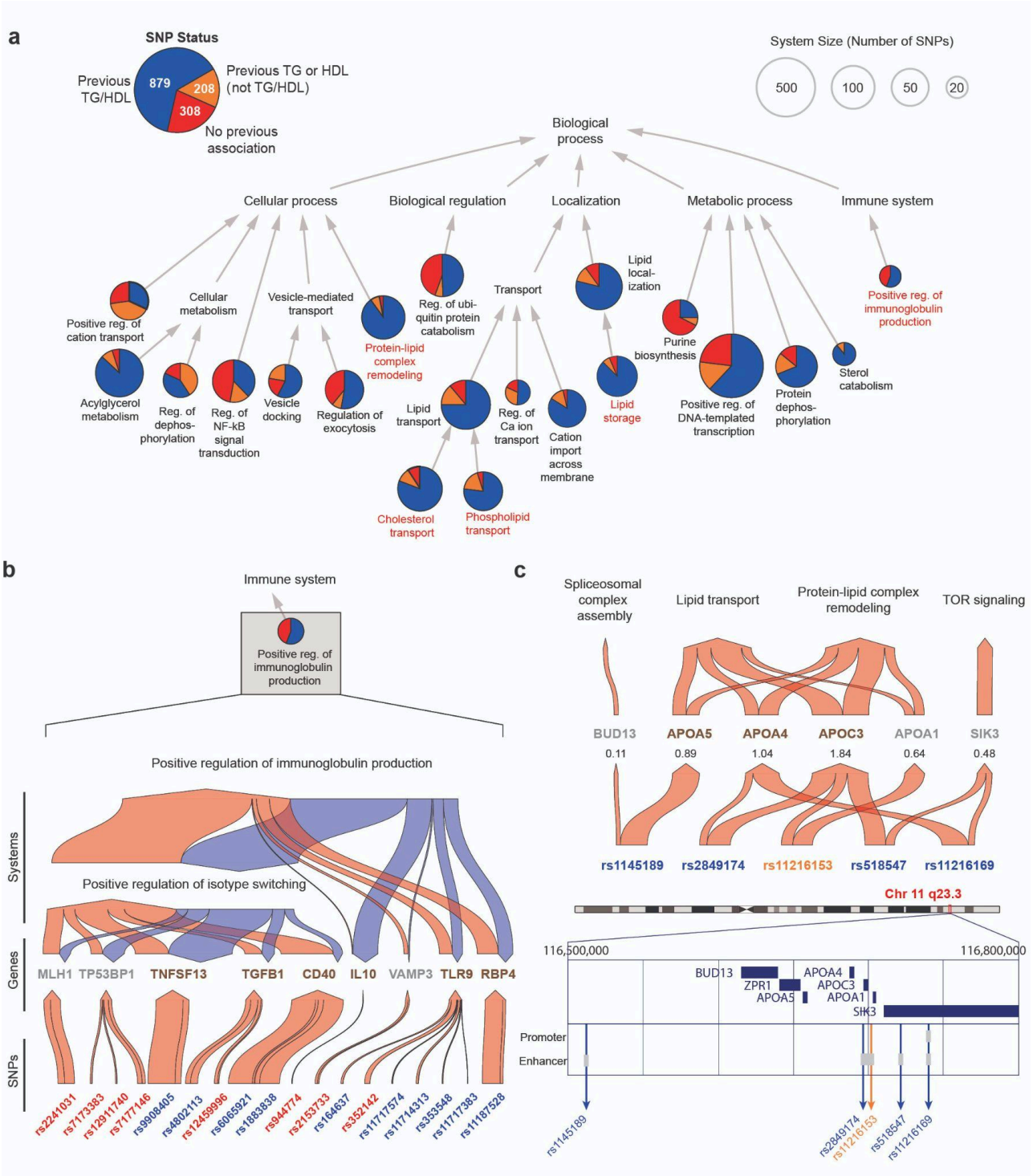

Fig. 3: **Multigenic Systems Determining Prediction of TG/HDL**. **a**, Hierarchy of important multigenic systems. Top 20 systems are represented as circular nodes, with size proportional to the number of genes annotated to the system. Arrows represent involvement in broader systems (Gene Ontology "is_a" 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**, 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. **c**, Genetic information flow at chromosomal locus 11q23.3. Top: genetic information flow visualized similarly to panel b, 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 – 166.8 MB expanded to detail gene open reading frames (dark blue rectangles) and regulatory regions (promoters or enhancers; light gray).

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 (**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 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. 5c**) 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, Supplementary Table 1**). 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[58,59], supporting a functional biochemical interaction.

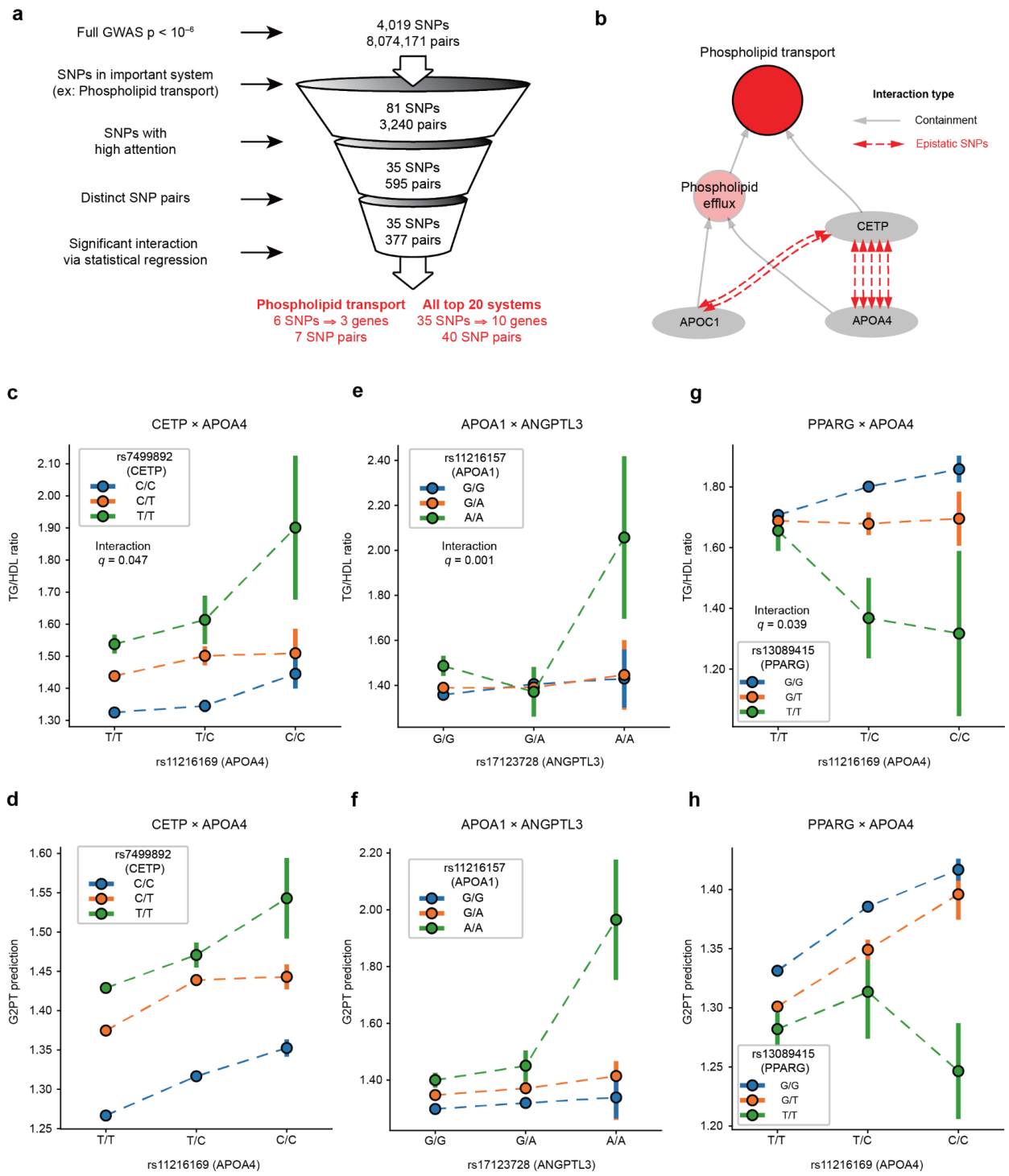

Fig. 4: Epistatic Interactions Identified by Attention-Based Epistasis Search. **a**, Pipeline for detecting epistasis. Interacting SNP pairs are identified through a progressive multi-step screening process based on the system under study (here, Phospholipid transport), model attention and validation 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 13089415 (APOA4) on (g) measured or (h) predicted TG/HDL ratio. Display as per panel (c).

Beyond phospholipid transport, we broadly applied this same epistasis screening procedure across all 20 systems most important to G2PT predictions (**Fig. 4a**). This broader screen identified 40 SNP-SNP pairs with significant epistasis (**Supplementary Table 1**). 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*, 1qp1.3) loci; here phenotypic impact was seen predominantly in the combination of homozygous minor states for both variants (**Fig. 4e,f**). 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**).

Discussion

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[60,61]. When applied to TG/HDL, a surrogate marker for insulin resistance, this architecture outperforms genome-wide PRS methods (**Fig. 2**) despite being constrained to a few thousand loci. Detailed examination of the model identifies a core set of biological systems driving phenotypic predictions which integrate signals from hundreds of genes, including expected and unexpected factors (**Fig. 3**). Further analysis of these systems identifies numerous epistatic interactions underlying regulation of TG/HDL, which we confirm are used by the model to make predictions (**Fig. 4**). The ability to learn nonlinear gene-by-gene interactions is of general interest, as it moves beyond the capabilities of current PRS models.

Interpretation of the G2PT model 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.

This model interpretation procedure revealed high attention on the unexpected system of 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 TG or HDL, others such as RBP4 are known causal biomarkers of insulin resistance in humans[62,63], but through distinct mechanisms from immunoglobulin production. Although few studies have directly linked insulin resistance to immunoglobulin production, 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[64]. These observations provide a potential molecular mechanism for our finding in humans.

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[65], 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*[66], 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*[67].

Despite the promising use of Transformers to approach genotype-phenotype questions, our study also points to some current limitations. First, computing of model attention is an expensive operation[29] with substantial training time and data required to reach convergence. In our study, G2PT was allocated 4 A30 GPUs over approximately 168 hours of training. Further scaling of this compute time and hardware could enable increases in the number of SNPs considered, faster convergence, larger model architectures or deeper hyperparameter explorations, with enhanced ability to capture complex genotype–phenotype relationships. We also experienced challenges arising from our use of the 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[68], GO Causal Activity Models (GO-CAMS)[69] or maps of biological structures and systems resolved directly from ‘omics data[70–72]. Regardless, G2PT has immediate application to the genetic analysis of diverse phenotypes of interest, including those related to multigenic diseases such as T2D, 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.

Materials and Methods

GWAS Data

Human genotype and phenotype data were obtained from the UK Biobank[44], in which participants of Caucasian ancestry had been genotyped with SNP arrays and characterized for TG and HDL levels in millimoles per liter (mmol/L). The UK Biobank study was approved by the Research Ethics Committee, and informed consent was obtained from all participants. Analysis of UK Biobank data was conducted under application numbers 51436 and 26041. 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 matrix. We performed a Bayesian logistic regression analysis using BOLT-LMM[73] to identify SNPs that statistically associate with $\log_2$(TG/HDL), while including sex, age, and the top 10 principal components as covariates. Significantly associated SNPs were then retained based on various p-value thresholds from 10^{-5} to 10^{-8} (**Fig. 2**).

SNP-to-Gene Mapping

Significantly associated SNPs were mapped to (potentially multiple) protein-coding genes using the union of (1) associations from cS2G[46], (2) eQTL associations from GTEx v7[45], or (3) the nearest gene in genomic coordinates (Genome Reference Consortium hg19 assembly[74]). For the eQTL SNP-to-gene mappings, we used adipose subcutaneous, adipose visceral omentum, liver, pancreas, adrenal gland, muscle skeletal, and uterus tissue types[42].

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[75] 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 graph contained three types of nodes, representing SNPs, genes and systems, and three types of directed edges, representing SNP→gene mappings, gene→system annotations, and subsystem→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: 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.

Hierarchical Graph Transformer

The hierarchical graph transformer (HiGT) is a modified version of a Graph Attention Transformer[29,76] that leverages a hierarchical knowledge graph (see above section 'Gene Ontology Knowledge Hierarchy'). For each node *i* with embedding state E_i, the HiGT function transforms this embedding by incorporating effects from graph neighbors *j* (Eqn. 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 (Eqn. 2) scaled by an attention weight α (Eqn. 3). In particular:

$$\text{HiGT}(E_i, \{E_j \mid j \in neighbors_i\}) = \text{FFNN}\left(\text{LN}\left(\sum_j \text{MultiHeadAttn}(E_i, E_j)W^O\right)\right) \quad (1)$$

$$\text{MultiHeadAttn}(E_i, E_j) = \text{Concat}(Attn_1(E_i, E_j), \cdots, Attn_h(E_i, E_j), \cdots, Attn_n(E_i, E_j)) \quad (2)$$

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

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

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 the embedding was set to d=64. The number of heads was set to n=4. This HiGT function is 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 (eqn. 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:

$$E_i = E_i + \sum_z \text{HiGT}_z(E_i, \{E_j | j \in \text{source}_i\}) \; \forall i \in \text{genes}, \; \forall j \in \text{SNPs}_z \quad (5)$$

Different HiGT modules are applied based on the zygosity of the SNP alteration, z $\in$ {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:

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

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

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

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*:

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

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

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

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 function 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} = (sex, age, PC1, …, PC10) onto an embedding of size *d* using a Multi-Layer Perceptron, yielding a participant embedding P^{embed}.

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

where m = 12 is 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 (eqn. 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[77] 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 $\hat{Y}$, a prediction of phenotype Y, the $\log_2$(TG/HDL) phenotype. Altogether, this layer can be formulated as:

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

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

Model Training and Comparative Evaluation

Nested cross-validation was used to train and robustly evaluate model performance in genotype-phenotype prediction. In each fold, data were split into training, validation, and test sets in a 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 (via grid search), 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 TG/HDL values, optimized with decoupled weight decay regularization (AdamW and L2 regularization)[78]. Hyperparameters were optimized through a grid search, including for G2PT optimal p-value threshold and training epoch; 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.

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 TG/HDL ratios, 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 '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. Second, 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 vs. bottom 10% of individuals ranked by system attention). Third, SNP pairs (*SNP1*, *SNP2*) were tested for a statistical interaction in prediction of the phenotype according to a combinatorial linear model[79,80]:

$$y = \alpha_1 \cdot SNP1 + \alpha_2 \cdot SNP2 + \alpha_{12} \cdot SNP1 \times SNP2 + \beta \cdot \text{sex} + \gamma \cdot \text{age} \quad (11)$$

where y is the $\log_2$(TG/HDL) phenotype. 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 ($\alpha_{12} \neq 0$) 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.

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. For this study, functional genomic annotations used for SNP-to-gene mapping were obtained in November 2023 from https://alkesgroup.broadinstitute.org/cS2G. Additionally, eQTL data were sourced from GTEx v7, accessible at https://www.gtexportal.org/home/downloads/adult-gtex/qtl. GO data were downloaded from https://release.geneontology.org/2023-07-27/index.html.

Code Availability

The source code is available at https://github.com/idekerlab/G2PT.

Acknowledgements

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).

Competing Interest Declaration

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.

References

1. Suzuki, K. *et al.* Genetic drivers of heterogeneity in type 2 diabetes pathophysiology. *Nature* **627**, 347–357 (2024).
2. Aragam, K. G. *et al.* Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants. *Nat. Genet.* **54**, 1803–1815 (2022).
3. Chen, Y. *et al.* Genome-wide association meta-analysis identifies 17 loci associated with nonalcoholic fatty liver disease. *Nat. Genet.* **55**, 1640–1650 (2023).
4. Schughart, K. & Williams, R. W. *Systems Genetics: Methods and Protocols*. (Methods in Molecular Biology, 2018).
5. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322**, 881–888 (2008).
6. Lewis, C. M. & Vassos, E. Polygenic risk scores: from research tools to clinical instruments. *Genome Med.* **12**, 44 (2020).
7. Choi, S. W., Mak, T. S.-H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.* **15**, 2759–2772 (2020).
8. Wray, N. R. *et al.* From Basic Science to Clinical Application of Polygenic Risk Scores: A Primer. *JAMA Psychiatry* **78**, 101–109 (2021).
9. Hahn, S.-J., Kim, S., Choi, Y. S., 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* **86**, 104383 (2022).
10. International Schizophrenia Consortium *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
11. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* doi:10.1038/s41588-018-0183-z.
12. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
13. Schmidhuber, J. Deep learning in neural networks: An overview. *arXiv [cs.NE]* (2014).
14. Sarker, I. H. Deep learning: A comprehensive overview on techniques, taxonomy, applications and

research directions. *SN Comput. Sci.* **2**, 420 (2021).

15. Mienye, I. D. & Swart, T. G. A comprehensive review of deep learning: Architectures, recent advances, and applications. *Information (Basel)* **15**, 755 (2024).
16. van Hilten, A. *et al.* GenNet framework: interpretable deep learning for predicting phenotypes from genetic data. *Commun Biol* **4**, 1094 (2021).
17. Li, Y. *et al.* DeepGWAS: Enhance GWAS Signals for Neuropsychiatric Disorders via Deep Neural Network. *Res Sq* (2023) doi:10.21203/rs.3.rs-2399024/v1.
18. Sigala, R. E. *et al.* Machine Learning to Advance Human Genome-Wide Association Studies. *Genes* **15**, (2023).
19. Miao, J. *et al.* Valid inference for machine learning-assisted genome-wide association studies. *Nat. Genet.* **56**, 2361–2369 (2024).
20. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
21. Lin, Z. *et al.* Evolutionary-scale prediction of atomic level protein structure with a language model. *Synthetic Biology* (2022).
22. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
23. Kim, C. *et al.* Transparent medical image AI via an image-text foundation model grounded in medical literature. *Nat Med* **30**, 1154–1165 (2024).
24. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
25. Cui, H. *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* **21**, 1470–1480 (2024).
26. Dalla-Torre, H. *et al.* Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nat. Methods* **22**, 287–297 (2025).
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.* **25**, bbad438 (2023).

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, A. M. Structured Attention Networks. *arXiv [cs.CL]* (2017).

32. Gene Ontology Consortium *et al.* The Gene Ontology knowledgebase in 2023. *Genetics* **224**, iyad031 (2023).

33. Azodi, C. B., Tang, J. & Shiu, S.-H. Opening the Black Box: Interpretable Machine Learning for Geneticists. *Trends Genet.* **36**, 442–455 (2020).

34. Kamal, M. S., Dey, N., Chowdhury, L., Hasan, S. I. & Santosh, K. C. Explainable AI for Glaucoma Prediction Analysis to Understand Risk Factors in Treatment Planning. *IEEE Trans. Instrum. Meas.* **71**, 1–9 (2022).

35. Watson, D. S. Interpretable machine learning for genomics. *Hum. Genet.* **141**, 1499–1513 (2022).

36. Qiu, W. *et al.* Interpretable machine learning prediction of all-cause mortality. *Commun. Med.* **2**, 125 (2022).

37. Susnjak, T. & Griffin, E. Towards clinical prediction with transparency: An explainable AI approach to survival modelling in residential aged care. *bioRxiv* (2024) doi:10.1101/2024.01.14.24301299.

38. Lundberg, S. M. *et al.* Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng* **2**, 749–760 (2018).

39. Reaven, G. Insulin resistance and coronary heart disease in nondiabetic individuals. *Arterioscler. Thromb. Vasc. Biol.* **32**, 1754–1759 (2012).

40. Orgel, E. & Mittelman, S. D. The links between insulin resistance, diabetes, and cancer. *Curr. Diab. Rep.* **13**, 213–222 (2013).

41. Oliveri, A. *et al.* Comprehensive genetic study of the insulin resistance marker TG:HDL-C in the UK Biobank. *Nat. Genet.* **56**, 212–221 (2024).

42. 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.* **15**, 8068 (2024).

43. Goodfellow, I., Bengio, Y. & Courville, A. Deep Learning. *MIT Press* https://mitpress.mit.edu/9780262035613/deep-learning/ (2021).

44. 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.* **12**, e1001779 (2015).

45. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

46. Gazal, S. *et al.* Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. *Nat. Genet.* **54**, 827–836 (2022).

47. Privé, F., Vilhjálmsson, B. J., Aschard, H. & Blum, M. G. B. Making the most of clumping and thresholding for polygenic scores. *Am. J. Hum. Genet.* **105**, 1213–1221 (2019).

48. Zou, H. & Hastie, T. Addendum: Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.* **67**, 768–768 (2005).

49. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *arXiv [cs.LG]* (2016).

50. Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X. & Sham, P. C. Polygenic scores via penalized regression on summary statistics. *Genet. Epidemiol.* **41**, 469–480 (2017).

51. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger. *Bioinformatics* **36**, 5424–5431 (2021).

52. 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.* **318**, 51–54 (2012).

53. 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.* **151**, 42–50 (2008).

54. Wang, F. *et al.* Apolipoprotein A-IV: a protein intimately involved in metabolism. *J. Lipid Res.* **56**, 1403–1418 (2015).

55. Dziembowski, A. *et al.* Proteomic analysis identifies a new complex required for nuclear pre-mRNA retention and splicing. *EMBO J.* **23**, 4847–4856 (2004).

56. Karczewski, K. J. *et al.* Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. *Cell Genom.* **2**, 100168 (2022).

57. Dornbos, P. *et al.* Evaluating human genetic support for hypothesized metabolic disease genes. *Cell Metab* **34**, 661–666 (2022).

58. Andraski, A. B. *et al.* The distinct metabolism between large and small HDL indicates unique origins of

human apolipoprotein A4. *JCI Insight* **8**, (2023).

59. de Grooth, G. J. *et al.* A review of CETP and its relation to atherosclerosis. *J. Lipid Res.* **45**, 1967–1974 (2004).
60. Yun, S., Jeong, M., Kim, R., Kang, J. & Kim, H. J. Graph transformer networks. *Adv. Neural Inf. Process. Syst.* **32**, (2019).
61. Shehzad, A. *et al.* Graph Transformers: A Survey. *arXiv [cs.LG]* (2024).
62. Yang, Q. *et al.* Serum retinol binding protein 4 contributes to insulin resistance in obesity and type 2 diabetes. *Nature* **436**, 356–362 (2005).
63. Graham, T. E. *et al.* Retinol-binding protein 4 and insulin resistance in lean, obese, and diabetic subjects. *N. Engl. J. Med.* **354**, 2552–2563 (2006).
64. Winer, D. A. *et al.* B cells promote insulin resistance through modulation of T cells and production of pathogenic IgG antibodies. *Nat. Med.* **17**, 610–617 (2011).
65. Graham, M. J. *et al.* Cardiovascular and metabolic effects of ANGPTL3 antisense oligonucleotides. *N. Engl. J. Med.* **377**, 222–232 (2017).
66. 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* **3**, 100339 (2023).
67. 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.* **78**, 523–530 (2009).
68. Gillespie, M. *et al.* The reactome pathway knowledgebase 2022. *Nucleic Acids Res.* **50**, D687–D692 (2022).
69. Thomas, P. D. *et al.* Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems. *Nat. Genet.* **51**, 1429–1433 (2019).
70. Qin, Y. *et al.* A multi-scale map of cell structure fusing protein images and interactions. *Nature* **600**, 536–542 (2021).
71. Zheng, F. *et al.* Interpretation of cancer mutations using a multiscale map of protein systems. *Science* **374**, eabf3067 (2021).
72. Schaffer, L. V. *et al.* Multimodal cell maps as a foundation for structural and functional genomics.

Nature 1–10 (2025).

73. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
74. Homo sapiens genome assembly GRCh37. *NCBI* https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.13/.
75. Yu, M. K. *et al.* DDOT: A Swiss Army Knife for Investigating Data-Driven Biological Ontologies. *Cell Syst* **8**, 267–273.e3 (2019).
76. Veličković, P. *et al.* Graph Attention Networks. *arXiv [stat.ML]* (2017).
77. Ye, T. *et al.* Differential Transformer. *arXiv [cs.CL]* (2024).
78. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. *arXiv [cs.LG]* (2017) doi:10.48550/ARXIV.1711.05101.
79. Phillips, P. C. Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.* **9**, 855–867 (2008).
80. Niel, C., Sinoquet, C., Dina, C. & Rocheleau, G. A survey about methods dedicated to epistasis detection. *Front. Genet.* **6**, 285 (2015).