Mechanistic genotype-phenotype translation using hierarchical transformers

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

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 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. Department of Computer Science and Engineering, University of California at San Diego, La Jolla CA 92093
8. Department of Bioengineering, University of California at San Diego, La Jolla CA 92093

† Correspondence to: 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 a lack of mechanistic understanding and widespread epistatic interactions. Recently, Transformer models have emerged as a powerful general-purpose architecture in machine learning, 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 functions, 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 learns to predict this trait via high attention to genetic variants underlying 24 functions, including immune response and cholesterol transport, with accuracy exceeding state-of-the-art. It implicates unexpected epistatic interactions, including those among APOC1 and CETP. This work positions Hierarchical Transformers as a general approach to functionally interpret polygenic risk. The source code is available at https://github.com/idekerlab/G2PT.

1 Introduction

Common diseases such as type 2 diabetes, cardiovascular disorders, and fatty liver are highly polygenic and physiologically heterogeneous, involving complex networks of interactions within and among multigenic molecular functions [1–4]. In these complex diseases, examination of single genetic variants, or even single genes, has had limited utility. Rather, progress has been made systematically identifying associated genetic variants through genome-wide association studies (GWAS), and then combining these SNP-phenotype associations using methodologies collectively termed Polygenic Risk Scores (PRS) [5–8]. PRS methods have been applied to predict phenotypes in a wide range of common multigenic diseases [6, 9–11] but have two major limitations: 1) they model loci additively and thus miss functional dependencies among (genetic epistasis), and 2) they emit a risk estimate for an individual but cannot distinguish the perturbed molecular and physiological pathways underlying that persons set of risk driving SNPs which if understood could be used to gain biological understanding and guide treatment [12].

Recent developments in deep learning [13–23] offer significant opportunities to advance the current framework, as they have the capacity to model both complex epistatic interactions and knowledge of molecular mechanisms. In particular, the Transformer [24] has emerged as a central state-of-the-art modeling approach in diverse fields including natural language understanding, generation of realistic text, photographs and video, and automated software programming [25–27]. It has also shown strong potential to address longstanding problems in the biomedical sciences [28–32], including prediction of 3D protein structures or expression patterns from primary genome sequence data [20]. The Transformer architecture is known for its central use of an “attention mechanism” [24], which allows a neural network to increase the weights of certain relationships and downgrade others depending on context, similar to how humans selectively focus on certain aspects of a sentence or scene depending on the task at hand [24, 33, 34]. Although it can depend on the modeling task, analyzing where a model directs its attention has potential utility in understanding the rules and logic underlying its predictions, a process known as model interpretation [24, 35–37]. The ability to clearly interpret a model is also important for promoting transparency, trust, and fair decision-making, all of which are critical in clinical practice [38–42].

Here we describe the Genotype-to-Phenotype Transformer (G2PT), a hierarchical 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 computing attention across embedded representations of genes and a hierarchy of multigenic functions. As proof-of-concept, we apply G2PT to reveal a constellation of genetic factors that govern TriGlyceride / HDL-cholesterol ratio (TG/HDL), a central readout of metabolic function and risk for diabetes and cardiovascular disorders [43–46] (**Fig. 1b**). The result is a predictive genomic model that offers mechanistic interpretability and facilitates the discovery of numerous epistatic interactions among genes and regulatory regions.

2 Results

2.1 G2PT Model Overview.

The G2PT framework models the states of biological entities, including variants, genes, multigenic functions, 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, 48]. Positions in the embedding (i.e. the states of each entity) are governed by a Hierarchical Transformer (HiTR), a deep neural network that models bidirectional flow of information across the hierarchy of entities. Such information flow includes the effects of variants on the states of genes (SNP-gene interactions), the effects of altered genes on multigenic functions and superfunctions (gene-function and function-function interactions), and the reciprocal influences of functions on the states of their component functions and genes (reverse interactions, **Fig. 1a**, **Methods**). Based on the collection of variants comprising an individual’s genotype, the HiTR 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 functions is used to predict phenotype.

2.2 Using G2PT to Model a Metabolic Phenotype.

As proof-of-concept, we used G2PT to study human metabolism, focusing on the TG/HDL ratio as a model metabolic phenotype (**Fig. 1b**). Human subjects, along with their corresponding genotypic variants (covering 203,126 Single Nucleotide Polymorphisms, or SNPs) and matching TG/HDL values (**Supplementary Fig. 1**), were obtained from a cohort of 423,888 participants profiled in the UK Biobank [49, 50]. SNPs were mapped to their nearest gene (hg37 reference genome coordinates, **Methods**), and genes were mapped to a hierarchy of multigenic functions condensed from Biological Process terms recorded in the Gene Ontology (**Fig. 1b**) [51].

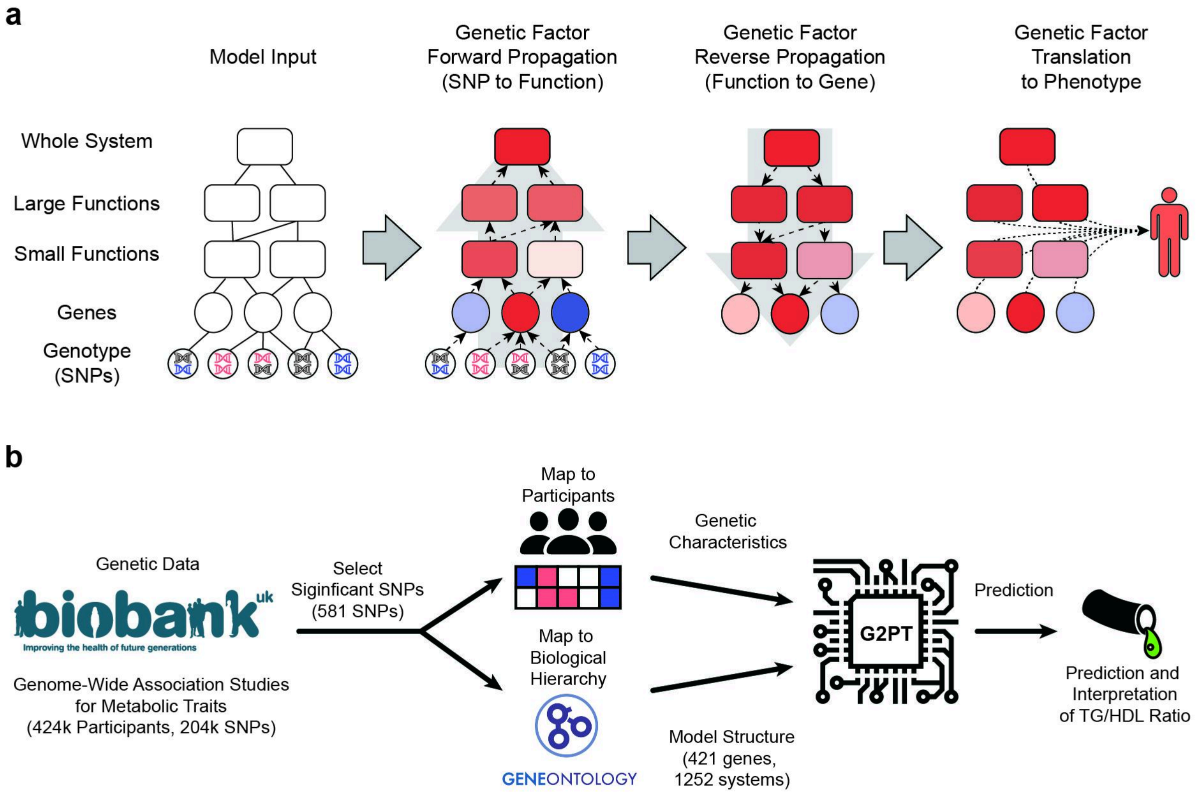
a
Model Input
Genetic Factor Forward Propagation (SNP to Function)
Genetic Factor Reverse Propagation (Function to Gene)
Genetic Factor Translation to Phenotype
Whole System
Large Functions
Small Functions
Genes
Genotype (SNPs)
b
Genetic Data
biobank uk
Improving the health of future generations
Genome-Wide Association Studies for Metabolic Traits (424k Participants, 204k SNPs)
Select Siginficant SNPs (581 SNPs)
Map to Participants
Map to Biological Hierarchy
GENEONTOLOGY
Genetic Characteristics
Model Structure (421 genes, 1252 systems)
G2PT
Prediction
Prediction and Interpretation of TG/HDL Ratio

Fig. 1: G2PT Workflow and Application to TG/HDL Ratio. a, G2PT workflow. 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 functions (top layers). The presence of SNP minor alleles modifies the embedding states of downstream genes and multigenic functions (forward propagation). Conversely, state changes in functions influence the states of sub-functions and genes they contain (reverse propagation). Finally, all gene and function states are integrated to predict phenotype. **b,** Proof of concept via prediction of triglyceride/high-density-lipoprotein ratio (TG/HDL). Genotypic data and corresponding metabolic traits for over 430,000 participants are extracted 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 multi-genic functions defined by Gene Ontology terms. This information is used by G2PT to predict the TG/HDL phenotype.

Using this information, G2PT models were trained to translate an individual's pattern of SNPs to a prediction of TG/HDL. Training and evaluation were carried out in the robust framework of five-fold nested cross-validation [52, 53], in which the population is divided into separate training, validation, and test samples using 60/20/20 splits (**Methods**). As for current PRS models [54], 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 75 SNPs, identified at a starting threshold of $p = 10^{-8}$, to 2088 SNPs, identified at a relaxed threshold of $p = 10^{-2}$ (**Fig. 2**). Separate G2PT models were trained on each of these inputs. The entire training procedure required approximately 48 hours using 4 NVIDIA A30 Graphics Processing Units (GPUs, **Methods**).

Following training, we assessed the performance of G2PT models in predicting TG/HDL levels for held-out individuals in the test set (samples not included in any aspect of model training or hyperparameter tuning). For individuals with a high TG/HDL ratio (defined as TG/HDL ≥ 2.3 following standard guidelines [55]), G2PT could categorize these individuals at an odds ratio of 3.10, relative to current PRS methodology which yielded an odds ratio of 1.95 (**Fig. 2a**). We found that G2PT achieved its highest predictive accuracy when considering SNPs selected at a p-value threshold of 10^{-5} ($R^2 = 0.144$, **Fig. 2b**). We also benchmarked this performance against several alternative machine learning models, including XGBoost [56] and the ElasticNet architecture (**Methods**), finding that the performance of G2PT outcompeted both of these alternatives (**Fig. 2**, $p < 10^{-4}$ by paired t-tests).

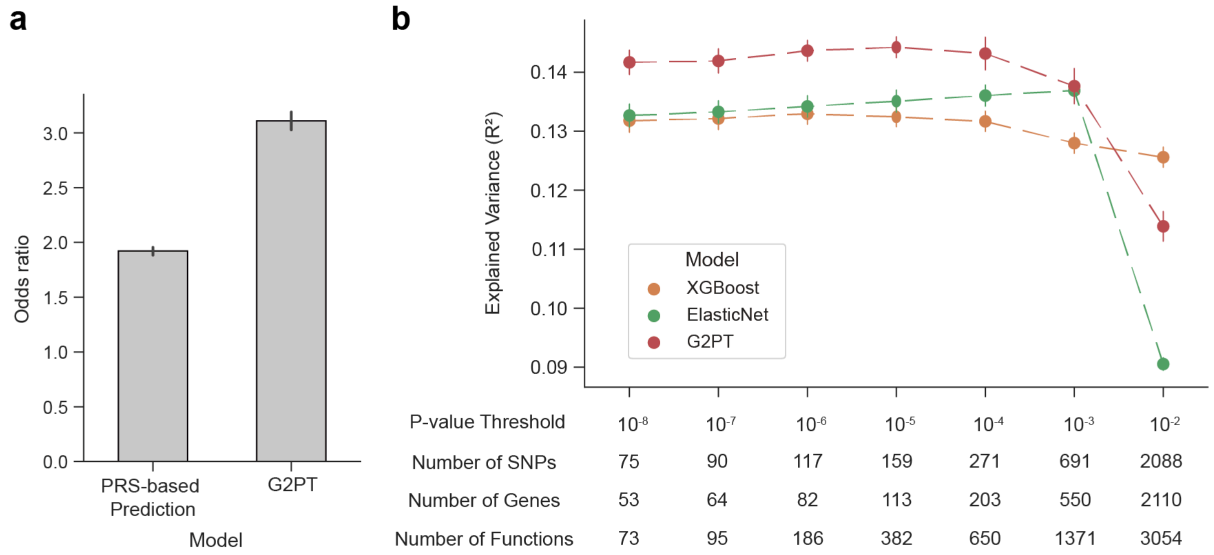

Fig. 2: Analyzing Predictive Performance of G2PT Against Baselines. **a**, Odds ratio performance in categorizing individuals with high TG/HDL (see text). The performance is shown for a current PRS approach (BOLT-LMM) and G2PT. **b**, Predictive performance of G2PT against other machine learning models (XGBoost and ElasticNet) as measured by explained variance (R^2) and scanned across an expanding range of P-value thresholds for selection of significant SNPs. The number of SNPs, genes, and functions at each threshold is provided. Both panels: bars/points represent mean performance and error bars represent 95% confidence intervals over five folds of cross validation.

2.3 Transformer Attention Reveals Biological Mechanisms Underlying Phenotype.

We next studied the model's attention to genes and functions in predicting the TG/HDL phenotype (**Methods**). As a positive control, we examined the Apolipoprotein A-V (APOA5) gene, a known regulator of TG/HDL [57] which we expected should be given high attention by the G2PT framework. We indeed found a substantial number of subjects for which G2PT gave high attention to APOA5, and that these subjects tended to have significantly higher predicted TG/HDL levels (**Supplementary Fig. 2a**). Further inspection revealed a predictive SNP rs45515495 located upstream of this gene with a minor allele frequency (MAF) of 6%. Similar high attention was observed for the hierarchy of molecular functions in which this gene is involved (e.g., Sterol homeostasis, **Supplementary Fig. 2b**).

Altogether across all predictions for the 420K+ individuals, we identified significant model attention on 24 multigenic functions (importance score > 0.25, **Methods**) covering a total of 191 SNPs linked to 123 genes. These functions spanned a variety of biological processes, including HDL particle remodeling, lipid localization, and lipopolysaccharide signaling, all of which had been previously linked to TG/HDL ratio [58, 59]. We also observed unexpected processes pertaining to macrophage and leukocyte differentiation (**Fig. 3a-b**). As for the 123 genes, 89 of these had been previously associated with TG/HDL, TG, or HDL-cholesterol in the GWAS catalog [60], whereas the remaining 34 gene associations had not been previously reported (**Fig. 3c**, **Supplementary Table 1**).

We compared the G2PT importance scores with the functional enrichment scores provided by MAGMA [61] for the same dataset. We observed moderate agreement between these results (Pearson $r = 0.30$), particularly in enrichment of functions related to lipid and lipoprotein metabolism (e.g. high-density particle remodeling and reverse cholesterol transport, **Supplementary Fig. 3**). On the other hand, we observed a number of functions which G2PT had scored strongly but MAGMA had not, including immune regulatory and hematopoietic functions (**Supplementary Fig. 3**).

2.4 Genetic Variants Underlying Leukocyte Differentiation Impact TG/HDL.

We next inspected leukocyte differentiation, an unexpected function identified by G2PT in prediction of TG/HDL phenotype (**Fig. 4a,b**). Within this function we noted the Tribbles homolog 1 (TRIB1) gene, which harbored multiple SNPs with TG/HDL association, including one with an extremely high level of significance (rs2954021, $p < 10^{-200}$). This gene had been previously linked to lipid metabolism but not leukocyte differentiation [62–65]. A second gene assigned to leukocyte differentiation, protein kinase C alpha (PRKCA), was given high G2PT attention based on a single SNP (rs117524772, p = 3.5x10^{-7}, **Fig. 4b**) which was only moderately associated with TG/HDL (i.e. failing a strict genome-wide significance threshold of $p < 10^{-8}$). Beyond these genes, leukocyte differentiation was impacted by informative variants from five other gene loci, such as the Retinoic Acid Receptor Alpha (RARA) and H2.0-Like Homeobox (HLX) genes, both of which are well known drivers of leukocyte differentiation that had not been identified in previous GWAS of TG/HDL or related phenotypes.

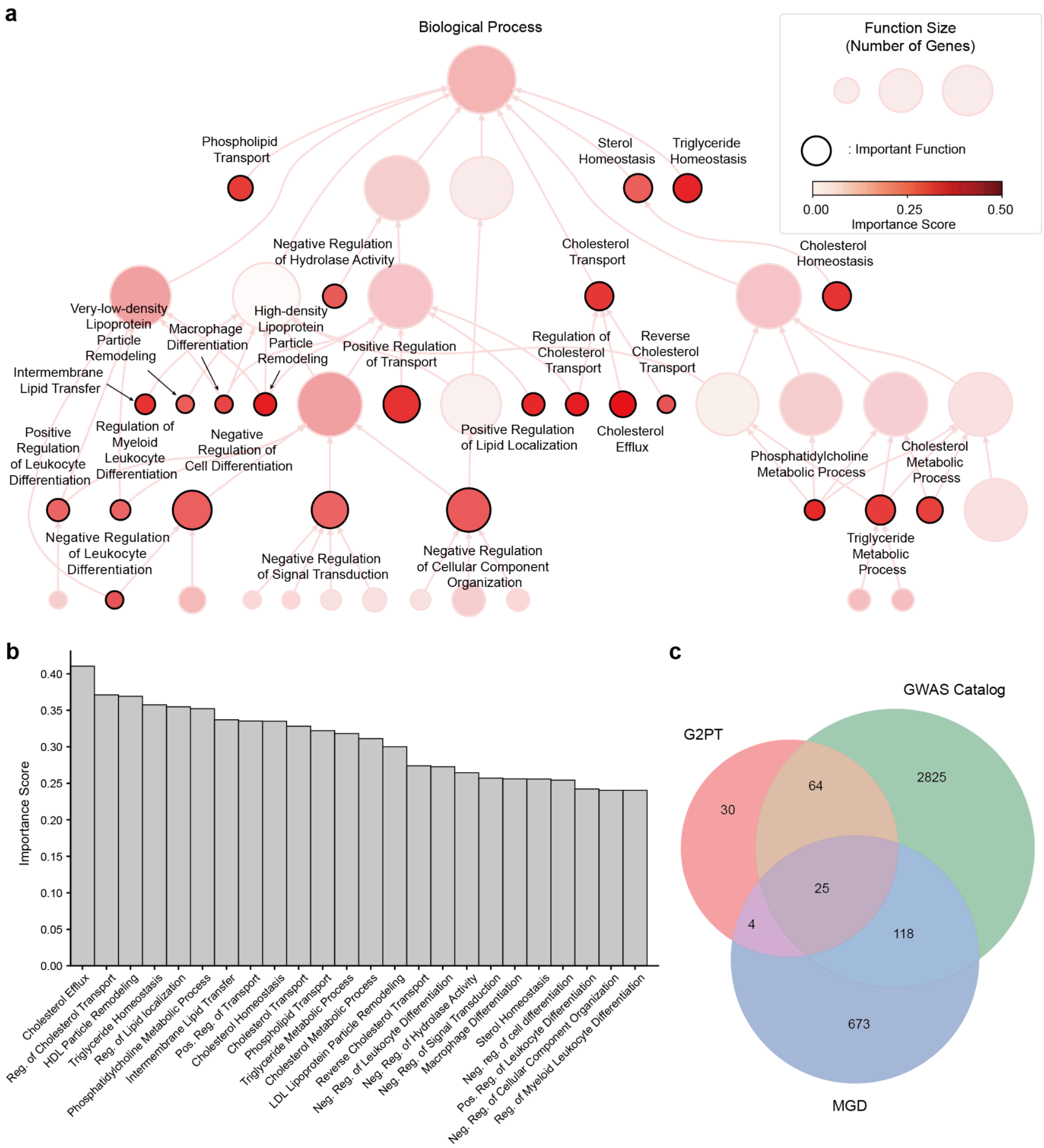

Fig. 3: **Important Functions in Prediction of TG/HGL**. **a**, Hierarchy of important functions extracted from the GO Biological Process database. Functions are represented as circles, with the size of each circle proportional to the number of genes assigned to that function. Arrows represent involvement in a broader function ("is_a") or containment of one function by another ("part_of"). Color intensity represents the importance score, which captures the attention given to that function across the human individuals considered by the G2PT model (**Methods**). **b**, Bar plot showing the importance scores of the top 24 multigenic functions (GO Biological Process) identified by the G2PT model. **c**, Venn diagram illustrating the overlap of genes returned by G2PT (red) versus GWAS Catalog (green) or MGD (blue). Numbers in each section represent the count of genes shared by the corresponding results. GWAS: Genome Wide Association Studies of TG, HDL, or TG/HDL. MGD: Mouse Genome Database.

2.5 G2PT Uncovers an Epistatic Interaction Between CETP and APOC1.

Within the function of HDL Particle Remodeling—one of the highest scoring important functions (**Fig. 3b**), which involves enlargement or reduction of HDL particle size—two SNPs that were particularly impactful for phenotype prediction were rs1800775, a SNP on chromosome band 16q13 associated with the cholesteryl ester transfer protein (CETP), and rs483082, a SNP on 19q13 linked to the apolipoprotein C-I (APOC1) (**Fig. 4c**). Among individuals for which G2PT placed high attention on HDL Particle Remodeling, the states of these two SNPs were significantly correlated, indicative of genetic epistasis (**Supplementary Fig. 4**, Fisher's exact test $p < 10^{-26}$). In particular, individuals homozygous for the rs483082 reference allele showed a marked underrepresentation for presence of the minor allele for rs1800775 (**Fig. 4d**). Notably, the APOC1 SNP was linked to an increase in TG/HDL ($p < 10^{-48}$), the CETP SNP was linked to a decrease ($p < 10^{-85}$), and in cases where the minor alleles for both SNPs were present or absent simultaneously, the average TG/HDL ratio across participants remained unchanged ($p = 0.66$) (**Fig. 4e**). These findings suggested that these SNPs act in an opposing, compensatory fashion to combinatorially modulate the TG/HDL phenotype. Using combinatorial linear modeling [66, 67] (**Methods**, **Supplementary Fig. 4**), this epistatic interaction was revealed to be statistically significant (Bonferroni $q < 10^{-4}$), along with 6 other SNP-SNP epistatic interactions among the 24 important functions (**Methods**, **Supplementary Table 2**).

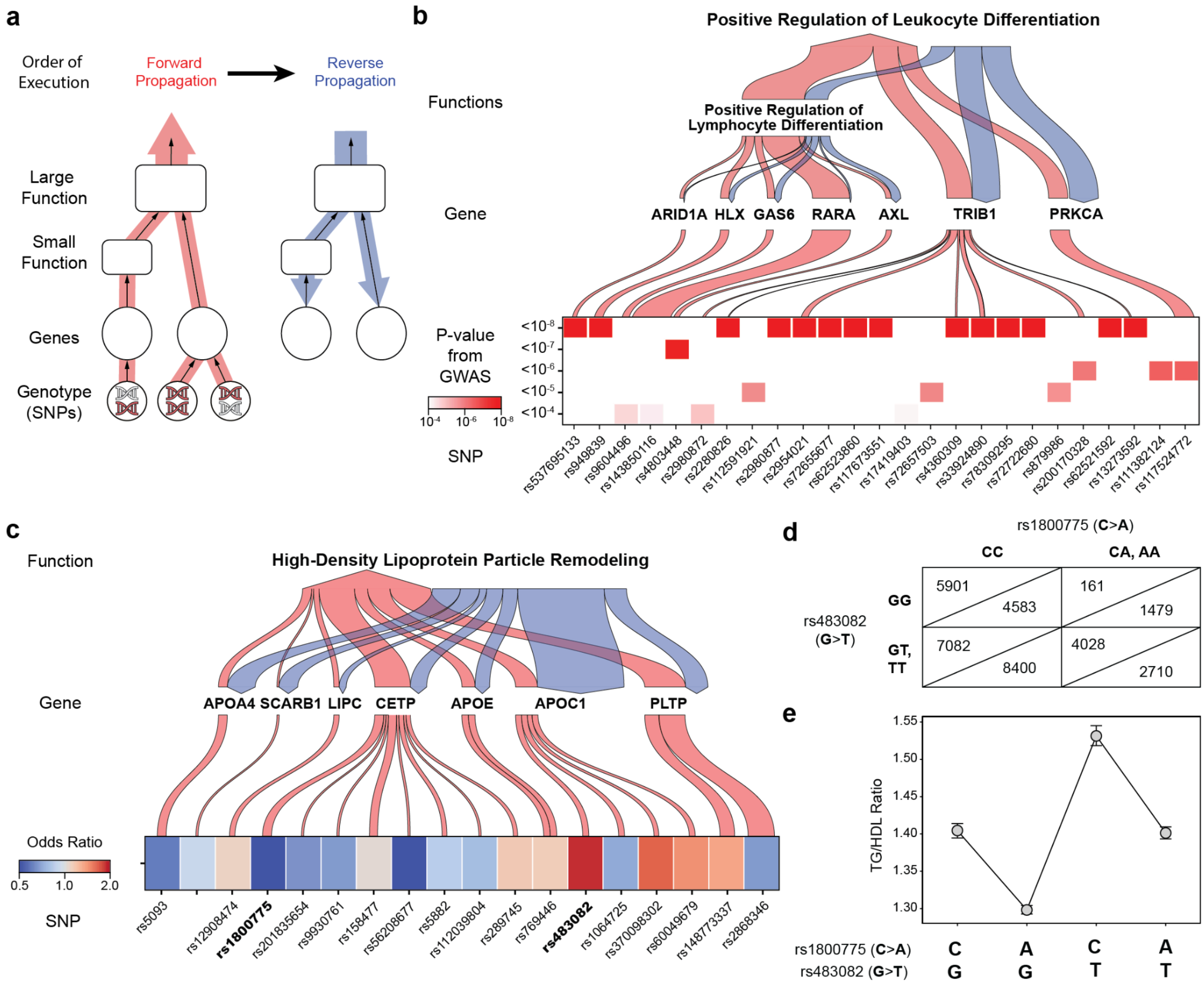

Fig. 4: Convergence of SNPs on Multigenic Functions. a, Illustration of the bidirectional flow of information through the functional hierarchy. Forward propagation (red arrows) represents genetic information flow from SNPs to functions, while reverse propagation (blue arrows) indicates information flow from functions back to genes. Forward and reverse propagation are sequential steps during model execution. **b**, Specific example of (a), visualizing flow of information through functions related to leukocyte differentiation. Arrow width is proportional to the amount of attention given by the Transformer model. Below the diagram is the significance from GWAS summary statistics of each relevant SNP, with darker red implying greater GWAS significance. **c**, Flow of genetic information from SNPs through key genes involved in HDL particle remodeling. The heatmap shows the odds ratio of each SNP being present in the function receiving high attention, where an odds ratio of 1 implies no significance in the presence or absence. SNPs in bold (rs1800775, rs483082) have a strong epigenetic

interaction as explored in panels d-e. **d**, Contingency table showing combinations of allelic states for the epistatic SNPs (rs1800775, rs483082). The observed number of individuals with each combination is shown in the upper triangles, whereas the expected numbers (assuming no interaction) are shown in the bottom triangles. **e**, TG/HDL ratios for all four combinations of alleles from the interacting pair of SNPs. The circles represent mean values and error bars represent 95% confidence intervals.

2.6 Validation by Mouse Gene Disruptions.

To corroborate the implicated genes and functions against independent supporting evidence, we examined data from the Mouse Genome Database (MGD), which records nearly 364,000 mouse phenotypes in response to gene knock-outs (KO) or other genetic perturbations. Focusing on the 123 genes important to G2PT for predicting TG/HDL ratio, we found that genetic disruptions in 29 of these genes had been documented to cause mouse phenotypes related to circulating TG or HDL cholesterol levels (**Fig. 5**), representing highly significant enrichment (hypergeometric test $q < 10^{-2}$ for both TG and HDL; **Fig. 3c**). Despite having MGD evidence, four of these genes — zinc finger E-box binding homeobox 1 (ZEB1), peroxisome proliferator activated receptor delta (PPARD), G protein subunit alpha 11 (GNA11), and RNA-binding protein Raly (RALY) — had not been previously linked to TG or HDL by previous GWAS results (**Fig. 3c**). Further inspection showed that all four genetic loci harbored SNPs with marginal, but not genome-wide, association to TG/HDL (**Supplementary Fig. 5**), explaining why they might have been overlooked earlier. In addition to TG and HDL, the 123-gene set was also enriched for phenotypes related to lipodystrophy, type-II diabetes, liver metabolism, and immune cell and hematopoietic phenotypes (hypergeometric $q < 10^{-2}$ for all five MGD phenotypes. **Fig. 5**).

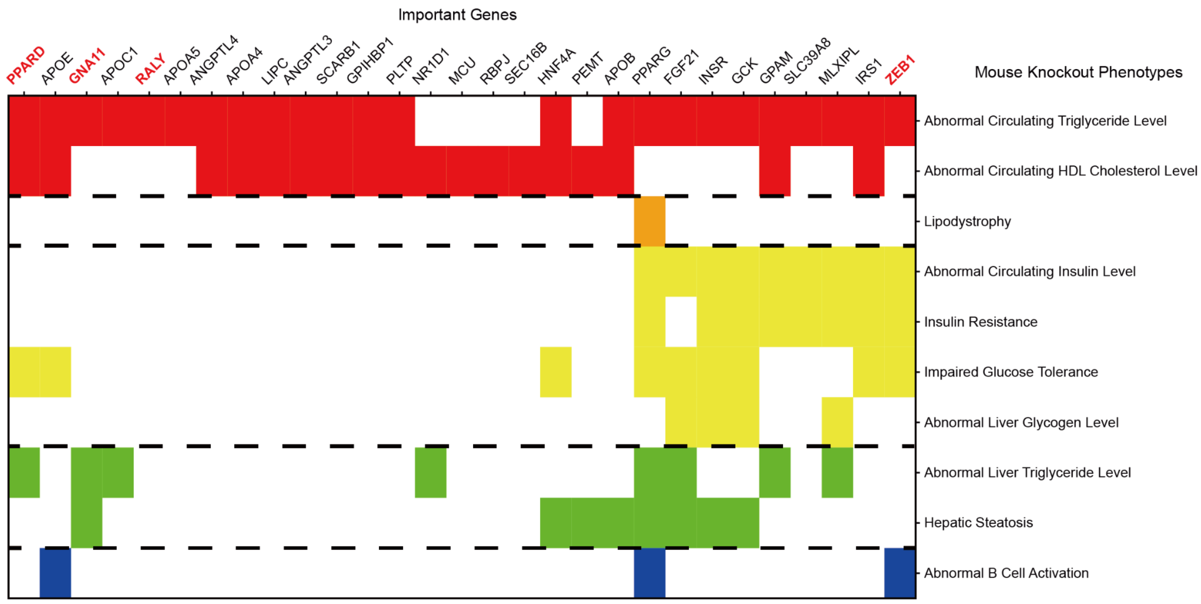

Fig. 5. **Exploration of Important Genes and Functions in Knock-out Mice**. Associations between genes and phenotypes based on experimental data from the Mouse Genome Database (MGD). The heatmap displays the important 29 genes with TG or HDL related phenotypes in MGD (columns) and their significant enrichment in 11 knockout phenotypes (rows). Gene/phenotype effects (for which gene knockout alters the phenotype) are marked by colored squares, with color denoting the class of phenotype: red for TG and HDL homeostasis, orange for adipose-related diseases, yellow for diabetes-related phenotypes, green for liver metabolism phenotypes, and blue for immune and hematopoietic phenotypes. The genes highlighted in red have not been previously associated with TG, HDL, or TG/HDL ratio in the GWAS catalog.

3 Discussion

In this exploration of Transformer models in genotype-phenotype translation, we have found that G2PT achieves better than state-of-the-art performance in prediction of TG/HDL, a model phenotype (**Fig. 2**). One reason for this favorable performance may relate to the multi-head attention mechanism used by Transformer models, which can be very adept at identification of informative interactions among input features [24]. Whether Transformer attention weights offer a means of model interpretation has been a topic of some debate [35, 68]. Here we demonstrate one means by which attention can indeed inform interpretation, by separating the computation of attention into genetic factor propagation and genetic factor translation (**Fig. 1a**). During genetic factor propagation, multi-head attention is used to propagate the effects of SNPs across the knowledge hierarchy of genes and functions. During subsequent genetic factor translation, single-head attention is used to quantify the impacts of these altered genes and functions on an individual's phenotype. The initial use of multi-headed attention enables the model to be informed by diverse interactions, whereas the later use of single-headed attention enables a single score to prioritize the biological mechanisms that underlie a phenotype, making attention interpretable.

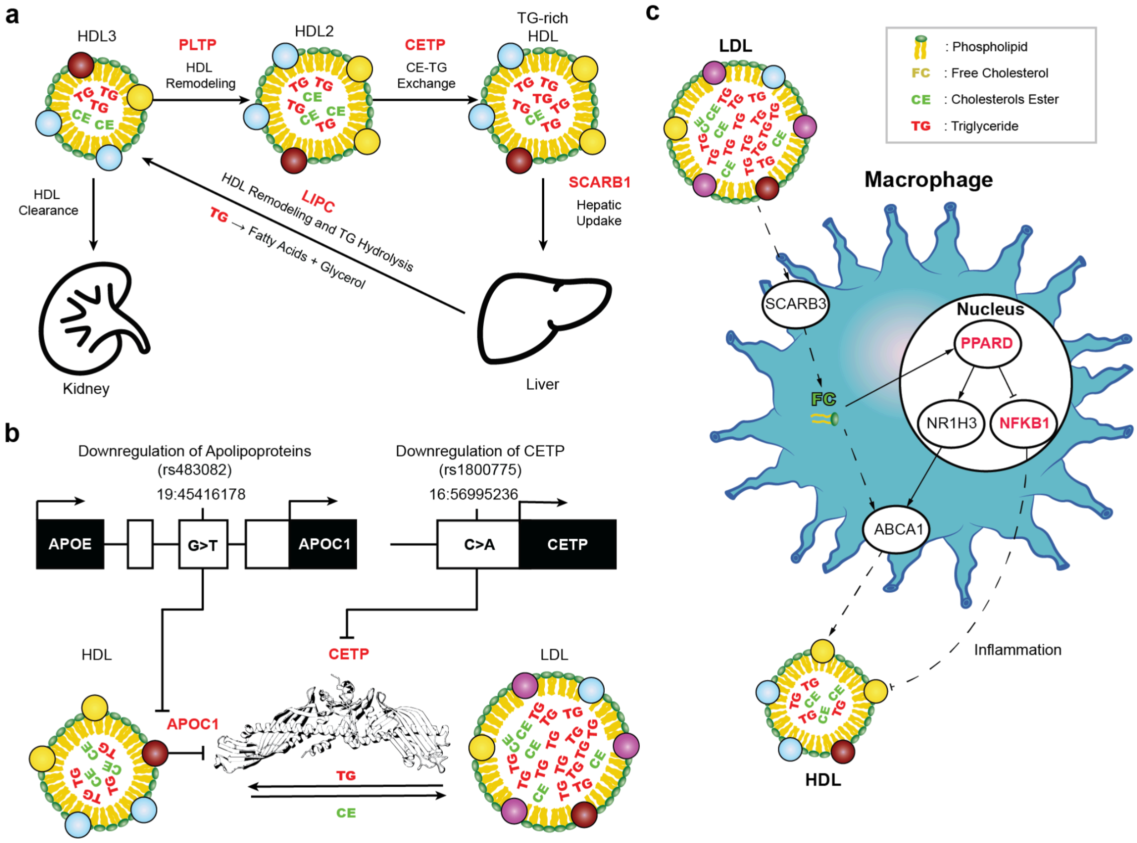

Fig. 6: Biological Models of Important Functions Towards Prediction of TG/HDL. a, HDL particle remodeling pathway. HDL particles are enlarged by phospholipid transfer protein (PLTP). Cholesteryl ester transfer protein (CETP) facilitates the exchange of cholesterol esters (CE) with triglycerides (TG) between HDL and LDL, making HDL TG-rich. Scavenger receptor class B member 1 (SCARB1) mediates hepatic uptake of TG-rich HDL, where hepatic lipase (LIPC) processes HDL into smaller particles for renal clearance and hydrolyzes TG to fatty acids and glycerol. **b**, Interaction between APOE/APOC1 and CETP. SNP rs483082 reduces APOC1 expression via a CTCF binding site, while SNP rs1800775 decreases CETP expression from its promoter. APOC1 inhibits CETP activity; therefore, downregulation of APOC1 leads to increased CETP activity. **c**, Role of macrophages in HDL regulation. Macrophages take up low-density lipoprotein (LDL) and internalize lipids such as free cholesterols (FC) and phospholipids through scavenger receptor class B member 3 (SCARB3). These activate peroxisome proliferator-activated receptor delta (PPARD) and liver X receptor alpha (NR1H3), which lead to export of lipids and apolipoproteins, facilitating assembly of HDL. PPARD also inhibits Nuclear Factor Kappa B Subunit 1 (NFKB1), thereby suppressing inflammatory signaling and increasing HDL levels in the blood. All panels: Red bold text denote important genes for G2PT prediction.

Another notable aspect of the G2PT architecture is the bidirectional flow of genetic information across biological scales. During the genetic factor propagation phase, G2PT not only transmits the effects of variants upwards in scale to impact genes and their collective functions, but it reverses this flow by enabling the states of functions to impact how variants in specific genes within that function are incorporated. Reverse propagation captures the biological context in which genes and variants operate and promotes cross-talk among multiple genetic variants that may have conditional interrelationships. For example, during forward propagation, the embedding of PRKCA was informed by its variant allele type only (**Methods**, **Supplementary Fig. 6a**). Following reverse propagation, the PRKCA embedding was updated in the context of the relatively high importance of its parent function (due to the convergence of this and other genetic variants), which increased its own importance for TG/HDL prediction (**Supplementary Fig. 6b**). Reverse propagation proved very effective in our genotype-phenotype translation task, as it allowed the model to identify PRKCA, a gene involved in hematopoietic functions, as important towards governing the TG/HDL phenotype (**Fig. 3a-b**, **Fig. 4b, Supplementary Fig. 3**). While this functional association had not been made previously (e.g. using tools such as MAGMA), it aligns with the known role of macrophages [69] in HDL-mediated cholesterol transport [70, 71] as well as results from previous population genetic studies which have reported associations between leukocyte counts and lipid traits [72–74].

Further examination of G2PT attention revealed a notable epistatic interaction among genetic variants underlying the genes CETP and APOC1 (**Fig. 4c-e**). This interaction involves a SNP in the promoter region of CETP (rs1800775), which a prior study found is able to downregulate CETP expression [75]. Another SNP is located between APOE and APOC1 on a binding site for the CCCTC-binding factor (CTCF), which has been observed to cause downregulation of APOC1, an inhibitor of CETP activity [76–78] (**Fig. 6a, b**). Plausibly, downregulation of CETP leads to elevated HDL, while downregulation of APOC1 leads to decreased HDL. Under this model, simultaneous downregulation of APOC1 and CETP by the presence of both SNPs would result in an epistatic interaction that stabilizes HDL levels (**Fig. 4e**).

Of the 34 genes implicated by G2PT that did not have previous association to TG/HDL, 4 had outside support via genetic disruptions in mice: PPARD, GNA11, RALY, and ZEB1 (**Fig. 3c**, **Fig. 5a**). Peroxisome proliferator-activated receptors are well-known for their role in lipid metabolism [79]. Like its better studied paralog PPARA, PPARD likely regulates TG/HDL ratio by modulating ATP-binding cassette transporter 1 (ABCA1) through liver X receptor alpha (NR1H3) in macrophages [80]. Unlike PPARA, PPARD indirectly influences HDL levels through its involvement in inflammatory pathways, specifically by interacting with Nuclear Factor Kappa B Subunit 1 (NFKB1) [81] (**Fig. 6c**). Both PPARD and NFKB1 were identified as significant genes by G2PT, underscoring their potential roles in lipid metabolism and inflammation. Notably, PPARD was missed in previous GWAS, likely due to the low minor allele frequency of the SNP (MAF = 3%, association p-value p = 4.3x10^{-8}, **Supplementary Fig. 5**) or the position of the SNP within an intron rather than coding region.

Despite the promising use of Transformers to approach genotype-phenotype questions, our study also points to some current limitations. First, computing attention is an expensive operation [24] with substantial training time and data required to reach convergence. In our study, G2PT required four A30 GPUs over approximately 40 hours of training. There are also challenges introduced by using the Gene Ontology as a prior, as it includes many groups of functions [82] with nearly identical sets of genes, and it has a natural bias towards well-studied functions. Some of these issues may arise because GO has been manually constructed rather than computationally derived [51]. An important step moving forward will thus be to explore models that leverage alternative knowledge structures, such as Reactome [83], or maps of biological structures and functions derived directly from ‘omics data [84, 85]. Regardless, G2PT has immediate application to the genetic analysis of diverse phenotypes of interest, including those related to multigenic diseases such as type-II diabetes, autism, aging, or cancer. More generally, this work presents a template for constructing interpretable Transformer architectures for application to other deep learning challenges.

4 Materials and Methods

4.1 Genome-Wide Association Study (GWAS) for Lipid Traits.

This study uses human genotype data from the UK Biobank [49]. Participants of Caucasian ancestry were genotyped with SNP arrays and had values recorded for their HDL and TG levels [86] based on millimole per liter (mmol/L). The ratios of TG-to-HDL were then log2 transformed. Participant SNPs were imputed utilizing the Michigan Imputation Server [87] guided by the 1000 Genomes Project Phase 3 reference panel [88]. SNPs with an R^2 value greater than 0.2 underwent Linkage Disequilibrium (LD) pruning [89, 90], retaining a total of 203,126 SNPs. We performed a Bayesian logistic regression analysis utilizing BOLT-LMM software to explore the association between SNPs and TG/HDL ratios, while including sex, age, and the top 10 principal components as covariates. The significantly associated SNPs were then selected based on a GWAS p-value threshold to build machine learning models.

4.2 G2PT Model.

G2PT uses a Hierarchical Transformer (HiTR) to integrate and distribute genotypic information across different levels in a gene function hierarchy (here, Gene Ontology Biological Process) (**Fig. 1a**). This model propagates the effects of genetic variations on biological states of genes and functions, then translates these altered states to predict phenotypes (**Fig. 1a-b)**. Specific details are as follows.

Hierarchical Transformer (HiTR). HiTR is a modified version of the Transformer [24] that leverages a graph *H* representing a gene function hierarchy. Nodes of *H* represent biological entities, including SNPs, genes and biological functions (also called "systems" below). Edges represent hierarchical functional relationships, including annotation of genes to functions and involvement of a function in a broader parent function (i.e. "is_a" and "part_of" relations defined in the Gene Ontology). For some entity *i* within *H*, we define E_i as its embedding state, and C_i as the subset of other entities in *H* connected to *i*. The embedding state of each entity *j* connected to *i* is defined as $\{E_j | j \in C_i \subset H\}$. The objective of the HiTR is to "update" the embedding of *i* by considering the effect from each connected entity *j* in C_i. The embedding update is defined as the following:

$$E_i^{'} = E_i + HiTR(E_i, \{E_j | j \in C_i \subset H\}) \tag{1}$$

The HiTR creates an embedding used for updating E_i by computing a weighted sum of linear projections of each E_j:

$$\text{HiTR}(E_i, \{E_j | j \in C_i \subset H\}) = \sum_{j \in C_i} \alpha_{ij} E_j W^v, \text{where } \alpha_{ij} = \frac{(E_i W^q)(E_j W^k)^T}{\sqrt{d_k}} \tag{2}$$

W^q, W^k, W^v are learnable weight matrices encoding the central concepts of query, key and value used in the Transformer model. The interactions between entity *i* and the set of connected entities *j* are represented as attention values α_{ij}. The parameter d_k encodes the size of the hidden dimension of embedding (for scaling). The updated embedding, $E^{'}_i$, is computed by adding the HiTR results to the original embedding, E_i, thus producing an embedding incorporating effects from connected entities in the hierarchy.

Genetic Factor Propagation Phase. G2PT models the bidirectional effects of genetic alterations using the following modules: SNP2Gene, Gene2Sys, Sys2Env, Env2Sys, and Sys2Gene. Each module plays a role in propagating changes through different layers of the function hierarchy. The SNP2Gene, Gene2Sys, and Sys2Env modules propagate the effects of genetic variation upward towards the root of hierarchy. Env2Sys and Sys2Gene then reverse the genetic factor propagation toward lower layers of the hierarchy and back to genes. All of these modules use multi-headed attention.

Overall, the complete aggregation of changes upon a single function (system) can be formulated as:

$$F_i^{altered} = F_i + \text{Gene2Sys}(F_i) + \text{Sys2Env}(F_i^{\text{Gene2Sys}}) + \text{Env2Sys}(F_i^{\text{Gene2Sys}\rightarrow\text{Sys2Env}}) \tag{3}$$

F_i is the original embedding of the function *i*. $\text{Gene2Sys}(F_i)$ updates F_i by computing the HiTR results from its genes, $\text{Sys2Env}(F_i)$ further updates F_i by computing the HiTR results from its sub-functions, and $\text{Env2Sys}(F_i)$ finalizes the update of F_i by computing the HiTR results from its super-functions Additionally, the alteration of a single gene can be formulated as:

$$G_i^{altered} = G_i + \text{SNP2Gene}(G_i) + \text{Sys2Gene}(G_i) \tag{4}$$

G_i is the original embedding of the gene *i*. $\text{SNP2Gene}(G_i)$ updates G_i by computing the HiTR results from embeddings of its SNPs, and $\text{Sys2Gene}(G_i)$ further updates G_i by computing the HiTR results from its functions including gene *i*.

Genetic Factor Translation Phase. G2PT creates an initial embedding, *P*, for each participant using their sex and age covariates and projecting these values through linear layers. G2PT then uses the finalized embedding states of functions (systems) and genes from the genetic factor propagation phase to update the participant embedding in two modules: Sys2Pheno and Gene2Pheno. Sys2Pheno uses equation (1) to update *P* across all function embeddings, while Gene2Pheno uses equation (1) to update *P* across all gene embeddings. To maximize interpretability, both modules only use one attention head. The embeddings from Sys2Pheno and Gene2Pheno are concatenated and projected through a final layer to predict phenotype. Altogether, this layer can be formulated as:

$$Y^{pred} = W^{pred}(concat(Sys2Pheno(P, F^{altered}), Gene2Pheno(P, G^{altered}))) \tag{5}$$

W^{pred} represents the learnable prediction weights. The output of this projection is a predicted value for the phenotype. A detailed description of each module can be found in **Supplementary Methods**.

4.3 Model Training and Performance Evaluation.

Nested cross-validation was employed to robustly evaluate model performance in mapping genotype to phenotype. In each fold, data were split into training, validation, and test sets in a 3:1:1 ratio, with the training set used for selecting significant SNPs and fitting model parameters, the validation set used for tuning model hyperparameters, and the test set comprising held-out samples for independent evaluation of performance. BOLT-LMM was applied to assess the genome-wide marginal significance of genotype-phenotype association for each SNP, enabling us to pre-select SNPs used in model training. With SNPs selected through specific p-value thresholds, the genotypes of participants are represented by vectors of SNPs, where each SNP is encoded as 0 to represent the homozygous reference allele, 1 the heterozygous major/minor allele, and 2 the homozygous minor allele. Using these notations for representing zygosity of a SNP, we constructed a participant-by-SNP matrix to train XGBoost and ElasticNet models using hyperparameters optimized through a grid search. To build the G2PT model, the selected SNPs were mapped to their nearest protein-coding genes, and the resulting gene mappings were used to prune the Gene Ontology (GO) [51] hierarchy with the DDOT package [91].

4.4 Scoring the Importance of Genes and Functions.

During the genetic factor translation phase, G2PT assigns attention scores that quantify the effect of genes and functions towards predicting an individual's phenotype (see above section 4.2). To assess the importance of these genes and functions across the population, we trained the model with the whole population and optimal hyperparameters identified through nested cross-validation (see above section 4.3). Recognizing sex as a confounding factor, we calculated Pearson correlations between the individual attention scores 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 function and gene.

4.5 Detection of Epistatic Interactions in Important Functions.

For each important function, we first used a chi-square test to identify SNPs that exhibited significant differences in frequency within the subset of individuals with high function attention relative to the overall population. Next, we tested the SNPs assigned to genes annotated to the important function for potential epistatic interactions. Epistatic interaction was defined as significant co-occurrence or mutual exclusivity using Fisher's exact test. Finally, to measure the effect of a pair of interacting SNPs on phenotype, we constructed a combinatorial linear model as follows [66, 67]:

$$y = \alpha_1 \cdot SNP1 + \alpha_2 \cdot SNP2 + \alpha_{12} \cdot SNP1 \times SNP2 + \beta \cdot \text{sex} + \gamma \cdot \text{age} \tag{6}$$

Each α represents an effect size of the SNP, which is estimated from data, and SNP1×SNP2 denotes the interaction term between SNP1 and SNP2. β and γ are the effect sizes of the sex and age covariates, respectively. This function models the significance of the epistatic interaction with respect to the phenotype and is corrected through the Bonferroni procedure (**Supplementary Fig. 4**). For each statistical test, we used an adjusted p-value threshold of 10^{-2} to filter out insignificant SNPs and epistatic SNP pairs.

4.6 Validation of Important Functions by Mouse Gene Disruption Data.

We extracted the mammalian phenotype ontology from the `mp.owl` file (https://www.informatics.jax.org/downloads/reports/mp.owl) and mapped genes to phenotypes using mammalian phenotype annotations [92]. We extended gene associations through the mammalian phenotype ontology by linking genes associated with child phenotypes to all their parent phenotypes. Statistical enrichment was assessed using a hypergeometric test with q-values corrected by Benjamini-Hochberg.

5 References

1. Brüning, J.C., Winnay, J., Bonner-Weir, S., Taylor, S.I., Accili, D., Kahn, C.R.: Development of a novel polygenic model of NIDDM in mice heterozygous for IR and IRS-1 null alleles. Cell. 88, 561–572 (1997).
2. Uma Jyothi, K., Reddy, B.M.: Gene-gene and gene-environment interactions in the etiology of type 2 diabetes mellitus in the population of Hyderabad, India. Meta Gene. 5, 9–20 (2015).
3. Doria, A., Patti, M.-E., Kahn, C.R.: The emerging genetic architecture of type 2 diabetes. Cell Metab. 8, 186–200 (2008).
4. Abd El-Aziz, T.A., Mohamed, R.H.: LDLR, ApoB and ApoE genes polymorphisms and classical risk factors in premature coronary artery disease. Gene. 590, 263–269 (2016).
5. Boyle, E.A., Li, Y.I., Pritchard, J.K.: An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. 169, 1177–1186 (2017).
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., Lin, T., Austin, J., McGrath, J.J., Hickie, I.B., Murray, G.K., Visscher, P.M.: From Basic Science to Clinical Application of Polygenic Risk Scores: A Primer. JAMA Psychiatry. 78, 101–109 (2021).
9. Wilson, P.W.F., Meigs, J.B., Sullivan, L., Fox, C.S., Nathan, D.M., D'Agostino, R.B., Sr: Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study: The Framingham offspring study. Arch. Intern. Med. 167, 1068–1074 (2007).
10. Hippisley-Cox, J., Coupland, C., Vinogradova, Y., Robson, J., Minhas, R., Sheikh, A., Brindle, P.: Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. BMJ. 336, 1475–1482 (2008).
11. 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).
12. Milne, R.L., Kuchenbaecker, K.B., Michailidou, K., Beesley, J., Kar, S., Lindström, S., Hui, S., Lemaçon, A., Soucy, P., Dennis, J., Jiang, X., Rostamianfar, A., Finucane, H., Bolla, M.K., McGuffog, L., Wang, Q., Aalfs, C.M., ABCTB Investigators, Adams, M., Adlard, J., Agata, S., Ahmed, S., Ahsan, H., Aittomäki, K., Al-Ejeh, F., Allen, J., Ambrosone, C.B., Amos, C.I., Andrulis, I.L., Anton-Culver, H., Antonenkova, N.N., Arndt, V., Arnold, N., Aronson, K.J., Auber, B., Auer, P.L., Ausems, M.G.E.M., Azzollini, J., Bacot, F., Balmaña, J., Barile, M., Barjhoux, L., Barkardottir, R.B., Barrdahl, M., Barnes, D., Barrowdale, D., Baynes, C., Beckmann, M.W., Benitez, J., Bermisheva, M., Bernstein, L., Bignon, Y.-J., Blazer, K.R., Blok, M.J., Blomqvist, C., Blot, W., Bobolis, K., Boeckx, B., Bogdanova, N.V., Bojesen, A., Bojesen, S.E., Bonanni, B., Børresen-Dale, A.-L., Bozsik, A., Bradbury, A.R., Brand, J.S., Brauch, H., Brenner, H., Bressac-de Paillerets, B., Brewer, C., Brinton, L., Broberg, P., Brooks-Wilson, A., Brunet, J., Brüning, T., Burwinkel, B., Buys, S.S., Byun, J., Cai, Q., Caldés, T., Caligo, M.A., Campbell, I., Canzian, F., Caron, O., Carracedo, A., Carter, B.D., Castelao, J.E., Castera, L., Caux-Moncoutier, V., Chan, S.B., Chang-Claude, J., Chanock, S.J., Chen, X., Cheng, T.-Y.D., Chiquette, J., Christiansen, H., Claes, K.B.M., Clarke, C.L., Conner, T., Conroy, D.M., Cook, J., Cordina-Duverger, E., Cornelissen, S., Coupier, I., Cox, A., Cox, D.G., Cross, S.S., Cuk, K., Cunningham, J.M., Czene, K., Daly, M.B., Damiola, F., Darabi, H., Davidson, R., De Leeneer, K., Devilee, P., Dicks, E., Diez, O., Ding, Y.C., Ditsch, N., Doheny, K.F., Domchek, S.M., Dorfling, C.M., Dörk, T., Dos-Santos-Silva, I., Dubois, S., Dugué, P.-A., Dumont, M., Dunning, A.M., Durcan, L., Dwek, M., Dworniczak, B., Eccles, D., Eeles, R., Ehrencrona, H., Eilber, U., Ejlertsen, B., Ekici, A.B., Eliassen, A.H., EMBRACE, Engel, C., Eriksson, M., Fachal, L., Faivre, L., Fasching, P.A., Faust, U., Figueroa, J., Flesch-Janys, D., Fletcher, O., Flyger, H., Foulkes, W.D., Friedman, E., Fritschi, L., Frost, D., Gabrielson, M., Gaddam, P., Gammon, M.D., Ganz, P.A., Gapstur, S.M., Garber, J., Garcia-Barberan, V., García-Sáenz, J.A., Gaudet, M.M., Gauthier-Villars, M., Gehrig, A., GEMO Study Collaborators, Georgoulias, V., Gerdes, A.-M., Giles, G.G., Glendon, G., Godwin, A.K., Goldberg, M.S., Goldgar, D.E., González-Neira, A., Goodfellow, P., Greene, M.H., Alnæs, G.I.G., Grip, M., Gronwald, J., Grundy, A., Gschwantler-Kaulich, D., Guénel, P., Guo, Q., Haeberle, L., Hahnen, E., Haiman, C.A., Håkansson, N., Hallberg, E., Hamann, U., Hamel, N., Hankinson, S., Hansen, T.V.O., Harrington, P., Hart, S.N., Hartikainen, J.M., Healey, C.S., HEBON, Hein, A., Helbig, S., Henderson, A., Heyworth, J., Hicks, B., Hillemanns, P., Hodgson, S., Hogervorst, F.B., Hollestelle, A., Hooning, M.J., Hoover, B., Hopper, J.L., Hu, C., Huang, G., Hulick, P.J., Humphreys, K., Hunter, D.J., Imyanitov, E.N., Isaacs, C., Iwasaki, M., Izatt, L., Jakubowska, A., James, P., Janavicius, R., Janni, W., Jensen, U.B., John, E.M., Johnson, N., Jones, K., Jones, M., Jukkola-Vuorinen, A., Kaaks, R., Kabisch, M., Kaczmarek, K., Kang, D., Kast, K., kConFab/AOCS Investigators, Keeman, R., Kerin, M.J., Kets, C.M., Keupers, M., Khan, S., Khusnutdinova, E., Kiiski, J.I., Kim, S.-W., Knight, J.A., Konstantopoulou, I., Kosma, V.-M., Kristensen, V.N., Kruse, T.A., Kwong, A., Lænkholm, A.-V., Laitman, Y., Lalloo, F., Lambrechts, D., Landsman, K.,

Lasset, C., Lazaro, C., Le Marchand, L., Lecarpentier, J., Lee, A., Lee, E., Lee, J.W., Lee, M.H., Lejbkowicz, F., Lesueur, F., Li, J., Lilyquist, J., Lincoln, A., Lindblom, A., Lissowska, J., Lo, W.-Y., Loibl, S., Long, J., Loud, J.T., Lubinski, J., Luccarini, C., Lush, M., MacInnis, R.J., Maishman, T., Makalic, E., Kostovska, I.M., Malone, K.E., Manoukian, S., Manson, J.E., Margolin, S., Martens, J.W.M., Martinez, M.E., Matsuo, K., Mavroudis, D., Mazoyer, S., McLean, C., Meijers-Heijboer, H., Menéndez, P., Meyer, J., Miao, H., Miller, A., Miller, N., Mitchell, G., Montagna, M., Muir, K., Mulligan, A.M., Mulot, C., Nadesan, S., Nathanson, K.L., NBSC Collaborators, Neuhausen, S.L., Nevanlinna, H., Nevelsteen, I., Niederacher, D., Nielsen, S.F., Nordestgaard, B.G., Norman, A., Nussbaum, R.L., Olah, E., Olopade, O.I., Olson, J.E., Olswold, C., Ong, K.-R., Oosterwijk, J.C., Orr, N., Osorio, A., Pankratz, V.S., Papi, L., Park-Simon, T.-W., Paulsson-Karlsson, Y., Lloyd, R., Pedersen, I.S., Peissel, B., Peixoto, A., Perez, J.I.A., Peterlongo, P., Peto, J., Pfeiler, G., Phelan, C.M., Pinchev, M., Plaseska-Karanfilska, D., Poppe, B., Porteous, M.E., Prentice, R., Presneau, N., Prokofieva, D., Pugh, E., Pujana, M.A., Pylkäs, K., Rack, B., Radice, P., Rahman, N., Rantala, J., Rappaport-Fuerhauser, C., Rennert, G., Rennert, H.S., Rhenius, V., Rhiem, K., Richardson, A., Rodriguez, G.C., Romero, A., Romm, J., Rookus, M.A., Rudolph, A., Ruediger, T., Saloustros, E., Sanders, J., Sandler, D.P., Sangrajrang, S., Sawyer, E.J., Schmidt, D.F., Schoemaker, M.J., Schumacher, F., Schürmann, P., Schwentner, L., Scott, C., Scott, R.J., Seal, S., Senter, L., Seynaeve, C., Shah, M., Sharma, P., Shen, C.-Y., Sheng, X., Shimelis, H., Shrubsole, M.J., Shu, X.-O., Side, L.E., Singer, C.F., Sohn, C., Southey, M.C., Spinelli, J.J., Spurdle, A.B., Stegmaier, C., Stoppa-Lyonnet, D., Sukiennicki, G., Surowy, H., Sutter, C., Swerdlow, A., Szabo, C.I., Tamimi, R.M., Tan, Y.Y., Taylor, J.A., Tejada, M.-I., Tengström, M., Teo, S.H., Terry, M.B., Tessier, D.C., Teulé, A., Thöne, K., Thull, D.L., Tibiletti, M.G., Tihomirova, L., Tischkowitz, M., Toland, A.E., Tollenaar, R.A.E.M., Tomlinson, I., Tong, L., Torres, D., Tranchant, M., Truong, T., Tucker, K., Tung, N., Tyrer, J., Ulmer, H.-U., Vachon, C., van Asperen, C.J., Van Den Berg, D., van den Ouweland, A.M.W., van Rensburg, E.J., Varesco, L., Varon-Mateeva, R., Vega, A., Viel, A., Vijai, J., Vincent, D., Vollenweider, J., Walker, L., Wang, Z., Wang-Gohrke, S., Wappenschmidt, B., Weinberg, C.R., Weitzel, J.N., Wendt, C., Wesseling, J., Whittemore, A.S., Wijnen, J.T., Willett, W., Winqvist, R., Wolk, A., Wu, A.H., Xia, L., Yang, X.R., Yannoukakos, D., Zaffaroni, D., Zheng, W., Zhu, B., Ziogas, A., Ziv, E., Zorn, K.K., Gago-Dominguez, M., Mannermaa, A., Olsson, H., Teixeira, M.R., Stone, J., Offit, K., Ottini, L., Park, S.K., Thomassen, M., Hall, P., Meindl, A., Schmutzler, R.K., Droit, A., Bader, G.D., Pharoah, P.D.P., Couch, F.J., Easton, D.F., Kraft, P., Chenevix-Trench, G., García-Closas, M., Schmidt, M.K., Antoniou, A.C., Simard, J.: Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. Nat. Genet. 49, 1767–1778 (2017).

13. Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P.T., Gross, S.S., Dorfman, L., McLean, C.Y., DePristo, M.A.: A universal SNP and small-indel variant caller using deep neural networks. Nat. Biotechnol. 36, 983–987 (2018).
14. Ma, J.Z., Yu, M.K., Fong, S., Ono, K., Sage, E., Demchak, B., Sharan, R., Ideker, T.: Using deep learning to model the hierarchical structure and function of a cell. Nat. Methods. 15, 290–+ (2018).
15. Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., Telenti, A.: A primer on deep learning in genomics. Nat. Genet. 51, 12–18 (2019).
16. Kuenzi, B.M., Park, J., Fong, S.H., Sanchez, K.S., Lee, J., Kreisberg, J.F., Ma, J.Z., Ideker, T.: Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells. Cancer Cell. 38, 672–+ (2020).
17. Grinberg, N.F., Orhobor, O.I., King, R.D.: An evaluation of machine-learning for predicting phenotype: studies in yeast, rice, and wheat. Mach. Learn. 109, 251–277 (2020).
18. Spiga, O., Cicaloni, V., Dimitri, G.M., Pettini, F., Braconi, D., Bernini, A., Santucci, A.: Machine learning application for patient stratification and phenotype/genotype investigation in a rare disease. Brief. Bioinform. 22, (2021). https://doi.org/10.1093/bib/bbaa434.
19. Cheng, C.-Y., Li, Y., Varala, K., Bubert, J., Huang, J., Kim, G.J., Halim, J., Arp, J., Shih, H.-J.S., Levinson, G., Park, S.H., Cho, H.Y., Moose, S.P., Coruzzi, G.M.: Evolutionarily informed machine learning enhances the power of predictive gene-to-phenotype relationships. Nat. Commun. 12, 5627 (2021).
20. Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J.R., Grabska-Barwinska, A., Taylor, K.R., Assael, Y., Jumper, J., Kohli, P., Kelley, D.R.: Effective gene expression prediction from sequence by integrating long-range interactions. Nat. Methods. 18, 1196–1203 (2021).
21. Medvedev, A., Mishra Sharma, S., Tsatsorin, E., Nabieva, E., Yarotsky, D.: Human genotype-to-phenotype predictions: Boosting accuracy with nonlinear models. PLoS One. 17, e0273293 (2022).
22. Tonner, P.D., Pressman, A., Ross, D.: Interpretable modeling of genotype-phenotype landscapes with state-of-the-art predictive power. Proc. Natl. Acad. Sci. U. S. A. 119, e2114021119 (2022).
23. Theodoris, C.V., Xiao, L., Chopra, A., Chaffin, M.D., Al Sayed, Z.R., Hill, M.C., Mantineo, H., Brydon, E.M., Zeng, Z., Liu, X.S., Ellinor, P.T.: Transfer learning enables predictions in network biology. Nature. 618, 616–624 (2023).
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need. arXiv e-prints. arXiv:1706.03762 (2017).

25. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, http://arxiv.org/abs/1810.04805, (2018).
26. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, http://arxiv.org/abs/2010.11929, (2020).
27. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language Models are Few-Shot Learners, http://arxiv.org/abs/2005.14165, (2020).
28. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D.: Highly accurate protein structure prediction with AlphaFold. Nature. 596, 583–589 (2021).
29. Cui, H., Wang, C., Maan, H., Wang, B.: scGPT: Towards Building a Foundation Model for Single-Cell Multi-omics Using Generative AI, https://www.biorxiv.org/content/biorxiv/early/2023/05/01/2023.04.30.538439, (2023). https://doi.org/10.1101/2023.04.30.538439.
30. Yang, F., Wang, W., Wang, F., Fang, Y., Tang, D., Huang, J., Lu, H., Yao, J.: scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. Nature Machine Intelligence. 4, 852–866 (2022).
31. Consens, M.E., Dufault, C., Wainberg, M., Forster, D., Karimzadeh, M., Goodarzi, H., Theis, F.J., Moses, A., Wang, B.: To Transformers and Beyond: Large Language Models for the Genome, http://arxiv.org/abs/2311.07621, (2023).
32. Ji, Y., Zhou, Z., Liu, H., Davuluri, R.V.: DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics. 37, 2112–2120 (2021).
33. Bahdanau, D., Cho, K., Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate, http://arxiv.org/abs/1409.0473, (2014).
34. Kim, Y., Denton, C., Hoang, L., Rush, A.M.: Structured Attention Networks, http://arxiv.org/abs/1702.00887, (2017).
35. Wiegreffe, S., Pinter, Y.: Attention is not not Explanation, http://arxiv.org/abs/1908.04626, (2019).
36. Manning, C.D., Clark, K., Hewitt, J., Khandelwal, U., Levy, O.: Emergent linguistic structure in artificial neural networks trained by self-supervision. Proc. Natl. Acad. Sci. U. S. A. 117, 30046–30054 (2020).
37. Hao, Y., Dong, L., Wei, F., Xu, K.: Self-Attention Attribution: Interpreting Information Interactions Inside Transformer. AAAI. 35, 12963–12971 (2021).
38. Azodi, C.B., Tang, J., Shiu, S.-H.: Opening the Black Box: Interpretable Machine Learning for Geneticists. Trends Genet. 36, 442–455 (2020).
39. 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).
40. Watson, D.S.: Interpretable machine learning for genomics. Hum. Genet. 141, 1499–1513 (2022).
41. Qiu, W., Chen, H., Dincer, A.B., Lundberg, S., Kaeberlein, M., Lee, S.-I.: Interpretable machine learning prediction of all-cause mortality. Commun. Med. 2, 125 (2022).
42. Susnjak, T., Griffin, E.: Towards clinical prediction with transparency: An explainable AI approach to survival modelling in residential aged care, https://www.medrxiv.org/content/10.1101/2024.01.14.24301299v1, (2024). https://doi.org/10.1101/2024.01.14.24301299.
43. Assmann, G., Schulte, H.: Relation of high-density lipoprotein cholesterol and triglycerides to incidence of atherosclerotic coronary artery disease (the PROCAM experience). Am. J. Cardiol. 70, 733–737 (1992).
44. Burchfiel, C.M., Laws, A., Benfante, R., Goldberg, R.J., Hwang, L.J., Chiu, D., Rodriguez, B.L., Curb, J.D., Sharp, D.S.: Combined effects of HDL cholesterol, triglyceride, and total cholesterol concentrations on 18-year risk of atherosclerotic disease. Circulation. 92, 1430–1436 (1995).
45. Kim, J., Shin, S.-J., Kim, Y.-S., Kang, H.-T.: Positive association between the ratio of triglycerides to high-density lipoprotein cholesterol and diabetes incidence in Korean adults. Cardiovasc. Diabetol. 20, 183 (2021).
46. Yuge, H., Okada, H., Hamaguchi, M., Kurogi, K., Murata, H., Ito, M., Fukui, M.: Triglycerides/HDL cholesterol ratio and type 2 diabetes incidence: Panasonic Cohort Study 10. Cardiovasc. Diabetol. 22, 308 (2023).
47. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space, http://arxiv.org/abs/1301.3781, (2013).
48. Grover, A., Leskovec, J.: node2vec: Scalable Feature Learning for Networks. KDD. 2016, 855–864 (2016).

49. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., Collins, R.: 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).
50. DeForest, N., Wang, Y., Zhu, Z., Dron, J.S., Koesterer, R., Natarajan, P., Flannick, J., Amariuta, T., Peloso, G.M., Majithia, A.R.: 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).
51. Gene Ontology, Consortium: The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. 49, D325–D334 (2021).
52. Dagasso, G., Yan, Y., Wang, L., Li, L., Kutcher, R., Zhang, W., Jin, L.: Comprehensive-GWAS: a pipeline for genome-wide association studies utilizing cross-validation to assess the predictivity of genetic variations. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 1361–1367. IEEE (2020).
53. Nicholls, H., Ng, F.L., Watson, D., Jacobsen, J., Warren, H., Cacheiro, P., Smedley, D., Munroe, P., Caulfield, M., Cabrera, C., Barnes, M.: Post-GWAS machine learning prioritizes key genes regulating blood pressure, https://www.researchsquare.com/article/rs-2402775/v1, (2023). https://doi.org/10.21203/rs.3.rs-2402775/v1.
54. Loh, P.-R., Tucker, G., Bulik-Sullivan, B.K., Vilhjálmsson, B.J., Finucane, H.K., Salem, R.M., Chasman, D.I., Ridker, P.M., Neale, B.M., Berger, B., Patterson, N., Price, A.L.: Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat. Genet. 47, 284–290 (2015).
55. Lamarche, B., Després, J.P., Moorjani, S., Cantin, B., Dagenais, G.R., Lupien, P.J.: Triglycerides and HDL-cholesterol as risk factors for ischemic heart disease. Results from the Québec cardiovascular study. Atherosclerosis. 119, 235–245 (1996).
56. Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System, http://arxiv.org/abs/1603.02754, (2016).
57. Kersten, S.: Role and mechanism of the action of angiopoietin-like protein ANGPTL4 in plasma lipid metabolism. J. Lipid Res. 62, 100150 (2021).
58. Kathiresan, S., Melander, O., Guiducci, C., Surti, A., Burtt, N.P., Rieder, M.J., Cooper, G.M., Roos, C., Voight, B.F., Havulinna, A.S., Wahlstrand, B., Hedner, T., Corella, D., Tai, E.S., Ordovas, J.M., Berglund, G., Vartiainen, E., Jousilahti, P., Hedblad, B., Taskinen, M.-R., Newton-Cheh, C., Salomaa, V., Peltonen, L., Groop, L., Altshuler, D.M., Orho-Melander, M.: Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nat. Genet. 40, 189–197 (2008).
59. Jadhav, K.S., Bauer, R.C.: Trouble with Tribbles-1. Arterioscler. Thromb. Vasc. Biol. 39, 998–1005 (2019).
60. Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., Groza, T., Güneş, O., Hall, P., Hayhurst, J., Ibrahim, A., Ji, Y., John, S., Lewis, E., MacArthur, J.A.L., McMahon, A., Osumi-Sutherland, D., Panoutsopoulou, K., Pendlington, Z., Ramachandran, S., Stefancsik, R., Stewart, J., Whetzel, P., Wilson, R., Hindorff, L., Cunningham, F., Lambert, S.A., Inouye, M., Parkinson, H., Harris, L.W.: The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. Nucleic Acids Res. 51, D977–D985 (2023).
61. de Leeuw, C.A., Mooij, J.M., Heskes, T., Posthuma, D.: MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput. Biol. 11, e1004219 (2015).
62. Richardson, T.G., Sanderson, E., Palmer, T.M., Ala-Korpela, M., Ference, B.A., Davey Smith, G., Holmes, M.V.: Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable Mendelian randomisation analysis. PLoS Med. 17, e1003062 (2020).
63. Sinnott-Armstrong, N., Tanigawa, Y., Amar, D., Mars, N., Benner, C., Aguirre, M., Venkataraman, G.R., Wainberg, M., Ollila, H.M., Kiiskinen, T., Havulinna, A.S., Pirruccello, J.P., Qian, J., Shcherbina, A., FinnGen, Rodriguez, F., Assimes, T.L., Agarwala, V., Tibshirani, R., Hastie, T., Ripatti, S., Pritchard, J.K., Daly, M.J., Rivas, M.A.: Genetics of 35 blood and urine biomarkers in the UK Biobank. Nat. Genet. 53, 185–194 (2021).
64. Koskeridis, F., Evangelou, E., Said, S., Boyle, J.J., Elliott, P., Dehghan, A., Tzoulaki, I.: Pleiotropic genetic architecture and novel loci for C-reactive protein levels. Nat. Commun. 13, 6939 (2022).
65. Oliveri, A., Rebernick, R.J., Kuppa, A., Pant, A., Chen, Y., Du, X., Cushing, K.C., Bell, H.N., Raut, C., Prabhu, P., Chen, V.L., Halligan, B.D., Speliotes, E.K.: Comprehensive genetic study of the insulin resistance marker TG:HDL-C in the UK Biobank. Nat. Genet. 56, 212–221 (2024).
66. 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).
67. Niel, C., Sinoquet, C., Dina, C., Rocheleau, G.: A survey about methods dedicated to epistasis detection. Front. Genet. 6, 285 (2015).
68. Jain, S., Wallace, B.C.: Attention is not explanation. arXiv preprint arXiv:1902.10186. (2019).
69. Rönnstrand, L.: Signal transduction via the stem cell factor receptor/c-Kit. Cell. Mol. Life Sci. 61,

2535–2548 (2004).

70. Lazar, M.A.: Progress in cardiovascular biology: PPAR for the course. Nat. Med. 7, 23–24 (2001).
71. Pownall, H.J., Rosales, C., Gillard, B.K., Gotto, A.M., Jr: High-density lipoproteins, reverse cholesterol transport and atherogenesis. Nat. Rev. Cardiol. 18, 712–723 (2021).
72. Lai, Y.C., Woollard, K.J., McClelland, R.L., Allison, M.A., Rye, K.-A., Ong, K.L., Cochran, B.J.: The association of plasma lipids with white blood cell counts: Results from the Multi-Ethnic Study of Atherosclerosis. J. Clin. Lipidol. 13, 812–820 (2019).
73. Tucker, B., Sawant, S., McDonald, H., Rye, K.-A., Patel, S., Ong, K.L., Cochran, B.J.: The association of serum lipid and lipoprotein levels with total and differential leukocyte counts: Results of a cross-sectional and longitudinal analysis of the UK Biobank. Atherosclerosis. 319, 1–9 (2021).
74. Groenen, A.G., Bazioti, V., van Zeventer, I.A., Chen, L., Groot, H.E., Balder, J.-W., Zhernakova, A., van der Harst, P., Rimbert, A., Kuivenhoven, J.A., Fu, J., Westerterp, M.: Large HDL particles negatively associate with leukocyte counts independent of cholesterol efflux capacity: A cross sectional study in the population-based LifeLines DEEP cohort. Atherosclerosis. 343, 20–27 (2022).
75. Dachet, C., Poirier, O., Cambien, F., Chapman, J., Rouis, M.: New functional promoter polymorphism, CETP/-629, in cholesteryl ester transfer protein (CETP) gene related to CETP mass and high density lipoprotein cholesterol levels: role of Sp1/Sp3 in transcriptional regulation. Arterioscler. Thromb. Vasc. Biol. 20, 507–515 (2000).
76. Clark, D., Skrobot, O.A., Adebiyi, I., Susce, M.T., de Leon, J., Blakemore, A.F., Arranz, M.J.: Apolipoprotein-E gene variants associated with cardiovascular risk factors in antipsychotic recipients. Eur. Psychiatry. 24, 456–463 (2009).
77. Nazarian, A., Philipp, I., Culminskaya, I., He, L., Kulminski, A.M.: Inter- and intra-chromosomal modulators of the APOE ε2 and ε4 effects on the Alzheimer’s disease risk. Geroscience. 45, 233–247 (2023).
78. GTEx Consortium: Erratum: Genetic effects on gene expression across human tissues. Nature. 553, 530 (2018).
79. van Raalte, D.H., Li, M., Pritchard, P.H., Wasan, K.M.: Peroxisome proliferator-activated receptor (PPAR)-: A pharmacological target with a promising future. Pharm. Res. 21, 1531–1538 (2004).
80. Sprecher, D.L., Massien, C., Pearce, G., Billin, A.N., Perlstein, I., Willson, T.M., Hassall, D.G., Ancellin, N., Patterson, S.D., Lobe, D.C., Johnson, T.G.: Triglyceride:high-density lipoprotein cholesterol effects in healthy subjects administered a peroxisome proliferator activated receptor delta agonist. Arterioscler. Thromb. Vasc. Biol. 27, 359–365 (2007).
81. Barish, G.D., Atkins, A.R., Downes, M., Olson, P., Chong, L.-W., Nelson, M., Zou, Y., Hwang, H., Kang, H., Curtiss, L., Evans, R.M., Lee, C.-H.: PPARdelta regulates multiple proinflammatory pathways to suppress atherosclerosis. Proc. Natl. Acad. Sci. U. S. A. 105, 4271–4276 (2008).
82. Thomas, P.D.: The Gene Ontology and the meaning of biological function. Methods Mol. Biol. 1446, 15–24 (2017).
83. Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., Griss, J., Sevilla, C., Matthews, L., Gong, C., Deng, C., Varusai, T., Ragueneau, E., Haider, Y., May, B., Shamovsky, V., Weiser, J., Brunson, T., Sanati, N., Beckman, L., Shao, X., Fabregat, A., Sidiropoulos, K., Murillo, J., Viteri, G., Cook, J., Shorser, S., Bader, G., Demir, E., Sander, C., Haw, R., Wu, G., Stein, L., Hermjakob, H., D’Eustachio, P.: The reactome pathway knowledgebase 2022. Nucleic Acids Res. 50, D687–D692 (2022).
84. Qin, Y., Huttlin, E.L., Winsnes, C.F., Gosztyla, M.L., Wacheul, L., Kelly, M.R., Blue, S.M., Zheng, F., Chen, M., Schaffer, L.V., Licon, K., Bäckström, A., Vaites, L.P., Lee, J.J., Ouyang, W., Liu, S.N., Zhang, T., Silva, E., Park, J., Pitea, A., Kreisberg, J.F., Gygi, S.P., Ma, J., Harper, J.W., Yeo, G.W., Lafontaine, D.L.J., Lundberg, E., Ideker, T.: A multi-scale map of cell structure fusing protein images and interactions. Nature. 600, 536–542 (2021).
85. Zheng, F., Kelly, M.R., Ramms, D.J., Heintschel, M.L., Tao, K., Tutuncuoglu, B., Lee, J.J., Ono, K., Foussard, H., Chen, M., Herrington, K.A., Silva, E., Liu, S.N., Chen, J., Churas, C., Wilson, N., Kratz, A., Pillich, R.T., Patel, D.N., Park, J., Kuenzi, B., Yu, M.K., Licon, K., Pratt, D., Kreisberg, J.F., Kim, M., Swaney, D.L., Nan, X., Fraley, S.I., Gutkind, J.S., Krogan, N.J., Ideker, T.: Interpretation of cancer mutations using a multiscale map of protein systems. Science. 374, eabf3067 (2021).
86. Richardson, T.G., Leyden, G.M., Wang, Q., Bell, J.A., Elsworth, B., Davey Smith, G., Holmes, M.V.: Characterising metabolomic signatures of lipid-modifying therapies through drug target mendelian randomisation. PLoS Biol. 20, e3001547 (2022).
87. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P.-R., Iacono, W.G., Swaroop, A., Scott, L.J., Cucca, F., Kronenberg, F., Boehnke, M., Abecasis, G.R., Fuchsberger, C.: Next-generation genotype imputation service and methods. Nat. Genet. 48, 1284–1287 (2016).
88. 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., Abecasis, G.R.: A global reference for human genetic variation. Nature. 526, 68–74 (2015).

89. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C.: PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575 (2007).
90. Gaunt, T.R., Rodríguez, S., Day, I.N.M.: Cubic exact solutions for the estimation of pairwise haplotype frequencies: implications for linkage disequilibrium analyses and a web tool “CubeX.” BMC Bioinformatics. 8, 1–9 (2007).
91. Yu, M.K., Ma, J., Ono, K., Zheng, F., Fong, S.H., Gary, A., Chen, J., Demchak, B., Pratt, D., Ideker, T.: DDOT: A Swiss Army Knife for Investigating Data-Driven Biological Ontologies. Cell Syst. 8, 267–273.e3 (2019).
92. Smith, C.L., Eppig, J.T.: The mammalian phenotype ontology: enabling robust annotation and comparative analysis. Wiley Interdiscip. Rev. Syst. Biol. Med. 1, 390–399 (2009).