

Title: Human Plasma Proteomics Links Modifiable Lifestyle Exposome to Disease Risk

Authors: Shakson Isaac¹, Randall J. Ellis¹, Alexander Gusev^{2,3,4}, Venkatesh L. Murthy⁵, Miriam S. Udler^{6,7,8}, Chirag J. Patel^{1*}

*Correspondence to: chirag_patel@hms.harvard.edu

Affiliations:

¹Department of Biomedical Informatics, Harvard Medical School, Boston, MA. 02215

²Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, 02215, USA.

³Division of Genetics, Brigham & Women's Hospital, Boston, MA, USA.

⁴The Eli and Edythe L. Broad Institute, Cambridge, MA, 02142, USA.

⁵Department of Internal Medicine and Frankel Cardiovascular Center, University of Michigan, Ann Arbor, MI, USA

⁶Department of Medicine, Harvard Medical School, Boston, MA, 02115, USA

⁷Programs in Metabolism and Medical & Population Genetics, The Broad Institute of MIT and Harvard, Cambridge, MA, 02142, USA

⁸Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, 02114, USA

Containing: 3987 words for the main text and 632 words for the main figure legends, excluding methods and supplementary materials. 6 Main Figures, 7 Supplementary Figures

Keywords: plasma proteomics; lifestyle factors; exposome; gene-by-environment; disease association; interventions; GLP1; biobank; precision medicine

Abstract

Environmental exposures influence disease risk, yet their underlying biological mechanisms remain poorly understood. We present the Human Exposomic Architecture of the Proteome (HEAP), a framework and resource integrating genetic, exposomic, and proteomic data to uncover how lifestyle influences disease through plasma proteins. Applying HEAP to 2,686 proteins in 53,014 UK Biobank participants, we identified over 11,000 exposure–protein associations across 135 lifestyle exposures. Exposures explained a substantial portion of proteomic variation, with 9% of proteins more influenced by lifestyle than genetics. Mediation analyses across 270 diseases revealed proteins linking exposures to disease risk; for instance, IGFBP1 and IGFBP2 mediated the effects of exercise and diet on type 2 diabetes. These findings were supported by concordant proteomic shifts in interventional studies of exercise and GLP1 agonists, underscoring therapeutic relevance. HEAP provides a resource for advancing disease prevention and precision medicine by revealing mechanisms through which lifestyle shapes human health.

Introduction

Cumulative lifestyle exposures¹ such as diet, exercise, and smoking are well-recognized contributors to disease risk, but their precise biological mechanisms remain poorly understood². Circulating plasma proteomics offer a powerful tool to uncover these mechanisms, as plasma proteins reflect key biological processes, including lipid and hormone transport, enzymatic activity, signaling, and immune functions³. Fluctuations in protein expression are linked to disease mechanisms and serve as biomarkers for chronic conditions, covering metabolic disorders, cardiovascular disease, and neurodegenerative diseases⁴⁻⁶. Uncovering the effects of modifiable exposures and their interactions with genetics is crucial for developing effective prevention strategies, assessing interventions, and improving health outcomes.

Large-scale biobank studies leveraging high-throughput plasma proteomics now enable systematic investigation of the mechanisms linking genetics, the exposome¹, and disease. For example, the largest genetic study of the plasma proteome, using the UK Biobank, cataloged over 14,000 protein quantitative trait loci (pQTLs) and 1,100 gene-by-environment (GxE) interactions across nearly 3,000 plasma proteins in approximately 50,000 individuals^{7,8}. Exposome-wide association studies (ExWAS) have traditionally identified lifestyle, chemical, and social factors associated with disease risk^{9,10}, but often without consideration of causal biological mediators. Recent studies have begun integrating omic profiling to assess exposure effects^{11,12}; however, these are often limited by smaller sample sizes, model mis-specifications, and lack of consideration for multiple exposures simultaneously, leading to a fragmented literature of associations¹³⁻¹⁵. With large-scale biobanks now storing genetic, exposomic, and molecular data at the individual level, we can begin to disentangle how genetics and exposures—and their cumulative burden—drive protein expression and mediate disease risk. We hypothesize that if the lifestyle exposome is causal for disease risk, its effects must be mediated by biological responses, such as changes in protein expression.

Our study recognizes the complex interplay between genetics and environmental exposures in shaping the proteome and its potential impact in disease. To scalably identify genetic and exposomic

influences on protein expression, we introduce the **H**uman **E**xposomic **A**rchitecture of the **P**roteome (HEAP), a statistical machine-learning framework that models associations between lifestyle exposures and plasma proteins in a biobank setting, accounting for genetics, demographics, and comorbidities. HEAP integrates polygenic scores (PGS) and the contribution of multiple exposures via polyexposure scores (PXS) within a penalized regression approach, enabling the assessment of how both genetics and exposures interact to shape the proteome. To mitigate confounding and model bias common in biobank studies¹⁶, HEAP compares exposure-protein associations across multiple model specifications, including adjustments for demographics, population structure, comorbidities, and medication history. Lastly, HEAP quantifies the extent to which genetics and exposures mediate disease risk through specific proteins.

We applied HEAP to UK Biobank data, characterizing the joint contribution of genetics and lifestyle exposures across 2,686 plasma proteins and 135 exposures. We replicated key associations in independent interventional cohorts with circulating proteomics to evaluate potential causal links. Overall, HEAP provides a comprehensive resource for exploring protein mediators of lifestyle-driven disease risk, offering insights to guide targeted interventions and advance personalized medicine.

Results

We analyzed individual-level lifestyle components of the exposome, namely smoking, alcohol use, dietary habits, vitamin consumption, and exercise behavior across 53,014 UK Biobank (UKB) participants. To account for latent factors influencing lifestyle behavior, specifically socioeconomic status, we incorporated aggregated geographical and social determinant data from the UKB. To investigate how lifestyle exposures influence the proteome, we developed the **H**uman **E**xposomic **A**rchitecture of the **P**roteome (HEAP). HEAP performs four main functions across various model specifications to address confounding and model bias (**Figure 1, Table 1**). First, HEAP computes the partitioned variance, R^2 , of genetic (G), exposure (E), and polygenic gene-by-environment (GxE) effects on the expression of each protein in the measured proteome ($n=2,686$ proteins) using a polygenic score (PGS)¹⁷ and constructed polyexposure score (PXS) derived via LASSO regression^{10,18}. Second, HEAP identifies significant

associations between individual exposures and proteins, along with GxE interactions. Third, HEAP identifies protein mediators that link exposures to disease. Finally, HEAP evaluates exposure-protein and protein-disease associations using proteomic shifts observed in interventional studies. For further details on these functions, please refer to the **Methods** section.

Genetic and Exposomic Variation in the Plasma Proteome

Exposomic R^2 values ($E:R^2$) ranged from 0 to 14% (median 0.5%), indicating that aggregate lifestyle exposures explain a modest proportion of the variation in plasma protein expression. Genetic R^2 values ($G:R^2$), in comparison, ranged from 0 to 82% (median 3.6%) (**Figure 2**). Integrating across all models which jointly considered the exposome and genetics, we identified approximately 9% ($n = 242$) of 2,686 plasma proteins had higher variance explained from the exposome versus genetics (i.e., $R^2_{\text{exposures}} > R^2_{\text{genetics}}$) (**Figure 2A**). R^2 values varied across models, with standard errors ranging from 0 to $1e-3$ for genetics and 0 to 0.01 for exposures (**Figure 2B-C; Supplementary Table 1**). A nested cross-validation procedure confirmed that total, exposomic, and genetic R^2 estimates were stable across train-test splits, while gene-by-environment (GxE) R^2 was not (**Supplementary Figure 1**).

Overall, more proteins had higher total genetic versus exposomic variance explained. Specifically, 2,021 proteins had genetic R^2 values greater than 0.01, while 926 proteins had exposomic R^2 values above the same threshold (**Figure 2D**). We performed gene set enrichment analysis (GSEA)¹⁹ to identify GO²⁰ pathways associated with exposures and genetics using ranked protein lists based on $E:R^2$ and $G:R^2$, respectively. Lifestyle exposures were enriched for pathways related to lipid homeostasis (adj.p = $6.2e-3$), cholesterol efflux (adj.p = $4.6e-3$), and cargo receptor activity (adj.p = $2.9e-4$). Genetics were enriched for pathways related to adaptive immune response (adj.p = $7.9e-6$), interferon-gamma production (adj.p = $4.2e-4$), and immune receptor activity (adj.p = $6.8e-7$) (**Figure 2E, Supplementary Table 2**).

To further investigate $E:R^2$, we partitioned the $E:R^2$ by lifestyle exposure categories, such as exercise, diet, alcohol use, and smoking (**Figure 2F, Supplementary Table 3**). Among the top 70 proteins

($E:R^2 > 0.05$), a substantial proportion with high $E:R^2$ were explained by exercise-related exposures, reaching up to R^2 of 0.1. Exercise-related proteins LEP, FABP4, and GDF15 regulate pathways involved in feeding homeostasis and energy expenditure^{21–23}. Diet-related proteins FGF21, GDF15, and REN play key roles in lipid metabolism, body weight, and blood pressure regulation^{23–25}. Alcohol-related proteins CEACAM16, MAMDC4, and FCAMR are involved in cell adhesion and immune regulation^{26,27}. Smoking-related proteins CXCL17 and PRSS8 contribute to mucosal immunity and fluid homeostasis^{28,29}.

Approximately 500 proteins showed low explained variability by either genetics or exposures ($R^2 < 0.01$). Of these, 42 were explained by age, sex, or body mass index (BMI) with $R^2 > 0.01$ (**Figure 2G**). A catalog of each protein's R^2 due to demographics, population structure, and medications is provided (**Supplementary Figure 2, Supplementary Table 4**).

Exposomic and Polygenic Gene-by-Environment Associations to Plasma Proteome

We next identified exposomic associations across the plasma proteome. Using an 80/20 train-test split, we detected 11,177 significant exposomic (E) associations (eg., smoking, activity, and diet) that replicated in both the training and test sets under Bonferroni correction ($p\text{-value} < 7 \times 10^{-8}$) (**Figure 3A, Supplementary Table 5**). The effect sizes of significant E associations were consistent, with pairwise correlations between effect sizes emerging from different specifications greater than 0.95 (**Figure 3B**). However, the minimal specification of age & sex showed weaker correlations compared to specifications adjusting for BMI, population structure, medications, and regional income (**Figure 3C**).

We observed robust trends in the magnitude of effect sizes across different exposure types and duration of exposure. Smoking and alcohol exhibit larger effect sizes compared to exercise, diet, and vitamins, with deprivation indices showing the smallest effect (**Figure 3D**). Longer exposure durations, such as frequent alcohol intake, daily exercise, and daily smoking, had larger effect sizes across the proteome. For example, individuals with daily alcohol intake exhibit stronger associations than those who consume alcohol infrequently (**Figure 3E**). To illustrate the extent of strong and significant exposomic

associations, we highlighted associations such as LEP and Strenuous Sports, and IGFBP1 and Number of Active Days which have independent and additive effects from genetics (**Figure 3F-G**).

At the tissue and pathway level, significant effects ($FDR < 0.05$) were observed for many exposures. Using protein sets derived from GTEx³⁰ and Reactome³¹ (See “Methods” for details), we found tissue-specific effects and corresponding pathway enrichments across lifestyle exposures (**Figure 3H, Supplementary Table 6-7**). Exercise exposures such as strenuous sports and number of days of vigorous physical activity showed positive enrichments in skeletal ($adj.p = 2.2e-7$) and cardiac ($adj.p = 1.5e-6$) muscle. Pathway enrichments included striated muscle contraction ($adj.p = 3.1e-5$) and extracellular matrix (ECM) proteoglycans ($adj.p = 2.6e-4$). Smoking exposure showed positive enrichments in the lung ($adj.p = 3.0e-8$) and negative enrichment in the brain ($adj.p = 1.9e-7$), with pathways including innate immune response ($adj.p = 1.6e-5$) and glycosaminoglycan metabolism ($adj.p = 3.9e-2$), respectively. Alcohol usage had spleen ($adj.p = 5.3e-9$) and whole blood ($adj.p = 3.7e-6$) enrichments, with pathway enrichment in fibrin clot formation ($adj.p = 1.9e-2$) and peptide hormone metabolism ($adj.p = 2.0e-2$). Diet such as fresh fruit intake had tissue-specific effects in the liver ($adj.p = 6.6e-9$) and brain ($adj.p = 1.7e-3$) with pathway enrichments in peptide ligand-binding receptors ($adj.p = 2.6e-2$) and ECM proteoglycans ($adj.p = 2.5e-2$).

We identified 54 significant polygenic gene-by-environment associations (GxE) that replicated in the train and test sets under Bonferroni correction ($p < 7 \times 10^{-8}$) (**Figure 4A, Supplementary Table 8**). For GxE associations, significant effect sizes were consistent across specifications (correlation > 0.90), with the main deviation occurring under the minimal specification of age & sex, similar to E associations (**Figure 4B**). Additionally, we observe greater effect sizes with exposure categories of alcohol, smoking, exercise, and diet compared to deprivation indices (**Figure 4C**). We highlighted GxE associations such as a crossover GxE interaction between APOF and alcohol intake (**Figure 4D**).

Proteins Mediate Effects of Exposures to Disease Outcomes

We applied mediation analysis to evaluate whether proteins mediate the relationship between exposure and incident disease. To quantify the influence of the aggregate genetics and the exposome, we used the PGS and PXS for each protein, respectively. We applied g-computation³² to partition the indirect effects (mediation strength) of genetics and exposures. We used a Cox regression model to assess the indirect and direct effects of incident disease. We developed the GEM statistic, which is the log ratio of the indirect effects of exposures and genetics across all disease labels (See “Methods” for details). Briefly, ‘indirect effects’ are relationships that link exposures and genetics to disease outcomes via a protein mediator, while ‘direct effects’ are relationships that link exposures and genetics to disease outcomes independent of the mediator.

Summarizing across 270 disease indicators, we identified 156 proteins mediating a greater influence of lifestyle exposures on disease risk than genetics ($GEM > 0$) (**Figure 5A, Supplementary Figure 3A, Supplementary Table 9-10**). Examples include LEP ($GEM = 1.30$), IGFBP1 ($GEM = 0.53$), IGFBP2 ($GEM = 0.31$), and WFDC2 ($GEM = 0.36$). GDF15 and WFDC2 had significant hazard ratios for over 140 disease labels, reflecting their pleiotropic effects. Across the proteome, we found a positive Pearson correlation ($r = 0.14$, $p = 6e-13$) between the number of significant disease associations and GEM (**Supplementary Figure 3B**), suggesting that pleiotropic proteins – proteins influencing multiple disease phenotypes – are more likely to be influenced by lifestyle exposures than by common genetic variation.

We highlight specific examples of significant protein mediation across three diseases: type 2 diabetes, emphysema, and mental disorders due to alcohol usage. In all three diseases, plasma proteins yielded significant hazard ratios (HR) and demonstrated high discriminative value (c-index) in a Cox proportional hazards model (**Figure 5B**). For type 2 diabetes (T2D), we observe that certain proteins – such as IGFBP1, IGFBP2, FGF21, and CKB – demonstrate marginally stronger indirect effects from exposures, while proteins IL18R1 and IGSF9 are strongly influenced by genetics (**Figure 5C**). Combining evidence from univariate associations, we found that exercise behaviors such as strenuous sports, minutes of intense exercise, and days of general activity are associated with increased IGFBP1

expression leading to reduced risk for T2D. In contrast, diets high in meat intake (e.g. beef, poultry, and processed meats) are associated with lowered IGFBP2 expression leading to increased risk for T2D.

For emphysema, proteins such as CXCL17, LAMP3, WFDC2, and TNF were strongly associated with future disease risk, with some having roughly equal contributions from both genetic and exposure indirect effects. Specifically, current smoking status was positively associated with WFDC2, LAMP3, and CXCL17 which, when elevated, are associated with a higher risk for emphysema. These proteins serve as biomarkers for reduced lung function^{29,33,34}, linking smoking to an increased risk of emphysema and chronic obstructive pulmonary disease (**Figure 5D**).

For mental disorders via alcohol usage, we identified MAMDC4 and CEACAM16 as the most predictive proteins. Both have large genetic and exposure indirect effects and large partitioned R^2 with alcohol usage ($R^2 > 0.04$). Current alcohol use was associated with increased MAMDC4 and decreased CEACAM16 expression, both of which are linked to increased mental disorder risk (**Figure 5E**).

Validation of Exposomic Associations Across Independent Interventional Cohorts

To examine the robustness of our lifestyle-proteome disease associations, we compared our HEAP lifestyle findings with proteomic changes in interventional studies on exercise and GLP1 agonists. We utilized plasma proteomic data from the HERITAGE^{35,36} cohort (n = 654), a 20-week endurance training intervention, to validate HEAP exercise associations. Additionally, we incorporated serum proteomic data from 68-week GLP1 agonist treatment in STEP1 (n = 1133) and STEP2 (n = 595) randomized control trials (RCTs)³⁷⁻³⁹ to evaluate correlations between proteomic changes related to GLP1 agonist use with HEAP lifestyle associations. Both interventions help mitigate potential time-varying confounding unaccounted for in cross-sectional studies.

We correlated significant (adj.p < 0.05) protein associations in each intervention with maximally adjusted HEAP lifestyle associations. The UKB lifestyle factors strenuous sports (r = 0.81, adj.p = 0.02), swimming and cycling (r = 0.52, adj.p = 0.03), and duration of vigorous activity (r = 0.74, adj.p = 0.006) had positive correlations with HERITAGE, validating exercise associations we observed in the UKB

(**Figure 6A, Supplementary Figure 4**). Compared to the GLP1 agonist interventions in STEP1 and STEP2, the UKB lifestyle factors strenuous sports (STEP1: $r = 0.69$, $\text{adj.p} = 2.5\text{e-}3$; STEP2: $r = 0.76$, $\text{adj.p} = 8.5\text{e-}3$), swimming and cycling (STEP1: $r = 0.51$, $\text{adj.p} = 8.1\text{e-}5$; STEP2: $r = 0.54$, $\text{adj.p} = 1.6\text{e-}3$), fresh fruit intake (STEP1: $r = 0.51$, $\text{adj.p} = 9.9\text{e-}3$; STEP2: $r = 0.71$, $\text{adj.p} = 1.2\text{e-}3$), and never smoking (STEP1: $r = 0.35$, $\text{adj.p} = 8.1\text{e-}3$; STEP2: $r = 0.41$, $\text{adj.p} = 4.6\text{e-}3$) were positively correlated, implying ‘healthy’ lifestyle behaviors and GLP1 agonist effects are similar (**Figure 6A**). Additionally, former daily tobacco smokers – participants who quit or reduced smoking from daily usage – (STEP1: $r = 0.56$, $\text{adj.p} = 9.9\text{e-}3$; STEP2: $r = 0.69$, $\text{adj.p} = 1.1\text{e-}3$) were positively correlated with GLP1 agonist effects, suggesting that smoking cessation may induce proteomic changes similar to GLP1 agonist treatment (**Supplementary Figure 4**).

We highlight concordance between exposures and interventional studies. Frequency of vigorous activity in UKB and post exercise effects in HERITAGE showed similar direction of effects for key proteins like LEP and IGFBP1 (**Figure 6B**). Proteins MYOM2 and ACTN2, involved in muscle stabilization⁴⁰, exhibited increased expression in response to exercise. Fresh fruit intake and GLP1 agonist showed concordant effects on proteins like GUSB and DSG2, GUSB breaks down glycosaminoglycans in lysosomes⁴¹ while DSG2 are a component of desmosomes in epithelial cells and cardiomyocytes⁴² (**Figure 6C**). However, some proteins, like FABP4 and APOM, differed in direction between the UKB and interventional studies, even after several covariate adjustments, likely reflecting unmeasured confounding and bias inherent in the cross-sectional design of the UKB.

We assessed type 2 diabetes risk by triangulating protein mediation effects, exposure associations in the UKB, and interventions (HERITAGE, STEP1, STEP2). We generated a network of the top 12 proteins protective ($\text{HR} < 0.75$) or deleterious ($\text{HR} > 1.75$) for T2D, using GLP1 agonist effects to triangulate causal protein-disease relationships. GLP1 agonists decreased T2D risk factors like HbA1c (STEP1: -5.1 mmol/mol ; STEP2: -18.7 mmol/mol) and BMI (STEP1: -5.8 kg/m^2 ; STEP2: -2.6 kg/m^2) over a 68 week period^{37,38}. Protein expression protective against T2D increased with GLP1 agonists, and vice versa (**Figure 6D**). Exposures linked to “healthy” lifestyle choices generally align with GLP1 agonist

and HERITAGE effects. For example, vigorous exercise and GLP1 agonist effects increased IGFBP1 expression, decreasing T2D risk. Aggregate lifestyle exposure modulates proteins like IGFBP1 (GEM = 0.53) and IGFBP2 (GEM = 0.31) implicated in T2D pathophysiology, while proteins with strong genetic effects are effectively modulated by GLP1 agonists (**Figure 6D**). The network shows that exercise and GLP1 agonists may synergistically act through proteins like IGFBP1, MXRA8, and APOF, while exposures like alcohol and moderate beef intake act antagonistically with GLP1 agonists through proteins like IGFBP2, APOF, and ADIPOQ (**Figure 6D**).

Discussion

In this study, we constructed the HEAP resource and employed it to explore how the lifestyle exposome contributes to the human plasma proteome. Analyzing data from over 50,000 UK Biobank participants, we partitioned the variance of 2,686 plasma proteins into genetic, exposomic, and gene-by-environment (GxE) associations. Our findings demonstrate that while genetic factors and other demographic influences account for the majority of proteomic variation, lifestyle factors significantly modulate protein expression that are subsequently related to disease outcomes. We identified robust exposome-protein associations in the UK Biobank and replicated key associations across interventional cohorts. Additionally, we discovered polygenic GxE interactions reflecting the complex relationship between genetics and exposures in shaping the plasma proteome. Importantly, several proteins were found to link environmental exposures to disease, with examples related to type 2 diabetes, emphysema, and mental disorders highlighted. The HEAP framework and these initial findings provide a foundation for future studies aimed at identifying actionable biomarkers and causal links between the exposome, proteome, and phenome.

Our study adds to the growing literature on human plasma proteomic variation and its utility in medicine. Previous studies found that cis- and trans-pQTL variation account for a similar proportion of the plasma proteome as non-genetic factors (e.g. sex, BMI, and physiological factors)⁴³. Our study extends this by examining *cumulative* lifestyle factors through a polyexposure score^{10,18}, which aggregates

the explained variance from multiple lifestyle exposures. We demonstrate that, in some cases, these lifestyle factors account for as much variability in protein expression as genetic factors. We found that the connections between aggregated lifestyle exposure and disease outcomes can be explained by these proteins, providing potential biological mechanistic insight. Lastly, we postulate that proteins with low total explained variance are due to measurement error, protein concentrations near limit of detection, and unadjusted for time-varying confounding variables⁴⁴.

Additionally, prior studies have identified variance QTL and GxE effects at specific loci on the plasma proteome⁸. Our study advances this understanding by considering polygenic GxE effects on the exposome. We found few polygenic GxE signals that were consistent across a train-test split of the biobank cohort data. We attribute this to the sample size needed to establish appropriate power for interaction-based analyses compared to main effects (**Supplementary Figure 5**). However, despite this, we argue that identifying robust polygenic GxE effects can offer personalized insights into disease risk. For example, we found that alcohol intake frequency effects on APOF expression differs based on an individual's genetic contribution to APOF expression. Increased APOF expression is protective for T2D as it may inhibit lipid transfer events with LDL (i.e. inhibiting cholesteryl ester transfer protein activity)⁴⁵. Therefore, we hypothesize, people that fall within a certain polygenic score range for APOF are likely tolerable to alcohol intake frequency.

A key distinction between the genome and the exposome includes its time-varying nature and potential for confounding. While a subset of the lifestyle exposome can be tested in a randomized setting (e.g., physical activity or dietary interventions), others cannot (e.g., smoking). In contrast, genotypes are randomly distributed at birth⁴⁶. By integrating multiple modeling scenarios, dose-response relationships, mediation, and validation, we claim *HEAP* will be useful in the analytic armamentarium for biobank researchers to prioritize exposure factors in disease.

Throughout this study, we applied criteria⁴⁷ to strengthen our confidence in associations between lifestyle exposures and plasma proteins. We demonstrated the strength and scale of significant associations across the proteome, with standardized beta effects on average ranging from 0.1 - 0.28. We

found that these effects persist even with conditioning on indices of socioeconomic status, which contributed little to plasma proteome variation. To address potential confounding from demographics and comorbidities, we established the consistency of these associations across several model scenarios. We identified significant, monotonically increasing dose-response relationships based on the duration of exposures, like alcohol intake. We pinpointed tissue- and pathway- specific effects using GTEx³⁰ and Reactome³¹, providing biological plausibility for connections between lifestyle exposure and plasma proteins. Finally, we quantified how the impact of aggregate genetics and exposures on future disease risk is mediated by proteins.

We validated exposome-protein links and examined how exposome-protein-disease links align with interventional cohort data from HERITAGE³⁶ and STEP1/STEP2³⁹. Specifically, we compared the concordance of effects estimated in the UKB with these cohorts. We validated exercise associations with plasma proteomics in the HERITAGE cohort and explored effects of lifestyle exposures in relation to GLP1 agonist treatment in STEP1/STEP2 randomized controlled trials. We suggest that identifying shared effects between lifestyle exposures and GLP1 agonists could reveal synergistic and antagonistic interactions, ultimately guiding more effective drug usage. Additionally, GLP1 agonists affect a broad spectrum of proteins associated with T2D risk³⁹. For instance, targeting proteins with higher genetic liability ($GEM < 0$) effectively. This may reflect the strong efficacy of GLP1 agonist treatment on cardiometabolic traits and beyond⁴⁸ compared to lifestyle interventions. This triangulation of observational findings with interventional studies highlights HEAP's potential to identify biological mechanisms linking lifestyle exposures to clinical interventions. Moving forward, we advocate for applying HEAP to investigate the mechanisms of lifestyle exposures across other diseases, utilizing the power of plasma proteomic profiling.

Despite the comprehensive nature of our analysis, there are important limitations. For many exposures, the reliance on cross-sectional data from the UK Biobank limits our ability to establish causality. To mitigate this, we enhanced the generalizability of our findings by incorporating train-test splits, cross validation, and bootstrapping procedures throughout the HEAP framework. Future work

incorporating longitudinal and intervention-based studies with exposomic and plasma proteomic data will help address this limitation. Further, the UK Biobank has sample bias, with an overrepresentation of healthy and higher socioeconomic status individuals of European descent⁴⁹. Future studies should assess the effects of cumulative exposures in more diverse populations to improve comparability and inference of exposomic variability. Furthermore, lifestyle exposures are a small subset of the larger exposome which includes other potentially interacting dimensions such as chemical and microbiome¹. Future studies should assess the effects of exposures across other dimensions to derive a more complete picture of how exposures drive biological mechanisms.

Attaining a causal relationship between lifestyle exposures and disease is challenging. In observational settings, biases in sample collection, confounding, and model misspecification can all contribute to inaccurate representations. By examining omic modalities, such as plasma proteomics, we can explore the biological pathways that mediate these associations, providing a better understanding of the mechanisms through which lifestyle exposures contribute to disease progression. We present a novel approach to explain how lifestyle exposures can contribute to disease. HEAP serves as a new tool in the discovery armamentarium accumulating evidence linking exposures and disease. Considering future studies, we anticipate that HEAP will extend to other omic modalities, such as metabolomics, lipidomics, transcriptomics, with integration of relevant interventional studies and clinical trials to facilitate the creation of a comprehensive exposome-omic-disease database. Ultimately, our exposome-proteome atlas provides a resource for identifying biomarkers that can guide targeted health interventions, addressing the complex interactions between genetics, exposures, and disease.

Online Methods

UK Biobank Cohort

The UK Biobank (UKB)⁵⁰ contains 500k individuals with genetic, exposomic, and omics data. Individuals range from 40-70 years of age with genetic, health-record, exposomic, imaging, clinical, metabolomic, and proteomic measurements.

We used UKB plasma proteomics for this study. Approximately, 50k individuals contained measurements across 3000 protein biomarkers that ranged across disease phenotypes: cardiometabolic, neurological, inflammation-related, and oncology-related⁵⁰. Normalized protein expression (NPX) values were obtained from UKB which were built from selected quality control measures and normalization procedures that accounted for batches⁷.

We used exposomic measurements from the UKB from individual-level survey data. The measurements include lifestyle factors such as alcohol consumption, diet, exercise, smoking usage, internet usage, and vitamins. We incorporated neighborhood-level data of income and multiple deprivation indices which aggregate local income, employment, education, and housing to encompass potential latent features of lifestyle. We used imputed genotypes from UKB as previously described⁵¹ to generate genetic scores.

HERITAGE Cohort

We used plasma proteomic profiling from the HERITAGE cohort – which included 654 individuals of both European and African ancestry – to correlate findings with UKB exposure-protein associations³⁶. The HERITAGE cohort is a longitudinal intervention study, in which participants completed a 20-week endurance training program, exercising three times per week under supervision³⁵. Inclusion criteria required participants to be healthy (i.e. no cardiometabolic disease) and sedentary for three months prior to the exercise intervention. The study profiled 4,914 proteins using the SOMAscan platform and collected data at two time points: one before the program began and one 24 hours after the program ended³⁶. These measurements were recorded at a fasted and rested state. We utilized the

summarized statistics of relative changes in protein concentration from before to after the exercise regimen.

GLP1 Agonist Randomized Clinical Trials: STEP1 and STEP2 Cohorts

We used serum proteomic profiling from the STEP1 and STEP2 phase 3 randomized, double blind, placebo controlled clinical trials³⁹. STEP1 included participants with obesity, while STEP2 included participants with type 2 diabetes^{37,38}. Both trials had participants take 2.4 mg of semaglutide, a GLP1 receptor agonist, for 68 weeks to assess the safety and utility of semaglutide for weight loss and glycemic control in these respective populations. The fasting serum samples of 1956 individuals were profiled for ~6400 proteins using the SOMAscan assay v4.1 at baseline and after treatment³⁹. Summarized statistics of proteomic effects were utilized to capture the relative change in protein concentration from baseline to week 68 with semaglutide 2.4 mg treatment compared to placebo.

Preprocessing Exposomic Features

We preprocessed 135 exposomic measurements across multiple lifestyle exposure categories: 6 in alcohol features, 50 in weekly diet features, 38 in exercise frequency features, 16 in smoking features, 17 in vitamin features, 2 in internet usage features, and 6 in deprivation indices. These features included a mixture of continuous, ordinal, unordered categorical, and binary features from survey questionnaire data. Ordinals are handled with contrasts, while unordered categorical and binary features are one-hot (dummy) encoded. Exposomic measurements with at least 80% complete cases were considered for the analysis, and any participants with missing data were excluded.

Preprocessing Plasma Proteins

We utilized normalized protein expression accounting for batch effects and technical factors. For full details on quality control measures and repeatability of plasma protein expression, refer to Sun et al.,

2023. We excluded proteins that did not have a built genetic score and had greater than 80% missingness.

We considered complete cases for each protein in the regression analyses.

Covariate Selection

Several covariate specifications were performed (**Table 1**). These specifications were chosen based on prior knowledge of factors that influence protein expression variability. Common covariate specifications were used for the plasma proteins including age, sex, age², age * sex, age² * sex, BMI, and fasting time⁷. We utilized a polygenic score (PGS) so covariates related to population structure such as assessment center and the top 20 genetic PCs were added. Further adjustments involved common medications such as cholesterol lowering drugs, insulin, blood pressure drugs, and hormone replacement therapy. In total, we accounted for age, sex, body mass index (BMI), fasting time, population structure, and current medication usage. A batch effect covariate was not included for this analysis since the UKB proteomics were normalized with batch correction⁷.

Protein Polygenic Score Usage

We used polygenic scores (PGS) for each plasma protein in the UKB. A comprehensive study of genetic scores across the proteome, metabolome, and transcriptome has been conducted across multiple cohorts¹⁷. These polygenic scores, defined by calculated weights, were stored in the OmicsPred¹⁷ database which includes scores built for UKB participants of multi-ancestry. These polygenic scores were constructed using a Bayesian ridge regression on individual-level genotype data under a train-test split. The scores were validated on external cohorts, such as the INTERVAL⁵² cohort, which profiled the same proteins as the UKB. In our analysis, we treated the polygenic scores similar to a covariate to account for their potential influence on the relationship between exposomic factors and plasma proteins.

Protein Polyexposure Score Creation

We built polyexposure scores (PXS) using a lasso model where each normalized plasma protein (P) is regressed onto the PGS, every exposure feature ($E_{1:H}$), and selected covariates ($C_{1:K}$) is optimized: $P \sim PGS + E_{1:H} + PGS * E_{1:H} + C_{1:K}$. This model is evaluated under a 10-fold nested cross validation procedure with a 90/10 train-test split for each protein. The outer loop evaluated the generalizability of the constructed PXS using the R^2 metric across folds. The inner loop identified the optimal shrinkage parameter, λ_{min} , for the lasso model. To minimize data leakage, scaling and normalization parameters were calculated using the training data for each fold and applied to the test set.

The PXS is constructed using non-zero feature weights, β_i , from the lasso model. The PXS and GxE interaction (GIS) score are calculated as: $PXS = \sum_{i=1}^h \beta_i * E_i$ and $GIS = \sum_{i=1}^h \beta_i * PGS * E_i$. Out-of-fold (OOF) predictions were used to generate the finalized PXS score for every protein in the dataset, minimizing overfitting when the PXS is utilized in downstream analysis.

Variance Decomposition (Partitioned R^2)

We quantified the partitioned R^2 of G, E, and GxE using an ANOVA on the linear model: $P \sim PGS + PXS + GIS + C_{1:K}$, where the PXS and GIS are the scores built under the lasso procedure. The partitioned R^2 values reflected how much of the variance in each protein is explained by genetics (G), exposures (E), gene-by-environment (GxE), and covariates ($C_{1:K}$) providing a quantifiable measure of the independent and joint effects of genetics and the exposome.

Stability of the R^2 estimate was tested by constructing PXS and GIS scores under lasso models of several covariate specifications. From deriving the mean estimates of R^2 across covariate specifications we derived a classification of proteins driven by genetics ($R^2_{genetics} > R^2_{exposures}$) or exposures ($R^2_{genetics} < R^2_{exposures}$).

For variance decomposition of each exposure category, we recomputed the PXS using predefined categories (e.g. Exercise, Diet, Smoking, Alcohol, etc.). The PXS of a category is defined as $PXS_c = \sum_{i=1}^C \beta_i * E_i$, where C is the number of exposures assigned to the pre-defined category.

Univariate Exposomic and Polygenic GxE Associations

We assessed the effects of individual exposures via univariate E and polygenic GxE association tests to obtain effect sizes and significance of individual exposures. We constructed univariate associations under a 80/20 train-test split using a linear model for each protein: $P \sim PGS + E_i + PGS * E_i + C_{1:k}$. The beta for E_i is the effect of an exposure on a protein and the beta for $PGS * E_i$ is the polygenic GxE interaction effect. For all exposures, standardized betas and their respective p-value were cataloged. For multiple-hypothesis testing, a Bonferroni correction was applied based on the number of preprocessed exposomic features and proteins tested (see above Methods, “Preprocessing Exposomic Features”).

We declared an effect to be validated if we found an association with a significant p-value in both the train and test set. We checked the stability of estimates by running models under several covariate specifications. We performed Pearson correlation tests of validated effects observed across the specifications. We utilized linear mixed models to account for standard errors across specifications when estimating trends of validated effects for individual exposures. From this, we assessed the effect size, direction, significance, and pattern of exposure-protein associations.

Mediation Analysis

We quantified the extent to which proteins mediate the relationships between exposure, genetics, and disease. We assumed the following temporality for mediation analysis where genetics and exposures impact circulating proteins leading to development of disease. The genetic component is defined since birth and the exposure component is defined by past self-reported events recorded in surveys. The proteins are measured at baseline and incident (future) disease is captured via ICD10 codes following this protein measurement. Therefore, each protein acts as a mediator between the genetic and exposomic determinants of disease. We use mediation analysis to quantify the indirect and direct effects. Where indirect effects are relationships between an exposure/genetics and disease outcome that occur through

the mediator, protein. While direct effects are relationships that link exposures/genetics to the disease outcome independent of the mediator, protein.

Mediation Framework

We utilized the PGS and PXS for each protein to capture aggregate genetic and exposure effects. This enabled stable estimates for mediation, where a strong association between predictor and mediator reduces confounding and bias from unaccounted mediators⁵³. Mediation analysis involved two regression models: a mediator model and outcome model. We used a linear regression to define the mediator model: $P \sim \text{PGS} + \text{PXS} + C$. The mediator model regressed the protein, P, onto predictors, PGS and PXS, and covariates, C. We used a cox regression to capture the time-varying disease outcomes defined by: $D \sim P + \text{PGS} + \text{PXS} + C$. This outcome model regressed the disease outcome, D, onto P, PGS, PXS, and C. We ran these models under several covariate specifications to check for stability of estimates as we removed potential confounding. Additionally, we excluded disease codes when the number of cases is less than 100, to improve the stability and generalization of estimates obtained.

G-computation

We applied g-computation to obtain the indirect and direct effects^{32,54}. In g-computation, simulated counterfactual scenarios are used to estimate causal effects⁵³. This potential outcomes framework provides mediation strength estimates for PGS and PXS. The procedure first fits the mediator and outcome models under the data. Next, counterfactuals for the mediator are simulated by varying the predictor, and outcomes are then simulated based on these changes in the predictor and mediator. The natural indirect effect (NIE, referred to as mediation strength) measures how changes in the mediator, protein, resulting from ± 1 standard deviation (SD) shifts in the predictor (PGS or PXS) affect the disease outcome. The natural direct effect (NDE) measures the impact of ± 1 SD shifts in the predictor (PGS or PXS) on the disease outcome while holding the mediator, protein, constant. We obtained bootstrapped confidence intervals for the NIE and NDE.

GEM statistic

We developed the GEM statistic to quantify a protein's overall influence by genetics and exposures across all diseases. Using the g-computation method, we calculated both the indirect and direct effects for genetics and exposures. For each protein, we obtained the indirect effects of genetics (I_g) and exposures (I_e) across 270 disease labels. We calculated the GEM statistic as the log of the slope (I_e/I_g) from regressing the indirect effects of exposures and genetics. We classified proteins using GEM as follows: $GEM > 0$ if exposures dominate, $GEM < 0$ if genetics dominate, and $GEM = 0$ if both contribute equally.

To verify this logic, we applied the Baron and Kenny concept to mediation⁵⁵, which states under a system of two linear models, the indirect effect (IE) is the product of coefficients between the predictor (B_p) in the mediator model and the mediator (B_m) in the outcome model, $IE = B_p * B_m$. So for one disease, if the indirect effects between genetics and exposures through a protein mediator is exactly the same $B_{p_e} * B_m = B_{p_g} * B_m$, then the ratio of their coefficients (B_{p_e} / B_{p_g}) is 1. Across multiple diseases, the value of B_m differs based on the strength of protein-disease association. Therefore, we chose to use the slope from the regression of the indirect effects ($I_e = a_0 + b_1 * I_g + \epsilon$) to compute GEM, as the ratio could lead to inaccurate estimates when the protein mediator effects are near 0.

Enrichment Analysis

We applied enrichment analysis to identify pathways and tissues impacted by the lifestyle exposome. We used rank ordered significant exposure-protein association effect sizes (i.e. t-statistic) to find biological pathways and tissues upregulated or downregulated for each exposure (e.g. alcohol, diet, exercise, smoking, etc.). We used the gene set enrichment analysis, GSEA¹⁹, approach with parameters, $minGGsize = 10$, $maxGGsize = 500$, and FDR correction to obtain pathways and tissues significantly enriched ($FDR < 0.05$) in exposure-protein associations. We utilized resources GO^{20,56} and Reactome^{31,57} for pathway enrichment, and GTEx³⁰ for tissue enrichment.

Tissue-specific Exposomic Effects

We defined tissue-specific protein sets to tissue enrichment analysis using GTEx. GTEx is a resource measuring gene expression from 838 deceased donors with genotype information across 54 tissues, to assign plasma proteins to specific tissues³⁰. Using the gene expression defined by GTEx, we classified a gene in a tissue set if the fold change for a specific tissue is 4 times that of the mean expression of other tissues. This was determined through referencing ways plasma proteins are defined to organ sets^{58,59}. We extended this analysis to retain tissue-specific results and to balance out organs that have multiple tissues, such as the brain which is divided into multiple brain regions.

Observational and Interventional Cohort Comparisons

We compared the effect sizes of exposure-protein associations with circulating proteomics of intervention studies. Exercise intervention study HERITAGE^{35,36} and GLP1 agonist RCTs STEP1 and STEP2^{37–39}. For the HERITAGE study, reported log fold change (logFC) differences between pre- vs post-training proteomic signatures were used for comparison with UKB effect sizes. Approximately 250 plasma proteins were able to be cross-referenced between the HERITAGE SomaScan and UKB Olink platforms. For the STEP1 and STEP2 RCTs, the summary statistics of the serum protein longitudinal effects from both the STEP1 and STEP2 trials were used for comparison with UKB effect sizes. Approximately 2.5k proteins were cross-referenced between the STEP1/2 SomaScan and UKB Olink platforms.

We correlated the exposure-protein effect sizes found in UKB to effects observed in each intervention study. We calculated Pearson correlations and p-values adjusted for FDR using the Benjamini-Hochberg procedure for each intervention separately. We created an initial network with significant proteins protective ($HR < 0.75$) or deleterious ($HR > 1.75$) for a disease. Using associations from exposures in the UKB and interventions (i.e. exercise and GLP1 agonist), we pruned this network down to the significant protein links to exposures, interventions, and disease. We incorporated GEM

scores for each protein into the network to describe when a protein exhibits greater exposomic influence than genetic influence. We triangulated links between exposures, protein, and disease from visualizing these associations across cohorts.

Power Analysis and Curves

We created power curves to assess the ability to replicate findings across cohorts. We used the range of partitioned R^2 for each exposure association and the amount of covariates utilized to define an appropriate effect size to assess the sample size at which 80% power is reached.

Coding Software and Availability

We used the R environment, R 4.2.1, and the following R packages `Glmnet`⁶⁰, `Coxph`⁶¹, `ClusterProfiler`⁶² to perform our analysis. Code used to generate all results and visualizations for this paper are included here <https://github.com/shakson-isaac/HEAP>. Code used to generate the website and visualizations are included here <https://github.com/shakson-isaac/HEAPweb>.

Resource

We created a resource containing significant associations generated by HEAP, hosted at <https://heap.bio>. This resource provides a comprehensive overview of HEAP, including summary statistics and visualizations, all presented in a dynamic, interactive, and user-friendly format.

Data Availability

UK Biobank data is available for researchers through <https://www.ukbiobank.ac.uk/> (Project ID 52887). Summarized statistics and information on data access to proteomics in the HERITAGE, STEP1, and STEP2 cohorts are described in their respective papers^{36,39}. We will submit weights for the protein polyexposure scores after peer review.

Author Contributions

S.I. was responsible for the concept, statistical design, undertaking, data analysis, manuscript construction, and editing. R.E., A.G., V.M., and M.U. were responsible for statistical design and manuscript construction. C.J.P. was responsible for the concept, statistical design, undertaking, data analysis, manuscript construction, and editing oversight for all sections.

Acknowledgements

This work was supported by grants from the National Institutes of Environmental Health Sciences (NIEHS) R01ES032470, National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) R01DK137993. National Institutes of Environmental Health Sciences (NIEHS) U24ES036819, Advanced Research Projects Agency for Health ARPA-H D24AC00345.

References

1. Vermeulen, R., Schymanski, E. L., Barabási, A.-L. & Miller, G. W. The exposome and health: Where chemistry meets biology. *Science* **367**, 392–396 (2020).
2. Nyberg, S. T. *et al.* Association of healthy lifestyle with years lived without major chronic diseases. *JAMA Intern. Med.* **180**, 760–768 (2020).
3. Anderson, N. L. & Anderson, N. G. The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteomics* **1**, 845–867 (2002).
4. Deng, Y.-T. *et al.* Atlas of the plasma proteome in health and disease in 53,026 adults. *Cell* **188**, 253–271.e7 (2025).
5. Gadd, D. A. *et al.* Blood protein assessment of leading incident diseases and mortality in the UK Biobank. *Nat. Aging* **4**, 939–948 (2024).
6. Carrasco-Zanini, J. *et al.* Proteomic signatures improve risk prediction for common and rare diseases. *Nat. Med.* **30**, 2489–2498 (2024).
7. Sun, B. B. *et al.* Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* **622**, 329–338 (2023).
8. Hillary, R. F. *et al.* Systematic discovery of gene-environment interactions underlying the human plasma proteome in UK Biobank. *Nat. Commun.* **15**, 7346 (2024).
9. Patel, C. J., Bhattacharya, J. & Butte, A. J. An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. *PLoS One* **5**, e10746 (2010).
10. He, Y. *et al.* Comparisons of polyexposure, polygenic, and clinical risk scores in risk prediction of type 2 diabetes. *Diabetes Care* **44**, 935–943 (2021).
11. Maitre, L. *et al.* Multi-omics signatures of the human early life exposome. *Nat. Commun.* **13**, 7024 (2022).
12. Perry, A. S. *et al.* Proteomics, human environmental exposure, and cardiometabolic risk. *Circ. Res.* **135**, 138–154 (2024).

13. Ioannidis, J. P. A. Why most published research findings are false. *PLoS Med.* **2**, e124 (2005).
14. Patel, C. J. & Ioannidis, J. P. A. Studying the elusive environment in large scale. *JAMA* **311**, 2173 (2014).
15. Manrai, A. K., Ioannidis, J. P. A. & Patel, C. J. Signals among signals: Prioritizing nongenetic associations in massive data sets. *Am. J. Epidemiol.* **188**, 846–850 (2019).
16. Khoury, M. J. & Ioannidis, J. P. A. Big data meets public health. *Science* **346**, 1054–1055 (2014).
17. Xu, Y. *et al.* An atlas of genetic scores to predict multi-omic traits. *Nature* **616**, 123–131 (2023).
18. He, Y. *et al.* Prediction and stratification of longitudinal risk for chronic obstructive pulmonary disease across smoking behaviors. *Nat. Commun.* **14**, 8297 (2023).
19. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).
20. Gene Ontology Consortium *et al.* The Gene Ontology knowledgebase in 2023. *Genetics* **224**, (2023).
21. Friedman, J. M. Leptin and the endocrine control of energy balance. *Nat. Metab.* **1**, 754–764 (2019).
22. Furuhashi, M., Saitoh, S., Shimamoto, K. & Miura, T. Fatty acid-binding protein 4 (FABP4): Pathophysiological insights and potent clinical biomarker of metabolic and cardiovascular diseases. *Clin. Med. Insights Cardiol.* **8**, 23–33 (2014).
23. Groarke, J. D. *et al.* Ponesemab for the treatment of cancer cachexia. *N. Engl. J. Med.* **391**, 2291–2303 (2024).
24. Wang, D. *et al.* GDF15: emerging biology and therapeutic applications for obesity and cardiometabolic disease. *Nat. Rev. Endocrinol.* **17**, 592–607 (2021).
25. Prieto, M. C., Gonzalez, A. A., Visniauskas, B. & Navar, L. G. The evolving complexity of the collecting duct renin-angiotensin system in hypertension. *Nat. Rev. Nephrol.* **17**, 481–492 (2021).
26. Zheng, J. *et al.* Carcinoembryonic antigen-related cell adhesion molecule 16 interacts with alpha-tectorin and is mutated in autosomal dominant hearing loss (DFNA4). *Proc. Natl. Acad. Sci. U. S. A.* **108**, 4218–4223 (2011).
27. Shibuya, A. *et al.* Fc alpha/mu receptor mediates endocytosis of IgM-coated microbes. *Nat.*

- Immunol.* **1**, 441–446 (2000).
28. Aggarwal, S., Dabla, P. K. & Arora, S. Prostate: An epithelial sodium channel regulator. *J. Biomark.* **2013**, 179864 (2013).
 29. Burkhardt, A. M. *et al.* CXCL17 is a major chemotactic factor for lung macrophages. *J. Immunol.* **193**, 1468–1474 (2014).
 30. GTEx Consortium *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
 31. Milacic, M. *et al.* The reactome pathway knowledgebase 2024. *Nucleic Acids Res.* **52**, D672–D678 (2024).
 32. Wang, A. & Arah, O. A. G-computation demonstration in causal mediation analysis. *Eur. J. Epidemiol.* **30**, 1119–1127 (2015).
 33. Dougherty, G. W. *et al.* Recessively inherited deficiency of secreted WFDC2 (HE4) causes nasal polyposis and bronchiectasis. *Am. J. Respir. Crit. Care Med.* **210**, 63–76 (2024).
 34. Lunding, L. P. *et al.* LAMP3 deficiency affects surfactant homeostasis in mice. *PLoS Genet.* **17**, e1009619 (2021).
 35. Sarzynski, M. A. *et al.* The HERITAGE Family Study: A review of the effects of exercise training on cardiometabolic health, with insights into molecular transducers. *Med. Sci. Sports Exerc.* **54**, S1–S43 (2022).
 36. Robbins, J. M. *et al.* Plasma proteomic changes in response to exercise training are associated with cardiorespiratory fitness adaptations. *JCI Insight* **8**, (2023).
 37. Wilding, J. P. H. *et al.* Once-weekly semaglutide in adults with overweight or obesity. *N. Engl. J. Med.* **384**, 989–1002 (2021).
 38. Davies, M. *et al.* Semaglutide 2·4 mg once a week in adults with overweight or obesity, and type 2 diabetes (STEP 2): a randomised, double-blind, double-dummy, placebo-controlled, phase 3 trial. *Lancet* **397**, 971–984 (2021).
 39. Maretty, L. *et al.* Proteomic changes upon treatment with semaglutide in individuals with obesity.

- Nat. Med.* **31**, 267–277 (2025).
40. Chiu, C. *et al.* Mutations in alpha-actinin-2 cause hypertrophic cardiomyopathy: a genome-wide analysis. *J. Am. Coll. Cardiol.* **55**, 1127–1135 (2010).
 41. Naz, H. *et al.* Human β -glucuronidase: structure, function, and application in enzyme replacement therapy. *Rejuvenation Res.* **16**, 352–363 (2013).
 42. Schlipp, A. *et al.* Desmoglein-2 interaction is crucial for cardiomyocyte cohesion and function. *Cardiovasc. Res.* **104**, 245–257 (2014).
 43. Carrasco-Zanini, J. *et al.* Mapping biological influences on the human plasma proteome beyond the genome. *Nat. Metab.* **6**, 2010–2023 (2024).
 44. Eldjarn, G. H. *et al.* Large-scale plasma proteomics comparisons through genetics and disease associations. *Nature* **622**, 348–358 (2023).
 45. Morton, R. E., Liu, Y. & Izem, L. ApoF knockdown increases cholesteryl ester transfer to LDL and impairs cholesterol clearance in fat-fed hamsters. *J. Lipid Res.* **60**, 1868–1879 (2019).
 46. Ioannidis, J. P. A., Loy, E. Y., Poulton, R. & Chia, K. S. Researching genetic versus nongenetic determinants of disease: a comparison and proposed unification. *Sci. Transl. Med.* **1**, 7ps8 (2009).
 47. Hill, A. B. The environment and disease: Association or causation? *Proc. R. Soc. Med.* **58**, 295–300 (1965).
 48. Xie, Y., Choi, T. & Al-Aly, Z. Mapping the effectiveness and risks of GLP-1 receptor agonists. *Nat. Med.* **31**, 951–962 (2025).
 49. Fry, A. *et al.* Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).
 50. 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).
 51. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
 52. Moore, C. *et al.* The INTERVAL trial to determine whether intervals between blood donations can be

- safely and acceptably decreased to optimise blood supply: study protocol for a randomised controlled trial. *Trials* **15**, 363 (2014).
53. Imai, K., Keele, L. & Yamamoto, T. Identification, inference and sensitivity analysis for causal mediation effects. *Stat. Sci.* **25**, 51–71 (2010).
 54. Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66**, 688–701 (1974).
 55. Baron, R. M. & Kenny, D. A. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* **51**, 1173–1182 (1986).
 56. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
 57. Joshi-Tope, G. *et al.* Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.* **33**, D428–32 (2005).
 58. Oh, H. S.-H. *et al.* Organ aging signatures in the plasma proteome track health and disease. *Nature* **624**, 164–172 (2023).
 59. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
 60. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
 61. Therneau, T. M. & Grambsch, P. M. *Modeling Survival Data: Extending the Cox Model*. (Springer Science & Business Media, 2000).
 62. Xu, S. *et al.* Using clusterProfiler to characterize multiomics data. *Nat. Protoc.* **19**, 3292–3320 (2024).

Tables

Table 1. Model Covariate Specifications Applied in UK Biobank.

Label	Covariate Specification
Model 1	Age, Sex
Model 2	Age, Sex, BMI, Fasting Time
Model 3	Age, Sex, Age * Sex, Age ² , Age ² * Sex, BMI, Fasting Time, Assessment Center, 20 Genetic PCs
Model 4	Age, Sex, Age * Sex, Age ² , Age ² * Sex, BMI, Fasting Time, Assessment Center, 20 Genetic PCs, Medications (Blood Pressure, Hormone Replacement, Insulin, Cholesterol Lowering, Contraceptives)
Model 5	Age, Sex, Age * Sex, Age ² , Age ² * Sex, BMI, Fasting Time, Assessment Center, 40 Genetic PCs, Medications (Blood Pressure, Hormone Replacement, Insulin, Cholesterol Lowering, Contraceptives)
Model 6	Age, Sex, Age * Sex, Age ² , Age ² * Sex, BMI, Fasting Time, Assessment Center, 40 Genetic PCs, Medications (Blood Pressure, Hormone Replacement, Insulin, Cholesterol Lowering, Contraceptives), Multiple Deprivation Indices

Figures

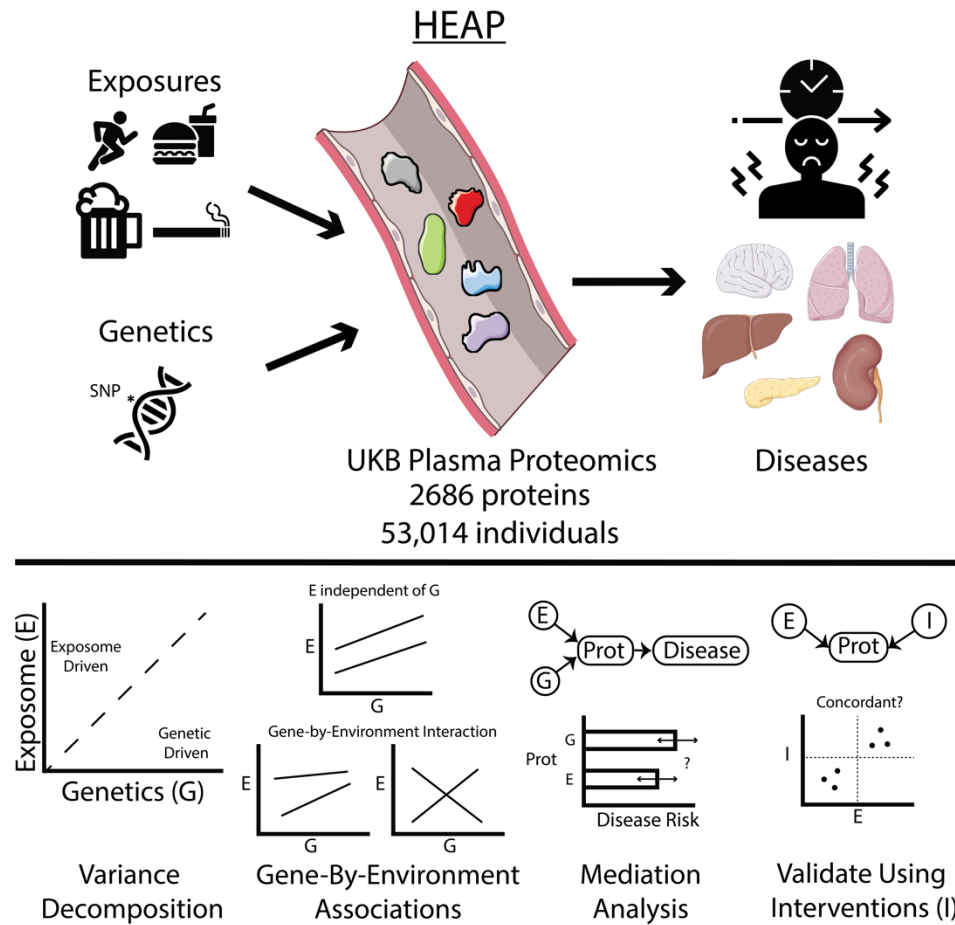


Figure 1. Overview of *HEAP* framework. *HEAP* characterizes how lifestyle exposures and genetics influence plasma proteins relevant to a wide range of disease outcomes. *HEAP* achieves this through four key modules: variance decomposition, gene-by-environment association, mediation analysis, and validation using interventions.

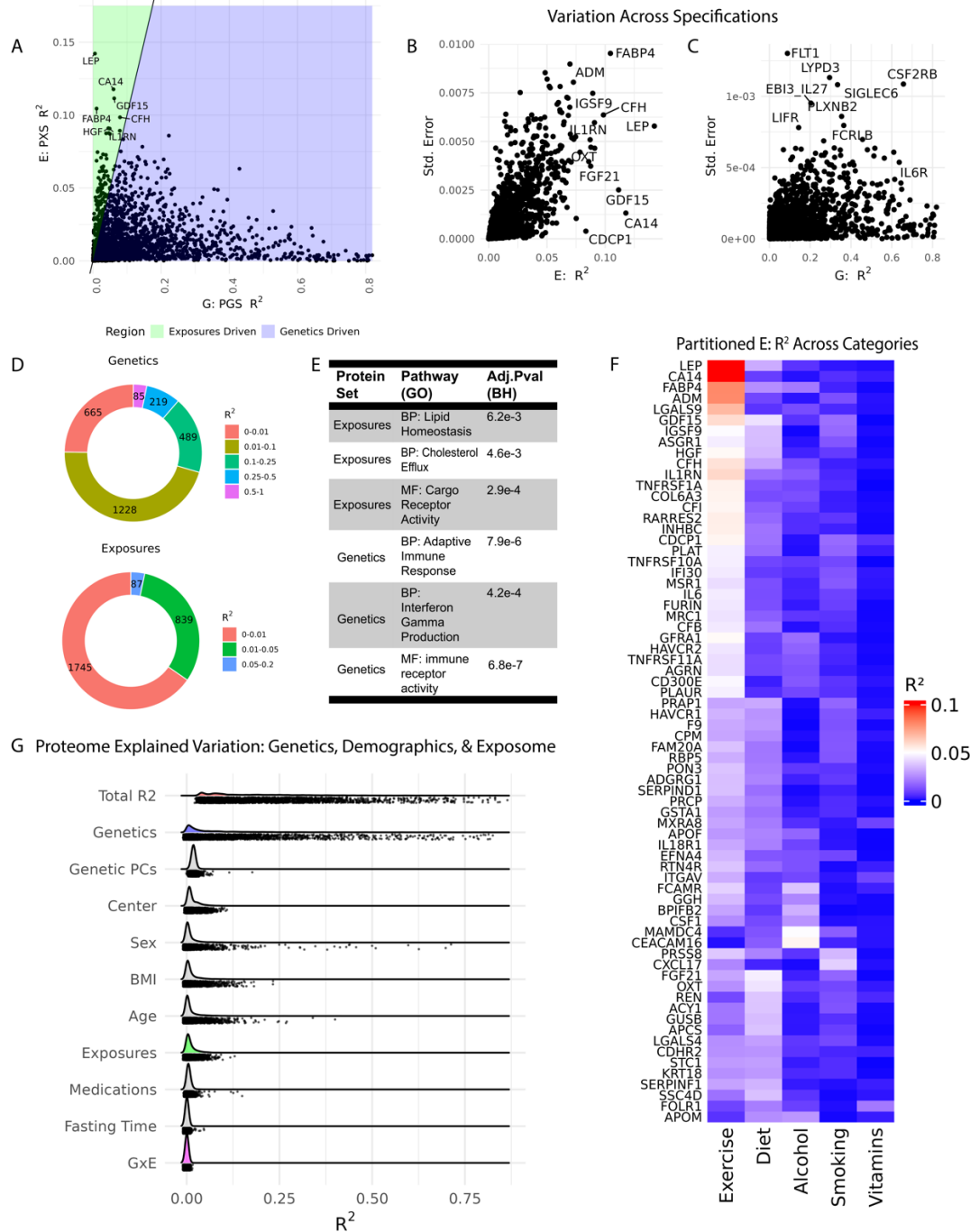


Figure 2. Spectrum of Genetic vs Exposure Driven Proteins. **A.)** Scatter plot of mean PGS and PXS test set R^2 for each plasma protein across specifications with $R_G^2 = R_E^2$ line separating exposome vs genetics driven proteins. **B-C.)** Scatter plots of variation of R_G^2 and R_E^2 across all covariate specifications with standard error shown. **D.)** Pie charts show distribution of PGS and PXS R^2 for each protein. **E.)** Table of significant GO pathways via GSEA with ranked protein lists by exposures (R_E^2) and genetics (R_G^2). **F.)** Heatmap of partitioned R^2 across exposure categories with row hierarchical clustering. **G.)** Distributions and individual points of explained variation, R^2 , across genetics, exposome, and covariates used.

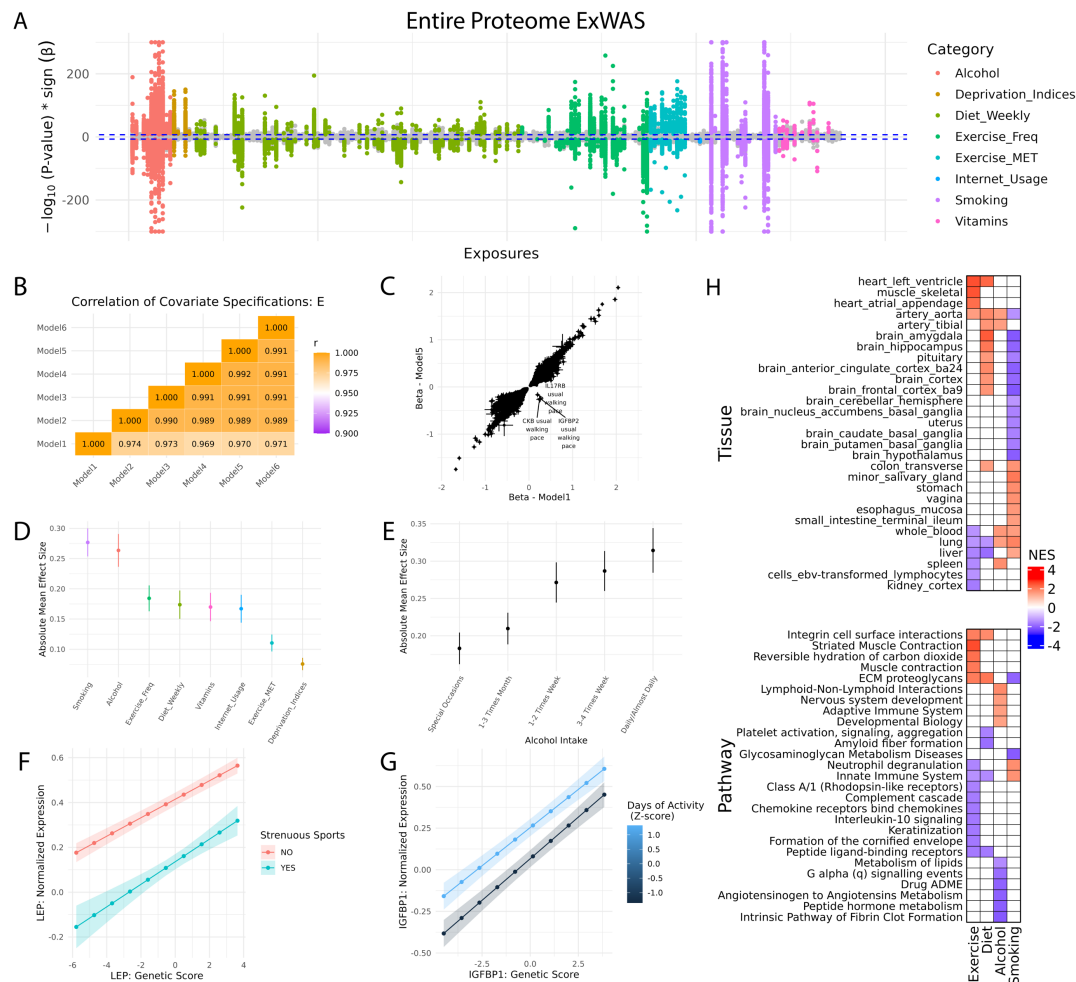


Figure 3. Proteome Exposomic Associations. A.) Miami plot of exposure-protein associations across all exposure categories tested using maximal specification. Horizontal dotted line notes the Bonferroni correction threshold ($p\text{-val} < 7 \times 10^{-8}$). Colored dots indicate successful replication across the 80/20 train/test split. **B.)** Pearson correlations of significant associations across covariate specifications. **C.)** Scatter plot with Beta estimates of exposures across minimal specification and maximal specification with 95% confidence interval error bars highlighting flipped associations. **D-E.)** Absolute mean effect sizes across covariate specifications with 95% confidence intervals using pooled standard errors for each **D.)** exposure category and **E.)** alcohol intake frequency. **F-G.)** Exposomic associations conditionally independent of genetics with 95% confidence intervals. **F.)** LEP and Strenuous Sports **G.)** IGFBP1 and summed days of physical activity. **H.)** Significant ($FDR < 0.05$) tissue and pathway specific effects of representative exposure categories.

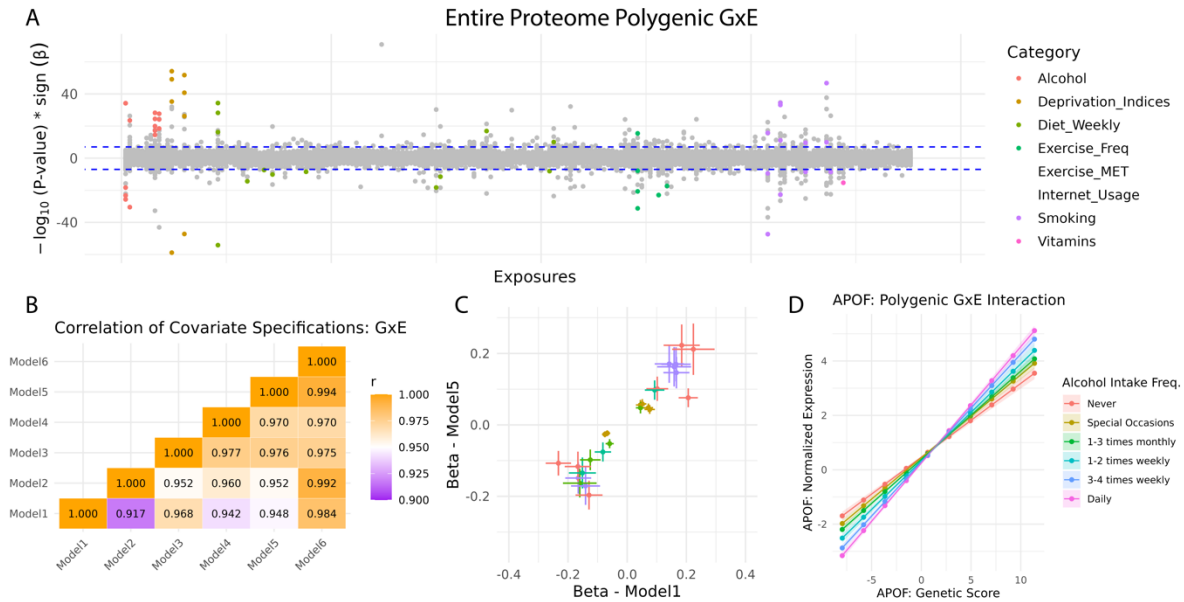


Figure 4. Proteome Polygenic Gene-by-Environment (GxE) Associations. A.) Miami plot of polygenic GxE protein associations across all exposure categories tested using maximal specification. Horizontal dotted line notes the Bonferroni correction threshold ($p\text{-val} < 7 \times 10^{-8}$). Colored dots indicate successful replication across the 80/20 train/test split. **B.)** Pearson correlations of significant associations across covariate specifications. **C.)** Scatter plot with Beta estimates of GxE associations across minimal specification and maximal specification with 95% confidence interval error bars colored by exposure category. **D.)** Polygenic GxE plot of Alcohol intake frequency and polygenic score of APOF with 95% confidence intervals shown.

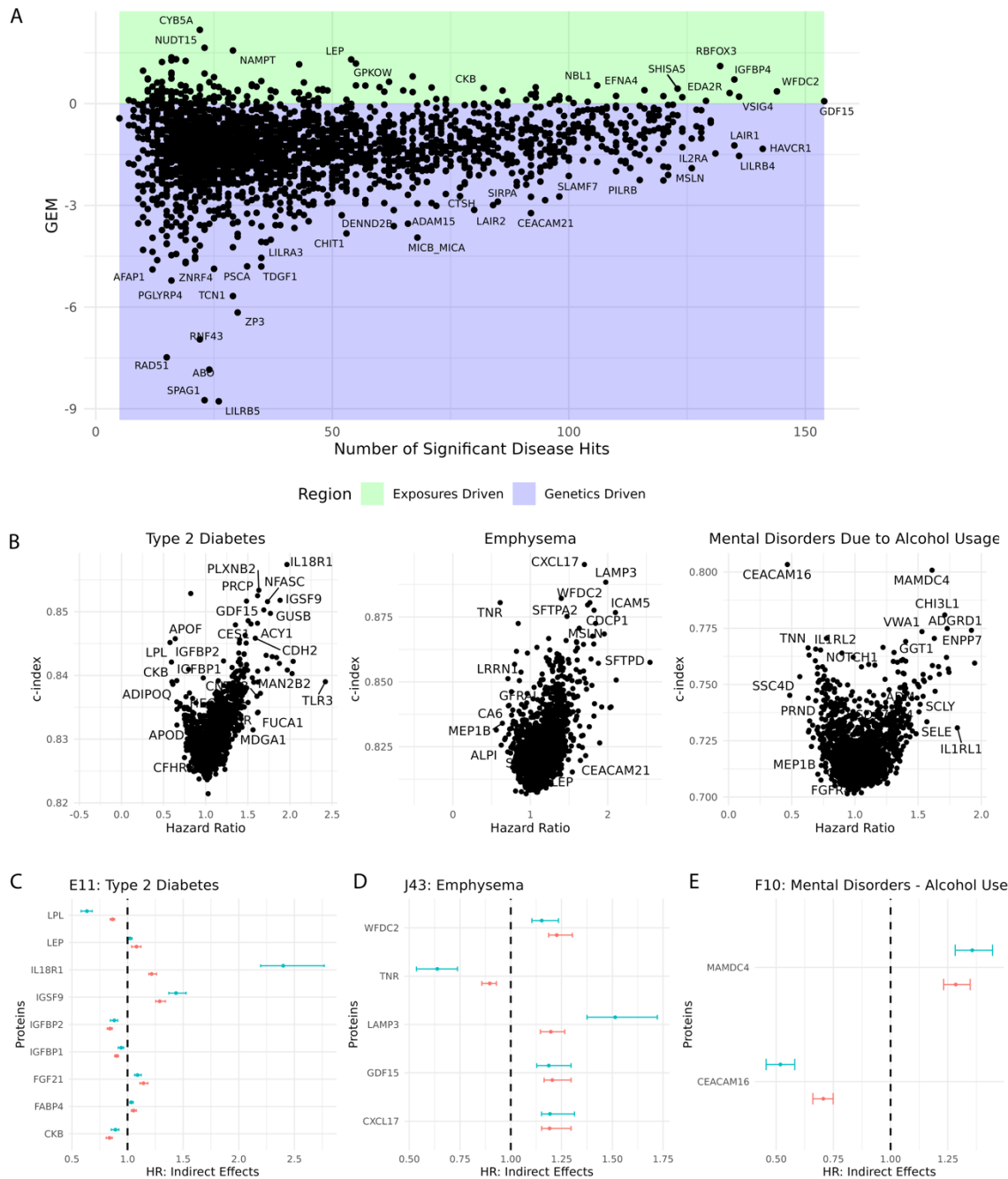


Figure 5. Proteins Mediate Effects of Genetics and Exposures on Disease Risk. A.) Scatter plot of GEM statistic and number of significant disease hits for each protein with GEM = 0 line separating exposome vs genetic driven protein mediation using maximal specification. **B.)** Volcano plots showing proteome-wide associations to selected diseases: Type 2 Diabetes, Emphysema, and Mental Disorders due to Alcohol. Hazard ratio and c-index of the protein is reported for the maximal specification adjustment. **C-E.)** Plots denoting the exposure- and genetic-mediated hazard ratios with bootstrapped 95% confidence intervals from 1000 resamples across **C.)** Type 2 Diabetes, **D.)** Emphysema, and **E.)** Mental Disorders Due to Alcohol Use.

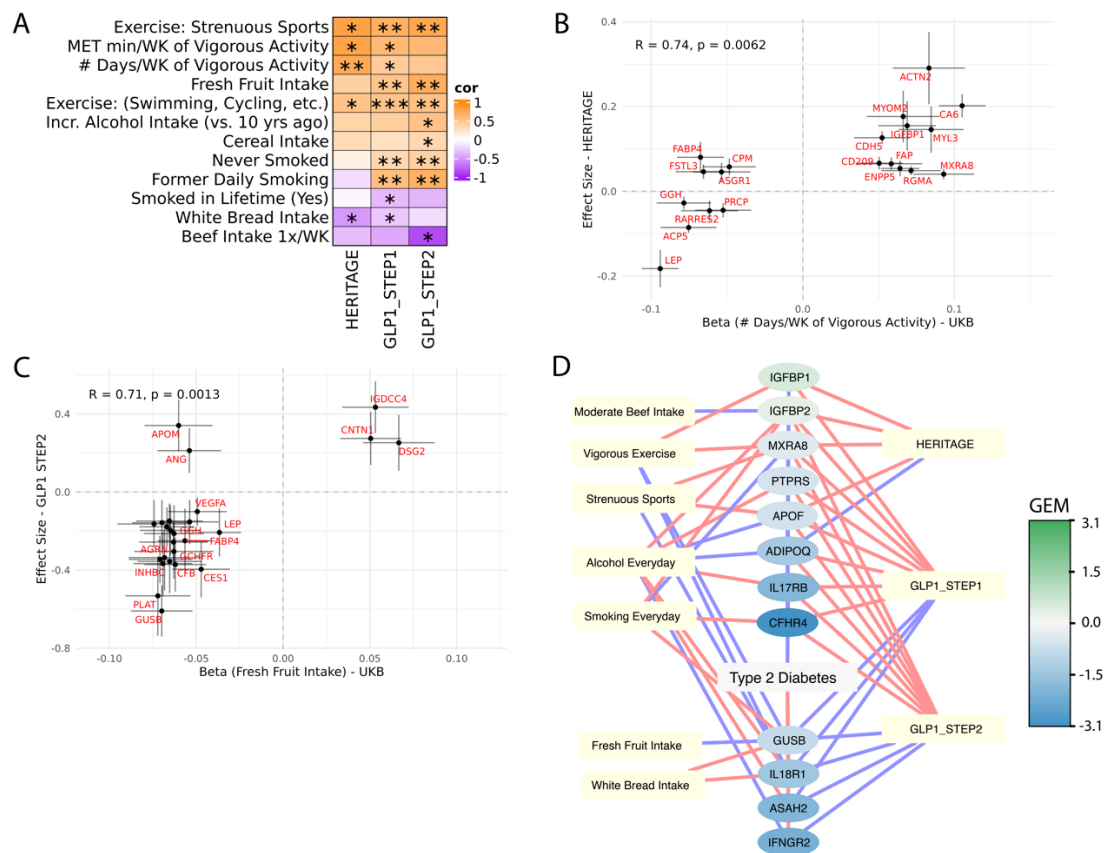


Figure 6. Replication of Exposomic Associations with Interventional Cohorts. A.) Heatmap of Pearson correlations between significant proteomic signatures for HEAP (UKB) under maximal specification, HERITAGE exercise intervention study, and GLP1 agonist STEP1/STEP2 trials (FDR < 0.05). Significant correlations (FDR < 0.05) are denoted by asterisks marked as: “*” adj.p < 0.05, “**” adj.p < 0.01, “***” adj.p < 0.001. **B-C.)** Scatter plot of HEAP associations with interventions: HERITAGE or STEP1/2. Pearson correlation and Benjamini-Hochberg p-value shown. **B.)** UKB days of vigorous activity vs HERITAGE. **C.)** UKB fresh fruit intake vs GLP1 agonist. **D.)** Network of associations between exposures, interventions, proteins, and type 2 diabetes. Red links are positive associations, blue links are negative associations, and green-blue gradient filled nodes indicate the GEM score of a protein.

Supplementary Table Legends

Supplementary Table 1. R^2 of genetics, exposures, and gene-by-environment interactions across the UKB plasma proteome.

Supplementary Table 2. Summary statistics of GO pathway enrichment analysis of genetic and exposure ranked protein lists.

Supplementary Table 3. R^2 of lifestyle exposome categories (e.g. exercise, diet, smoking, exercise, etc.).

Supplementary Table 4. R^2 of covariates listed in model specifications.

Supplementary Table 5. Summary statistics of lifestyle exposure associations with the plasma proteome from model specification 5.

Supplementary Table 6. Summary statistics of GTEX-defined tissue enrichment analysis across all lifestyle exposures.

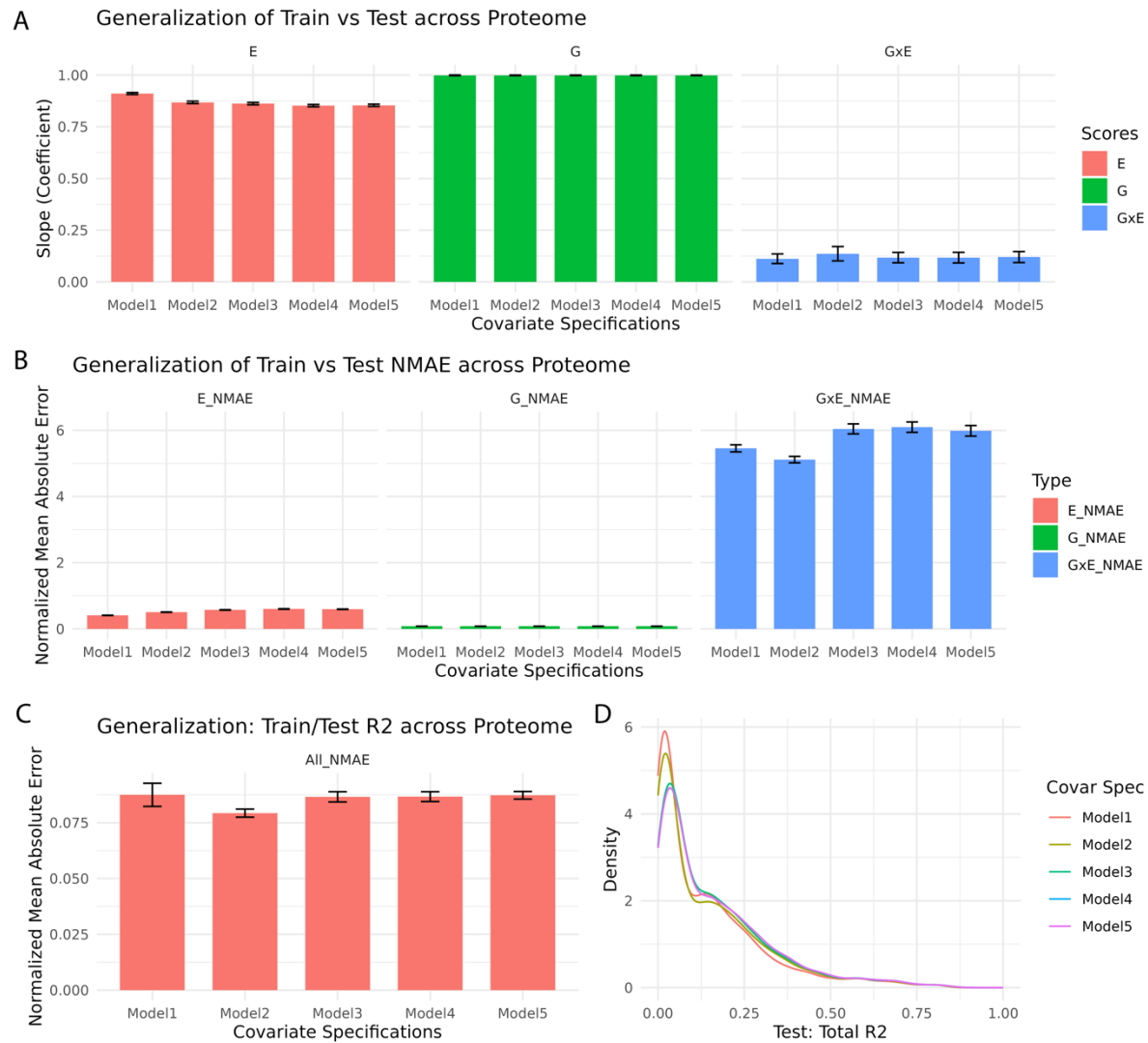
Supplementary Table 7. Summary statistics of Reactome pathway enrichment analysis across all lifestyle exposures.

Supplementary Table 8. Summary statistics of polygenic gene-by-environment associations with the plasma proteome from model specification 5.

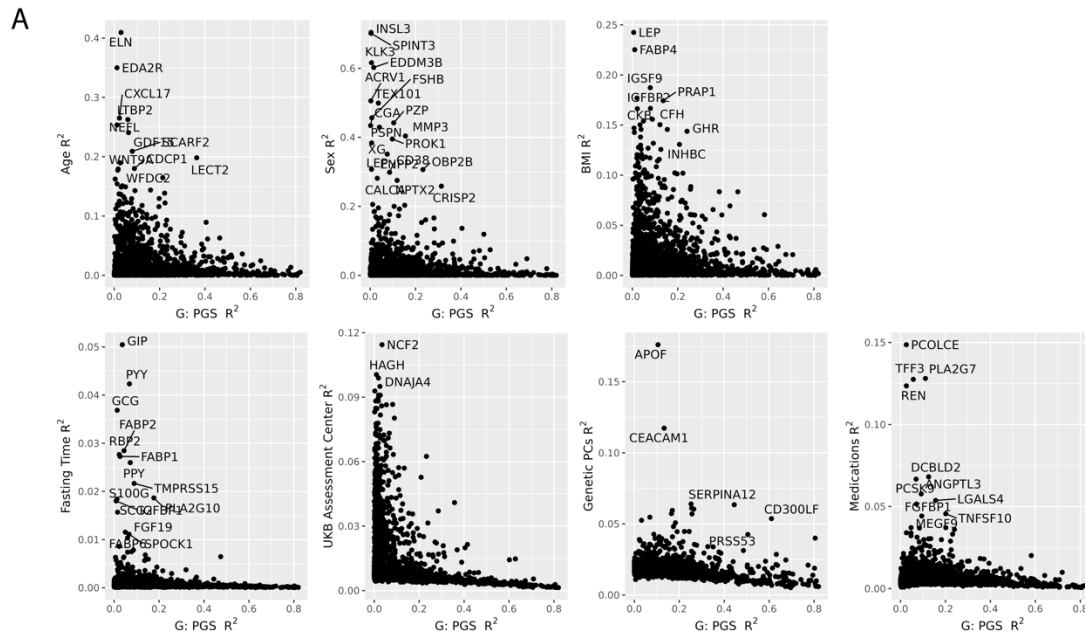
Supplementary Table 9. GEM statistics of plasma proteins of the maximal model specification.

Supplementary Table 10. Summary statistics of g-computation mediation of plasma proteins of the maximal model specification.

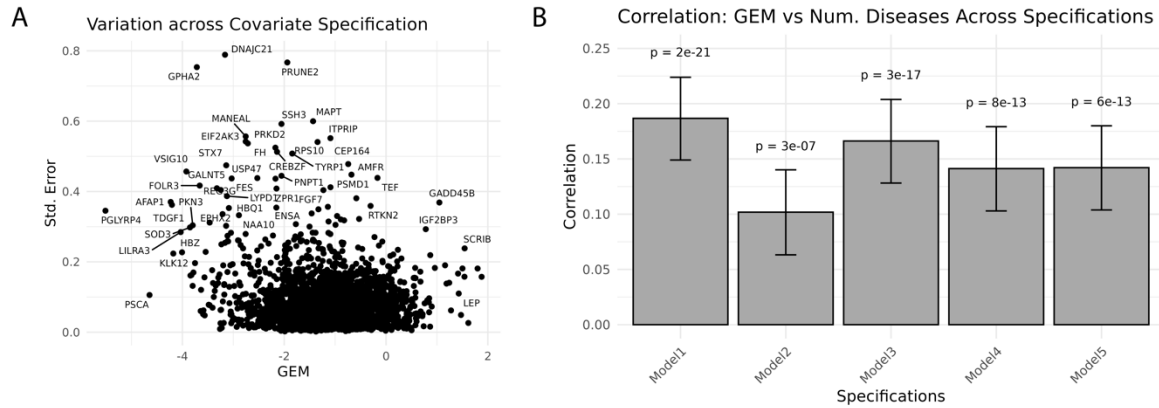
Supplementary Figures



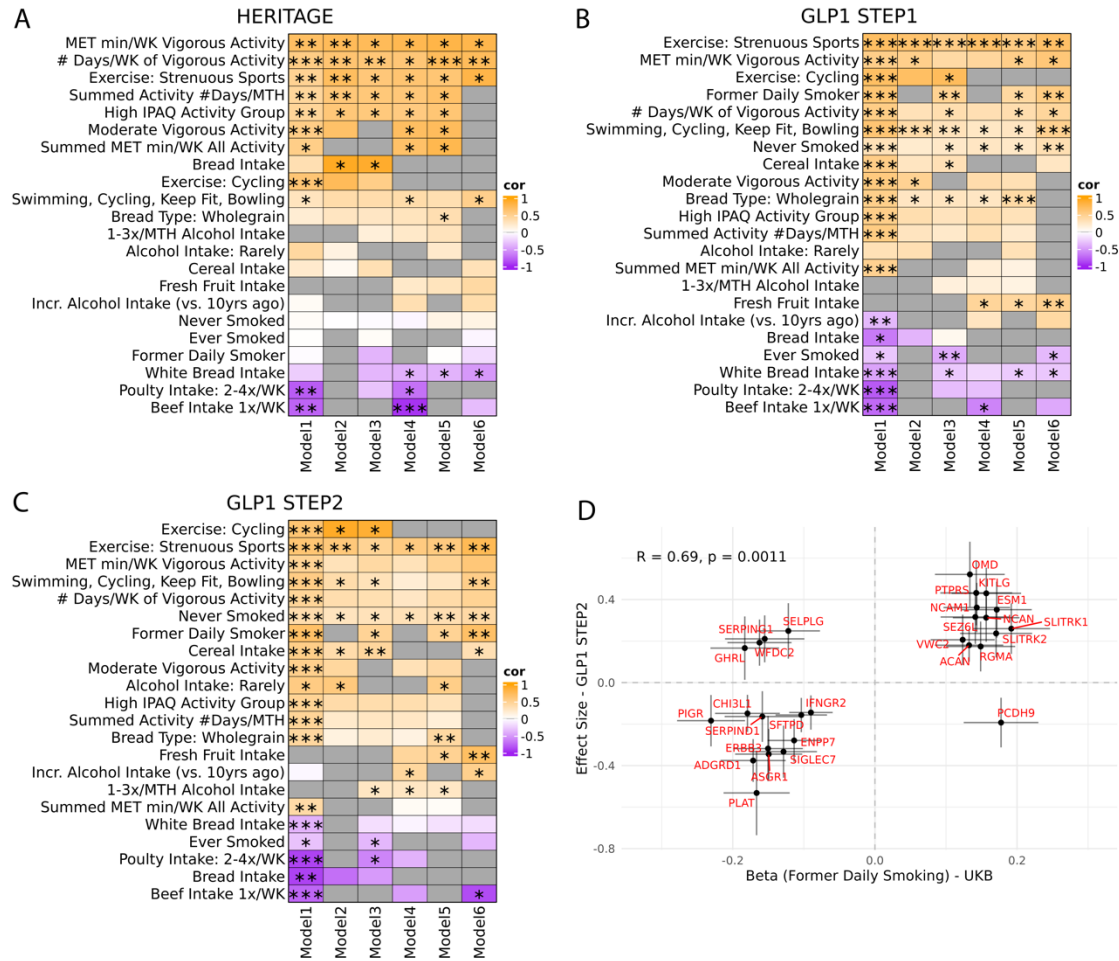
Supplementary Figure 1. Generalizability of R^2 estimates. **A.)** Plot of the slope between train and test R^2 estimates across the proteome for the polyexposure (E), polygenic (G), and gene-by-environment (GxE) scores for each covariate specification. **B.)** Normalized mean absolute error split by E, G and GxE and for the **C.)** entire model's R^2 . **D.)** Density of total R^2 's by covariate specification.



Supplementary Figure 2. R^2 estimates for covariates. **A.)** Scatter plot of R^2 estimates for covariates (e.g. Age, Sex, BMI, etc.) vs polygenic scores for each protein.

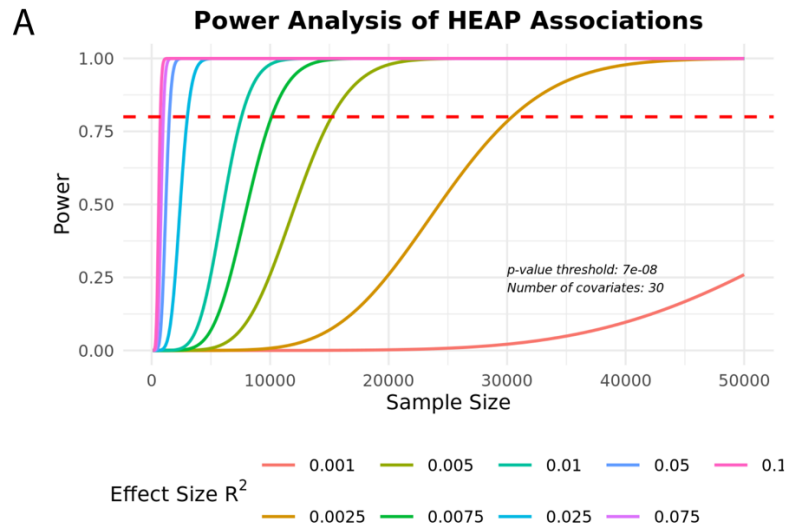


Supplementary Figure 3. *Variability in GEM estimates across specifications and correlation with pleiotropy.* **A.)** Scatter plot of GEM statistics with standard error of the estimate across specifications. **B.)** Pearson correlations of GEM statistics and a protein's number of significant disease associations across specifications. 95% confidence interval and adjusted p-value shown.



Supplementary Figure 4. Correlations of HEAP associations with Interventional Cohorts. A-C.)

Heatmap of Pearson correlations between significant proteomic signatures for HEAP under differing covariate specifications, HERITAGE exercise intervention study, and GLP1 agonist STEP1/STEP2 trials (FDR < 0.05). Significant correlations (FDR < 0.05) are denoted by asterisks marked as: “*” adj.p < 0.05, “**” adj.p < 0.01, “***” adj.p < 0.001. Correlation of proteomic signatures with **A.)** HERITAGE **B.)** GLP1 STEP1, and **C.)** GLP1 STEP2. **D.)** Scatter plot showing correlations between associations found in HEAP: former daily smoking vs GLP1 agonist in STEP2 cohort. Pearson correlation and Benjamini-Hochberg p-value shown.



Supplementary Figure 5. Power to Detect E and $G \times E$ Associations. A.) Power curves across the range of R^2 detected in HEAP for both exposure-protein and polygenic gene-by-environment associations.