Nucleic Acids Research, 2026, **54**, D1222–D1233
https://doi.org/10.1093/nar/gkaf1058
Advance access publication date: 29 October 2025
Database issue

ProteoNexus: an integrative database to characterize genetic architecture, estimate mediation effects, and construct and evaluate prediction models of the plasma proteome

Kaixin Shao[1], **Zixin Luo**[1], **Peng Huang**[2,*], **Sheng Yang**[1,*]

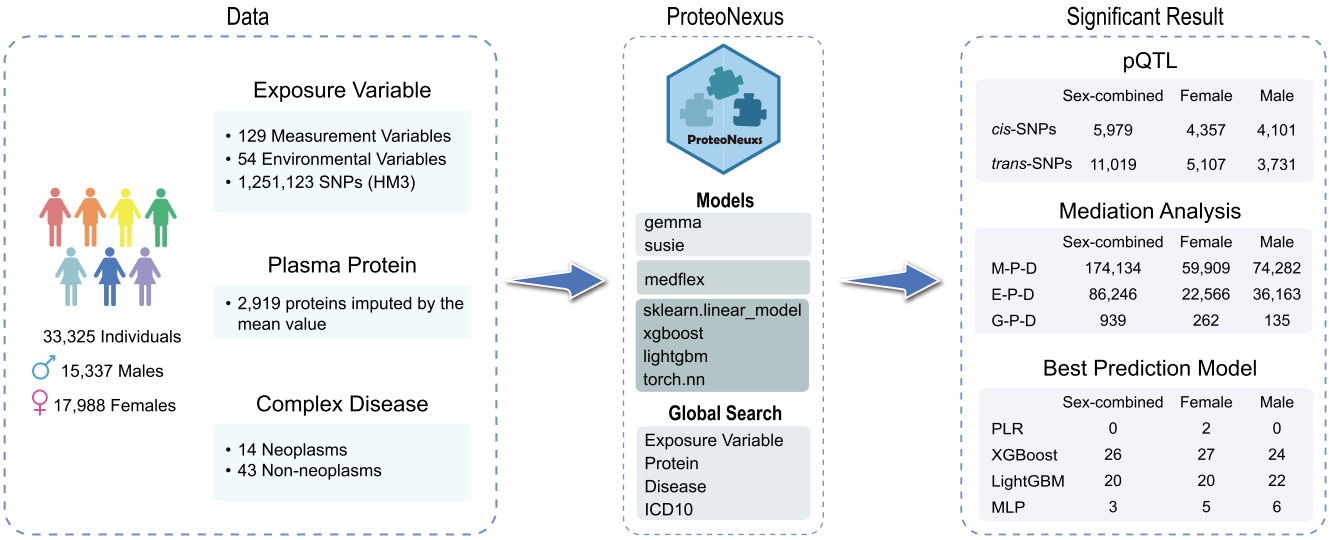
[1]Department of Biostatistics, Centre for Global Health, School of Public Health, Nanjing Medical University, Nanjing, Jiangsu 211166, China
[2]Department of Epidemiology, Key Laboratory of Public Health Safety and Emergency Prevention and Control Technology of Higher Education Institutions in Jiangsu Province, Centre for Global Health, School of Public Health, National Vaccine Innovation Platform, Nanjing Medical University, Nanjing, Jiangsu 211166, China

*To whom correspondence should be addressed. Email: yangsheng@njmu.edu.cn
Correspondence may also be addressed to Peng Huang. Email: huangpeng@njmu.edu.cn

Abstract

Proteins are biological effectors that mediate the effects of exposures on diseases and serve as predictors for constructing high-performance disease prediction models. However, an integrative, sex-specific proteomic resource using a biobank-scale dataset remains unavailable. Here, we introduce ProteoNexus, a database featuring a standardized best-practice pipeline integrating protein pQTLs mapping, mediation analysis, and risk prediction. Following stringent quality control, ProteoNexus comprises three categories of exposures: 129 measurement-based variables, 54 environmental variables, 1 251 123 single-nucleotide polymorphisms (SNPs), and 57 incident diseases among 33 325 European participants. ProteoNexus identifies 16 998 putative causal pQTLs, of which 5 979 are *cis*-pQTLs and 11 019 are *trans*-pQTLs in the combined-sex dataset, while 9 464 and 7 832 pQTLs were identified in the female and male datasets, respectively. Using a two-step screening strategy, ProteoNexus identifies 308 325, 144 975, and 1 336 significant pathways caused by measurement-based variables, environmental variables, and SNPs, respectively, followed by enrichment analysis of proteins associated with these exposures. With 21 optimized parameters for four machine learning algorithms, ProteoNexus provides an online analysis module that enables users to analyze their own proteomic data. Users can search for results by protein, reported disease, ICD-10 code, or exposure, with accompanying summary statistics for each query. ProteoNexus is freely accessible at https://www.proteonexus.com/.

Graphical abstract



Received: August 11, 2025. **Revised:** September 16, 2025. **Accepted:** September 29, 2025

Introduction

Plasma proteins are biological effectors that bridge the gap between complex traits, environmental exposures, and the genome to complex disease incidence [1, 2]. Recent high-throughput proteomics technology profiles circulating proteins at a population scale [3–5] and elucidates their mediator role in disease etiology [6–10]. For example, based on two-step proteome-wide Mendelian randomization (MR), Yoshiji *et al*. identified *COL6A3* as a positive mediator of the effect of body mass index (BMI) on increased coronary artery disease risk [8]. Integration of genome-wide association study (GWAS), a useful tool for identifying single-nucleotide polymorphisms (SNPs) associated with complex traits, with proteomics may facilitate identification of protein quantitative trait loci (pQTLs) [3–5, 11–13]. For example, the UK Biobank Pharma Proteomics Project (UKB-PPP) identified >14 000 pQTLs in common variants and 5 400 pQTLs in rare variants by measuring ∼3000 plasma proteins in nearly 50 000 participants [3, 4]. In addition, sex-specificity can be identified in the genetic architecture of complex traits and plasma levels of most proteins [14–17]. Sex also acts as an effect modifier when assessing protein effects [7]. Proteomics-driven mediation analyses in individual-level datasets and whole-genome pQTL analysis in a sex-specificity framework remain absent.

Concurrently, proteome risk scores for complex diseases derived from large-scale proteomics have exhibited higher prediction performance with or without clinical variables [1, 18]. Paired with the accessibility of proteomics data in biobank-scale data, such as UKB-PPP [3, 4] and FinnGen [5], a variety of studies have used machine learning to facilitate prediction model construction for diseases, such as major cardiovascular events with *Boruta* [18], peripheral artery disease for type II diabetes (T2D) patients with lasso [19], and Parkinson's disease with light gradient boosting machine (LightGBM) [20]. You *et al*. developed a multilayer perceptron neural network model to construct proteomic risk scores, which had superior or equivalent predictive performance compared to established clinical indicators across nearly all endpoints for 45 diseases [21]. Both the restricted set of model types [22, 23] and the lack of sophisticated hyperparameter optimization frameworks severely limit the potential for developing robust, generalizable prediction models.

To address these issues, we introduce ProteoNexus (https://www.proteonexus.com/), an integrative platform that characterizes the genetic architecture of 2 919 plasma proteins, quantifies the mediating roles of proteins between 129 measurement-based variables, 54 environmental variables, and 1 251 123 SNPs and 57 diseases, and constructs predictive models for 57 disease incidences (Fig. 1A). To minimize potential bias, we performed rigorous quality control (QC) for the UKB-PPP data (Application ID: 144904), including participant selection, onset disease identification, and covariable inclusion. In the pQTL module, we applied linear regression and fine mapping to define putative causal SNPs [24]. In the mediation modules, based on two-step screening, we estimated the direct effect (DE) and indirect effect (IE) and their corresponding 95% confidence intervals (CIs) with the Delta method [25]. In the prediction module, we implemented and optimized four distinct machine learning algorithms within a tree-structured Parzen estimator (TPE) framework and provided an online analytic function with the uploaded proteomic data to facilitate simple and user-friendly analysis of the risk prediction. All analytical modules provide sex-specific results, dynamic visualization capabilities, and functional enrichment analyses, thereby offering researchers an accessible, end-to-end resource for exploring plasma proteome mediation and constructing optimized predictive models for complex diseases (Fig. 1B).

Materials and methods

UKB data curation and processing

The UKB is a population-based cohort study that enrolled >500 000 individuals between 2006 and 2010 [26]. Proteomic profiling was performed on blood plasma samples from the UKB-PPP with 53 016 participants [3]. In brief, the Olink Explore 3072 platform was used to quantify 2 941 protein analytes, representing 2 923 unique proteins.

To ensure the robustness of ProteoNexus, we performed strict data QC. In the participant QC, we excluded participants: (i) not of European (EUR) ancestry; (ii) with relatedness; (iii) with gender incongruence; (iv) without genetic analysis; (v) without genetic principal components (PCs); (vi) without BMI information; and (vii) with missing data of >30% protein [17, 27–29]. For the QC of the proteomic data, proteins with missing values in >20% of the individuals were excluded. The remaining missing data were then imputed with the protein-specific mean value [19]. In the SNP QC, we excluded the SNPs (i) with a minor allele frequency <0.01; (ii) with a Hardy-Weinberg equilibrium P-value < 10^{-7}; (iii) with >5% missing data ($Pm > 0.05$); (iv) duplicated; (v) not Hapmap3 (HM3) [27, 29]. The analytic dataset comprised 33 325 participants, consisting of 17 988 females and 15 337 males, with 2 919 proteins and 1 251 123 HM3 SNPs (Supplementary Fig. S1).

To enhance the accuracy and generalization of ProteoNexus, we defined three types of exposure variables, considered incident diseases as the outcome, and included appropriate covariables. The exposure variables, referring to $\boldsymbol{E}$, included: (i) measurement-based variables: 57 anthropometric, 65 blood and urine tests, one psychosocial factor, and six other variables (Supplementary Table S1); (ii) environmental variables: 22 residential air pollution, nine greenspace and coastal proximity, five sociodemographic, and 18 healthy life variables (Supplementary Table S2); and (iii) genetic variables. Specifically, we used three variables at baseline to assess the SES of each participant at the individual level, including family income level (data field: p738_i0), education qualification (data field: p6138_i0), and employment status (data field: p6142_i0). Following the same procedure in [17], we included four integrative factors collected at baseline, including "no current smoking," "regular physical activity," "healthy diet," and "no alcohol consumption," to define "healthy lifestyle." If an individual has three or more indexes marked as "Yes" or "High," we defined his/her healthy lifestyle as 1 [17]. Next, cancer outcomes were sourced from the cancer registry [International Classification of Diseases (ICD) codes], whereas noncancer diseases were sourced from first-occurrence traits available in the UKB. The first-occurrence disease was defined as the integration of ICD-10 codes with self-reported disease (field ID: p20001 and p20007 for self-reported cancer definition, p40006 and p40008 for cancer diagnosis, and p20002 and p20009 for self-reported non-cancer definition). We involved 57 diseases, which comprised 14 neoplasms and

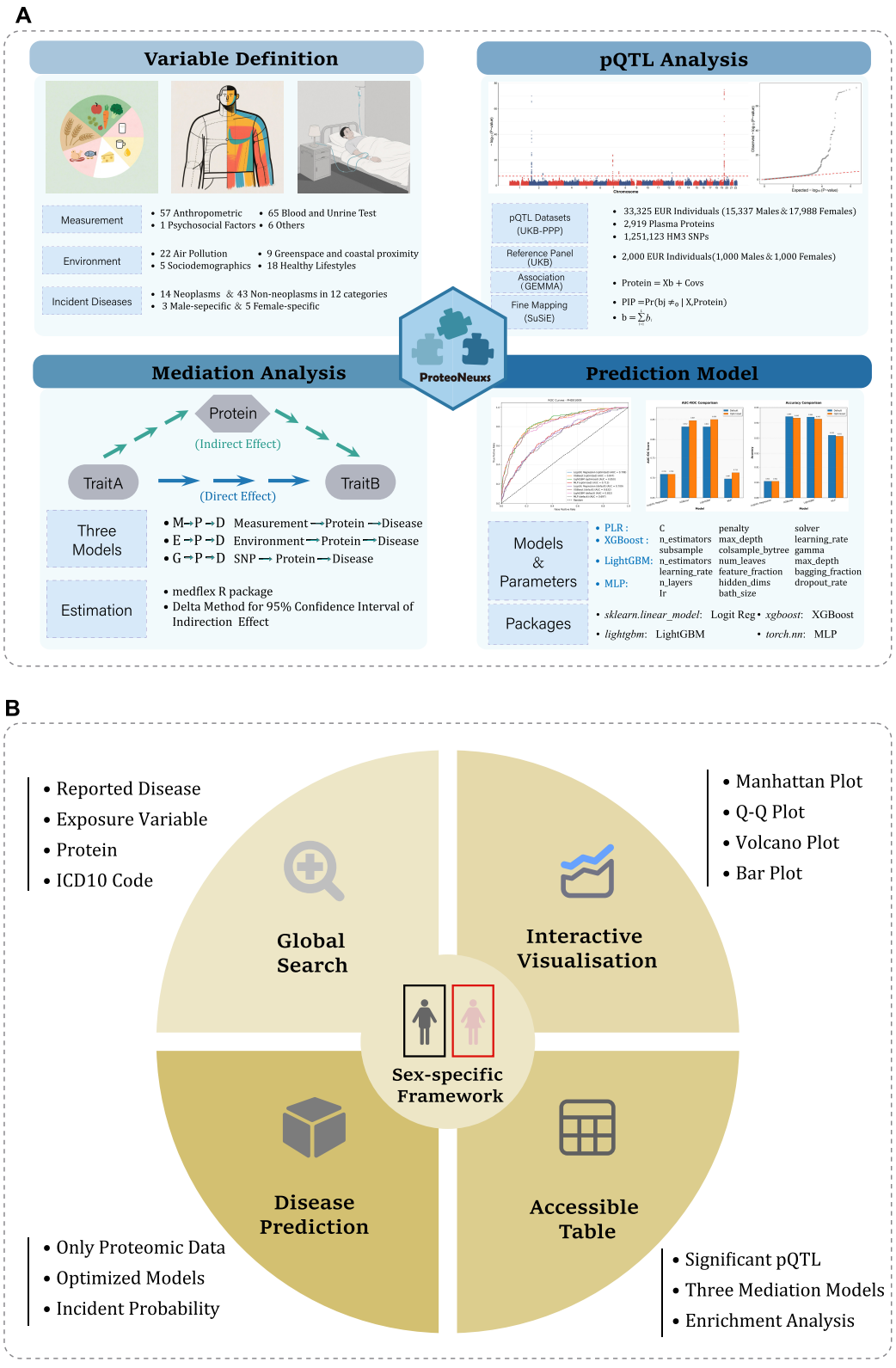


Figure 1. Workflow and design of ProteoNexus. (**A**) Based on a strict QC and variable identification for UKBB data, ProteoNexus characterizes the genetic architecture of 2 919 plasma proteins by identifying the significant and putative causal SNPs. It performs mediation analysis between 129 measurement-based variables, 54 environmental variables, 1 251 123 SNPs, and 57 diseases, along with enrichment analysis. It optimizes four machine learning algorithms to construct predictive models for 57 disease incidences. (**B**) ProteoNexus consists of four functional modules: global search, interactive visualization, disease prediction, and accessible table. We performed the analysis and prediction models in the sex-specific framework.

43 non-neoplasms (Supplementary Table S3) [17]. In covariable selection, we included age when attended assessment center (field ID: p21022), sex (field ID: p31 and p22001), site (field ID: p54) for adjusting the environmental variable, array (field ID: p22000), BMI at baseline (field ID: p21001_i0), ethnic background (field ID: p21000), genetically related variables (field ID: p22018, p22010, and p22020), and the top 18 genetic PCs (field ID: p22009_a1~p22009_a18) to adjust for population stratification [3, 17, 30].

pQTL estimation

In the whole-genome pQTL analysis, we defined the putative causal pQTL with marginal effect and posterior inclusion probability (PIP) (Fig. 1A). We first calculated marginal effect size for the 33 325 participants by fitting standard linear regression using the GEMMA software [31]. For each plasma protein, we fitted a linear regression model to remove the effects of sex, age, site, BMI, array, and the top 18 PCs and obtained residuals (Equation 1):

$$\boldsymbol{P} \sim \beta_G \boldsymbol{G} + \boldsymbol{\beta}_c \boldsymbol{C}, \quad (1)$$

where $\boldsymbol{P}$ was the expression of protein, $\boldsymbol{G}$ showed the genotype data, β_G indicated the marginal effect for SNPs, $\boldsymbol{C}$ indicated the covariables, and $\boldsymbol{\beta}_c$ represented the effect for covariables. Furthermore, for summary statistics of each protein, we used SuSiE to estimate the PIP of each SNP within any credible set [32]. Following our previous paper [17], we used preprocessed EUR reference panels with 2 000 EUR UKB individuals (1 000 females and 1 000 males) with HM3 SNPs. We used the block-wise linkage disequilibrium (LD) matrix [33]. For SuSiE, we used default configuration: maximum number of effects (L) = 10, target credible set (CS) coverage = 0.95, and prior variance = 50. For a specific protein, we defined *cis*-SNPs within 1 Mb of transcription start site or transcription end site of the coding gene [3]. We cataloged pQTLs with the following annotations: chromosome, SNP ID, genomic position (GRCh37), observed and missing sample sizes, minor and major alleles, allele frequency, effect size, standard error, P-value, PIP, and the family-wise error rate (FWER) from a Bonferroni correction. We also produced Manhattan and Q–Q plots. Putative causal pQTLs were defined as variants with FWER < 0.05 and PIP > 0.8. For the per-protein FWER adjustment, we set the number of hypothesis tests (K) equal to the number of SNPs tested for that protein. All pQTL summary statistics are available for download, enabling further custom analyses by users.

Mediation analyses

To investigate the mediator role of proteins, we evaluated three types of pathways connecting exposures to disease outcomes via protein levels: (i) M-P-D [Measurement-based variable → Protein(s) → Disease], (ii) E-P-D [Environmental variable → Protein(s) → Disease], and (iii) G-P-D [Genetic variant → Protein(s) → Disease]. To ensure the accuracy and robustness of the mediation model, we employed a two-step statistical approach in the FDR framework (Fig. 2).

In Step 1, we fitted three kinds of TE models, including M–D, E–D, and G–D (Equation 2):

$$logit\,(\boldsymbol{D} = 1) \sim \beta_{ED} \boldsymbol{E} + \boldsymbol{\beta}_{CD} \boldsymbol{C}, \quad (2)$$

where $\boldsymbol{D}$ indicated the disease incidence, $\boldsymbol{E}$ denoted measurement-based, environmental, or genetic variables

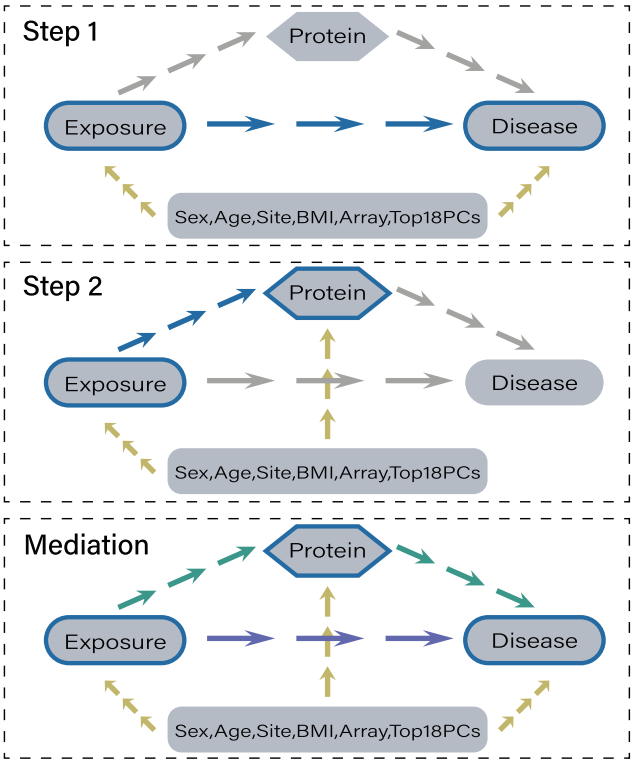


Figure 2. Schematic diagram for the mediation analysis in ProteoNexus. Step 1 is the TE model. Step 2 is the association between all proteins and one specific exposure variable. We only fit the mediation analysis after the two-step screening. The arrows with five colors represent different types of effect: (i) blue arrow: TE or effect of exposure variable to proteins; (ii) yellow arrow: confounding effect; (iii) green arrow: IE; (iv) purple arrow: DE; (v) gray arrow: no effects.

(i.e. SNP), β_{ED} represented the effect of exposure variable, and $\boldsymbol{\beta}_{CD}$ represented the effect for covariables. For a specific disease, we only retained associations with significant total effect (TE). In the M–D and E–D models, we fitted the logit model to estimate the total effect and defined its significance with a nominal P-value $< .05$. In the G–D model, we fitted linear models in GEMMA to estimate each SNP's total effect on disease. Statistical significance was assessed using a Bonferroni-corrected FWER of 0.05. For the per-SNP correction, K was defined as the number of diseases tested.

In Step 2, we investigated the association between all proteins and a specific exposure variable, fitting the M–P, E–P, and G–P models (Equation 3):

$$\boldsymbol{P} \sim \beta_{EP} \boldsymbol{E} + \boldsymbol{\beta}_{CP} \boldsymbol{C}, \quad (3)$$

where β_{EP} represented the effect of exposure variable and $\boldsymbol{\beta}_{CP}$ was the effect for covariables. For a specific exposure variable, we only collected the protein with significant association. For M–P and E–P, we fitted the linear model to estimate the associations and defined its significance with FDR in the BH procedure < 0.05. We provided a volcano plot to indicate the relationship between plasma proteins and exposure variables. For the significant proteins, we performed pathway enrichment analysis with two databases: Gene Ontology (GO) [34] and Kyoto Encyclopedia of Genes and Genomes (KEGG) [35] using the clusterprofiler package [36, 37] and employed a bar plot to show the top significant pathways. In the G–P model, we used GEMMA to fit linear models estimating pQTL ef-

fects. Variants were deemed significant if Bonferroni-corrected FWER < 0.05 and PIP > 0.8; for per–SNP correction, K was set to the number of proteins tested.

For the mediation model, we estimated DE and IE for three types of pathways using the *medflex* R package, referring to M–P–D, E–P–D, and G–P–D [25]. Different from the bootstrap strategy for CI estimation in the mediation R package, we used the Delta method to estimate 95% CIs. Specifically, besides the pathways defined by the two above models, we provided the pathways with significant indirect effect size (FDR in the BH procedure < 0.05). For the significant pathways, we offered the effect size, SE, P-value, and 95% CIs for DE and IE, respectively.

Prediction model construction and evaluation

We not only provided pQTL and mediation analysis but also integrated an online prediction module for incident diseases using baseline proteomic profiles. We trained four types of classification models for each disease: (i) penalized logistic regression (PLR) with L1/L2 regularization, (ii) extreme gradient boosting (XGBoost), (iii) LightGBM, and (iv) multi-layer perceptron (MLP) neural network (Supplementary Table S4). To avoid overfitting and maximize generalizability, we randomly split the cohort into a training set (80%) and a hold-out test set (20%) stratified by disease status. Model hyperparameters were tuned via five-fold cross-validation on the training set using the TPE approach (implemented in the Optuna framework) to efficiently explore the hyperparameter space. The optimization objective was to maximize the area under the receiver operating characteristic curve (AUC). The final models selected were those with the highest cross-validated AUC. We evaluated performance on the test set, reporting metrics including AUC and, for reference, accuracy at the optimal threshold. Specifically, for diseases analyzed in the sex-combined dataset, we nonetheless optimized hyperparameters separately within each sex.

To ensure security and privacy, we have implemented a suite of data protection measures. All data exchanged between the user's browser and our servers is protected by transport layer security (TLS) encryption, thereby mitigating interception risks. User sessions are executed within logically isolated environments, ensuring that uploaded data and computational processes remain segregated from those of other users and precluding any unauthorized access. Throughout the analytical workflow, files reside exclusively in memory or temporary storage; no residual data persists on our servers beyond the active session. Upon session termination or completion of the analysis, all data and temporary files are irreversibly and immediately destroyed. Furthermore, the system operates without collecting personally identifiable information and deliberately refrains from logging any user-specific metadata, such as IP addresses or access timestamps, thereby preserving total anonymity.

Design of ProteoNexus

ProteoNexus organizes its results across various web interfaces, including Home, pQTL, Mediation M-P-D, Mediation E-P-D, Mediation G-P-D, Global Search, Prediction Models and About. The pQTL web interface features an analytical workflow, dual input fields for protein specification and sex selection, a comprehensive table displaying 13 key properties of putative causal SNPs, and two integrated visualizations. Users must specify both the protein name and sex parameter simultaneously. Complete pQTL analysis results are publicly accessible through Science Data Bank (ScienceDB), accompanied by metadata in JSON format. The three mediation analysis interfaces present dedicated workflows, input fields for exposure variables and disease specification, tabulated results for Step 1, Step 2, and mediation models, along with corresponding visualizations. Disease traits are systematically categorized into 12 groups according to ICD-10 classification. The Global Search interface features a unified query system supporting keyword searches by gene symbol (or protein name), exposure, or disease, providing summary statistics and visualizations for three mediation analyses, quantification of mediators (proteins), and significant mediation results. The Prediction Models interface incorporates a disease specification field, visualizations for predictive performance assessment, and a Streamlit-based file upload component for the user's proteomics data.

Moreover, ProteoNexus implements three categories of interactive data visualization: rapid filtering and sorting of tabulated data, such as significant pQTLs; zoom functionality for detailed examination of visualizations, such as volcano plots depicting protein-exposure associations; and downloadable figures and datasets for subsequent offline analysis. The platform's functional modules accommodate three distinct sex-specificity options for comprehensive analysis.

Implementation

ProteoNexus is implemented as a dynamic web application to maximize accessibility. The front-end interface, developed with Streamlit, a Python-based web framework, renders interactive charts and tables directly in the browser. The back end comprises precomputed result tables and visualizations, including pQTL outputs, mediation results, and model coefficients, and processes user queries through Python APIs. In addition, the prediction server hosts all pretrained models and automatically applies the optimal model, selected according to the user's specifications, to the uploaded data. It facilitates external validation and clinical translation of proteomic risk scores. The site is deployed on a cloud server and can be accessed via HTTPS without login.

Results

Overview of ProteoNexus

ProteoNexus is an integrative and user-friendly database designed to identify the causal pathways between exposure variables and diseases mediated by proteins and to construct the prediction model with proteomic data (Fig. 1). The database includes carefully curated UKB-PPP individual-level data, including participant selection, protein filtration, disease onset identification, covariable selection, and SNP QC. In the pQTL analysis, ProteoNexus uncovered 34 294 putative causal pQTLs regulating 2 256 different proteins, across the sex-combined, female, and male datasets. ProteoNexus provides detailed results for the pQTLs of each protein, including a table of putative causal pQTLs, Manhattan plots, and a corresponding Q–Q plot, all of which are available for download. In the mediation analysis conducted in three different scenarios, ProteoNexus thoroughly defines the significant pathways for 308 325 M–P–D, 144 975 E–P–D, and 1 336 G–P–D using a two-step screening process. When specifying the expo-

sure variable, sex and disease, ProteoNexus progressively provides the detailed information for the diseases with significant TE, the detailed information for the significant proteins and their enrichment analysis results, and the detailed information for mediation analysis. Along with visualization for summary results, the platform supports a keyword query system for the mediation analysis in the Global Search interface. Trait-based searches can be conducted using the Reported Trait and ICD-10 code. In the prediction model, we optimized four algorithms and provided components for the user's proteomic data. These features make ProteoNexus a powerful tool for investigating the functions of plasma proteins for complex diseases.

Details of ProteoNexus

ProteoNexus investigates pQTL with association tests and fine-mapping analysis, including *cis*- and *trans*-SNPs for the sex-combined and sex-specific datasets. We defined 2 183 (74.79%), 1 953 (66.91%), and 1 772 (60.71%) protein expression levels regulated by SNPs, respectively. Interestingly, we identified 1 528 and 2 061 shared proteins regulated by *cis*-SNPs and *trans*-SNPs among the three datasets. We also identified 58 and 75 sex-specific proteins regulated by *cis*-SNPs and *trans*-SNPs among the three datasets (Fig. 3A and B). For example, *DXO*, one of the sex-specific proteins, showed two putative causal *trans*-pQTLs in the female dataset, whereas it was regulated only by five *cis*-SNPs in the male dataset. The sex-combined dataset contains 5 979 putative causal *cis*-SNPs and 11 019 putative causal *trans*-SNPs, respectively. The two sex-specific datasets include 4 357 and 4 101 putative causal *cis*-SNPs and 5 107 and 3 731 putative causal *trans*-SNPs, respectively (Fig. 3C and D). Among three datasets, we detected 401, 233, and 214 proteins with the number of putative causal *cis*-SNPs >5, respectively (Fig. 3E). In addition, we detected 731, 268, and 190 proteins with the number of putative causal *trans*-SNPs >5, respectively (Fig. 3F). In sex-shared investigation, the CS size for each protein ranges from 222 (*TNPO1*) to 516 (*DSG4*) (median: 317). In sex-specific investigation, the CS size for each protein ranges from 726 (*SLC16A1*) to 905 (*KIT*) (median: 831).

ProteoNexus provides the result of three kinds of mediation analysis, including M–P–D, E–P–D, and G–P–D for sex-combined and sex-specific situations. For the sex-combined dataset, ProteoNexus features 174 134 M–P–D, 86 246 E–P–D, and 939 G–P–D pathways mediated by 2 138, 1 804, and 635 proteins, respectively. For M–P–D, we defined 160 138 (91.96%), 8 419 (4.83%), 2 050 (1.18%), and 3 527 (2.03%) pathways for anthropometric, blood and urine test, psychosocial factors, and other (Fig. 4A and B). One thousand two hundred eighty-two proteins have been identified as mediators of the pathway from impedance of whole body, manual entry to Type II diabetes, while only one protein (*PRAP1*) has been identified to mediate the pathway from pulse rate to Type II diabetes ($\beta_{DE} = 0.013$, $\beta_{IE} = 0.010$). For E–P–D, we defined 19 203 (22.27%), 3 378 (3.92%), 18 169 (21.07%), and 45 496 (52.75%) pathways for residential air pollution, greenspace and coastal proximity, sociodemographic, and healthy lifestyle (Fig. 4C and D). Nine hundred sixty proteins have been identified as mediators of the pathway from current tobacco smoking to acute kidney failure, whereas one protein (*REG4*) has been identified to mediate the pathway from no smoke to colorectal cancer ($\beta_{DE} = -0.190$, $\beta_{IE} = -0.039$). For G–P–D, we defined 216 (16.17%) and 1 120 (83.83%) pathways for *cis*-SNPs and *trans*-SNPs (Fig. 4E and F). For sex-specificity, ProteoNexus identified 59 909 M–P–D, 22 566 E–P–D, and 262 G–P–D pathways mediated by 1 850, 1 456, and 224 proteins for female and 74 282 M–P–D, 36 163 E–P–D, and 135 G–P–D pathways mediated by 1 745, 1 342, and 107 proteins for male (Fig. 4A–F). For example, the E-P-D pathway from healthy diet to coronary artery disease was found to involve 536, 440, and 0 proteins in sex-combined, female, and male datasets, respectively.

ProteoNexus provides four optimized prediction models, including PLR, XGBoost, LightGBM, and MLP, for each disease (Fig. 4G). The prediction models, on average, attained strong discriminative ability, with AUC > 0.75 for 26 of the 57 diseases. Specifically, across sex-specificity, XGBoost is the single best-performing algorithm for the largest number of outcomes (26 diseases in sex-combined, 27 diseases in female, and 24 diseases in male), followed by LightGBM (20 diseases in sex-combined, 20 diseases in female, and 22 diseases in male) and MLP (three diseases in sex-combined, five diseases in female, and six diseases in male). For example, in T2D, an XGBoost model achieved AUC $= 0.884$, substantially outperforming a model using age, sex, BMI, and family history [21]. In acute renal failure, proteomic predictors reached AUC $= 0.807$, indicating high discriminative power. Notably, 32 of 49 non-sex-specific diseases saw improved AUC in one or both sex-specific models compared to the sex-combined model. For instance, the AUC for predicting kidney failure increased by 4.05% (from 0.766 to 0.797) when using a female-specific model and by 3.71% (from 0.781 to 0.810) with a male-specific model, reflecting sex differences in proteomic risk markers.

Case study: three analytic modules of dementia using ProteoNexus

To demonstrate the analytical capabilities and biological insights afforded by ProteoNexus, we conducted a comprehensive investigation of onset dementia (ICD-10: F00, F01, F02, F03, G30, and G31). This case study illustrates how ProteoNexus helps biological investigation by starting from the Global Search interface and providing three kinds of mediation analysis: M–P–D, E–P–D, and G–P–D, accompanied by corresponding pQTLs and optimized prediction models.

In the Global Search interface, when specifying "dementia," ProteoNexus presents 2 769 statistically significant pathways, comprising 1 872 M–P–D, 849 E–P–D, and 48 G–P–D pathways and involving 290 protein mediators (Fig. 5A and B). ProteoNexus also outputs the top 10 mediators with the highest frequency across all pathways, notably *NEFL* ($n = 43$), *GDF15* ($n = 35$), *BCAN* ($n = 32$), *HPGDS* ($n = 31$), *REN* ($n = 30$), and *FAP* ($n = 29$) (Fig. 5C). Unless sex is explicitly specified, all results reported below refer to the sex-combined dataset.

In the M–P–D interface, ProteoNexus quantifies the effects of 22 measurement-based variables mediated by 234 different proteins, of which 20 anthropometric variables are selected, underscoring the central role of body composition and physical function in dementia susceptibility. The associations between variables related to muscle strength and adiposity and dementia are mediated by the largest number of proteins, such as hand-grip strength (left) by 124 proteins and waist-to-hip ratio (WHR) associated with 126 proteins. We

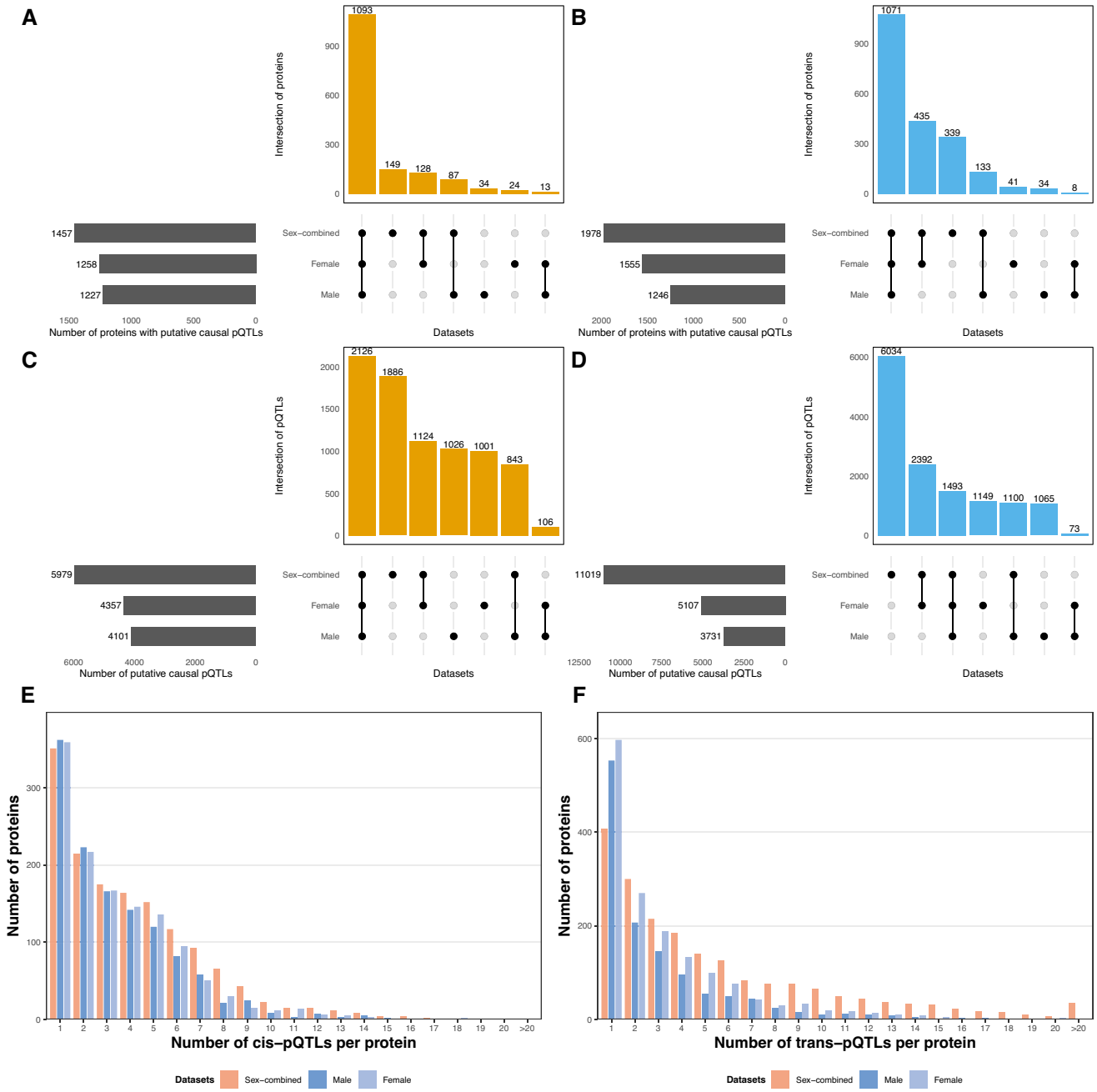


Figure 3. Data statistics of pQTL. (**A**) The upset plot shows the number of proteins with putative causal *cis*-SNPs among the three datasets. (**B**) The upset plot shows the number of proteins with putative causal *trans*-SNPs among the three datasets. (**C**) The upset plot shows the number of putative causal *cis*-SNPs among the three datasets. (**D**) The upset plot shows the number of putative causal *trans*-SNPs among the three datasets. (**E**) The bar plot presents the distribution of the number of *cis*-pQTLs per protein. (**F**) The bar plot presents the distribution of the number of *trans*-pQTLs per protein.

use WHR as an example to show the two-step screening for the mediation study. In Step 1, WHR is associated with 31 diseases, such as six diseases in the "Diseases of the circulatory system" category and five in the "Mental and behavioral disorders" category. Specifically, the TE between WHR and dementia is significant ($\beta_{TE} = 1.257,\ P = 2.71 \times 10^{-2}$). In Step 2, 2 213 proteins are associated with WHR (Fig. 5D), such as *ERBB2* ($\beta = 2.373,\ FDR = 1.68 \times 10^{-157}$). In the mediation analysis, ProteoNexus detects 126 mediators, such as *TNFRSF10B* ($\beta_{IE} = 0.164,\ 95\%\ CI: 0.096 \sim 0.231,\ FDR = 4.30 \times 10^{-4}$) (Fig. 5E). Notably, *TNFRSF10B* significantly mediates the effects of both hand-grip strength (left) ($\beta_{IE} = -0.002,\ 95\%\ CI: -0.004 \sim -0.001,\ FDR = 0.006$) and hand-grip strength (right) on dementia ($\beta_{IE} = -0.002,\ 95\%\ CI: -0.003 \sim -0.001,\ FDR = 0.007$) in female dataset, corroborating findings previously reported in a mouse model [38].

In the E-P-D interface, ProteoNexus demonstrates that environmental and lifestyle exposures are mediated by 201 different proteins, including four healthy lifestyle factors, three residential air pollution variables, two greenspace and coastal proximity variables, and two sociodemographic vari-

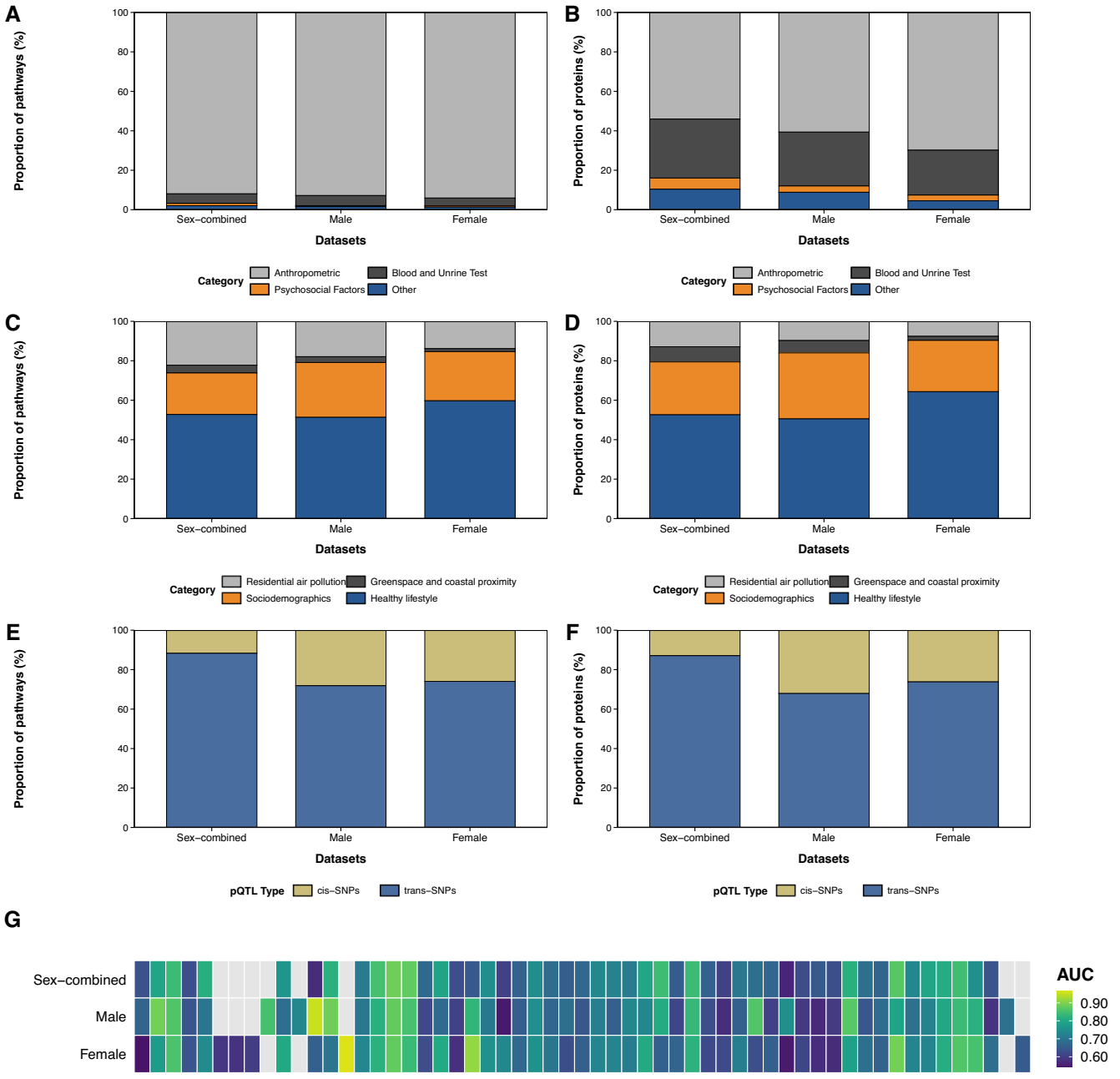


Figure 4. Data statistics of mediation analysis for three kinds of exposure variables and prediction models in three sex-specific datasets. (**A**) The stacked bar plot shows the proportion of pathways among the four categories of measurement-based variables in M–P–D. (**B**) The stacked bar plot shows the proportion of proteins in four categories of measurement-based variables in M–P–D. (**C**) The stacked bar plot shows the proportion of pathways for four categories of measurement-based variables in E–P–D. (**D**) The stacked bar plot shows the proportion of proteins in four categories of measurement-based variables in E–P–D. (**E**) The stacked bar plot shows the proportion of pathways of *cis*-SNP and *trans*-SNP in G–P–D. (**F**) The stacked bar plot shows the proportion of *cis*-SNP and *trans*-SNP proteins in G–P–D. (**G**) The heatmap illustrates the highest AUC values of 57 disease incidences.

ables. Consistent with the findings of a previous study [39], the association between sociodemographic factors and dementia is mediated by the largest number of proteins: 155 protein mediators for employment status and 148 mediators for the Townsend deprivation index (TDI). We used TDI as an example to show the two-step screening for the mediation study. In Step 1, TDI is associated with 34 diseases, such as six diseases in the "Diseases of the circulatory system" category and five in the "Mental and behavioral disorders" category. Specifically, the TE between TDI and dementia is significant ($\beta_{TE} = 0.060,\ P = 1.28 \times 10^{-7}$). In Step 2, 1 617 proteins are associated with TDI, such as *CXCL17* ($\beta = 0.0485,\ 95\%\ CI:\ 0.045 \sim 0.052,\ FDR = 2.80 \times 10^{-160}$). Mapping to gene symbols, we defines 2 082 GO pathways, including 1 792 Biological Process, 107 Cellular Component and 183 Molecular Function, and 54 KEGG pathways belonging to six categories (Fig. 5F). Among them, ProteoNexus defines 49 pathways directly related to neuron, such as neuroinflammatory response (GO: 0 150 076, $FDR = 3.04 \times 10^{-9}$) and regulation of neurogenesis (GO: 0 050 767, $FDR = 1.02 \times 10^{-7}$). In the mediation analysis, ProteoNexus detects 148 mediators in sex-combined

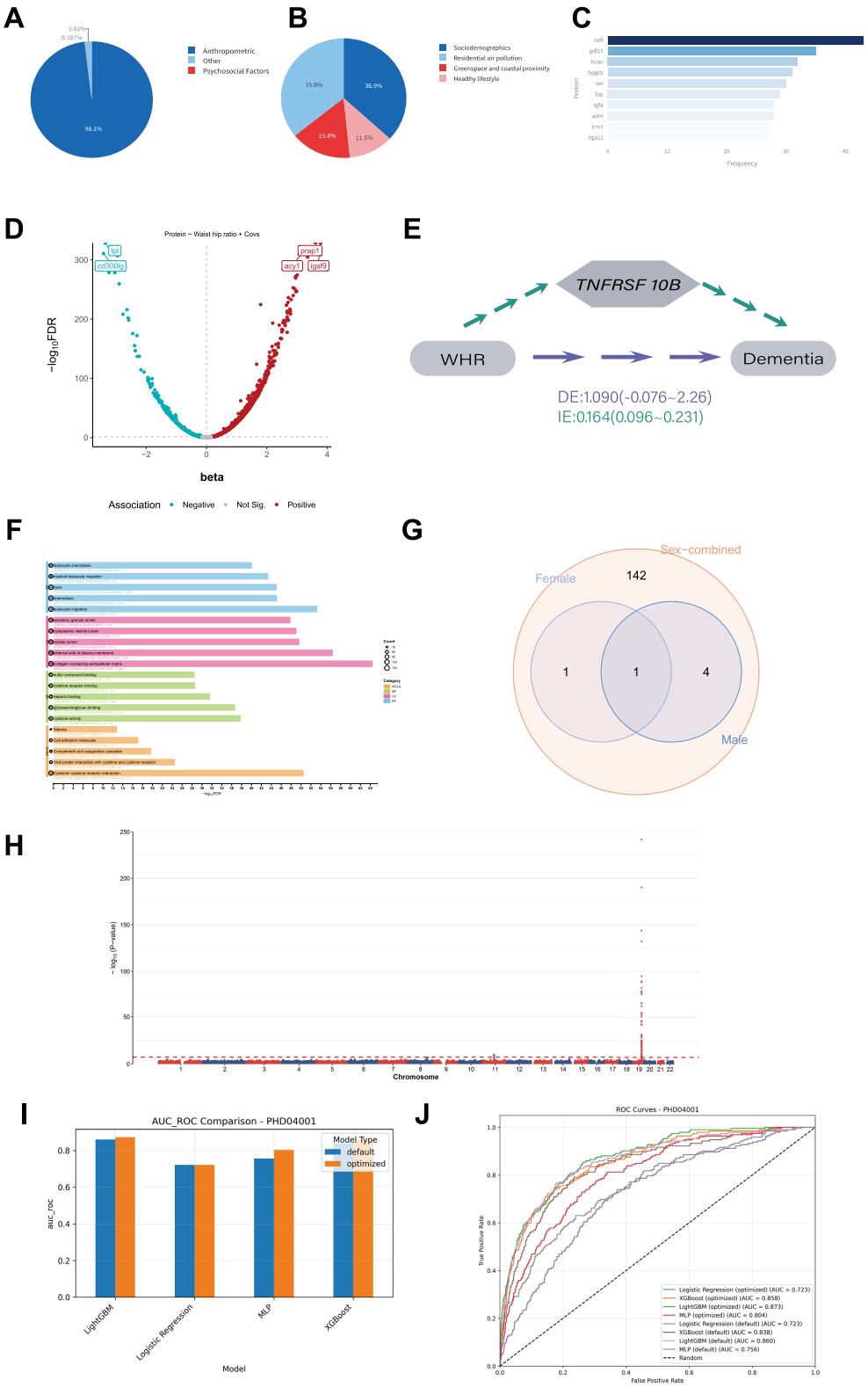


Figure 5. Summary of dementia results using ProteoNexus. (**A**) The pie chart shows the proportion of the M–P–D. (**B**) The pie chart shows the proportion of the E–P–D. (**C**) The bar plot presents the frequency of the top ten proteins. (**D**) The volcano plot illustrates the associations between plasma proteins and WHR. The 2 213 proteins that lie above the dotted line were identified as having an FDR of <0.05 in the BH procedure. Red and blue dots show the up- and down-regulated proteins. The diagram illustrates the mediation result for the WHR-*TNFRSF10B*-dementia pathway. (**F**) The bar plot presents the top pathways from BP, CC, and MF in GO and KEGG. (**G**) The Venn diagram displays the logical relationships between datasets. (**H**) The Manhattan plot presents the pQTL of *APOE*. (**I**) The bar plot shows a comparison of the four models with default and optimized parameters. The AUC plot indicates the prediction performance of the models.

dataset, such as *GDF15* ($\beta_{IE} = 0.013$, 95% $CI: 0.009 \sim 0.017$, $FDR < 10^{-5}$). Notably, *GDF15* also mediates this pathway in both female and male datasets (*male dataset*: $\beta_{IE} = 0.012$, 95% $CI: 0.006 \sim 0.017$, $FDR = 5.37 \times 10^{-3}$; *female dataset*: $\beta_{IE} = 0.015$, 95% $CI: 0.010 \sim 0.020$, $FDR = 3.00 \times 10^{-5}$) (Fig. 5G). Previous population-based studies also defined *GDF15* as a biomarker associated with dementia [40, 41].

In the G-P-D interface, ProteoNexus investigates 19 putative causal SNPs associated with dementia that are mediated by 30 proteins. Among them, there are seven *cis*-SNPs of *APOE*: rs157580, rs4420638, rs7412, rs157582, rs1160985, rs2075650, and rs8105340, which are consistent with previous genetic studies [42, 43]. Specifically, ProteoNexus defines four different SNPs that exhibit sex specificity and are mediated by *APOE*. For example, the rs7412 polymorphism, encoding the *APOE* ϵ2 isoform, acts as a mediator in the female dataset ($\beta_{IE} = -0.440$, 95% $CI: -0.571 \sim -0.309$, $FDR = 1.81 \times 10^{-9}$), whereas no significant IE is detected in males. These observations indicate that even canonical *APOE* risk alleles exhibit sex-contingent penetrance, suggesting that hormonal milieu, sex-specific transcriptional regulation, or differential post-translational processing modulate their functional consequences [44, 45]. The Manhattan plot depicts a dense cluster of genome-wide significant pQTLs on chromosome 19 centered on *APOE* (Fig. 5H).

In the prediction models interface, ProteoNexus presents prediction performance comparisons between optimized and default models using ROC and AUC. If a user uploads their data, ProteoNexus returns risk estimates generated by optimized models. In the sex-combined cohort, the optimized LightGBM model achieves AUC = 0.873, up from 0.860 with the default configuration (Fig. 5I–J). In the female cohort, the optimized XGBoost improved AUC from 0.798 to 0.837, whereas the optimized LightGBM model increased AUC from 0.825 to 0.851 in the male cohort.

Benchmarking to existing analysis and platform

Current resources primarily concentrate on either basic prediction models or the provision of download links for plasma proteome research. The deCODE, UKB-PPP, and ARIC provide only summary statistics of whole-genome and *cis*-pQTL analyses, respectively [3, 5, 11]. The resource created by Deng *et al.* delineates associations between proteins and diseases using Cox regression, establishes causal relationships through MR, and constructs prediction models exclusively with LightGBM using default settings [22]. Gadd *et al.* created a web server to investigate the association between individual proteins and 23 age-related diseases and mortality using Cox regression analysis [46]. Based on a protein genetic association approach, PWAS hub explored the association between genes and 819 complex disorders solely through GWAS data and annotation information [16]. Integrating pQTL and GWAS summary statistics, Wu *et al.* proposed BILSS, a novel PWAS method, and provided a resource to explore causal proteins for 700 traits in different ancestries [47]. Additionally, there is currently neither a database that investigates the mediation effects of the proteome for exposure variables and complex diseases nor an online server that facilitates simple and user-friendly analysis of optimized prediction models for incident diseases using proteomic data.

Discussion

Here, we developed ProteoNexus, an integrative and interactive resource that combines the genetic architecture, causal mediation pathways, and predictive capacity of the human plasma proteome. We do not claim formal causal inference; mediation analyses are performed with the *medflex* R package under linear and generalized linear models and are interpreted accordingly. Based on strict data QC procedures and screening strategies, ProteoNexus leverages a sex-specific framework to profile a multi-faceted landscape of the pathways from genetic predispositions and exposure factors to the onset of diseases mediated by proteins. Based on machine learning and deep learning approaches, ProteoNexus delivers optimized predictive models. Using dementia analysis results as an example, we demonstrate detailed biostatistical insights for the three modules. For pQTL mapping, we used standard linear regression rather than linear mixed models to preserve end-to-end model-family compatibility with mediation analyses (via *medflex* and the Delta method), restricting to 33 325 genetically inferred European-ancestry participants and adjusting for technical covariates and the top 18 genetic PCs to mitigate population structure and relatedness.

A key strength of ProteoNexus lies in its integrative framework, which contrasts with existing resources that typically focus on a single analytical dimension, such as providing pQTL summary statistics or standalone predictive models. By harmonizing genetic association, mediation analysis, and risk prediction within a single platform built on prospectively collected data, ProteoNexus facilitates a more holistic understanding of disease etiology. The strict adherence to temporal causality strengthens the validity of our findings for both mediation and prediction. To maintain a uniform modeling framework across modules, incident outcomes in Step 1 were analyzed with logistic regression rather than Cox proportional hazards models, and we recommend evaluating prioritized signals with Cox models in follow-up analyses. The extensive sex-stratified analyses throughout the database address a critical gap, revealing widespread sexual dimorphism in the genetic regulation of proteins and their role in disease, which has significant implications for developing sex-specific biomarkers and therapeutic strategies [16, 17]. Model usage and sample-selection differences relative to the UKB-PPP interactive portal, including our use of linear regression instead of linear mixed models and the restriction to European-ancestry participants, may yield discrepancies in the number and identity of reported pQTL signals, and this context should guide interpretation of reliability and novelty. These choices approximate the confounding control typically afforded by mixed models while enabling coherent propagation from pQTL mapping into mediation analysis at scale.

Looking forward, the field of proteomics is poised for significant expansion, which will create opportunities to address these limitations and build upon the foundation established by ProteoNexus. UKB has announced a landmark project to measure ~5 400 proteins in samples from all 500 000 participants, including 100 000 longitudinal samples collected up to 15 years apart. The availability of this vastly large-scale dataset will provide unprecedented statistical power to discover novel pQTLs using in- and cross-ancestry fine mapping [48], refine mediation pathways, and enhance the accuracy of predictive models. Additionally, the inclusion of longitudinal proteomic data will enable the study of protein dynamics over time, of-

fering insights into the trajectory of disease development and the aging process. Future iterations of ProteoNexus will aim to incorporate these expanded datasets, which include more diverse populations, to improve the trans-ancestry portability of our findings. Furthermore, subsequent analyses will be designed to include the sex chromosomes to provide a complete and unbiased assessment of the genetic factors contributing to sex differences in disease.

ProteoNexus represents a significant step forward in translating large-scale proteogenomic data into actionable biological knowledge and epidemiological applications. In addition, its architecture lays the groundwork for the development of additional computational platforms that cater to the analysis needs of a wide variety of omics data, such as metabolomics and image-derived phenotypes [49, 50].

Acknowledgements

This study has been conducted using UK Biobank resource under Application Number 144904. We are also grateful to the participants and study staff of UK Biobank. The computational resources generously provided by the High Performance Computing Center of Nanjing Medical University are greatly appreciated.

Author Contributions Kaixin Shao (Visualization, Methodology, Validation, Software, Writing—original draft), Zixin Luo (Visualization), Peng Huang (Conceptualization, Validation), and Sheng Yang (Conceptualization, Methodology, Formal analysis, Writing—review & editing).

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

This work was supported by the Natural Science Foundation of China [No. 82273741 to S.Y. and 82173585 to P.H.] and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD). Funding to pay the Open Access publication charges for this article was provided by the Natural Science Foundation of China.

Data availability

All relevant data are available through the ProtoNexus website (https://www.proteonexus.com/). Code for ProteoNexus is on Github (https://github.com/KaishinShaw/ProteoNexus) and Zenodo (https://zenodo.org/records/16759106).

References

1. Williams SA, Kivimaki M, Langenberg C *et al.* Plasma protein patterns as comprehensive indicators of health. *Nat Med* 2019;25:1851–7. https://doi.org/10.1038/s41591-019-0665-2
2. Miller GW, Consortium BE, Bennett LM *et al.* Integrating exposomics into biomedicine. *Science* 2025;388:356–8. https://doi.org/10.1126/science.adr0544
3. Sun BB, Chiou J, Traylor M *et al.* Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* 2023;622:329–38. https://doi.org/10.1038/s41586-023-06592-6
4. Dhindsa RS, Burren OS, Sun BB *et al.* Rare variant associations with plasma protein levels in the UK Biobank. *Nature* 2023;622:339–47. https://doi.org/10.1038/s41586-023-06547-x
5. Ferkingstad E, Sulem P, Atlason BA *et al.* Large-scale integration of the plasma proteome with genetics and disease. *Nat Genet* 2021;53:1712–21. https://doi.org/10.1038/s41588-021-00978-w
6. Wang B, Pozarickij A, Mazidi M *et al.* Comparative studies of 2168 plasma proteins measured by two affinity-based platforms in 4000 Chinese adults. *Nat Commun* 2025;16:1869. https://doi.org/10.1038/s41467-025-56935-2
7. Qin H, Tromp J, Maaten JMt *et al.* Clinical and proteomic risk profiles of new-onset heart failure in men and women. *JACC: Heart Fail* 2025;13:435–49. https://doi.org/10.1016/j.jchf.2024.09.022
8. Yoshiji S, Lu T, Butler-Laporte G *et al.* Integrative proteogenomic analysis identifies COL6A3-derived endotrophin as a mediator of the effect of obesity on coronary artery disease. *Nat Genet* 2025;57:345–57. https://doi.org/10.1038/s41588-024-02052-7
9. Carrasco-Zanini J, Wheeler E, Uluvar B *et al.* Mapping biological influences on the human plasma proteome beyond the genome. *Nat Metab* 2024;6:2010–23. https://doi.org/10.1038/s42255-024-01133-5
10. Beydoun MA, Beydoun HA, Hu Y-H *et al.* Mediating and moderating effects of plasma proteomic biomarkers on the association between poor oral health problems and brain white matter microstructural integrity: the UK Biobank study. *Mol Psychiatry* 2025;30:388–401. https://doi.org/10.1038/s41380-024-02678-3
11. Zhang J, Dutta D, Köttgen A *et al.* Plasma proteome analyses in individuals of European and African ancestry identify cis-pQTLs and models for proteome-wide association studies. *Nat Genet* 2022;54:593–602. https://doi.org/10.1038/s41588-022-01051-w
12. Suhre K, McCarthy MI, Schwenk JM. Genetics meets proteomics: perspectives for large population-based studies. *Nat Rev Genet* 2021;22:19–37. https://doi.org/10.1038/s41576-020-0268-2
13. Pietzner M, Wheeler E, Carrasco-Zanini J *et al.* Mapping the proteo-genomic convergence of human diseases. *Science* 2021;374:eabj1541. https://doi.org/10.1126/science.abj1541
14. Koprulu M, Wheeler E, Kerrison ND *et al.* Sex differences in the genetic regulation of the human plasma proteome. *Nat Commun* 2025;16:4001. https://doi.org/10.1038/s41467-025-59034-4
15. Bernabeu E, Canela-Xandri O, Rawlik K *et al.* Sex differences in genetic architecture in the UK Biobank. *Nat Genet* 2021;53:1283–9. https://doi.org/10.1038/s41588-021-00912-0
16. Zucker R, Kelman G, Linial M. PWAS Hub: exploring gene-based associations of complex diseases with sex dependency. *Nucleic Acids Res* 2025;53:D1132–43. https://doi.org/10.1093/nar/gkae1125
17. Yang S, Ye X, Ji X *et al.* PGSFusion streamlines polygenic score construction and epidemiological applications in biobank-scale cohorts. *Genome Med* 2025;17:77. https://doi.org/10.1186/s13073-025-01505-w
18. Royer P, Björnson E, Adiels M *et al.* Large-scale plasma proteomics in the UK Biobank modestly improves prediction of major cardiovascular events in a population without previous cardiovascular disease. *Eur J Prev Cardiol* 2024;31:1681–9. https://doi.org/10.1093/eurjpc/zwae124
19. Yu H, Zhang J, Qian F *et al.* Large-scale plasma proteomics improves prediction of peripheral artery disease in individuals with type 2 diabetes: a prospective cohort study. *Diabetes Care* 2025;48:381–9. https://doi.org/10.2337/dc24-1696
20. Gan Y-H, Ma L-Z, Zhang Y *et al.* Large-scale proteomic analyses of incident Parkinson's disease reveal new pathophysiological insights and potential biomarkers. *Nat Aging* 2025;5:642–57. https://doi.org/10.1038/s43587-025-00818-0
21. You J, Guo Y, Zhang Y *et al.* Plasma proteomic profiles predict individual future health risk. *Nat Commun* 2023;14:7817. https://doi.org/10.1038/s41467-023-43575-7
22. Deng Y-T, You J, He Y *et al.* Atlas of the plasma proteome in health and disease in 53,026 adults. *Cell* 2025;188:253–71. https://doi.org/10.1016/j.cell.2024.10.045

23. Park H, Norby FL, Kim D *et al.* Proteomic signatures for risk prediction of atrial fibrillation. *Circulation* 2025;152:217–29.https://doi.org/10.1161/CIRCULATIONAHA.124.073457
24. Li Z, Zhou X. Towards improved fine-mapping of candidate causal variants. *Nat Rev Genet* 2025. https://doi.org/10.1038/s41576-025-00869-4
25. Steen J, Loeys T, Moerkerke B *et al.* medflex: an R package for flexible mediation analysis using natural effect models. *J Stat Soft* 2017;76:1–46. https://doi.org/10.18637/jss.v076.i11
26. Sudlow C, Gallacher J, Allen N *et al.* UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015;12:e1001779. https://doi.org/10.1371/journal.pmed.1001779
27. Yang S, Zhou X. Accurate and scalable construction of polygenic scores in large biobank data sets. *Am Hum Genet* 2020;106:679–93. https://doi.org/10.1016/j.ajhg.2020.03.013
28. Yang S, Zhou X. PGS-server: accuracy, robustness and transferability of polygenic score methods for biobank scale studies. *Briefings Bioinf* 2022;23:bbac039. https://doi.org/10.1093/bib/bbac039
29. Cao C, Zhang S, Wang J *et al.* PGS-Depot: a comprehensive resource for polygenic scores constructed by summary statistics based methods. *Nucleic Acids Res* 2024;52:D963–71. https://doi.org/10.1093/nar/gkad1029
30. Privé F, Luu K, Blum MGB *et al.* Efficient toolkit implementing best practices for principal component analysis of population genetic data. *Bioinformatics* 2020;36:4449–57.https://doi.org/10.1093/bioinformatics/btaa520
31. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* 2012;44:821–4. https://doi.org/10.1038/ng.2310
32. Zou Y, Carbonetto P, Wang G *et al.* Fine-mapping from summary data with the "Sum of Single Effects" model. *PLoS Genet* 2022;18:e1010299. https://doi.org/10.1371/journal.pgen.1010299
33. Berisa T, Pickrell JK. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* 2016;32:283–5. https://doi.org/10.1093/bioinformatics/btv546
34. The Gene Ontology Consortium. The gene ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res* 2019;47:D330–8.https://doi.org/10.1093/nar/gky1055
35. Kanehisa M, Goto S. KEGG: kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 2000;28:27–30. https://doi.org/10.1093/nar/28.1.27
36. Wu T, Hu E, Xu S *et al.* clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *The Innovation* 2021;2:100141. https://doi.org/10.1016/j.xinn.2021.100141
37. Yang S, Zhou X. SRT-server: powering the analysis of spatial transcriptomic data. *Genome Med* 2024;16:18. https://doi.org/10.1186/s13073-024-01288-6
38. Burgaletto C, Platania CBM, Di Benedetto G *et al.* Targeting the miRNA-155/TNFSF10 network restrains inflammatory response in the retina in a mouse model of Alzheimer's disease. *Cell Death Dis* 2021;12:905. https://doi.org/10.1038/s41419-021-04165-x
39. Oh HS-H, Le Guen Y, Rappoport N *et al.* Plasma proteomics links brain and immune system aging with healthspan and longevity. *Nat Med* 2025;31:2703–11. https://doi.org/10.1038/s41591-025-03798-1
40. Blew CO, Duggan MR, Joynes CM *et al.* Multi-cohort analyses link plasma GDF15 with dementia, brain atrophy, and plasma biomarkers. *Alzheimers Dement* 2025;20:e086953. https://doi.org/10.1002/alz.086953
41. Chen C, Paolillo EW, Saloner R *et al.* GDF15 as a prognostic biomarker of cognitive decline in frontotemporal dementia. *Alzheimers Dement* 2024;20:e089335. https://doi.org/10.1002/alz.089335
42. Sun X, Mews M, Wheeler NR *et al.* Preliminary insights from a multi-ancestry TWAS in Alzheimer's Disease in African and European populations. *Alzheimers Dement* 2024;20:e092593. https://doi.org/10.1002/alz.092593
43. Lu M, Zhang Y, Yang F *et al.* TWAS Atlas: a curated knowledgebase of transcriptome-wide association studies. *Nucleic Acids Res* 2023;51:D1179–87. https://doi.org/10.1093/nar/gkac821
44. Xu X, Kwon J, Yan R *et al.* Sex differences in apolipoprotein E and Alzheimer Disease pathology across ancestries. *JAMA Netw Open* 2025;8:e250562. https://doi.org/10.1001/jamanetworkopen.2025.0562
45. Ungar L, Altmann A, Greicius MD. Apolipoprotein E, gender, and Alzheimer's disease: an overlooked, but potent and promising interaction. *Brain Imaging and Behavior* 2014;8:262–73. https://doi.org/10.1007/s11682-013-9272-x
46. Gadd DA, Hillary RF, Kuncheva Z *et al.* Blood protein assessment of leading incident diseases and mortality in the UK Biobank. *Nat Aging* 2024;4:939–48. https://doi.org/10.1038/s43587-024-00655-7
47. Zhang W, Zhang X, Qiu C *et al.* An atlas of genetic effects on the monocyte methylome across European and African populations. medRxiv, https://doi.org/10.1101/2024.08.12.24311885, 14 August 2025, preprint: not peer reviewed.
48. Gao B, Zhou X. MESuSiE enables scalable and powerful multi-ancestry fine-mapping of causal variants in genome-wide association studies. *Nat Genet* 2024;56:170–9. https://doi.org/10.1038/s41588-023-01604-7
49. Smith SM, Douaud G, Chen W *et al.* An expanded set of genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nat Neurosci* 2021;24:737–45. https://doi.org/10.1038/s41593-021-00826-4
50. Ritchie SC, Surendran P, Karthikeyan S *et al.* Quality control and removal of technical variation of NMR metabolic biomarker data in ~120,000 UK Biobank participants. *Sci Data* 2023;10:64. https://doi.org/10.1038/s41597-023-01949-y

Received: August 11, 2025. **Revised:** September 16, 2025. **Accepted:** September 29, 2025