

<https://doi.org/10.1038/s43856-025-00964-x>

Multi-modal analyses of proteomic measurements associated with type 2 diabetes from the Project Baseline Health Study



Alessandra Breschi¹✉, Yuliang Wang¹, Sarah Short¹, Wilman Luk¹, David Erani², Pouya Kheradpour¹, Peter Cimermanic¹, Gary J. Tong¹, Jean Philippe Martin¹, Manway Liu¹, Lulu Cao¹, Daniel Liu¹, Raneer Chatterjee³, Lydia Coulter Kwee⁴, Thomas M. Snyder¹, Andrew Han¹, Katherine Drake¹ & Charles C. Kim¹

Abstract

Background Understanding diabetes at the molecular level can help refine diagnostic approaches and personalized treatment efforts.

Methods We generated proteomic data from plasma collected from participants enrolled in the longitudinal observational cohort study Project Baseline Health Study (PBHS) (evaluated cohort, $n = 738$, 27.9% of the total PBHS cohort), and integrated those data with information from their medical history and laboratory tests to determine diabetes status. We then identified biomarker proteins associated with diabetes status.

Results Here we identify 87 differentially expressed proteins in people with diabetes compared to those without diabetes, 71 of which show higher expression. This proteomic profile, integrated with clinical data into a logistic regression model, can discriminate diabetes status with over 85% balanced accuracy.

Conclusions Our approach indicates that proteomic data can enhance diabetes phenotyping, showing potential for marker-based stratification of diabetes diagnosis. These results suggest that a holistic molecular-clinical approach to diagnosis might help personalize treatments or interventions for people with diabetes.

Plain language summary

Diabetes is a complex disease in which people's blood sugar levels become too high. People are diagnosed and monitored using conventional blood tests. We took a group of people, analyzed their blood proteins, and used computational methods to match blood protein profiles with clinical information about who had diabetes. We could thus classify individuals in detail; we could identify people who may have blood protein profiles resembling people with diabetes, even if they do not have a diabetes diagnosis. Our method could be further developed to improve the identification of people at higher risk of developing diabetes.

Approximately 28 million people have been diagnosed with type 2 diabetes (T2D) in the US, with an additional 8.5 million people estimated to be undiagnosed¹. Current diagnostic criteria for diabetes and prediabetes involve measuring blood glucose and percentage of glycated hemoglobin (HbA1c)² to determine whether levels are above the “normal references” of 99 mg/dL and 5.6%, respectively. Common T2D phenotypes include insulin resistance and hyperglycemia, but, in the entirety of its pathology, T2D is a complex disease often associated with other systemic alterations, such as obesity, lipid metabolism alterations, hypertension, chronic inflammation, and endothelial damage³. Therefore, identifying markers to refine the stratification of diabetes phenotypes could, in turn, improve the personalization of intervention strategies.

The Project Baseline Health Study (PBHS) is a prospective, multi-center, longitudinal study including participants with diverse backgrounds and representative of a wide spectrum of health states. The PBHS cohort included participants in the United States, mainly California and North Carolina, who were followed for up to 4 years with regular annual visits. During the study, longitudinal data were collected, enabling multiple deep-phenotyping modalities, including medical history, clinical laboratory tests, molecular and digital profiling⁴. Previous research has analyzed the PBHS cohort to identify clinical characteristics of diabetes and prediabetes⁵.

The objective of our analyses was to expand the clinical characterization of T2D in the PBHS cohort by integrating proteomic and clinical profiling to identify plasma proteins associated with diabetes. We performed

¹Verily Life Sciences, South San Francisco, CA, USA. ²Onduo, P.C., Newton, MA, USA. ³Department of Medicine, Duke University School of Medicine, Durham, NC, USA. ⁴Duke Molecular Physiology Institute, Duke University School of Medicine, Durham, NC, USA. ✉e-mail: ale.breschi@gmail.com

enrichment analysis, network analysis, and transcriptomics analysis to determine which pathways were altered in participants with diabetes compared to participants with normoglycemia. Finally, we performed unsupervised and supervised machine learning modeling that combined proteomic and clinical data to assess whether the integration of molecular and clinical data could improve diabetes phenotyping compared to either modality alone. We identify differential plasma proteomic profiles for T2D and prediabetes states consistent with known features of T2D pathophysiology, based on known protein functions. The logistic regression model resulting from combining these profiles with clinical features can discriminate diabetic disease status with high accuracy without relying on traditional molecular measures such as HbA1c. We also identify normoglycemic participants and participants with prediabetes with metabolic, physical, and cardiovascular features resembling T2D.

Methods

Participants: the Project Baseline Health Study

The PBHS is a longitudinal cohort study approved by both a central Institutional Review Board (the WCG IRB; approval tracking number 20170163, work order number 1-1506365-1) and IRBs at each of the participating institutions: Stanford University, Duke University, and the California Health and Longevity Institute. This study included participants who met all PBHS eligibility criteria (key criteria were US residency and age ≥ 18 years), and all participants provided informed consent. A full description of study procedures has been previously reported⁴.

During the study visits, questionnaires collected participants' medical history information (spanning multiple disease areas including immune, metabolic and cardiovascular, mental health, neurological, infectious, and musculoskeletal) and biological samples were collected and bio-banked. Samples collected include whole blood, plasma, serum, stool, saliva, tears, urine, and facial swabs. Blood and urine samples were also submitted for standard clinical laboratory analysis, including complete blood count. Participants also underwent echocardiography and wore a Verily Study Watch (Verily Life Sciences, South San Francisco, California), which recorded acceleration data via an onboard inertial measurement unit (IMU) with a 30 Hz 3-axis accelerometer. Data included in this analysis were collected between 2017 and 2022. Participants were followed for up to 4 years.

Analyzable cohort

The analyzable cohort for this study consisted of 698 participants in the PBHS with available proteomic data and who maintained the same diagnosis throughout the study (unless otherwise noted).

The portion of this study involving data modeling included the sub-cohort of participants with complete clinical data available to enable the analysis.

Availability of proteomic data. For the present analysis, we pooled together proteomic data available from several PBHS participant subsets, corresponding to several substudies for which first-visit plasma samples were processed through mass spectrometry: first, a pilot study including a sequential set of 91 participants enrolled at study initiation. Second, a T2D pilot substudy including 330 participants with self-reported prediabetes/T2D and matched participants with normoglycemia; participants were matched 1:1 based on demographics and overall physical health (specifically based on sex at birth, age, race, blood pressure, resting pulse rate, respiratory rate, average daily step count); participants with cancer, infectious or autoimmune disease were excluded from this substudy. Third, a liver injury substudy including 154 participants with hepatic conditions and 1:1 matched controls, following the same matching strategy as in the T2D pilot study; liver injury was identified as self-reported hepatic disease, such as cirrhosis or hepatitis, or altered lab tests, such as elevated bilirubin. Fourth and last, an exploratory substudy including 394 participants randomly selected from the entire PBHS cohort.

The substudies were initiated during the initial PBHS enrollment period; thus, participants were selected based on their health status at study start. Participants were monitored for up to 4 years, and several changes in their health status were observed. After further clinical data review and accounting for health status changes, the following participants were selected for this study: 1 participant with T2D, 8 participants with prediabetes and 62 participants with normoglycemia from the first pilot study; 105 participants with T2D, 49 participants with prediabetes and 93 participants with normoglycemia from the second (T2D) pilot substudy; 32 participants with T2D, 16 participants with prediabetes and 71 participants with normoglycemia from the third (liver injury) substudy; and 17 participants with T2D, 37 participants with prediabetes and 247 participants with normoglycemia from the fourth (exploratory) substudy.

Diagnosis at study start and follow-up. We integrated two sources of information: self-reported diabetes status and results from clinical tests for HbA1c, fasting blood glucose, fasting blood glucose (FBG), and non-fasting blood glucose (nFBG) performed at the baseline study visit. Participants with pre-existing diagnoses of T2D or prediabetes, including those with HbA1c or blood glucose values outside of the disease's clinical range at study start, were classified according to the pre-existing diagnosis (assuming these may reflect cases of successful disease management). Participants without a diagnosis for T2D or prediabetes could be classified as "with T2D" or "with prediabetes" if their HbA1c or blood glucose was in the diabetic or prediabetic clinical range at study start and at the following yearly visit (diabetes defined as HbA1c $\geq 6.5\%$, or FBG ≥ 126 mg/dL or random blood glucose [RBG] ≥ 200 mg/dL; prediabetes defined as HbA1c between 5.7% and 6.4%, or FBG between 100 and 125 mg/dL)². All participants without initial T2D diagnosis who were classified as "with T2D" had HbA1c $\geq 6.5\%$. To monitor the maintenance of a given diagnosis or the occurrence of progression events to T2D or prediabetes, we followed study measurements of HbA1c and blood glucose. When HbA1c or blood glucose test results shifted to the diabetic or prediabetic clinical range for at least 2 study visits at any point, diagnoses were updated. Self-reports of initiation of diabetes medications while on the study were marked as progression events to prediabetes or T2D, depending on the indication of the medication. We considered the following diabetes medications: metformin, pioglitazone, glimepiride, exenatide, canagliflozin, empagliflozin, dapagliflozin, dulaglutide, glimepiride, glibenclamide, glipizide, glucagon, chlorpropamide, glyburide, sitagliptin, saxagliptin, linagliptin, alogliptin, semaglutide, liraglutide, and insulin. With these criteria, we observed 242 participants who developed prediabetes, 42 participants with an initial prediabetes diagnosis who reverted their HbA1c and FBS to normoglycemia, 37 participants who progressed from prediabetes to T2D, and 7 participants with normoglycemia who developed T2D. Of these, 101, 17, 15, and 4 participants had plasma samples processed through our proteomic pipeline, respectively.

Normoglycemic participants reporting taking diabetes medications for the treatment of another condition, such as polycystic ovary syndrome, were excluded from the analysis.

Proteomics

Experimental setup and data acquisition. Plasma was aliquoted from whole blood samples collected in K2 EDTA tubes, and plasma samples were processed through Verily Life Sciences' proprietary liquid chromatography-mass spectrometry (LC-MS) proteomic assay (For full details, see Supplementary Methods).

Proteomic analysis pipeline. According to the experimental design, each sample was processed as two technical replicates for each batch. The two technical replicates within the batch were injected in randomized non-consecutive order onto the LC-MS instrument. If the instrument performance was degrading during a batch, more than two replicates were processed. Custom code was used, unless specified otherwise.

Mass spectra were stored as proprietary Thermo Fisher .raw files. The spectra were analyzed to infer peptide and protein abundances (see processing steps in Supplementary Methods). Overall, 289 high-confidence protein groups were detected, with an average of 9 peptides mapping to a protein group and a median of 4 peptides.

Plasma contamination. To take into account potential biases due to different levels of plasma contamination at sample collection, contamination indices for erythrocytes, platelets, and coagulation were computed. Each contamination index was computed in each individual sample by summing the expression of the proteins in each contamination index protein signature⁶. Platelet and erythrocyte contamination was computed as the ratio of the sum of platelet and erythrocyte protein expression over the sum of all expressed proteins in each sample. Coagulation contamination was computed as the ratio of the sum of all expressed proteins over the sum of coagulation proteins in each sample.

The sample-specific contamination indices were added as confounding variables to the differential expression model.

Differential protein expression analysis

To identify differentially expressed (DE) proteins between individuals with T2D and normoglycemia, we built a linear model for each protein. The batch-corrected expression of each protein (see Supplementary Methods) was modeled as a function of the diabetic phenotype, accounting for the following potential confounding factors: sex, age, race, smoking status, presence of comorbidities, statin usage, hypertension medication usage, platelet contamination, erythrocyte contamination, and coagulation contamination. Participants self-reported as never smoking, formerly smoking, or currently smoking, which was mapped to a discrete variable in that order. The presence of self-reported comorbidities was added as a single model term. Comorbidities were: cancer, autoimmune diseases, excluding diabetes, infectious diseases, diverticulitis, pancreatitis, and pneumonia. The `ols()` function from the `statsmodels.formula.api` Python package was used to build the linear models. The *p* value associated with the diabetes phenotype coefficient was adjusted for multiple testing with the Benjamini–Hochberg correction⁷.

In addition, to test the stability of the DE proteins to changes in the sample composition, we built the linear models for 10 random subsets of 90% of the samples, allowing resampling across the subsets. Thus, for each protein, we obtained 10 false discovery rate (FDR)-adjusted *p* values, one for each of the random subsets. Finally, a protein was considered DE if the FDR-adjusted *p* value was less than 0.05 across all the random subsets.

Gene ontology (GO) term enrichment. The gene ontology (GO) annotation from January 2023 was used to compute GO term enrichment on the DE proteins. We limited the GO annotations to terms with experimental evidence, manual and electronic annotation or inferred from sequence or structural similarity, corresponding to the following evidence codes: EXP (inferred from experiment), IDA (inferred from direct assay), IPI (inferred from physical interaction), IMP (inferred from mutant phenotype), IGI (inferred from genetic interaction), IEP (inferred from expression pattern), TAS (traceable author statement), IC (inferred by curator), IEA (inferred from electronic annotation), ISS (inferred from sequence or structural similarity). Only GO terms with at least 3 proteins represented in our data were tested for enrichment. A hypergeometric test was performed to test the enrichment for each annotated GO term within the biological process and cellular component namespaces. Upregulated and downregulated proteins in individuals with T2D (compared to normoglycemic) were tested for GO enrichment separately. The *p* value was adjusted for multiple testing with the Benjamini–Hochberg correction⁷ separately for each namespace and each protein set. The list of all detected plasma proteins was used as the background set for the hypergeometric test.

A protein could be annotated with more than one GO term. To annotate the proteins uniquely with one GO term on a heatmap, the following custom GO slim terms were assigned in this order: lipid transport,

complement activation, blood coagulation, inflammatory response, and immune system process.

Protein analyses: protein–protein interaction (PPI) networks and tissue specificity. Protein–protein interactions (PPIs) were exported from the STRING database v11.5⁸. Only high-confidence interactions were included (minimum combined score of 500⁹). In addition, only PPIs between positively co-expressed DE proteins were included (Pearson's correlation coefficient between protein expression values across all participants with T2D and normoglycemia ≥ 0.2). The resulting PPI network was finally filtered to restrict to a core of at least 2 degrees for each node. This ensured a certain level of network connectivity.

Louvain's community detection algorithm¹⁰ was applied to the final PPI network. Each community was annotated with the custom GO slim categories described above. The Python package `networkx`¹¹ was used for network analysis.

We used the Genotype-Tissue Expression (GTEx) database to examine DE protein expression patterns. Because some genes in GTEx can be specific to multiple tissues¹², tissue-specific genes encoding for DE proteins were selected using increasingly stringent tissue-specificity thresholds (tissue-specificity score > 3 or > 4). In addition, the tissue assignment was deduplicated by assigning the gene to the tissue with the highest tissue-specificity score.

Single-cell liver RNA-seq analysis

Single-cell RNA-seq (scRNA-seq) data obtained from the liver of healthy donors were downloaded from the GSE185477 GEO study¹³. Liver cells from multiple healthy donors are pooled into the same dataset. The authors provided single-cell type annotation, normalized read counts at the single-cell level, and Uniform Manifold Approximation and Projection (UMAP) projection values. For each DE protein expressed in liver, bulk RNA-seq from GTEx (FPKM > 1), we computed the average Z-score of gene expression in the liver scRNA-seq dataset scaled across the set of hepatic cell types, namely hepatocytes, cholangiocytes, and stellate cells. UMAP values are taken directly from the original dataset.

Clustering of participants at study start

Clustering analysis was performed on 110 participants with prediabetes, 155 with diabetes, and 467 normoglycemic, with clinical and proteomic data at study start. Supervised principal component analysis (PCA) was performed on filtered clinical and proteomic features before clustering (see details below and in Supplementary Methods).

Clinical features included clinical and demographic variables (sex, age, and race). Self-reported race categories were Asian, Black or African American, Hispanic, white, or other. Clinical features measured from standard blood and urine tests and vitals were manually curated to remove redundancy and avoid missingness. To avoid collinearity in measurements, the manual curation removed results from laboratory measurements known to be clinically related to or derived from each other and confirmed to be correlated with each other in the current cohort (Pearson correlation > 0.8 , additional details in Supplementary Methods).

The selected clinical features were concatenated to the matrix of batch-corrected expression values of variance-filtered proteins (see Supplementary Methods), and used as input for PCA.

Different combinations of the number of principal components, clustering algorithms, and *k* number of clusters were evaluated by computing commonly used clustering metrics. Specifically, we tried the following combinations: First, a combination of 3, 10, 15, 30, or 50 principal components, based on the percentage of variance explained by each number of components. Second, a combination of *k*-means clustering, hierarchical clustering with Ward clustering and Euclidean distance, and hierarchical clustering with average clustering and Euclidean distance. Third combination of *k* = {2, 3, 4, 5, 6, 7, 8, 9, 10} clusters.

The clustering metrics we computed were: First, within-cluster sum-of-squares, it measures the variability of the observations within each cluster; in

general, a cluster that has a small sum of squares is more compact than a cluster that has a large sum of squares. Second, Silhouette score¹⁴; mean of the Silhouette Coefficients for each sample; the score is bounded between -1 for incorrect clustering and $+1$ for highly dense clustering; scores around zero indicate overlapping clusters; the score is higher when clusters are dense and well separated, which relates to a standard concept of a cluster. Third, the Calinski–Harabasz index¹⁵, the index is the ratio of the sum of between-clusters dispersion and of within-cluster dispersion for all clusters (where dispersion is defined as the sum of distances squared); a higher Calinski–Harabasz score relates to a model with better defined clusters. Fourth, adjusted Mutual Information (MI) index¹⁶, it is an adjustment of the MI score to account for chance; it accounts for the fact that the MI is generally higher for two clusterings with a larger number of clusters, regardless of whether there is actually more information shared. Fifth, Davies–Bouldin index¹⁷, this index signifies the average “similarity” between clusters, where the similarity is a measure that compares the distance between clusters with the size of the clusters themselves; zero is the lowest possible score; values closer to zero indicate a better partition.

Based on clustering performance across the different clustering metrics and interpretability, we selected k-means clustering on 30 principal components, with $k = 3$.

Clustering validation with orthogonal features

To evaluate if the clusters obtained were also related to features not included in the clustering input feature set (not related to blood work), we examined differences of these orthogonal features between diabetic-like and normoglycemic-like clusters within each clinical phenotype.

Methylation age. Methylated DNA was measured using the Illumina EPIC 850K array from DNA extracted from

frozen, stored whole blood collected at enrollment (see¹⁸ for details).

DNA methylation-derived ages were predicted using coefficients supplied by Horvath using a linear combination of the coefficients and the corresponding beta value in each sample¹⁹. An adjustment was made for non-adult age, as described in the corresponding manuscript. Missing values were filled in using a standard value provided by the authors¹⁹.

Physical activity data. As part of PBHS assessments, participants underwent standard physical performance challenges, including 6-min walk test, 10-m walk tests (fast pace and comfortable pace), 30-s chair stand. Moreover, the average number of daily steps during daily living was computed using the data collected from the Verily Study Watch²⁰. For each day of the week, Monday to Sunday, the median number of daily steps on that day of the week was computed over 90 days. Only days with at least 720 min of watch-wearing time were included in the median calculation. The medians were averaged to obtain an average daily step count.

Echocardiographic measurements. Each study site performed echocardiography with quality control by the Duke Clinical Research Institute Imaging Core Laboratory. Images were analyzed according to best practices and the American Society of Echocardiography recommendations for chamber quantification and assessment of diastolic dysfunction (detailed methods previously published²¹).

Machine learning models of type 2 diabetes

We built three T2D classification models using three different sets of input features for 155 participants with T2D and 467 participants with normoglycemia who had both clinical and proteomic data: clinical only, proteomics only, and clinical and proteomics combined (see details in Supplementary Methods).

Model interpretation with SHAP values

Feature importance for the diabetes prediction model using the combined dataset was assessed by analyzing the SHapley Additive exPlanations

(SHAP) values²² in the prediabetic population at study start. The SHAP values for the prediabetic participants were computed from the model trained on the entire cohort of normoglycemic and diabetic participants, as defined above. Examining the SHAP values associated with a model can reveal what features are driving the model prediction for each observation in the dataset.

To summarize the SHAP values of the protein features, SHAP values for groups of functionally related proteins were added together. This was possible because of the additive nature of SHAP values²². The groups of functionally related proteins were manually curated from GO term annotation and domain expert knowledge.

Results

Study population and molecular data generation

Of 2502 participants enrolled in the PBHS, 174 were initially excluded due to inconclusive reports for phenotypic assignment, and 78 due to having conditions incompatible with this analysis (latent autoimmune diabetes of adults, 2; type-1 diabetes, 20; history of gestational diabetes, 56) (Table 1, Supplementary Fig. 1A–C). Additionally, 5 participants reported developing gestational diabetes during the course of the study and were also excluded (Supplementary Fig. 1A–C).

Self-reported medical conditions were complemented with on-study laboratory results to identify participants with T2D and prediabetes. Specifically, participants without a T2D or prediabetes diagnosis were considered to have T2D or prediabetes if their HbA1c or blood glucose was in the diabetic or prediabetic clinical range at study start and at the following yearly visit (diabetes defined as HbA1c $\geq 6.5\%$, or FBG ≥ 126 mg/dL or RBG ≥ 200 mg/dL; prediabetes defined as HbA1c between 5.7% and 6.4%, or FBG between 100 and 125 mg/dL)². To monitor the consistency of a given diagnosis over time, or the occurrence of progression to T2D or prediabetes, we followed study measurements of HbA1c and blood glucose, updating diagnoses when HbA1c or blood glucose test results shifted to the diabetic or prediabetic clinical range for at least 2 study visits at any point (see “Methods” for additional details). After excluding those whose diagnoses shifted on study, the evaluable population consisted of 1915 participants, including 1319 with normoglycemia, 335 with prediabetes, and 263 with T2D (Supplementary Fig. 1A–C). Both the cohort with T2D and the cohort with prediabetes generally had higher proportions of participants who were male, older, self-reported as Black, with hypertension, and on hypertension medication. Cohort participants with T2D also had higher RBG and HbA1c levels than the overall population (Table 1).

In a complementary effort, LC-MS proteomics was performed on plasma samples collected at study start from 738 participants (Supplementary Fig. 1C). Of the eligible population above, plasma samples for proteomic analyses from the first study visit were available for 473 normoglycemic, 110 prediabetic, and 155 diabetic participants. Baseline differences between this “evaluated with proteomics” subcohort, the originating set of T2D, prediabetes, and normoglycemia cohorts, and the overall PBHS population were consistent throughout.

The quality of LC-MS data was assessed via commonly computed quality metrics. In particular, we observed a median coefficient of variation of 0.07 and an average of 20 missing proteins across all samples (Supplementary Fig. 2A, B). Moreover, we observed a high correlation between C-Reactive Protein quantified by LC-MS and by standard clinical blood test (Spearman’s correlation coefficient = 0.96, Supplementary Fig. 2C).

Participants with T2D had upregulation in inflammation-related proteins

To characterize the circulating proteome in participants with diabetes, we compared protein expression between plasma samples of participants with T2D and normoglycemia. After QC filtering (Methods), a total of 289 proteins were detected across all samples. Of these, we identified 87 DE proteins (Fig. 1a, b; Supplementary Data 1), after adjusting for demographic and clinical confounding variables, proteomic batch, and plasma contamination, and correcting for multiple testing and stability ($FDR \leq 0.05$

Table 1 | Demographic breakdown of study cohort

Characteristic		Total PBHS N = 2502	Evaluable, N = 1915			Evaluated with proteomics at start, n = 738		
			Normoglycemic n = 1319	Prediabetes n = 335	T2D n = 263	Normoglycemic n = 473	Prediabetes n = 110	T2D n = 155
Sex, n (%)	Male	1375 (55.0)	712 (54.0)	182 (54.3)	132 (50.2)	220 (46.5)	57 (51.8)	81 (52.3)
	Female	1127 (45.0)	607 (46.0)	153 (45.7)	131 (49.8)	253 (53.5)	53 (48.2)	74 (47.7)
Mean age, years (SD)		50.0 (17.2)	44.3 (16.5)	59.7 (14.5)	59.6 (12.7)	43.6 (15.3)	58.7 (13.6)	58.2 (12.4)
Age, n (%)	18–29	398 (15.9)	319 (24.2)	10 (3.0)	5 (1.9)	112 (23.7)	1 (0.9)	4 (2.6)
	30–39	451 (18.0)	317 (24.0)	28 (8.4)	18 (6.8)	110 (23.3)	11 (10.0)	10 (6.5)
	40–49	411 (16.4)	232 (17.6)	42 (12.5)	40 (15.2)	93 (19.7)	17 (15.5)	28 (18.1)
	50–59	442 (17.7)	186 (14.1)	75 (22.4)	60 (22.8)	76 (16.1)	25 (22.7)	36 (23.2)
	60–69	399 (15.9)	145 (11.0)	77 (23.0)	76 (28.9)	55 (11.6)	28 (25.5)	49 (31.6)
	70+	401 (16.0)	120 (9.1)	103 (30.7)	64 (24.3)	27 (5.7)	28 (25.5)	28 (18.1)
Race, n (%)	White	1590 (63.5)	883 (66.9)	200 (59.7)	148 (56.3)	310 (65.5)	64 (58.2)	87 (56.1)
	Black/African American	400 (16.0)	138 (10.5)	78 (23.3)	77 (29.3)	59 (12.5)	28 (25.5)	47 (30.3)
	Asian	272 (10.9)	147 (11.1)	38 (11.3)	22 (8.4)	58 (12.3)	12 (10.9)	12 (7.7)
	Hispanic	88 (3.5)	54 (4.1)	9 (2.7)	7 (2.7)	12 (2.5)	4 (3.6)	3 (1.9)
	Mixed	70 (2.8)	45 (3.4)	4 (1.2)	3 (1.1)	16 (3.4)	1 (0.9)	2 (1.3)
	American Indian/Alaskan	31 (1.2)	21 (1.6)	2 (0.6)	1 (0.4)	6 (1.3)	1 (0.9)	0
	Hawaiian/Pacific Islander	28 (1.1)	14 (1.1)	3 (0.9)	3 (1.1)	9 (1.9)	0	2 (1.3)
	Other	23 (0.9)	17 (1.3)	1 (0.3)	2 (0.8)	3 (0.6)	0	2 (1.3)
RBG, mean (SD)		97.7 (36.3)	87.4 (11.2)	93.8 (13.8)	150.3 (70.0)	88.0 (12.8)	95.7 (15.0)	153.0 (68.8)
HbA1c, mean (SD)		5.7 (1.0)	5.2 (0.3)	5.9 (0.2)	7.5 (1.8)	5.2 (0.3)	5.9 (0.3)	7.4 (1.7)
BMI, mean (SD)		28.4 (6.8)	26.8 (5.8)	29.2 (6.5)	34.3 (7.9)	27.2 (6.0)	30.7 (6.5)	34.6 (7.4)
Comorbidities, n (%)		988 (39.5)	498 (37.8)	145 (43.3)	104 (39.5)	171 (36.2)	39 (35.5)	39 (25.2)
Hypertension, n (%)		682 (27.3)	194 (14.7)	122 (36.4)	164 (62.4)	76 (16.1)	49 (44.5)	101 (65.2)
Hypertension medication, n (%)		596 (23.8)	146 (11.1)	106 (31.6)	164 (62.4)	56 (11.8)	40 (36.4)	99 (63.9)
Statins, n (%)		418 (16.7)	90 (6.8)	92 (27.5)	124 (47.1)	35 (7.4)	38 (34.5)	69 (44.5)

Summary statistics were computed for the entire PBHS cohort and for the PBHS participants with proteomic data generated from plasma collected during the initial study start visit. Clinical and medication status was evaluated at study start. Comorbidities were: cancer, autoimmune diseases, excluding diabetes, infectious diseases, diverticulitis, pancreatitis, and pneumonia.

BMI body mass index, HbA1c glycated hemoglobin, RBG random blood glucose, SD standard deviation, T2D type 2 diabetes.

across all 10 bootstrapped samples, Supplementary Fig. 3, see “Methods”). Seventy-one DE proteins (82%) showed higher expression levels in participants with diabetes, while only 16 (18%) had lower expression (Fig. 1c, d). GO enrichment analysis showed that proteins more abundant in participants with T2D were involved in the complement system (odds ratios 3.1 and 3.6, and adjusted *p* values 0.002 and 0.004 for classical and alternative complement activation pathways, respectively). No significant enrichment was found for proteins that were less abundant in participants with T2D, although many of these proteins were involved in lipid transport, especially high-density lipoproteins (HDLs) (Fig. 1b, Supplementary Fig. 4).

We analyzed PPIs from the STRING database^{8,23} for this set of DE proteins. There were four main complexes in the PPI network of DE proteins: two complement sub-complexes, a blood coagulation complex and an apolipoprotein complex, consistent with the GO enrichment results (Fig. 1c). Both upregulated (LDL) and downregulated (HDL) apolipoproteins were present in the same PPI community, since some of them, such as apolipoprotein C, exchange freely between lipoprotein complexes²⁴.

DE proteins are secreted by the liver and exhibit zonation expression patterns

The liver plays a large role in blood protein synthesis²⁵. Consistent with this, nearly all (48 out of 50, or 96%) of the DE proteins with liver expression (Fig. 2a) were also previously identified in peripheral blood secretome analysis²⁶. To further characterize a potential relationship of DE plasma

proteins with liver dysfunction in T2D, we investigated their spatial expression patterns using a single-cell RNA-seq liver atlas from healthy donors¹³.

Clustering reveals phenotypic profiles beyond diagnosis

We combined clinical and proteomic data to explore if we could identify participants with diabetic features based on clinical elements and protein abundance (that is, beyond HbA1c and blood glucose).

In particular, we focused on clinical features measured from standard blood tests and vitals, removing highly correlated features (Supplementary Fig. 5, “Methods”). As expected, several clinical features were significantly associated with diabetes diagnosis, with the top three after HbA1c and blood glucose being waist circumference, body mass index (BMI), and triglycerides (Supplementary Fig. 6, “Methods”). To account for technical sources of variation in protein abundance, proteins were filtered by the proportion of explained variance by each potential confounder (Supplementary Fig. 7, Supplementary Methods). PCA based on these selected clinical and protein features showed that participants followed a gradient in the projected UMAP space, rather than clearly defined phenotypic clusters (Fig. 3a). We observed that the first two principal components already explained ~25% of the variance, and the first 30 components captured over 60% of the variance (Supplementary Fig. 8A). Then, we applied several clustering algorithms, different numbers of principal components for dimensionality reduction before clustering, and different numbers of clusters *k* to explore how

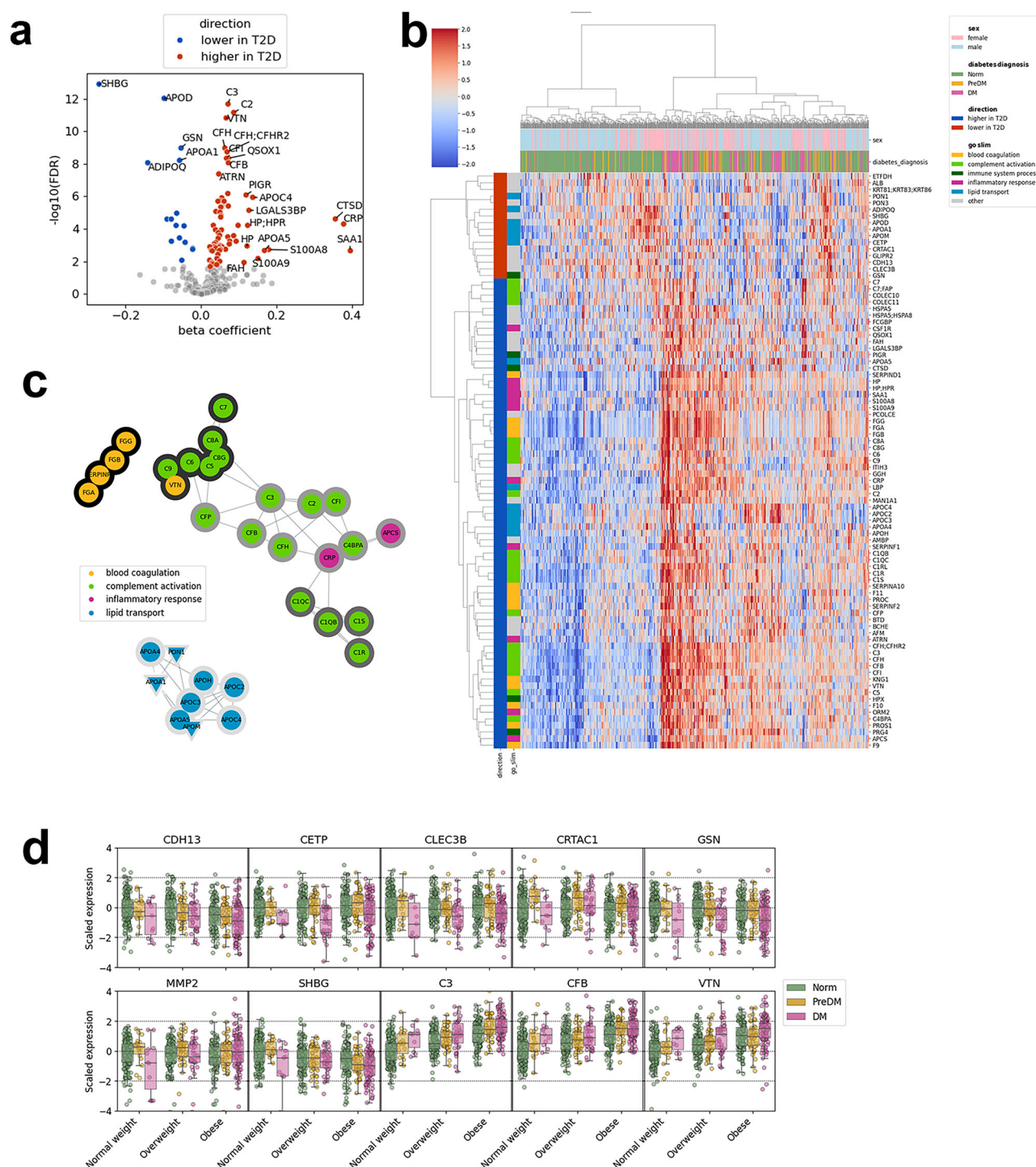


Fig. 1 | Differentially expressed proteins between diabetic and normoglycemic participants. **a** Volcano plot highlighting differentially expressed proteins. The x-axis refers to the phenotype-associated coefficient in the linear model for each protein. All samples are considered to compute the beta coefficient. **b** Heatmap showing the Z-score of batch-corrected expression of DE proteins. Hierarchical clustering was obtained with Euclidean distance and complete linkage. **c** PPI network of co-expressed DE proteins after 2-core filtering. Communities are detected with the Louvain algorithm. Nodes are colored based on the selected GO Slim. The border of the node is colored based on the assigned community. Circles: higher in

T2D; Triangles: lower in T2D. **d** Expression of DE proteins between diabetic and normoglycemic individuals with normal weight. Expression is scaled to normoglycemic median expression; line bars extend to $\pm 1.5 \times$ interquartile range (default seaborn boxplot parameter). [Sample sizes: Normal weight, $n = 231$ {Norm, 201; PreDM, 19; DM, 11}; overweight, $n = 226$ {Norm, 143; PreDM, 45; DM, 38}; obese, $n = 275$ {Norm, 123; PreDM, 46; DM, 106}]. DE differentially expressed, DM diabetes mellitus, FDR false discovery rate, GO gene ontology, Norm normoglycemic, PPI protein-protein interaction, preDM prediabetic, T2D type 2 diabetes.

unsupervised clusters relate to the clinically defined phenotypes (Supplementary Fig. 8, “Methods”). K-means clustering with 3 clusters ($k = 3$) resulted in good overlap between clusters and phenotypes, regardless of the number of principal components (Fig. 3b; Supplementary Fig. 8B, C) and

exhibited a good Silhouette profile (Fig. 3c). Based on the overlap between clusters and phenotypes, we assigned “normoglycemic-like”, “diabetes-like” and “intermediate” cluster labels (Fig. 3d). Interestingly, 105 (22%) and 178 (38%) participants with clinically defined normoglycemia were grouped

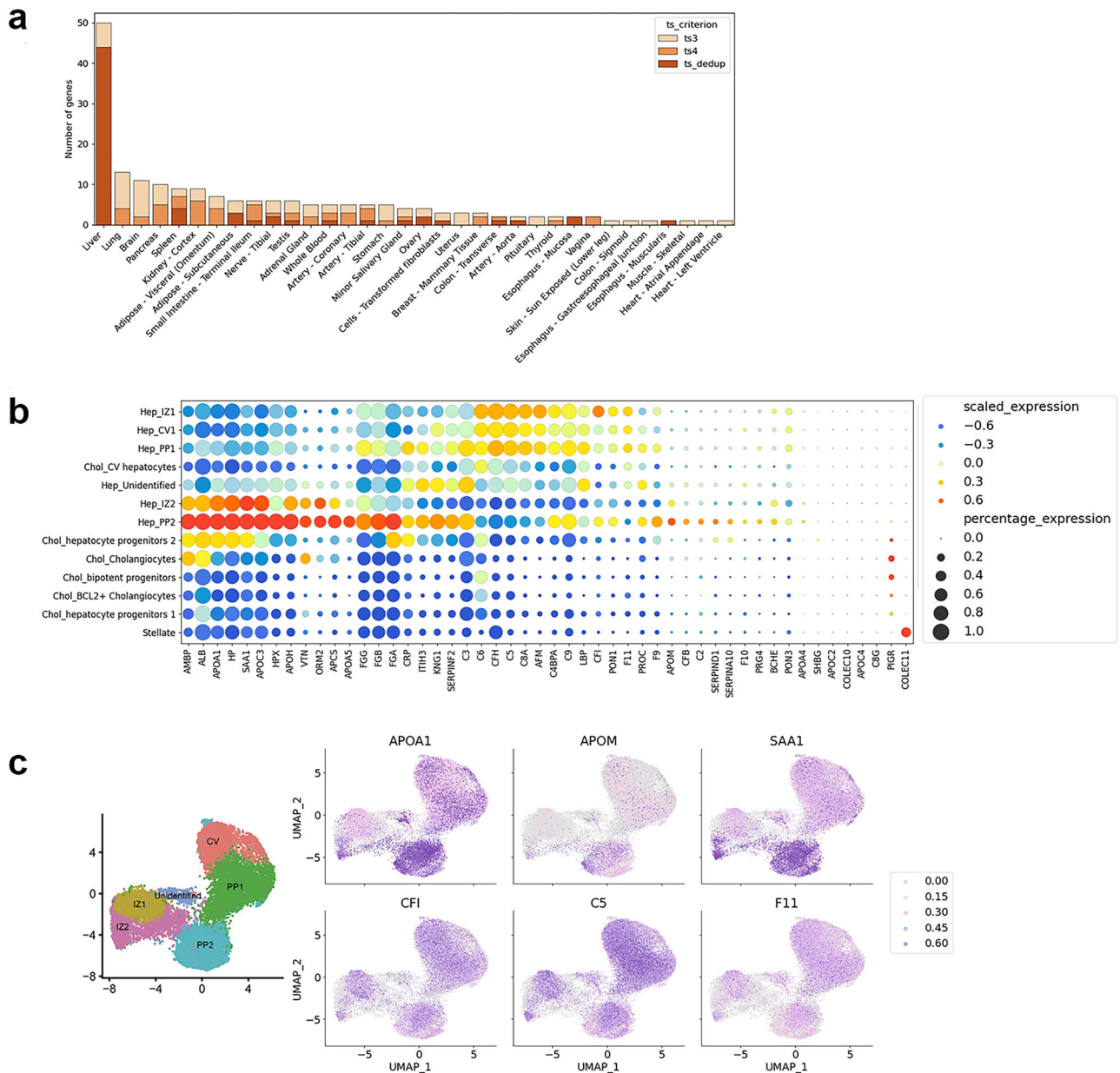


Fig. 2 | Transcriptomics analysis of DE proteins in the liver. a Distribution of tissue-specific DE proteins at the transcriptional level in the GTEx dataset. **b** Gene expression of DE proteins in scRNA-seq data from hepatocytes from different regions of the liver. Dot color is proportional to the scaled, normalized gene expression values. The size of the dots is proportional to the proportion of cells of a

given type expressing the gene. **c** UMAP of selected DE proteins showing their zonation patterns of gene expression¹³. Sample size: experiments in 50259 samples. chol cholangiocyte, DE differentially expressed, dedup duplicates removed, GTEx Genotype-Tissue Expression Project, hep hepatocyte, TS transcriptional, UMAP Uniform Manifold Approximation and Projection.

with the diabetes-like cluster and to the intermediate cluster, respectively (Fig. 3d).

To investigate which proteomic features are associated with the clusters, we were particularly interested in proteins that might already be altered in some, potentially undiagnosed, participants with normoglycemia. Within normoglycemic participants, we performed differential expression of plasma proteins between participants assigned to the normoglycemic-like and diabetes-like clusters. Out of the DE proteins identified above, we found 48 proteins, most of which were overexpressed in plasma samples of participants with normoglycemia assigned to the diabetes-like cluster ($FDR \leq 0.05$, $|\text{coefficient}| \geq 0.1$, Fig. 3e). Many of these proteins are involved in immune response, which might signal higher inflammation in a subset of normoglycemic participants.

Clusters are also associated with differences in physical performance and echocardiogram

To help demonstrate the potential clinical relevance of the clusters, we examined differences between the cluster groups at the metabolic, physical performance, and cardiac health level within each phenotype. We looked at the distribution of metabolic, physical performance and cardiac features across phenotypes and clusters for each sex, although tests for statistical significance were performed considering the two sexes together because of limited sample size (Fig. 4; Supplementary Fig. 9). Of these, biological age, predicted from DNA methylation assay, physical performance features, excluding pulse rate, and echocardiogram-derived features were not part of the clustering input features, thus representing an orthogonal validation to the cluster assignment.

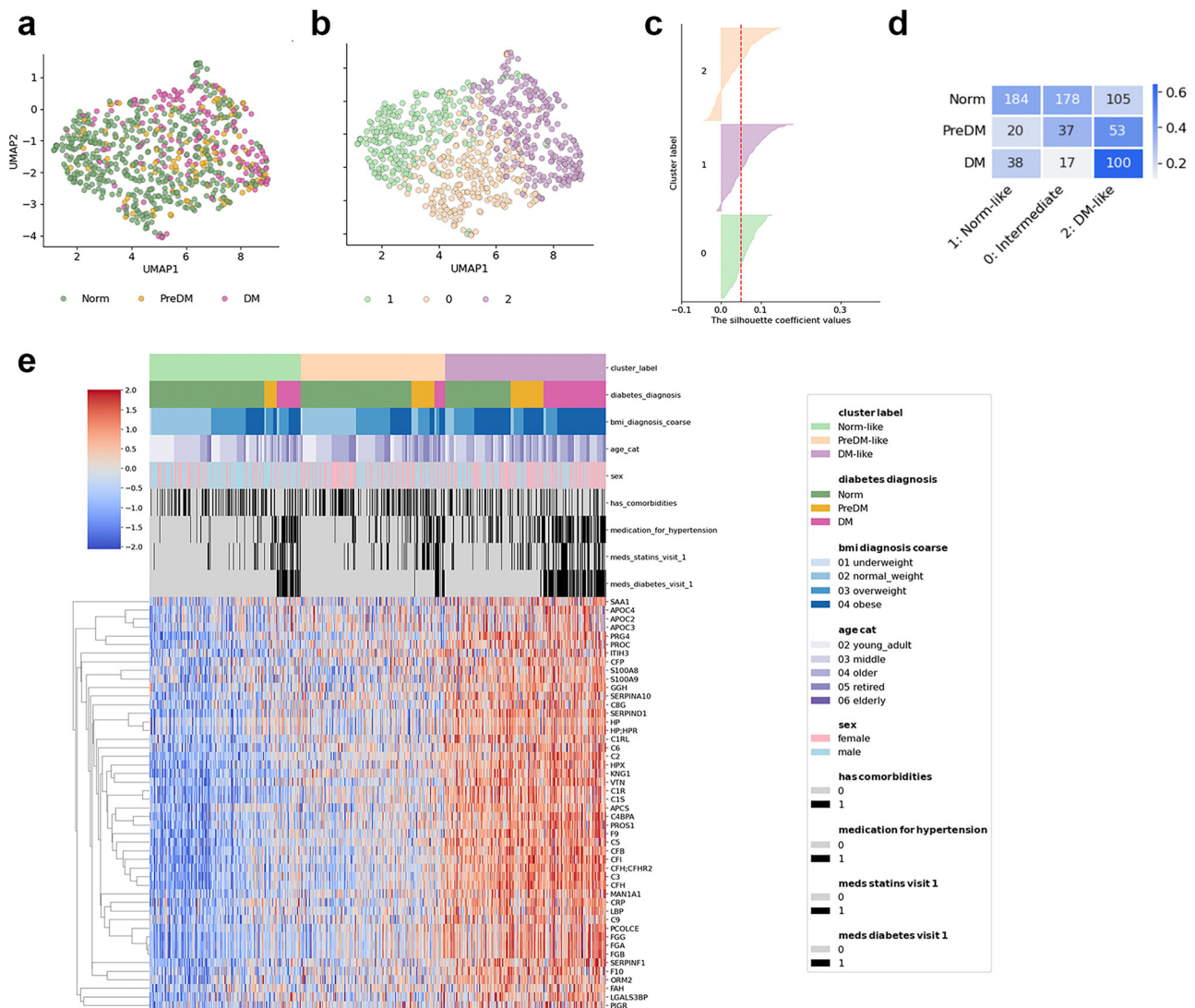


Fig. 3 | Clustering analysis of diabetic, prediabetic, and normoglycemic participants based on clinical and proteomic data. UMAP embeddings of selected proteins and clinical features after PCA, colored by diabetic status (a) and k-means cluster labels (b). c Silhouette profile of participants assigned to three clusters with k-means clustering on proteomic and clinical data projected in a 15-dimensional space with supervised PCA. d Number of participants assigned to each cluster, by diabetic status, colored by relative distribution across diabetic status within each cluster. e Normalized expression of DE proteins between normoglycemic and

diabetic participants and between participants assigned to the normoglycemic-like or the diabetic-like clusters. DE proteins are filtered at $FDR \leq 0.05$ and absolute coefficient ≥ 0.1 . Participants are sorted by cluster, diabetic status, BMI, diagnosis, and age, in this order. Sample size, $n = 732$. BMI body mass index, DE differentially expressed, DM diabetes mellitus, FDR false discovery rate, Norm normoglycemic, PCA principal component analysis, preDM prediabetic, UMAP Uniform Manifold Approximation and Projection.

Several metabolic features were significantly different across cluster groups. HbA1c had clinically minor, but statistically significant, differences between cluster groups especially within the T2D phenotype, suggesting that the proteins used to assign cluster groups may provide some additional value for further differentiating diabetes subtypes (Fig. 4a; Supplementary Fig. 9A). Highly significant differences in BMI and chronological age were observed between participants with normoglycemia assigned to the normoglycemic-like and diabetes-like clusters (Fig. 4a). We also observed significant trends in methylation age and systolic blood pressure within the group of patients with normoglycemia, where participants in the diabetes-like subgroup were older and had higher systolic blood pressure than in the normoglycemic-like subgroup (Fig. 4a).

Features associated with physical performance were also significantly different between participants classified as diabetes-like and normoglycemic-like, especially within the normoglycemic group (Fig. 4b). Since pulse rate was the only one of the physical performance features used for clustering, this finding suggests that the additional clinical and molecular

markers of diabetes we identified might also be related to physical performance overall. The average daily step count from the Verily Study Watch was higher for participants with T2D in the normoglycemic-like subgroup compared to those in the diabetes-like subgroup, albeit not significantly (Fig. 4b).

Finally, since diabetes is often associated with cardiovascular comorbidities²⁷, we compared the distribution of features derived from echocardiogram images between cluster groups for each phenotype. We focused on measurements related to left ventricular size and mitral valve blood flow, since alterations in these have been previously reported in patients with diabetes²⁸. Indeed, left ventricular mass and left ventricular septal thickness were significantly higher in participants in the diabetes-like normoglycemic subgroup compared to the normoglycemic-like normoglycemic subgroup (Fig. 4c). This could be a sign of ventricular hypertrophy, which is associated with hypertension²⁹ and is common in clinically diagnosed T2D patients³⁰. Measurements of mitral valve blood flow were significantly different in participants in the diabetes-like normoglycemic

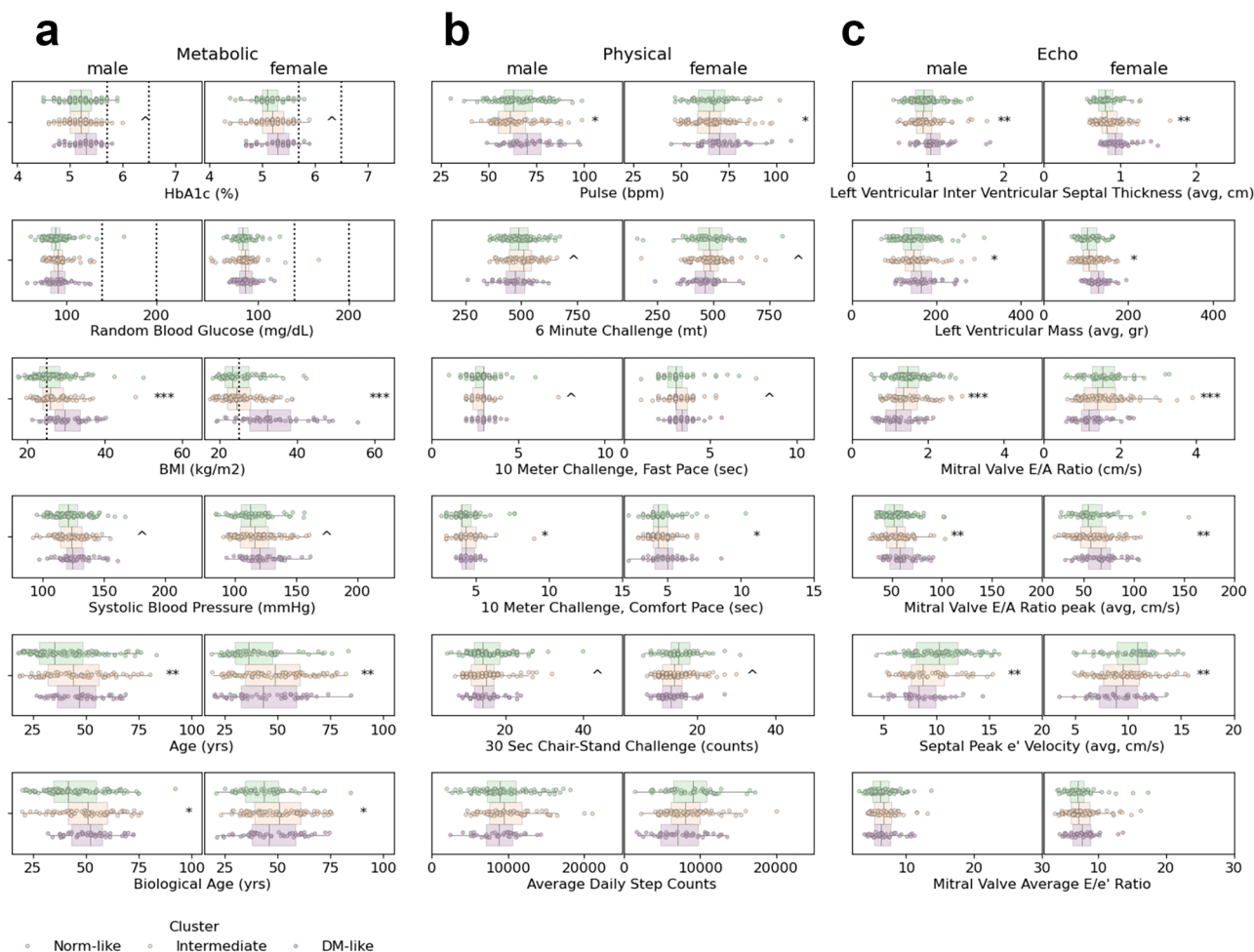


Fig. 4 | Differences in external features between participants with normoglycemia assigned to different clusters. FDR-adjusted p value comparing features between normoglycemic-like and diabetic-like clusters for each observed phenotype: \wedge : 0.01–0.05; *: 0.001–0.01; **: 0.0001–0.001; ***: <0.0001 (Mann–Whitney test; full series of exact p values in Supplementary Data 3). Because of sample size, female and male participants are pooled together for testing; line bars extend to $\pm 1.5 \times$ inter-quartile range (default seaborn boxplot parameter). Features are distributed into three main groups based on feature type: **a** metabolic features [Sample numbers: HbA1c, $n = 467$; Random Blood Glucose, $n = 467$; BMI, $n = 467$; Systolic Blood

Pressure, $n = 467$; Age, $n = 467$; Biological Age, $n = 438$]; **b** physical performance features [Sample numbers: Pulse, $n = 467$; 6-min Challenge, $n = 441$; 10-m Challenge, Fast Pace, $n = 399$; 10-m Challenge, Comfort Pace, $n = 392$; 30-sec Chair-Stand Challenge, $n = 459$; Average Daily Step Count, $n = 380$]; and **c** derived features from echocardiogram [Sample numbers: Left ventricular Inter-ventricular Septal Thickness, $n = 452$; Left Ventricular Mass, $n = 452$; Mitral Valve E/A Ratio, $n = 454$; Mitral Valve E/A Ratio Peak, $n = 454$; Septal peak e' Velocity, $n = 375$; Mitral Valve E/e' Ratio, $n = 373$]. BMI body mass index, FDR false discovery rate, HbA1c percentage of glycated hemoglobin.

subgroup compared to the normoglycemic-like subgroup: mitral valve E/A ratio and septal peak e' velocity were lower, while mitral valve E/A ratio peak was higher (Fig. 4c).

Clinical and proteomic features combined best discriminate diabetes status

Having observed significant differences at the clinical and molecular level between participants with T2D and normoglycemia, we sought to compare the ability of different feature sets to differentiate T2D from normoglycemia without using HbA1c or blood glucose in the models (these were initially used to refine the clinical diabetes phenotype and might lead to inflated performance of any model). We built three models using three sets of features: clinical features only, proteomic features only, and clinical and proteomic features combined (Supplementary Fig. 10A; “Methods”). We trained a ridge-logistic regression model on participants with normoglycemia ($n = 467$, 77%) and T2D ($n = 155$, 23%) with clinical and proteomic data to discriminate diabetes status in a repeated cross-validation setting for hyperparameter tuning (Supplementary Fig. 10B), and evaluated the performance using repeated nested cross-validation on the same dataset³¹ (Supplementary Fig. 10C). To

address the high data dimensionality, feature selection was performed as a preprocessing step inside the cross-validation pipelines for the proteomics only and combined datasets, with the number of selected features being a hyperparameter to tune (Methods). Clinical features were already filtered by excluding highly correlated features as described (“Methods” and Supplementary Methods).

We compared model performance between the datasets by testing the differences across several performance metrics within the repeated nested cross-validation setting (Fig. 5a). The distribution of model performance across several metrics was similar between the metrics computed within the repeated cross-validation and the repeated nested cross-validation settings (Fig. 5a). The model using the combined dataset performed best consistently across all the performance metrics, except for precision (Fig. 5a). In this context, precision, also called positive predictive value, measures the proportion of patients actually clinically defined as “with T2D” within the overall group of those classified “with T2D.” Therefore, lower precision would imply more individuals without T2D were classified as having T2D. This is consistent with our hypothesis that incorporating additional protein biomarkers could increase the resolution of T2D subtypes over the current clinical diagnosis of T2D.

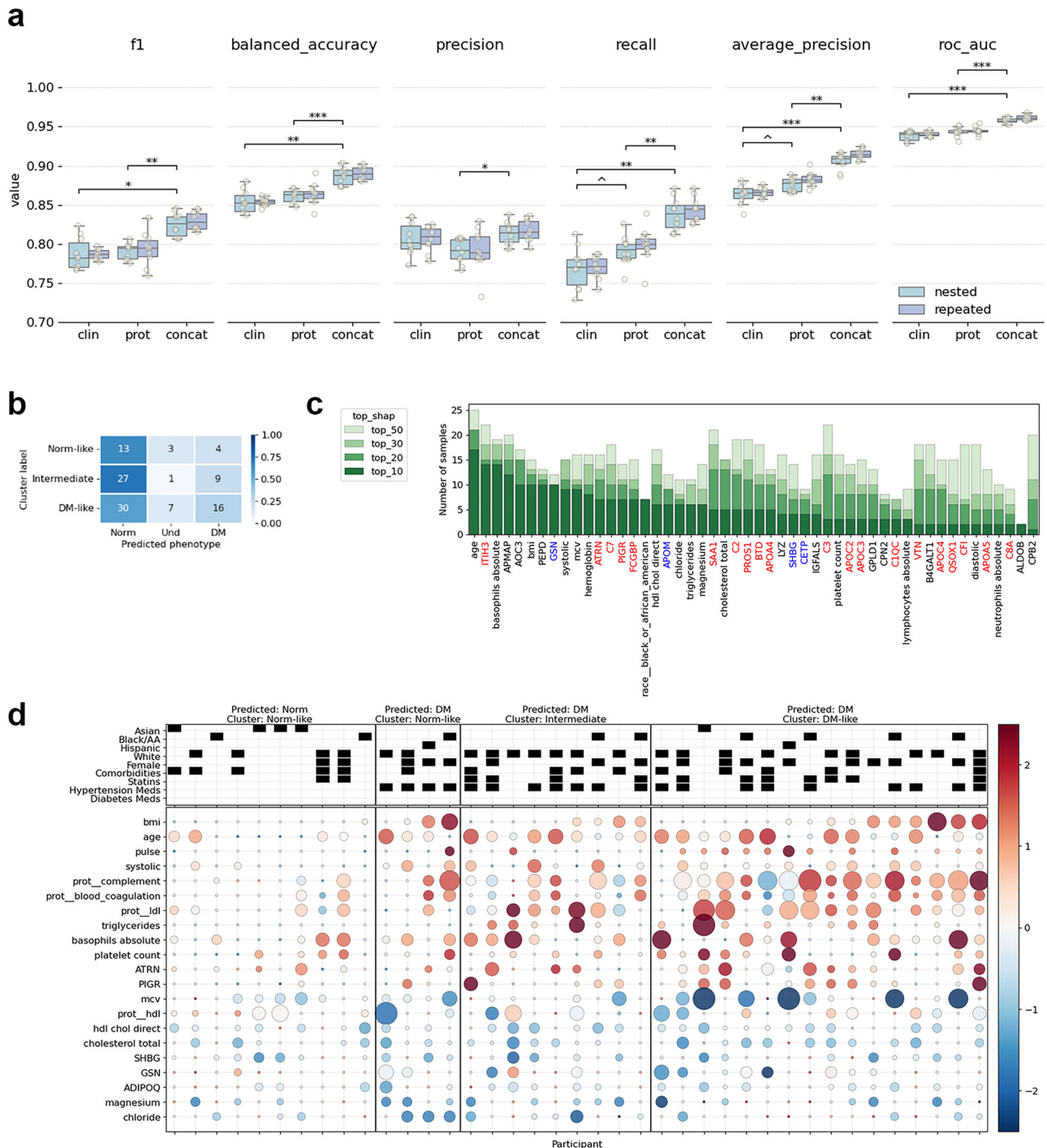


Fig. 5 | Machine learning classification of T2D. a Performance metrics of a ridge-logistic regression model discriminating diabetic status. Performance metrics are computed for each repeated CV (repeated) or for each repeated nested CV (nested) iteration. Differences between performance metrics of models trained with different datasets are tested with a one-sided Mann–Whitney test, comparing the concatenated datasets to clinical and proteomics only, correcting for multiple testing. Adjusted p value legend: \wedge : 0.01–0.05; $*$: 0.001–0.01; $**$: 0.0001–0.001; $***$: <0.0001 (full series of exact p values in Supplementary Data 3) [Sample size: 10 ML experiments]. **b** Confusion matrix of diabetes classification in prediabetics and cluster labels [Sample size: $n = 110$]. **c** Number of samples classified as diabetic where the feature has the highest SHAP value. Red: protein is DE and overexpressed in diabetics, Blue: protein is DE and underexpressed in diabetics. **d** Features contributing

to diabetes classification in participants with prediabetes, divided by cluster assignment. Ten participants with prediabetes classified as normoglycemic are also shown as controls. Circle size is proportional to the positive SHAP value for each participant and feature. Circle color is proportional to the feature value scaled across all participants with proteomic data [Sample size: $n = 39$]. Black/AA Black/African American, BMI body mass index, chol cholesterol, clin clinical, concat concatenated, CV cross-validation setting, DM diabetes mellitus, hdl high-density lipid, ldl low-density lipid, mcv mean corpuscular volume, ML machine learning, norm normoglycemic, preDM prediabetic, prot proteomic, prot-complement protein-complement, ROC AUC receiver operating characteristics area under the curve, SHAP SHapley Additive exPlanations, T2D type 2 diabetes, Und undecided.

To understand the relationship between diabetes classification and cluster assignment, and to inspect further which features are contributing to diabetes classification at the individual level, we applied the model selected with repeated cross-validation using the combined dataset to discriminate diabetes status for 110 participants with prediabetes throughout the study. Of these, 29 (26%) were classified “with T2D” by the model with probability higher than 0.6, while 70 were classified “with normoglycemia” with probability lower than 0.4 (Fig. 5b). Most of the participants classified “with T2D” were assigned to the intermediate or the diabetic-like clusters, with proportionally more participants assigned to the diabetic-like cluster, although not significantly (Fig. 5b).

To gain more insights into which features are contributing the most to discriminate diabetes status, we computed the SHAP values²² for all the features and counted how many times a feature had the highest ranking SHAP value across the 29 participants classified with T2D (Fig. 5c).

To investigate feature contribution at the individual level we examined the SHAP values²² for the 27 participants with prediabetes classified “with T2D” and 10 participants with prediabetes classified “with normoglycemia” with the lowest classification probability as control (Fig. 5d). Expectedly, known risk factors for diabetes, such as BMI, age and blood pressure, were positively contributing to T2D classification for some, but not all, individuals (Fig. 5d). Lipid measurements from a standard lipid panel were also associated with T2D classification. Specifically, triglycerides were positively contributing to T2D predictions, while HDL and total cholesterol were negatively contributing to diabetes predictions. Similarly, we noticed that lower magnesium and chloride were also contributing to T2D predictions, with magnesium and chloride deficits in T2D patients having been reported before^{32,33}.

Leveraging the additive nature of SHAP values²², we computed participant-level aggregated SHAP values for groups of functionally related proteins (see Supplementary Table 1 for the manually curated list of aggregated proteins). Consistently, with the differential protein expression results, complement, coagulation and LDL transport-related proteins showed positive contribution to diabetes classification in most participants with prediabetes, while HDL-related apolipoproteins showed negative contribution to diabetes classification in some participants with prediabetes (Fig. 5d). We also noted positive contributions to diabetes predictions for individual proteins, such as attractin protein (ATRN), which is involved in immune cell signaling³⁴, and PIGR, which is involved in inflammatory response and hepatic malignancy³⁵, and negative contributions for other individual proteins such as sex hormone-binding globulin (SHBG), whose levels have been shown to be inversely associated with diabetes risk³⁶ and adiponectin (ADIPOQ), also inversely associated with diabetes risk as well as obesity and insulin resistance³⁷.

Finally, whereas we could appreciate the same proteomic and clinical features associated with T2D across multiple participants, examining SHAP values at the participant level highlighted how the contribution of each feature to diabetes classification can vary between individuals. For example, qualitatively inspecting Fig. 5d, we could observe some people classified “with T2D” because of their high BMI and inflammatory markers, whereas for others it was their age and higher triglycerides, or their high inflammatory markers and lipid dysregulation.

Discussion

T2D, and metabolic disorders more broadly, are a complex set of diseases with varying pathological mechanisms and comorbidities across individuals, including obesity and insulin resistance. Current T2D diagnostic tools, such as HbA1c and FBG, rely on established thresholds to define normoglycemic, prediabetic, and diabetic phenotypes. Characterizing additional biomarkers from accessible tissues, such as blood, that can further stratify these phenotypes would be relevant across several applications, including identifying individuals who can most benefit earlier treatments on in their clinical course, such as treatment with GLP-1 agonist therapeutics (currently limited in their availability), and individuals who are at higher risk for developing diabetes sequelae, who may benefit from more intensive monitoring.

This analysis of a PBHS subcohort has identified differential plasma proteomic profiles for T2D and prediabetes states, which could enable a more refined stratification. Both based on their known functions and the direction of their perturbation in disease, the proteins in these profiles are consistent with known features of T2D pathophysiology. Moreover, the combination of these profiles with clinical features allowed the development of a logistic regression model that could discriminate diabetic disease status with high accuracy without relying on traditional molecular measures such as HbA1c. Our clustering/model also identified normoglycemic participants and participants with prediabetes that exhibit metabolic, physical, and cardiovascular features that resemble T2D, suggesting that our approach may be useful for further patient stratification and risk management.

This type of analysis was enabled by the availability of a unique research resource such as the PBHS cohort, consisting of deeply phenotyped individuals, both healthy and spanning multiple disease areas, including diabetes. The collection of multiple types of data, ranging from clinical to digital and molecular profiling, allows for an integrative characterization of diseases. This is particularly valuable for complex conditions, like T2D.

As part of PBHS, we generated one of the largest proteomic datasets, consisting of almost seven hundred individuals with a range of dysglycemia, including participants with diabetes, prediabetes, and normoglycemia. Comparing plasma proteins in people with diabetes and normoglycemic individuals revealed that inflammatory and blood coagulation markers are overexpressed in people with diabetes. This is consistent with the emerging role of systemic inflammation in the pathophysiology of T2D and associated metabolic disorders, which has generated increasing interest in inflammation as a target for intervention³⁸.

Proteins negatively associated with T2D, such as ADIPOQ, SHBG, and APOM, have also been reported in several other large-scale proteomic studies as negatively associated with obesity³⁹, newly diagnosed T2D^{39,40}, and incident T2D^{41–43}. A few complement proteins that were positively associated with T2D, such as CFH, C1S, and C5, were also shown to be positively associated with obesity³⁹ and T2D risk^{41–43}. This analysis, however, identified complement proteins and apolipoproteins positively associated with T2D beyond previous proteomic studies.

The complement system, originally viewed as a supportive first line of defense against microbial invaders, is increasingly being studied for its role in the initiation and progression of metabolic disorders, including obesity, insulin resistance, and T2D⁴⁴. Many individuals with T2D in the PBHS cohort were overweight or obese, which contributes to the overexpression of inflammatory markers in plasma, but we found that some complement proteins, including component 3, complement factor B and complement factor I, were also overexpressed in participants with T2D and normal weight. Our results identifying more complement proteins may be explained by the narrower focus of previous studies, with statistical models especially focused on identifying biomarkers of T2D incidence corrected for BMI^{41–43}.

The liver (mainly hepatocytes) is responsible for the biosynthesis of about 80–90% of plasma complement components⁴⁴. We found that, anatomically, most of the DE proteins in this study were liver-centric, a finding largely consistent with results of previous transcriptional analyses of microdissected liver tissue that reported overexpression of immune-related genes in the zone closer to the central vein²³ and pronounced zonation of active complement gene transcription, specifically in periportal and inter-portal hepatocytes¹³. Yet, some genes that were detected in liver biopsies from GTEx were not detected in the single-cell dataset, for example APOC2 and APOC4, both of which respond to metabolic cues in the liver by activation of transcription factors and nuclear hormone receptors⁴⁵. This may be due to these genes being expressed below the detection limit in single cells. Another explanation could be that these genes are detected in only some GTEx samples from donors with pre-existing conditions, such as diabetes.

Additionally, we found that proteins involved in blood coagulation and hemostasis were also overexpressed in the plasma of participants with T2D. Examples of these proteins included fibrinogen subunits alpha (FGA), beta (FGB), gamma (FGG), plasminogen (PLG), and plasmin inhibitor (SERPINF2)⁴⁶. Overexpression of hemostatic proteins in conjunction with

overexpression of inflammatory markers could represent a response to endothelial cell damage in blood vessels, as the metabolic burden of T2D, including insulin resistance, hyperglycemia and release of excess free fatty acids, along with other metabolic abnormalities affects vascular wall by a series of events including endothelial dysfunction, platelet hyperactivity, oxidative stress and low-grade inflammation⁴⁷. Indeed, it has been suggested that T2D and/or other cardiometabolic diseases can each cause reversible microvascular injury with accompanying dysfunction, which in time may or may not become irreversible and anatomically identifiable disease⁴⁸.

Altogether, the physiological observations related to the DE proteins suggest that the liver zone close to the central vein might be related to immune response and to overall inflammation, based on the signals from complement genes. Additional multi-omics studies can help elucidate the interrelation between T2D and liver dysfunction, particularly metabolic dysfunction-associated steatotic liver disease (previously known as NAFLD), including steatohepatitis (MASH; previously known as NASH)^{49,50}; and how they might be linked through inflammatory mechanisms such as complement activation⁵¹.

Clustering analysis of participants with normoglycemia, diabetes, and prediabetes based on clinical and proteomic features showed that 10% of normoglycemic participants had a clinico-molecular profile that resembled that of participants with T2D. At the proteomic level, these participants, mostly overweight and obese, consistently showed elevated levels of inflammatory and blood coagulation proteins. This suggests that measuring the presence of inflammatory and hemostatic pathways in plasma might help stratify within groups with seemingly similar levels of glycemic control. Participants such as these, normoglycemic by clinical standards but stratified closer to those with T2D, might be at high risk for diabetes, supporting the need for a holistic phenotypic assessment to properly diagnose diabetes or general metabolic dysregulation. Furthermore, normoglycemic participants in the diabetes-like cluster had, on average, poorer physical performance than the other normoglycemic participants and altered echocardiogram readouts indicative of left ventricular hypertrophy, which may also be linked to hypertension⁵². Somewhat conversely, the findings regarding physical activity levels recorded via wearable device indicated that participants with T2D in the normoglycemic-like subgroup, that is, with lower inflammatory markers, were more physically active.

In addition, several of our echocardiographic-related observations are consistent with prior reports establishing a relationship between echocardiographic abnormalities and T2D, particularly, abnormalities related to left ventricular size and mitral valve blood flow^{53,54}.

These results demonstrate the potential of utilizing proteomic markers alongside clinical features for improved T2D stratification. Building upon this work, future studies might incorporate additional metabolic measurements for a more comprehensive understanding of diabetic subtypes and other conditions leading to glycemic dysregulation. For example, oral glucose tolerance tests (OGTTs), C-peptide measurements, and insulin sensitivity tests might inform on the presence of other conditions, such as insulin resistance and beta cell dysfunction, and explore how protein features cluster with these distinct pathophysiologies. Indeed, lack of OGTT measurements and other markers of glycemic excursion, for example, from continuous glucose monitor data, is a limitation of the present study, as it would have allowed for further characterization of those participants with normoglycemia showing diabetic-like proteomic and clinical profiles.

Finally, we trained a machine learning model to discriminate diabetes status based on clinical and proteomic features and applied it to participants with prediabetes. The model trained on both clinical and proteomic features combined performed better than the models trained on clinical or proteomic features alone, achieving over 85% balanced accuracy. This performance is consistent with other clinical and/or molecular diabetes classifiers^{55,56}. To investigate the contribution of each feature to the model classification, we applied the model to participants with prediabetes and examined the SHAP values, which quantify how much a feature is contributing to diabetes classification for each individual. Consistent with the rest of the analysis, many participants with prediabetes who were classified

as “with T2D” by the model showed elevated levels of complement and hemostatic proteins. However, differences in feature contribution between individuals could also be appreciated, emphasizing the importance of assessing metabolic disorders in a holistic and personalized manner.

Despite revealing insights into the clinico-molecular characterization of T2D, we acknowledge some limitations of the study. Consistent with the complexity of diabetes and the known interrelationship between T2D and other metabolic disorders (like obesity, hypertension, or hypercholesterolemia), comorbidities were also observed in the PBHS cohort. PBHS was an observational study that was not designed to investigate factors associated with diabetes independently of other metabolic conditions. While the analysis was adjusted for demographic factors and medications used to treat hypercholesterolemia and hypertension, we did not adjust for other factors, such as BMI, because its known association with T2D would have reduced the power of the analysis. Similarly, at the time of this analysis, the follow-up time for participants in the PBHS was less than 4 years, relatively short to monitor the development of T2D; other studies designed to assess the risk of developing diabetes have analyzed more than 10 years of clinical records⁴². Moreover, participants engaged in the PBHS might have taken the initiative of medication and lifestyle changes that altered their metabolic status. Therefore, we focused our analysis on data from the initial visit of the study, and the machine learning model is a T2D classifier rather than a T2D risk predictor. Additionally, we point out that because of the heterogeneity in the T2D phenotype, the metabolic alterations observed in normoglycemic participants, and the potential confounding effect of disease treatment, together with the already mentioned incomplete fasting blood glucose measurement and the lack of OGTT results, the diagnosis to train the machine learning model and evaluate its performance is imperfect. This, together with the fact that the current diagnostic criteria for diabetes are sometimes not enough to correctly diagnose the disease⁵⁷, can lead to noisier model training and inexact estimates of performance. For instance, whereas lower levels of HDL are reportedly observed in T2D patients⁵⁸, higher total cholesterol is observed in T2D patients⁵⁹. Our ML model was potentially associating lower levels of total cholesterol with diabetes predictions because of statin use, prescribed to treat hypercholesterolemia. Similarly, the substantial overlap between T2D diagnosis and use of T2D medications prevented us from dissecting the impact of T2D medications on T2D classification. Finally, the findings herein were derived and evaluated within a single cohort, so further studies in independent cohorts will be an important step toward clinical application.

In conclusion, we have shown how a large-scale clinical cohort of deeply phenotyped participants across a health spectrum can be the source for integrative analyses that explore multiple layers of a complex disease. This holistic approach examines clinical and molecular features for each individual. In this case, we have presented a deep molecular characterization of the T2D continuum at the individual level, identifying differential proteomic profiles in those with normoglycemia, prediabetes, and T2D consistent with known pathophysiologic features of the disease. These results could serve as a foundation for further studies evaluating the utility of these molecular signatures in identifying patients at the highest risk for severe outcomes, as well as with the greatest likelihood to respond to specific interventions in a growing therapeutic arsenal.

Data availability

PBHS data cannot be made available directly on public repositories due to the terms of consent for the study (which required any specific data access to be approved by the study governance). Deidentified PBHS data corresponding to this study are available upon request for the purpose of examining its reproducibility. Interested investigators should direct requests to jsaiz@verily.com. Requests are subject to approval by PBHS governance. Source data for the graphs and charts in Figs. 1–5 in the main manuscript can be found in the file Supplementary Data 2.

Received: 19 October 2023; Accepted: 12 June 2025;

Published online: 03 July 2025

References

- Centers for Disease Control and Prevention National Diabetes Statistics Report. Prevalence of both diagnosed and undiagnosed diabetes. Available at <https://www.cdc.gov/diabetes/data/statistics-report/diagnosed-undiagnosed-diabetes.html>. Accessed 16 September 2023.
- Centers for Disease Control and Prevention Diabetes Basics. Diabetes tests. Available at <https://www.cdc.gov/diabetes/basics/getting-tested.html>. Accessed 16 September 2023.
- ElSayed, N. A. et al. 4. Comprehensive medical evaluation and assessment of comorbidities: standards of care in diabetes-2023. *Diab. Care* **46**, S49–S67 (2023).
- Arges, K. et al. The Project Baseline Health Study: a step towards a broader mission to map human health. *npj Digit. Med.* **3**, 1–10 (2020).
- Chatterjee, R. et al. Multi-dimensional characterization of prediabetes in the Project Baseline Health Study. *Cardiovasc. Diabetol.* **21**, 1–13 (2022).
- Geyer, P. E. et al. Plasma Proteome Profiling to detect and avoid sample-related biases in biomarker studies. *EMBO Mol. Med.* **11**, e10427 (2019).
- Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57**, 289–300 (1995).
- Szklarczyk, D. et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
- von Mering, C. et al. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* **33**, D433–7 (2005).
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **10**, P10008 (2008).
- Hagberg, A. A., Schult, D. A. & Swart, P. J. “Exploring Network Structure, Dynamics, and Function Using NetworkX.” In *Proc. 7th Python in Science Conference* (edited by Varoquaux, G., Vaught, T. & Millman, J.) 11–15 (SCIPY 08, Conference; Pasadena, CA, 2008).
- Yang, R. Y. et al. A systematic survey of human tissue-specific gene expression and splicing reveals new opportunities for therapeutic target identification and evaluation. Preprint at *bioRxiv* <https://www.biorxiv.org/content/10.1101/311563v1> (2018).
- Andrews, T. S. et al. Single-cell, single-nucleus, and spatial RNA sequencing of the human liver identifies cholangiocyte and mesenchymal heterogeneity. *Hepatol. Commun.* **6**, 821–840 (2022).
- Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
- Calinski, T. & Harabasz, J. A dendrite method for cluster analysis. *Commun. Stat. Theory Methods* **3**, 1–27 (1974).
- Vinh, N. X., Epps, J. & Bailey, J. Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.* **11**, 2837–2854 (2010).
- Davies, D. L. & Bouldin, D. W. A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1**, 224–227 (1979).
- Uchihara, B. et al. Accelerated epigenetic aging is associated with multiple cardiometabolic, hematologic, and renal abnormalities: a project baseline health substudy. *Circ. Genom. Precis. Med.* <https://doi.org/10.1161/CIRCGEN.122.003772> (2023).
- Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, 1–20 (2013).
- Popham, S. et al. An algorithm to classify real-world ambulatory status from a wearable device using multimodal and demographically diverse data: validation study. *JMIR Biomed. Eng.* **8**, e43726 (2023).
- Cauwenberghs, N. et al. Clinical and echocardiographic diversity associated with physical fitness in the Project Baseline Health Study: implications for heart failure staging. *J. Card. Fail.* <https://doi.org/10.1016/j.cardfail.2023.04.008> (2023).
- Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Proc. Advances in Neural Information Processing Systems* Vol. 30 (ed. Guyon, I. et al.) (Curran Associates, Inc., 2017).
- McEnerney, L. et al. Dual modulation of human hepatic zonation via canonical and non-canonical Wnt pathways. *Exp. Mol. Med.* **49**, e413 (2017).
- Feingold, K. R. Introduction to lipids and lipoproteins. in *Endotext* (ed. Feingold, K. R. et al.) (MDText.com, Inc., 2021).
- Schreiber, G. Synthesis and secretion of plasma proteins in the liver. *Pathology* **10**, 394 (1978).
- Uhlén, M. et al. The human secretome. *Sci. Signal.* **12**, 609 (2019).
- Cai, X. et al. Association between prediabetes and risk of all cause mortality and cardiovascular disease: updated meta-analysis. *BMJ* **370**, m2297 (2020).
- Palmieri, V. et al. Effect of type 2 diabetes mellitus on left ventricular geometry and systolic function in hypertensive subjects: Hypertension Genetic Epidemiology Network (HyperGEN) study. *Circulation* **103**, 102–107 (2001).
- Aronow, W. S. Hypertension and left ventricular hypertrophy. *Ann. Transl. Med.* **5**, 310 (2017).
- Mohan, M. et al. Left ventricular hypertrophy in diabetic cardiomyopathy: a target for intervention. *Front. Cardiovasc. Med.* **8**, 746382 (2021).
- Krstajic, D., Buturovic, L. J., Leahy, D. E. & Thomas, S. Cross-validation pitfalls when selecting and assessing regression and classification models. *J. Cheminform.* **6**, 10 (2014).
- Barbagallo, M., & Dominguez, L. J. Magnesium and type 2 diabetes. *World J. Diab.* **6**, 1152–1157 (2015).
- Khan, R. N., Saba, F., Kausar, S. F. & Siddiqui, M. H. Pattern of electrolyte imbalance in Type 2 diabetes patients: Experience from a tertiary care hospital. *Pak. J. Med. Sci. Q.* **35**, 797–801 (2019).
- Duke-Cohan, J. S. et al. Attractin (DPPT-L), a member of the CUB family of cell adhesion and guidance proteins, is secreted by activated human T lymphocytes and modulates immune cell interactions. *Proc. Natl. Acad. Sci. USA* **95**, 11336–11341 (1998).
- Sphyrin, N. & Mani, S. A. plgR: frenemy of inflammation, EMT, and HCC progression. *J. Natl. Cancer Inst.* **103**, 1644–1645 (2011).
- Aroda, V. R. et al. Circulating sex hormone binding globulin levels are modified with intensive lifestyle intervention, but their changes did not independently predict diabetes risk in the Diabetes Prevention Program. *BMJ Open Diab. Res. Care* **8**, e001841 (2020).
- Tsalamandris, S. et al. The role of inflammation in diabetes: current concepts and future perspectives. *Eur. Cardiol.* **14**, 50–59 (2019).
- Shim, K., Begum, R., Yang, C. & Wang, H. Complement activation in obesity, insulin resistance, and type 2 diabetes mellitus. *World J. Diab.* **11**, 1–12 (2020).
- Zaghloul, S. B. et al. Revealing the role of the human blood plasma proteome in obesity using genetic drivers. *Nat. Commun.* **12**, 1–13 (2021).
- Gummesson, A. et al. Longitudinal plasma protein profiling of newly diagnosed type 2 diabetes. *eBioMedicine* **63**, 103147 (2021).
- Ngo, D. et al. Proteomic profiling reveals biomarkers and pathways in type 2 diabetes risk. *JCI Insight* **6**, e144392 (2021).
- Rooney, M. R. et al. Proteomic predictors of incident diabetes: results from the Atherosclerosis Risk in Communities (ARIC) Study. *Diab. Care* **46**, 733–41 (2023).
- Cronjé, H. T. et al. Plasma proteomic risk markers of incident type 2 diabetes reflect physiologically distinct components of glucose-insulin homeostasis. *Diabetes* **72**, 666–73 (2023).
- Qin, X. & Gao, B. The complement system in liver diseases. *Cell. Mol. Immunol.* **3**, 333–340 (2006).
- Wolska, A. et al. Apolipoprotein C-II: new findings related to genetics, biochemistry, and role in triglyceride metabolism. *Atherosclerosis* **267**, 49–60 (2017).
- Kaur, R., Kaur, M. & Singh, J. Endothelial dysfunction and platelet hyperactivity in type 2 diabetes mellitus: molecular insights and therapeutic strategies. *Cardiovasc. Diabetol.* **17**, 121 (2018).

47. Horton, W. B. & Barrett, E. J. Microvascular dysfunction in diabetes mellitus and cardiometabolic disease. *Endocr. Rev.* **42**, 29–55 (2021).
48. Kaze, A. D. et al. Microvascular disease and incident heart failure among individuals with type 2 diabetes mellitus. *J. Am. Heart Assoc.* **10**, e018998 (2021).
49. Gastaldello, A. & Cusi, K. From NASH to diabetes and from diabetes to NASH: mechanisms and treatment options. *JHEP Rep.* **1**, 312–328 (2019).
50. Tanase, D. M. et al. The intricate relationship between type 2 diabetes mellitus (T2DM), insulin resistance (IR), and nonalcoholic fatty liver disease (NAFLD). *J. Diab. Res.* **2020**, 3920196 (2020).
51. Guo, Z. et al. The role of complement in nonalcoholic fatty liver disease. *Front. Immunol.* **13**, 1017467 (2022).
52. Ismail, T. F. et al. Hypertensive heart disease—the imaging perspective. *J. Clin. Med. Res.* **12**, <https://doi.org/10.3390/jcm12093122> (2023).
53. Kozakova, M., Morizzo, C., Fraser, A. G. & Palombo, C. Impact of glycemic control on aortic stiffness, left ventricular mass and diastolic longitudinal function in type 2 diabetes mellitus. *Cardiovasc. Diabetol.* **16**, 78 (2017).
54. Zhang, Y. et al. Atrioventricular coupling and left atrial abnormality in type 2 diabetes mellitus with functional mitral regurgitation patients verified by cardiac magnetic resonance imaging. *Cardiovasc. Diabetol.* **21**, 100 (2022).
55. Iparraguirre-Villanueva, O., Espinola-Linares, K., Flores Castañeda, R. O. & Cabanillas-Carbonell, M. Application of machine learning models for early detection and accurate classification of type 2 diabetes. *Diagnostics* **13**, <https://doi.org/10.3390/diagnostics13142383> (2023).
56. Agliata, A. et al. Machine learning as a support for the diagnosis of type 2 diabetes. *Int. J. Mol. Sci.* **24**, <https://doi.org/10.3390/ijms24076775> (2023).
57. Owora, A. H. Commentary: Diagnostic validity and clinical utility of HbA1c tests for type 2 diabetes mellitus. *Curr. Diab. Rev.* **14**, 196–199 (2018).
58. Farbstein, D. & Levy, A. P. HDL dysfunction in diabetes: causes and possible treatments. *Expert Rev. Cardiovasc. Ther.* **10**, 353–361 (2012).
59. Rhee, E.-J., Han, K., Ko, S.-H., Ko, K.-S. & Lee, W.-Y. Increased risk for diabetes development in subjects with large variation in total cholesterol levels in 2,827,950 Koreans: a nationwide population-based study. *PLoS ONE* **12**, e0176615 (2017).

Acknowledgements

The authors wish to thank PBHS participants, researchers, and study staff. The Project Baseline Health Study and this analysis were funded by Verily Life Sciences, South San Francisco, California. Verily Life Sciences is the funding source for the PBHS and was responsible for data collection. The authors were fully responsible for the data analysis and interpretation presented herein and the writing of this article. The following individuals: A.B. and K.D. had access to the raw data. The authors had access to the

full dataset for the study, reviewed and approved the final manuscript for submission.

Author contributions

Study concept and design: A.B., Y.W., K.D., C.C.K. Data collection: Verily Life Sciences. Data analysis and interpretation: A.B., Y.W., K.D., C.C.K., S.S., W.L., P.K., P.C., G.J.T., J.P.M., M.L., L.C., D.L., T.M.S., A.H. Draft writing and review: all. Draft approval for submission: all.

Competing interests

A.B., Y.W., S.S., W.L., P.K., P.C., G.T., J.P.M., M.L., L.C., D.L., T.S., A.H., K.D., C.C.K. declare the following competing interests: employment and equity ownership in Verily Life Sciences. D.E. declares the following competing interests: employment and equity ownership in Onduo, P.C. L.C. and L.C.K. declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at

<https://doi.org/10.1038/s43856-025-00964-x>.

Correspondence and requests for materials should be addressed to Alessandra Breschi.

Peer review information *Communications Medicine* thanks Julie Van, Ronald Ma, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>

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

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

© The Author(s) 2025