Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Prediagnostic Plasma Proteomic Signatures and Risk Stratification for Incident Chronic Kidney Disease in Type 2 Diabetes

Wenhong Liu
School of Basic Medical Sciences, Zhejiang Chinese Medical University
Yuxin Hong
School of Pharmaceutical Sciences, Zhejiang Chinese Medical University
Hao Yan
Research Center for Life Sciences Computing, Zhejiang Lab
Zongxin Meng
Jinhua Institute, Zhejiang Chinese Medical University
Huan Zhang
Research Center for Life Sciences Computing, Zhejiang Lab
Yue Yuan
Research Center for Life Sciences Computing, Zhejiang Lab
Zhiwei Xu
Jinhua Institute, Zhejiang Chinese Medical University
Zhaoxiang Wang
Research Center for Life Sciences Computing, Zhejiang Lab
Hui Wang
wh@zcmu.edu.cn

Hangzhou Medical College

Research Article

Keywords: Chronic kidney disease, Mendelian randomization, Plasma proteomics, Protein risk score, Risk stratification, Type 2 diabetes, UK Biobank

Posted Date: June 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10043065/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Chronic kidney disease (CKD) is a major complication of type 2 diabetes (T2D), but the circulating protein state that precedes clinically recognised CKD remains incompletely defined. We aimed to determine whether baseline plasma proteomics identifies a biologically interpretable prediagnostic protein signature and improves risk stratification for incident CKD in adults with T2D who were CKD-free at baseline.

Methods

We analyzed 1,715 adults with type 2 diabetes who were free of chronic kidney disease at baseline and had plasma proteomic profiling in the UK Biobank Pharma Proteomics Project. During a median 14.6-year follow-up, Cox models evaluated 2,920 baseline plasma proteins for incident chronic kidney disease. Baseline protein patterns were compared across future chronic kidney disease timing groups. Functional enrichment and protein–protein interaction network analyses characterised biological pathways. For prediction modelling, training-set Cox screening and least absolute shrinkage and selection operator regression selected candidate proteins for Cox models as selected proteins or a protein risk score. Exploratory Mendelian randomisation, colocalisation and translational annotation prioritised candidate proteins. The National Health and Nutrition Examination Survey and the broader UK Biobank cohort were used to contextualize the population burden and cardiorenal-metabolic relevance of chronic kidney disease in type 2 diabetes.

Results

During follow-up, 262 participants developed incident chronic kidney disease, and 449 proteins were associated with chronic kidney disease risk after false discovery rate correction. Chronic kidney disease-associated proteins showed graded baseline abundance patterns by future chronic kidney disease timing and were enriched in immune activation, receptor signalling, cytokine-related and extracellular matrix pathways, with network hubs including TNF, EGFR and IL10. Adding selected proteins or a protein risk score to clinical variables improved the C-index from 0.715 to 0.732 and 0.741, respectively. Exploratory prioritisation integrated association, prediction, genetic and translational evidence.

Conclusions

Baseline plasma proteomics captured a biologically organized prediagnostic chronic kidney disease risk state within the cardiorenal-metabolic context of type 2 diabetes. These findings support further evaluation of plasma proteomics for earlier CKD risk refinement in adults with T2D before clinically recognised CKD.

Research Insights

What is already known about this subject?

- CKD risk remains heterogeneous among adults with type 2 diabetes.
- Whether plasma proteomics captures a prediagnostic CKD state remains unclear.

What is the key research question?

- Can baseline plasma proteomics identify a biologically interpretable prediagnostic signature and improve risk stratification for incident CKD in adults with type 2 diabetes who are CKD-free at baseline?

What is new?

- Among 1,715 chronic kidney disease-free adults with type 2 diabetes, 449 plasma proteins were associated with incident chronic kidney disease.
- These proteins showed graded baseline abundance by future chronic kidney disease timing, including CKD occurring more than 10 years later.
- The signature implicated immune, receptor signalling and extracellular matrix pathways, improved risk stratification beyond clinical variables and supported exploratory candidate protein prioritisation.

How might this study influence clinical practice?

- After external validation, a prediagnostic plasma protein profile may help identify adults with type 2 diabetes who warrant closer kidney surveillance before chronic kidney disease is clinically recognised.

Introduction

Type 2 diabetes is a major driver of chronic kidney disease (CKD), but kidney disease progression among adults with T2D is clinically heterogeneous. Some individuals remain free of clinically recognised CKD for many years, whereas others develop CKD despite apparently similar baseline clinical profiles. Although estimated glomerular filtration rate(eGFR) and urinary albumin-to-creatinine ratio(UACR) are essential for CKD detection and staging, they may not fully capture earlier systemic biological activity that precedes clinically recognised CKD [1]. Although eGFR and UACR are central to CKD detection and staging, they may not fully capture biological activity that precedes clinically recognised CKD. This gap is clinically important because type 2 diabetes-related CKD represents both a substantial prevalent burden and a long-term risk state, motivating approaches that can identify biologically high-risk patients before overt CKD develops.

Plasma proteins may reflect biological processes involved in diabetic kidney disease (DKD), including immune activation, vascular injury, extracellular matrix remodelling and cardiorenal dysfunction [2].

Population-scale proteomic studies have shown that selected protein signatures can improve disease prediction beyond routine clinical information [3]. In people with diabetes, protein risk scores have improved incident CKD prediction when added to clinical models [4]. However, improved prediction alone does not establish whether CKD-free adults with type 2 diabetes already carry a biologically organised prediagnostic protein state. It also remains unclear whether such a state varies according to future CKD timing, or whether it maps to pathways and candidate proteins relevant to mechanistic and translational follow-up.

We therefore first quantified CKD burden among adults with type 2 diabetes in the National Health and Nutrition Examination Survey (NHANES) and estimated type 2 diabetes -associated incident CKD risk in the broader UK Biobank cohort to frame the clinical need. We then investigated whether baseline plasma proteomics could identify a biologically interpretable prediagnostic signature of incident CKD in adults with type 2 diabetes who were free of CKD at baseline. We further examined whether this signature was related to future CKD timing, pathway enrichment, protein-protein interaction network hubs, model-based risk stratification, and exploratory genetic and translational protein prioritisation.

RESEARCH DESIGN AND METHODS

Study Population

UK Biobank was used for the prospective and proteomic analyses. UK Biobank is a population-based prospective cohort that enrolled more than 500,000 participants between 2006 and 2010, with baseline assessment and follow-up through linked health records [5]. Because UK Biobank is a volunteer cohort, its participants may not be fully representative of all adults with type 2 diabetes in the general population. Type 2 diabetes was identified using the Eastwood et al. algorithm, which incorporates diabetes diagnosis, diabetes type, medication use, and related clinical information [6]. For the prospective analysis, eligible participants were free of CKD at baseline and had defined type 2 diabetes status. For the proteomic analysis, the cohort was further restricted to participants with type 2 diabetes, available baseline plasma proteomic data, no CKD at baseline, and ascertainable CKD follow-up. The final proteomic cohort included 1,715 participants and was randomly divided into a training set of 1,372 participants and an independent test set of 343 participants for prediction modelling (Supplementary Figure S1).

NHANES 2011–2018 was used to estimate CKD burden among adults with type 2 diabetes. Adult mobile examination center participants were included if they had sufficient information to define type 2 diabetes and CKD status. Type 2 diabetes was defined by self-reported physician-diagnosed diabetes, current insulin or glucose-lowering medication use, or glycated haemoglobin (HbA1c) ≥ 48 mmol/mol (6.5%) [7]. Pregnant participants and those likely to have type 1 diabetes were excluded. An 8-year examination weight was constructed for the combined survey cycles according to NHANES analytic guidance.

Ascertainment of CKD

In NHANES, eGFR was calculated from serum creatinine using the 2021 CKD-EPI creatinine equation [8]. Prevalent CKD was defined as eGFR < 60 mL/min/1.73 m² or UACR ≥ 30 mg/g [1].

In UK Biobank, baseline CKD was defined by eGFR < 60 mL/min/1.73 m², UACR ≥ 30 mg/g, prior CKD records, or insufficient information to exclude CKD at baseline. Participants with baseline CKD were excluded. Incident CKD after baseline was ascertained through linked hospital inpatient records and death registers using International Classification of Diseases, 10th Revision (ICD-10) codes N18, N18.1, N18.2, N18.3, N18.4, N18.5, and N18.9.

Measurements of Plasma Proteins

Plasma proteomic data were obtained from the UK Biobank Pharma Proteomics Project [2]. Protein abundance was measured using the Olink Explore 3072 platform and reported as normalized protein expression values. The platform quantified 2,941 protein analytes representing 2,923 unique proteins. After excluding proteins with > 30% missingness in the proteomic cohort, 2,920 proteins remained for proteome-wide Cox screening. Proteins were annotated using gene symbols.

Covariates

Baseline covariates were obtained from questionnaires, physical measurements, laboratory assays, and linked health records. Covariates included age at recruitment, sex, body mass index (BMI), systolic blood pressure, HbA1c, high-density lipoprotein (HDL) cholesterol, low-density lipoprotein (LDL) cholesterol, triglycerides, total cholesterol, Townsend deprivation index, smoking status, alcohol intake, eGFR, UACR, and age at first diagnosis of diabetes. Sex was recorded in UK Biobank and NHANES and included as a covariate where applicable. Race or ethnicity information was used for description or adjustment when available and was interpreted as a social rather than biological construct.

Statistical Analysis

Continuous variables were summarised as mean ± SD or median (interquartile range), and categorical variables as n (%). NHANES analyses incorporated examination weights, strata, and primary sampling units to account for the complex survey design. Weighted CKD prevalence and 95% CIs were estimated among adults with type 2 diabetes, with sensitivity analyses using alternative CKD definitions and missing-data handling approaches (Supplementary Tables S1–S3).

In the UK Biobank prospective analysis, Cox proportional hazards models estimated incident CKD risk associated with type 2 diabetes among participants free of CKD at baseline. Follow-up time was used as the time scale. The unadjusted model estimated this association without covariate adjustment. The age- and sex-adjusted model additionally included age and sex. The multivariable-adjusted model further

adjusted for body mass index, smoking status, alcohol intake, Townsend deprivation index, systolic blood pressure, total cholesterol, HDL cholesterol, and triglycerides. The kidney-adjusted model additionally included eGFR and UACR. Sensitivity analyses included stricter baseline CKD exclusion criteria, lag analyses excluding early CKD events, complete-case analyses, and same-sample model comparisons.

Among UK Biobank participants with type 2 diabetes and no CKD at baseline, each protein was tested separately using Cox proportional hazards regression. Model 0 included the protein only. Model 1 adjusted for age and sex. Model 2 further adjusted for eGFR, UACR, BMI, systolic blood pressure, alcohol intake, smoking status, Townsend deprivation index, triglycerides, LDL cholesterol, HDL cholesterol, HbA1c, and total cholesterol. Hazard ratios were reported per 1-SD higher protein abundance. Multiple testing was controlled using the Benjamini-Hochberg false discovery rate and Bonferroni correction. Proteins with FDR-adjusted P < 0.05 in Model 2 were carried forward for biological analyses.

Baseline protein abundances were summarised across three future CKD groups: no incident CKD, incident CKD after > 10 years, and incident CKD within ≤ 10 years. Functional enrichment analyses tested Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways using FDR-significant proteins against the background of all 2,920 screened proteins. Protein–protein interaction analysis was performed using the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) in Homo sapiens, restricted to high-confidence interactions with a combined score ≥ 0.700 [9].

To prevent data leakage, all feature selection for prediction modelling was performed exclusively within the training set, separate from the full-cohort proteome-wide discovery analysis. The clinical model was based on a SCORE2-Diabetes–derived variable set [10]. The original SCORE2-Diabetes equation was not applied, and model coefficients were re-estimated for incident CKD. Proteins associated with incident CKD in training-set Cox screening at FDR-adjusted P < 0.05 were retained as candidate protein predictors. The least absolute shrinkage and selection operator (LASSO) regression with fivefold cross-validation was applied to the standardised candidate proteins in the training set using incident CKD status as the binary outcome, to further select protein predictors and derive protein weights. The selected protein set and LASSO-derived weights were applied unchanged to the independent test set.

The protein risk score was calculated as: $PRS_k = \sum_{j=1}^{p} \beta_j Z_{kj}$ where PRS_k is the PRS for participant k, p denotes the number of selected proteins, β_j is LASSO-derived coefficient for protein j, and Z_{kj} is the standardised abundance of protein j in participant k. Positive coefficients indicated higher predicted CKD risk, whereas negative coefficients indicated lower predicted risk.

Four Cox-based prediction models were fitted in the training set and evaluated in the same independent test set: the clinical model, the clinical plus selected-proteins model, the clinical plus PRS model, and the protein-only model. Incremental predictive value was assessed using ΔC-index, net reclassification improvement (NRI), and integrated discrimination improvement (IDI), with the clinical model as the

reference. Model discrimination was assessed using Harrell's C-index, time-dependent AUC curves, and fixed-time-window ROC analyses across different prediction horizons.

For exploratory risk stratification, 10-year predicted risk was calculated for each model in the independent test set. Data-derived cutoffs were identified using X-tile to categorise participants into low-risk and high-risk groups. Kaplan-Meier curves were used to visualise CKD-free survival, and log-rank tests and univariable Cox models were used to compare groups. Because the cutoffs were derived within the test set, these analyses were considered exploratory and were not used as primary evidence of model performance. Prediction-model sensitivity analyses included repeated random training-test splits, complete-case analysis, 2-year lag analysis, Bonferroni-threshold feature selection, bootstrap internal validation, and calibration assessment.

For genetic and translational follow-up, the 59 LASSO-selected proteins were intersected with the 449 FDR-significant proteins from the full-cohort Cox analysis, yielding 53 overlapping proteins for Mendelian randomisation, colocalisation, and translational annotation. Protein quantitative trait loci (pQTLs) were obtained from the deCODE plasma proteome study, which measured 4,907 aptamers in 35,559 Icelanders [11]. Olink proteins were mapped to deCODE protein targets by gene symbol, with UniProt identifiers used for confirmation when available. Candidate pQTL instruments were selected using $P < 1 \times 10^{-5}$ and F-statistic > 10; and independent instruments were retained by linkage disequilibrium clumping.

Outcome summary statistics for CKD in type 2 diabetes were obtained from a published genome-wide association study [12]. European-ancestry outcome summary statistics were prioritised when available. Exposure and outcome data were harmonized, and inverse-variance weighted Mendelian randomisation was used as the primary method when multiple instruments were available. Weighted median, MR-Egger, heterogeneity tests, MR-Egger intercept testing, single-SNP analysis, and leave-one-out analyses were performed when applicable. Cis-pQTL and colocalisation analyses used variants within ± 1 Mb of the corresponding protein-coding gene, and PP.H4 was used to summarise evidence for a shared causal variant. GREP ICD and ATC annotation outputs were used to summarise disease, drug, and cardiorenal or metabolic annotation support. These genetic and translational analyses were exploratory.

All statistical tests were two-sided, and $P < 0.05$ was considered statistically significant unless otherwise specified. Statistical analyses were performed using R (version 4.5.2) and Python (version 3.12.13).

RESULTS

Population-level CKD burden and Type 2 diabetes-associated incident CKD risk

We quantified CKD burden among adults with type 2 diabetes in NHANES 2011–2018 and assessed the prospective association between type 2 diabetes and incident CKD in the broader UK Biobank cohort

(Fig. 1A–C). In NHANES, 3,621 adults with type 2 diabetes and complete survey-design information were included, of whom 1,380 had CKD. The weighted CKD prevalence was 34.70% (95% CI, 32.46%–37.00%) (Supplementary Tables S1–S3).

In UK Biobank, the prospective analysis included 445,242 participants free of CKD at baseline, among whom 23,976 developed incident CKD during follow-up. Type 2 diabetes was associated with higher incident CKD risk in the unadjusted model (HR, 4.71; 95% CI, 4.56–4.86), after adjustment for age and sex (HR, 3.60; 95% CI, 3.48–3.72), the multivariable-adjusted model (HR, 2.07; 95% CI, 2.00–2.15), and the kidney-adjusted model (HR, 2.23; 95% CI, 2.15–2.31) (Supplementary Tables S4–S5).

Baseline characteristics of the CKD-free Type 2 diabetes proteomic cohort

A total of 1,715 participants with type 2 diabetes, baseline plasma proteomic data, and no CKD at baseline were included in the incident CKD analysis (Table 1). During a median follow-up of 14.6 years (interquartile range, 11.4–15.5 years), 262 participants (15.3%) developed incident CKD and 1,453 did not develop CKD. Compared with participants who remained CKD-free, those who developed incident CKD were older and had higher BMI, HbA1c, and UACR levels, but lower baseline eGFR, LDL-C, and total cholesterol levels (Table 1; Supplementary Table S6).

Table 1
Baseline characteristics of the participants (n = 1715)

Characteristic	Total	Without incident CKD	With incident CKD	*P*-value[a]
Participants (%)	1715 (100.0)	1453 (84.7)	262 (15.3)	
Sex, n (%)				1.000
Male	1065 (62.1)	902 (62.1)	163 (62.2)	
Female	650 (37.9)	551 (37.9)	99 (37.8)	
Age, years	60.1 ± 7.0	59.7 ± 7.1	62.2 ± 5.9	< 0.001
BMI(kg/m^2)	31.3 ± 5.9	31.1 ± 5.8	32.5 ± 6.0	< 0.001
SBP(mmHg)	142.7 ± 17.9	142.6 ± 17.6	142.9 ± 19.1	0.826
Townsend deprivation index	-0.2 ± 3.5	-0.3 ± 3.5	0.1 ± 3.6	0.057
Smoking state, n (%)				0.241
Current	196 (11.4)	160 (11.0)	36 (13.7)	
No current	1519 (88.6)	1293 (89.0)	226 (86.3)	
Alcohol consumption, n (%)				0.625
Current	1422 (82.9)	1208 (83.1)	214 (81.7)	
No current	293 (17.1)	245 (16.9)	48 (18.3)	
HbA1c (mmol/mol)	51.5 ± 13.4	51.2 ± 13.4	53.2 ± 13.0	0.025
HDL-C(mmol/L)	1.2 ± 0.3	1.2 ± 0.3	1.2 ± 0.3	0.045
LDL-C(mmol/L)	2.7 ± 0.8	2.7 ± 0.8	2.5 ± 0.7	< 0.001
TG (mmol/L)	2.1 ± 1.1	2.0 ± 1.1	2.2 ± 1.3	0.051
TC (mmol/L)	4.4 ± 1.0	4.5 ± 1.0	4.3 ± 0.9	< 0.001
eGFR (mL/min/1.73 m^2)	94.3 ± 12.7	96.0 ± 11.6	84.7 ± 14.5	< 0.001
UACR (mg/g)	12.3 ± 6.9	11.9 ± 6.7	13.8 ± 7.3	0.005

Values are mean ± SD for continuous variables and n (%) for categorical variables.

[a] *P* values were calculated using Welch's t test for continuous variables and chi-square test for categorical variables.

Prediagnostic plasma protein signature associated with incident CKD in CKD-free adults with type 2 diabetes

Among 2,920 plasma proteins, 449 were associated with incident CKD after FDR correction, including 413 positively and 36 inversely associated proteins (Fig. 1D; Supplementary Tables S7–S8). Of these, 162 also met the Bonferroni threshold.

The strongest positive associations were observed for WFDC2 (HR per 1-SD increment, 1.90; 95% CI, 1.65–2.17; FDR P = 4.82×10^{-19}), IGFBP4 (HR, 1.73; 95% CI, 1.51–1.97; FDR P = 1.41×10^{-14}), COL6A3 (HR, 1.75; 95% CI, 1.52–2.01; FDR P = 9.58×10^{-14}), PGF (HR, 1.46; 95% CI, 1.32–1.60; FDR P = 5.62×10^{-13}), and PIK3IP1 (HR, 1.68; 95% CI, 1.47–1.93; FDR P = 1.87×10^{-12}). The strongest inverse associations included TCL1A (HR, 0.75; 95% CI, 0.66–0.85; FDR P = 9.77×10^{-5}), AFM (HR, 0.77; 95% CI, 0.68–0.88; FDR P = 1.18×10^{-3}), XPNPEP2 (HR, 0.78; 95% CI, 0.68–0.88; FDR P = 1.44×10^{-3}), NELL1 (HR, 0.79; 95% CI, 0.70–0.89; FDR P = 2.35×10^{-3}), and NPHS1 (HR, 0.77; 95% CI, 0.68–0.88; FDR P = 2.35×10^{-3}).

Prediagnostic proteins showed graded patterns according to future CKD timing

Participants were grouped according to future CKD timing: no incident CKD during follow-up (n = 1,453), incident CKD after > 10 years (n = 109), or incident CKD within ≤ 10 years (n = 153) (Fig. 1E). The top positively and inversely associated proteins exhibited corresponding graded abundance trends across these groups (Supplementary Tables S9–S10).

Functional enrichment and protein interaction patterns of CKD-associated proteins

Functional enrichment analysis of the 449 FDR-significant proteins linked the prediagnostic signature to receptor-related molecular functions, extracellular matrix structural constituent terms, and cytokine–cytokine receptor interaction pathways (Fig. 1F; Supplementary Table S11). Additional enriched terms included extracellular matrix, cell surface, positive regulation of cell activation, positive regulation of leukocyte activation, cell–cell adhesion, and ossification. High-confidence STRING protein–protein interaction analysis identified interconnected modules annotated as TNF inflammatory signalling, adaptive immune activation, innate immune signalling, IL10 cytokine regulation, extracellular matrix remodelling, and EGFR or tissue remodelling (Fig. 1G; Supplementary Tables S12–S13). Hub proteins included TNF, EGFR and IL10.

Protein-informed prediction models for incident CKD

Feature selection in the training set yielded 59 selected proteins (Supplementary Table S14). In the independent test set, protein-informed models showed higher discrimination than the clinical comparator (Fig. 2A). The C-index was 0.715 for the clinical model, 0.732 for the clinical plus selected-

proteins model, 0.741 for the clinical plus PRS model, and 0.722 for the protein-only model. Compared with the clinical comparator, the protein-informed models showed positive ΔC-index values and numerically higher NRI and IDI estimates (Supplementary Table S15).

Time-dependent AUC analyses showed higher discrimination for protein-informed models across follow-up, with similar patterns in fixed 10-year and > 10-year prediction analyses (Fig. 2B–D; Supplementary Table S15). For exploratory visualisation of risk separation, participants in the test set were categorised into low-risk and high-risk groups using data-derived X-tile cutoffs based on predicted 10-year risk. Kaplan-Meier curves showed separation of CKD-free survival between groups across models, with higher predicted-risk groups showing greater incident CKD risk (Fig. 2E; Supplementary Table S16). Because the cutoffs were derived within the test set, these risk-stratification analyses were interpreted as exploratory.

Across repeated train-test splits and bootstrap internal validation, protein-informed models generally showed higher discrimination than the clinical model. Repeated split, bootstrap, and calibration analyses generally supported the main model comparison, whereas complete-case, 2-year lag, and Bonferroni-threshold analyses were interpreted cautiously because of smaller samples, fewer events, or model instability (Supplementary Table S15).

Exploratory genetic and translational prioritisation

To prioritise proteins supported by both full-cohort association evidence and training-set model selection, the 449 FDR-significant proteins from the full-cohort Cox analysis were intersected with the 59 proteins selected during training-set model development. This yielded 53 overlapping proteins for exploratory Mendelian randomisation, colocalisation, and translational annotation analyses (Fig. 3A; Supplementary Table S17).

MR analyses identified FDR-supported associations for CLEC7A (OR, 1.28; 95% CI, 1.10–1.49; FDR P = 0.042) and NPPB (OR, 0.76; 95% CI, 0.65–0.90; FDR P = 0.042). Follow-up cis-pQTL colocalisation did not provide strong evidence for shared causal variants. Disease and drug-target annotation showed limited coverage overall but highlighted REN, NPPB, and TNFSF11 as proteins with cardiorenal or therapeutic relevance (Supplementary Table S17). These analyses were exploratory.

Discussion

The principal finding of this study is that baseline plasma proteomic profiling identified a broad and biologically interpretable prediagnostic protein signature associated with future incident CKD in adults with type 2 diabetes who were free of CKD at baseline. Population-level analyses contextualised this clinical need: approximately one-third of adults with type 2 diabetes in NHANES had CKD, and type 2 diabetes remained associated with a more than twofold higher risk of incident CKD in the broader UK Biobank cohort after multivariable adjustment. The strongest positive and inverse signals also showed graded baseline abundance patterns according to future CKD timing. These timing-related patterns suggest that baseline proteomic perturbation was not limited to participants who developed CKD shortly

after baseline, but was also detectable, although less prominently, among those who developed CKD more than 10 years later. These findings support the central premise of this study that adults with type 2 diabetes may already exhibit a systemic plasma protein risk state before CKD is clinically recognised.

Existing proteomic studies have shown that circulating proteins can improve CKD prediction in people with diabetes and have highlighted inflammatory proteins in advanced DKD [4, 13]. The contribution of the present study is not simply to show another protein-based prediction gain. Rather, by focusing on CKD-free adults with type 2 diabetes, mapping baseline protein gradients by future CKD timing, integrating pathway enrichment and protein-protein interaction network hubs, and incorporating exploratory genetic and translational prioritisation, this analysis connects risk stratification with biological interpretation and protein prioritisation before CKD becomes clinically recognised.

Several proteins most strongly associated with higher CKD risk linked this prediagnostic state to plausible mechanisms involving tubular stress, extracellular matrix remodelling, and immune activation. WFDC2, which encodes human epididymis protein 4 (HE4), showed the strongest positive association in our analysis. This finding is biologically plausible because HE4 is upregulated in fibrotic kidneys and has been shown to inhibit protease-mediated collagen degradation, thereby linking WFDC2 to extracellular matrix accumulation and renal fibrosis [14]. In type 2 diabetes, circulating profibrotic proteins, including WFDC2, have also been associated with early progressive renal decline among individuals with and without albuminuria, supporting its relevance to diabetes-related kidney risk [15]. COL6A3 further connects the CKD-associated proteomic signature to extracellular matrix turnover through the collagen VI and endotrophin pathway. Endotrophin is a collagen VI–derived matrikine implicated in fibroinflammatory remodelling, and urinary endotrophin has been reported to predict CKD progression [16]. IGFBP4 supports a complementary role in immune and complement activation, as well as podocyte injury, because recent DKD evidence suggests that IGFBP2 and IGFBP4 can activate complement signalling in macrophages and promote podocyte apoptosis [17]. Together, the elevation of WFDC2, COL6A3, and IGFBP4 suggests that the prediagnostic CKD risk state in type 2 diabetes may involve converging processes of tubular injury, matrix remodelling, and immune-complement pathways.

The proteins associated with lower CKD risk may offer a complementary view of kidney integrity and systemic protection. NPHS1 encodes nephrin, a key component of the podocyte slit diaphragm and glomerular filtration barrier; urinary nephrin has been studied as a biomarker of podocyte injury and early diabetic nephropathy [18]. However, because our study measured plasma rather than urinary or kidney tissue NPHS1, the inverse association should be interpreted cautiously. Lower circulating NPHS1 should not be equated with urinary nephrin shedding or kidney tissue nephrin depletion. One possibility is that circulating NPHS1 reflects podocyte-associated protein turnover or reserve, although this interpretation remains speculative and requires validation using paired plasma, urine, and kidney tissue measurements. AFM also deserves attention among the inverse signals. Higher serum afamin concentrations have recently been associated with higher eGFR, less albuminuria, and lower future kidney failure risk in a CKD cohort [19]. In our study, the inverse association between AFM and incident CKD is consistent with the possibility that lower AFM abundance may mark loss of a favourable

systemic or renal risk profile. These inverse signals suggest that CKD risk stratification may depend not only on the accumulation of injury-associated proteins but also on reduced abundance of proteins related to kidney integrity or systemic protection.

Plasma proteomic information also provided incremental value for model-based CKD risk stratification beyond clinical variables. The C-index was 0.715 for the clinical model, 0.732 for the clinical plus selected-proteins model, and 0.741 for the clinical plus protein-risk-score model; the protein-only model showed comparable discrimination, with a C-index of 0.722. These findings suggest that plasma proteins captured risk information not fully represented by routine clinical variables. Therefore, the clinically relevant interpretation is the incremental value of adding protein information to clinical variables, especially through a streamlined protein risk score model, rather than replacing established clinical assessment. Further calibration assessment and external validation are needed before clinical implementation.

The potential clinical role of plasma proteomics is risk refinement rather than replacement of eGFR or UACR, which remain central to CKD detection, staging, and monitoring [1]. The CKD burden observed in NHANES and the elevated incident CKD risk associated with type 2 diabetes in UK Biobank highlight the need for earlier and more individualized risk assessment. Protein profiling may help identify adults with type 2 diabetes whose biological risk appears higher than suggested by routine clinical variables alone. Future studies should evaluate whether smaller targeted protein panels can support clinically feasible CKD risk refinement in type 2 diabetes. Such risk refinement may be particularly relevant in diabetes care, where CKD contributes substantially to long-term kidney and cardiovascular vulnerability.

The exploratory genetic and translational analyses provided an additional prioritisation layer but did not establish causality or therapeutic relevance. MR highlighted CLEC7A and NPPB, but follow-up colocalisation did not provide strong evidence for shared causal variants. The lack of strong colocalisation suggests that these MR findings may be influenced by linkage disequilibrium or distinct causal variants within the same locus. Therefore, CLEC7A and NPPB should be viewed as biologically plausible, hypothesis-generating prioritisation signals rather than definitive causal targets. CLEC7A is biologically plausible because Dectin-1 has been implicated in macrophage-mediated kidney inflammation and fibrosis, including pathways involving neutrophil recruitment, TGF-β1 secretion, and renal dysfunction [20]. NPPB is better interpreted as part of a broader cardiorenal stress axis rather than a kidney-specific signal; NT-proBNP has been shown to improve prediction of cardiorenal complications in type 2 diabetes, consistent with the closely intertwined nature of diabetic kidney and heart disease [21].

Several limitations should be considered. The observational design cannot establish causality, and proteins were measured at a single baseline time point; therefore, timing patterns should not be interpreted as within-person proteomic changes. Incident CKD was identified using linked health records and may miss milder or undiagnosed disease. The UK Biobank is a volunteer cohort with limited ancestral and socioeconomic diversity, and may not represent all adults with type 2 diabetes. In addition,

plasma protein measurements do not allow the tissue or cellular sources of the observed signals to be determined. Although all four prediction models were evaluated in the same independent test set, the models require external validation, calibration, assay standardisation, and prospective evaluation before clinical implementation. Risk stratification analyses based on data-derived cutoffs should also be interpreted as exploratory. Similarly, the MR, colocalisation, and translational annotation analyses were intended for exploratory prioritisation and should not be interpreted as definitive evidence of causality or therapeutic efficacy.

Conclusion

In summary, this study identified a biologically interpretable prediagnostic plasma protein risk state for incident CKD among adults with type 2 diabetes who were free of CKD at baseline. This signature was linked to future CKD timing, pathway and network biology, and protein-informed risk stratification beyond clinical variables. These findings support further validation of targeted plasma protein panels for earlier CKD risk refinement in type 2 diabetes before clinical translation.

List of abbreviations

CKD, chronic kidney disease
eGFR, estimated glomerular filtration rate
FDR, false discovery rate
HbA1c, glycated haemoglobin
LASSO, least absolute shrinkage and selection operator
MR, Mendelian randomisation

PPI, protein–protein interaction
PRS, protein risk score
UACR, urinary albumin-to-creatinine ratio

Declarations

Acknowledgments. The authors thank the participants, investigators and staff of the UK Biobank and the National Health and Nutrition Examination Survey for their valuable contributions to this study. The authors also acknowledge the investigators who generated and shared the publicly available proteomic and genome-wide association summary statistics used in the exploratory Mendelian randomisation analyses.

Ethics Approval. The UK Biobank study was approved by the North West Multi-Centre Research Ethics Committee in 2011 (11/NW/0382) and renewed in 2016 (16/NW/0274) and 2021 (21/NW/0157). All participants provided written informed consent before enrolment. The present research was conducted using UK Biobank data under Application Number 584010, in accordance with the World Medical Association Declaration of Helsinki and the UK Biobank Ethics and Governance Framework. NHANES

protocols were approved by the National Center for Health Statistics Research Ethics Review Board, and all NHANES participants provided written informed consent.

Data and Resource Availability. UK Biobank and UK Biobank Pharma Proteomics Project data are available to approved researchers through UK Biobank. NHANES data are publicly available from the National Center for Health Statistics. Summary results supporting the findings of this study are provided in the supplementary tables. Analytic code is available from the corresponding author upon reasonable request, subject to institutional and data-use restrictions.

Funding. This work was supported by the Zhejiang Provincial “Pioneer” and “Leading Goose” Research and Development Program of China (2026C01049 and 2023C03041), the National Natural Science Foundation of China (82274239), and the Zhejiang Provincial High-Level Talents Special Support Program for Scientific and Technological Innovation Leaders (2023R5255).

Competing interests. The authors declare that they have no competing interests.

Author contributions. WL, YH, HY, ZW and HW conceived the study idea and designed the study. WL, YH, and HY analyzed and interpreted the data and drafted the manuscript. ZM, HZ, YY, and ZX contributed to data acquisition, data analysis, and interpretation of the results. All authors critically revised the manuscript and approved the final version. HW and ZW are the guarantors of this work and, as such, had full access to all the data used in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

References

1. Kidney Disease: Improving Global Outcomes CKDWG: KDIGO 2024 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease. *Kidney Int* 2024, 105(4S):S117-S314.
2. Sun BB, Chiou J, Traylor M, Benner C, Hsu YH, Richardson TG, Surendran P, Mahajan A, Robins C, Vasquez-Grinnell SG *et al*: Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* 2023, 622(7982):329-338.
3. Carrasco-Zanini J, Pietzner M, Davitte J, Surendran P, Croteau-Chonka DC, Robins C, Torralbo A, Tomlinson C, Grunschlager F, Fitzpatrick N *et al*: Proteomic signatures improve risk prediction for common and rare diseases. *Nat Med* 2024, 30(9):2489-2498.
4. Ye Z, Zhang Y, Zhang Y, Yang S, He P, Liu M, Zhou C, Gan X, Huang Y, Xiang H *et al*: Large-Scale Proteomics Improve Prediction of Chronic Kidney Disease in People With Diabetes. *Diabetes Care* 2024, 47(10):1757-1763.
5. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M *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(3):e1001779.
6. Eastwood SV, Mathur R, Atkinson M, Brophy S, Sudlow C, Flaig R, de Lusignan S, Allen N, Chaturvedi N: Algorithms for the Capture and Adjudication of Prevalent and Incident Diabetes in UK Biobank.

PLoS One 2016, 11(9):e0162388.

7. American Diabetes Association Professional Practice Committee for D: 2. Diagnosis and Classification of Diabetes: Standards of Care in Diabetes-2026. *Diabetes Care* 2026, 49(Supplement_1):S27-S49.
8. Inker LA, Eneanya ND, Coresh J, Tighiouart H, Wang D, Sang Y, Crews DC, Doria A, Estrella MM, Froissart M *et al*: New Creatinine- and Cystatin C-Based Equations to Estimate GFR without Race. *N Engl J Med* 2021, 385(19):1737-1749.
9. Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, Gable AL, Fang T, Doncheva NT, Pyysalo S *et al*: The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res* 2023, 51(D1):D638-D646.
10. Group SC-DW, the ESCCRC: SCORE2-Diabetes: 10-year cardiovascular risk estimation in type 2 diabetes in Europe. *Eur Heart J* 2023, 44(28):2544-2556.
11. Ferkingstad E, Sulem P, Atlason BA, Sveinbjornsson G, Magnusson MI, Styrmisdottir EL, Gunnarsdottir K, Helgason A, Oddsson A, Halldorsson BV *et al*: Large-scale integration of the plasma proteome with genetics and disease. *Nat Genet* 2021, 53(12):1712-1721.
12. van Zuydam NR, Ahlqvist E, Sandholm N, Deshmukh H, Rayner NW, Abdalla M, Ladenvall C, Ziemek D, Fauman E, Robertson NR *et al*: A Genome-Wide Association Study of Diabetic Kidney Disease in Subjects With Type 2 Diabetes. *Diabetes* 2018, 67(7):1414-1427.
13. Niewczas MA, Pavkov ME, Skupien J, Smiles A, Md Dom ZI, Wilson JM, Park J, Nair V, Schlafly A, Saulnier PJ *et al*: A signature of circulating inflammatory proteins and development of end-stage renal disease in diabetes. *Nat Med* 2019, 25(5):805-813.
14. LeBleu VS, Teng Y, O'Connell JT, Charytan D, Muller GA, Muller CA, Sugimoto H, Kalluri R: Identification of human epididymis protein-4 as a fibroblast-derived mediator of fibrosis. *Nat Med* 2013, 19(2):227-231.
15. Ihara K, Skupien J, Kobayashi H, Md Dom ZI, Wilson JM, O'Neil K, Badger HS, Bowsman LM, Satake E, Breyer MD *et al*: Profibrotic Circulating Proteins and Risk of Early Progressive Renal Decline in Patients With Type 2 Diabetes With and Without Albuminuria. *Diabetes Care* 2020, 43(11):2760-2767.
16. Rasmussen DGK, Fenton A, Jesky M, Ferro C, Boor P, Tepel M, Karsdal MA, Genovese F, Cockwell P: Urinary endotrophin predicts disease progression in patients with chronic kidney disease. *Sci Rep* 2017, 7(1):17328.
17. Liang J, Huang Y, Peng D, Xie Y, Liu Y, Lu X, Xu J: IGFBP2 and IGFBP4 interact to activate complement pathway in diabetic kidney disease. *Ren Fail* 2025, 47(1):2440528.
18. Kandasamy Y, Smith R, Lumbers ER, Rudd D: Nephrin - a biomarker of early glomerular injury. *Biomark Res* 2014, 2:21.
19. Kollerits B, Kotsis F, Schneider MP, Schultheiss UT, Weissensteiner H, Schonherr S, Forer L, Meiselbach H, Wanner C, Eckardt KU *et al*: Association of Serum Afamin Concentrations With Kidney

Failure in Patients With CKD: Findings From the German CKD Cohort Study. *Am J Kidney Dis* 2025, 85(4):432-441 e431.

20. Ye S, Huang H, Xiao Y, Han X, Shi F, Luo W, Chen J, Ye Y, Zhao X, Huang W *et al*: Macrophage Dectin-1 mediates Ang II renal injury through neutrophil migration and TGF-beta1 secretion. *Cell Mol Life Sci* 2023, 80(7):184.
21. Ma RCW, Tam CHT, Hou Y, Lau ESH, Ozaki R, Lui JNM, Chow E, Kong APS, Huang C, Ng ACW *et al*: NT-proBNP improves prediction of cardiorenal complications in type 2 diabetes: the Hong Kong Diabetes Biobank. *Diabetologia* 2025, 68(2):342-356.

Figures

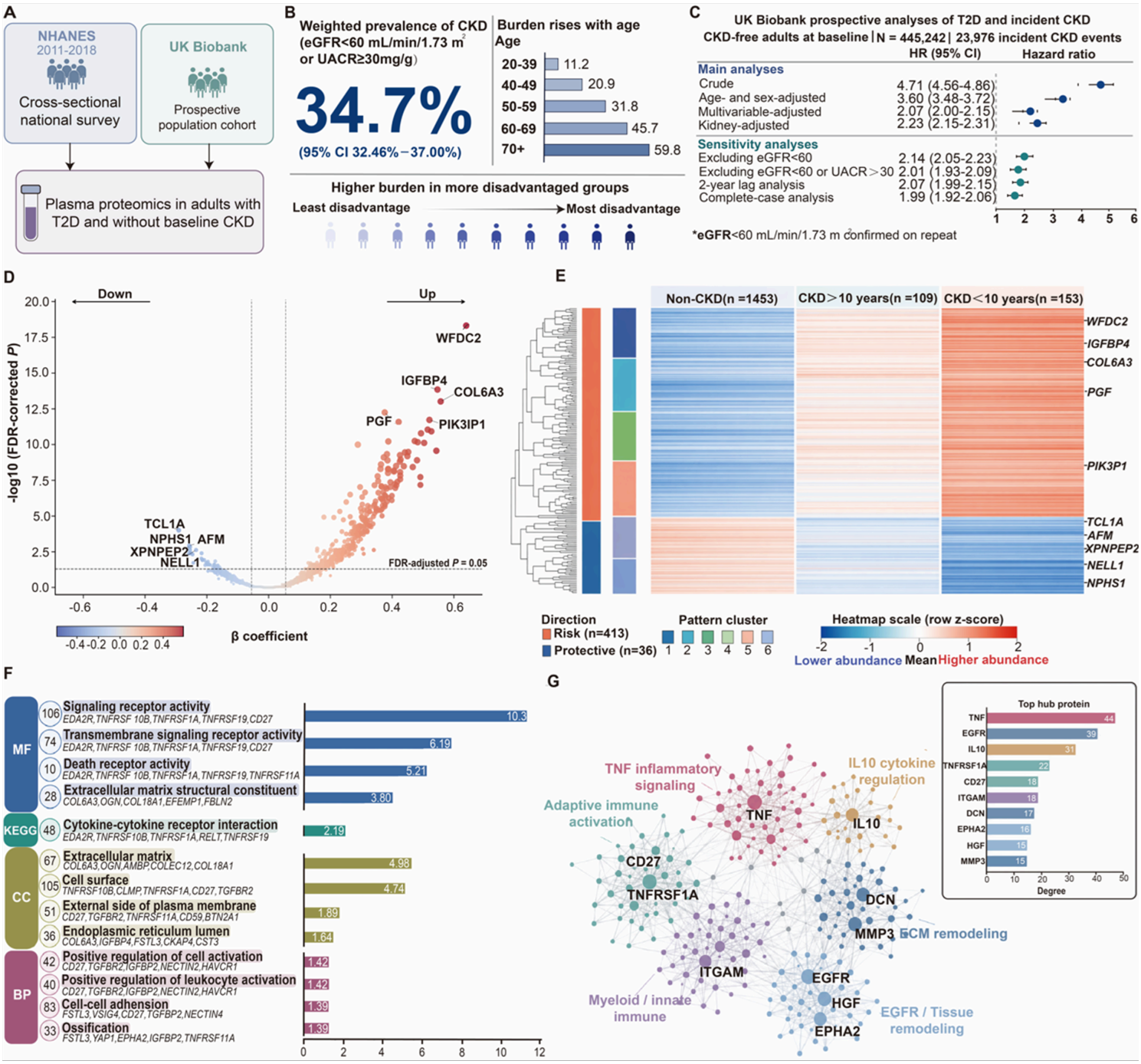


Figure 1

Population-level CKD burden, incident CKD risk, and prediagnostic proteomic signatures in adults with type 2 diabetes.

A. Study overview, including NHANES cross-sectional analyses, UK Biobank prospective analyses, and plasma proteomic analyses among adults with type 2 diabetes (T2D) without baseline CKD. B. Weighted CKD prevalence among adults with T2D in NHANES 2011–2018, overall and by age and socioeconomic disadvantage. CKD was defined as eGFR <60 mL/min/1.73 m² or UACR ≥30 mg/g. C. Association between T2D status and incident CKD in UK Biobank participants free of CKD at baseline. Hazard ratios and 95% CIs are shown for the unadjusted, age- and sex-adjusted, multivariable-adjusted, kidney-

adjusted, and sensitivity analyses. D. Volcano plot of protein-wide Cox associations between baseline plasma protein abundance and incident CKD. The x-axis shows Cox regression β coefficients, and the y-axis shows −log10 FDR-adjusted P values. Selected top positively and inversely associated proteins are labelled. E. Heatmap of baseline abundance patterns for CKD-associated proteins across future CKD timing groups: no incident CKD, incident CKD after >10 years, and incident CKD within ≤10 years. Rows represent CKD-associated proteins, and colors indicate row-standardised abundance. F. Functional enrichment results for FDR-significant CKD-associated proteins across selected molecular function, KEGG, cellular component, and biological process terms. Bar length represents −log10 FDR-adjusted *P*values. G. High-confidence protein–protein interaction network of CKD-associated proteins, with major modules and hub proteins labelled. CKD, chronic kidney disease; eGFR, estimated glomerular filtration rate; FDR, false discovery rate; KEGG, Kyoto Encyclopedia of Genes and Genomes; NHANES, National Health and Nutrition Examination Survey; T2D, type 2 diabetes; UACR, urinary albumin-to-creatinine ratio.

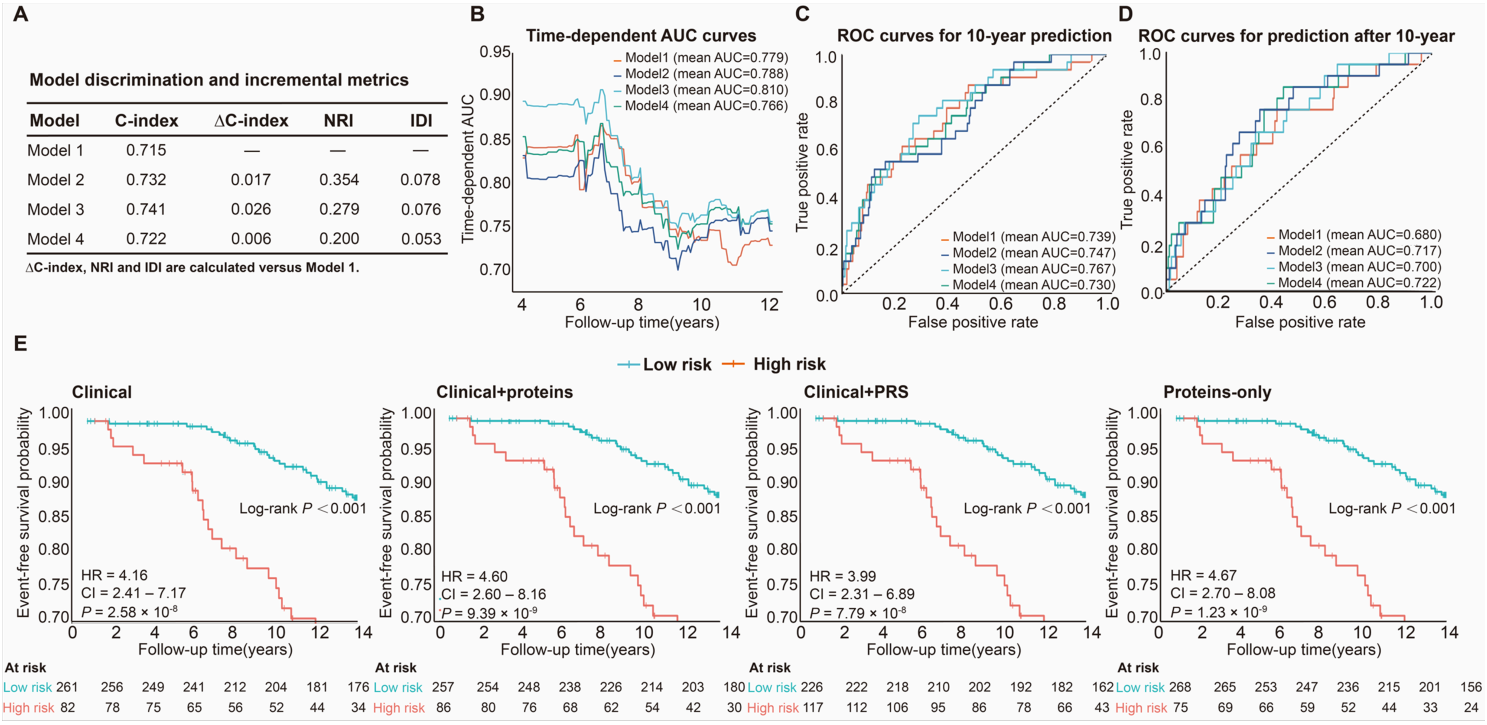


Figure 2

Protein-informed risk stratification models for incident CKD in adults with type 2 diabetes.
A. Model discrimination and incremental performance metrics in the independent test set. Four models were evaluated: Model 1, clinical model; Model 2, clinical plus selected-proteins model; Model 3, clinical plus protein risk score model; and Model 4, protein-only model. ΔC-index, NRI, and IDI were calculated versus the clinical model. B. Time-dependent AUC curves across follow-up. C. ROC curves for 10-year prediction. D. ROC curves for prediction after 10 years. E. Kaplan-Meier curves showing CKD-free survival in low-risk and high-risk groups defined by model-predicted 10-year risk in the independent test set. Cutoffs were derived using X-tile and used for exploratory visualisation. AUC, area under the curve; CKD, chronic kidney disease; IDI, integrated discrimination improvement; NRI, net reclassification improvement; PRS, protein risk score; ROC, receiver operating characteristic.

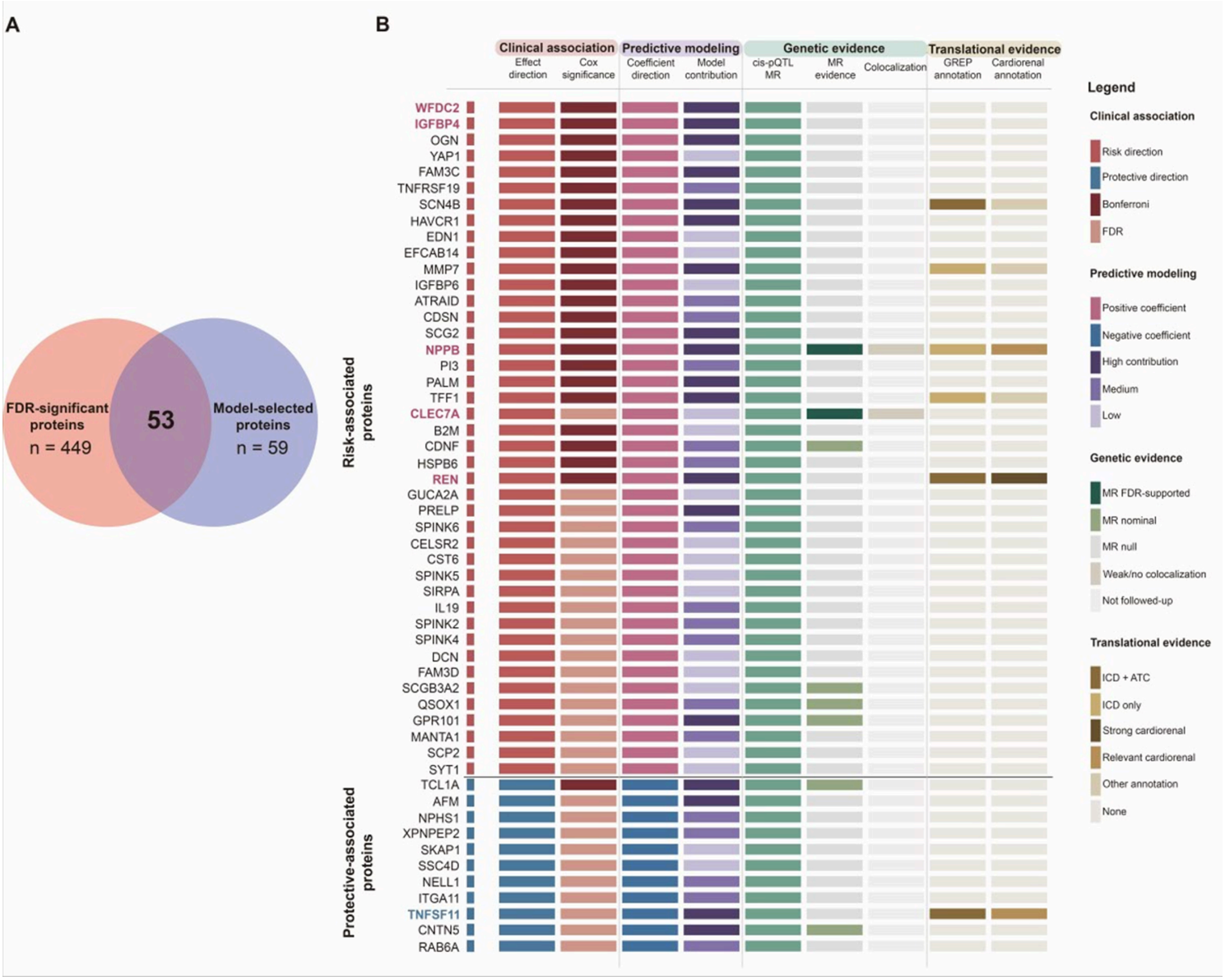


Figure 3

Exploratory prioritisation of overlapping CKD-associated and model-selected proteins.

A. Intersection of FDR-significant CKD-associated proteins from the full-cohort Cox analysis and proteins selected during training-set model development. Fifty-three overlapping proteins were prioritised for exploratory downstream analyses. B. Integrated prioritisation matrix for the 53 overlapping proteins, summarizing clinical association direction and significance, predictive modelling contribution, genetic evidence from Mendelian randomisation and colocalisation analyses, and translational evidence from disease and drug-target annotation. Proteins are grouped according to risk-associated and protective-associated directions. CKD, chronic kidney disease; FDR, false discovery rate; MR, Mendelian randomisation; pQTL, protein quantitative trait locus. These analyses were intended for hypothesis generation and protein prioritisation rather than definitive causal inference.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- GraphicalAbstract.tiff
- SupplementaryTablesS1S17.xlsx
- SupplementFigure.docx