

A glycosylation-related gene signature for diabetic nephropathy: dual-panel development, cross-platform and cross-compartment external validation, immune deconvolution, and characterization of an ECM–glycosaminoglycan module

Chen Yaqi

yaqicn@163.com

Nanyang Medical College

Li Tongping

First Affiliated Hospital of Henan College of Classical Chinese Medicine

Research Article

Keywords: diabetic nephropathy, diabetic kidney disease, glycosylation, ECM–glycosaminoglycan module, FN1, transcriptomics, external validation, LASSO, immune deconvolution

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11305812/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

Diabetic nephropathy (DN) is a leading cause of end-stage renal disease, and available markers incompletely capture disease activity. Protein glycosylation is increasingly implicated in diabetic kidney injury, but the cross-platform reproducibility of glycosylation-centric transcriptional modules remains unclear.

Methods A protein-coding glycosylation-related gene (GRG) list ($n = 812$; Gene Ontology and UniProt glycosylation terms, filtered against NCBI Gene type_of_gene annotation) was constructed. Differentially expressed genes (DEGs) in glomeruli from the discovery cohort GSE96804 (41 DN vs. 20 controls) were intersected with GRGs. Two exploratory LASSO panels (performance-first lambda.min; parsimony-first lambda.1se) were derived by univariate logistic regression, random forest, and LASSO, and assessed by ROC analysis, calibration, decision curve analysis, and leak-free repeated cross-validation with every supervised step re-estimated within each fold. External validation used GSE30528 (glomeruli, 9 DKD vs. 13 controls) and GSE30529 (tubulointerstitium, 10 DKD vs. 12 controls). Immune infiltration was profiled by nu-support-vector-regression (nu-SVR) deconvolution of the official LM22 matrix.

Results Of 610 DEGs, 27 protein-coding GRDEGs (15 up-, 12 down-regulated) remained after the coding-gene filter. GRDEGs were markedly enriched in glycosaminoglycan and heparin binding, extracellular matrix, and integrin/focal adhesion signalling. The 7-gene lambda.min panel outperformed the 5-gene lambda.1se panel on all four evaluation layers. The apparent training AUC was 0.9988 (saturated); leak-free out-of-fold AUCs were 0.9565 ± 0.0210 (lambda.min; optimism 0.0423) and 0.9561 ± 0.0236 (lambda.1se; optimism 0.0415). External AUCs were 0.778 ($p = 0.033$) and 0.683 ($p = 0.156$) for the lambda.min panel and 0.735 ($p = 0.071$) and 0.625 ($p = 0.339$) for the lambda.1se panel. Neither composite score exceeded its best single-gene member in either cohort (best single genes: LPL 0.983, FN1 0.923, EGF 0.967, FN1 0.958). FN1 was up-regulated and significant in all three cohorts (AUC 0.911/0.923/0.958). SELL tracked the neutrophil signal at the expression layer ($r \approx 0.84$) and the proportion layer ($r = 0.31$ – 0.47 across all three cohorts).

Conclusions We characterize a glycosylation-centric ECM–GAG transcriptional module in DN with FN1 as a cross-cohort hub. The composite risk scores were exploratory, did not outperform single genes, and are not a validated diagnostic model; prospective multi-cohort validation is required.

Background

Diabetic nephropathy (DN; diabetic kidney disease, DKD) affects approximately 40% of individuals with diabetes and is the leading cause of end-stage renal disease worldwide [1]. The global burden continues to accelerate, with projections indicating sustained growth of DN-attributed disability and mortality through 2045 [1, 2]. Although renal biopsy remains the diagnostic gold standard, it is invasive and rarely performed; consequently, there is intense interest in molecular markers that capture disease activity non-

invasively, including circulating microRNAs such as miRNA-377 [3], and recent work suggests that currently used clinical and histopathological markers still leave substantial diagnostic gaps [4, 5].

Protein glycosylation — the enzymatic attachment of glycan chains to proteins — regulates protein folding, stability, ligand recognition, and extracellular matrix (ECM) organization, and has emerged as an important dimension of kidney pathobiology [6]. In diabetes, altered N-glycosylation profiles of circulating immunoglobulin G predict incident nephropathy and other microvascular complications [7, 8], and aberrant serum protein glycosylation can distinguish DN from non-diabetic renal diseases [9]. At the tissue level, the glomerular glycocalyx and heparan sulfate proteoglycans are early targets of diabetic injury, and glycocalyx-protective strategies delay DN progression experimentally [10, 11]. Despite this convergent evidence, the reproducibility of glycosylation-centric transcriptional modules in the kidney across platforms and renal compartments has not been established; a recent prognostic study of glycosylation-related genes in DN did not address cross-cohort or cross-compartment generalization [12].

Public transcriptomic cohorts now permit such analyses. GSE96804 provides glomerular expression profiles from 41 patients with DN and 20 non-diabetic controls on a uniform Affymetrix platform [13]; transcription-regulatory relationships in this cohort have since been mapped by differential co-expression and differential regulation analysis [14]. The paired cohorts GSE30528 (glomeruli) and GSE30529 (tubulointerstitium) allow cross-platform and cross-compartment testing [15]. Machine-learning approaches — univariate logistic filtering, random forest (RF) ranking, and LASSO penalization — have become standard for compact biomarker panel construction in DN [16–19].

In the present study, we (i) constructed a reproducible, protein-coding GRG list from Gene Ontology and UniProt; (ii) identified GRDEGs between DN and control glomeruli; (iii) characterized their functions through GO and KEGG enrichment; (iv) derived two exploratory LASSO risk-score panels ($\lambda_{\min}$ and λ_{1se}) through three-step machine learning; (v) evaluated discrimination, calibration, clinical utility, and selection stability in the training set — including a leak-free repeated cross-validation in which every supervised step was re-estimated within each fold — and in two external cohorts spanning platforms and renal compartments; and (vi) profiled immune infiltration by nu-SVR deconvolution with the official LM22 matrix, benchmarked against rank-based scoring, and related it to the signature genes. The study is framed as a characterization of the DN ECM–glycosaminoglycan module; the risk scores are reported as exploratory, dual-panel adjuncts rather than a validated diagnostic model.

Methods

Study design and data sources

This study used only publicly available, de-identified data from the Gene Expression Omnibus (GEO) [20] and therefore required no additional ethical approval. The discovery/training cohort was GSE96804 (platform GPL17586, Affymetrix Human Gene 2.0 ST Array), comprising 41 glomerular samples from

patients with DN and 20 glomerular samples from the unaffected portion of tumor nephrectomies; group labels were verified from the series metadata [13]. Two external validation cohorts from the same original study design were used: GSE30528 (platform GPL571, Affymetrix U133A; 9 DKD vs. 13 control glomeruli) and GSE30529 (GPL571; 10 DKD vs. 12 control tubulointerstitial samples) [15]. All three cohorts were downloaded as series matrices with their phenotypic tables (Table 1).

Table 1
Datasets and study design

Role	GEO accession	Compartment	Groups	Platform	Genes (post-mapping)	Source
Discovery / training	GSE96804	Glomeruli	41 DN vs. 20 control	GPL17586 (HuGene-2.0 ST)	30,906	Pan et al. [13]
External validation 1	GSE30528	Glomeruli	9 DKD vs. 13 control	GPL571 (U133A)	14,289	Woroniecka et al. [15]
External validation 2	GSE30529	Tubulointerstitium	10 DKD vs. 12 control	GPL571 (U133A)	14,289	Woroniecka et al. [15]

Construction of the glycosylation-related gene (GRG) list

A reproducible GRG list was assembled as the union of: (i) Gene Ontology (GO) [21] biological process and molecular function terms and all their descendants covering protein N-linked and O-linked glycosylation, glycoprotein biosynthesis, aminoglycan/glycosaminoglycan (GAG) metabolism, glycosyltransferase activity, glycosidase activity, GAG binding, and aminoglycan metabolism (807 genes), and (ii) reviewed human UniProt entries [22] annotated with the keywords Glycosyltransferase (KW-0328, n = 232), Glycosidase (KW-0326, n = 78), and Congenital Disorder of Glycosylation (KW-0900, n = 59). The union (831 genes) was then filtered against the NCBI Gene type_of_gene annotation to retain protein-coding genes only, excluding non-coding transcripts (e.g., microRNAs), because the Gene Ontology and UniProt branches do not themselves distinguish coding from non-coding loci. The final list contained 812 unique protein-coding genes (GO branch 789; UniProt branch 343) (Additional file 1).

Preprocessing

Series matrices were parsed, and probes were mapped to official gene symbols using platform annotations (100% mapping for GPL17586; 95.3%, 21,225/22,277 probes, for GPL571). Multiple probes per gene were collapsed by mean of repeated probes (limma avereps). Data distributions were inspected before and after limma normalizeBetweenArrays quantile normalization [23]; GSE96804 was already on the log2 scale, whereas GSE30528/30529 contained author-level median-centered values, so no

additional log transformation was applied. After normalization, the analysis matrices contained 30,906 genes $\times$ 61 samples (GSE96804) and 14,289 genes $\times$ 22 samples (GSE30528/30529).

Identification of DEGs and GRDEGs

Differential expression (DN vs. control) in the training set was computed with limma [23]. DEGs were defined as $|\log_2 \text{ fold change}| > 1$ with Benjamini–Hochberg adjusted $p < 0.05$. GRDEGs were defined as the intersection of DEGs with the 812-gene protein-coding GRG list, which by construction contains no non-coding transcripts.

Functional enrichment

GRDEGs were subjected to GO (biological process, cellular component, molecular function) and KEGG [24] enrichment with clusterProfiler [25]. Significance thresholds followed adj. $p < 0.05$ and $q < 0.25$. KEGG pathway–gene mappings were obtained from the KEGG REST API (rest.kegg.jp).

Construction of the exploratory risk-score panels

Three sequential machine-learning steps were applied in the training cohort. First, each GRDEG was tested by univariate logistic regression, and genes with $p < 0.05$ were retained for odds-ratio (OR) visualization as a forest plot. Second, a random forest (ntree = 1000, random seed 2024) [26] ranked retained genes by mean decrease in Gini, and the top 10 genes advanced. Third, LASSO logistic regression (glmnet, alpha = 1) [27] with 10-fold cross-validation selected two panels under two pre-specified rules: the performance-first lambda.min panel and the parsimony-first lambda.1se panel (with type.measure = "deviance", the glmnet binomial default). Selection stability of both panels was quantified by 1000 bootstrap resamples in which the entire selection pipeline (random forest ranking followed by LASSO) was re-run in each resample, and gene selection frequencies were recorded. The risk score was defined as $\text{RiskScore} = \sum (\beta_i \times \text{expression}_i)$ using the LASSO coefficients of each panel.

Model evaluation in the training set

Discrimination was quantified by ROC analysis with DeLong confidence intervals (pROC) [28]. A nomogram was constructed with the rms package [29], and internal calibration was assessed by bootstrap resampling ($B = 300$). Decision curve analysis (DCA) quantified net benefit across threshold probabilities 0.01–0.99 versus treat-all and treat-none strategies [30]. Group differences in RiskScore were tested with the Mann–Whitney U test. Functional coherence of the model genes was evaluated by GOSemSim [31] using Wang's semantic similarity method: for each model gene, the mean GO-BP semantic similarity to the remaining model genes ("friends" analysis) was computed. To avoid optimism from re-using the discovery set, a leak-free repeated stratified 5-fold cross-validation (100 repeats) was additionally performed in which every supervised step — limma differential expression, intersection with

the GRG list, random forest ranking, and LASSO fitting — was re-estimated inside each training fold, and out-of-fold (OOF) AUCs were aggregated. The difference between the apparent (in-sample) AUC and the mean OOF AUC is reported explicitly as the optimism, and is interpreted as a lower bound on the true generalization gap because the OOF evaluation shares the same 61 samples, platform, and batch. All AUC confidence intervals were derived from the DeLong method implemented in pROC.

External validation

The training-derived LASSO coefficients were applied, without re-fitting, to the quantile-normalized expression matrices of GSE30528 and GSE30529, for both the lambda.min 7-gene panel and the lambda.1se 5-gene panel (all panel genes are represented on GPL571). ROC analysis, bootstrap calibration ($B = 200$), and DCA were repeated in each cohort; the nomograms shown for the validation cohorts illustrate the training-cohort panel applied to each cohort. Single-gene ROC curves and directional consistency (mean difference DKD vs. control versus the training log2FC) were computed for all panel genes in all cohorts, and pairwise Spearman correlations among panel genes were calculated in the training set.

Immune infiltration analysis

Immune deconvolution was performed with the CIBERSORT nu-support-vector-regression algorithm [32, 33] using the official LM22 expression matrix (547 genes $\times$ 22 cell types; obtained from the [mdozmorov/Immuno_notes](#) mirror of the Newman et al. resource), applied after quantile normalization (QN = TRUE). For the median-centered GPL571 matrices, the author-supplied scale was retained and no additional transformation was applied. The ssGSEA scoring of an LM22-derived binary marker matrix (gseapy) [34, 35] was retained only as a rank-based comparator, because rank-based enrichment systematically inflates gene–signature correlations when the gene itself is a member of the signature. Deconvolved fractions (not enrichment scores) were used for all primary immune statistics: Spearman correlations (i) among the 22 cell-type fractions, with Benjamini–Hochberg correction across the 231 pairwise tests, (ii) between each panel gene and each cell type (154 tests), with Benjamini–Hochberg correction across that family, and (iii) a zero-abundance audit identifying cell types absent from renal tissue. A decontamination control was included by re-running the nu-SVR after removing SELL from LM22.

Statistical analysis

Analyses were performed in R 4.5.3 (2026-03-11) and Python 3.12.3. All tests were two-sided, and $p < 0.05$ was considered significant. Multiple testing was controlled by Benjamini–Hochberg correction within each pre-specified family.

Software and package versions

Package versions were as follows. R 4.5.3: limma 3.66.0, clusterProfiler 4.18.4, enrichplot 1.30.4, GOSemSim 2.36.0, GO.db 3.22.0, org.Hs.eg.db 3.22.0, AnnotationDbi 1.72.0, glmnet 5.0, randomForest 4.7–1.2, pROC 1.19.1, rms 8.1-1, e1071 1.7–17, preprocessCore 1.72.0, pheatmap 1.0.13, ggplot2 4.0.3, RCircos 1.2.2 and ggVennDiagram 1.5.7; Python 3.12.3: gseapy 1.3.1, pandas 3.0.5, numpy 2.5.3, scipy 1.18.1, scikit-learn 1.9.1 and matplotlib 3.11.2. Immune deconvolution used the CIBERSORT R implementation with the official LM22 signature matrix, with its quantile-normalisation step re-implemented in base R and the original licence header retained. The complete analysis pipeline is provided as Additional file 2.

Use of large language models

During the preparation of this work, the authors used Claude (Anthropic), a large language model, to assist in generating and debugging the analysis code presented in Additional file 2. The tool was not used to generate or modify the data, to perform or direct the statistical analyses, or to interpret the results. All code produced with this assistance was executed and inspected by the authors and was validated against the expected outputs. The authors reviewed and edited the content as needed and take full responsibility for the integrity of the analyses and for the content of the publication.

Results

Data processing and identification of GRDEGs

After quantile normalization, sample distributions were homogeneous within each cohort (Fig. 1). Limma analysis of the training set (41 DN vs. 20 controls) identified 610 DEGs (330 up-regulated, 280 down-regulated) at $|\log_2\text{FC}| > 1$ and $\text{adj. } p < 0.05$ (Fig. 2a). Intersecting these DEGs with the 812-gene protein-coding GRG list yielded 27 GRDEGs (15 up-regulated, 12 down-regulated; Fig. 2b; Table 2); the single non-coding hit, MIR644A, was excluded by the coding-gene filter and is analyzed separately as a sensitivity item (Discussion). The most strongly dysregulated GRDEGs included FN1 ($\log_2\text{FC} = 4.00$, $\text{adj. } p = 9.1 \times 10^{-8}$), VCAN (2.17), CCL21 (2.72), PTN (1.53), and ANGPTL3 (– 1.52), KNG1 (– 1.91), EGF (– 1.51), GBA3 (– 1.46), and APOH (– 1.45). Hierarchical clustering of the top 20 GRDEGs separated DN from control samples (Fig. 2c). Chromosomal mapping distributed the 27 GRDEGs across 16 chromosomes, with chromosome 1 contributing the largest number ($n = 5$; Fig. 2d).

Table 2

The 27 protein-coding glycosylation-related differentially expressed genes (GRDEGs) MIR644A (non-coding) was removed by the coding-gene filter and is reported as a sensitivity item.

Gene	log2FC	adj. P	Direction	Gene	log2FC	adj. P	Direction
FN1	4.003	9.10E-08	Up	POSTN	1.406	2.37E-04	Up
SLIT3	1.415	1.21E-07	Up	PRELP	1.049	4.81E-04	Up
LPL	-1.271	2.44E-07	Down	THBS2	1.017	5.21E-04	Up
FBN1	1.248	1.88E-06	Up	ANGPTL3	-1.518	5.44E-04	Down
ADAMTS1	1.261	3.93E-06	Up	CFH	1.342	7.48E-04	Up
EGF	-1.506	4.27E-06	Down	VCAN	2.167	1.25E-03	Up
SELL	-1.094	1.25E-05	Down	KNG1	-1.913	1.76E-03	Down
ADAMTS5	1.102	2.43E-05	Up	CCL19	1.809	2.69E-03	Up
APOH	-1.452	5.97E-05	Down	LYVE1	1.428	3.66E-03	Up
CCL21	2.718	6.72E-05	Up	UGT2B7	-1.406	8.15E-03	Down
FUT6	-1.025	1.16E-04	Down	CHI3L1	-1.137	8.65E-03	Down
PCOLCE2	-1.006	1.25E-04	Down	NELL1	1.496	1.45E-02	Up
GBA3	-1.463	1.41E-04	Down	IGF1	-1.379	3.17E-02	Down
PTN	1.529	1.89E-04	Up	—	—	—	—

Up-regulated: n = 15; down-regulated: n = 12. adj. P, Benjamini–Hochberg adjusted.

Functional enrichment of GRDEGs

Enrichment analysis of the 27 protein-coding GRDEGs returned 258 significant GO-BP, 19 GO-CC, and 30 GO-MF terms, and 4 KEGG pathways (Fig. 3a–e). The dominant molecular functions were glycosaminoglycan binding (19/27 genes, adj. $p = 1.0 \times 10^{-28}$) and heparin binding (17/27 genes), with contributors including FN1, SLIT3, LPL, FBN1, ADAMTS1, SELL, ADAMTS5, APOH, PTN, POSTN, PRELP, THBS2, CFH, VCAN, and KNG1. The dominant cellular component was the extracellular matrix. Significant biological processes included positive regulation of PI3K/AKT signaling, regulation of glycoprotein biosynthesis, and very-low-density lipoprotein particle remodeling (LPL/APOH/ANGPTL3). KEGG terms comprised integrin signaling, cholesterol metabolism, focal adhesion, and cytoskeleton in muscle cells (Fig. 3e), consistent with an ECM–integrin–focal adhesion axis dominated by glycosylated matrix proteins and proteoglycans.

Three-step machine learning identifies two exploratory risk-score panels

All 27 protein-coding GRDEGs were significant in univariate logistic regression (all $p < 0.05$; Additional file 3; Fig. 4a). Random forest ranking (OOB error rate 4.9%) placed FBN1, FN1, LPL, SLIT3, CCL21, SELL, EGF, THBS2, PCOLCE2, and ADAMTS1 in the top 10 by mean decrease in Gini (Fig. 4b). LASSO cross-validation yielded two panels (Fig. 4c, d): the performance-first lambda.min panel (lambda = 0.0145) comprising FBN1, FN1, LPL, SLIT3, CCL21, SELL, and EGF, and the parsimony-first lambda.1se panel (lambda = 0.0641) comprising FN1, LPL, SLIT3, SELL, and EGF. Bootstrap selection frequencies (1000 resamples) were 0.91 (LPL), 0.78 (SELL), 0.62 (SLIT3), 0.53 (EGF), 0.53 (CCL21), 0.53 (FBN1), and 0.48 (FN1) at lambda.min; at lambda.1se only LPL (0.92), SELL (0.78), FN1 (0.66), SLIT3 (0.64), and EGF (0.56) exceeded 0.5 (Additional file 4). The lambda.1se criterion therefore excluded FBN1 and CCL21, the two members with the lowest lambda.min selection frequencies (0.53 and 0.53, marginally above FN1 at 0.48), so the two criteria converge on a five-gene parsimonious subset while retaining FBN1 and CCL21 only in the performance-first panel. The lambda.min risk score was:

$$\text{RiskScore (lambda.min)} = 0.7914 \times \text{Exp(FBN1)} + 0.1335 \times \text{Exp(FN1)} - 1.3057 \times \text{Exp(LPL)} + 1.1038 \times \text{Exp(SLIT3)} + 0.0228 \times \text{Exp(CCL21)} - 1.8746 \times \text{Exp(SELL)} - 0.6488 \times \text{Exp(EGF)}$$

(Table 3). Positive weights (FBN1, FN1, SLIT3, CCL21) were assigned to up-regulated GRDEGs and negative weights (LPL, SELL, EGF) to down-regulated GRDEGs, mirroring the direction of differential expression. The lambda.1se risk score was $0.1763 \times \text{Exp(FN1)} - 0.6382 \times \text{Exp(LPL)} + 0.4519 \times \text{Exp(SLIT3)} - 0.9842 \times \text{Exp(SELL)} - 0.3518 \times \text{Exp(EGF)}$.

Table 3 The lambda.min seven-gene LASSO risk score			
Gene	log2FC direction	LASSO coefficient	Contribution
FBN1	Up	+ 0.7914	Positive
FN1	Up	+ 0.1335	Positive
LPL	Down	-1.3057	Negative
SLIT3	Up	+ 1.1038	Positive
CCL21	Up	+ 0.0228	Positive
SELL	Down	-1.8746	Negative
EGF	Down	-0.6488	Negative

Performance of the risk-score panels in the training set

In the training set, the lambda.min risk score separated DN from controls with an apparent AUC of 0.9988 (95% CI 0.9954–1.0000; Fig. 5b), and DN samples carried substantially higher scores (mean - 3.37 vs. -11.20; Mann–Whitney $p = 3.5 \times 10^{-10}$; Fig. 5f). A nomogram integrating the seven genes is shown in Fig. 5a. Bootstrap calibration ($B = 300$) showed close agreement between predicted probabilities and observed proportions (Fig. 5c), and DCA demonstrated positive net benefit across a

wide range of threshold probabilities, exceeding both treat-all and treat-none strategies (Fig. 5d; Additional file 5). Because the training AUC corresponds to 819 of 820 case-control pairs and is therefore saturated, the primary performance estimate is the leak-free repeated cross-validation: OOF AUC = 0.9565 ± 0.0210 for the lambda.min panel and 0.9561 ± 0.0236 for the lambda.1se panel, with an explicit optimism of 0.0423 and 0.0415, respectively, which forms a lower bound on the true generalization gap (Additional file 6; see Methods and Discussion). In the GOSemSim "friends" analysis restricted to the lambda.min panel, FN1 exhibited the highest mean GO-BP semantic similarity (0.462), followed by EGF (0.416) and CCL21 (0.413), indicating that FN1 is the most functionally coherent component of the panel (Fig. 5e); this functional-coherence evidence — together with its cross-cohort differential expression (below) — is the basis for the FN1 hub narrative, not its LASSO selection, whose bootstrap frequency (0.48) reflects the collinearity with FBN1 ($r = 0.90$) discussed below.

External validation across platforms and renal compartments

Without re-fitting, the lambda.min 7-gene panel achieved an AUC of 0.778 (95% CI 0.580–0.976) in the cross-platform glomerular cohort GSE30528 (Mann–Whitney $p = 0.033$) and an AUC of 0.683 (95% CI 0.452–0.915) in the cross-compartment tubular cohort GSE30529 ($p = 0.156$; Figs. 6b and 7b; Table 4). The lambda.1se 5-gene panel achieved 0.735 ($p = 0.071$) and 0.625 ($p = 0.339$), respectively (Additional file 7); the parsimony panel did not reach nominal significance in either cohort, so the performance-first panel is the primary reported panel and both panels are reported side by side (Table 4; Additional file 8). Nomograms, calibration curves, and DCA for both cohorts are shown in Figs. 6a–d and 7a–d (net-benefit tables in Additional files 9 and 10); calibration errors were small in both cohorts.

Table 4
Discrimination of the composite risk scores and of individual panel genes

Cohort	Predictor	AUC	95% CI
GSE96804 (training)	Risk score (lambda.min, 7 genes)	0.9988	0.9954–1.0000
GSE96804 (training)	Risk score (lambda.1se, 5 genes)	0.9976	0.9917–1.0000
GSE96804 (leak-free repeated CV, mean $\pm$ SD)	OOF AUC, lambda.min	0.9565	SD 0.0210
GSE96804 (leak-free repeated CV, mean $\pm$ SD)	OOF AUC, lambda.1se	0.9561	SD 0.0236
GSE96804 (training)	FBN1	0.928	0.859–0.997
GSE96804 (training)	FN1	0.911	0.835–0.987
GSE96804 (training)	SLIT3	0.929	0.866–0.992
GSE96804 (training)	LPL	0.906	0.821–0.991
GSE96804 (training)	CCL21	0.888	0.802–0.974
GSE96804 (training)	SELL	0.854	0.749–0.958
GSE96804 (training)	EGF	0.885	0.800–0.971
GSE30528 (glomeruli, cross-platform)	Risk score (lambda.min, 7 genes)	0.778	0.580–0.976
GSE30528 (glomeruli, cross-platform)	Risk score (lambda.1se, 5 genes)	0.735	0.521–0.950
GSE30528	FBN1	0.761	0.552–0.969
GSE30528	FN1	0.923	0.799–1.000
GSE30528	SLIT3	0.615	0.367–0.864
GSE30528	LPL	0.983	0.943–1.000
GSE30528	CCL21	0.872	0.710–1.000
GSE30528	SELL	0.718	0.484–0.952
GSE30528	EGF	0.735	0.517–0.953
All validation panels use the frozen training-cohort coefficients; every panel gene is represented on GPL571, so no missing-member adjustment is required. OOF, out-of-fold. The repeated-CV rows report the mean OOF AUC across 100 repeats, with every supervised step (including differential expression) re-estimated within each fold. In both external cohorts, the best single-gene marker exceeds both composite scores.			

Cohort	Predictor	AUC	95% CI
GSE30529 (tubules, cross-compartment)	Risk score (lambda.min, 7 genes)	0.683	0.452–0.915
GSE30529 (tubules, cross-compartment)	Risk score (lambda.1se, 5 genes)	0.625	0.381–0.869
GSE30529	FBN1	0.875	0.723–1.000
GSE30529	FN1	0.958	0.887–1.000
GSE30529	SLIT3	0.500	0.234–0.766
GSE30529	LPL	0.950	0.867–1.000
GSE30529	CCL21	0.600	0.333–0.867
GSE30529	SELL	0.942	0.824–1.000
GSE30529	EGF	0.967	0.905–1.000
All validation panels use the frozen training-cohort coefficients; every panel gene is represented on GPL571, so no missing-member adjustment is required. OOF, out-of-fold. The repeated-CV rows report the mean OOF AUC across 100 repeats, with every supervised step (including differential expression) re-estimated within each fold. In both external cohorts, the best single-gene marker exceeds both composite scores.			

Single-gene analysis demonstrated that LPL (AUC 0.983 and 0.950), FN1 (0.923 and 0.958), and EGF (0.735 and 0.967) remained strong, directionally consistent discriminators, that FBN1 (0.761 and 0.875) also transferred well, whereas SELL (0.718 and 0.942), CCL21 (0.872 and 0.600), and SLIT3 (0.615 and 0.500) performed heterogeneously (Fig. 8a, c, d; Table 4; Additional file 11). Directionally, 13 of the 15 gene-by-cohort comparisons matched the training log2FC (Additional file 12); the two discordant comparisons both involved SELL, which was down-regulated in DN glomeruli but up-regulated in the tubular compartment. In the training set, every panel gene individually achieved an AUC of 0.854–0.929 (Fig. 8a).

Notably, in neither external cohort did either composite score exceed its best single-gene member: in GSE30528 the best single gene was LPL (AUC 0.983) versus 0.778 and 0.735 for the lambda.min and lambda.1se composite scores, and in GSE30529 the best single gene was EGF (0.967) versus 0.683 and 0.625, respectively (Table 4). Thus the weighted composite added no discriminative value over the best individual marker in external data, and the single-gene and module-level results, rather than the composite, carry the transferable signal.

Pairwise correlation analysis revealed a tightly coordinated ECM–chemokine module (FN1–FBN1 $r = 0.848$; FN1–SLIT3 $r = 0.871$; FN1–CCL21 $r = 0.799$) opposed to an EGF-centered axis (FN1–EGF $r = -0.742$; FBN1–EGF $r = -0.752$) (Fig. 8b). The near-collinearity of FN1 and FBN1 ($r = 0.848$) explains the panel instability under resampling: LASSO tends to select one member of a near-collinear pair essentially

at random, which is why the "same pipeline, different draw" instability reported above arises, and why the ECM–GAG module – rather than any individual member – is the stable biological unit.

Immune infiltration landscape and its relationship with the signature

Primary immune results are based on nu-SVR deconvolution of the official LM22 matrix (Fig. 9; the cell–cell correlation matrices under both methods are provided in Additional files 13 and 14, with the ssGSEA version retained as a comparator). A zero-abundance audit showed that 10 of the 22 cell types (e.g., naive CD4 T cells, follicular helper T cells, M0 macrophages, eosinophils) received zero fraction in $\geq 50\%$ of samples, consistent with their absence from renal tissue; these types were not interpreted further (Additional file 15). Pairwise correlations among the 22 cell-type fractions were not uniformly positive (96 positive vs. 135 negative of 231 pairs; range -0.67 to 0.58), demonstrating that the uniformly positive structure seen under rank-based scoring was a method artifact; a principal-component audit showed PC1 explaining only 17% of variance and unrelated to disease group ($r = -0.04$, $p = 0.76$), and the correlation structure did not collapse after partialling out PC1 (69 of 231 partial correlations at $q < 0.05$; Additional file 16). Across the 154 model gene–cell type pairs, 37 reached $q < 0.05$ under nu-SVR, whereas the same 154 pairs analysed with the rank-based ssGSEA comparator yielded 113 significant pairs; because both analyses used an identical sample size, correlation statistic and correction family, this discrepancy is itself direct evidence that rank-based scoring inflates apparent significance (Additional files 17 and 18).

The SELL–neutrophil association was reproduced in all three cohorts at the proportion layer ($r = 0.308$, $p = 0.016$, $q = 0.060$ in GSE96804; $r = 0.428$, $p = 0.047$ in GSE30528; $r = 0.473$, $p = 0.026$ in GSE30529; decontaminated LM22 without the SELL row gave $0.327/0.434/0.503$), and at the expression layer by a neutrophil marker score excluding SELL ($r = 0.839$ in GSE96804; 0.562 and 0.579 in the two partition cohorts) (Fig. 9c; Table 5; Additional file 19). ECM-panel members showed the opposite pattern (FBN1 – 0.750 , FN1 – 0.714 , SLIT3 – 0.730 , CCL21 – 0.624 versus neutrophil fraction), and LPL/EGF were positive ($0.563/0.619$), jointly depicting a neutrophil-high/ECM-low axis in a subset of DN glomeruli. The three estimators measure three different quantities – rank-based co-enrichment (0.846 , method-inflated and retired), expression-layer co-expression ($0.839/0.562/0.579$, direct evidence), and mixture proportions (0.308 – 0.473 , conservative primary estimate) – and agree in sign (three-method concordance table in Additional file 20); however, the proportion-level estimate did not survive FDR control in the discovery cohort ($q = 0.060$) and is therefore reported as a directionally consistent, hypothesis-generating signal rather than as a significant association.

Table 5

Model gene-immune cell type Spearman correlations in the training set (nu-SVR deconvolution)

Gene	Immune cell type	r	p
CCL21	Macrophages M2	0.810	2.64E-15
FN1	Macrophages M2	0.806	4.45E-15
SLIT3	Macrophages M2	0.751	3.20E-12
FBN1	Neutrophils	-0.750	3.34E-12
FBN1	Macrophages M2	0.747	4.57E-12
SLIT3	Neutrophils	-0.730	2.48E-11
FN1	Neutrophils	-0.714	1.06E-10
EGF	Macrophages M2	-0.703	2.70E-10
CCL21	Neutrophils	-0.624	7.82E-08
SELL	Macrophages M1	-0.623	8.49E-08
SELL	Neutrophils	0.308	1.58E-02
Spearman correlations between panel genes and nu-SVR deconvolved cell-type fractions (official LM22) in the training set (n = 61). The ten largest-magnitude pairs are shown, together with the SELL-neutrophil pair, which is retained because it is the focus of the three-method concordance analysis. Two-sided p values are given. Benjamini-Hochberg correction across the complete 154-pair family (Additional file 17) left 37 pairs significant at $q < 0.05$ (smallest surviving $r = 0.321$). The SELL-neutrophil pair ($r = 0.308$, $p = 0.016$) did not survive FDR correction ($q = 0.060$) and is therefore reported as a direction-consistent, hypothesis-generating signal rather than as a significant association; it is free of signature-membership autocorrelation, since removing the SELL row from LM22 gave $r = 0.327$.			

Discussion

In this study, we constructed a glycosylation-centric transcriptional portrait of DN and characterized a protein-coding GRDEG set whose dominant signal is an ECM-GAG module (FBN1, FN1, SLIT3, LPL, EGF, and partners). Two LASSO panels – the performance-first lambda.min panel (FBN1, FN1, LPL, SLIT3, CCL21, SELL, EGF) and the parsimony-first lambda.1se panel (FN1, LPL, SLIT3, SELL, EGF) – were derived and tested side by side; because the lambda.min panel outperformed the lambda.1se panel on all four evaluation layers, it is the primary reported panel, while the lambda.1se panel documents the parsimony-first alternative (panel-by-panel trade-off in Additional file 21). Four findings merit emphasis: the dominance of a GAG-binding/ECM module among GRDEGs; the robust, directionally consistent performance of LPL, FN1, and EGF as single-gene markers across all three cohorts; the fact that the weighted composite score did not exceed its best single-gene member in external data; and the embedding of the module in a neutrophil-associated immune axis quantified at both the expression and the proportion layer.

The functional landscape of the 27 protein-coding GRDEGs was dominated by glycosaminoglycan and heparin binding and by the extracellular matrix. This is pathobiologically consistent: the glomerular glycocalyx and heparan sulfate chains are early casualties of diabetic injury, and glycocalyx-preserving interventions delay DN in experimental models [10, 11]. Ten GRDEGs encode classical secreted or membrane-associated ECM glycoproteins (FN1, FBN1, ADAMTS1/5, POSTN, PRELP, THBS2, VCAN, KNG1, CHI3L1), mirroring the glomerulosclerosis and tubulointerstitial fibrosis that define progressive DN [4]. FN1 — the top-ranked gene by log2FC (4.00) and a positive component of the final panel — is a canonical mediator of diabetic matrix expansion: AGEs induce fibronectin upregulation via sphingosine kinase 1 [36], fibronectin–ECM receptor interaction blockade attenuates mesangial activation [37], and FN1 has previously been identified as an immune-infiltration-related biomarker for DN [38]. Its top position in the GOSemSim friends analysis (mean semantic similarity 0.462) further indicates that FN1 acts as the functional hub of the panel.

The behavior of EGF deserves specific comment. EGF was down-regulated in DN glomeruli (log2FC = -1.51) and entered the panel with a negative weight, so a low EGF expression raises the predicted DN probability. This mirrors the established biomarker literature: urinary EGF, largely of tubular origin, declines with tubular atrophy and interstitial fibrosis, and low urinary EGF predicts CKD progression [39, 40], whereas higher levels reflect active tubular repair [41]. Concurrently, EGFR signaling in residual tubules can promote fibrosis [42], reconciling the apparent paradox that loss of tubular EGF output marks severity while misdirected EGFR activation drives scarring. Our data — where EGF is among the most transferable single-gene markers (AUC 0.735–0.967 across all three cohorts) — support the view that EGF-axis integrity is a discriminating feature of DN tissue state.

An immune–chemokine axis was also captured. CCL21 was the most up-regulated chemokine-encoding GRDEG (log2FC = 2.72), and SELL was down-regulated with a strongly negative LASSO weight. Both genes structure leukocyte trafficking through the CCR7/chemokine system [43], and urinary L-selectin (the SELL product) tracks renal inflammatory activity in lupus nephritis [44]. SELL tracked the neutrophil signal in all three cohorts at both the expression layer (marker score excluding SELL: $r = 0.839/0.562/0.579$) and the proportion layer (nu-SVR: $r = 0.308/0.428/0.473$), whereas CCL21, SLIT3, and FN1 tracked M2 macrophage and dendritic-cell signatures and EGF did so inversely. The higher rank-based ssGSEA estimate for SELL–neutrophils ($r = 0.846$) was method-inflated and is reported only as a comparator. An immune-infiltration component in DN pathogenesis is increasingly recognized [16–19, 45], and our results extend it by showing that the glycosylation-related transcriptome is coupled to it — plausibly through heavily glycosylated chemoattractants and selectin ligands.

LPL contributed a negative weight and was the single most transferable marker in validation (AUC 0.983/0.950). Together with APOH and ANGPTL3, it anchored the VLDL/lipoprotein remodeling term in the enrichment analysis, pointing to a lipid-metabolic dimension of the DN glycosylation program. This connects to recent evidence linking the ANGPTL–O-GlcNAc axis to DN [46] and to altered lipoprotein lipase mass in chronic kidney disease [47], and it suggests that glycosylation-dependent lipid handling in the diabetic kidney is a coherent molecular module rather than an artifact of enrichment.

Two features explain why the composite score failed to exceed its best single-gene member in external cohorts. First, the LASSO coefficients were fitted on log2-scale intensities from GPL17586 and applied without re-fitting to the median-centered values of GPL571, so the weighted sum of the composite score is subject to a scale mismatch that a single-gene comparison — which depends only on rank order and is therefore invariant to monotone rescaling — largely escapes. Second, *SELL*, the only panel gene whose direction of change was discordant between renal compartments (down in glomeruli, up in tubules; 2 of 15 gene-by-cohort comparisons), carries the largest absolute coefficient in both panels (-1.875 in the $\lambda_{\min}$ panel and -0.984 in the λ_{1se} panel), so a single reversed gene receives the greatest weight in the composite score. These observations identify score calibration rather than a failure of the underlying biology as the limiting factor for cross-cohort transfer of the weighted composite, and they argue for single-gene or module-level reporting in this setting.

MIR644A was excluded from the final protein-coding GRG list and from the primary model (Methods), and is reported here only as a sensitivity item. Its original inclusion traced to GO annotations linking it to the regulation of glycoprotein biosynthesis (GO:0009101/GO:0010559); miR-644, the microRNA it encodes, has so far been characterized only within the circPFBP/miR-644/ADAMTSL5 regulatory axis in gastric cancer [48], and no renal function has been assigned to it. MIR644A is represented on GPL17586 (probe TC20000241.hg.1, reliably detected, log2 median 7.0) but has no probe on GPL571 and therefore cannot be validated across platforms. Removing this non-coding transcript and re-running the entire selection pipeline left the training AUC unchanged (819/820, a saturated metric), but changed the panel composition (the excluded member was replaced by *FBN1*) and slightly reduced external performance, from 0.795/0.708 to 0.778/0.683. Reassuringly, all seven genes of the revised panel are represented and verifiable on both platforms, whereas the pre-audit panel contained one member that could not be; the analysis therefore does not depend on the non-coding transcript. Separately, *SLIT3* — a secreted heparin-binding glycoprotein implicated in fibrosis via *SLIT3/ROBO1* signaling [49] and in angiogenesis via *Robo4* [50] — has not been functionally characterized in DN; its strong co-expression with *FN1* ($r = 0.871$) suggests participation in the matrix–angiogenesis crosstalk of the diabetic glomerulus, although its effect did not replicate in the tubular compartment.

The performance profile follows a clinically plausible gradient, and the optimism must be stated explicitly: the apparent training AUC (0.9988; 819/820 case–control pairs, i.e., saturated) exceeds the leak-free out-of-fold estimate for the same panel (0.9565) by 0.0423, and this internal optimism is a lower bound on the true generalization gap, because the external AUCs obtained without any re-fitting — 0.778 in the same-compartment cross-platform cohort and 0.683 across renal compartments for the $\lambda_{\min}$ panel — fall 0.22–0.32 below the apparent value, i.e., 5–8 times the internal optimism. The external AUCs therefore represent a more realistic performance ceiling, and the tubular cohort did not reach nominal significance. That *LPL* and *FN1* individually maintained AUCs of 0.92–0.98 in validation, and *EGF* reached 0.967 in the tubular cohort, while neither composite score exceeded its best single-gene member, suggests that a simplified single-marker or module-level strategy is preferable to a weighted composite for translation, a question we suggest testing in future designs.

This study has limitations. First, all cohorts are retrospective public microarray datasets; the validation series were author-normalized (median-centered), so cross-cohort application of frozen coefficients involves a scale mismatch — within-cohort z-scored sensitivity analysis reduced the glomerular AUC from 0.778 to 0.727 and the tubular AUC from 0.683 to 0.633, and cohort-internal standardization is recommended for any future application. Second, the 20 controls are the unaffected portions of tumor nephrectomies and may carry cancer-associated transcriptional changes; this potential confounder is inherent to the discovery design. Third, the validation cohorts are small (n = 22 each), yielding wide confidence intervals, and the tubular cohort did not reach nominal significance for either panel. Fourth, the two validation cohorts are partitions of the same study (Woroniecka et al.) and are therefore described throughout as cross-platform and cross-compartment rather than as independent replications. Fifth, the immune results at the proportion layer rely on LM22 nu-SVR deconvolution, whose absolute fractions are estimates, not counts; rank-based ssGSEA was retained only as a comparator because it systematically inflates gene–signature correlations (0.846 vs. 0.308 for SELL–neutrophils). Sixth, SELL showed direction heterogeneity between the glomerular training set (down) and the tubular compartment (up), and the near-collinearity of FN1 and FBN1 ($r = 0.848$) renders individual member selection within the ECM module unstable under resampling — hence the dual-panel, module-level framing. Seventh, the risk scores are cross-sectional classifiers, not prognostic models: no longitudinal outcomes or clinical covariates (eGFR, albuminuria) were available, and causality cannot be inferred. Eighth, MIR644A could not be validated externally (no GPL571 probe) and no orthogonal (qPCR/protein) assay was performed. Prospective, multi-center validation with matched biopsy histology, protein-level assays, and longitudinal renal outcomes is the necessary next step.

Conclusions

We identified 27 protein-coding glycosylation-related differentially expressed genes in DN glomeruli and characterized an ECM–GAG module (FBN1, FN1, SLIT3, and partners) together with a lipid/chemokine axis (LPL, CCL21, SELL, EGF) that is directionally consistent across a cross-platform glomerular cohort and a cross-compartment tubular cohort. FN1 — up-regulated and significant in all three cohorts (AUC 0.911/0.923/0.958) and the most functionally coherent member of the module — emerges as the cross-cohort hub. Two exploratory risk scores (a performance-first 7-gene panel and a parsimony-first 5-gene panel) discriminated DN in the cross-platform cohort, but the tubular comparison did not reach nominal significance, and neither composite score exceeded its best single-gene member; accordingly, the scores are offered as hypothesis-generating adjuncts to the module characterization rather than as a validated diagnostic model, and MIR644A is reported as a sensitivity item only. Prospective, multi-cohort validation remains the necessary next step.

Abbreviations

AUC area under the curve

BH Benjamini

Hochberg

CI confidence interval
DCA decision curve analysis
DEG differentially expressed gene
DKD diabetic kidney disease
DN diabetic nephropathy
ECM extracellular matrix
GAG glycosaminoglycan
GEO Gene Expression Omnibus
GO Gene Ontology
GO
BP Gene Ontology biological process
GO
CC Gene Ontology cellular component
GO
MF Gene Ontology molecular function
GRDEG glycosylation
related differentially expressed gene
GRG glycosylation
related gene
KEGG Kyoto Encyclopedia of Genes and Genomes
LASSO least absolute shrinkage and selection operator
LM22 22
cell-type leukocyte signature matrix
nu
SVR nu-support vector regression
OOF out
of-fold
OR odds ratio
RF random forest
ROC receiver operating characteristic
ssGSEA single
sample gene set enrichment analysis
VLDL very
low-density lipoprotein.

Declarations

Ethics approval and consent to participate

This study used only publicly available, de-identified data from the Gene Expression Omnibus (accessions GSE96804, GSE30528, GSE30529). No new human participants or human tissue were involved. Ethical approval and consent to participate were therefore not required, and no institutional review board approval was obtained.

Consent for publication

Not applicable.

Availability of data and materials

The datasets analysed during the current study are available in the Gene Expression Omnibus repository under accessions GSE96804 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE96804>), GSE30528 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30528>), and GSE30529 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30529>). All other data generated or analysed during this study are included in this published article and its additional files; the complete analysis pipeline is provided as Additional file 2.

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' contributions

CY designed the study, constructed the GRG list, performed all bioinformatic and statistical analyses, interpreted the results, and drafted the manuscript. LT contributed to the clinical interpretation of the findings and critically revised the manuscript. Both authors read and approved the final manuscript.

Acknowledgements

The authors thank the investigators and participants of the original GSE96804, GSE30528, and GSE30529 studies.

References

1. Lin S, Liu R, Zhong G, et al. Burden and trends of diabetic kidney disease in East Asia, 1990–2038: an analysis of the Global Burden of Disease Study 2023. *Diabetes Metab Syndr*. 2025;19:103333. 10.1016/j.dsx.2025.103333.
2. Jiang L, Jia J, Chen T, et al. High BMI-attributed diabetic kidney disease in China versus globally, 1990–2021, with projections to 2045: divergent trends and an accelerating burden. *PLoS ONE*. 2025;20:e0338118. 10.1371/journal.pone.0338118.
3. Ponnudhali D, Lakshmiprabha S, Dhas PK, Nagarajan P. Circulating miRNA-377 as an epigenetic regulator and diagnostic biomarker in the early detection and progression of diabetic nephropathy. *J Pharm Bioallied Sci*. 2026;18(3):230–2. 10.4103/jpbs.jpbs_127_26.
4. Guan Y, Huang J, Ma J, et al. Research progress on pathogenesis of diabetic nephropathy and intervention with traditional Chinese medicine. *J Multidiscip Healthc*. 2026;19:619464. 10.2147/JMDH.S619464.
5. Owusu-Asante H, Darko SN, Senu E, et al. Podocyte markers as diagnostic biomarkers of diabetic nephropathy: a case-control comparison to haemato-biochemical markers in a Ghanaian population. *Endocrinol Diabetes Metab*. 2026;9:e70240. 10.1002/edm2.70240.
6. Ling Y, Cai F, Su T, et al. Glycosylation in kidney diseases. *Precis Clin Med*. 2025;8(3):pbaf017. 10.1093/pcmedi/pbaf017.
7. Memarian E, Heijmans R, Slieker RC, et al. IgG N-glycans are associated with prevalent and incident complications of type 2 diabetes. *Diabetes Metab Res Rev*. 2023;39:e3685. 10.1002/dmrr.3685.
8. Martinez Escobedo C, Birukov A, Heijmans R, et al. IgG N-glycosylation and incident neuropathy and nephropathy in people with type 2 diabetes: a pooled analysis of prospective cohort studies. *Diabetes Care*. 2026;49:1714–9. 10.2337/dc26-0620.
9. Liu M, Yu H, Zhang D, et al. Alteration of glycosylation in serum proteins: a new potential indicator to distinguish non-diabetic renal diseases from diabetic nephropathy. *RSC Adv*. 2018;8:38872–82. 10.1039/C8RA06832A.
10. Zhang B, Bu C, Wang Q, et al. Low molecular weight heparin promotes the PPAR pathway by protecting the glycocalyx of cells to delay the progression of diabetic nephropathy. *J Biol Chem*. 2024;300:107493. 10.1016/j.jbc.2024.107493.
11. Wu L, Zhang X, Zhao J, et al. The therapeutic effects of marine sulfated polysaccharides on diabetic nephropathy. *Int J Biol Macromol*. 2024;277:129269. 10.1016/j.ijbiomac.2024.129269.
12. Li X, Sun T, Li X, et al. Glycosylation-related genes and prognostic signatures in diabetic nephropathy. *Int J Gen Med*. 2025;18:7075–94. 10.2147/IJGM.S545376.
13. Pan Y, Jiang S, Hou Q, et al. Dissection of glomerular transcriptional profile in patients with diabetic nephropathy: SRGAP2a protects podocyte structure and function. *Diabetes*. 2018;67:717–30. 10.2337/db17-0755.
14. Shi JS, Qiu DD, Le WB, et al. Identification of transcription regulatory relationships in diabetic nephropathy. *Chin Med J*. 2018;131:2886–90. 10.4103/0366-6999.246063.

15. Woroniecka KI, Park AS, Mohtat D, et al. Transcriptome analysis of human diabetic kidney disease. *Diabetes*. 2011;60:2354–69. 10.2337/db10-1181.
16. Wu X, Guo B, Chang X, et al. Identification and validation of glycolysis-related diagnostic signatures in diabetic nephropathy: a study based on integrative machine learning and single-cell sequence. *Front Immunol*. 2024;15:1427626. 10.3389/fimmu.2024.1427626.
17. Liu H, Fu Q, Chen J. Transcriptomic profiling combined with machine learning and Mendelian randomization identifies diagnostic biomarkers and immune infiltration patterns in diabetic kidney disease. *Molecules*. 2026;31:1390. 10.3390/molecules31091390.
18. Duan S, Zhou Q, Shi Y, et al. Immune infiltration and mitochondrial function in diabetic kidney disease: WGCNA and machine learning identified hub genes with clinical validation. *Int J Mol Sci*. 2026;27:4696. 10.3390/ijms27114696.
19. Lan F, Zhao J, Liang D, et al. Comprehensive analysis of cuproptosis-related ceRNA network and immune infiltration in diabetic kidney disease. *Heliyon*. 2024;10:e35700. 10.1016/j.heliyon.2024.e35700.
20. Clough E, Barrett T. The Gene Expression Omnibus database. *Methods Mol Biol*. 2016;1374:93–110. 10.1007/978-1-4939-3578-9_5.
21. Gene Ontology Consortium. The Gene Ontology knowledgebase in 2023. *Genetics*. 2023;224:iyad031. 10.1093/genetics/iyad031.
22. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res*. 2023;51:D523–31. 10.1093/nar/gkac1052.
23. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43:e47. 10.1093/nar/gkv007.
24. Kanehisa M, Furumichi M, Sato Y, et al. KEGG: biological systems database as a model of the real world. *Nucleic Acids Res*. 2025;53:D672–7. 10.1093/nar/gkae909.
25. Wu T, Hu E, Xu S, et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation*. 2021;2:100141. 10.1016/j.xinn.2021.100141.
26. Breiman L. Random forests. *Mach Learn*. 2001;45:5–32. 10.1023/A:1010933404324.
27. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33:1–22. 10.18637/jss.v033.i01.
28. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S + to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:77. 10.1186/1471-2105-12-77.
29. Harrell FE Jr. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. 2nd ed. Cham: Springer; 2015.
30. Vickers AJ, Van Calster B, Wynants L, Steyerberg EW. Decision curve analysis: confidence intervals and hypothesis testing for net benefit. *Diagn Progn Res*. 2023;7:11. 10.1186/s41512-023-00148-y.
31. Yu G, Li F, Qin Y, et al. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*. 2010;26:976–8. 10.1093/bioinformatics/btq064.

32. Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods*. 2015;12:453–7. 10.1038/nmeth.3337.
33. Chen B, Khodadoust MS, Liu CL, et al. Profiling tumor infiltrating immune cells with CIBERSORT. *Methods Mol Biol*. 2018;1711:243–59. 10.1007/978-1-4939-7493-1_12.
34. Barbie DA, Tamayo P, Boehm JS, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*. 2009;461:108–12. 10.1038/nature08460.
35. Hanzelmann S, Castelo R, Guinney J. *BMC Bioinformatics*. 2013;14:7. 10.1186/1471-2105-14-7. GSVA: gene set variation analysis for microarray and RNA-seq data.
36. Chen C, Gong W, Li C, et al. Sphingosine kinase 1 mediates AGEs-induced fibronectin upregulation in diabetic nephropathy. *Oncotarget*. 2017;8:78660–76. 10.18632/oncotarget.20205.
37. Wang J, Zhang Q, Li S, et al. Low molecular weight fucoidan alleviates diabetic nephropathy by binding fibronectin and inhibiting ECM-receptor interaction in human renal mesangial cells. *Int J Biol Macromol*. 2020;150:304–14. 10.1016/j.ijbiomac.2020.02.087.
38. Wang Y, Zhao M, Zhang Y. Identification of fibronectin 1 (FN1) and complement component 3 (C3) as immune infiltration-related biomarkers for diabetic nephropathy using integrated bioinformatic analysis. *Bioengineered*. 2021;12:5386–401. 10.1080/21655979.2021.1960766.
39. Ju W, Nair V, Smith S, et al. Tissue transcriptome-driven identification of epidermal growth factor as a chronic kidney disease biomarker. *Sci Transl Med*. 2015;7:316ra193. 10.1126/scitranslmed.aac7071.
40. Keller F, Denicolo S, Leierer J, et al. Association of urinary epidermal growth factor, fatty acid-binding protein 3, and vascular cell adhesion molecule 1 levels with the progression of early diabetic kidney disease. *Kidney Blood Press Res*. 2024;49(1):1013–25. 10.1159/000542267.
41. Ono K, Maeshima A, Nagayama I, et al. Urinary epidermal growth factor level as a noninvasive indicator of tubular repair in patients with acute kidney injury. *Diagnostics*. 2024;14:947. 10.3390/diagnostics14090947.
42. Harris RC, Zhang MZ. The role of the epidermal growth factor receptor in kidney tubulointerstitial fibrosis. *Semin Nephrol*. 2026;151690. 10.1016/j.semnephrol.2026.151690.
43. Lai J, Shan H, Cui S, et al. Bioinformatics analysis reveals CCR7 as a potential biomarker for predicting CKD progression. *Medicine*. 2024;103:e33705. 10.1097/MD.00000000000033705.
44. Shen Y, Vanarsa K, Yin Z, et al. Urine L-selectin reflects clinical and histological renal disease activity and treatment response in lupus nephritis across multi-ethnicity. *Front Immunol*. 2023;14:1200167. 10.3389/fimmu.2023.1200167.
45. Lin J, Lin Y, Li X, et al. Uncovering the role of anoikis-related genes in modulating immune infiltration and pathogenesis of diabetic kidney disease. *J Inflamm Res*. 2024;17:4975–91. 10.2147/JIR.S446752.
46. Pandey A, Badruddeen, Akhtar J, et al. Impact of angiotensin-like 4 (ANGPTL4) protein and O-linked beta-N-acetylglucosamine (O-GlcNAc) in diabetic nephropathy. *Diabetes Res Clin Pract*. 2026;235:113211. 10.1016/j.diabres.2026.113211.

47. Hitsumoto T. Preheparin serum lipoprotein lipase mass as a coronary risk factor in patients with chronic kidney disease. *Cardiol Res.* 2026;17(1):1–9. 10.14740/cr2192.
48. Kong S, Tian S, Wang Z, et al. Circular RNA circPFKP suppresses the proliferation and metastasis of gastric cancer cell via sponging miR-644 and regulating ADAMTSL5 expression. *Bioengineered.* 2022;13:12326–37. 10.1080/21655979.2022.2073001.
49. Song G, Zhan Y, Ma J, et al. Epicardial adipose tissue-secreted ANGPTL2 drives obesity-induced atrial fibrosis via the SLIT3/ROBO1 and Wnt5a signaling axes. *Cell Signal.* 2026;148:112823. 10.1016/j.cellsig.2026.112823.
50. Chen X, He X, Lin X, et al. Regulation of H-type angiogenesis and permeability in the subchondral bone of osteoarthritis: the role of Slit3 and the Robo4/Rac1-GTP/ ROS axis. *Front Cell Dev Biol.* 2026;14:1787805. 10.3389/fcell.2026.1787805.