



OPEN

## Identifying myoglobin as a mediator of diabetic kidney disease: a machine learning-based cross-sectional study

Ruoru Wu<sup>1,2</sup>, Zhihao Shu<sup>3</sup>, Fei Zou<sup>2</sup>, Shaoli Zhao<sup>3</sup>, Saolai Chan<sup>2</sup>, Yaxian Hu<sup>2</sup>, Hong Xiang<sup>4</sup>, Shuhua Chen<sup>5</sup>, Li Fu<sup>6</sup>, Dongsheng Cao<sup>6</sup> & Hongwei Lu<sup>1,3</sup>✉

In view of the alarming increase in the burden of diabetes mellitus (DM) today, a rising number of patients with diabetic kidney disease (DKD) is forecasted. Current DKD predictive models often lack reliable biomarkers and perform poorly. In this regard, serum myoglobin (Mb) identified by machine learning (ML) may become a potential DKD indicator. We aimed to elucidate the significance of serum Mb in the pathogenesis of DKD. Electronic health record data from a total of 728 hospitalized patients with DM (286 DKD vs. 442 non-DKD) were used. We developed DKD ML models incorporating serum Mb and metabolic syndrome (MetS) components (insulin resistance and  $\beta$ -cell function, glucose, lipid) while using SHapley Additive exPlanation (SHAP) to interpret features. Restricted cubic spline (RCS) models were applied to evaluate the relationship between serum Mb and DKD. Serum Mb-mediated renal function impairment induced by MetS components was verified by causal mediation effect analysis. The area under the receiver operating characteristic curve of the DKD machine learning models incorporating serum Mb and MetS components reached 0.85. Feature importance analysis and SHAP showed that serum Mb and MetS components were important features. Further RCS models of DKD showed that the odds ratio was greater than 1 when serum Mb was > 80. Serum Mb showed a significant indirect effect in renal function impairment when using MetS components such as HOMA-IR, HGI and HDL-C/TC as a reason. Moderately elevated serum Mb is associated with the risk of DKD. Serum Mb may mediate MetS component-caused renal function impairment.

Diabetic kidney disease (DKD) is a common microvascular complication of diabetes<sup>1,2</sup>, and 50% of patients with type 2 diabetic mellitus (T2DM) will eventually develop DKD<sup>3</sup>. Novel biomarker discovery and accurate prediction models will aid DKD intervention and reduce disease burden<sup>4,5</sup>. However, currently available predictive models often lack reliable biomarkers and perform poorly<sup>6-9</sup>.

Components of metabolic syndrome (MetS), a cluster of risk indicators reflecting conditions such as insulin resistance (IR), dyslipidemia, and comorbidities (i.e. prothrombotic and proinflammatory states)<sup>10</sup>, have been widely shown to be associated with cardiovascular disease events<sup>11,12</sup>. MetS components are also used for DKD risk prediction, but their predictive power is not ideal. For example, based on the ADVANCE study (2012) of 18,517 T2DM patients, the COX proportional hazards regression approach indicated that baseline blood glucose was not a good predictor of DKD, whereas baseline estimated glomerular filtration rate (eGFR) and urinary albumin-creatinine ratio (ACR) dominated risk prediction<sup>13</sup>.

To identify any potential DKD-related indicators, we performed machine learning (ML)-based data analysis to evaluate all relevant indicators in 51,866 hospitalized patients with T2DM. Interestingly, we found serum myoglobin (Mb), a 17 kDa monomeric hemoglobin<sup>14</sup> that present in human myocardial and skeletal muscle

<sup>1</sup>Health Management Center, The Third Xiangya Hospital of Central South University, Changsha 410013, Hunan, People's Republic of China. <sup>2</sup>Xiangya School of Medicine, Central South University, Changsha 410013, Hunan, People's Republic of China. <sup>3</sup>Department of Cardiology, The Third Xiangya Hospital, Central South University, No. 138, Tongzi, Changsha 410013, Hunan, People's Republic of China. <sup>4</sup>Center for Experimental Medicine, The Third Xiangya Hospital of Central South University, Changsha 410013, Hunan, People's Republic of China. <sup>5</sup>Department of Biochemistry, School of Life Sciences, Central South University, Changsha 410013, Hunan, People's Republic of China. <sup>6</sup>Xiangya School of Pharmaceutical Sciences, Central South University, Changsha 410013, People's Republic of China. ✉email: hongweilu@csu.edu.cn

tissues to facilitate the diffusion of O<sub>2</sub> from the blood to muscle mitochondria<sup>15</sup>, as an important feature when excluding renal function indicators (Fig. S1). Since this is only a preliminary analysis finding and the influences of missing values and related diseases are not considered, further in-depth analysis is needed. From a molecular mechanism perspective, subclinical elevations in serum Mb have recently been observed in DM and chronic kidney disease (CKD)<sup>16</sup>, which may be explained by diabetic muscle damage and abnormal glucose and lipid metabolism, suggesting that MetS components may be contributing to elevated serum Mb levels<sup>17–20</sup>. Further, tissue damage caused by Mb has been confirmed at the molecular level. For example, studies have shown that Mb interacts with glucose to form end-of-glycosylation end products (AGEs), which ultimately lead to intramolecular and/or intermolecular cross-linking of long-lived proteins, causing chronic inflammation<sup>21–23</sup>. However, the importance of serum Mb in DKD has not been evaluated, nor has serum Mb been investigated as a risk factor in models of DKD. Moreover, the pathological role of serum Mb as a mediator in MetS component-induced renal function impairment has not been reported.

Therefore, to clarify the possible pathophysiological significance of serum Mb in DKD, we used a cross-sectional study design and obtained a total of 51,866 electronic health records (EHRs) of hospitalized patients with T2DM in the Third Xiangya Hospital of Central South University. We established DKD ML models incorporating serum Mb and simultaneously evaluated the models' performance and feature importance to verify the importance of serum Mb in DKD. Logistic regression-based restricted cubic spline (RCS) models were applied to further assess the relationship between serum Mb and DKD. We explored the causal relationship between MetS components, serum Mb, and DKD using causal mediation analysis, aiming to test the hypothesis that serum Mb mediates the induction of renal function impairment by MetS components.

## Materials and methods

**Clinical data sources.** After obtaining the data access authorization from the Information Network Center of the Third Xiangya Hospital, we retrospectively collected the EHRs of 51,866 hospitalized patients from the Hospital Information System (HIS). The Third Xiangya Hospital of Central South University is located in Changsha City, Hunan Province. It is a grade A tertiary general hospital serving the middle reaches of the Yangtze River. This study was approved by the IRB of the Third Xiangya Hospital of Central South University (No: 21112). All data had been completely anonymized prior to our access to the HIS. Because the patients' personally identifiable information had been effectively protected, the IRB of the Third Xiangya Hospital of Central South University waived the requirement for informed consent. The present study was conducted in accordance with the guidelines of the Declaration of Helsinki.

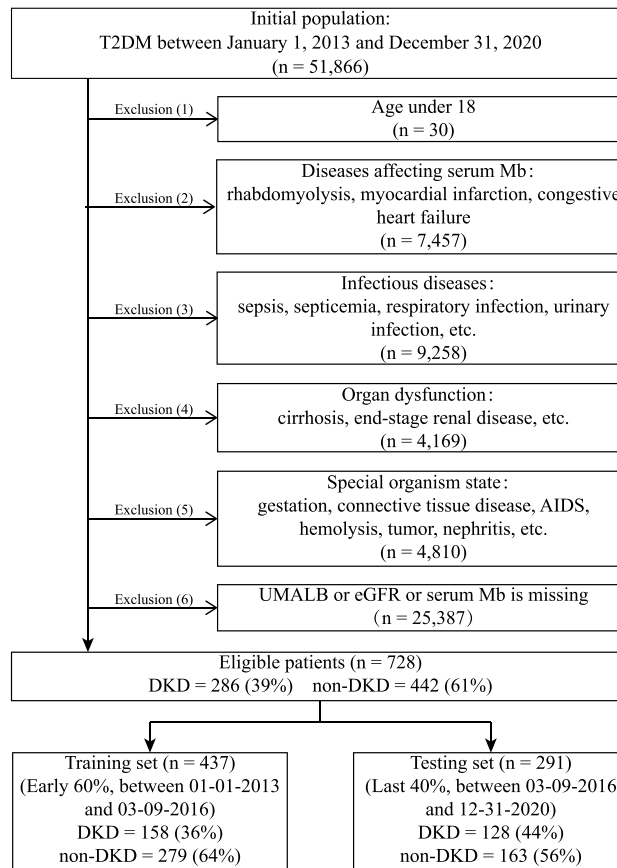
**Study population.** The EHRs of 51,866 hospitalized patients with T2DM in the Third Xiangya Hospital from January 1, 2013 to December 31, 2020 were collected. Exclusion criteria: (1) under 18 years of age ( $n = 30$ ), (2) diseases affecting Mb ( $n = 7457$ ), (3) infectious diseases ( $n = 9258$ ), (4) organ dysfunction ( $n = 4169$ ), (5) special body status ( $n = 4810$ ), and (6) missing urine microalbumin (UMALB), eGFR and/or serum Mb data ( $n = 25,387$ ). Finally, 728 patients (DKD = 286, non-DKD = 442) were eligible for the study. Eligible patients were divided into training set (60%, earlier than March 9, 2016) and testing set (40%, later than March 9, 2016), including 437 (DKD = 158, non-DKD = 279) and 291 patients (DKD = 128, non-DKD = 163), respectively, based on hospitalization date. The inclusion and exclusion criteria of the study population are presented in Fig. 1, and the demographic and clinical features are available in Table 1.

**Data preparation.** From the EHRs of T2DM, we extracted a total of 157 features in the following 7 domains: demographic information, hospitalization-related information, general information, laboratory tests, diagnosis, medication, and others. The data domain categories and descriptions are listed in Table 2. The details of feature calculation are shown in Table S1.

Diagnoses were organized using the ICD -9 and ICD-10 hierarchies. These diseases were diagnosed by experienced doctors who were licensed to practice medicine. T2DM was diagnosed according to the World Health Organization (WHO) criteria (1999)<sup>24</sup>. DKD was diagnosed by the NKF/KDOQI (2007) classification<sup>25</sup>. Serum Mb was tested by Mb-LATEX CN from Denka Seiken (Reference range 0–70 ng/mL).

The components of MetS have the following three sets of features: IR and  $\beta$ -cell function were evaluated by fasting C-peptide, 120 min C-peptide, Homeostatic Model Assessment of Insulin Resistance<sup>26</sup> (HOMA-IR), Gutt index<sup>27</sup>, triglyceride-glucose (TyG) index<sup>28</sup>, fasting insulin to glucose ratio<sup>29</sup> (I/G 0 min), 120 min insulin to glucose ratio<sup>29</sup> (I/G 120 min), insulinogenic index<sup>30</sup> (IGI), IGI to HOMA-IR ratio<sup>31</sup> (IGI/HOMA-IR), and Homeostasis Model of Assessment for beta-cell function<sup>32</sup> (HOMA-BETA). Glucose metabolism indicators included fasting glucose, 120 min glucose, glycated hemoglobin A1c (GHBA1C), and hemoglobin glycation index<sup>33</sup> (HGI). Lipid metabolism was assessed by total cholesterol (TC), triglycerides (TG), low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), HDL-C/TC, and TG/HDL-C ratio<sup>34</sup>.

**ML algorithms.** Two popular ML methods, extreme gradient boosting (XGBoost) and random forest (RF), were used to build DKD classification models based on EHR data. RF, first proposed by Breiman in 2001<sup>35</sup>, is an ensemble of unpruned classification or regression trees created using bootstrap samples of training data and random feature selection in tree induction. XGBoost is an efficient and scalable implementation of the gradient boosting framework<sup>36</sup>. It is used to develop models in a sequential stagewise fashion like other boosting methods and generalize them by allowing optimization of arbitrary differentiable loss functions. RF and XGBoost were applied using the randomForest (4.6–14)<sup>37</sup> and xgboost (1.3.2.1)<sup>38</sup> R packages, respectively. All ML methods used grid search and fivefold cross-validation for parameter adjustment. The final hyperparameters for the DKD models are listed in Table S2.



**Figure 1.** Inclusion and exclusion criteria of study population. Electronic health records (EHRs) of 51,866 hospitalized patients with type 2 diabetic mellitus (T2DM) were collected. Patients meeting the following exclusion criteria were removed: (1) under 18 years of age ( $n = 30$ ), (2) diseases affecting serum myoglobin (Mb) ( $n = 7457$ ), (3) infectious diseases ( $n = 9258$ ), (4) organ dysfunction ( $n = 4169$ ), (5) special body status ( $n = 4810$ ), and (6) missing urine microalbumin (UMALB), estimated glomerular filtration rate (eGFR) and/or serum Mb data ( $n = 25,387$ ). Finally, 728 patients were eligible for our study. The training set and testing set sample sizes were 437 and 291, respectively.

**Model validation.** The evaluation indicators of the confusion matrix, including sensitivity (SE), specificity (SP), positive predictive value (PPV), negative predictive value (NPV), recall, F1, and balanced accuracy (BA), were used to analyze the discrimination ability of DKD ML models. The probability scores output by the models were used to calculate the area under the receiver operating characteristic (ROC) curve (AUC). These statistical parameters are defined as follows:

$$SE = \frac{TP}{TP + FN},$$

$$SP = \frac{TN}{TN + FP},$$

$$PPV = \frac{TP}{TP + FP},$$

$$NPV = \frac{TN}{TN + FN},$$

$$\text{Recall} = \frac{TP}{TP + FN},$$

Feature	Non-DKD (n = 442)	DKD (n = 286)	P
Male	232 (52.5)	161 (56.3)	0.352
Age	60.00 [52.00, 68.75]	63.00 [54.00, 70.00]	0.013
BMI (kg/m <sup>2</sup> )	23.74 [22.11, 25.95]	24.22 [22.04, 26.14]	0.577
mean SBP (mmHg)	130.29 [120.19, 139.00]	139.57 [129.91, 148.86]	<0.001
mean DBP (mmHg)	77.34 [72.16, 83.00]	78.38 [74.00, 83.78]	0.040
mean HR (beats/minute)	77.00 [74.63, 79.25]	77.07 [74.43, 80.86]	0.204
Hypertension	217 (49.1)	208 (72.7)	<0.001
Hyperlipidemia	132 (29.9)	83 (29.0)	0.873
Diabetic retinopathy	123 (27.8)	165 (57.7)	<0.001
Diabetic neuropathy	231 (52.3)	207 (72.4)	<0.001
Coronary heart disease	64 (14.5)	56 (19.6)	0.087
ACCI	5.00 [4.00, 7.00]	6.00 [5.00, 8.00]	<0.001
FRS	14.00 [11.00, 16.00]	15.00 [12.00, 17.00]	<0.001
UMALB (mg/L)	14.60 [8.35, 28.00]	203.35 [41.64, 721.58]	<0.001
Scr (μmol/L)	68.00 [58.00, 80.00]	96.50 [72.00, 141.25]	<0.001
eGFR (ml/min/1.73m <sup>2</sup> )	96.26 [78.26, 115.96]	66.76 [41.16, 92.65]	<0.001
BUN (mmol/L)	5.39 [4.48, 6.44]	6.56 [5.27, 8.53]	<0.001
Mb (ng/mL)	48.65 [37.70, 65.52]	80.47 [56.92, 120.67]	<0.001
CK (U/L)	79.00 [57.00, 111.00]	89.00 [62.00, 132.00]	0.014
CK-MB (U/L)	16.00 [13.00, 20.38]	18.00 [13.00, 22.00]	0.084
LDH (U/L)	195.00 [171.00, 225.00]	203.00 [177.00, 247.00]	0.013
<b>20 MetS components</b>			
Fasting C-peptide (ng/mL)	1.48 [0.90, 2.12]	1.71 [1.10, 2.46]	<0.001
120 min C-peptide (ng/mL)	2.49 [1.39, 4.43]	2.72 [1.58, 4.20]	0.316
HOMA-IR	1.87 [0.98, 3.21]	1.96 [1.08, 3.48]	0.228
Gutt index	133.69 [119.65, 160.31]	133.05 [120.73, 155.43]	0.811
TyG index	6.91 [6.38, 7.50]	7.03 [6.45, 7.50]	0.412
I/G 0 min	0.15 [0.10, 0.25]	0.18 [0.12, 0.30]	0.002
I/G 120 min	0.20 [0.12, 0.37]	0.22 [0.14, 0.37]	0.472
IGI	0.25 [0.09, 0.62]	0.24 [0.11, 0.58]	0.700
IGI/HOMA-IR	0.16 [0.04, 0.42]	0.15 [0.04, 0.44]	0.982
HOMA-BETA	21.03 [9.66, 46.36]	25.87 [12.59, 49.02]	0.044
Fasting glucose (mmol/L)	8.14 [6.30, 10.75]	8.00 [6.12, 10.38]	0.272
120 min glucose (mmol/L)	11.60 [8.70, 14.10]	11.40 [8.53, 14.64]	0.898
GHBA1C (%)	8.55 [7.00, 10.20]	8.10 [6.80, 9.70]	0.033
HGI	0.25 [- 0.91, 1.79]	- 0.01 [- 1.06, 1.33]	0.022
TC (mmol/L)	4.45 [3.84, 5.29]	4.49 [3.80, 5.41]	0.867
TG (mmol/L)	1.58 [1.03, 2.10]	1.62 [1.16, 2.20]	0.133
LDL-C (mmol/L)	2.30 [1.85, 2.82]	2.22 [1.90, 2.74]	0.668
HDL-C (mmol/L)	1.15 [0.98, 1.34]	1.10 [0.99, 1.30]	0.343
HDL-C/TC (%)	0.25 [0.21, 0.31]	0.24 [0.21, 0.29]	0.430
TG/HDL ratio (%)	1.35 [0.86, 2.17]	1.47 [0.93, 2.15]	0.326

**Table 1.** Demographic and clinical features. Data are presented as median [interquartile range] for continuous variables or number (%) for categorical variables. P-values for differences between groups were based on the Wilcoxon test for continuous variables and the chi-square test for categorical variables. Details of features calculation are shown in Table S1. *Non-DKD* Diabetes without any kidney disease, *SBP* systolic blood pressure, *DPB* diastolic blood pressure, *HR* heart rate, *ACCI* age-adjusted Charlson Comorbidity Index, *FRS* Framingham Risk Score, *UMALB* urinary albumin excretion, *eGFR* estimated glomerular filtration rate, *Scr* serum creatinine, *BUN* urea, *Mb* myoglobin, *CK* creatine kinase, *CK-MB* creatine kinase Isoenzyme, *LDH* lactate dehydrogenase, *HOMA-IR* Homeostatic Model Assessment of Insulin Resistance, *TyG index* triglyceride-glucose index, *I/G 0 min* fasting Insulin to glucose ratio, *I/G 120 min* 120 min Insulin to glucose ratio, *IGI* insulinogenic index, *IGI/HOMA-IR* IGI to HOMA-IR ratio, *HOMA-BETA* Homeostasis Model of Assessment for beta-cell function, *GHBA1C* glycated hemoglobin A1c, *HGI* hemoglobin glycation index, *TC* total cholesterol, *TG* triglycerides, *LDL-C* low-density lipoprotein cholesterol, *HDL-C* high-density lipoprotein cholesterol and *TG/HDL ratio* triacylglycerol to high density lipoprotein cholesterol ratio.

Domain	Descriptions	Data type	Number of features
Demographics	Gender, age	Binary, numeric	2
Hospitalization characteristics	Hospitalization date, leave date, hospital day, admission department, discharge department	Date, numeric, integer	5
General information	Height, weight, BMI, mean SBP, SD SBP, max SBP, min SBP, mean DBP, SD DBP, max DBP, min DBP, mean HR, SD HR, max HR, min HR	Numeric	15
Laboratory tests	GHBA1C, fasting glucose, fasting insulin, fasting C-peptide, 120 min glucose, 120 min insulin, 120 min C-peptide, TC, TG, LDL-C, HDL-C, HDL-C/TC ratio, HGI, TyG index, TG/HDL ratio, HOMA-IR, HOMA-BETA, IGI, IGI/HOMA-IR, I/G 0 min, I/G 120 min, Gutt index, WBC, Hb, PLT, MCV, MCH, MCHC, Neu, Lym, Lym%, Neu%, CK, LDH, CK-MB, Mb, troponin, BNP, UMALB, BUN, Scr, UA, eGFR, C-reactive protein, hypersensitive C-reactive protein, NLR, PLR	Numeric, integer	47
Diagnoses	T2DM, hyperlipidemia, DKD, CKD, stages of CKD, renal anemia, renal bone disease, DR, stages of DR, DN, diabetic foot, CAD, heart failure, stages of heart failure, extremity arteriosclerosis, carotid atherosclerosis, renal artery stenosis and sclerosis, cerebral arteriosclerosis, retinal arteriosclerosis, acute MI, old MI, congestive heart failure, stroke, end-stage renal disease, sepsis, septicemia, HT, stages of HT, peptic ulcer, fatty liver disease, cirrhosis, viral hepatitis, chronic bronchitis, COPD, asthma, arrhythmia, atrial flutter and atrial fibrillation, pregnancy period, hyperthyroidism, hypothyroidism, hyperparathyroidism, hashimoto's thyroiditis, brain tumors, breast tumors, lung tumors, digestive system tumors, urinary system tumors, thyroid tumors, tumors of the male reproductive system, tumors of the female reproductive system, leukemia, lymphoma, metastasis, respiratory system infection, urinary system infection, rheumatoid or connective tissue disease, other cerebrovascular diseases or TIA, other liver diseases, HIV/AIDS, hemiplegia, Alzheimer's disease, rhabdomyolysis, hemolysis, nephritis, infection, tumors	Numeric, binary	66
Medications	Diuretics, beta-blockers, calcium-channel blockers, ACE inhibitor, angiotensin-receptor blocker, vasodilators, statins, fibrates, other lipid-lowering agents, sulfonyleureas, glinides, biguanides, thiazolidinediones, alpha-glucosidase inhibitors, DPP-4 inhibitor, SGLT-2 inhibitor, insulin, antihypertensive agents, lipid-lowering agents, hypoglycemic agents	Binary	20
Others	ACCI, FRS	Integer	2

**Table 2.** Data domain categories and descriptions. *SBP* systolic blood pressure, *DPB* diastolic blood pressure, *HR* heart rate, *GHBA1C* glycosylated hemoglobin A1c, *TC* total cholesterol, *TG* triglycerides, *LDL-C* low-density lipoprotein cholesterol, *HDL-C* high-density lipoprotein cholesterol, *WBC* white blood cell count, *Hb* hemoglobin, *PLT* platelet count, *MCV* mean corpuscular volume, *MCH* mean corpuscular hemoglobin, *MCHC* mean corpuscular hemoglobin concentration, *Neu* neutrophils, *Lym* lymphocyte, *Neu%* percentage of Neu, *Lym%* percentage of Lym, *CK* creatine kinase, *LDH* lactate dehydrogenase, *CK-MB* creatine kinase isoenzyme, *Mb* myoglobin, *UMALB* urinary albumin excretion, *BUN* urea, *Scr* serum creatinine, *UA* uric acid, *HGI* hemoglobin glycation index, *TyG index* triglyceride-glucose index, *TG/HDL ratio* triacylglycerol to high density lipoprotein cholesterol ratio, *HOMA-IR* homeostatic model assessment of insulin resistance, *IGI* insulinogenic index, *IGI/HOMA-IR* IGI to HOMA-IR ratio, *HOMA-BETA* homeostasis model of assessment for beta-cell function, *I/G 0 min* fasting insulin to glucose ratio, *I/G 120 min* 120 min insulin to glucose ratio, *eGFR* estimated glomerular filtration rate, *NLR* neutrophil-to-lymphocyte ratio, *PLR* platelet-to-lymphocyte ratio, *CKD* chronic kidney disease, *DR* diabetic retinopathy, *DN* diabetic neuropathy, *CAD* coronary artery disease, *MI* myocardial infarction, *HT* hypertension, *ACCI* age-adjusted Charlson comorbidity index, *FRS* Framingham risk score.

$$F1 = \frac{2 \times PPV \times recall}{PPV + recall},$$

$$BA = \frac{SE + SP}{2},$$

where true positive (TP) is the number of correctly classified objects, true negative (TN) is the number of correctly misclassified objects, false positive (FP) is the number of incorrectly classified objects, and false negative (FN) is the number of incorrectly misclassified objects.

**Feature evaluation and model interpretation.** Feature importance of RF was estimated by (1) mean decrease in prediction accuracy without the feature in the model and (2) mean decrease in the Gini index, a measure of impurity of the dataset (i.e. risk of misclassifying data), by including the feature. For both parameters, the higher the score, the more important the feature.

We evaluated feature contributions to model predictions using an interpretability method called SHapley Additive exPlanation (SHAP). SHAP is based on the Shapley value, a concept introduced in the 1950s to measure each player's contribution to a collaborative game<sup>39</sup>. In 2017, Lundberg and Lee proposed a broadly applicable SHAP method to interpret a variety of models, including regression and classification<sup>40,41</sup>. For binary classification in XGBoost, the output of the model is the log odds ratio, that is, the SHAP values sum up to the log loss of the model for each sample. The function is defined as follows

$$z = \varphi_0 + \sum_{i=1}^M \varphi_i x_i',$$

where  $z$  is the log odds ratio,  $x'$  is a simplified binary feature vector containing values 1 or 0 for observed and missing features,  $M$  is the number of simplified input features, and  $\phi_i$  is the attribution value of the  $i$ th molecular descriptor or the SHAP value. In probability theory,  $z$  is the corresponding value of probability  $p$  after a certain transformation, and the transformation function is defined as follows

$$z = \log \frac{p}{1-p},$$

where  $p$  varies monotonically in the range (0, 1) and  $z$  varies from infinitesimal to infinity. A feature is predicted to be a risk factor if its  $z$ -value is greater than 0. The larger the  $z$ -value, the more likely it is a risk factor rather than a protective factor. Therefore, the SHAP method can help us understand the model more intuitively. The SHAPforxgboost (0.1.1) and fastshap (0.0.5) R packages were applied in the analysis.

**Decision curve analysis (DCA).** We used DCA to assess the clinical utility of ML models. A decision curve is generated by plotting net benefit against a range of threshold probabilities. Decision curves include benefits and harms on the same scale, so they can be directly compared, supporting comparisons between models. The net benefit is calculated as

$$\text{Netbenefit} = \frac{\text{Truepositive}}{N} - \frac{\text{Truenegative}}{N} \times \frac{P_t}{1-P_t},$$

where  $N$  is the total sample size and  $P_t$  represents the threshold probability. Two extreme strategies, classifying for all and classifying for none, were added for reference. The R package rmda (1.6) was used.

**RCS models.** The association between serum Mb and DKD was evaluated on a continuous scale with RCS curves based on logistic regression models. To balance best fit and overfitting, the number of knots was chosen empirically to be 4. The serum Mb associated with the lowest risk of DKD was the concentration with the lowest odds ratio on the spline curve. Odds ratios were adjusted for gender, age, BMI, hyperlipidemia, hypertension, age-adjusted Charlson comorbidity index (ACCI), and hospitalization date.

**Causal mediation analysis.** Causal mediation analysis identifies potential pathways to explain the observed association between a given exposure and a particular outcome<sup>42</sup>, examining how a third intermediate variable, the mediator, is related to the observed outcome relationship. By using a counterfactual framework, causal mediation analysis allows us to decompose the total effect (TE) into direct effects (DE) and indirect effects (IE). In our hypothetical model, we defined TE as the effect of MetS components on renal function impairment, DE as the effect of MetS components on renal function impairment after removing the individual effect of Mb on renal function impairment, and IE as the effect of MetS components on renal function impairment specially associated with Mb. A directed acyclic graph (DAG) illustrating mediation process is shown in Fig. 5A. Through the causal mediation analysis, we estimated the mediation effect of serum Mb based on logistic regression in all eligible patients, DKD or non-DKD subgroups. According to previous DKD investigators<sup>43,44</sup>, we adjusted and standardized the MetS components, Mb, UMALB, and eGFR for 7 variables: gender, age, BMI, hyperlipidemia, hypertension, ACCI, and hospitalization date. The proportion of serum Mb-mediated renal function impairment due to the MetS components was calculated as

$$\text{Prop.Mediated(average)} = \frac{\log(IE)}{\log(IE) + \log(DE)},$$

Causal mediation analysis was performed using the R package mediation (4.5.0).

**Statistical analysis.** R software (version 4.0.3) was used as a statistical analysis tool in this study. Data are presented as median [interquartile range] for continuous variables or number (%) for categorical variables. Differences between groups were based on the Wilcoxon test for continuous variables and the chi-square test for categorical variables. Relationships between MetS components, serum Mb, and DKD were assessed with the Spearman's correlation coefficient ( $r_s$ ), controlling for 7 variables: gender, age, BMI, hyperlipidemia, hypertension, ACCI, and admission date. All statistical significance was defined as  $P < 0.1$ . Missingness was addressed by setting missing categorical data to 0, whereas for numerical data, imputation was completed using the predictive mean matching (PMM) strategy.

**Bias.** Observer bias may arise from the data on patients hospitalized in a grade A tertiary general hospital, as these patients may have more severe than average DM or DKD. Besides, selection bias caused by exclusion of UMALB missing individuals brings DKD rate elevation but comparable cardiovascular risks, leading to limited representation (Table S3). However, the significance and possible mediator role of serum Mb in DKD, as a common rule, was mainly focused in our study, and we believe that the bias did not greatly affect our conclusion.

Models	SE	SP	PPV	NPV	Recall	F1	BA	AUC
XGBoost-88	0.719	0.853	0.793	0.794	0.719	0.754	0.786	0.864
RF-88	0.688	0.908	0.854	0.787	0.688	0.762	0.798	0.884
XGBoost-83	0.523	0.865	0.753	0.698	0.523	0.618	0.694	0.806
RF-83	0.477	0.920	0.824	0.691	0.477	0.604	0.698	0.824

**Table 3.** Performance of machine learning algorithms in testing set (n = 291).

## Results

**Participant characteristics.** The clinical characteristics of the participants are presented in Table 1. There were no significant between-group differences in age, sex, and BMI. However, the DKD group showed significantly more diabetes complications such as diabetic retinopathy and diabetic neuropathy. Importantly, serum Mb was significantly elevated in the DKD group, consistent with previous studies<sup>45</sup>.

**DKD model performance.** To verify and evaluate the significance of serum Mb in DKD, we built DKD ML models incorporating serum Mb, taking advantage of its freedom from collinearity in processing high-dimensional EHR data on disease phenotypes and clinical outcomes<sup>46–48</sup>, aiming to discover important features overlooked by traditional methods. We first accessed the model performance to ensure that the following feature importance analysis and model interpretation are accurate and reliable.

After removing features with a missing ratio of more than 40% or zero variance, a total of 88 important features (including renal function) were available for DKD modeling to obtain high-accuracy DKD classification models. Two popular ML algorithms, XGBoost and RF algorithms, have been applied, named XGBoost-88 and RF-88, respectively. The model performances in the testing set are summarized in Table 3 and the ROC curves are shown in Fig. 2A. Overall, both approaches performed satisfactorily (e.g. all AUCs > 0.85). Clearly, the best classifier based on the RF algorithm was able to achieve AUC values of 0.884, BA values of 0.798, and PPV values of 0.854 in the testing set.

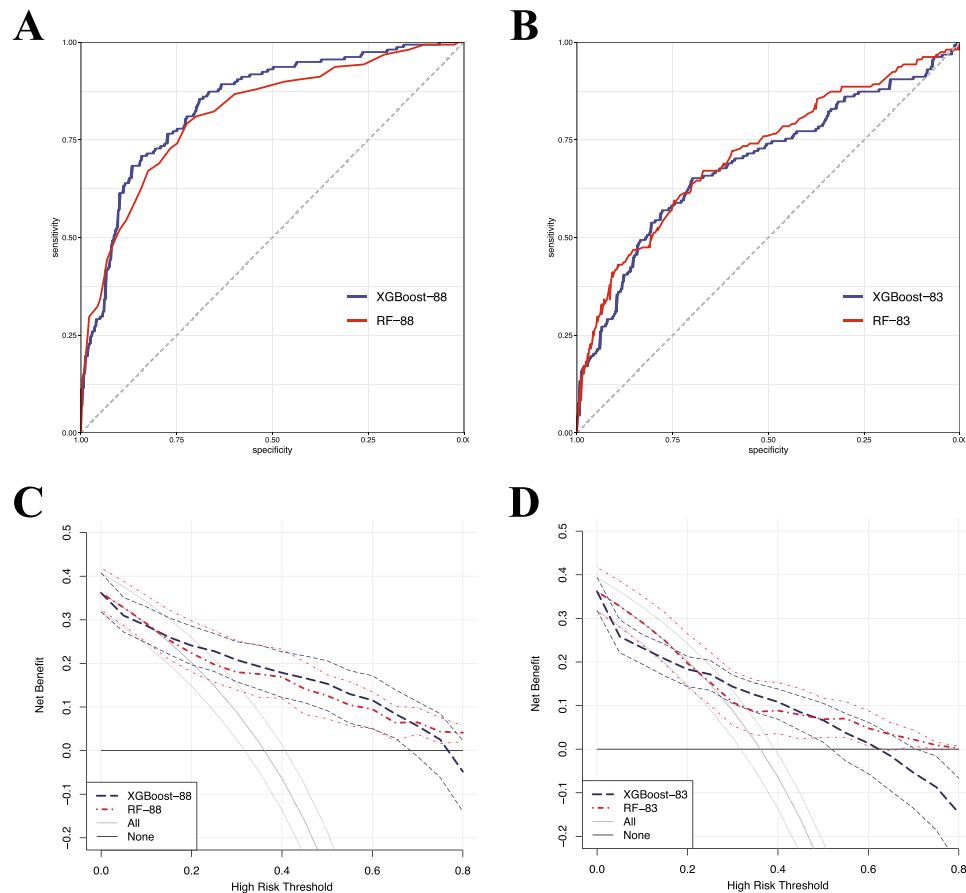
We also applied XGBoost and RF algorithms to renal function exclusion dataset (contained 83 features and removed UMALB, serum creatinine (Scr), eGFR, blood urea nitrogen (BUN), and serum uric acid (SUA)), aiming to find the features hidden by renal function features. XGBoost-83 and RF-83 were named respectively. Model performances in the testing set are summarized in Table 3 and ROC curves are shown in Fig. 2B. On the whole, the performances of both approaches were relatively satisfactory (e.g. all AUCs > 0.80). Compared with the model including renal function features, BA decreased by about 0.1, SP did not change much, and SE decreased by 0.2 in the renal-function-excluded-model, indicating that renal function indicators are crucial for identifying DKD, while other features are important for ruling out DKD. SE sensitivity, SP specificity, PPV positive predictive value, NPV negative predictive value, BA balanced accuracy, AUC area under curve.

**DCA for ML models.** To assess the clinical utility of several ML models we built, DCA was applied to these ML models (Fig. 2C,D). For both XGBoost/RF-88 and XGBoost/RF-83, the net benefits of two models were similar across the range of threshold probabilities. When the threshold probability was between 0.4 and 0.6, both models outperformed the reference strategies. These results of DCA show the consistency and efficiency of our models and are consistent with our model performance data (Table 3).

**Feature importance analysis and model interpretation.** Next, to identify important features hidden from indicators of renal function and attest the importance of serum Mb for DKD, feature importance analysis and model interpretation were applied, the reliability of which has been proved by model performance and DCA.

The top 30 important variables for RF-88 and RF-83 are presented in Fig. 3A,B, and two estimators, mean decrease accuracy and mean decrease Gini, have been applied to measure feature importance. In RF-88, UMALB, eGFR, and Scr were listed as top-ranked features. It's worth noting that serum Mb consistently ranked in the front. Meanwhile, MetS components such as I/G 120 min, I/G 0 min, HOMA-BETA, HOMA-IR, TyG index, Gutt index, 120 min C-peptide (reflecting IR and  $\beta$ -cell function), HGI, fasting glucose (reflecting glucose metabolism), and LDL-C, TG/HDL ratio (reflecting lipid metabolism) were on the list. Intriguingly, in RF-83, serum Mb emerged as the most important feature under all estimators. Meanwhile, MetS components such as 120 min C-peptide, fasting C-peptide, IGI, IGI/HOMA-IR, HOMA-BETA, HOMA-IR, I/G 0 min, TyG index (reflecting IR and  $\beta$ -cell function), fasting glucose (reflecting glucose metabolism), LDL-C, HDL-C (reflecting lipid metabolism) were relatively important features.

To overcome the black-box nature of XGBoost and RF algorithms and better understand the prediction models, model interpretation based on XGBoost-88 correct classifications was carried out using the SHAP approach. The top 30 important features identified by SHAP are shown in Fig. 3C. We found that the renal function features UMALB, Scr, and eGFR remained dominant. Serum Mb (ranked 13th) was recognized as an important feature. MetS components, including HOMA-IR, fasting C-peptide, 120 min C-peptide, IGI/HOMA-IR, HOMA-BETA, I/G 0 min, Gutt index (reflecting IR and  $\beta$ -cell function), 120 min glucose, fasting glucose, HGI (reflecting glucose metabolism), and HDL-C (reflecting lipid metabolism) were important features. These results are consistent with the feature importance analysis described above. We noted that higher serum Mb, worsening renal functions (i.e. higher UMALB and Scr and lower eGFR), more diabetes comorbidities (i.e. diabetic neuropathy and diabetic



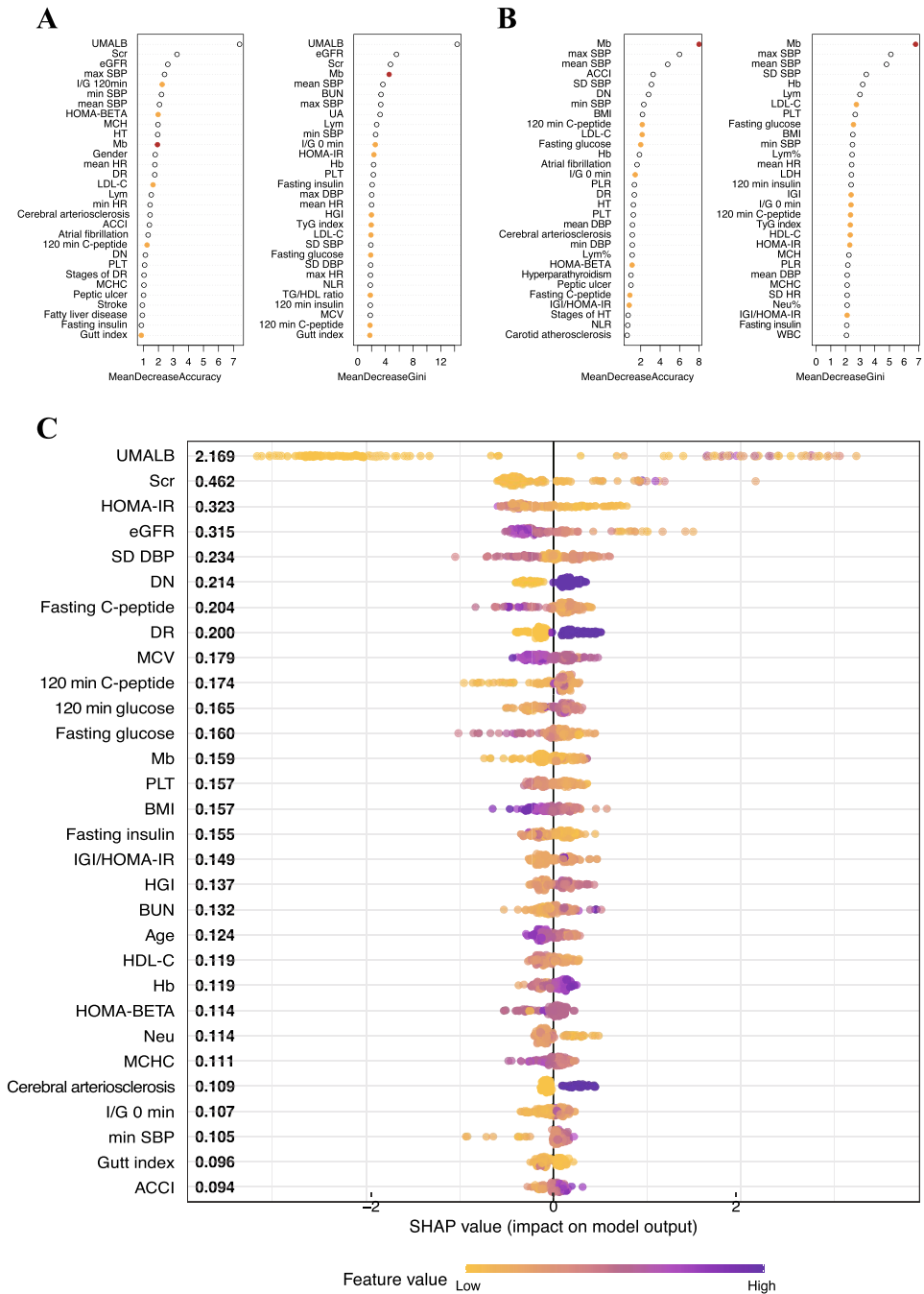
**Figure 2.** Receiver operating characteristic (ROC) curves and decision curve analysis (DCA) results in the testing set. (A,B) ROC curves of XGBoost and RF models based on 88 features and 83 features, respectively. (C,D) DCA for classifying DKD in T2DM patients by machine learning (ML) models based on 88 features and 83 features, respectively. The net benefit and 95% confidence interval for the ML models are shown with blue and red dashed lines. Two extreme strategies, classifying for all and classifying for none, are represented by solid grey and solid black lines.

retinopathy), worsening IR and  $\beta$ -cell function (higher IGI/HOMA-IR, AC, HOMA-BETA, and I/G 0 min), and worse glucose control (e.g. higher HGI and AGLU) were associated with DKD.

Our results suggest that in addition to renal function features, serum Mb is a hidden important feature in DKD, and that elevated serum Mb is associated with DKD classification. Meanwhile, MetS components also play a key role in helping to build ML decision-making algorithms in DKD.

**RCS identifies the association between serum Mb and DKD.** By using ML algorithm-based variable importance analysis and SHAP, we verified the importance of serum Mb for DKD discrimination, supporting serum Mb as a potential important feature of DKD. Next, we further evaluated the association between serum Mb and DKD using RCS on a continuous scale (Fig. 4A,B). We found that the odds ratio was greater than 1 when serum Mb was greater than about 80, indicating that moderately elevated serum Mb is associated with DKD risk. Similar results were seen in male and female patients. However, when adjusting for eGFR, serum Mb was not significant associated with DKD across the range (Fig S2), indicating a strong correlation between serum Mb and eGFR.

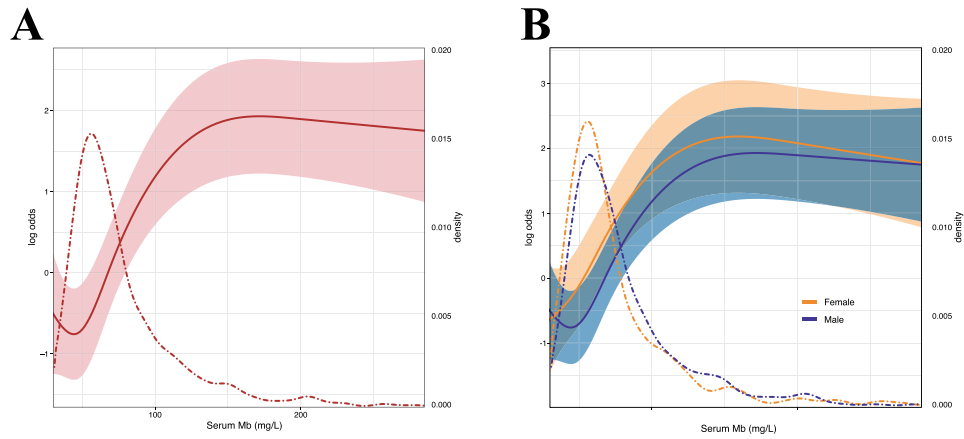
**Causal mediation analysis reflects serum Mb as a possible mediator of MetS component-induced renal function impairment.** Based on the aforementioned ML variable importance analysis, SHAP approach, and RCS results, we noticed that serum Mb and MetS components were closely related to DKD. However, studies have shown that serum Mb is elevated in the case of metabolism disorders, and Mb may directly or indirectly cause tissue damage. Thus, we systematically evaluated whether serum Mb mediates renal impairment associated with the MetS components (20 indicators), and the most significant results are shown in Fig. 5. The spearman correlation coefficient between MetS components and among MetS components, serum Mb, and renal function indicators are shown in Fig. S3 and Table S4, indicating that each component was irreplaceable. Renal function impairment (or DKD severity) was evaluated by eGFR and UMALB.



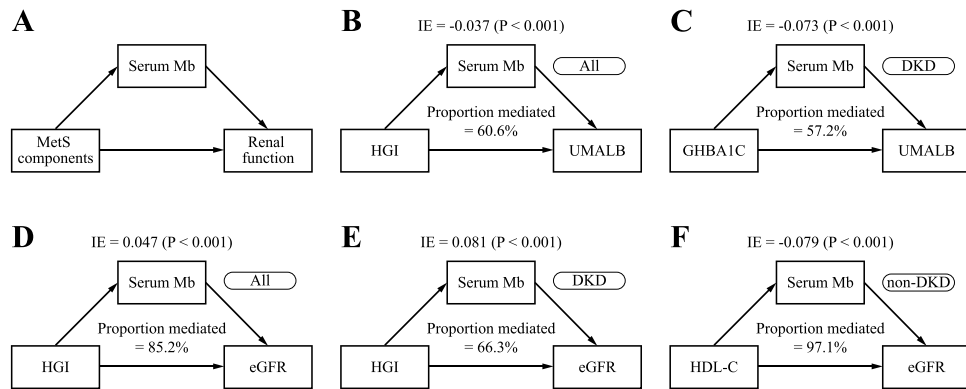
**Figure 3.** Feature importance analysis and model interpretation results. **(A,B)** Feature importance was estimated by mean decrease accuracy (left) and mean decreased Gini index (right) based on RF-88 and RF-83, respectively. Serum Mb and metabolic syndrome (MetS) components are indicated by red and orange dots, respectively. **(C)** Global and local importance for top 30 important features detecting by SHapley Additive exPlanation (SHAP).

Significant IE and TE were found in 8 out of 20 causal mediation models when evaluating renal impairment using eGFR across all observations. Six indicators (HGI, 120 min glucose, fasting glucose, HOMA-IR, IGI/HOMA-IR and IGI in descending order of proportion mediated, the same below) reflecting IR,  $\beta$ -cell function, and glucose metabolism in the DKD subgroup and two indicators (HDL-C, HDL-C/TC) reflecting lipid metabolism in the non-DKD subgroup showed significant IE and TE (see Table 4 for significant results and Table S5 for full results).

Significant IE and TE were found in 3 out of 20 causal mediation models when evaluating renal impairment using UMALB across all observations. 5 indicators (Gutt index, I/G 0 min, 120 min glucose, GHBA1C, and HGI)



**Figure 4.** Restricted cubic spline models for log odds ratios of DKD versus serum Mb on a continuous scale. Analyses were adjusted for gender, age, BMI, hyperlipidemia, hypertension, ACCI, and hospitalization date. **(A)** The red solid line is the multivariable adjusted log odds ratio, and the shaded area shows the 95% confidence interval derived from restricted cubic spline regressions with four knots. The red dashed curve represents the distribution density of the population with different serum Mb levels. **(B)** The orange and blue solid lines are the multivariable adjusted log odds ratio for female and male subgroups, respectively, and the shaded areas show the 95% confidence intervals derived from restricted cubic spline regressions with four knots. The orange and blue dashed curves show the distribution density of populations with different serum Mb levels in female and male subgroups, respectively.



**Figure 5.** Serum Mb mediates the relationship between MetS components and renal function. **(A)** A directed acyclic graph (DAG) illustrating the mediation process. **(B–F)** Results of causal mediation analysis of 5 significant indirect effects (IE) for serum Mb in all eligible patients, DKD or non-DKD subgroups. **(B,C)** Significant IE seen in all eligible patients and DKD subgroup using UMALB as an indicator of renal function. **(D,E,F)** Significant IE seen in all eligible patients, DKD and non-DKD subgroups, respectively, using eGFR as an indicator of renal function.

reflecting IR,  $\beta$ -cell function, and glucose metabolism in the DKD subgroup and no indicator in the non-DKD subgroup showed significant IE and TE (see Table 5 for significant results and Table S6 for full results).

Based on the above systematic evaluation, we support that serum Mb is a potential mediator of MetS component-induced renal function impairment. Generally, compared with UMALB, an indicator of renal function impairment, eGFR has more significant IE and TE as well as higher mediated proportion, indicating serum Mb as a mediator mainly participates in the decline of eGFR, that is, the later phase of DKD progression. The results also suggest that lipid metabolism-induced renal function impairment through serum Mb is important in the non-DKD subgroup, whereas IR,  $\beta$ -cell function, and glucose metabolism are important in the DKD subgroup. These may indicate different causes of elevated serum Mb, and therefore different glycemic and/or lipid control strategies should be employed.

### Discussion

**Serum Mb is a potential risk factor for DKD.** Current DKD prediction models based on large cohorts of T2DM patients, such as ADVANCE study (2012)<sup>13</sup>, Diabetes Cohort Study (DCS) (2013)<sup>49</sup>, and National Diabetes Register (NDR) (2011)<sup>44</sup>, are generally limited by lack of biomarkers and poor performance. A recently

Group	MetS components	Number of patients (%)	P <sub>IE</sub>	IE [95% CI]	DE [95% CI]	TE [95% CI]	Proportion mediated (%)
All	Fasting C-peptide	728 (100.0)	0.02	[- 0.07, - 0.01]	[- 0.30, - 0.16]	[- 0.35, - 0.20]	13.5
	HOMA-IR		<0.001	[0.02, 0.08]	[0.09, 0.22]	[0.15, 0.28]	22.1
	IGI		<0.001	[- 0.10, - 0.03]	[- 0.32, - 0.19]	[- 0.39, - 0.26]	17.6
	IGI/HOMA-IR		0.02	[- 0.07, 0.00]	[- 0.17, - 0.04]	[- 0.21, - 0.08]	20.6
	Fasting glucose		<0.001	[0.02, 0.09]	[0.10, 0.21]	[0.15, 0.28]	23.9
	120 min glucose		0.04	[0.00, 0.07]	[0.00, 0.11]	[0.03, 0.16]	39.5
	HGI		<0.001	[0.03, 0.11]	[- 0.05, 0.07]	[0.01, 0.15]	85.2
	HDL-C/TC		<0.001	[- 0.10, - 0.03]	[- 0.08, 0.05]	[- 0.15, 0.00]	81.4
DKD	HOMA-IR	286 (39.3)	<0.001	[0.04, 0.15]	[0.05, 0.27]	[0.12, 0.39]	32.1
	IGI		<0.001	[- 0.16, - 0.03]	[- 0.38, - 0.18]	[- 0.48, - 0.25]	19.7
	IGI/HOMA-IR		0.08	[- 0.14, 0.00]	[- 0.26, - 0.02]	[- 0.31, - 0.07]	25.4
	Fasting glucose		<0.001	[0.05, 0.17]	[0.06, 0.27]	[0.15, 0.40]	34.6
	120 min glucose		<0.001	[0.04, 0.15]	[- 0.01, 0.17]	[0.07, 0.27]	50.1
	HGI		<0.001	[0.05, 0.20]	[- 0.05, 0.14]	[0.07, 0.26]	66.3
Non-DKD	HDL-C	442 (60.7)	<0.001	[- 0.13, - 0.03]	[- 0.10, 0.08]	- 0.17, 0.00]	97.1
	HDL-C/TC		<0.001	[- 0.10, - 0.02]	[- 0.20, - 0.02]	[- 0.25, - 0.08]	36.3

**Table 4.** DE, IE, and TE of MetS components on renal function impairment (eGFR) mediated through serum Mb (Significant). *TE* total effect, *DE* direct effects, *IE* indirect effects, *CI* confidence interval.

Group	MetS components	Number of patients (%)	P <sub>IE</sub>	IE [95% CI]	DE [95% CI]	TE [95% CI]	Proportion mediated (%)
All	Fasting C-peptide	728 (100.0)	0.08	[0.00, 0.05]	[0.06, 0.23]	[0.07, 0.27]	13.3
	I/G 0 min		<0.001	[0.02, 0.08]	[0.01, 0.15]	[0.05, 0.20]	32.9
	HGI		<0.001	[- 0.07, - 0.01]	[- 0.07, 0.04]	[- 0.11, 0.00]	60.6
DKD	Gutt index	286 (39.3)	0.10	[- 0.01, 0.12]	[- 0.03, 0.17]	[0.01, 0.22]	40.2
	I/G 0 min		<0.001	[0.02, 0.12]	[- 0.01, 0.22]	[0.04, 0.28]	39.8
	120 min glucose		<0.001	[- 0.11, - 0.03]	[- 0.25, - 0.08]	[- 0.32, - 0.14]	25.8
	GHBA1C		<0.001	[- 0.13, - 0.04]	[- 0.15, 0.03]	[- 0.23, - 0.04]	57.2
	HGI		<0.001	[- 0.11, - 0.03]	[- 0.16, 0.02]	[- 0.23, - 0.04]	49.2

**Table 5.** DE, IE, and TE of MetS components on renal function impairment (UMALB) mediated through serum Mb (Significant). *TE* total effect, *DE* direct effects, *IE* indirect effects, *CI* confidence interval.

observed relationship between elevated serum Mb and DKD suggests that serum Mb may be a risk factor for DKD<sup>43,45</sup>. However, to our knowledge, no studies have evaluated serum Mb as a risk factor for DKD prediction, nor has serum Mb been used as a risk factor for DKD model construction, partially due to limitations in methods to address collinearity.

We first verified the importance of serum Mb in DKD. To find those features that might be overlooked, the ML algorithms were chosen because they are not constrained by collinearity when processing high-dimensional EHR data. Apart from common DKD-related features, such as renal function features and fasting glucose, we also specially developed ML models including serum Mb. Our DKD models had comparable performance to other models, achieving an AUC of 0.85 in the presence of renal function features and 0.80 after removal of renal function features. Further, in our study, feature importance analysis and SHAP have recognized serum Mb as an important feature. Equally important, RCS found that moderately elevated serum Mb was associated with the risk of DKD, which is consistent with previous findings<sup>17–20</sup>. Therefore, we believe that serum Mb is a potential risk factor for DKD and deserves more attention and further study.

**Serum Mb may mediate MetS component-induced renal function impairment.** Current DKD management focuses solely on blood pressure control, glycemic and lipid control, weight loss, and diet<sup>50–52</sup>, in other words, we emphasize the strong association between MetS components and DKD, but more efforts are needed to find appropriate treatment for DKD<sup>53,54</sup>. Recently, elevated Mb have been observed in DM and DKD. In a 200 controlled study of 188 DM patients, serum Mb levels in DM patients were elevated and were much higher in DM patients with metabolic syndrome<sup>16</sup>. In a cross-sectional study of DKD, serum Mb was found to be positively correlated with DKD in a model adjusted for age, gender, cardiovascular risk factors, glycemic control, and blood sugar<sup>45</sup>. In another study, serum Mb was found to be related to CKD, and elevated serum Mb levels were related to advanced stage of CKD<sup>43</sup>. Interestingly, the elevation of serum Mb cannot be explained by Statins

uses<sup>55</sup>. The cause of elevated serum Mb may be due to abnormal glucose and lipid metabolism and diabetic muscle tissue damage<sup>17,18</sup>. In diabetic rats, liraglutide and insulin increased carnitine palmitoyl transferase 1 (CPT1) expression, decreased Mb, and improved cardiovascular function<sup>19</sup>. In a skeletal muscle-specific (Cpt1bm<sup>-/-</sup>) mouse model, a 25% fat diet resulted in increased intramuscular lipids and serum Mb and decreased mouse vitality<sup>20</sup>.

Further, at the molecular level, Mb interacts with glucose to form AGEs<sup>33,56</sup>, which are active contributors to cellular stress and chronic inflammation<sup>34,57</sup>. Also, oxidative modification of Mb induced by stress-related protein carbonyl compounds is an important cause for the development of microvascular and macrovascular complications in diabetes<sup>21–23</sup>.

These cumulative evidence indicate that serum Mb has a mediator role in DKD. Confirming the role of serum Mb in DKD progression may provide a thorough comprehension of DKD pathogenesis and a promising therapeutic target for DKD. However, as far as we know, the causal mediation relationship between serum Mb and MetS components and renal function impairment has not been reported.

In our study, feature importance analysis, SHAP, and RCS revealed that MetS components and elevated serum Mb were closely related to DKD. In causal mediation analysis, significant IE and TE were found using either UMALB or eGFR as indicators of renal function impairment, indicating serum Mb as a mediator in MetS component-induced renal function impairment. Thus, we believe that more studies on the role of serum Mb in DKD, based on multilevel research designs and emerging strategies, will provide insights into the pathogenesis of DKD and reveal new treatment targets for DKD.

**Study Limitations.** Our findings should be further verified in cohort studies, as the strength of the evidence we generated is limited by the cross-sectional study design. External validity was difficult to perform in our study because serum Mb was not a commonly used indicator in DKD studies. Further, it is necessary to gain insights into the pathophysiological mechanisms at the cellular and molecular levels to reveal the causal relationship between MetS components, serum Mb, and renal function impairment.

### Data availability

The datasets generated during and/or analyzed during the current study are not publicly available usage due to the other ongoing projects but are available from the corresponding author on reasonable request.

Received: 18 March 2022; Accepted: 28 November 2022

Published online: 10 December 2022

### References

1. Tanabe, K., Wada, J. & Sato, Y. Targeting angiogenesis and lymphangiogenesis in kidney disease. *Nat. Rev. Nephrol.* **16**(5), 289–303. <https://doi.org/10.1038/s41581-020-0260-2> (2020).
2. Nusinovici, S. *et al.* Retinal microvascular signs and risk of diabetic kidney disease in asian and white populations. *Sci. Rep.* **11**(1), 4898. <https://doi.org/10.1038/s41598-021-84464-7> (2021).
3. Thomas, M. C. *et al.* Diabetic kidney disease. *Nat. Rev. Dis. Primers* **1**(July), 1–20. <https://doi.org/10.1038/nrdp.2015.18> (2015).
4. Alicic, R. Z., Rooney, M. T. & Tuttle, K. R. Diabetic kidney disease: Challenges, progress, and possibilities. *Clin. J. Am. Soc. Nephrol.* **12**(12), 2032–2045. <https://doi.org/10.2215/CJN.11491116> (2017).
5. Macisaac, R. J., Ekinci, E. I. & Jerums, G. Markers of and risk factors for the development and progression of diabetic kidney disease. *Am. J. kidney Dis. Off J. Natl. Kidney Found* **63**(2 Suppl 2), S39–62. <https://doi.org/10.1053/j.ajkd.2013.10.048> (2014).
6. Tziomalos, K. & Athyros, V. G. Diabetic nephropathy: New risk factors and improvements in diagnosis. *Rev. Diabet. Stud.* **12**(1–2), 110–118. <https://doi.org/10.1900/RDS.2015.12.110> (2015).
7. Colhoun, H. M. & Marcovecchio, M. L. Biomarkers of diabetic kidney disease. *Diabetologia* **61**(5), 996–1011. <https://doi.org/10.1007/s00125-018-4567-5> (2018).
8. Jiang, W. *et al.* Establishment and validation of a risk prediction model for early diabetic kidney disease based on a systematic review and meta-analysis of 20 cohorts. *Diabetes Care* **43**(4), 925–933. <https://doi.org/10.2337/dc19-1897> (2020).
9. Slieker, R. C. *et al.* Performance of prediction models for nephropathy in people with type 2 diabetes: Systematic review and external validation study. *BMJ* **374**, n2134. <https://doi.org/10.1136/bmj.n2134> (2021).
10. Cornier, M.-A. *et al.* The metabolic syndrome. *Endocr. Rev.* **29**(7), 777–822. <https://doi.org/10.1210/er.2008-0024> (2008).
11. Wu, S. H., Liu, Z. & Ho, S. C. Metabolic syndrome and all-cause mortality: A meta-analysis of prospective cohort studies. *Eur. J. Epidemiol.* **25**(6), 375–384. <https://doi.org/10.1007/s10654-010-9459-z> (2010).
12. Mottillo, S. *et al.* The metabolic syndrome and cardiovascular risk a systematic review and meta-analysis. *J. Am. Coll. Cardiol.* **56**(14), 1113–1132. <https://doi.org/10.1016/j.jacc.2010.05.034> (2010).
13. Jardine, M. J. *et al.* Prediction of kidney-related outcomes in patients with type 2 diabetes. *Am. J. kidney Dis. Off J. Natl. Kidney Found* **60**(5), 770–778. <https://doi.org/10.1053/j.ajkd.2012.04.025> (2012).
14. Berenbrink, M. Myoglobin's old and new clothes: From molecular structure to integrated function and evolution. *J. Exp. Biol.* **213**(Pt 16), 2711–2712. <https://doi.org/10.1242/jeb.048918> (2010).
15. Wittenberg, B. A. & Wittenberg, J. B. Myoglobin-mediated oxygen delivery to mitochondria of isolated cardiac myocytes. *Proc. Natl. Acad. Sci. U. S. A.* **84**(21), 7503–7507. <https://doi.org/10.1073/pnas.84.21.7503> (1987).
16. Odum, E. P. & Young, E. E. Elevated cardiac troponin I, creatine kinase and myoglobin and their relationship with cardiovascular risk factors in patients with type 2 diabetes. *Diabetes Metab. Syndr.* **12**(2), 141–145. <https://doi.org/10.1016/j.dsx.2017.12.001> (2018).
17. Miyamoto, T., Watanabe, K., Fukuda, K. & Moritani, T. Near-infrared spectroscopy of vastus lateralis muscle during incremental cycling exercise in patients with type 2 diabetes. *Phys. Ther. Res.* **23**(1), 23–30. <https://doi.org/10.1298/ptr.E9984> (2020).
18. Gildea, N. *et al.* Influence of type 2 diabetes on muscle deoxygenation during ramp incremental cycle exercise. *Respir. Physiol. Neurobiol.* **269**, 103258. <https://doi.org/10.1016/j.resp.2019.103258> (2019).
19. Huang, Q. *et al.* Incremental effect of liraglutide on traditional insulin injections in rats with type 2 diabetes mellitus by maintaining glycolipid metabolism and cardiovascular function. *Exp. Ther. Med.* **17**(3), 1863–1869. <https://doi.org/10.3892/etm.2019.7148> (2019).
20. Warfel, J. D. *et al.* A low fat diet ameliorates pathology but retains beneficial effects associated with CPT1b knockout in skeletal muscle. *PLoS ONE* **12**(12), e0188850. <https://doi.org/10.1371/journal.pone.0188850> (2017).

21. Ghelani, H., Razmovski-Naumovski, V., Pragada, R. R. & Nammi, S. Attenuation of glucose-induced myoglobin glycation and the formation of advanced glycation end products (AGEs) by (R)- $\alpha$ -lipoic acid in vitro. *Biomolecules* <https://doi.org/10.3390/biom8010009> (2018).
22. Raupbach, J., Ott, C., Koenig, J. & Grune, T. Proteasomal degradation of glycated proteins depends on substrate unfolding: Preferred degradation of moderately modified myoglobin. *Free Radic. Biol. Med.* **152**, 516–524. <https://doi.org/10.1016/j.freeradbiomed.2019.11.024> (2020).
23. You, Y., Liu, F., Du, K.-J., Wen, G.-B. & Lin, Y.-W. Structural and functional alterations of myoglobin by glucose-protein interactions. *J. Mol. Model.* **20**(7), 2358. <https://doi.org/10.1007/s00894-014-2358-6> (2014).
24. Alberti, K. G. & Zimmet, P. Z. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: Diagnosis and classification of diabetes mellitus provisional report of a WHO consultation. *Diabet. Med.* **15**(7), 539–553. [https://doi.org/10.1002/\(SICI\)1096-9136\(199807\)15:7%3c539::AID-DIA668%3e3.0.CO;2-S](https://doi.org/10.1002/(SICI)1096-9136(199807)15:7%3c539::AID-DIA668%3e3.0.CO;2-S) (1998).
25. KDOQI. KDOQI clinical practice guidelines and clinical practice recommendations for diabetes and chronic kidney disease. *Am. J. kidney Dis. Off J. Natl. Kidney Found* **49**(2 Suppl 2), S12–154. <https://doi.org/10.1053/j.ajkd.2006.12.005> (2007).
26. Mingrone, G. *et al.* Metabolic surgery versus conventional medical therapy in patients with type 2 diabetes: 10-year follow-up of an open-label, single-centre, randomised controlled trial. *Lancet* **397**(10271), 293–304. [https://doi.org/10.1016/S0140-6736\(20\)32649-0](https://doi.org/10.1016/S0140-6736(20)32649-0) (2021).
27. Rynders, C. A. *et al.* Comparing simple insulin sensitivity indices to the oral minimal model postexercise. *Med. Sci. Sports Exerc.* **48**(1), 66–72. <https://doi.org/10.1249/MSS.0000000000000728> (2016).
28. Xie, Y. *et al.* Temporal relationship between body mass index and triglyceride-glucose index and its impact on the incident of hypertension. *Nutr. Metab. Cardiovasc. Dis.* **29**(11), 1220–1229. <https://doi.org/10.1016/j.numecd.2019.07.003> (2019).
29. von Post-Skagegård, M., Vessby, B. & Karlström, B. Glucose and insulin responses in healthy women after intake of composite meals containing cod-, milk-, and soy protein. *Eur. J. Clin. Nutr.* **60**(8), 949–954. <https://doi.org/10.1038/sj.ejcn.1602404> (2006).
30. Lemieux, P. *et al.* Effects of 6-month vitamin D supplementation on insulin sensitivity and secretion: A randomised, placebo-controlled trial. *Eur. J. Endocrinol.* **181**(3), 287–299. <https://doi.org/10.1530/EJE-19-0156> (2019).
31. Monti, L. D. *et al.* Decreased diabetes risk over 9 year after 18-month oral L-arginine treatment in middle-aged subjects with impaired glucose tolerance and metabolic syndrome (extension evaluation of L-arginine study). *Eur. J. Nutr.* **57**(8), 2805–2817. <https://doi.org/10.1007/s00394-017-1548-2> (2018).
32. McKay, D. L., Eliasziw, M., Chen, C. Y. O. & Blumberg, J. B. A pecan-rich diet improves cardiometabolic risk factors in overweight and obese adults: A randomized controlled trial. *Nutrients* <https://doi.org/10.3390/nu10030339> (2018).
33. Hempe, J. M. *et al.* The hemoglobin glycation index identifies subpopulations with harms or benefits from intensive treatment in the ACCORD trial. *Diabetes Care* **38**(6), 1067–1074. <https://doi.org/10.2337/dc14-1844> (2015).
34. Block, G. *et al.* Diabetes prevention and weight loss with a fully automated behavioral intervention by email, web, and mobile phone: A randomized controlled trial among persons with prediabetes. *J. Med. Internet Res.* **17**(10), e240. <https://doi.org/10.2196/jmir.4897> (2015).
35. Breiman, L. Random forests. *Mach. Learn.* **45**(1), 5–32. <https://doi.org/10.1023/A:1010933404324> (2001).
36. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Association for Computing Machinery, 2016).
37. Liaw, A. & Wiener, M. Classification and regression by random forest. *Forest.* **23** (2001).
38. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, 2016). doi:<https://doi.org/10.1145/2939672.2939785>.
39. Eisenman, R. L. A profit-sharing interpretation of Shapley value for N-person games. *Behav. Sci.* **12**(5), 396–398. <https://doi.org/10.1002/bs.3830120506> (1967).
40. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30**(2), 4766–4775 (2017).
41. Lundberg, S.M., Lee, S.-I. Consistent feature attribution for tree ensembles. <http://arxiv.org/abs/1706.06060> (2017).
42. Richiardi, L., Bellocco, R. & Zugna, D. Mediation analysis in epidemiology: Methods, interpretation and bias. *Int. J. Epidemiol.* **42**(5), 1511–1519. <https://doi.org/10.1093/ije/dyt127> (2013).
43. Lenglet, A. *et al.* Prognostic implications of plasma myoglobin levels in patients with chronic kidney disease. *Int. J. Artif. Organs* **35**(11), 959–968. <https://doi.org/10.5301/ijao.5000124> (2012).
44. Afghahi, H. *et al.* Risk factors for the development of albuminuria and renal impairment in type 2 diabetes—the Swedish National Diabetes Register (NDR). *Nephrol. Dial. Transplant. Off Publ. Eur. Dial. Transpl. Assoc.—Eur. Ren. Assoc.* **26**(4), 1236–1243. <https://doi.org/10.1093/ndt/gfq535> (2011).
45. Carlsson, A. C. *et al.* Growth differentiation factor 15 (GDF-15) is a potential biomarker of both diabetic kidney disease and future cardiovascular events in cohorts of individuals with type 2 diabetes: A proteomics approach. *Ups. J. Med. Sci.* **125**(1), 37–43. <https://doi.org/10.1080/03009734.2019.1696430> (2020).
46. Lemley, K. V. Machine learning comes to nephrology. *J. Am. Soc. Nephrol.* **30**(10), 1780–1781. <https://doi.org/10.1681/ASN.2019070664> (2019).
47. Dimopoulos, A. C. *et al.* Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk. *BMC Med. Res. Methodol.* **18**(1), 179. <https://doi.org/10.1186/s12874-018-0644-1> (2018).
48. Cowie, M. R. *et al.* Electronic health records to facilitate clinical research. *Clin. Res. Cardiol.* **106**(1), 1–9. <https://doi.org/10.1007/s00392-016-1025-6> (2017).
49. Elley, C. R. *et al.* Derivation and validation of a renal risk score for people with type 2 diabetes. *Diabetes Care* **36**(10), 3113–3120. <https://doi.org/10.2337/dc13-0190> (2013).
50. Liew, A. *et al.* Asian pacific society of nephrology clinical practice guideline on diabetic kidney disease. *Nephrology* **25**(Suppl 2), 12–45. <https://doi.org/10.1111/nep.13785> (2020).
51. Doshi, S. M. & Friedman, A. N. Diagnosis and management of type 2 diabetic kidney disease. *Clin. J. Am. Soc. Nephrol.* **12**(8), 1366–1373. <https://doi.org/10.2215/CJN.11111016> (2017).
52. Barrett, E. J. *et al.* Diabetic microvascular disease: An endocrine society scientific statement. *J. Clin. Endocrinol. Metab.* **102**(12), 4343–4410. <https://doi.org/10.1210/jc.2017-01922> (2017).
53. Stanton, R. C. Clinical challenges in diagnosis and management of diabetic kidney disease. *Am. J. kidney Dis. Off J. Natl. Kidney Found* **63**(2 Suppl 2), S3–21. <https://doi.org/10.1053/j.ajkd.2013.10.050> (2014).
54. Lytvyn, Y., Bjornstad, P., van Raalte, D. H., Heerspink, H. L. & Cherney, D. Z. I. The new biology of diabetic kidney disease—mechanisms and therapeutic implications. *Endocr. Rev.* **41**(2), 202–231. <https://doi.org/10.1210/endo/bnz010> (2020).
55. Yasuda, G. *et al.* Safety and efficacy of fluvastatin in hyperlipidemic patients with chronic renal disease. *Ren. Fail.* **26**(4), 411–418. <https://doi.org/10.1081/jdi-120039826> (2004).
56. Banerjee, S., Maity, S. & Chakraborti, A. S. Methylglyoxal-induced modification causes aggregation of myoglobin. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **155**, 1–10. <https://doi.org/10.1016/j.saa.2015.10.022> (2016).
57. Goh, S.-Y. & Cooper, M. E. Clinical review: The role of advanced glycation end products in progression and complications of diabetes. *J. Clin. Endocrinol. Metab.* **93**(4), 1143–1152. <https://doi.org/10.1210/jc.2007-1817> (2008).

## Acknowledgements

We would like to thank the Information Network Center of the Third Xiangya Hospital for providing the clinical data.

## Author contributions

R.W., Z.S., H.L., D.C., F.Z., S.Z., S.C. conceived and designed the study. R.W., Z.S., F.Z., S.Z., S.C., Y.H., H.X. coordinated data collection. R.W., F.Z., Z.S., S.Z., L.F. conducted the analyses. R.W., Z.S., F.Z., H.L., L.F. wrote the manuscript. All co-authors critically revised the manuscript.

## Funding

This work was supported by the National Natural Science Foundation of China (grant no. 81870352, 81970252), the Key Research and Development Project of Hunan Province (grant no.2020SK2087, 2019SK2041), the Undergraduate Training Program for Innovation and Entrepreneurship of Hunan Province (S2020105330657).

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-022-25299-8>.

**Correspondence** and requests for materials should be addressed to H.L.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

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



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

© The Author(s) 2022