nature portfolio

Peer Review File

AI-based multiomics profiling reveals complementary omics contributions to personalized prediction of cardiovascular disease

Corresponding Author: Professor Qingpeng Zhang

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors introduce a deep learning framework with a multitask objective to predict risks of nine common cardiovascular diseases. The model uses multilayer perceptron (MLP) based architecture and predicts binary outcome of each individual disease. It takes proteomic and metabolic biomarkers of 24,308 participants as inputs and is trained and evaluated through a 10-fold cross validation schema. By evaluations including risk stratification, Harrel's C-index, calibration plot and net benefit curve, the authors show that multiomics signatures can not only be applied as individual predictors, but enhance CVD risk predictions when combined with other clinical predictors.

Major comments:
1. The proposed deep learning model is trained using 24,308 total participants data on 9 cardiovascular diseases, where the number of all cardiovascular outcomes accounts for 4,213 and some diseases have very limited data points (e.g. 123 for AAA). It is very likely that training with such imbalanced data will lead to a biased model. The limited sample size for selected outcomes also raises concerns about the potential risk of overfitting especially if high embedding dimensions are applied.

2. Although the authors showed the superior performance of their deep learning derived multiomics scores in comparison with combinations of other predictors, additional analysis is needed to prove the necessity of using neural networks to process multiomic data since simpler machine learning models (e.g. XGBoost, random forest) can also introduce non-linearity of input features. The authors may consider building simpler machine learning models to generate alternative multiomic scores and compare with the deep learning derived scores.

3. Some details of data composition and processing need to be added to verify the validity of model development and improve the clarity of the manuscript. The composition of the study population needs to be clarified. For example, the reported total number of CVD outcomes does not match the sum of outcomes of all individual CVD. Additional clarification is needed to explain the proportion of individuals with multiple CVDs and how the incidence rates are calculated. In addition, the authors didn't mention how dataset splitting was performed in training and testing and how individuals with multiple CVDs are handled in the model training and evaluation. If such cases were treated as independent data points, a single participant with multiple CVDs may appear in both training and testing sets, resulting in data leakage and inflation of model performance.

Minor comments:
1. Additional details of the deep learning model architecture and loss function need to be added. For example, does the model output a unique score for each CVD (9 scores in total)? How is the overall loss calculated? Also, the authors should include their final choice of model hyperparameters to improve reproducibility of the model.
2. In Figure 3, it would be clearer if all subplots use the same x-axis range in each sub-figure.
3. There are some redundant texts in the manuscript. For example, the authors currently repeatedly explain meanings of all CVD abbreviations and population breakdown in all figure legends. Instead, it may be better if the authors use a separate section (e.g. in Methods) to explain their evaluation setup.
4. Some figures need to be adjusted. For example, the texts in Figure 5 c & d are too small to read. Supplementary figures

13 & 14 are not completely displayed.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The authors undertake Genomic, Transcriptomic, and Proteomic profiling of a large cohort of patients from the UK biobank and through AI deep learning models have developed genomic, transcriptomic, and proteomic risk scores for the prediction of 9 cardiovascular disease outcomes. This study reports that such these multiomic risk scores are superior to existing risk scores based on clinical and biochemical data both individually and in combination. This is a great pilot dataset, but it needs true external validation in an independent cohort, in order to be mature for presentation as definitive.

Major Comments
- The authors helpfully provide a list of proteins and metabolites whose levels increase and decrease the risk of cardiovascular disease. However with NTpro-BNP coming out as the top hit in all 9 cardiovascular diseases is there risk that these results have been confounded? High NTProBNP contributes towards increased CAD risk for example, but does this merely represent the interaction or comorbidity of different types of CVD such as CAD and HF? On closer inspection, the directionality and effect sizes of the proteins are quite similar visually between CVD conditions – are the authors therefore suggesting that the multiomic signatures of 9 different cardiovascular diseases are largely the same?
- Similarly, LDL metabolite fractions and total lipids / cholesterol are not strongly predictive of CAD (which is unusual in a healthy cohort, most of whom are not taking statins). Does this trend persist when the 20% of the cohort on lipid-lowering therapies are removed from the cohort? It would be helpful if the authors could comment on this observation and why this might be the case.
- The authors report that “the genetic risk of CVDeath was measured using the PRS for CAD since there is no available GWAS summary statistics for CVDeath”. Is there any reason the author’s couldn’t run their own CVDeath GWAS rather than applying the CAD GWAS results to a non-identical outcome.
- The authors describe the demographics of the whole cohort but not the demographics of the individual training and validation cohorts. Could the authors please provide these along with statistical testing to show that such cohorts are not statistically different. This is especially important given the absence of an external validation cohort.
- On trying to access the online demo of the model, the weblink appeared to be broken. Please could the authors provide a working link to allow reviewers and readers to access the online demo.
-The main issue with this work is that is used 10% of the cohort for validation, so this is considered internal (not external) validation, and the models are expected to suffer from overfitting.

Minor Comments
- In the demographics it would also be helpful to know the numbers of participants on statin / antihypertensive therapies
- Minor grammatical error in supplementary figure 1 caption, lines 115-116. “Here, we defined the follow-up duration as the period from the date of baseline assessment to the date of incident any cardiovascular diseases…” doesn’t make sense grammatically. Perhaps the intended line is “the date of any incident cardiovascular diseases”
- Minor grammatical error in 2.3, paragraph 2, line 2: “(i.e., 95% CIs of delta C-index not included zero)”. “Included” should be changed to “including”
- Supplementary figure 10 appears to have a light blue figure legend marker for ‘clinical scores’ which seems to be redundant as it does not feature in the figure outside the legend.

(Remarks on code availability)
m/a

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have addressed all our concerns. We have no more comments.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The authors have made some reasonable changes to the manuscript generally. The inclusion of a geographically separate

cohort, whilst an improvement on the original study design, still does not offer quite the same level of robust validation as a true external validation cohort. On balance, I feel that the manuscript has sufficient merrit to warrant publication of the existing material with the following further revisions:

The manuscript states, under the limitations section, 'Finally, since the majority of UKB population are European ancestry and white ethnicity, the generalizability of our findings needs to be validated in ethnically diverse populations to ensure broader applicability. Please would the authors kindly submit a revision that mentions explicitly the lack of a proper external validation cohort (rather than just citing a need for ethnic diverse populations).

The authors discuss the clinical utility and application of the model in their discussion section. As some point in this discussion it should be clearly stated that the signatures are not clinically generalisable or applicable until they have undergone external validation.

With these revisions I would be satisfied to recommend publication.

(Remarks on code availability)

Dear Editor and Reviewers,

Thank you so much for reviewing our paper and the opportunity to revise our paper! We really appreciate the helpful comments and constructive suggestions! We have substantially revised the manuscript according to the review comments. All revised portions are marked in red in the track-change version of the manuscript.

Specific comments and our responses are detailed below.

REVIEWER 1'S COMMENTS

Remarks to the Author:

COMMENT 1-1. *The proposed deep learning model is trained using 24,308 total participants data on 9 cardiovascular diseases, where the number of all cardiovascular outcomes accounts for 4,213 and some diseases have very limited data points (e.g. 123 for AAA). It is very likely that training with such imbalanced data will lead to a biased model. The limited sample size for selected outcomes also raises concerns about the potential risk of overfitting especially if high embedding dimensions are applied.*

RESPONSE 1-1. Thank you very much for the insightful comments. We agree with the reviewer that class imbalance and potential overfitting are challenges when developing clinical prediction models. To address these concerns, we significantly revamped the study by incorporating more data and introducing sample- and task-level weighting in the loss function of our CardiOmicScore framework.

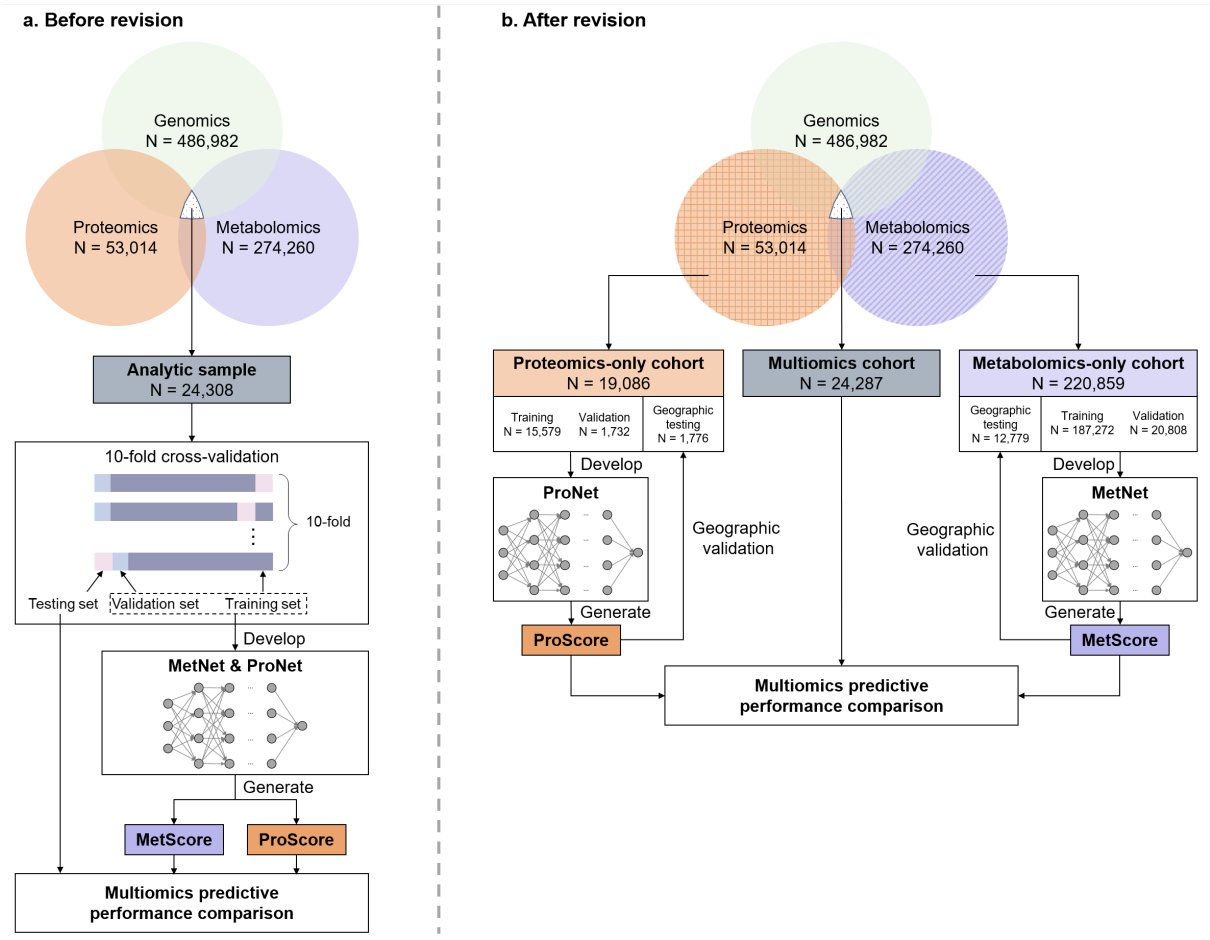


Figure R1: Comparison of study framework before and after revision.

Specifically, at the data level, we revised our data partitioning strategy (**Figure R1**) to leverage a much larger portion of the UK Biobank data for model development (i.e., training and validation sets in the Metabolomics-only and Proteomics-only cohorts), followed by a geographic validation on their respective testing sets, while keeping the Multiomics cohort as a completely non-overlapping dataset to assess the added predictive value of integrating different omics data (**Figure 1** and **Supplementary Figure 1**). For our two deep learning models within our CardiOmicScore framework, MetNet and ProNet, the development datasets (training + validation sets) now include 208,080 and 17,311 participants, respectively. This has substantially increased the number of incident cardiovascular diseases (CVDs) for model training.

Despite this increase, the proportion of incident ventricular arrhythmias and abdominal aortic aneurysm remains very low (e.g., 0.8% and 0.4% in our training cohort, respectively). Training reliable models for such rare events is indeed challenging. Therefore, we removed

these two diseases and cardiovascular death from our study to focus on the six most common CVDs: coronary artery disease (CAD), stroke, heart failure (HF), atrial fibrillation (AF), peripheral artery disease (PAD), and venous thromboembolism (VTE).

At the modeling level, we implemented two technical improvements in our CardiOmicScore framework to address the remaining class imbalance among the six CVDs:

- Sample-level weighting: For each disease prediction task, we calculated a pos_weight parameter as the ratio of negative samples to positive samples. This parameter was used in the loss function to apply a heavier penalty for misclassifying positive cases (i.e., individuals who develop a disease), forcing the model to pay closer attention to these rare events.
- Task-level weighting: We assigned a unique weight to each of the six disease prediction tasks, where the weight was equal to the inverse of the number of positive samples for that disease. This strategy ensures that tasks with fewer positive cases (e.g., PAD) contribute more proportionally to the overall loss function, preventing the model from being dominated by the most common diseases.

To provide strong evidence against overfitting and to rigorously assess the generalizability of our models, we implemented a multi-stage evaluation process that goes beyond standard internal validation (**Figure 1**). First, we evaluated the performance of MetScore and ProScore generated by our CardiOmicScore framework on their respective geographic testing sets (N = 12,779 for MetNet and N = 1,775 for ProNet), following best practices from recent large-scale UKB studies (Argentieri et al., 2024; Zhang et al., 2024; Julkunen & Rousu, 2025). Subsequently, the Multiomics cohort was used to evaluate the incremental benefit of multi-omics integration. We found that despite significant differences in several clinical characteristics between the training and geographic testing sets, both MetScore and ProScore demonstrated strong predictive performance in the geographic testing sets, confirming their generalizability across distinct UK populations (**Supplementary Figure 6**). Furthermore, this superior performance was replicated in the Multiomics cohort. This two-stage strategy provides a robust assessment of model performance and external validity.

The revised parts are copied below:
In the main text:

Introduction Section, **Fifth** Paragraph: "These CVDs include coronary artery disease (CAD), stroke, heart failure (HF), atrial fibrillation (AF), peripheral artery disease (PAD), and venous thromboembolism (VTE)—conditions with the highest global burden[1]."

Results Section, **Study population** Part: "A two-stage, individual-level data partitioning strategy was employed to rigorously evaluate our model's performance and generalizability (see sub-section 4.2 and Supplementary Figure 1). First, we excluded participants who had neither metabolomics nor proteomics data available. Then, we divided the remaining participants into the Metabolomics-only (N = 220,859; those with metabolomics but not proteomics data), Proteomics-only (N = 19,086; those with proteomics but not metabolomics data), and Multiomics (N = 24,287; those with genomics, metabolomics, and proteomics data) cohorts based on omics availability. The Metabolomics-only and Proteomics-only cohorts were used within our CardiOmicScore framework to develop two deep learning models, MetNet and ProNet, respectively. The Multiomics cohort was reserved as an untouched validation set to assess the added predictive value of integrating different omics data.

Following best practices in recent UKB studies[39–41], each of these development cohorts was further split into training/validation sets (England and Wales) and a geographic testing set (Scotland). Specifically, the Metabolomics-only cohort included 187,272 participants in the training set, 20,808 in the validation set, and 12,779 in the geographic testing set, while the Proteomics-only cohort comprised 15,579, 1,732, and 1,775 participants in the training, validation, and geographic testing sets, respectively. Baseline characteristics were comparable between training and validation sets in both cohorts, whereas significant differences were observed between training and geographic testing sets (Supplementary Tables 2–3), thus allowing a robust assessment of the models' regional generalizability.

Baseline characteristics and incident cases during follow-up were broadly consistent across all study datasets. The median age of participants at baseline was between 56.0 and 58.0 years. The proportion of males ranged from 42.8% to 45.8%. The prevalence of baseline medication use was comparable, with 15.8% to 18.1% of participants receiving lipid-lowering medication and 10.5% to 11.2% on antihypertensive medication. The distribution of outcomes was also similar, with 83.2–85.7% of participants remaining free of any incident CVD, 10.6–11.8% developing one CVD, and 3.7–5.0% developing multiple CVDs. Detailed

baseline characteristics and follow-up information for each cohort are provided in Supplementary Tables 1–3 and Supplementary Figure 2."

Results Section, **Advancing cardiovascular risk prediction with the power of omics information** Part: "Among the three omics scores, ProScore showed the highest performance for all CVDs with the C-index ranging from 0.69 (95% CI: 0.67–0.71) for VTE to 0.82 (95% CI: 0.80–0.84) for PAD (Figure 3a). The C-index of MetScore spanned from 0.64 (95% CI: 0.61–0.66) for VTE to 0.74 (95% CI: 0.71–0.76) for PAD. Importantly, the strong performance of ProScore and MetScore was replicated in their respective geographic testing sets, underscoring their generalizability across distinct UK populations (Supplementary Figure 6). In contrast, PRS provided the most limited predictive capacity (C-index range: 0.52–0.60). Finally, the discriminative performance increased with more clinical predictors included in the CPH models for all diseases (Figure 3a)."

Methods Section, **Data partition and imputation** Part: "To rigorously evaluate model performance and ensure generalizability, we adopted a two-stage individual-level data partition strategy, designed to approximate external validation while fully leveraging UKB data (Supplementary Figure 1). First, we excluded participants without both metabolomics and proteomics data. The remaining participants were then split into three cohorts based on omics data availability: a Metabolomics-only cohort (N = 220,859; those with metabolomics but not proteomics data), a Proteomics-only cohort (N = 19,086; those with proteomics but not metabolomics data), and a Multiomics cohort (N = 24,287; those with genomics, metabolomics, and proteomics data). In our CardiOmicScore framework, we used the Metabolomics-only and Proteomics-only cohorts to develop the respective deep learning models (MetNet and ProNet) and to generate the corresponding omics scores (MetScore and ProScore). To enable geographic validation, we further split each of these two cohorts into a training/validation set (England and Wales) and a geographic testing set (Scotland) according to the recruitment regions, following approaches commonly used in recent UKB studies[39–41]. The Multiomics cohort was kept untouched throughout model training and was subsequently used as an additional validation cohort to assess the incremental benefit of integrating multiple omics data types for risk prediction.

Categorical variables were one-hot encoded and continuous variables were standardized by the mean and standard deviation. We used the K-nearest neighbors algorithm (scikit-learn

v1.3.2 package)[65], setting the number of neighbors to five, to impute missing values for continuous variables. Categorical variables were imputed with mode. For metabolomics or proteomics data, imputation and standardization of continuous variables were performed using parameters derived exclusively from the training set. The fitted preprocessing models were then applied to the corresponding validation and geographic testing sets, as well as the Multiomics cohort. To derive imputation models and standardization parameters for subsequent application to the Multiomics cohort, we constructed two cohorts from all available individuals with relevant data: one for clinical information (N = 427,225) and one for genomics (N = 412,797)."

Methods Section, **Ascertainment of cardiovascular diseases** Part: "The six CVDs analyzed in this study included coronary artery disease (CAD), stroke, heart failure (HF), atrial fibrillation (AF), peripheral artery disease (PAD), and venous thromboembolism (VTE)."

Methods Section, **Deep learning models** Part: "We developed two multitask deep neural networks, named MetNet and ProNet, to derive the MetScore and ProScore for six CVDs, with 168 metabolites and 2,920 proteins as the input, respectively. MetNet was trained on the training set from the Metabolomics-only cohort, and ProNet on the training set from the Proteomics-only cohort, while their respective validation sets were used to monitor model fitting and select the optimal checkpoints. The final model was first evaluated on the geographic testing set to assess generalizability across regions. Subsequently, the Multiomics cohort, which remained untouched throughout model development, was used for a further performance evaluation."

Methods Section, **Deep learning models** Part: "To address the class imbalance in multitask prediction, we applied both sample-level and task-level weighting (Supplementary Methods). At the sample level, the pos_weight parameter was calculated as the ratio of the number of negative samples to positive samples for each task, increasing the penalty for misclassifying positive cases and encouraging the model to pay more attention to rare events. At the task level, each disease was assigned a weight equal to the inverse of its number of positive samples, ensuring that tasks with fewer positive cases contribute proportionally more to the overall loss. The final loss was computed as the mean of the weighted losses across all tasks."

In the Supplementary Materials:

Supplementary Methods Section, **Sample-level and task-level weighting** Part: "In our multi-task learning framework, the loss function incorporates both sample-level weighting and task-level weighting to account for variations in data distribution and task importance.

At the sample level, we applied the positive class weight used in PyTorch's BCEWithLogitsLoss:

$$pos\_weight_t = \frac{N_t - P_t}{P_t}$$

where N_t and P_t denote the number of samples and positive samples in task t $(t = 1, \ldots, T)$, respectively. The weighted binary cross-entropy loss for task t is:

$$\mathcal{L}_t^{sample} = -\frac{1}{N_t}\sum_{t=1}^{N_t}\left[pos\_weight_t \cdot y_i^{(t)} \cdot \log\sigma(x_i^{(t)}) + (1 - y_i^{(t)}) \cdot \log\left(1 - \sigma(x_i^{(t)})\right)\right]$$

where $x_i^{(t)}$ is the model's predicted logit of sample i for task t; $y_i^{(t)} \in \{0,1\}$ is the binary label of sample i in task t; $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function. This increases the penalty for misclassifying positive samples, encouraging the model to focus on rare events.

At the task level, we assigned each task t a weight:

$$\omega_t = \frac{1/P_t}{\sum_{k=1}^{T}(1/P_k)}$$

where T is the number of tasks. Tasks with fewer positive samples thus contribute proportionally more to the total loss, preventing them from being dominated by more prevalent diseases. For the ProNet, the task-level weights are 0.06 for coronary artery disease (CAD), 0.19 for stroke, 0.14 for heart failure (HF), 0.07 for atrial fibrillation (AF), 0.38 for peripheral artery disease (PAD), and 0.15 for venous thromboembolism (VTE). The MetNet model uses a comparable weighting scheme with values of 0.06 for CAD, 0.20 for stroke, 0.15 for HF, 0.07 for AF, 0.36 for PAD, and 0.16 for VTE.

The total weighted loss is:

$$\mathcal{L}_{total} = \sum_{t=1}^{T}\omega_t \cdot \mathcal{L}_t^{sample}$$."

The updated figures are copied below:

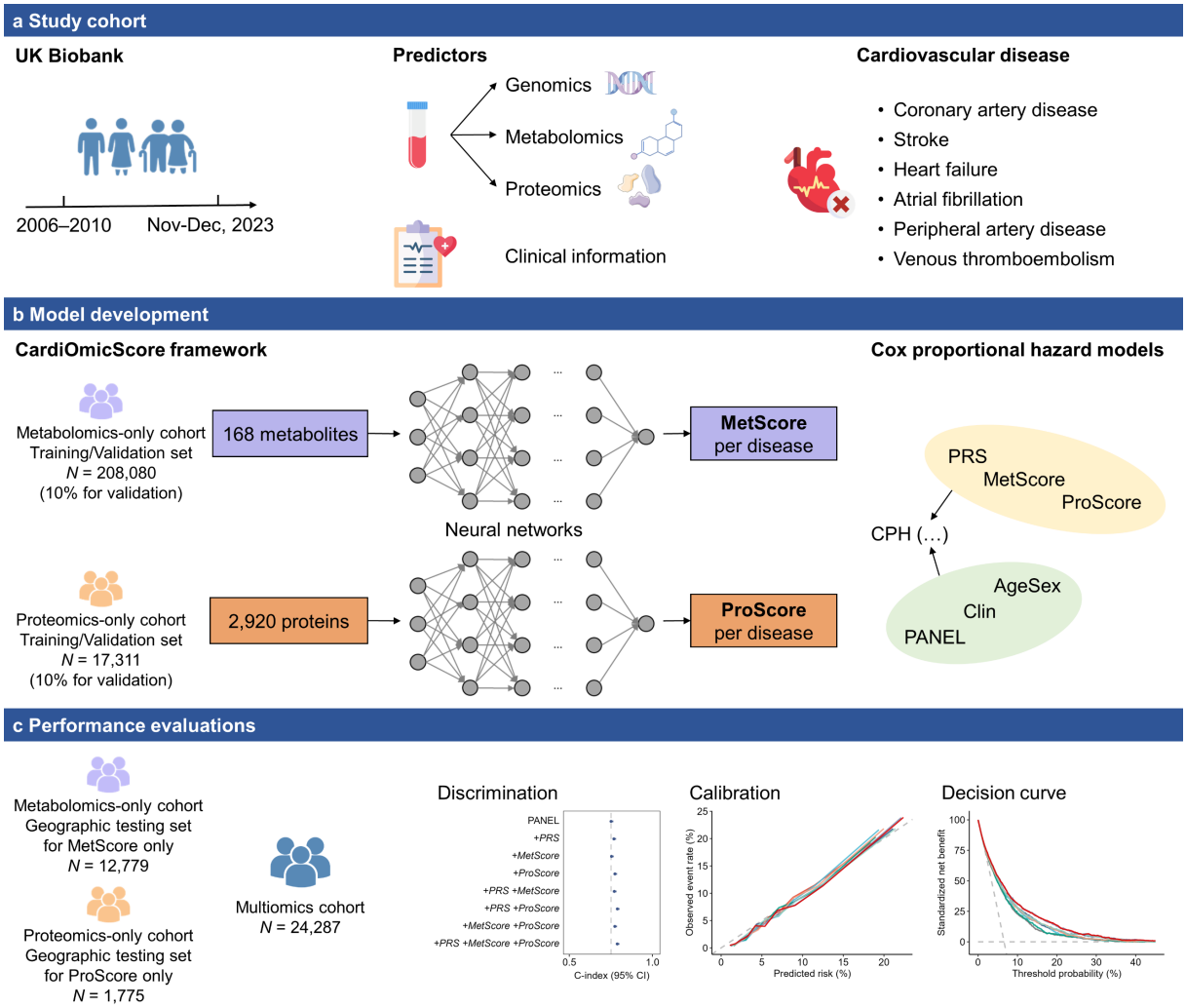
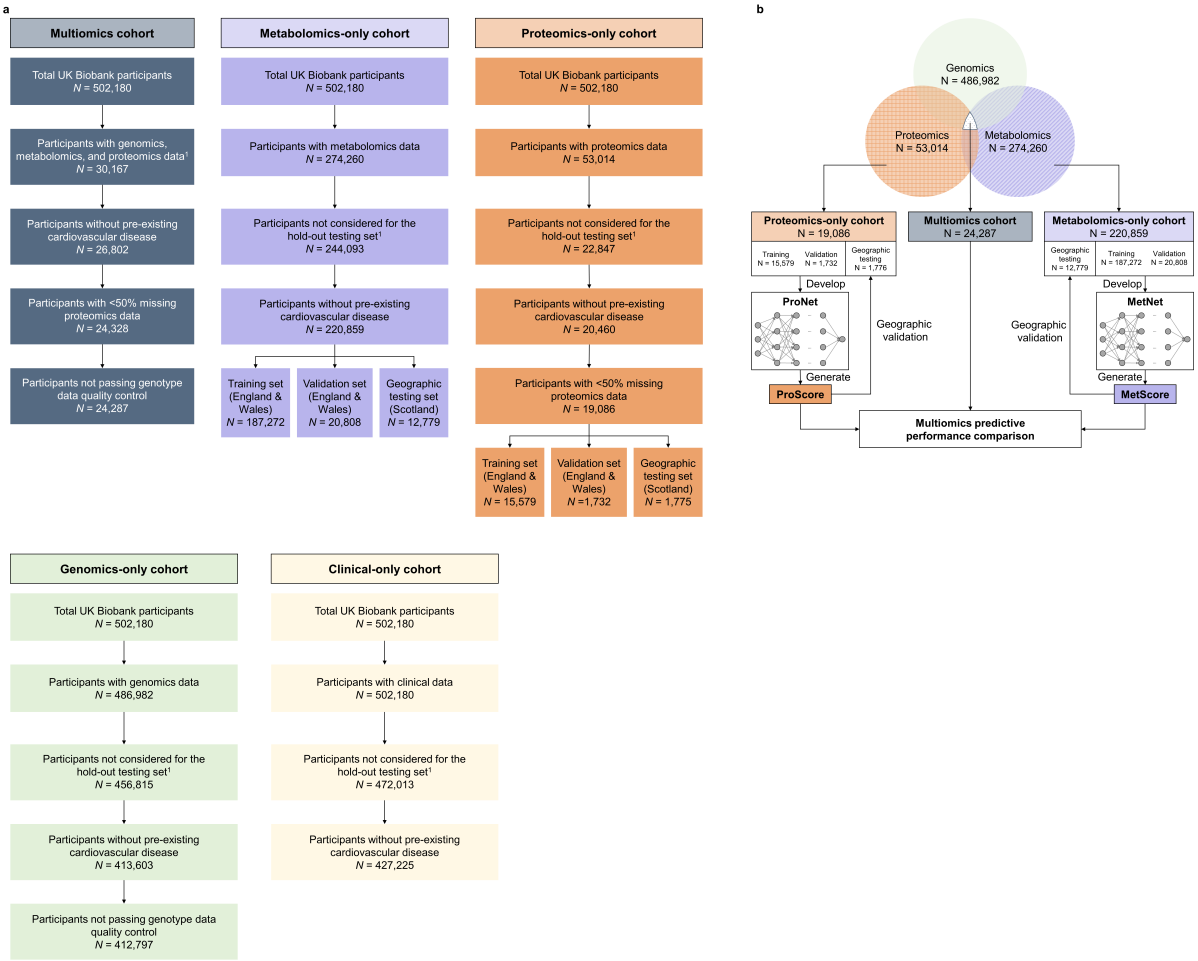
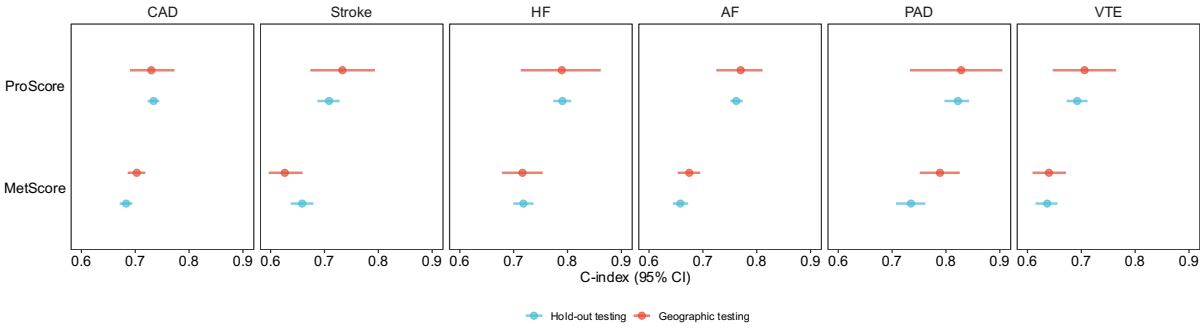


Figure 1: Study design. **a.** We extracted data from the UK Biobank, including clinical, genomic, metabolomic, and proteomic predictors, as well as six cardiovascular diseases (CVDs) defined by self-reported diagnoses, hospital episode statistics, and death records. **b.** A two-stage data partition strategy was applied: the Multiomics cohort, which remained untouched throughout model training and was used as an independent validation set to evaluate the added value of integrating multiple omics data types, and separate training, validation, and geographic testing sets within the Metabolomics-only and Proteomics-only cohorts for model development. Our CardiOmicScore framework utilized two multitask deep neural networks to predict the risk of six CVDs from 168 metabolites and 2,920 proteins separately, generating MetScore and ProScore for each CVD. We trained Cox proportional hazard (CPH) models on various combinations of polygenic risk score (PRS), MetScore, ProScore, and three predefined clinical predictor sets (i.e., AgeSex, Clin, and PANEL). **c.** Model performance was first evaluated for MetScore and ProScore in their respective geographic testing sets, and subsequently for all feature combinations in the Multiomics cohort. Performance metrics included Harrell's C-index, calibration plots, and net benefit curves, with confidence intervals estimated using 1,000 bootstrap resamples.



Supplementary Figure 1: Two-stage individual-level data partitioning strategy. **a.** Sample exclusion flowchart. Genotype data quality control excluded participants with discordant self-reported and genetic sex, sex chromosome aneuploidy, and individuals having more than ten third-degree genetic relatives. **b.** Study framework.
[1]"Participants not considered for the Multiomics cohort" in the Metabolomics-only, Proteomics-only, Genomics-only, and Clinical-only cohorts refer to those with available metabolomics, proteomics, genomics, or clinical data, respectively, after excluding the 30,167 individuals with all three omics data, who were eligible for inclusion in the Multiomics cohort.



Supplementary Figure 6: Discriminative performance of MetScore and ProScore. Discriminative performance was evaluated separately for MetScore and ProScore in both the geographic testing and the Multiomics cohort. The Multiomics cohort included 24,287

participants, while the geographic testing sets included 12,779 for MetScore and 1,775 for ProScore. Performance was assessed using Harrell's C-index with 95% confidence intervals estimated by 1,000 bootstrap resamples.

References:

Argentieri, M. A., Xiao, S., Bennett, D., Winchester, L., Nevado-Holgado, A. J., Ghose, U., Albukhari, A., Yao, P., Mazidi, M., Lv, J., Millwood, I., Fry, H., Rodosthenous, R. S., Partanen, J., Zheng, Z., Kurki, M., Daly, M. J., Palotie, A., Adams, C. J., … van Duijn, C. M. (2024). Proteomic aging clock predicts mortality and risk of common age-related diseases in diverse populations. *Nature Medicine*, 1–11. https://doi.org/10.1038/s41591-024-03164-7

Julkunen, H., & Rousu, J. (2025). Comprehensive interaction modeling with machine learning improves prediction of disease risk in the UK Biobank. *Nature Communications*, *16*(1), 6620. https://doi.org/10.1038/s41467-025-61891-y

Zhang, S., Wang, Z., Wang, Y., Zhu, Y., Zhou, Q., Jian, X., Zhao, G., Qiu, J., Xia, K., Tang, B., Mutz, J., Li, J., & Li, B. (2024). A metabolomic profile of biological aging in 250,341 individuals from the UK Biobank. *Nature Communications*, *15*(1), 8081. https://doi.org/10.1038/s41467-024-52310-9

COMMENT 1-2. *Although the authors showed the superior performance of their deep learning derived multiomics scores in comparison with combinations of other predictors, additional analysis is needed to prove the necessity of using neural networks to process multiomic data since simpler machine learning models (e.g. XGBoost, random forest) can also introduce non-linearity of input features. The authors may consider building simpler machine learning models to generate alternative multiomic scores and compare with the deep learning derived scores.*

RESPONSE 1-2. Thank you so much for the valuable suggestion. To benchmark our CardiOmicScore framework, we conducted a sensitivity analysis comparing it to several machine learning algorithms, including XGBoost, LightGBM, random forest, and logistic regression. We followed an identical pipeline to develop alternative MetScore and ProScore versions for a fair comparison. The results show that the omics scores generated by the CardiOmicScore framework consistently outperformed those derived from the alternative algorithms (**Supplementary Figure 17**). This finding suggests that while traditional machine

learning models can capture non-linear features, the CardiOmicScore framework's deep learning architecture, as a model with high representational capacity, was more effective at extracting predictive signals from the high-dimensional metabolomic and proteomic data. We copied the modifications in the revised manuscript below:

In the main text:

Results Section, **Sensitivity analyses** Part: "Moreover, our CardiOmicScore framework proved superior to conventional machine learning methods for generating omics scores. Specifically, MetScore and ProScore derived from MetNet and ProNet consistently outperformed analogous scores developed using XGBoost, LightGBM, random forest, and logistic regression (Supplementary Figure 17)."

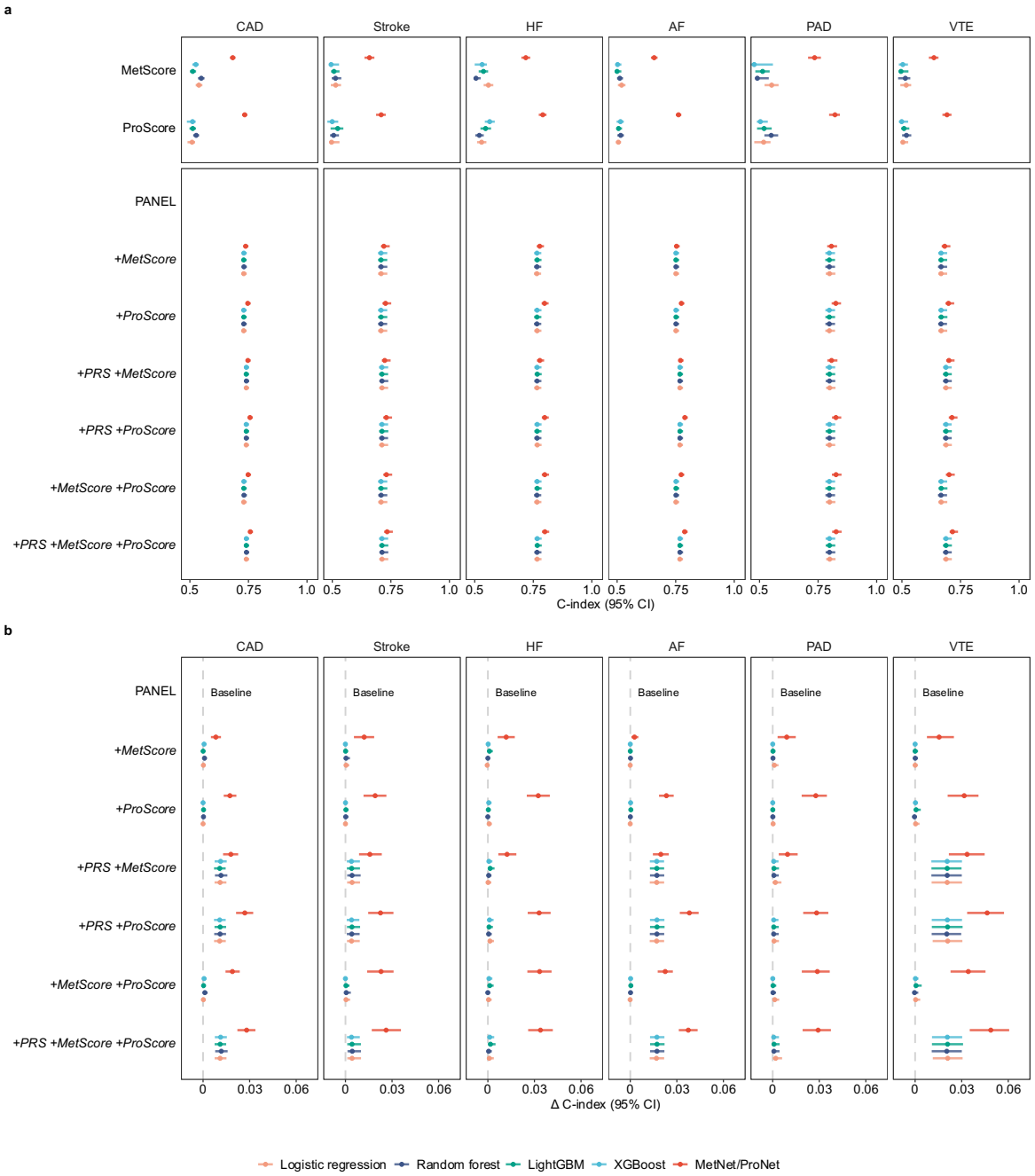
Discussion Section, **AI-identified biomarkers offer biological insights** Part: "Moreover, our CardiOmicScore framework outperformed conventional machine learning algorithms, underscoring the unique strength of the multitask deep learning architecture to model complex, non-linear interactions in high-dimensional omics data."

Methods Section, **Statistical analyses** Part: "Fifth, to demonstrate the advantage of our CardiOmicScore framework for high-dimensional omics data, we developed alternative MetScore and ProScore using several statistical machine learning algorithms, including Extreme Gradient Boosting (XGBoost, xgboost v2.0.3 package), LightGBM (lightgbm v4.4.0 package), random forest (scikit-learn v1.3.2 package), and logistic regression (scikit-learn v1.3.2 package), following an identical development pipeline (Supplementary Methods and Supplementary Table 16). We then compared the predictive utility of these scores against those generated by MetNet and ProNet."

In the Supplementary Materials:

Supplementary Methods Section, **Development of machine learning models** Part: "We developed alternative MetScore and ProScore using four widely-used machine learning algorithms: Extreme Gradient Boosting (XGBoost), LightGBM, random forest, and logistic regression. We trained a separate, independent model for each of the six cardiovascular diseases (CVDs) using all individual metabolomic or proteomic biomarkers as input features. Data splitting and preprocessing followed the same procedures as in the main analysis. For each of the 24 models (4 algorithms × 6 CVDs), we conducted a separate hyperparameter

optimization using a random search of 50 combinations. The optimal hyperparameter set was selected based on the combination that achieved the highest area under the receiver operating characteristic curve (AUC) on the validation set. Key hyperparameters were tuned for each model (Supplementary Table 15). For XGBoost and LightGBM, this included the number of estimators, learning rate, maximum tree depth, and subsampling ratios, while for random forest, we tuned the number of estimators, maximum tree depth, maximum features per split, and minimum samples for splits and leaves. For logistic regression, we tuned the inverse of regularization strength. To address class imbalance, we employed a sample weighting strategy. For XGBoost and LightGBM, we used the scale_pos_weight parameter; specifically, this was calculated as the ratio of the number of negative cases to positive cases in the training data. For random forest and logistic regression, the class_weight='balanced' setting was used. After identifying the optimal hyperparameters, a final model was retrained on the complete training dataset. To prevent overfitting in XGBoost and LightGBM, early stopping was applied during the tuning phase, halting training if the validation AUC did not improve for 20 consecutive rounds. The resulting risk predictions from these models constituted the alternative MetScores and ProScores used for comparison.”



Supplementary Figure 17: Performance comparison of CardiOmicScore framework against other statistical machine learning models. **a.** Discriminative performance trained on various predictor sets. MetScore and ProScore were estimated using different algorithms, including logistic regression, random forest, LightGBM, XGBoost, and MetNet/ProNet. Discriminative performance was evaluated by the Harrell's C-index. **b.** Differences in discriminative performance by adding omics information to clinical predictor sets. Data are presented as point estimates with 95% confidence intervals estimated from 1,000 bootstrap resamples.

Supplementary Table 15: Hyperparameters for machine learning models.

Parameters	Values
XGBoost	
Maximum depth	3, 4, 5, 6, 7, 8
Minimum child weight	1, 3, 5, 7
Percentage of subsample per tree	0.6, 0.7, 0.8, 0.9, 1.0
Percentage of features per tree	0.6, 0.7, 0.8, 0.9, 1.0
Gamma	0, 0.1, 0.2, 0.5, 1.0
Learning rate	0.01, 0.05, 0.1, 0.2
Number of boosting rounds	1000
LightGBM	
Maximum depth	3, 5, 7, -1
Minimum child samples	20, 30, 50
Percentage of subsample per tree	0.6, 0.8, 1.0
Percentage of features per tree	0.6, 0.8, 1.0
Number of leaves	15, 31, 63
Learning rate	0.01, 0.05, 0.1
Number of boosting rounds	1000
Random forest	
Maximum depth	5, 7, 10, None
Maximum features per split	Square root, Logarithm (base 2)
Minimum samples for splits	2, 5, 10
Minimum samples for leaves	1, 2, 4
Number of trees	100, 300, 500
Logistic regression	
Inverse of regularization strength	0.001, 0.01, 0.1, 1, 10, 100

COMMENT 1-3. *Some details of data composition and processing need to be added to verify the validity of model development and improve the clarity of the manuscript. The composition of the study population needs to be clarified. For example, the reported total number of CVD outcomes does not match the sum of outcomes of all individual CVD. Additional clarification is needed to explain the proportion of individuals with multiple CVDs and how the incidence rates are calculated. In addition, the authors didn't mention how dataset splitting was performed in training and testing and how individuals with multiple CVDs are handled in the model training and evaluation. If such cases were treated as independent data points, a single participant with multiple CVDs may appear in both training and testing sets, resulting in data leakage and inflation of model performance.*

RESPONSE 1-3. Thank you very much for the thoughtful comments and suggestions. We revised the **Methods** section to improve clarity and provide the requested details on data processing methods. First, we clarified that our individual-level data partition strategy

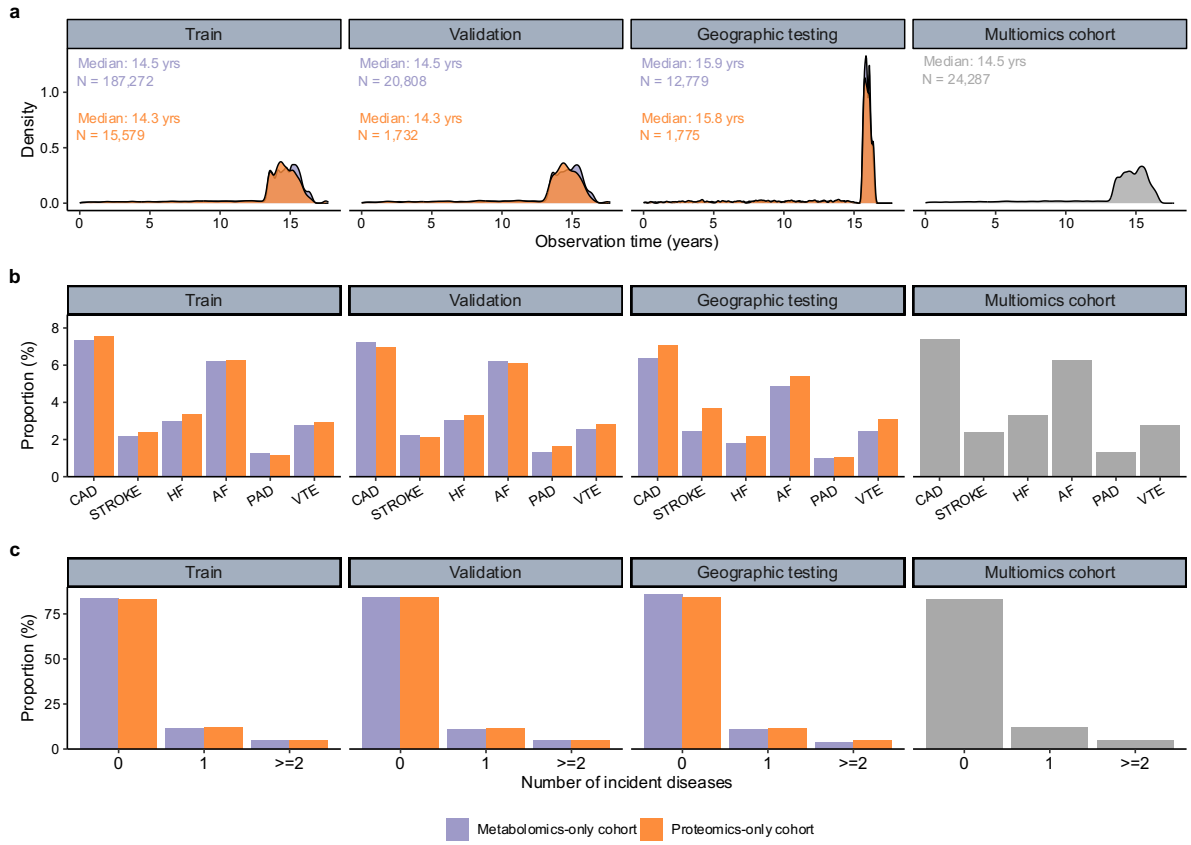
ensures that a single participant cannot appear in both training and testing sets, thus preventing data leakage (also refer to **RESPONSE 1-1**). Furthermore, we expanded our description of the multitask design, emphasizing that each participant is included only once in the dataset but may contribute to the loss function for multiple diseases if they experience several events, following the methodology of recent UKB studies (Buergel et al., 2022; You et al., 2023). This approach also explains that the difference between the total number of events and the sum of individual disease events arises because one person can contribute as a case to multiple diseases. Finally, we added details on how incidence rates were calculated in the **Methods** section, reported the specific proportion of individuals with multiple CVDs to the **Results** section, and updated the **Supplementary Figure 2** to present the follow-up duration and distribution of incident CVD cases.

The updated parts and figures are copied below:
In the main text:
Results Section, **Study population** Part: "Baseline characteristics and incident cases during follow-up were broadly consistent across all study datasets. The median age of participants at baseline was between 56.0 and 58.0 years. The proportion of males ranged from 42.8% to 45.8%. The prevalence of baseline medication use was comparable, with 15.8% to 18.1% of participants receiving lipid-lowering medication and 10.5% to 11.2% on antihypertensive medication. The distribution of outcomes was also similar, with 83.2–85.7% of participants remaining free of any incident CVD, 10.6–11.8% developing one CVD, and 3.7–5.0% developing multiple CVDs. Detailed baseline characteristics and follow-up information for each cohort are provided in Supplementary Tables 1–3 and Supplementary Figure 2."

Methods Section, **Deep learning models** Part: "Our multitask design enables the simultaneous modeling of multiple CVDs, accounting for individuals who experience several events over time (e.g., CAD followed by stroke or HF). In this setup, each participant contributes simultaneously to the prediction tasks for all six diseases. If a participant experiences multiple incident diseases, they generate multiple non-zero losses but are only included once in the dataset. This design avoids data duplication while fully leveraging all available incident disease information[17,26]. It also reflects the clinical reality in which multimorbidity is increasingly common, and where the most useful risk prediction models are those capable of discriminating the outcome of interest in the presence of co-occurring diseases[27]."

Methods Section, **Statistical analyses** Part: "Incidence proportion for each specific CVD was calculated as the number of new cases for that disease divided by the total number of participants at risk at baseline."

In the Supplementary Materials:



Supplementary Figure 2: Follow-up duration and distribution of incident cardiovascular disease (CVD) cases across datasets. **a.** Distribution of follow-up duration. Here, we defined the follow-up duration as the period from the date of baseline assessment to the date of any incident cardiovascular diseases, death, loss to follow-up, or end of available registry follow-up (November 30 2023 for England & Wales and December 31 2023 for Scotland), whichever came first. Text annotations provide the median observation time and the total number of participants for each dataset. **b.** Proportion of participants with incident cases of each CVD. **c.** Proportion of participants grouped by the number of incident CVD.

References:

Buergel, T., Steinfeldt, J., Ruyoga, G., Pietzner, M., Bizzarri, D., Vojinovic, D., Upmeier zu Belzen, J., Loock, L., Kittner, P., Christmann, L., Hollmann, N., Strangalies, H., Braunger, J. M., Wild, B., Chiesa, S. T., Spranger, J., Klostermann, F., van den Akker, E. B., Trompet, S., … Landmesser, U. (2022). Metabolomic profiles predict

individual multidisease outcomes. *Nature Medicine*, *28*(11), 2309–2320. https://doi.org/10.1038/s41591-022-01980-3

You, J., Guo, Y., Zhang, Y., Kang, J.-J., Wang, L.-B., Feng, J.-F., Cheng, W., & Yu, J.-T. (2023). Plasma proteomic profiles predict individual future health risk. *Nature Communications*, *14*(1), 7817. https://doi.org/10.1038/s41467-023-43575-7

COMMENT 1-4. *Additional details of the deep learning model architecture and loss function need to be added. For example, does the model output a unique score for each CVD (9 scores in total)? How is the overall loss calculated? Also, the authors should include their final choice of model hyperparameters to improve reproducibility of the model.*

RESPONSE 1-4. Thank you so much for these valuable suggestions. We revised the "Model development" subsection in the manuscript and added a "Sample-level and task-level weighting" part in the Supplementary Methods to improve the clarity and reproducibility of our CardiOmicScore framework. First, we clarified the model architecture and outputs in the **Methods** section, specifying that the model consists of a shared network and six parallel, disease-specific networks (one for each CVD), which generate a unique MetScore or ProScore for each CVD. Second, we expanded the description of the loss function in the **Methods** section, detailing our sample- and task-level weighting strategy to address class imbalance and clarifying how the final loss is computed, with the full mathematical formulas provided in the **Supplementary Methods** (also refer to **RESPONSE 1-1**). Finally, to enhance reproducibility, the optimal hyperparameter combinations are provided in the **Supplementary Table 15**, and a diagram of the model architecture is available in **Supplementary Figure 21**. All code used for model development and analysis has been made publicly available at Github: https://github.com/YanLuoCityU/cardiomicscore.

The revised parts are copied below:

Methods Section, **Deep learning models** Part: "We developed two multitask deep neural networks, named MetNet and ProNet, to derive the MetScore and ProScore for six CVDs, with 168 metabolites and 2,920 proteins as the input, respectively. MetNet was trained on the training set from the Metabolomics-only cohort, and ProNet on the training set from the Proteomics-only cohort, while their respective validation sets were used to monitor model fitting and select the optimal checkpoints. The final model was first evaluated on the

geographic testing set to assess generalizability across regions. Subsequently, the Multiomics cohort, which remained untouched throughout model development, was used for a further performance evaluation.

MetNet and ProNet had similar model architectures, consisting of a shared network and six parallel, disease-specific networks (one for each CVD) (Supplementary Figure 21). The shared neural network (denoted as "shared multilayer perceptron [MLP]") included multiple fully connected layers, each with a nonlinear activation function, dropout, and batch normalization. The output of the shared network (i.e., shared representation) is a high-dimension representation, capturing the common features for all CVDs. The disease-specific network, comprising a disease-specific MLP and a predictor MLP, was designed to learn disease-specific features and predict the risk of individual diseases (i.e., whether the disease will occur or not). The original metabolomic/proteomic biomarkers were fed into the disease-specific multi-layer MLP with nonlinear activation functions, dropout, and batch normalization to obtain a disease-specific representation. The shared and disease-specific representations were then concatenated and passed on to the predictor MLP that included linear layers followed by nonlinear activation functions, dropout, and batch normalization before the final single-output layer to generate the MetScore or ProScore for each disease. These scores were subsequently used as predictors in CPH models to assess their utility for cardiovascular risk prediction. Binary cross entropy was used as the loss function for each disease. To address the class imbalance in multitask prediction, we applied both sample-level and task-level weighting (Supplementary Methods). At the sample level, the pos_weight parameter was calculated as the ratio of the number of negative samples to positive samples for each task, increasing the penalty for misclassifying positive cases and encouraging the model to pay more attention to rare events. At the task level, each disease was assigned a weight equal to the inverse of its number of positive samples, ensuring that tasks with fewer positive cases contribute proportionally more to the overall loss. The final loss was computed as the mean of the weighted losses across all tasks.

Our multitask design enables the simultaneous modeling of multiple CVDs, accounting for individuals who experience several events over time (e.g., CAD followed by stroke or HF). In this setup, each participant contributes simultaneously to the prediction tasks for all six diseases. If a participant experiences multiple incident diseases, they generate multiple non-zero losses but are only included once in the dataset. This design avoids data duplication

while fully leveraging all available incident disease information[17,26]. It also reflects the clinical reality in which multimorbidity is increasingly common, and where the most useful risk prediction models are those capable of discriminating the outcome of interest in the presence of co-occurring diseases[27].

We performed random hyperparameter searches separately for MetNet and ProNet using their respective training and validation sets. Each model was trained and evaluated across 100 randomly sampled hyperparameter configurations. Hyperparameters included the hidden layer architecture, dropout rate, and activation function for all MLPs (Supplementary Table 15). The optimal hyperparameter combination for each model, provided in Supplementary Table 15, was selected based on the highest average Harrell's C-index across all six CVDs on the validation set. Once the optimal architecture was determined, we trained the final models five times using different random seeds to account for stochastic variability in the training process. The ultimate MetScore and ProScore for each participant represent the average of the predictions from these five independent runs, ensuring the robustness of the final scores. The final models were trained using the Adam optimizer. All deep learning models were developed in Python v3.8.5 using Pytorch v1.11.0 package with hyperparameter tuning using Optuna v4.0.0 package[77,78]."

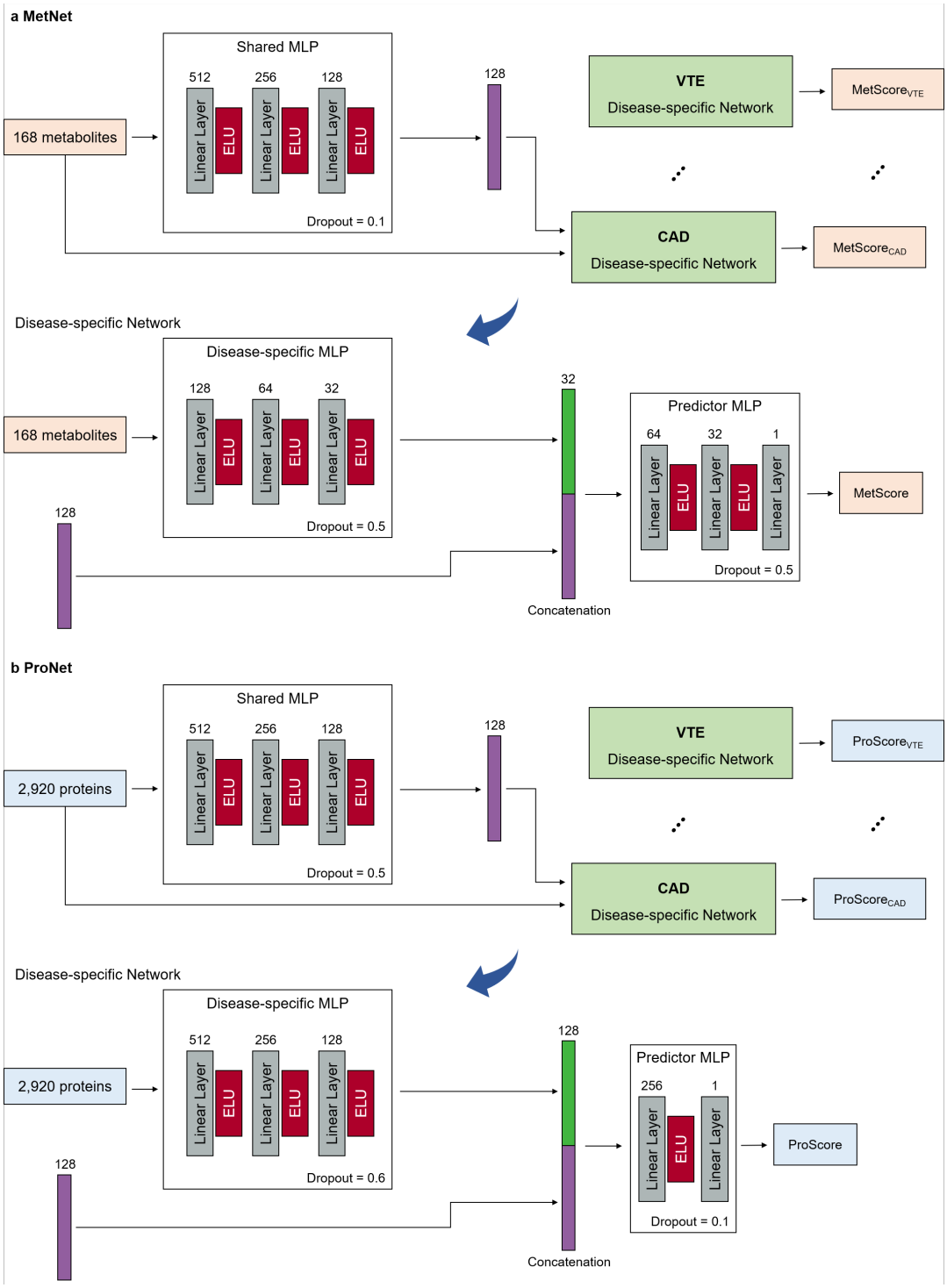
The updated figures and tables are copied below:

Supplementary Table 15: Hyperparameters for ProNet and MetNet.

Parameters	Search space	Final hyperparameters	
Share MLP		ProNet	MetNet
Hidden dimensions[1]	• [128, 64, 32] • [256, 128, 64, 32] • [512, 256, 128, 64, 32] • [256, 128, 64] • [512, 256, 128, 64] • [512, 256, 128]	[512, 256, 128]	[512, 256, 128]
Dropout	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7	0.5	0.1
Disease-specific MLP			
Hidden dimensions[1]	• [128, 64, 32] • [256, 128, 64, 32] • [512, 256, 128, 64, 32] • [256, 128, 64] • [512, 256, 128, 64] • [512, 256, 128]	[512, 256, 128]	[128, 64, 32]
Dropout	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7	0.6	0.5

Predictor MLP			
Hidden dimensions[2]	• [32] • [64, 32] • [128, 64, 32] • [256, 128, 64, 32] • [512, 256, 128, 64, 32] • [64] • [128, 64] • [256, 128, 64] • [512, 256, 128, 64] • [128] • [256, 128] • [512, 256, 128] • [256] • [512, 256]	[256]	[64, 32]
Dropout	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7	0.1	0.5
Activation	ReLU, LeakyReLU, ELU, Tanh	ELU	ELU
Batch size	128 for ProNet, 1024 for MetNet	128	1024
Learning rate	1×10^{-5} for ProNet, 1×10^{-4} for MetNet	1×10^{-5}	1×10^{-4}
Epochs	50	50	50

[1]Each list defines the neuron counts for successive hidden layers. The final value in each list represents the output dimension for that specific MLP block.
[2]The Predictor MLP was designed with a single output neuron (output dimension = 1) to generate the final scalar risk score.



Supplementary Figure 21: The architecture of MetNet (**a**) and ProNet (**b**). MLP = Multilayer perceptron, CAD = Coronary artery disease, VTE = Venous thromboembolism.

COMMENT 1-5. *In Figure 3, it would be clearer if all subplots use the same x-axis range in each sub-figure.*

RESPONSE 1-5. Thank you for pointing this out. We updated Figure 3 to use the same x-axis range for all subplots as requested. The revised figure is copied below:

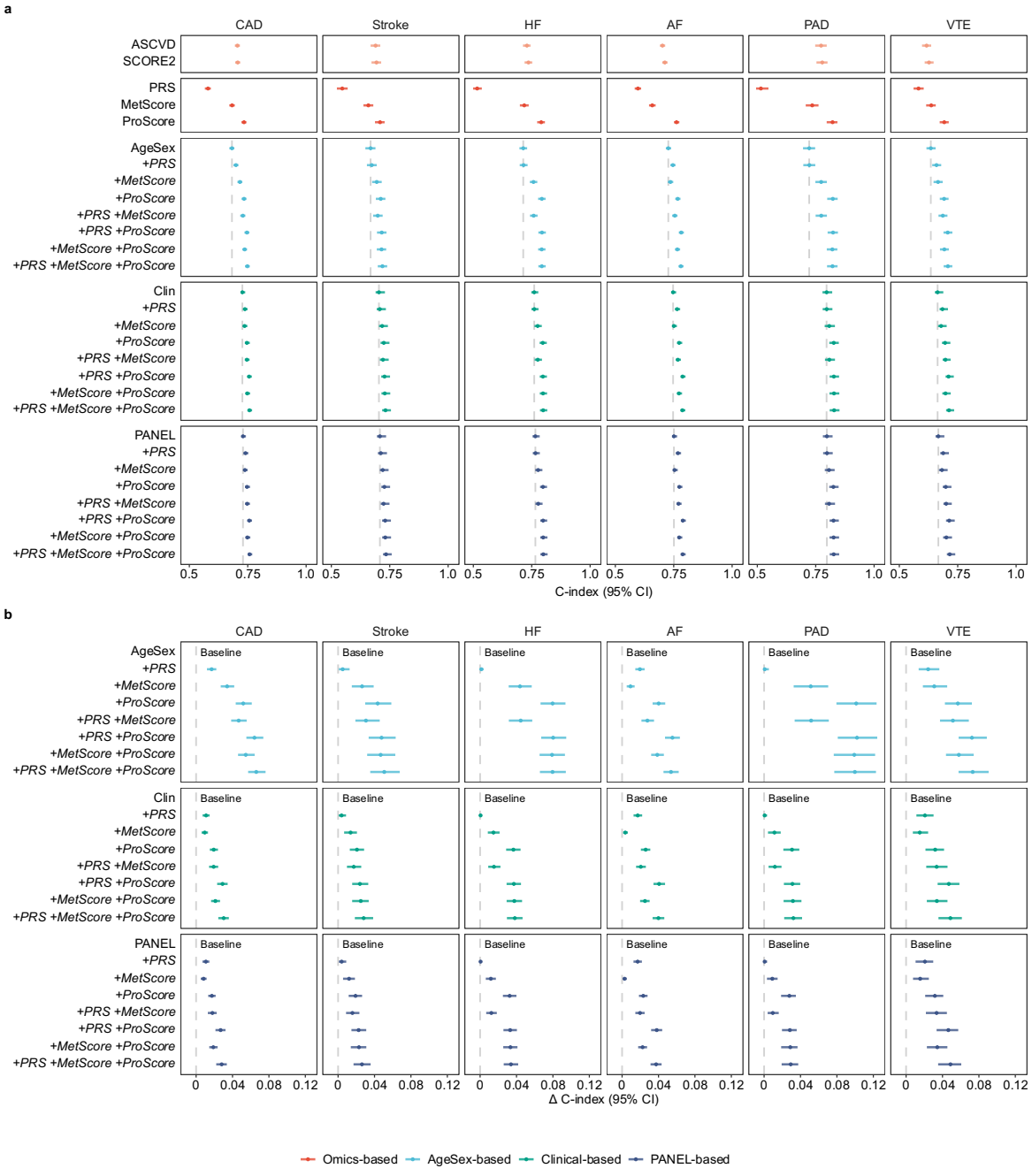


Figure 3: Predictive performance of multiomics for cardiovascular diseases. **a.** Discriminative performance of models trained on various predictor sets. Vertical dashed lines indicate the performance of the three clinical predictor sets. Discriminative performance was

evaluated by the Harrell's C-index. **b.** Differences in discriminative performance by adding omics information to clinical predictor sets. Data are presented as point estimates with 95% confidence intervals estimated from 1,000 bootstrap resamples.

COMMENT 1-6. *There are some redundant texts in the manuscript. For example, the authors currently repeatedly explain meanings of all CVD abbreviations and population breakdown in all figure legends. Instead, it may be better if the authors use a separate section (e.g. in Methods) to explain their evaluation setup.*

RESPONSE 1-6. Thank you for this valuable feedback. We followed your suggestion and removed the redundant text from the figure legends. The definitions of the CVD abbreviations and the number of incident cases and incidence rates for each cardiovascular disease are described in the **Methods** and **Results** sections, respectively, to avoid repetition. We copied the updated paragraphs in the main text below:

Results Section, **Study population** Part: "A two-stage, individual-level data partitioning strategy was employed to rigorously evaluate our model's performance and generalizability (see sub-section 4.2 and Supplementary Figure 1). First, we excluded participants who had neither metabolomics nor proteomics data available. Then, we divided the remaining participants into the Metabolomics-only (N = 220,859; those with metabolomics but not proteomics data), Proteomics-only (N = 19,086; those with proteomics but not metabolomics data), and Multiomics (N = 24,287; those with genomics, metabolomics, and proteomics data) cohorts based on omics availability. The Metabolomics-only and Proteomics-only cohorts were used within our CardiOmicScore framework to develop two deep learning models, MetNet and ProNet, respectively. The Multiomics cohort was reserved as an untouched validation set to assess the added predictive value of integrating different omics data."

Following best practices in recent UKB studies[39–41], each of these development cohorts was further split into training/validation sets (England and Wales) and a geographic testing set (Scotland). Specifically, the Metabolomics-only cohort included 187,272 participants in the training set, 20,808 in the validation set, and 12,779 in the geographic testing set, while the Proteomics-only cohort comprised 15,579, 1,732, and 1,775 participants in the training, validation, and geographic testing sets, respectively. Baseline characteristics were

comparable between training and validation sets in both cohorts, whereas significant differences were observed between training and geographic testing sets (Supplementary Tables 2–3), thus allowing a robust assessment of the models' regional generalizability.

Baseline characteristics and incident cases during follow-up were broadly consistent across all study datasets. The median age of participants at baseline was between 56.0 and 58.0 years. The proportion of males ranged from 42.8% to 45.8%. The prevalence of baseline medication use was comparable, with 15.8% to 18.1% of participants receiving lipid-lowering medication and 10.5% to 11.2% on antihypertensive medication. The distribution of outcomes was also similar, with 83.2–85.7% of participants remaining free of any incident CVD, 10.6–11.8% developing one CVD, and 3.7–5.0% developing multiple CVDs. Detailed baseline characteristics and follow-up information for each cohort are provided in Supplementary Tables 1–3 and Supplementary Figure 2."

Methods Section, **Ascertainment of cardiovascular diseases** Part: "The six CVDs analyzed in this study included coronary artery disease (CAD), stroke, heart failure (HF), atrial fibrillation (AF), peripheral artery disease (PAD), and venous thromboembolism (VTE). All diseases were ascertained based on self-reported diagnoses and operations, hospital episode statistics, and death records. Self-reported information was used only to determine the presence of CVDs at baseline. Detailed definitions are provided in Supplementary Table 5. Follow-up duration was calculated from the date of baseline assessment to the date of incident outcomes, death, loss to follow-up, or end of available registry follow-up (November 30 2023 for England & Wales and December 31 2023 for Scotland)[66], whichever came first."

COMMENT 1-7. *Some figures need to be adjusted. For example, the texts in Figure 5 c & d are too small to read. Supplementary figures 13 & 14 are not completely displayed.*

RESPONSE 1-7. Thank you for the comment. We corrected all the issues you pointed out: the content from the former Figure 5c–5d was moved to a new Figure 6, and the supplementary figures you mentioned (renumbered as Supplementary Figures 10 and 11) were reformatted for full and clear display. Due to their large size, Supplementary Figures 10 and 11 are not shown here. The updated Figures 5–6 is copied below:

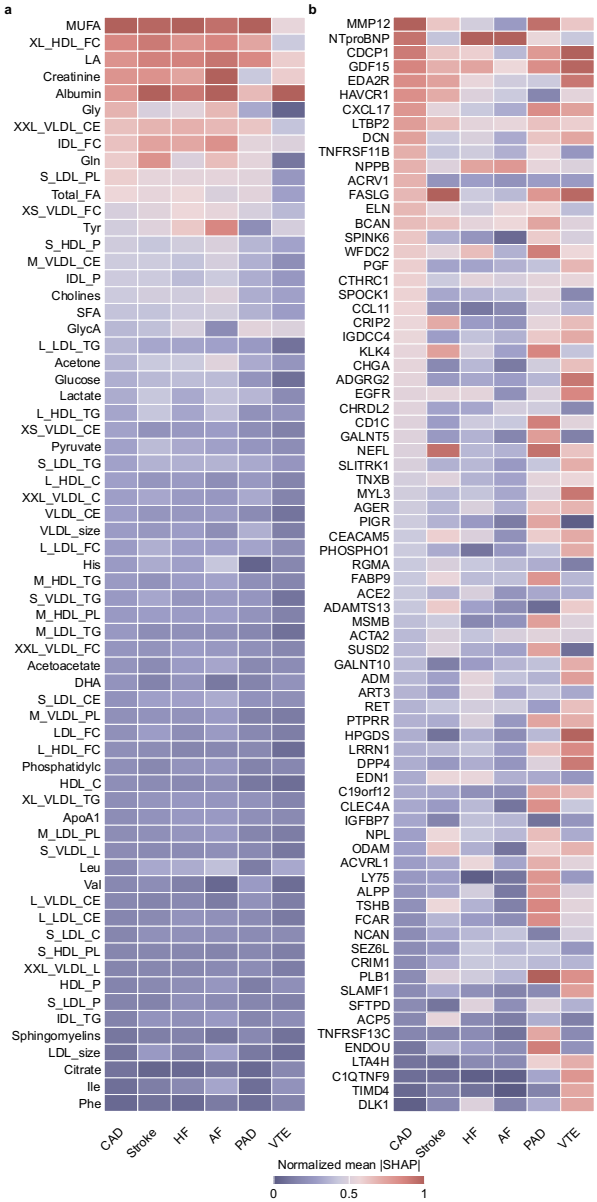


Figure 5: Importance of biomarkers in cardiovascular disease-specific MetScore (**a**) and ProScore (**b**). Importance was measured by mean absolute SHAP values. For each disease, the 50 most important metabolites and the 25 most important proteins were identified. The union of these features across all diseases yielded a total of 65 metabolites and 77 proteins.

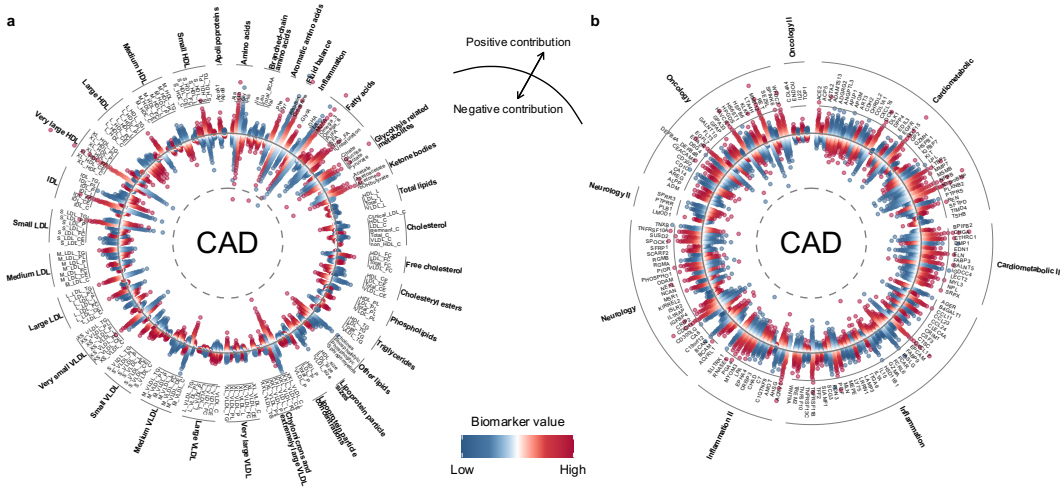


Figure 6: Metabolites (**a**) and proteins (**b**) are associated with cardiovascular disease risk. SHAP beeswarm plots are shown here for coronary artery disease as an example. Each individual is assigned a SHAP value for every protein and metabolite. To simplify the presentation, SHAP values for each biomarker were divided into percentiles, and the mean SHAP value and biomarker level were calculated for each percentile. Each dot represents one percentile. All 168 metabolites and the top-ranked 157 proteins are shown. Dark grey lines indicate zero SHAP value (i.e., no contribution). The farther a dot from this line, the stronger the absolute contribution to the MetScore or ProScore. Deviations toward the center and periphery represent negative and positive contributions. Red and blue colors indicate high and low levels of metabolites and proteins, respectively.

REVIEWER 2'S COMMENTS

Remarks to the Author:

COMMENT 2-1. *The authors undertake Genomic, Transcriptomic, and Proteomic profiling of a large cohort of patients from the UK biobank and through AI deep learning models have developed genomic, transcriptomic, and proteomic risk scores for the prediction of 9 cardiovascular disease outcomes. This study reports that such these multiomic risk scores are superior to existing risk scores based on clinical and biochemical data both individually and in combination. This is a great pilot dataset, but it needs true external validation in an independent cohort, in order to be mature for presentation as definitive.*

RESPONSE 2-1. Thank you very much for the positive feedback and for raising the crucial point about model validation. We agree that robust validation is essential for demonstrating the clinical potential of the risk model. Given the limited availability of external cohorts with comparable multi-omics data, we implemented a two-stage individual-level data partition strategy that goes beyond standard internal validation to rigorously assess the generalizability of our models (**Figure 1, Supplementary Figure 1,** and also refer to **RESPONSE 1-1**). This approach, consistent with best practices for large-scale studies within the UK Biobank (Argentieri et al., 2024; Zhang et al., 2024; Julkunen & Rousu, 2025), is intended to approximate the conditions of external validation. We first evaluated the performance of the MetScore and ProScore generated by our CardiOmicScore framework on a geographically distinct testing set (participants from Scotland) and subsequently performed a final evaluation of added value of multiomics integration on the Multiomics cohort, which was kept entirely separate from all development stages. Notably, even with significant differences in baseline clinical characteristics between the training and geographic testing sets, both MetScore and ProScore showed strong and reproducible predictive performance, which was further confirmed in the Multiomics cohort (**Supplementary Figure 6**). We believe this multi-stage approach provides robust evidence of our models' ability to generalize.

The revised parts are copied below:

In the main text:

Results Section, **Study population** Part: "A two-stage, individual-level data partitioning strategy was employed to rigorously evaluate our model's performance and generalizability (see sub-section 4.2 and Supplementary Figure 1). First, we excluded participants who had neither metabolomics nor proteomics data available. Then, we divided the remaining

participants into the Metabolomics-only (N = 220,859; those with metabolomics but not proteomics data), Proteomics-only (N = 19,086; those with proteomics but not metabolomics data), and Multiomics (N = 24,287; those with genomics, metabolomics, and proteomics data) cohorts based on omics availability. The Metabolomics-only and Proteomics-only cohorts were used within our CardiOmicScore framework to develop two deep learning models, MetNet and ProNet, respectively. The Multiomics cohort was reserved as an untouched validation set to assess the added predictive value of integrating different omics data.

Following best practices in recent UKB studies[39–41], each of these development cohorts was further split into training/validation sets (England and Wales) and a geographic testing set (Scotland). Specifically, the Metabolomics-only cohort included 187,272 participants in the training set, 20,808 in the validation set, and 12,779 in the geographic testing set, while the Proteomics-only cohort comprised 15,579, 1,732, and 1,775 participants in the training, validation, and geographic testing sets, respectively. Baseline characteristics were comparable between training and validation sets in both cohorts, whereas significant differences were observed between training and geographic testing sets (Supplementary Tables 2–3), thus allowing a robust assessment of the models' regional generalizability.

Baseline characteristics and incident cases during follow-up were broadly consistent across all study datasets. The median age of participants at baseline was between 56.0 and 58.0 years. The proportion of males ranged from 42.8% to 45.8%. The prevalence of baseline medication use was comparable, with 15.8% to 18.1% of participants receiving lipid-lowering medication and 10.5% to 11.2% on antihypertensive medication. The distribution of outcomes was also similar, with 83.2–85.7% of participants remaining free of any incident CVD, 10.6–11.8% developing one CVD, and 3.7–5.0% developing multiple CVDs. Detailed baseline characteristics and follow-up information for each cohort are provided in Supplementary Tables 1–3 and Supplementary Figure 2."

Results Section, **Advancing cardiovascular risk prediction with the power of omics information** Part: "Among the three omics scores, ProScore showed the highest performance for all CVDs with the C-index ranging from 0.69 (95% CI: 0.67–0.71) for VTE to 0.82 (95% CI: 0.80–0.84) for PAD (Figure 3a). The C-index of MetScore spanned from 0.64 (95% CI: 0.61–0.66) for VTE to 0.74 (95% CI: 0.71–0.76) for PAD. Importantly, the strong

performance of ProScore and MetScore was replicated in their respective geographic testing sets, underscoring their generalizability across distinct UK populations (Supplementary Figure 6). In contrast, PRS provided the most limited predictive capacity (C-index range: 0.52–0.60). Finally, the discriminative performance increased with more clinical predictors included in the CPH models for all diseases (Figure 3a)."

Methods Section, **Data partition and imputation** Part: "To rigorously evaluate model performance and ensure generalizability, we adopted a two-stage individual-level data partition strategy, designed to approximate external validation while fully leveraging UKB data (Supplementary Figure 1). First, we excluded participants without both metabolomics and proteomics data. The remaining participants were then split into three cohorts based on omics data availability: a Metabolomics-only cohort (N = 220,859; those with metabolomics but not proteomics data), a Proteomics-only cohort (N = 19,086; those with proteomics but not metabolomics data), and a Multiomics cohort (N = 24,287; those with genomics, metabolomics, and proteomics data). In our CardiOmicScore framework, we used the Metabolomics-only and Proteomics-only cohorts to develop the respective deep learning models (MetNet and ProNet) and to generate the corresponding omics scores (MetScore and ProScore). To enable geographic validation, we further split each of these two cohorts into a training/validation set (England and Wales) and a geographic testing set (Scotland) according to the recruitment regions, following approaches commonly used in recent UKB studies[39–41]. The Multiomics cohort was kept untouched throughout model training and was subsequently used as an additional validation cohort to assess the incremental benefit of integrating multiple omics data types for risk prediction.

Categorical variables were one-hot encoded and continuous variables were standardized by the mean and standard deviation. We used the K-nearest neighbors algorithm (scikit-learn v1.3.2 package)[65], setting the number of neighbors to five, to impute missing values for continuous variables. Categorical variables were imputed with mode. For metabolomics or proteomics data, imputation and standardization of continuous variables were performed using parameters derived exclusively from the training set. The fitted preprocessing models were then applied to the corresponding validation and geographic testing sets, as well as the Multiomics cohort. To derive imputation models and standardization parameters for subsequent application to the Multiomics cohort, we constructed two cohorts from all

available individuals with relevant data: one for clinical information ($N = 427{,}225$) and one for genomics ($N = 412{,}797$).”

The updated figures are copied below:

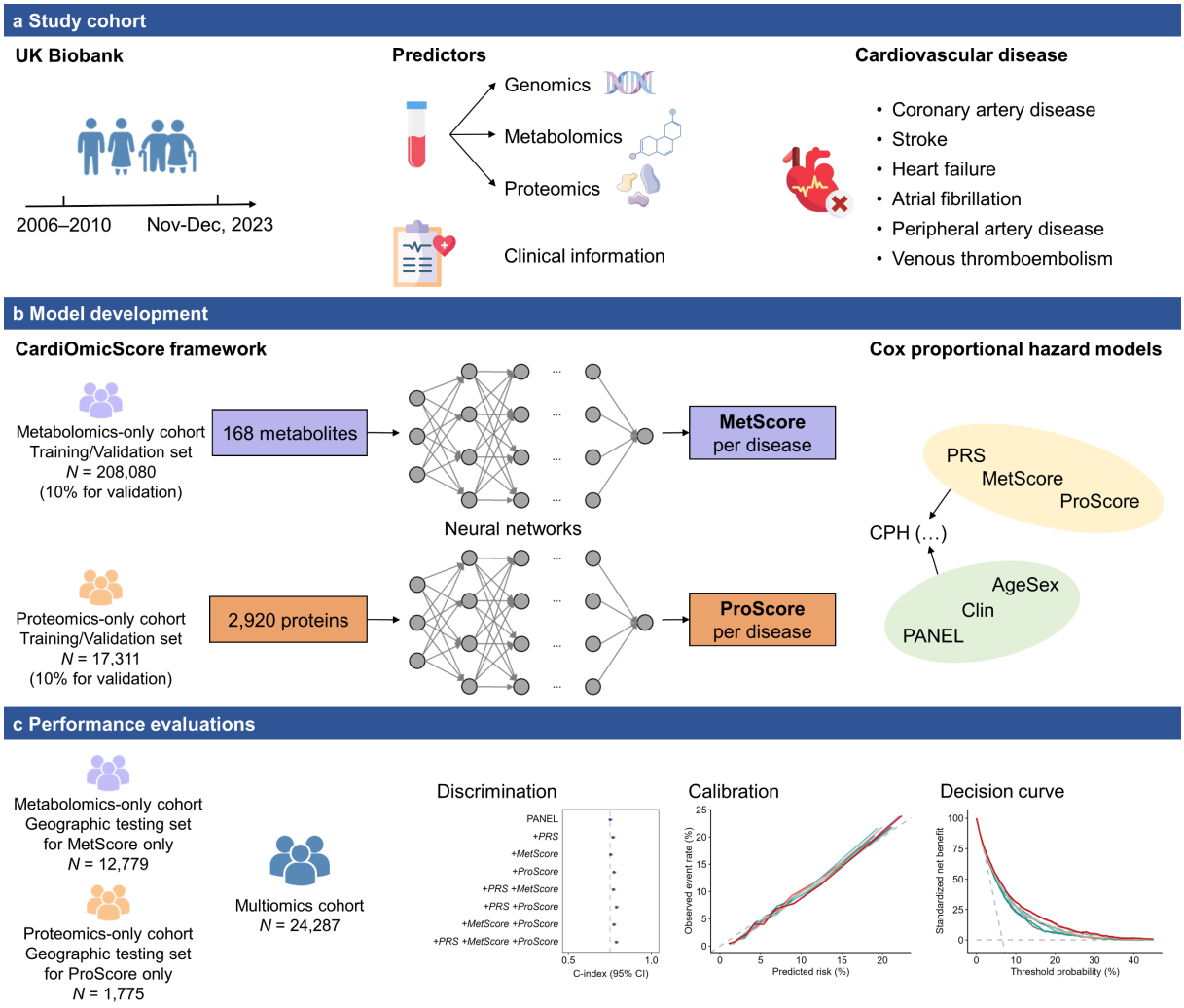
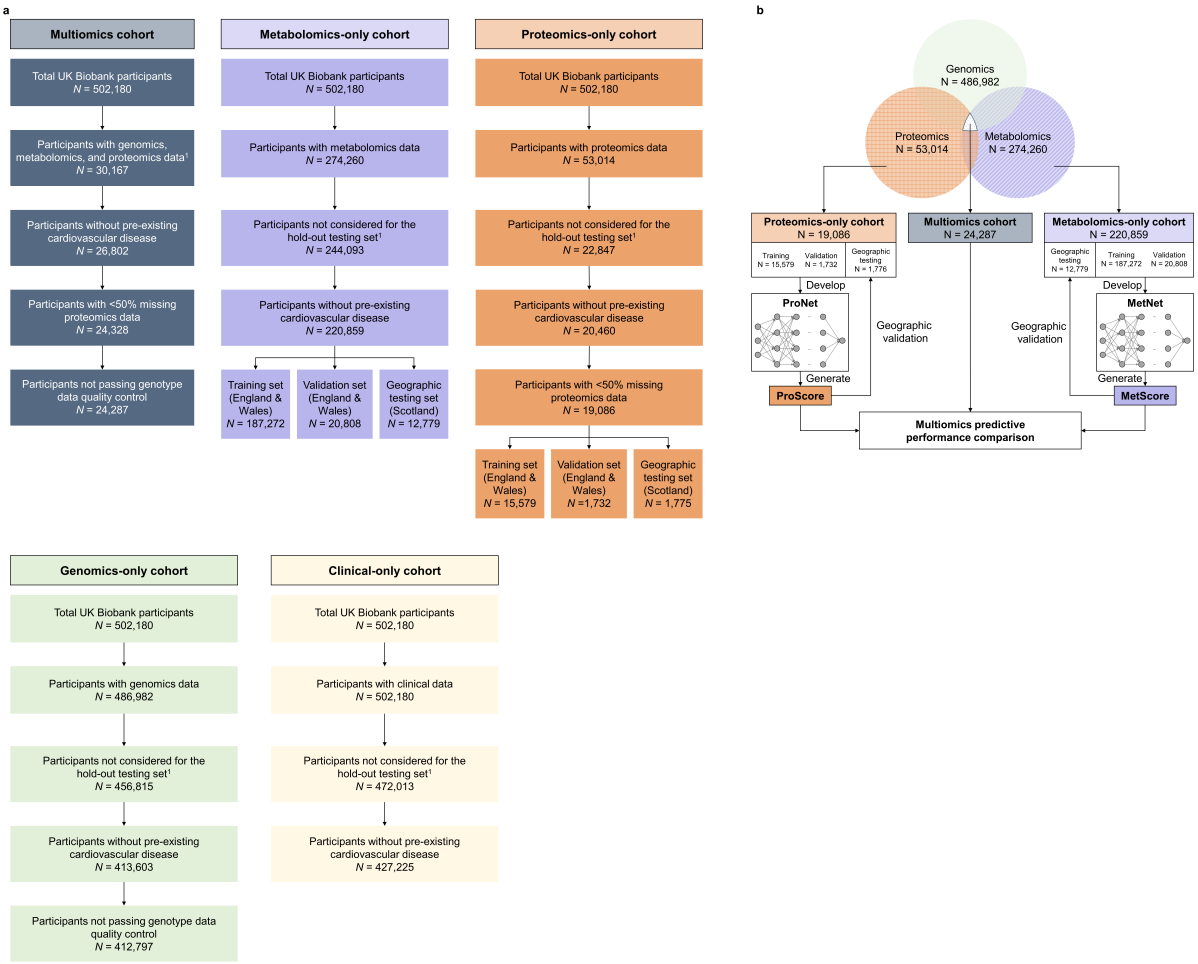
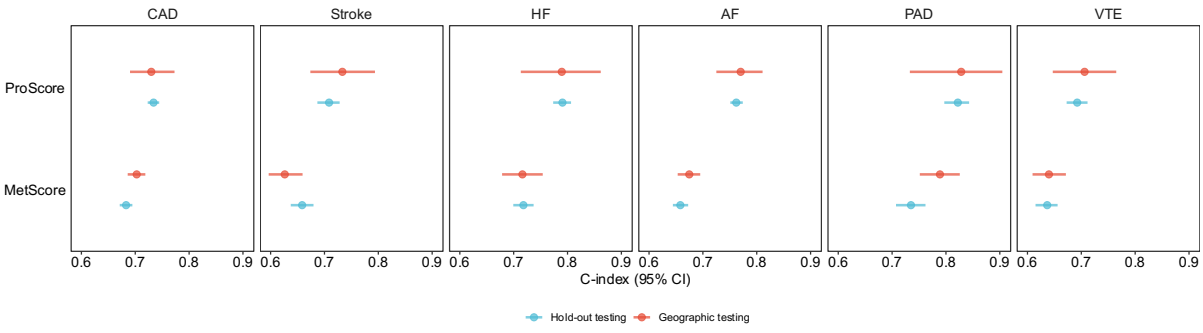


Figure 1: Study design. **a.** We extracted data from the UK Biobank, including clinical, genomic, metabolomic, and proteomic predictors, as well as six cardiovascular diseases (CVDs) defined by self-reported diagnoses, hospital episode statistics, and death records. **b.** A two-stage data partition strategy was applied: the Multiomics cohort, which remained untouched throughout model training and was used as an independent validation set to evaluate the added value of integrating multiple omics data types, and separate training, validation, and geographic testing sets within the Metabolomics-only and Proteomics-only cohorts for model development. Our CardiOmicScore framework utilized two multitask deep neural networks to predict the risk of six CVDs from 168 metabolites and 2,920 proteins separately, generating MetScore and ProScore for each CVD. We trained Cox proportional hazard (CPH) models on various combinations of polygenic risk score (PRS), MetScore, ProScore, and three predefined clinical predictor sets (i.e., AgeSex, Clin, and PANEL). **c.** Model performance was first evaluated for MetScore and ProScore in their respective geographic testing sets, and subsequently for all feature combinations in the Multiomics cohort. Performance metrics included Harrell’s C-index, calibration plots, and net benefit curves, with confidence intervals estimated using 1,000 bootstrap resamples.



Supplementary Figure 1: Two-stage individual-level data partitioning strategy. **a.** Sample exclusion flowchart. Genotype data quality control excluded participants with discordant self-reported and genetic sex, sex chromosome aneuploidy, and individuals having more than ten third-degree genetic relatives. **b.** Study framework.
[1]"Participants not considered for the Multiomics cohort" in the Metabolomics-only, Proteomics-only, Genomics-only, and Clinical-only cohorts refer to those with available metabolomics, proteomics, genomics, or clinical data, respectively, after excluding the 30,167 individuals with all three omics data, who were eligible for inclusion in the Multiomics cohort.



Supplementary Figure 6: Discriminative performance of MetScore and ProScore. Discriminative performance was evaluated separately for MetScore and ProScore in both the geographic testing and the Multiomics cohort. The Multiomics cohort included 24,287 participants, while the geographic testing sets included 12,779 for MetScore and 1,775 for ProScore. Performance was assessed using Harrell's C-index with 95% confidence intervals estimated by 1,000 bootstrap resamples.

References:

Argentieri, M. A., Xiao, S., Bennett, D., Winchester, L., Nevado-Holgado, A. J., Ghose, U., Albukhari, A., Yao, P., Mazidi, M., Lv, J., Millwood, I., Fry, H., Rodosthenous, R. S., Partanen, J., Zheng, Z., Kurki, M., Daly, M. J., Palotie, A., Adams, C. J., … van Duijn, C. M. (2024). Proteomic aging clock predicts mortality and risk of common age-related diseases in diverse populations. *Nature Medicine*, 1–11. https://doi.org/10.1038/s41591-024-03164-7

Julkunen, H., & Rousu, J. (2025). Comprehensive interaction modeling with machine learning improves prediction of disease risk in the UK Biobank. *Nature Communications*, *16*(1), 6620. https://doi.org/10.1038/s41467-025-61891-y

Zhang, S., Wang, Z., Wang, Y., Zhu, Y., Zhou, Q., Jian, X., Zhao, G., Qiu, J., Xia, K., Tang, B., Mutz, J., Li, J., & Li, B. (2024). A metabolomic profile of biological aging in 250,341 individuals from the UK Biobank. *Nature Communications*, *15*(1), 8081. https://doi.org/10.1038/s41467-024-52310-9

COMMENT 2-2. *The authors helpfully provide a list of proteins and metabolites whose levels increase and decrease the risk of cardiovascular disease. However with NTpro-BNP coming out as the top hit in all 9 cardiovascular diseases is there risk that these results have been confounded? High NTProBNP contributes towards increased CAD risk for example, but does this merely represent the interaction or comorbidity of different types of CVD such as CAD and HF? On closer inspection, the directionality and effect sizes of the proteins are quite similar visually between CVD conditions – are the authors therefore suggesting that the multiomic signatures of 9 different cardiovascular diseases are largely the same?*

RESPONSE 2-2. Thank you very much for the insightful comments. We agree that, in our initial analysis, the multiomic signatures were overly similar across different CVDs. Upon further investigation, we found that this issue was primarily driven by class imbalance, whereby model training was dominated by the most prevalent diseases (e.g., CAD, AF, and

HF). To address these concerns, we re-ran our analyses after implementing two significant changes. First, we refined our study's scope to the six most common CVDs, removing the two rarest outcomes (ventricular arrhythmias and abdominal aortic aneurysm) as their extremely low incidence rates (0.8% and 0.4%, respectively) were a primary source of model instability. Second, we incorporated both sample-level and task-level weighting strategies into CardiOmicScore framework to ensure that less frequent diseases contributed proportionally to model training (also refer to **RESPONSE 1-1**). In addition, to improve the stability and reliability of feature-importance estimates, we trained five models per score with different random seeds and calculated the final SHAP value for each biomarker by averaging the results across these five runs. These revisions have led to a much more nuanced and biologically plausible set of findings. For metabolites, there is a partial overlap in the top-ranking biomarkers across the six CVDs, reflecting shared metabolic pathways. In contrast, the important proteins are now more disease-specific. For example, GDF15 was important for CAD, HF, AF, and VTE, whereas KLK4 was specific for stroke and BCAN for AF (**Figures 5–6**). To more clearly illustrate these shared and disease-specific associations, we summarized the top-ranking metabolites and proteins for all six CVDs in a new table (**Supplementary Table 4**). Collectively, these results suggest that the revised model captures both common and disease-specific biology in a more reliable manner.

Results Section, **AI identifies disease-specific metabolites and proteins** Part: "We further utilized the Shapley Additive exPlanations (SHAP) method to examine the relationship of metabolites and proteins with each CVD. To capture global importance, we first calculated the mean absolute SHAP value for each feature and normalized these values within each outcome. Figure 5a-b display the top 65 metabolites and 77 proteins ranked by their global importance.

We observed a partial overlap in the top-ranking metabolites across the six CVDs. The most impactful metabolites across all CVDs included creatinine, albumin, glutamine, fatty acids such as linoleic acids (LA) and monounsaturated fatty acids (MUFA), and lipoprotein components such as free cholesterol in intermediate-density lipoprotein (IDL_FC), free cholesterol in very large high-density lipoprotein (XL_HDL_FC), cholesteryl esters in chylomicrons and extremely large very-low-density lipoprotein (XXL_VLDL_CE). Meanwhile, other biomarkers were important for specific diseases, such as tyrosine and leucine for HF and AF, and glycoprotein acetyls (GlycA) for PAD and VTE. For proteins, the

important biomarkers were more distinct between CVDs. We found that NT-proBNP and NPPB were important for CAD, HF, and AF, while GDF15 was important for CAD, HF, AF, and VTE. Other key associations included MMP12 for CAD and PAD, CDCP1 for CAD, HF, and VTE, EDA2R for CAD and stroke, NEFL for stroke and PAD, WFDC2 for HF and PAD, and FASLG for stroke and VTE. The most important disease-specific proteins also included KLK4 and CRIP2 for stroke; BCAN and ELN for AF; PLB1 and ENDOU for PAD; and HPGDS and ADGRG2 for VTE.

The SHAP analysis also revealed that while the direction of the top biomarker-disease associations was highly consistent across all six CVDs, the magnitude of their contributions varied. We used CAD as an example to illustrate the associations between plasma biomarkers and disease risks. Among metabolites, higher plasma levels of creatinine, GlycA, MUFA, tyrosine, glutamine, IDL_FC, and XL_HDL_FC were associated with an increased risk of CAD, while higher levels of albumin, LA, leucine, and XXL_VLDL_CE were associated with lower CAD risk (Figure 6a). For proteins, NT-proBNP, NPPB, GDF15, MMP12, CDCP1, EDA2R, NEFL, WFDC2, KLK4, CRIP2, ELN, positively influenced disease risk, while a negative association was observed between CAD risk and elevated levels of FASLG, BCAN, PLB1, ENDOU, HPGDS, and ADGRG2 (Figure 6b). SHAP beeswarm plots of metabolites and proteins for other diseases are provided in Supplementary Figure 10–11. Summary of top-ranking metabolites and proteins are presented in Supplementary Table 4."

Discussion Section, **AI-identified biomarkers offer biological insights** Part: "Evaluating the contributions of important proteins and metabolites to each CVD reveals underlying disease mechanisms and potential therapeutic targets. We confirmed the role of proteins such as NT-proBNP and NPPB, as well as metabolites such as albumin and creatinine, in CVD risk assessment. These biomarkers are well-established in epidemiological studies and are already part of routine clinical care[47–50]. However, our analyses revealed that ProScore outperformed any single proteins such as NT-proBNP and NPPB both as an individual predictor and when added to clinical models. Similarly, individual metabolic biomarkers such as creatinine, albumin, and traditional lipid-related biomarkers (e.g., total lipids or cholesterol), yielded lower predictive performance compared to MetScore. These findings suggest that the MetScore and ProScore may provide a more comprehensive representation of disease-related molecular mechanisms than single biomarkers, and thus highlight the importance of high-throughput metabolomics and proteomics in personalized CVD risk

prediction. Moreover, our CardiOmicScore framework outperformed conventional machine learning algorithms, underscoring the unique strength of the multitask deep learning architecture to model complex, non-linear interactions in high-dimensional omics data.

In addition to the established biomarkers, we identified several less-studied circulating metabolites and proteins associated with CVD risk, which were consistent with previous omics-based modeling studies[17,22–33]. We observed a positive association between future CVD and GlycA, a composite NMR-based signal and biomarker of systemic inflammation[51]. In line with earlier findings, increasing levels of MUFA[52] and tyrosine[53] were associated with a higher risk of CVDs, while we also confirmed the protective role of LA[52]. Interestingly, while remnant cholesterol is recognized as a causal risk factor for atherosclerotic cardiovascular disease (ASCVD)[54], we found that its components may have inconsistent associations with CVD risk: higher IDL_FC levels were associated with increased risk, whereas higher XXL_VLDL_CE levels were associated with lower risk. These opposing associations warrant future investigation, as they suggest that measuring IDL_FC and XXL_VLDL_CE may further stratify risk beyond traditional LDL-cholesterol levels.

Regarding proteins, our analysis identified a diverse spectrum of biomarkers, revealing pathways that are shared across CVDs and highly specific to individual diseases. Several proteins were associated with multiple CVDs, suggesting common underlying pathological processes. For example, the stress- and inflammation-induced cytokine GDF15 showed strong predictive value for CAD, HF, AF, and VTE, acting as a general marker of systemic stress and broad cardiovascular pathology[55,56]. Similarly, MMP12, an enzyme involved in extracellular matrix degradation and inflammatory responses, was associated with both CAD and PAD[57,58]. The apoptosis-related ligand FASLG, a key mediator of extrinsic cell death and vascular inflammation, was important for stroke and VTE[59]. In addition, the neuronal damage marker NEFL, reflecting systemic neuroaxonal injury, was predictive of both stroke and PAD, suggesting a mechanistic link between neural injury and peripheral vascular events[60]. Beyond these shared pathways, we also found several disease-specific proteins, underscoring the distinct biology of each condition. Key specific associations included such as CRIP2, a key regulator of cellular responses to ischemia, with stroke[61], and PLB1, involved in metabolic regulation, with PAD[62]. Taken together, the identification of biomarkers from such diverse pathways—covering cardiac stress, inflammation, neuronal injury, and matrix

remodeling—suggests that our CardiOmicScore framework can capture a multi-system biological signature of cardiovascular risk that extends far beyond traditional risk factors.

These key molecules not only serve as potential targets for new therapeutic interventions to prevent or reverse CVD progression, but also illuminate druggable pathways for repurposing existing medications—such as anti-inflammatory colchicine[63]. Our findings also suggest that based on individual omics profiling, clinicians may be able to tailor treatments targeting the underlying biological mechanisms of specific CVDs. For example, patients with high levels of inflammation markers like GlycA may benefit from treatments aimed at reducing systemic inflammation, while those with elevated proteins like MMP12, which is involved in extracellular matrix degradation, could be targeted with therapies focused on preventing the vascular remodeling that lead to atherosclerosis. Notably, our study provides only preliminary evidence for the predictive importance of circulating biomarkers, and further large cohort studies and clinical trials are needed to validate our findings."

Methods Section, **Deep learning models** Part: "To ensure stable and reliable interpretations, we generated SHAP values for each of the five models trained with different random seeds. The final SHAP value for each biomarker was calculated by averaging these results across the five runs."

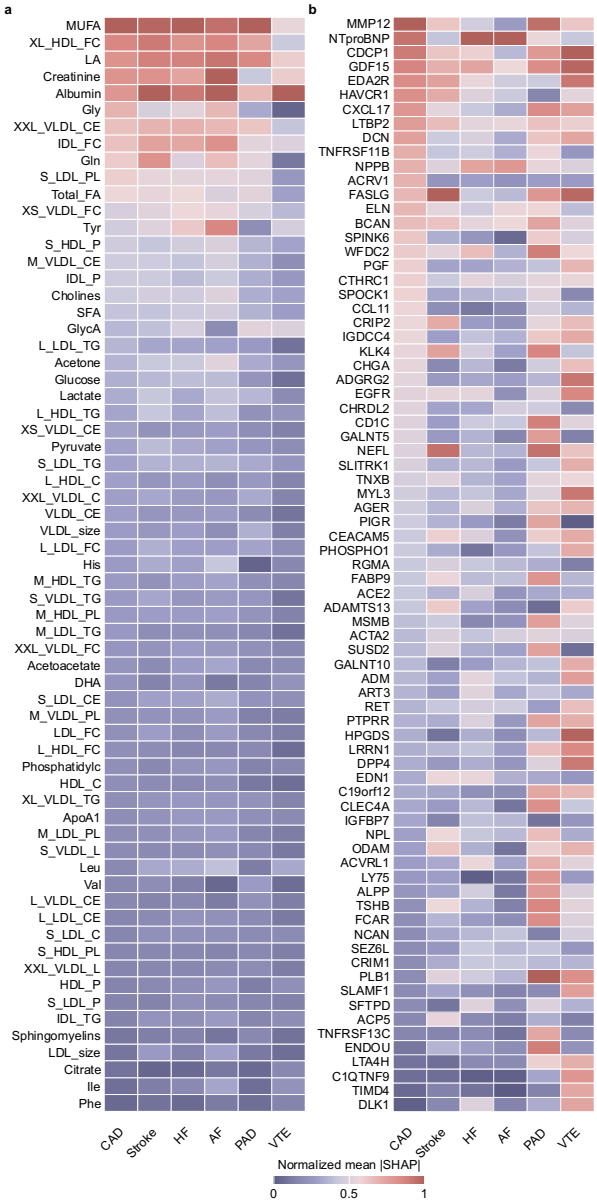


Figure 5: Importance of biomarkers in cardiovascular disease-specific MetScore (**a**) and ProScore (**b**). Importance was measured by mean absolute SHAP values. For each disease, the 50 most important metabolites and the 25 most important proteins were identified. The union of these features across all diseases yielded a total of 65 metabolites and 77 proteins.

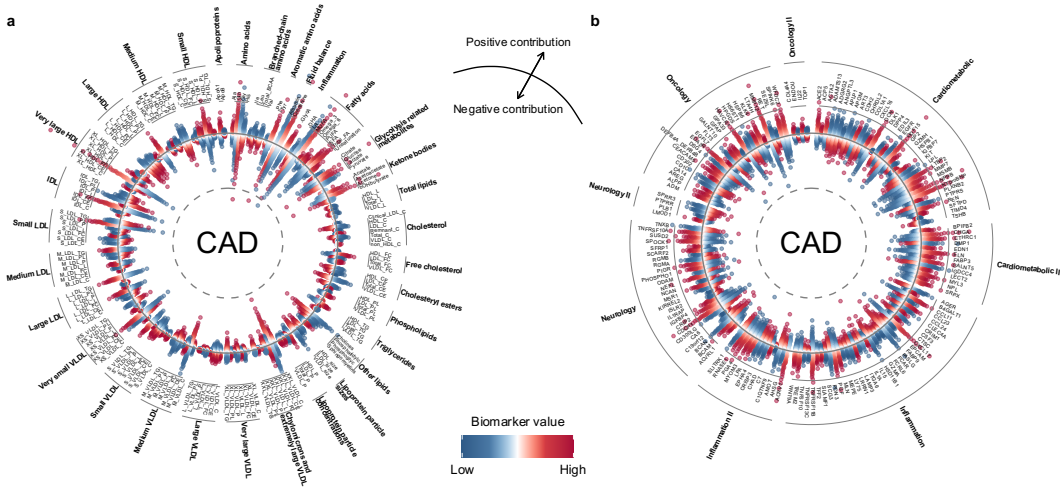


Figure 6: Metabolites (**a**) and proteins (**b**) are associated with cardiovascular disease risk. SHAP beeswarm plots are shown here for coronary artery disease as an example. Each individual is assigned a SHAP value for every protein and metabolite. To simplify the presentation, SHAP values for each biomarker were divided into percentiles, and the mean SHAP value and biomarker level were calculated for each percentile. Each dot represents one percentile. All 168 metabolites and the top-ranked 157 proteins are shown. Dark grey lines indicate zero SHAP value (i.e., no contribution). The farther a dot from this line, the stronger the absolute contribution to the MetScore or ProScore. Deviations toward the center and periphery represent negative and positive contributions. Red and blue colors indicate high and low levels of metabolites and proteins, respectively.

COMMENT 2-3. *Similarly, LDL metabolite fractions and total lipids / cholesterol are not strongly predictive of CAD (which is unusual in a healthy cohort, most of whom are not taking statins). Does this trend persist when the 20% of the cohort on lipid-lowering therapies are removed from the cohort? It would be helpful if the authors could comment on this observation and why this might be the case?*

RESPONSE 2-3. We thank the reviewer for the thoughtful comments. Following the reviewer's suggestion, we conducted a two-part sensitivity analysis among participants not taking lipid-lowering medication at baseline. We first refitted the MetNet model and re-calculated the SHAP-based feature importance rankings within this subgroup. Second, we specifically evaluated the predictive performance of traditional lipid-related biomarkers in CPH models for this subgroup and compared it to the full cohort. In the refitted MetNet model, we found that the top-ranked metabolites still included a mix of non-lipid and lipid markers (e.g., glycine, creatinine, GlycA) (**Supplementary Figure 18**), with seven of the top metabolites overlapping with our primary analysis, consistent with previous findings in the

UK Biobank (Buergel et al., 2022). Furthermore, while the predictive performance of some lipid markers for CAD did improve slightly in the participants not taking lipid-lowering medication, their added value to the comprehensive PANEL-based clinical model remained limited (**Supplementary Figures 19–20**). Taken together, these findings suggest that the confounding effect of lipid-lowering therapy is not the primary reason for the lower-than-expected ranking of traditional lipid markers in our model. Instead, it highlights the ability of our CardiOmicScore framework to capture more powerful predictive signals that go beyond what traditional lipid measures alone can offer.

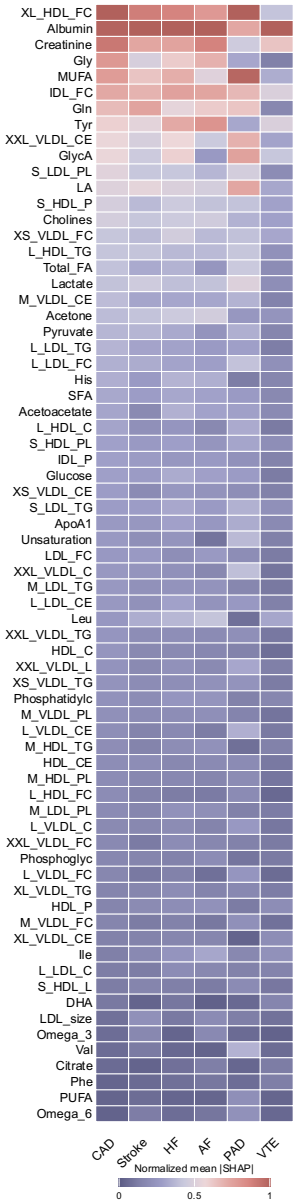
Results Section, **Sensitivity analyses** Part: "Finally, after refitting the MetNet model in the subgroup of participants not taking lipid-lowering medications, the most important metabolites included MUFA, glycine, creatinine, albumin, extra-large high-density lipoprotein free cholesterol, glutamine, IDL_FC, tyrosine, LA, and GlycA (Supplementary Figure 18). Seven of these metabolites overlapped with those highlighted in the main analysis. Furthermore, although the individual predictive performance of some traditional lipid-related biomarkers for CAD improved slightly after excluding participants on lipid-lowering medication, their addition still provided only limited improvement in predictive performance to the PANEL-based model. This finding was consistent in both the full Multiomics cohort and the subgroup excluding baseline users of lipid-lowering medication (Supplementary Figures 19–20)."

Discussion Section, **AI-identified biomarkers offer biological insights** Part: "Evaluating the contributions of important proteins and metabolites to each CVD reveals underlying disease mechanisms and potential therapeutic targets. We confirmed the role of proteins such as NT-proBNP and NPPB, as well as metabolites such as albumin and creatinine, in CVD risk assessment. These biomarkers are well-established in epidemiological studies and are already part of routine clinical care[47–50]. However, our analyses revealed that ProScore outperformed any single proteins such as NT-proBNP and NPPB both as an individual predictor and when added to clinical models. Similarly, individual metabolic biomarkers such as creatinine, albumin, and traditional lipid-related biomarkers (e.g., total lipids or cholesterol), yielded lower predictive performance compared to MetScore. These findings suggest that the MetScore and ProScore may provide a more comprehensive representation of disease-related molecular mechanisms than single biomarkers, and thus highlight the importance of high-throughput metabolomics and proteomics in personalized CVD risk

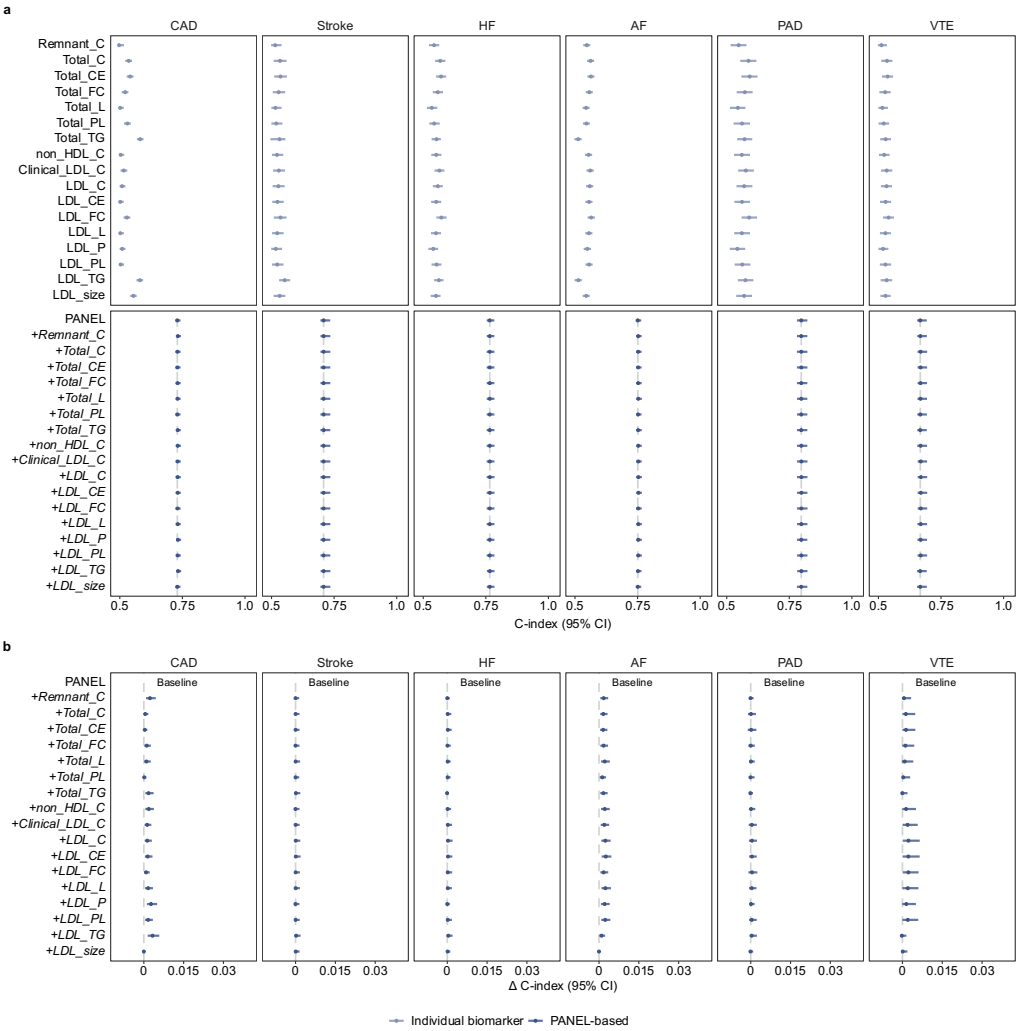
prediction. Moreover, our CardiOmicScore framework outperformed conventional machine learning algorithms, underscoring the unique strength of the multitask deep learning architecture to model complex, non-linear interactions in high-dimensional omics data."

Methods Section, **Statistical analyses** Part: "Finally, to investigate the potential effect of lipid-lowering treatment, we refitted the MetNet using the same architecture but trained only on participants not taking lipid-lowering medication. We then calculated the mean absolute SHAP values to identify the important metabolites within this subgroup. Additionally, using the same analytical procedure as our first sensitivity analysis, we evaluated the predictive performance of traditional lipid-related biomarkers in the Multiomics cohort after excluding baseline users of lipid-lowering medication, and compared it with their performance in the full set."

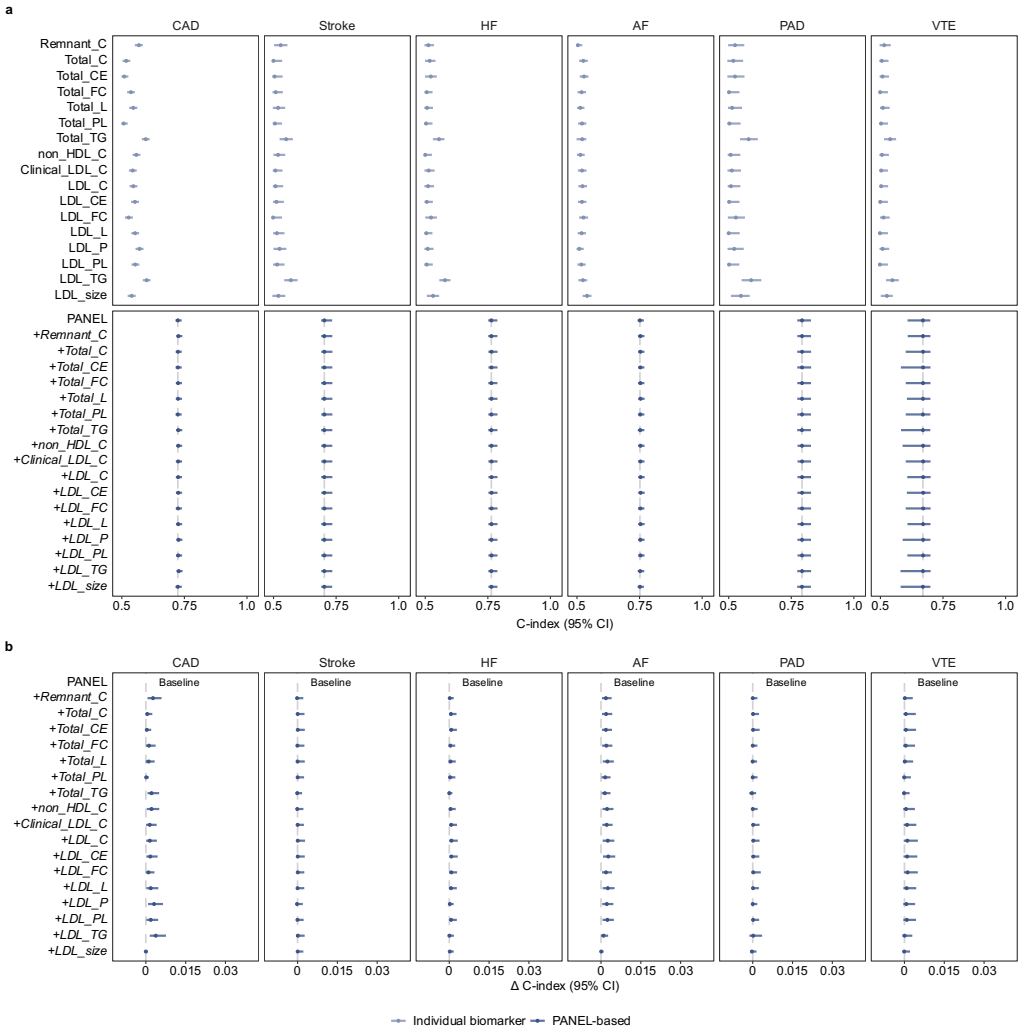
The added figures are copied below:



Supplementary Figure 18: Feature importance from MetNet refitted after excluding after excluding participants on lipid-lowering medication. The MetNet was refitted using an identical architecture on a subgroup of the Metabolomics-only cohort that excluded individuals on baseline lipid-lowering medication (N = 185,086). Feature importance was measured by calculating mean absolute SHAP values for the participants in the Multiomics cohort who were also not receiving lipid-lowering medication (N = 19,895). For each disease, the 50 most important metabolites were identified. The union of these features across all diseases yielded a total of 70 metabolites.



Supplementary Figure 19: Predictive performance of traditional lipid-related biomarkers for cardiovascular diseases. **a.** Discriminative performance for 17 lipid-related biomarkers. Vertical dashed lines indicate the performance of the PANEL set. Discriminative performance was evaluated by the Harrell's C-index. **b.** Differences in discriminative performance by adding each individual biomarker to the PANEL set. Data are presented as point estimates with 95% confidence intervals estimated from 1,000 bootstrap resamples.



Supplementary Figure 20: Predictive performance of traditional lipid-related biomarkers for cardiovascular diseases among participants not receiving lipid-lowering medication (*N* = 19,895). **a.** Discriminative performance for 17 lipid-related biomarkers. Vertical dashed lines indicate the performance of the PANEL set. Discriminative performance was evaluated by the Harrell's C-index. **b.** Differences in discriminative performance by adding each individual biomarker to the PANEL set. Data are presented as point estimates with 95% confidence intervals estimated from 1,000 bootstrap resamples.

COMMENT 2-4. *The authors report that "the genetic risk of CVDeath was measured using the PRS for CAD since there is no available GWAS summary statistics for CVDeath". Is there any reason the author's couldn't run their own CVDeath GWAS rather than applying the CAD GWAS results to a non-identical outcome.*

RESPONSE 2-4. Thank you very much for this important methodological point. We refined our analytical framework to focus on six common cardiovascular diseases (coronary artery disease, stroke, heart failure, atrial fibrillation, peripheral artery disease, and venous

thromboembolism) and removed CVDeath from our primary endpoints. Running a GWAS for CVDeath within our cohort was not feasible due to the limited number of events, which would result in insufficient power and unstable estimates. Similarly, ventricular arrhythmias and abdominal aortic aneurysm were excluded given their low incidence (0.8% and 0.4%, respectively). To still account for the influence of mortality, we performed competing risk analyses using Fine-Gray subdistribution hazard models for each of the six CVDs, treating all-cause death as a competing event, which confirmed that the predictive value of the omics scores was consistent with our main findings (**Supplementary Figure 16**). Finally, we added the PGS Catalog IDs for all six PRS to the **Methods** section to enhance reproducibility. The revised manuscript contains the following change:

Introduction Section, **Fifth** Paragraph: "Unlike traditional models, both MetNet and ProNet utilize a multitask architecture to learn high-dimensional representations from metabolomics and proteomics data, generating disease-specific risk scores, termed MetScore and ProScore, for a comprehensive range of CVDs. These CVD include coronary artery disease (CAD), stroke, heart failure (HF), atrial fibrillation (AF), peripheral artery disease (PAD), and venous thromboembolism (VTE)—conditions with the highest global burden[1]."

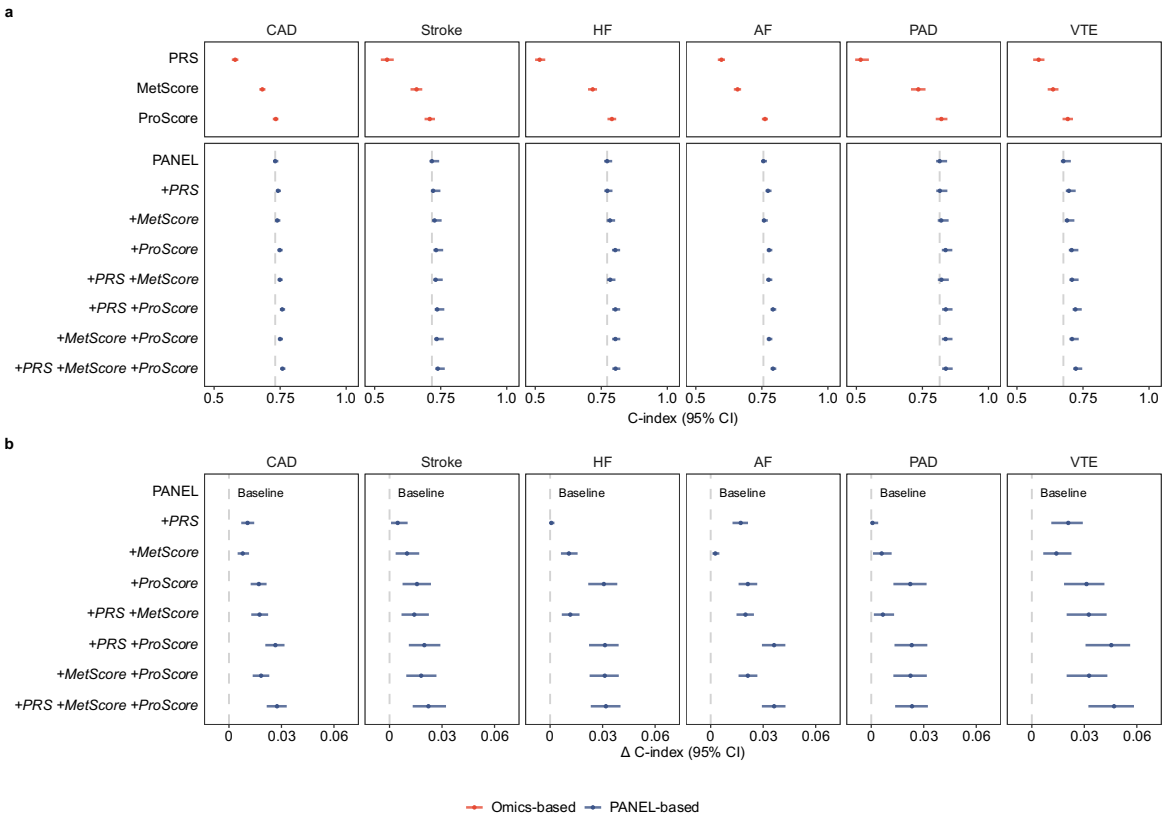
Results Section, **Statistical analyses** Part: "We also found that the predictive value of the omics scores remained unchanged after excluding incident cases that occurred within the first two years of follow-up (Supplementary Figure 15) or accounting for the competing risk of death (Supplementary Figure 16)."

Methods Section, **Ascertainment of cardiovascular diseases** Part: "The six CVDs analyzed in this study included coronary artery disease (CAD), stroke, heart failure (HF), atrial fibrillation (AF), peripheral artery disease (PAD), and venous thromboembolism (VTE)."

Methods Section, **Multiomics data** Part: "Effect sizes of SNP-disease associations were collected from published PRSs available in the PGS Catalog[13,69–73]. We included 235 SNPs for CAD (PGS003438)[13], 63 SNPs for stroke (PGS005230)[69], 38 SNPs for HF (PGS003969)[70], 154 SNPs for AF (PGS004905)[71], 19 SNPs for PAD (PGS005158)[72], and 297 SNPs for VTE (PGS000753)[73]. Full lists of SNPs are provided in Supplementary Tables 6–11."

Methods Section, **Statistical analyses** Part: "Fourth, to account for the competing risk of death, we fitted Fine-Gray subdistribution hazard models, treating all-cause death as a competing event (cmprsk v2.2.11 R package)."

The added figure is copied below:



Supplementary Figure 16: Predictive performance of multiomics for cardiovascular diseases after considering competing risk of death. **a.** Discriminative performance trained on various predictor sets. Vertical dashed lines indicate the performance of the PANEL set. Discriminative performance was evaluated by the Harrell's C-index. **b.** Differences in discriminative performance by adding omics information to the PANEL set. All-cause death was treated as a competing event. The number of primary (CVD-specific) and competing (death) events were as follows: CAD (1,788 primary, 1,592 competing), stroke (584 primary, 1,791 competing), HF (798 primary, 1,694 competing), AF (1,524 primary, 1,627 competing), PAD (324 primary, 1,879 competing), and VTE (679 primary, 1,727 competing). Data are presented as point estimates with 95% confidence intervals estimated from 1,000 bootstrap resamples.

COMMENT 2-5. *The authors describe the demographics of the whole cohort but not the demographics of the individual training and validation cohorts. Could the authors please provide these along with statistical testing to show that such cohorts are not statistically different. This is especially important given the absence of an external validation cohort.*

RESPONSE 2-5. Thank you for this valuable suggestion to provide a more detailed comparison of the study cohorts. We performed the requested analyses and added the results to the manuscript (**Supplementary Tables 1–3**, **Supplementary Figure 2**). The statistical comparisons confirm the need of our splitting strategy. As expected, there were no significant differences in baseline characteristics between the training and validation sets. In contrast, we observed significant differences between the training and geographic testing sets. This heterogeneity was by design, as the geographic split (England/Wales vs. Scotland) was intended to provide a robust test of our models' regional generalizability before the final evaluation. We have copied the relevant methods and results below (**Supplementary Tables 2–3** are omitted due to their length):

In the main text:

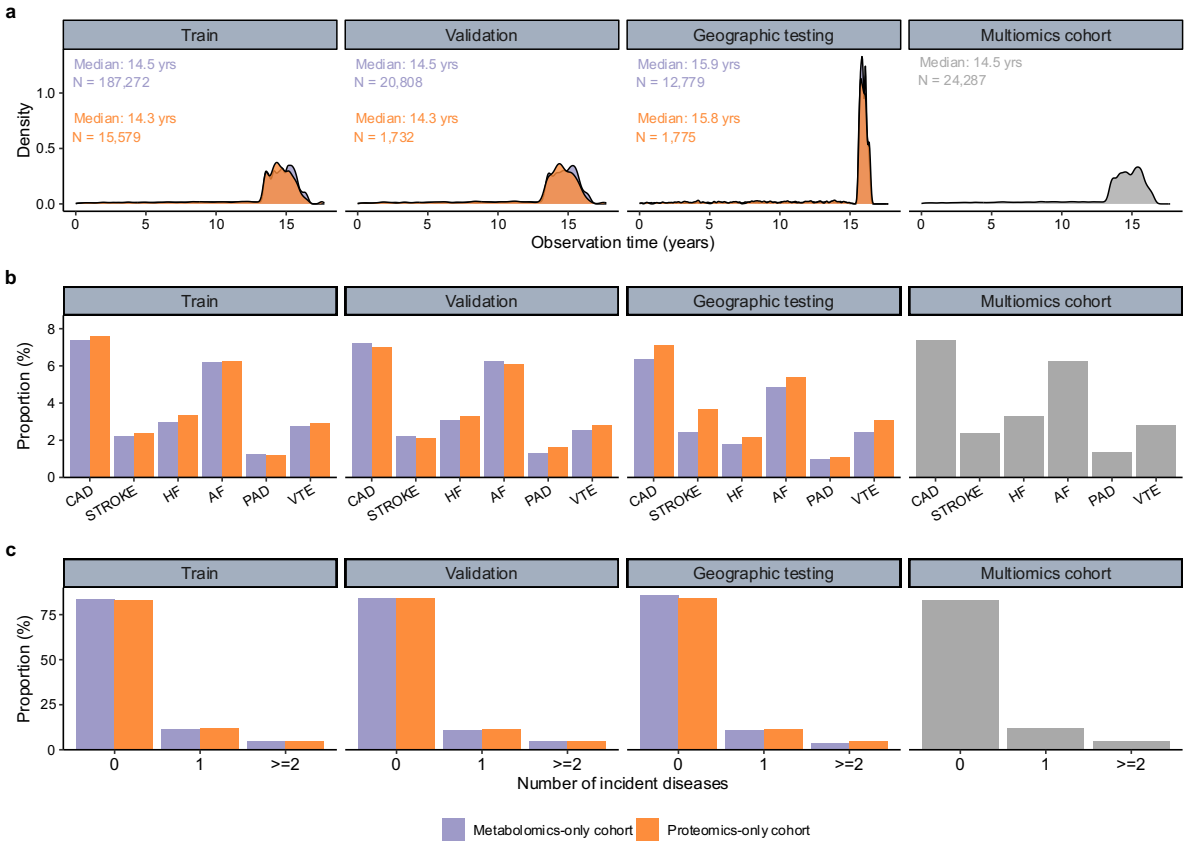
Results Section, **Study population** Part: "A two-stage, individual-level data partitioning strategy was employed to rigorously evaluate our model's performance and generalizability (see sub-section 4.2 and Supplementary Figure 1). We first divided participants into the Metabolomics-only (N = 220,859), Proteomics-only (N = 19,086), and Multiomics (N = 24,287) cohorts based on omics availability. The Metabolomics-only and Proteomics-only cohorts were used within our CardiOmicScore framework to develop two deep learning models, MetNet and ProNet, respectively. The Multiomics cohort was reserved as an untouched validation set to assess the added predictive value of integrating different omics data.

Following best practices in recent UKB studies[39–41], each of these development cohorts was further split into training/validation sets (England and Wales) and a geographic testing set (Scotland). Specifically, the Metabolomics-only cohort included 187,272 participants in the training set, 20,808 in the validation set, and 12,779 in the geographic testing set, while the Proteomics-only cohort comprised 15,579, 1,732, and 1,775 participants in the training, validation, and geographic testing sets, respectively. Baseline characteristics were comparable between training and validation sets in both cohorts, whereas significant differences were observed between training and geographic testing sets (Supplementary Tables 2–3), thus allowing a robust assessment of the models' regional generalizability.

Baseline characteristics and incident cases during follow-up were broadly consistent across all study datasets. The median age of participants at baseline was between 56.0 and 58.0 years. The proportion of males ranged from 42.8% to 45.8%. The prevalence of baseline medication use was comparable, with 15.8% to 18.1% of participants receiving lipid-lowering medication and 10.5% to 11.2% on antihypertensive medication. The distribution of outcomes was also similar, with 83.2–85.7% of participants remaining free of any incident CVD, 10.6–11.8% developing one CVD, and 3.7–5.0% developing multiple CVDs. Detailed baseline characteristics and follow-up information for each cohort are provided in Supplementary Tables 1–3 and Supplementary Figure 2."

Methods Section, **Statistical analyses** Part: "Baseline characteristics were summarized using median (IQR) for continuous variables or numbers (percentages) for categorical variables. Chi-squared tests were used for categorical variables, and Mann–Whitney U tests were used for non-normally distributed continuous variables to compare baseline characteristics between the training and validation sets, and between the training and geographic testing sets."

The updated figures and tables are copied below:



Supplementary Figure 2: Follow-up duration and distribution of incident cardiovascular disease (CVD) cases across datasets. **a.** Distribution of follow-up duration. Here, we defined the follow-up duration as the period from the date of baseline assessment to the date of any incident cardiovascular diseases, death, loss to follow-up, or end of available registry follow-up (November 30 2023 for England & Wales and December 31 2023 for Scotland), whichever came first. Text annotations provide the median observation time and the total number of participants for each dataset. **b.** Proportion of participants with incident cases of each CVD. **c.** Proportion of participants grouped by the number of incident CVD.

Supplementary Table 1: Baseline characteristics of participants in the Multiomics cohort.

Characteristics[1]	Multiomics cohort (*N* = 24,287)
Demographic information	
Age (years)	58.0 (50.0, 63.0)
Male	10,666 (43.9%)
Ethnicity	
White	22,735 (93.7%)
Black	521 (2.1%)
Asian	421 (1.7%)
Others	583 (2.4%)
Townsend deprivation index	-2.1 (-3.7, 0.5)
Healthy lifestyles	
Current smoking	2,467 (10.2%)
Daily alcohol intake	4,876 (20.1%)
Healthy sleep	17,785 (73.3%)
Physical activity	15,259 (81.3%)
Healthy diet	13,775 (56.8%)

Social connection	22,124 (91.2%)
Family disease history	
Family history of heart disease	10,179 (41.9%)
Family history of stroke	6,361 (26.2%)
Family history of hypertension	11,551 (47.6%)
Family history of diabetes	5,236 (21.6%)
Disease and medication history	
Hypertension	6,446 (26.5%)
Diabetes	1,161 (4.8%)
Lipid-lowering medication	4,392 (18.1%)
Antihypertensive medication	2,717 (11.2%)
Physical measurements	
Systolic blood pressure (mmHg)	136.5 (124.5, 149.5)
Diastolic blood pressure (mmHg)	82.0 (75.5, 89.0)
Height (cm)	167.5 (161.0, 175.0)
Weight (kg)	75.7 (66.0, 86.4)
Waist circumference (cm)	89.0 (80.0, 98.0)
Waist-hip ratio	0.9 (0.8, 0.9)
Body mass index (kg/m^2)	26.6 (24.1, 29.6)
Blood count	
Leukocyte count (10^9 cells/Litre)	6.6 (5.6, 7.8)
Lymphocyte count (10^9 cells/Litre)	1.9 (1.5, 2.3)
Monocyte count (10^9 cells/Litre)	0.5 (0.4, 0.6)
Neutrophill count (10^9 cells/Litre)	4.0 (3.2, 4.9)
Eosinophill count (10^9 cells/Litre)	0.1 (0.1, 0.2)
Basophill count (10^9 cells/Litre)	0.02 (0.00, 0.04)
Platelet count (10^9 cells/Litre)	249.3 (214.8, 288.7)
Haemoglobin concentration (grams/decilitre)	14.1 (13.3, 15.0)
Haematocrit percentage (%)	40.9 (38.6, 43.4)
Cardiovascular disease	
Coronary artery disease	1,795 (7.4%)
Stroke	584 (2.4%)
Heart failure	800 (3.3%)
Atrial fibrillation	1,525 (6.3%)
Peripheral artery disease	325 (1.3%)
Venous thromboembolism	679 (2.8%)

[1]Data are presented as median (interquartile range) or frequency (%).

COMMENT 2-6. *On trying to access the online demo of the model, the weblink appeared to be broken. Please could the authors provide a working link to allow reviewers and readers to access the online demo.*

RESPONSE 2-6. We thank the reviewer for identifying the issue with the weblink to our online demo and apologize for the inconvenience. The previous hosting platform experienced a technical issue that was beyond our control. To ensure stable, long-term access, we have now migrated the online demo to GitHub Pages. The restored link is now available at:

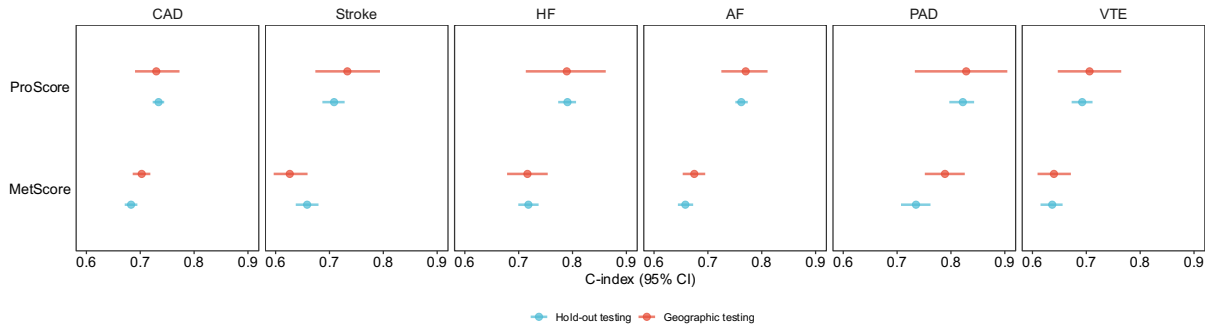
https://yanluocityu.github.io/cardiomicscore-website/. The manuscript has been updated with this new address.

Data and code availability Section: "The data was obtained from UK Biobank (https://www.ukbiobank.ac.uk/). Code associated with our paper is available at https://github.com/YanLuoCityU/cardiomicscore. An online demo is available at https://yanluocityu.github.io/cardiomicscore-website/.

COMMENT 2-7. *The main issue with this work is that is used 10% of the cohort for validation, so this is considered internal (not external) validation, and the models are expected to suffer from overfitting.*

RESPONSE 2-7. Thank you so much for the comment. To provide strong evidence against overfitting and to rigorously assess generalizability, we implemented a two-stage individual-level data partition strategy, which is designed to approximate the conditions of external validation (also refer to **RESPONSE 2-1**). First, we tested our models on a geographic testing set (participants from Scotland), which was completely separate from our training data (from England and Wales). Subsequently, we performed a final evaluation on a completely untouched Multiomics cohort to demonstrate the improvement in predictive performance by integrating different omics data types. Although the training and geographic testing sets exhibited significant differences in several clinical characteristics, both MetScore and ProScore achieved strong predictive performance, and this performance was highly consistent with the results from the Multiomics cohort, providing robust evidence for their generalizability (**Supplementary Figure 6**).

We copied the Supplementary Figure 6 below:



Supplementary Figure 6: Discriminative performance of MetScore and ProScore. Discriminative performance was evaluated separately for MetScore and ProScore in both the geographic testing and the Multiomics cohort. The Multiomics cohort included 24,287 participants, while the geographic testing sets included 12,779 for MetScore and 1,775 for ProScore. Performance was assessed using Harrell's C-index with 95% confidence intervals estimated by 1,000 bootstrap resamples.

COMMENT 2-8. *In the demographics it would also be helpful to know the numbers of participants on statin / antihypertensive therapies.*

RESPONSE 2-8. Thank you for this helpful suggestion. We added the proportion of participants on lipid-lowering and antihypertensive therapies to the **Result** section: "The prevalence of baseline medication use was comparable, with 15.8% to 18.1% of participants receiving lipid-lowering medication and 10.5% to 11.2% on antihypertensive medication."

COMMENT 2-9. *Minor grammatical error in supplementary figure 1 caption, lines 115-116. "Here, we defined the follow-up duration as the period from the date of baseline assessment to the date of incident any cardiovascular diseases…" doesn't make sense grammatically. Perhaps the intended line is "the date of any incident cardiovascular diseases".*

RESPONSE 2-9. Thank you so much for pointing out this error. We corrected the sentence in the figure caption, and the figure is now Supplementary Figure 2. The revised sentence is: "Here, we defined the follow-up duration as the period from the date of baseline assessment to the date of any incident cardiovascular diseases, death, loss to follow-up, or end of available registry follow-up (November 30 2023 for England & Wales and December 31 2023 for Scotland), whichever came first."

COMMENT 2-10. *Minor grammatical error in 2.3, paragraph 2, line 2: "(i.e., 95% CIs of delta C-index not included zero)". "Included" should be changed to "including".*

RESPONSE 2-10. We thank the reviewer for spotting this error. We have corrected the sentence in the **Methods** section, changing "included" to "including".

COMMENT 2-11. *Supplementary figure 10 appears to have a light blue figure legend marker for 'clinical scores' which seems to be redundant as it does not feature in the figure outside the legend.*

RESPONSE 2-11. Thank you for the suggestion. We removed the redundant "clinical scores" marker from the figure legend as requested. The revised figure is shown below:

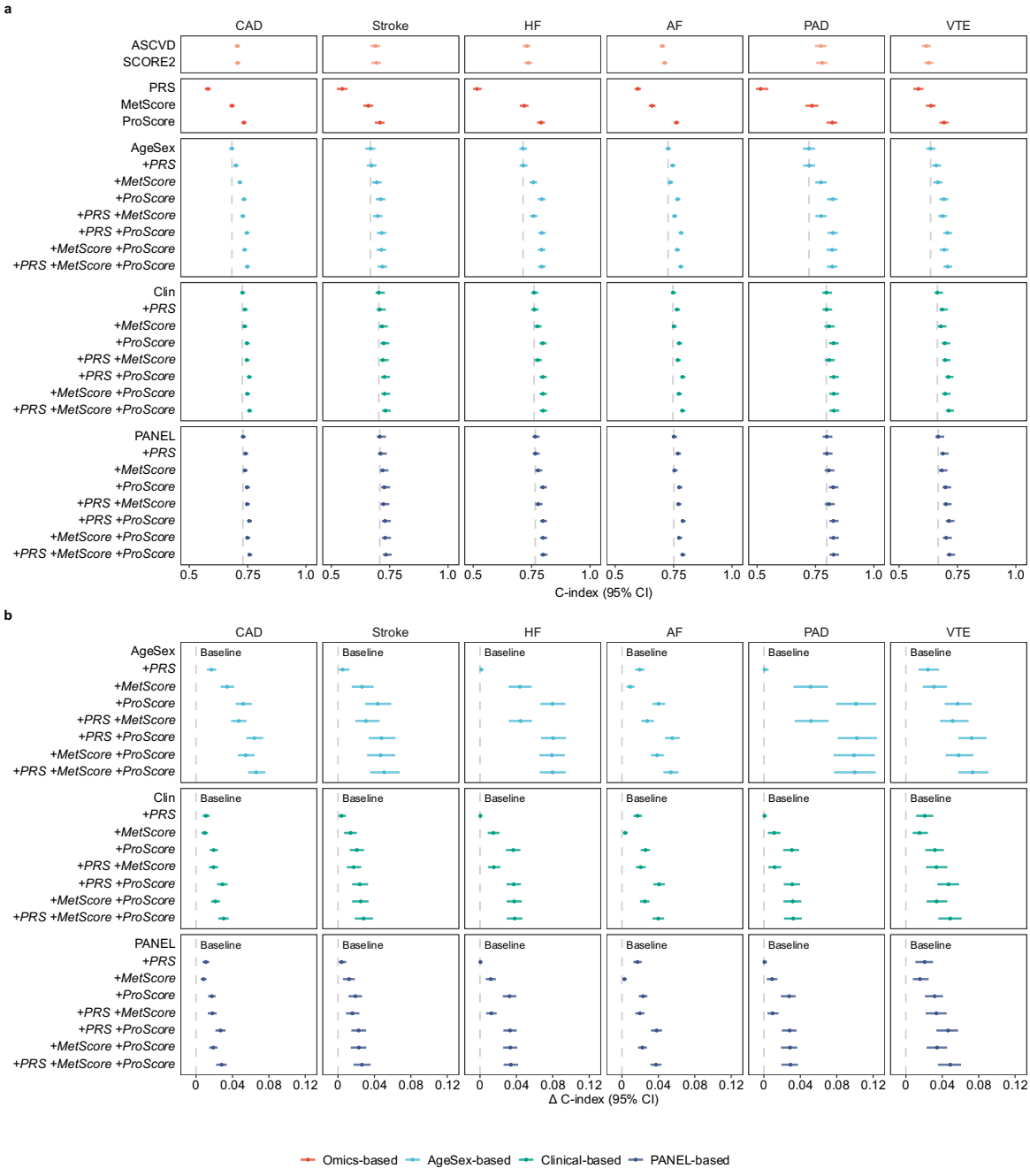


Figure 3: Predictive performance of multiomics for cardiovascular diseases. **a.** Discriminative performance of models trained on various predictor sets. Vertical dashed lines indicate the performance of the three clinical predictor sets. Discriminative performance was evaluated by the Harrell's C-index. **b.** Differences in discriminative performance by adding omics information to clinical predictor sets. Data are presented as point estimates with 95% confidence intervals estimated from 1,000 bootstrap resamples.

Dear Editor and Reviewers,

Thank you so much for taking the time to review this manuscript. We really appreciate all your comments and suggestions! We hope that the revised manuscript could meet the requirements for Nature Communications. All revised portions are marked in red in the revised manuscript.

The main comments and our specific responses are detailed below.

REVIEWER 2'S COMMENTS

Remarks to the Author:

COMMENT 2-1. *The authors have made some reasonable changes to the manuscript generally. The inclusion of a geographically separate cohort, whilst an improvement on the original study design, still does not offer quite the same level of robust validation as a true external validation cohort. On balance, I feel that the manuscript has sufficient merrit to warrant publication of the existing material with the following further revisions:*

The manuscript states, under the limitations section, 'Finally, since the majority of UKB population are European ancestry and white ethnicity, the generalizability of our findings needs to be validated in ethnically diverse populations to ensure broader applicability. Please would the authors kindely submit a revision that mentions explicitly the lack of a proper external validation cohort (rather than just citing a need for ethnic diverse populations).

RESPONSE 2-1. Thank you very much for the positive feedback. We fully agree that a geographically separated cohort does not provide the same level of robustness as a true independent external validation cohort. Following your recommendation, we have revised the limitations section to explicitly acknowledge the lack of a proper external validation cohort. We copied the updated parts in the main text below:

Discussion Section, **Twelfth** Paragraph: "Finally, although we conducted geographic validation to assess model generalizability, our study still lacks an independent external validation."

COMMENT 2-2. *The authors discuss the clinical utility and application of the model in their discussion section. As some point in this discussion it should be clearly stated that the signatures are not clinically generalisable or applicable until they have undergone external validation.*

RESPONSE 2-2. We thank the reviewer for this important and insightful comment. Following your suggestion, we have revised the Discussion section to clearly state that the identified molecular signatures are not clinically generalizable or applicable until external validation has been performed. The updated paragraph is copied below:

Discussion Section, **Tenth** Paragraph: "Notably, our study provides only preliminary evidence for the predictive importance of circulating biomarkers; therefore, it must be clearly stated that these signatures are not clinically generalizable or applicable until they have undergone external validation."