Manifold Fitting Reveals Metabolomic Heterogeneity and Disease Associations in UK Biobank Populations

Bingjie Li[a,1], **Jiaji Su**[a,1], **Runyu Lin**[a,1], **Shing-Tung Yau**[b,2], **and Zhigang Yao**[a,2]

This manuscript was compiled on April 18, 2025

Nuclear magnetic resonance (NMR)-based metabolic biomarkers provide comprehensive insights into human metabolism; however, extracting biologically meaningful patterns from such high-dimensional data remains a significant challenge. In this study, we propose a manifold-fitting-based framework to analyze metabolic heterogeneity within the UK Biobank population, utilizing measurements of 251 NMR biomarkers from 212,853 participants. Initially, our method clusters these biomarkers into seven distinct metabolic categories that reflect the modular organization of human metabolism. Subsequent manifold fitting to each category unveils underlying low-dimensional structures, elucidating fundamental variations from basic energy metabolism to hormone-mediated regulation. Importantly, three of these manifolds clearly stratify the population, identifying subgroups with distinct metabolic profiles and associated disease risks. These subgroups exhibit consistent links with specific diseases, including severe metabolic dysregulation and its complications, as well as cardiovascular and autoimmune conditions, highlighting the intricate relationship between metabolic states and disease susceptibility. Supported by strong correlations with demographic factors, clinical measurements, and lifestyle variables, these findings validate the biological relevance of the identified manifolds. By utilizing a geometrically informed approach to dissect metabolic heterogeneity, our framework enhances the accuracy of population stratification and deepens our understanding of metabolic health, potentially guiding personalized interventions and preventive healthcare strategies.

metabolic manifolds | geometric decomposition | manifold fitting | disease risk prediction | blood metabolomics | population heterogeneity

Nuclear magnetic resonance (NMR)-based metabolomics is transforming our understanding of human metabolic health by enabling the high-throughput, simultaneous quantification of a broad spectrum of circulating metabolites—including lipids, amino acids, and glycolysis-related compounds—at a population scale. This approach provides a holistic and cost-effective snapshot of systemic metabolism, reflecting both genetic and environmental influences (1). Such NMR-derived metabolic signatures have been shown to correlate with a wide array of clinical outcomes. For instance, specific plasma metabolite profiles have been linked to early atherosclerotic changes in subclinical cardiovascular disease, thereby supporting earlier interventions and improved risk stratification (2). In addition, metabolic biomarkers identified through plasma profiling have enhanced the prediction of diabetic complications such as diabetic retinopathy, informing targeted screening protocols and individualized patient management (3).

The UK Biobank has integrated extensive phenotypic and genetic data on over half a million participants, and ongoing initiatives have expanded these resources to include NMR-based metabolic profiling (4). With these data in hand, researchers have the opportunity to move beyond studying individual biomarkers toward a deeper understanding of metabolic heterogeneity within large, diverse populations. A more comprehensive perspective on biomarkers enables the characterization of distinct metabolic categories and the exploration of how genetic, lifestyle, and environmental factors shape metabolic profiles. By harnessing these large-scale metabolomic datasets, it is possible to uncover potential new disease mechanisms, refine disease subtyping, and improve the precision of risk prediction models. In this way, NMR metabolomics can guide clinical decision-making and public health strategies—identifying individuals at elevated risk for specific diseases long before clinical symptoms manifest, and pointing toward targeted lifestyle or therapeutic interventions.

Significance Statement

This study is the first to utilize a manifold-fitting framework within NMR-based metabolomics to explore metabolic heterogeneity in the UK Biobank population. Our method clusters 251 metabolic biomarkers into seven distinct categories that reflect the modular organization of human metabolism. Applying manifold fitting reveals low-dimensional structures in each category, capturing crucial metabolic variations associated with diverse disease risks. Notably, fitted manifolds in three categories distinctly stratify the population, each identifying two subgroups with unique metabolic profiles linked to a broad spectrum of diseases, from metabolic complications to cardiovascular and autoimmune disorders. This nuanced stratification enhances our understanding of the interactions between metabolism and disease, potentially guiding personalized health interventions and advancing preventive medicine strategies.

Author affiliations: [a]Department of Statistics and Data Science, National University of Singapore, Singapore 117546, Singapore; [b]Yau Mathematical Sciences Center, Jingzhai, Tsinghua University, Beijing 100084, China

Z.Y. designed research; B.L., J.S., R.L. and Z.Y. performed research; B.L., J.S., and R.L. analyzed data; Z.Y. contributed research methodology; S.-T.Y. supervised the project; and B.L., J.S., R.L., S.-T.Y. and Z.Y. wrote the paper.

The authors declare no competing interest.

[1]B.L., J.S., and R.L. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: styau@tsinghua.edu.cn or zhigang.yao@nus.edu.sg.

Emerging research leveraging NMR-derived metabolic biomarkers has taken two primary analytical directions. On the one hand, population-level studies have been focused on building predictive and stratification models—often using supervised statistical and machine learning approaches—to estimate disease risks or classify individuals based on known outcomes (5). On the other hand, unsupervised techniques, such as clustering or factorization methods, have been employed to dissect metabolic heterogeneity by identifying latent subgroups with distinct biochemical profiles, independent of predefined clinical labels (6). Although these methods have provided valuable insights, several limitations remain. Many commonly used analytic techniques rely on linear assumptions or simple dimension reduction tools (e.g., principal component analysis), which may fail to capture the complex, nonlinear relationships that characterize high-dimensional metabolic data. Addressing these shortcomings will require adopting more sophisticated models capable of navigating nonlinear solution spaces, such as kernel-based methods (7), deep learning architectures (8), or advanced dimension reduction techniques (9–11), thereby enabling a more nuanced understanding of metabolic heterogeneity and its links to health and disease.

Despite the high-dimensional nature of NMR-based metabolomic measurements—which can encompass hundreds of metabolites—the underlying biochemical pathways governing these metabolites are constrained by the organism's metabolic pathways and regulatory networks (12–14). As a result, the observed metabolic variation likely resides on a lower-dimensional manifold embedded within the high-dimensional space (15, 16), making metabolomic data an ideal candidate for manifold fitting approaches (17–19). Indeed, manifold fitting, an invention that estimates the underlying manifold, has already demonstrated its capacity to capture the intrinsic geometry of complex, high-dimensional datasets, as evidenced by its successful application to modeling nonlinear data structures (20) and enhancing single-cell RNA sequencing analysis through improved clustering and visualization (21).

A key advantage of manifold fitting for metabolomic data analysis lies in its ability to reconstruct a smooth manifold directly in the ambient measurement space, thereby retaining all metabolically relevant information while filtering out measurement noise. The complex, nonlinear relationships between metabolites—shaped by substrate-product transformations, regulatory feedback loops, and pathway cross-talk—are likely to form a smooth manifold that could capture the key underlying degrees of freedom in cellular metabolism (22). Unlike traditional dimension reduction techniques, which often rely on transformations that risk losing information, modern manifold fitting methods (19–21) operate directly in the original feature space. Their flexible neighborhood definitions enable them to adapt to diverse metabolite distributions, faithfully representing the complexity and nuance of metabolic networks.

The aim of this study is to develop a comprehensive analytical method to elucidate population-level metabolic heterogeneity from multiple perspectives. Our approach builds on the manifold-fitting framework, capitalizing on the intrinsic properties of metabolic biomarkers. Given that metabolic biomarkers naturally form distinct categories–each characterized by coordinated patterns of variation reflecting specific biological processes and associated disease risks–we first implement an unsupervised clustering approach to divide the metabolites into seven categories based on their biological relevance and coordination patterns. This categorization enables a modular understanding of human metabolism, from basic energy metabolism to complex lipoprotein regulation. Applying manifold fitting to these categories reveals underlying low-dimensional structures that remained consistent across the population. Remarkably, we identify distinct population stratification patterns that demonstrate strong associations with disease risks and health outcomes. Through comprehensive analysis of these metabolically defined subgroups, we uncover not only their unique disease susceptibility profiles but also differential responses to lifestyle factors. This integrative approach allows us to bridge the gap between metabolic patterns and actionable health management strategies, particularly for high-risk populations. Specifically, by examining how lifestyle factors modify disease risks in metabolically vulnerable subgroups, we provide insights into targeted intervention strategies. To our knowledge, this work represents the first application of manifold fitting to large-scale metabolomic data, setting the stage for more precise population stratification while enabling a deeper understanding of the interplay between metabolic profiles, lifestyle factors, and disease risks that can guide personalized prevention efforts.

Results

A brief overview of our framework. We investigate metabolic heterogeneity in a large-scale population cohort ($n = 212,853$) from the UK Biobank with 251 NMR-measured metabolic biomarkers and examine the associations of these biomarkers with lifestyle factors and clinical outcomes. Our analytical framework comprises four sequential phases: metabolic biomarker clustering, manifold fitting for each biomarker category, heterogeneity visualization, and characterization of metabolically distinct subgroups in relation to health outcomes and lifestyle factors.

Following data acquisition, an unsupervised clustering approach is implemented for the 251 metabolic biomarkers (**Fig. 1A**). This clustering strategy is based on the biological premise that metabolic pathways exhibit varying degrees of interconnectedness, with some pathways exhibiting strong regulatory coupling while others operate independently (23). Hence, identifying categories of highly interconnected metabolites while maintaining inter-category independence will enhance the detection of underlying manifold structures. The number of metabolite categories is optimized through silhouette coefficient maximization, yielding seven distinct metabolic categories, which are denoted as **C1**–**C7** for simplicity (*Materials and Methods*). The composition of these categories are detailed in subsequent analyses. This modular decomposition of the metabolome enables multi-dimensional characterization of population heterogeneity across distinct metabolic pathways, facilitating the identification of new disease associations through pathway-specific analyses.

Subsequently, manifold fitting is performed on each metabolic category to explore their underlying low-dimensional structures (**Fig. 1B**), resulting in seven distinct manifolds, which are denoted as **M1**–**M7** correspondingly (*Materials and Methods*). Our objective is to identify principal

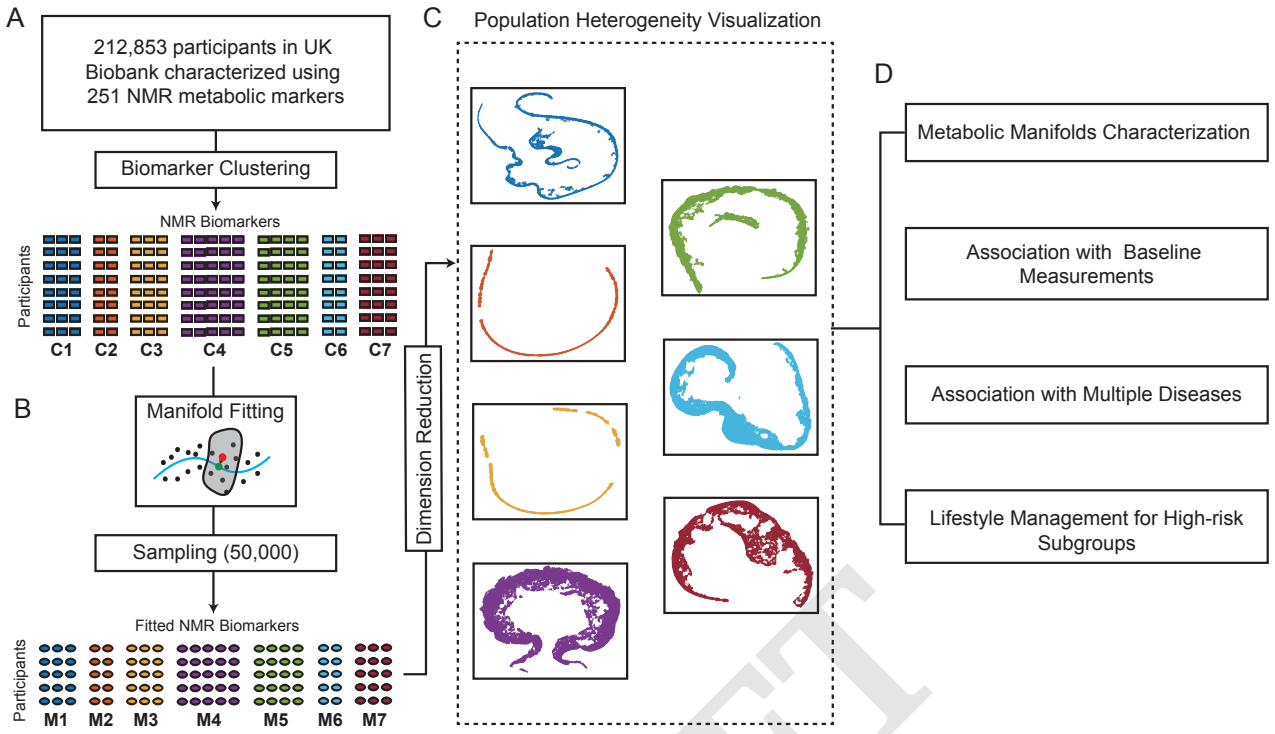

Fig. 1. Manifold-fitting-based framework for metabolic profiling and population heterogeneity in UK Biobank. (A) A total of 212,853 UK Biobank participants are characterized using 251 metabolic biomarkers. These biomarkers are divided into seven categories (**C1**–**C7**) based on their population-level associations. (B) Manifold fitting is applied to identify the underlying intrinsic structures. Seven manifolds (**M1**–**M7**) are extracted from each of the seven categories. (C) Population heterogeneity is visualized using UMAP dimension reduction. The visualization reveals distinct patterns across all seven manifolds. (D) Population health profiles are analyzed based on the heterogeneity discovered from the manifolds. These analyses encompass metabolic manifolds characterization, association with baseline measurements, association with multiple diseases, and lifestyle management for high-risk subgroups.

metabolic variations within the population while mitigating confounding variability that could be classified as noise. From a biological perspective, this noise stems from multiple sources: not only inherent technical measurement errors in the detection methodology and fluctuations in time-sensitive biomarkers induced by transient lifestyle factors during the measurement process (24), but also intrinsic variability from diverse complex biological processes within organisms, such as stochastic fluctuations in cellular signaling pathways, randomness in gene expression, molecular-level stochastic events, and inter-individual biological differences (25). Despite these multi-source noises, the manifold fitting procedure is able to minimize their impact on population heterogeneity characterization, leading to data with more pronounced structural features and well-defined distributions. The manifold fitting procedure reduces the impact of noise on population heterogeneity characterization, leading to more pronounced data structures and well-defined distributions.

After applying manifold fitting, we randomly select a subset of 50,000 participants for further dimension reduction and visualization using uniform manifold approximation and projection (UMAP, (11)). This process generates two-dimensional embeddings, resulting in seven distinct projections that capture population heterogeneity across various metabolic domains (**Fig. 1C**). Four of these projections (**M1**, **M2**, **M3**, **M5**) exhibit remarkable topological discontinuities, manifesting as well-defined, discrete population substructures. Subsequent density-based clustering analysis of the reduced-dimensional representations reveals robust population stratification into binary subgroups within **M1**, **M2**, and **M5**. The remaining metabolic manifolds demonstrate quasi-arc trajectories, suggesting continuous phenotypic variation along metabolic axes. The projection results are more informative compared to dimension reduction result of neural network based methods (see for *SI Appendix* Fig. S9).

The population heterogeneity can be applied to a variety of downstream health analyses (**Fig. 1D**), such as characterization of distinct metabolic manifold structures to establish metabolic reference states, investigation of associations with baseline clinical measurements to validate biological relevance, comprehensive analysis of relationships with multiple disease states to identify potential metabolic risk factors, and development of targeted lifestyle management strategies for identified high-risk subgroups (*Materials and Methods*). This systematic analytical framework enables a multi-perspective understanding of how metabolic heterogeneity intersects with health outcomes and lifestyle factors, potentially informing personalized intervention strategies.

Characterization of metabolic manifolds. The seven distinct metabolic manifolds we have identified contain different sets of metabolites. **Fig. 2A** illustrates the distribution of metabolic biomarkers across these manifolds. Notably, **M6** exhibits the highest density of relative lipoprotein lipid concentrations

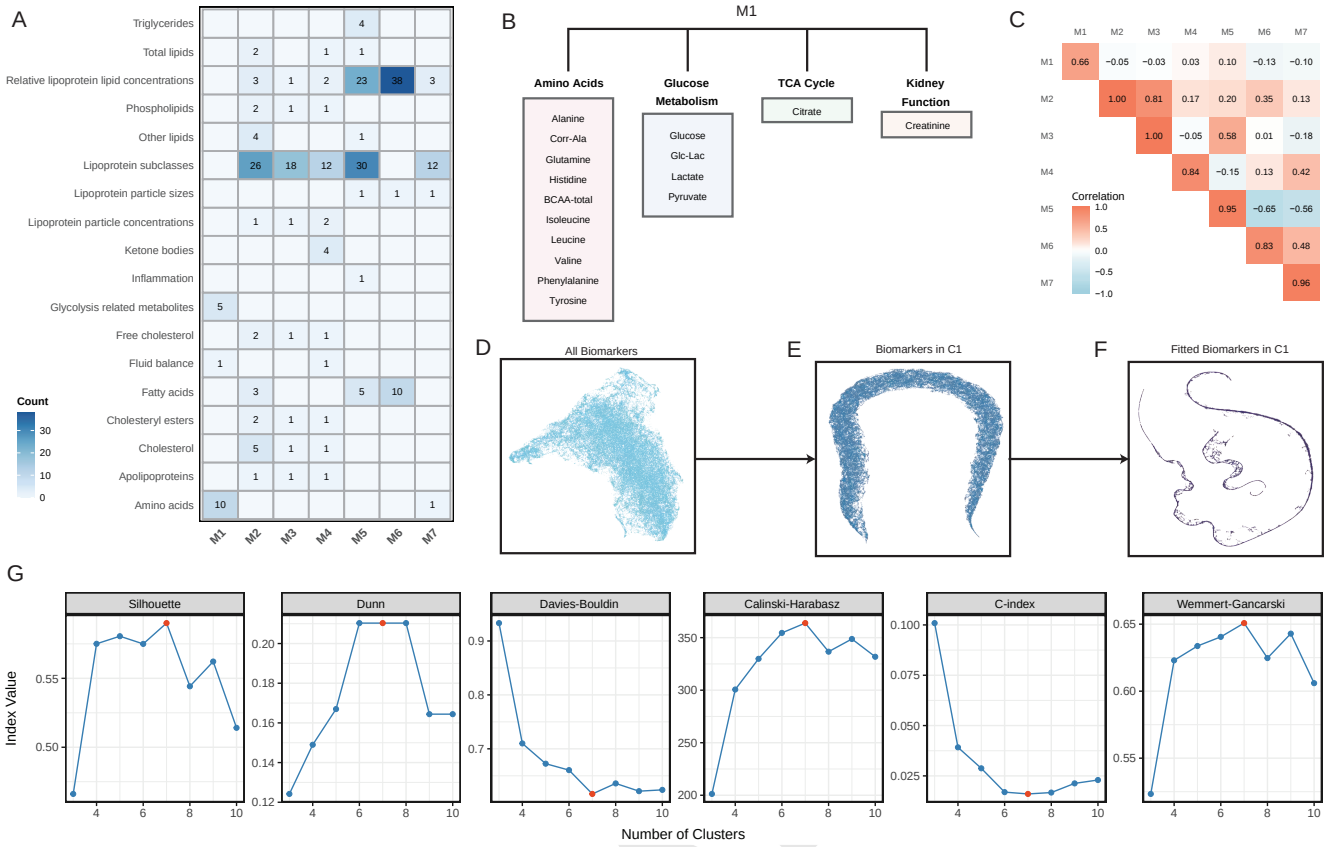

Fig. 2. Characterization of metabolic manifolds and their low-dimensional structure. (A) Heatmap showing the distribution of metabolic biomarkers across seven manifolds (**M1**–**M7**). (B) Detailed composition of **M1**, illustrating four distinct metabolic modules: amino acids, glucose metabolism, TCA cycle, and kidney function markers. (C) Mean Pearson correlation coefficients between biomarkers across different manifolds, demonstrating strong intra-manifold correlations and weak inter-manifold correlations. (D) Initial high-dimensional structure of all 251 biomarkers without clear pattern. (E) Emergence of low-dimensional characteristics in the first metabolite category (**C1**) after biomarker clustering. (F) Final manifold structure revealing two distinct subgroups after manifold fitting. (G) The optimal number of clusters identified by six clustering validation metrics. Higher values indicate better clustering for Silhouette score, Dunn index, Calinski-Harabasz index, and Wemmert-Gancarski index, while lower values are optimal for Davies-Bouldin index and C-index. All metrics consistently indicate seven clusters as optimal (red dots).

(contain 38 metabolic biomarkers, $n_b = 38$), while **M1** shows significant enrichment in amino acids ($n_b = 10$) and glycolysis-related metabolites ($n_b = 5$). Lipoprotein subclasses demonstrate heterogeneous distribution patterns, with substantial representation in **M2** ($n_b = 26$), **M3** ($n_b = 18$), **M5** ($n_b = 30$), and **M7** ($n_b = 12$). Here, we use **M1** as an example to demonstrate the biological rationale behind our metabolic marker categorization. The comprehensive description of metabolite compositions for all other manifolds can be found in *SI Appendix*, Table S2–S3.

Fig. 2B delineates the biological composition of **M1**, which comprises four distinct metabolic modules: amino acids, glucose metabolism, tricarboxylic acid (TCA) cycle, and kidney function biomarkers. The amino acid module encompasses ten key metabolites, including branched-chain amino acids, aromatic amino acids, and other essential and non-essential amino acids. The glucose metabolism module contains fundamental glycolytic intermediates and glucose-lactate ratio. The TCA cycle is represented by citrate, while kidney function is monitored through creatinine levels. **M1** captures a fundamental metabolic network centered on cellular energy metabolism and protein homeostasis, which reveals the intricate interplay between amino acid metabolism, glucose utilization, and energy production through the TCA cycle. The clustering of these metabolites suggests coordinated regulation of energy substrate utilization (26).

The biological validity of our biomarker clustering can be further substantiated through correlation analysis. **Fig. 2C** presents the mean Pearson correlation coefficients between biomarkers assigned to different manifolds. Global analysis reveals strong intra-manifold correlations while maintaining relatively weak inter-manifold correlations, supporting the modular organization of the metabolome. For instance, biomarkers within **M1** exhibit a high average correlation coefficient (0.66), whereas correlations between **M1** and other manifolds remain consistently low (less than 0.13).

The manifold fitting procedure substantially enhances the interpretability of the data. Initially, the 251 biomarkers do not show any obvious low-dimensional structure, as shown in the UMAP visualization **Fig. 2D**. However, after clustering the biomarkers, the first metabolite category (**C1**) begins to exhibit emerging low-dimensional characteristics, as depicted in the UMAP plot **Fig. 2E**, despite residual variance in other directions. Following manifold fitting, the data show pronounced directionality and segregate into two distinct, disconnected subgroups, as illustrated in **Fig. 2F**. Subsequent analyses have demonstrated that these subgroups exhibit significant differences in metabolic profiles, and

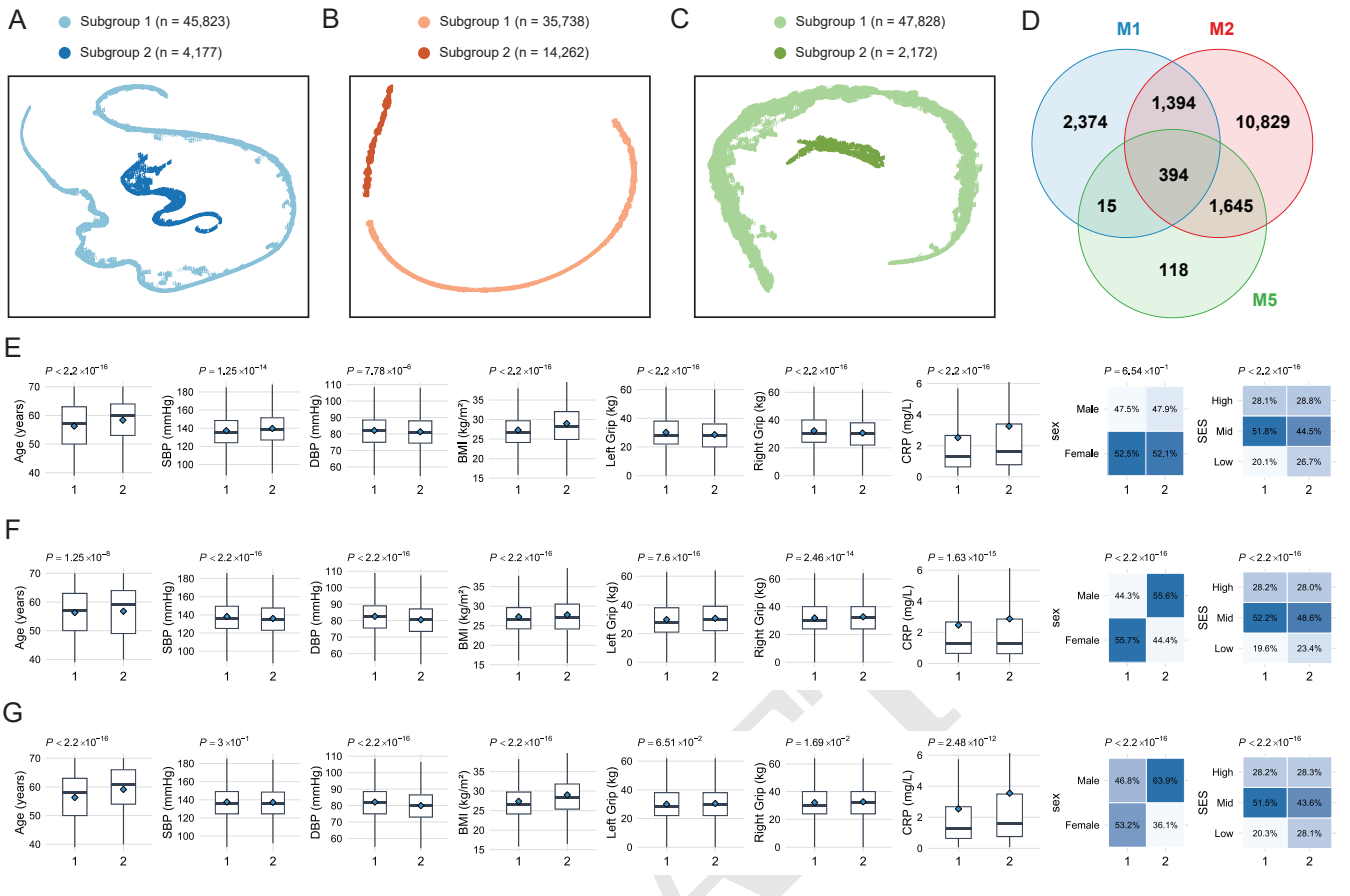

Fig. 3. Subgroups determined by metabolic manifold and associations with sociodemographic variables and baseline physical measurement. (A–C) Two-dimensional UMAP visualization of population stratification in metabolic manifolds (A: **M1**, B: **M2**, and C: **M5**). (D) Venn diagram showing the intersection patterns among subgroup 2 from **M1**, **M2**, and **M5**. (E–G) Distribution of sociodemographic variables and baseline physical measurement between subgroups from metabolic manifolds (E: **M1**, F: **M2**, and G: **M5**). For the boxplot, the central horizontal line represents the median value, while the lower and upper boundaries of the box denote the first and third quartiles, respectively. The whiskers extend to the most extreme data points not considered outliers, and the blue diamond indicates the mean value of the distribution. *P* is derived from Student's t-test and represents the probability of obtaining test results at least as extreme as the observed results under the null hypothesis, with $P < 0.05$ considered statistically significant.

disease risks. Analytical results for other metabolic manifolds are available in *SI Appendix*, Fig. S3–S5.

Fig. 2G shows the optimal number of clusters identified by six clustering validation metrics, including Silhouette score, the Dunn index (27), Davies-Bouldin index (28), Calinski-Harabasz index (29), C index (30), and Wemmert-Gancarski index (31). All six evaluation metrics indicate that seven clusters represent the optimal clustering structure, demonstrating the robustness and reliability of our clustering results.

Association between metabolic manifolds with demographics and physical measurements. Manifold fitting uncovers distinct population stratification patterns across three metabolic manifolds, each with unique topological properties in their low-dimensional representations:

- In **M1** (**Fig. 3A**), the larger subgroup ($n = 45,823$) encircles the smaller subgroup ($n = 4,177$). The larger subgroup naturally forms a flow pattern, while the smaller subgroup displays a serpentine structure with varying thickness along its length.

- **M2** (**Fig. 3B**) exhibits a simpler structural organization, with both subgroups arranged along curved trajectories. The larger subgroup ($n = 35,738$) and smaller subgroup ($n = 14,262$) form two isolated curves with continuously varying curvature.

- **M5** (**Fig. 3C**) shows a distribution pattern similar to **M1**, where the larger subgroup ($n = 47,828$) forms a circular pattern, with the smaller subgroup ($n = 2,172$) appearing as a distinct, compact structure within the circular formation.

The Venn diagram in **Fig. 3D** presents the number of overlapping participants among the minority subgroups from **M1**, **M2**, and **M5**. **M2** contains the largest exclusive subgroup ($n = 10,829$), followed by **M1** ($n = 2,374$) and **M5** ($n = 118$). Substantial overlap exists between **M2** and **M5** ($n = 1,645$), as well as between **M1** and **M2** ($n = 1,394$), while **M1** and **M5** share fewer individuals ($n = 15$). A core group of individuals ($n = 394$) exhibits metabolic characteristics common to all three manifolds, suggesting a distinct metabolic subgroup with features spanning multiple metabolic domains.

The three manifolds show consistent patterns in several key metabolic indicators (**Fig. 3E-G**). All demonstrate significant age and BMI differences between subgroups ($P <$

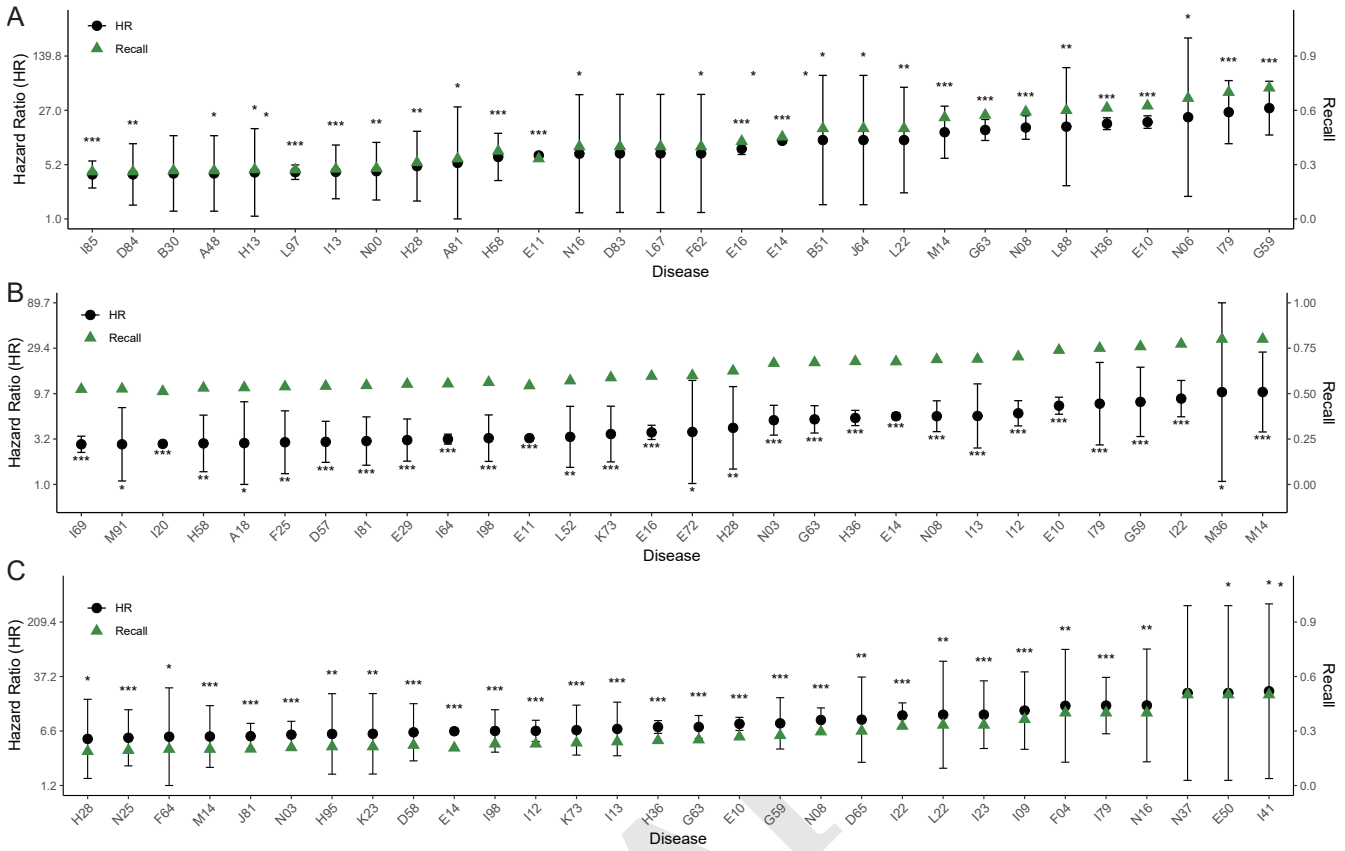

Fig. 4. Comparison of the hazard ratios (HRs) of diseases and their corresponding recall in three identified high-risk subgroups relative to their respective reference subgroups from metabolic manifolds (A: **M1**, B: **M2**, and C: **M5**). HRs represent the risk of disease incidence from baseline assessment (2007–2010) until right censoring. Each panel displays diseases sorted by descending HRs, with black circles and error bars representing the estimation of HRs and their 95% confidence intervals (CIs), respectively, and green triangles indicating the recall (percentage of all disease cases captured) of the high-risk subgroup among all patients with each disease. The number of asterisks $(*)$ represents the level of statistical significance: one asterisk $(*)$ indicates $P < 0.05$, two asterisks $(**)$ indicate $P < 0.01$, and three asterisks $(***)$ indicate $P < 0.001$.

2×10^{-16}), with Subgroup 2 consistently showing higher values. Blood pressure and grip strength measurements show systematic variations across manifolds, though with differing magnitudes. CRP levels are consistently elevated in Subgroup 2, suggesting a common inflammatory component. All manifolds display clear socioeconomic status (SES) stratification, indicating a robust link between metabolic profiles and social determinants of health.

Each manifold captures distinct aspects of metabolic heterogeneity. **M1** shows balanced gender distribution (47.5% vs 47.6% male) despite clear clinical differences (BMI: 26.64 vs 28.15 kg/m^2; CRP: 1.28 vs 1.63 mg/L). In contrast, **M2** exhibits the most pronounced gender disparity (46.8% vs 63.9% male) but more modest clinical differences (BMI: 26.65 vs 27.04 kg/m^2). **M5** has the strongest coupling between clinical and demographic differences, with the largest age gap (58 vs 61 years) and substantial BMI differences (26.68 vs 28.35 kg/m^2), alongside clear demographic stratification.

Association between metabolic manifolds with multiple diseases. Among the 923 diseases analyzed through Cox proportional hazards regression models, the three metabolic manifolds demonstrate significant positive associations (HR $>$ 1, $P < 0.05$) with 269 (29.1%), 287 (31.1%), and 298 (32.3%) diseases in **M1**, **M2**, and **M5**, respectively. For each manifold, we conducted separate survival analyses using membership in the high-risk cluster as the predictor variable and time-to-disease onset as the outcome. Protective effects (HR < 1, $P < 0.05$) are relatively rare, observed in only 7 (0.8%), 23 (2.5%), and 18 (1.9%) diseases. Despite representing small subgroups (8.4%, 28.5%, and 4.3% of the total population), these manifolds show high capture rates for severe conditions, with **M2** identifying the most diseases with high recall rates (81 diseases with recall $> 50\%$) compared to **M1** (15 diseases) and **M5** (6 diseases), suggesting its broader coverage of disease spectrum. Detailed results including all hazard ratios, confidence intervals, and P-values are available in *SI Appendix*, Table S4–S6.

M1, **M2**, and **M5**, each representing distinct population stratification patterns in their low-dimensional representations, demonstrate significant associations with both type 1 (E10; HR: 7.18–18.93) and type 2 diabetes (E11; HR: 3.24–6.89), accompanied by their microvascular complications including retinopathy (H36; HR: 5.31–18.02), nephropathy (N08; HR: 5.55–16.04), and neuropathy (G63; HR: 5.14–28.87). Here, hazard ratios (HRs) represent the risk of disease incidence from baseline assessment (2007–2010) until right censoring at the earliest of: date of disease diagnosis, date of death, date of loss to follow-up, or the last date of data collection. These associations are particularly robust in

conditions with large sample sizes, such as type 2 diabetes (E11; $n = 4,519$) and retinal disorders (H36; $n = 504$), as evidenced by their narrow CIs and highly significant P ($P < 0.001$). Correspondence and information between ICD-10 codes and disease names is available in *SI Appendix*, Table S1.

Despite these commonalities, each manifold captures distinct disease spectrums and risk gradients. **M1** (**Fig. 4A**), with its highest hazard ratios (HR > 15) for multiple conditions, predominantly highlights severe metabolic dysregulation and its complications, particularly in neurological (mononeuropathy (G59): HR = 28.87, 95% CI: 12.79–65.17, recall = 72.4%) and vascular disorders (arterial diseases (I79): HR = 25.56, 95% CI: 9.82–66.51, recall = 70.0%). In contrast, **M2** (**Fig. 4B**) exhibits a more moderate but broader risk profile (most HR < 6), distinctively emphasizing cardiovascular and autoimmune conditions, with notably elevated risks for arthropathies (M14; HR = 10.05, 95% CI: 3.77–26.77, recall = 80.0%) and subsequent myocardial infarction (I22; HR = 8.55, 95% CI: 5.47–13.34, recall = 77.3%). **M5** (**Fig. 4C**) demonstrates an intermediate pattern with a clear risk stratification, showing significant but generally lower hazard ratios compared to **M1**, while maintaining a comprehensive coverage of metabolic complications. Notably, the high recall rates (ranging from 60–80% for severe conditions) across all manifolds indicate their effectiveness in identifying the majority of high-risk patients despite representing relatively small subgroups. Even though the remaining five manifolds lack two- subgroup structures, they still exhibit population-specific patterns in disease onset risk, as demonstrated in *SI Appendix*, Fig. S1–S2.

These distinct risk patterns have important clinical implications for personalized prevention strategies. **M1** identifies a subgroup requiring aggressive complication prevention and intensive monitoring, particularly for neurological and vascular complications. **M2** suggests a need for comprehensive cardiovascular protection and autoimmune surveillance in its high-risk subgroup. **M5** indicates a requirement for stratified prevention strategies across multiple systems, with particular attention to the progression of metabolic syndrome. These findings highlight the potential of metabolic manifolds as complementary biomarkers for patient stratification, suggesting that their combined use might enable more precise risk prediction and personalized intervention strategies.

Associations between lifestyle and disease in high-risk metabolic subgroups. To translate our metabolic insights into actionable prevention strategies, we investigate the relationship between lifestyle factors and disease outcomes in high-risk populations identified through manifold analysis. Combining the high-risk subgroups of **M1**, **M2**, and **M5**, we have identified a comprehensive population that demonstrates increased susceptibility across multiple disease domains.

Our analysis focuses on three key lifestyle factors, including sleep patterns, physical activity and smoking status, and their associations with four major diseases: diabetes mellitus (E11, **Fig. 5A**), chronic ischemic heart disease (I25, **Fig. 5B**), chronic obstructive pulmonary disease (J44, **Fig. 5C**), and chronic kidney disease (N18, **Fig. 5D**). The results reveal consistent patterns of lifestyle-associated risk across these conditions, suggesting potential targets for intervention in metabolically vulnerable populations.

Sleep patterns emerge as a significant modifier of disease risk. Individuals with unhealthy sleep patterns show consistently higher disease incidence rates compared to those with healthy sleep patterns, with particularly pronounced differences in diabetes (14.48% vs 10.77%) and chronic kidney disease (9.36% vs 7.10%). This pattern suggests that sleep quality may be a crucial mediator of metabolic health, potentially through its effects on energy homeostasis and inflammatory pathways.

Physical activity levels demonstrate similarly striking associations with disease outcomes. The physically inactive group shows elevated risk across all studied conditions, with the most notable differences observed in diabetes (14.09% vs 8.90%) and ischemic heart disease (11.53% vs 9.72%). These findings underscore the importance of regular physical activity as a protective factor against metabolic dysfunction, even in populations with underlying metabolic vulnerability.

Smoking status shows the most dramatic impact on disease risk, particularly for respiratory conditions. The contrast is most striking for chronic obstructive pulmonary disease, where smokers show more than four-fold higher incidence rates compared to non-smokers (9.01% vs 2.03%). However, the impact of smoking extends beyond respiratory health, with elevated risks observed across all studied conditions, including diabetes (14.21% vs 9.65%) and chronic kidney disease (9.48% vs 6.14%).

Materials and Methods

UK Biobank cohort and data access. The UK Biobank is a comprehensive biomedical database (*SI Appendix*, A) that provides global access to data from approximately half a million participants aged between 40 and 69 years at baseline (32). This resource encompasses a wide array of health-related information, including questionnaire data on socio-economic status, lifestyle factors, and cognitive assessments, as well as measurements of heart and lung function, body size, and composition. Additionally, a variety of biochemical and imaging data are available.

The UK Biobank blood samples are collected at baseline in 22 assessment centres across the UK from 2007 to 2010. Protocols for handling and storing these samples are detailed in (33). Nightingale Health Plc. engages in the biomarker profiling of baseline plasma samples for the entire cohort. This profiling employs the Nightingale Health NMR biomarker platform (*SI Appendix*, B), the specifics of which are well-documented in (34, 35). The main procedural steps in the biomarker analysis and quality controlling are outlined in (4). This NMR metabolic biomarker dataset is made available to the research community via the UK Biobank as of March 2021. For this study, data are requested under UK Biobank project 146760 in March 2024. After the exclusion of incomplete data, a total of 251 NMR metabolic biomarkers from 212,853 participants are utilized. As the primary target of analysis, disease outcome is defined based on the first occurrence of 3-character ICD-10 codes using the hospital inpatient records from the UK Biobank. Additional lifestyle-related indicators, such as sleep and exercise levels, are also included in the analysis.

Refining data intrinsic structures. The proposed methodological framework employs all 251 NMR metabolic biomarkers to improve our understanding of their relationships and properties in a biological context, avoiding the selection of a subset. To cope with the high dimension and complex covariance structure among the biomarkers, we first partition them into several low-dimensional subspaces. Subsequently, we apply manifold fitting to each subspace to elucidate their intrinsic low-dimensional structures, facilitating further analysis. This process is illustrated in **Fig. 1A–C**.

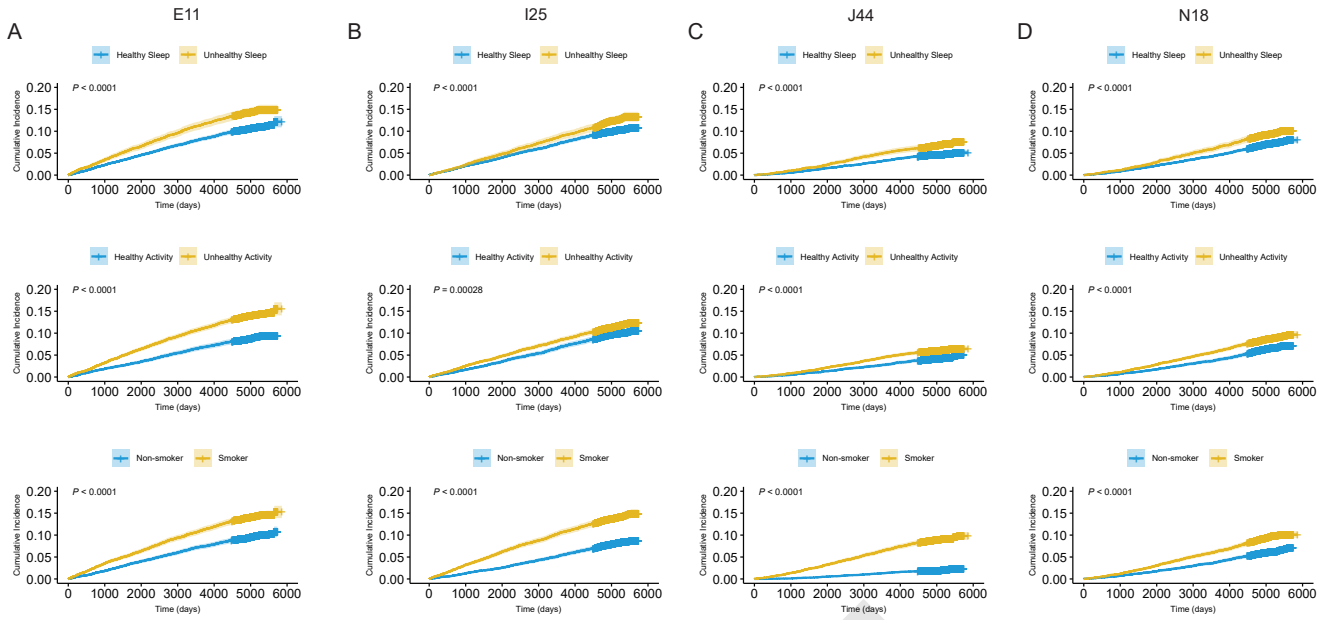

Fig. 5. Kaplan-Meier survival curves showing cumulative disease incidence stratified by sleep patterns, physical activity, and smoking status in metabolically vulnerable populations. Disease-specific incidence rates for (A) diabetes mellitus (E11), (B) chronic ischemic heart disease (I25), (C) chronic obstructive pulmonary disease (J44), and (D) chronic kidney disease (N18). Cross-mark represent right censoring marks indicating participants who were lost to follow-up, died from other causes, or reached the end of the study period without experiencing the event of interest.

Partitioning metabolic biomarkers into seven categories. The 251 metabolic biomarkers under study are involved in diverse metabolic pathways, with some directly measured and others calculated from these direct measurements. This variability contributes to a complex covariance structure, necessitating the partitioning of the biomarker space into multiple subspaces, or, in other words, cluster the biomarkers into multiple categories. Each biomarker is initially treated as a 212,853-dimensional vector, with pairwise distances calculated using a correlation metric. Dimension reduction and two-dimensional UMAP visualization reveal nonlinear dependencies and clustering features among the biomarkers (*SI Appendix*, Fig. S6). Following this, we recalculate a Manhattan distance matrix from the two-dimensional UMAP projections for hierarchical clustering using the average linkage method. We validate these clusters by calculating silhouette scores for cluster sizes ranging from 2 to 50, identifying the optimal size based on the highest average silhouette score. This process delineates seven distinct biomarker categories (**C1**–**C7**), each characterized by a low-dimensional nonlinear correlation structure or modeled around a latent manifold, exhibiting minimal inter-group correlations.

Fitting latent manifold in each subspace. The covariance matrices and two-dimensional visualizations for most metabolic biomarker categories exhibit strong nonlinear correlations within their respective subspaces. These dependencies imply that the relationships among biomarkers can be effectively refined and more accurately described using manifold fitting. This method leverages the structural features of the samples to enhance characterizations of principal variations by aligning the data towards a low-dimensional latent manifold. This alignment serves to filter out noise and irrelevant information, ensuring that the refined data remains consistent within the original space. The effectiveness of manifold fitting has been validated across various fields, and for more technical and theoretical details, refer to (19).

Take category **C1** as an instance, we represent the data in **C1** as an N by D_1 matrix $(x_{ij})_{N\times D_1}$, where N is the sample size, D_1 stands for the number of biomarkers in **C1**, and x_{ij} denotes the value of the jth biomarker in **C1** for the ith investigated individual. For each individual, let $x_i = (x_{i1}, \ldots, x_{iD_1})^\top$ be a D_1-dimensional vector in $\mathbb{R}^{D_1}$. The manifold fitting algorithm includes two primary steps: direction estimation and contraction estimation. Consider a sample point $z \in \{x_i : i = 1, \ldots, N\}$, the direction from z to the latent manifold is estimated using a reference point defined as:

$$\mu_z = \frac{\sum_{i=1}^N x_i \mathbf{1}_{\|x_i - z\| \le r_1}}{\sum_{i=1}^N \mathbf{1}_{\|x_i - z\| \le r_1}},$$

where $\|\cdot\|$ stands for the Euclidean norm and r_1 is a neighborhood radius . We then define a hyper-cylinder elongated along the vector $\mu_z - z$ by construct a projection matrix onto $\mu_z - z$ as:

$$\Pi_z = \frac{(\mu_z - z)(\mu_z - z)^\top}{\|\mu_z - z\|_2^2},$$

with which we decompose the vector $x_i - z$, for each x_i, into two directions:

$$u_i = \Pi_z(x_i - z), \quad v_i = x_i - z - u_i,$$

yielding the fitted version of z:

$$z' = \frac{\sum_{i=1}^N x_i \mathbf{1}_{\|u_1\| \le r_2, \|v_i\| \le r_1}}{\sum_{i=1}^N \mathbf{1}_{\|u_1\| \le r_2, \|v_i\| \le r_1}}, \qquad [1]$$

where $r_2 \gg r_1$ is another radius along the direction of $\mu_z - z$. As proposed by (19), with appropriately chosen parameters r_1 and r_2, z' is significantly closer to the latent manifold compared to z. This approach allows us to filter out less important information and improve subsequent analyses without prior knowledge of the dimension of the latent manifold.

Furthermore, two-dimensional UMAP visualizations of some categories of biomarkers represent uniformly distributed within a circular area, possibly due to the large variability in biomarker distribution ranges, possibly skewed by dominant components. To reduce these effects, we implement a rank-based transformation. Continuing to use **C1** as an example, the empirical cumulative distribution function for the jth biomarker is defined as:

$$\widehat{F}_j(t) = \frac{1}{N} \sum_{i=1}^N \mathbf{1}_{x_{ij} \le t},$$

where $\mathbf{1}_{x_{ij} \le t}$ is the indicator function of $x_{ij} \le t$. Subsequently, the raw observation x_{ij} is transformed to $y_{ij} \in [0, 1]$ by taking

$$y_{ij} = \widehat{F}_j(x_{ij}), \text{ for } i = 1, \ldots, N, \ j = 1, \ldots, D_1. \qquad [2]$$

The visualization of transformed data reveals stronger nonlinear dependencies in biomarker groups that are not characterized by the initial two-dimensional UMAP visualizations of the raw data. The manifold fitting is then performed on the dataset $\{y_{ij}\}$ and the resulting $\{y'_{ij}\}$ are transformed back to the measurement space using the quantile function

$$x'_{ij} = \widehat{F}_j^{-1}(y'_{i,j}) =: \inf\{t \in \mathbb{R} : \widehat{F}_j(t) \geq y'_{i,j}\}. \quad [3]$$

By applying the manifold fitting and transformation procedure to categories **C1**–**C7**, we generate the refined datasets on metabolic manifolds **M1**–**M7**, where the intrinsic geometric structures of the data are effectively refined. This entire process is encapsulated in the pseudo-code provided in Algorithm 1. Based on the initial visualizations, we determine that categories **C1** and **C4** do not require transformation. According to the recommendation from (19), the radii are set to be $r_1 = 5\sigma/\log_{10}(N)$ and $r_2 = 10\sigma\sqrt{\log_e(1/\sigma)/\log_{10}(N)}$, with a parameter σ. For the two categories without transformation, we σ fix as 1.5, while for the remaining transformed categories, σ is adjusted to $0.8\sqrt{D_j/D_1}$ for the jth category, where D_j denotes the number of biomarkers in that category.

Algorithm 1 Manifold fitting for each biomarker category

Input: Data set of interest (one of **C1**–**C7**): $\{x_i\}_{i=1}^N$; neighborhood radii r_1 and r_2.

Draw the two-dimensional visualization of $\{x_i\}_{i=1}^N$;
Determine if transformation is needed
if so **then**
    Calculate $z_{ij} \leftarrow \widehat{F}_j(x_{ij})$ with Eq. (2);
else
    Copy $z_{ij} \leftarrow x_{ij}$;
Update z'_i from z_i with Eq. (1), for $i = 1, \ldots, N$;
if so **then**
    Calculate $x'_{ij} \leftarrow \widehat{F}_j^{-1}(z'_{ij})$ with Eq. (3);
else
    Copy $x'_{ij} \leftarrow z'_{ij}$;

Return: Fitted set $\{x'_i\}_{i=1}^N$ (the corresponding one of **M1**–**M7**).

Heterogeneity analysis and phenotype associations. Following manifold fitting, we project each manifold (**M1**–**M7**) into two dimensions using UMAP. Due to the computational complexity, we randomly sample 50,000 participants to generate UMAP visualizations and conduct subsequent analyses in the main text; UMAP projections for the complete population are provided in *SI Appendix* Fig. S7–S8. Four of these projections (**M1**, **M2**, **M3**, **M5**) exhibit marked topological discontinuities, manifesting as well-defined, discrete population substructures. We then apply density-based spatial clustering of applications with noise (DBSCAN, (36)) to these reduced-dimensional representations, which reveals robust stratification into binary subgroups within manifolds **M1**, **M2**, and **M5**. For these three categories, we examine associations with demographic and clinical variables, including sex, age, body mass index (BMI), systolic and diastolic blood pressure, bilateral hand grip strength, and sleep quality metrics. For continuous variables, we assess between-group differences using two-tailed t-tests and visualize distributions through boxplots; for categorical variables, we compare frequency distributions using chi-squared tests.

Following population stratification, we conduct survival analysis on diseases recorded in the UK Biobank using Cox proportional hazards model. For each disease, we calculate the follow-up time as the interval between the baseline assessment date and either the date of the first disease diagnosis or the end of the follow-up period for censored cases. Disease status is determined using the International Classification of Diseases 10th revision (ICD-10) codes from UK Biobank health records. We construct Cox models incorporating subgroup assignments as predictors, with subgroup 2 designated as the high-risk group. For each disease, we compute hazard ratios with 95% CIs and corresponding P. To assess the practical utility of our stratification approach, we calculate the recall rate for each disease, which is defined as the proportion of total disease cases captured within subgroup 2. We specifically focus on diseases showing both statistical significance ($p < 0.05$) and clinical relevance (HR > 1), ranking them by HR magnitude to identify conditions most strongly associated with metabolic vulnerability. This comprehensive analysis enables us to identify both the statistical strength (through HRs and P) and practical significance (through recall rates) of the metabolic subgroup stratification in disease risk prediction.

For the high-risk subgroups identified through manifold fitting and clustering, we further investigate the relationship between lifestyle factors (*SI Appendix*, D) and disease outcomes (*SI Appendix*, C) through survival analysis. We first combine the high-risk subgroups from manifolds **M1**, **M2**, and **M5** to create a comprehensive high-risk subgroup. For each disease of interest, we exclude individuals who had the disease at baseline and calculate the follow-up time from the baseline assessment date to either the date of disease diagnosis or the end of follow-up period.

We stratify this high-risk population based on three lifestyle factors: sleep patterns (healthy/unhealthy sleep), physical activity levels (healthy/unhealthy activity), and smoking status (smoker/non-smoker). For each lifestyle factor, we generate Kaplan-Meier curves to visualize cumulative disease incidence over time, with differences between groups assessed using log-rank tests. We also calculate the crude incidence rates for each lifestyle category by dividing the number of incident cases by the total number of individuals in that category. This analysis enables us to quantify the potential impact of modifiable lifestyle factors on disease risk within metabolically vulnerable populations.

Discussion

Our study presents a new analytical framework for understanding metabolic heterogeneity in large populations through manifold fitting, offering several key advances in both methodological approaches and biological insights. By applying this framework to the UK Biobank cohort, we have demonstrated how complex metabolic relationships can be effectively captured and interpreted through low-dimensional manifolds, revealing distinct population substructures with meaningful clinical correlates.

Our approach offers several advantages over traditional metabolomic analysis methods. First, by operating directly in the original feature space, manifold fitting preserves the interpretability of metabolic measurements while effectively reducing noise and capturing underlying biological structure. Second, the identification of discrete subgroups within continuous metabolic variations provides a natural framework for patient stratification that could inform precision medicine approaches. Third, the strong associations between manifold-based subgroups and disease outcomes suggest that these metabolic patterns may serve as early indicators of disease risk, potentially enabling more targeted prevention strategies.

The clinical implications of our findings are particularly relevant for precision medicine. The distinct disease associations observed across different manifolds suggest that metabolic risk factors may cluster in patterns that are not captured by traditional clinical measurements. For instance, the identification of subgroups with elevated risk for specific disease types (metabolic complications, cardiovascular diseases, or autoimmune conditions) suggests the potential for more nuanced approaches to patient risk stratification and preventive care. Although some subgroups show substantial

overlap across manifolds, closer inspection reveals highly asymmetric population compositions, with certain manifolds capturing more specialized high-risk subpopulations. This hierarchical organization highlights how different models may uncover varying levels of metabolic vulnerability, from broader risk profiles to more severely affected individuals with amplified disease burdens.

Looking forward, our framework opens several promising avenues for future research. First, genetic analysis of the identified metabolic subgroups could provide crucial insights into the hereditary components of metabolic heterogeneity. By conducting genome-wide association studies within each manifold-defined subgroup, we could identify genetic variants that contribute to specific metabolic patterns (37). This could be particularly informative for understanding the genetic architecture of complex metabolic traits and their relationship to disease risk (38).

Further exploration could focus on the longitudinal stability of these metabolic manifolds and their potential as predictive biomarkers. Time-series analysis of metabolic profiles could reveal how individuals transition between different metabolic states and whether these transitions correlate with disease onset or progression. This could lead to the development of early warning systems for metabolic dysfunction and more precise timing of preventive interventions (39).

Data, Materials, and Software Availability. The Nightingale Health NMR biomarker data has been publicly available to the UK Biobank resource since Spring 2021 and can be accessed through the UK Biobank portal (https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=220). Approved researchers may access these data in accordance with the UK Biobank data-access protocol. Detailed information on the data access process is available at https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access.

The code of our analysis method, implemented in R and MATLAB, includes a demonstration of the pipeline, all necessary intermediate results, their corresponding final results, and all evaluation functions used in this study. This implementation is available at https://github.com/zhigang-yao/MF-Metabolomic-Heterogeneity (40). Additional results are provided in the SI Appendix.

ACKNOWLEDGMENTS. Z.Y. has been supported by the Singapore Ministry of Education Tier 2 grant (A-0008520-00-00 and A-8001562-00-00) and the Tier 1 grant (A8000987-00-00 and A-8002931-00-00) at the National University of Singapore; B.L. and J.S. are postdoctoral researchers supported by grant A-8001562-00-00. R.L. is a doctoral student supported by a Research Scholarship at the National University of Singapore.

1. P Soininen, et al., High-throughput serum nmr metabonomics for cost-effective holistic studies on systemic metabolism. *Analyst* **134**, 1781–1785 (2009).
2. PA Chevli, et al., Plasma metabolomic profiling in subclinical atherosclerosis: the diabetes heart study. *Cardiovasc. Diabetol.* **20**, 1–12 (2021).
3. S Yang, et al., Plasma metabolomics identifies key metabolites and improves prediction of diabetic retinopathy: Development and validation across multinational cohorts. *Ophthalmology* **131**, 1436–1446 (2024).
4. H Julkunen, et al., Atlas of plasma nmr biomarkers for health and disease in 118,461 individuals from the UK biobank. *Nat. Commun.* **14**, 604 (2023).
5. T Buergel, et al., Metabolomic profiles predict individual multidisease outcomes. *Nat. Medicine* **28**, 2309–2320 (2022).
6. W Zhang, et al., Classification of osteoarthritis phenotypes by metabolomics analysis. *BMJ open* **4**, e006286 (2014).
7. C Brouard, et al., Fast metabolite identification with input output kernel regression. *Bioinformatics* **32**, i28–i36 (2016).
8. P Sen, et al., Deep learning meets metabolomics: a methodological perspective. *Briefings Bioinforma.* **22**, 1531–1542 (2021).
9. M Belkin, P Niyogi, Using manifold stucture for partially labeled classification in *Advances in Neural Information Processing Systems*, eds. S Becker, S Thrun, K Obermayer. (MIT Press), Vol. 15, pp. 1–8 (2002).
10. L Van der Maaten, G Hinton, Visualizing data using T-SNE. *J. Mach. Learn. Res.* **9** (2008).
11. L McInnes, J Healy, J Melville, UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018) (Accessed on 20 May 2024).
12. ML Mo, BØ Palsson, Understanding human metabolic physiology: a genome-to-systems approach. *Trends Biotechnol.* **27**, 37–44 (2009).
13. K Suhre, C Gieger, Genetic variation in metabolic phenotypes: study designs and applications. *Nat. Rev. Genet.* **13**, 759–769 (2012).
14. JK Nicholson, ID Wilson, Understanding 'global' systems biology: Metabonomics and the continuum of metabolism. *Nat. Rev. Drug Discov.* **2**, 668–676 (2003).
15. MD Ritchie, et al., Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *The Am. J. Hum. Genet.* **69**, 138–147 (2001).
16. KH Liland, Multivariate methods in metabolomics–from pre-processing to dimension reduction and statistical analysis. *TrAC Trends Anal. Chem.* **30**, 827–841 (2011).
17. C Fefferman, S Ivanov, Y Kurylev, M Lassas, H Narayanan, Fitting a putative manifold to noisy data in *Conference on Learning Theory*, eds. B Sébastien, P Vianney, R Philippe. (PMLR), pp. 688–720 (2018).
18. C Fefferman, S Ivanov, M Lassas, H Narayanan, Fitting a manifold of large reach to noisy data. *J. Topol. Analysis* **0**, 1–82 (2023).
19. Z Yao, J Su, B Li, S-T Yau, Manifold fitting. *arXiv preprint arXiv:2304.07680* (2023) (Accessed on 10 June 2024).
20. Z Yao, J Su, S-T Yau, Manifold fitting with cyclegan. *Proc. Natl. Acad. Sci.* **121**, e2311436121 (2024).
21. Z Yao, B Li, Y Lu, S-T Yau, Single-cell analysis via manifold fitting: A framework for rna clustering and beyond. *Proc. Natl. Acad. Sci.* **121**, e2400002121 (2024).
22. R Steuer, et al., From structure to dynamics of metabolic pathways: application to the plant mitochondrial tca cycle. *Bioinformatics* **23**, 1378–1385 (2007).
23. R Steuer, J Kurths, O Fiehn, W Weckwerth, Observing and interpreting correlations in metabolomic networks. *Bioinformatics* **19**, 1019–1026 (2003).
24. A Izadpanah, et al., A short-term diet and exercise intervention ameliorates inflammation and markers of metabolic health in overweight/obese children. *Am. J. Physiol. Metab.* **303**, E542–E550 (2012).
25. M Kaern, TC Elston, WJ Blake, JJ Collins, Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Genet.* **6**, 451–464 (2005).
26. CB Newgard, Metabolomics and metabolic diseases: where do we stand? *Cell Metab.* **25**, 43–56 (2017).
27. JC Dunn, Well-separated clusters and optimal fuzzy partitions. *J. Cybern.* **4**, 95–104 (1974).
28. DL Davies, DW Bouldin, A cluster separation measure. *IEEE Transactions on Pattern Analysis Mach. Intell.* pp. 224–227 (1979).
29. T Caliński, J Harabasz, A dendrite method for cluster analysis. *Commun. Stat. Methods* **3**, 1–27 (1974).
30. FE Harrell Jr, KL Lee, DB Mark, Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Medicine* **15**, 361–387 (1996).
31. B Desgraupes, MB Desgraupes, Package 'clustercrit' (2018).
32. C Sudlow, et al., UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Med.* **12**, e1001779 (2015).
33. P Elliott, TC Peakman, The UK biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *Int. J. Epidemiol.* **37**, 234–244 (2008).
34. P Würtz, et al., Quantitative serum nuclear magnetic resonance metabolomics in large-scale epidemiology: a primer on-omic technologies. *Am. J. Epidemiol.* **186**, 1084–1096 (2017).
35. P Soininen, AJ Kangas, P Würtz, T Suna, M Ala-Korpela, Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. *Circ. Cardiovasc. Genet.* **8**, 192–206 (2015).
36. M Ester, HP Kriegel, J Sander, X Xu, A density-based algorithm for discovering clusters in large spatial databases with noise in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, eds. E Simoudis, J Han, U Fayyad. (AAAI Press), p. 226–231 (1996).
37. E Uffelmann, et al., Genome-wide association studies. *Nat. Rev. Methods Primers* **1**, 59 (2021).
38. H Zhong, X Yang, LM Kaplan, C Molony, EE Schadt, Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *The Am. J. Hum. Genet.* **86**, 581–591 (2010).
39. VP Mäkinen, et al., Longitudinal metabolomics of increasing body-mass index and waist-hip ratio reveals two dynamic patterns of obesity pandemic. *Int. J. Obes.* **47**, 453–462 (2023).
40. B Li, J Su, R Lin, S-T Yau, Z Yao, Github - zhigang-yao/mf-metabolomic-heterogeneity (2024) https://github.com/zhigang-yao/MF-Metabolomic-Heterogeneity, Deposited on 29 December 2024.