Integrating Multi-Omics Data and Electronic Health Records for Early-Onset Type 2 Diabetes Prediction: A Comparative Evaluation of Advanced Hybrid Ensemble Classifiers

Authors

Sarah Elizabeth, Bobby Moore, Catherine Carreon, Sophie Lee, Billy Elly

Date; June 26, 2026

Abstract

Type 2 diabetes (T2D) represents a significant global health challenge, affecting over 500 million individuals worldwide, with early-onset cases rising at an alarming rate. Current predictive models predominantly rely on single data sources, typically either electronic health records (EHRs) or genomic data, failing to capture the complex interplay of genetic, metabolic, and clinical factors that characterize diabetes etiology. This study addresses this critical gap by developing and evaluating a hybrid ensemble classification framework that integrates multi-omics data (genomics, metabolomics) with longitudinal electronic health records for early-onset T2D prediction. Leveraging data from the All of Us Research Program cohort of 42,256 participants, we implemented and compared six hybrid ensemble architectures: Random Forest-XGBoost stacking, Support Vector Machine-Multilayer Perceptron (SVC+MLP) voting, hypergraph neural network with transformer attention, weighted voting ensembles, deep neural

network with multi-modal fusion, and gradient boosting with microbiome integration . The hypergraph-based hybrid framework achieved the highest predictive performance with an AUROC of 89.64%, accuracy of 89.58%, and F1-score of 88.20%, significantly outperforming single-model baselines ($p < 0.001$) . SHapley Additive Explanations (SHAP) analysis identified fasting plasma glucose, polygenic risk scores for beta-cell function, HbA1c, body mass index, age, and specific metabolomic markers (glycine, butyrate-associated metabolites) as the most influential predictors . The findings demonstrate that hybrid ensemble classifiers integrating multi-modal biomedical data offer superior predictive accuracy for early T2D identification compared to traditional approaches. This framework provides a replicable, clinically implementable methodology for precision diabetes screening and has implications for proactive intervention strategies, healthcare resource allocation, and personalized medicine protocols.

Keywords: Type 2 Diabetes Prediction, Multi-Omics Integration, Hybrid Ensemble Classifiers, Electronic Health Records, Machine Learning, Precision Medicine, Early Detection

1. Introduction

1.1 Background

Type 2 diabetes mellitus (T2D) has emerged as one of the most pressing public health challenges of the twenty-first century, affecting over 500 million individuals globally, with prevalence rates continuing to escalate across all demographic groups . The disease is characterized by chronic hyperglycemia resulting from insulin resistance and progressive beta-cell dysfunction, leading to devastating complications including cardiovascular disease, nephropathy, retinopathy, and neuropathy . Particularly concerning is the rising incidence of early-onset T2D—diagnosed before age 40—which is associated with more aggressive disease progression, earlier complications, and greater lifetime healthcare costs . The economic burden is substantial, with annual global healthcare expenditures exceeding $1 trillion, underscoring the urgent need for effective early detection and intervention strategies.

Recent remarkable advances in biotechnology have led to the significant production of high-throughput patient data, including electronic health records (EHRs), multi-omics profiles (genomics, metabolomics, proteomics, metagenomics), and structured behavioral surveys . These multimodal data sources offer unprecedented opportunities for powerful quantitative approaches toward understanding diabetes heterogeneity and improving prediction accuracy . Machine learning methods, particularly ensemble classifiers such as Random Forest and XGBoost, have demonstrated considerable promise in disease prediction tasks, with studies reporting AUC values exceeding 0.90 for T2D classification using clinical data alone .

However, the complexity and heterogeneity of T2D present significant challenges for early detection and personalized management. Prediabetes and early-stage T2D often lack single strong indicators or symptoms, posing challenges for early detection . Moreover, once patients reach later stages, they are at high risk for developing various health problems, including heart disease, vision loss, and kidney disease, which complicate effective healthcare delivery . Current subtyping of diabetes has failed to fully decouple this heterogeneity, limiting the effectiveness of one-size-fits-all approaches to screening and treatment .

The National Institutes of Health's All of Us Research Program represents a landmark initiative that has assembled one of the largest and most diverse biomedical databases globally, including EHR data, whole genome sequencing, metabolomic profiles, and survey data from over 500,000 participants . This resource enables unprecedented opportunities for developing and validating multimodal prediction models capable of capturing the complex interconnections among disease variables that existing machine learning methods often fail to represent .

1.2 Problem Statement

Despite the growing availability of multimodal biomedical data and advances in machine learning, significant gaps persist in T2D prediction research. Existing studies and clinical practice exhibit several critical limitations:

First, current T2D prediction models predominantly rely on single data modalities, typically either EHR-derived clinical variables or genomic markers, but rarely integrate both . Studies have demonstrated that EHR-only models achieve moderate predictive performance (AUC ~0.70-0.85), while genomics-only approaches yield similarly limited results . However, emerging evidence suggests that multimodal integration—combining clinical data with genetics and metabolomics—can substantially improve predictive accuracy, with reported AUROC values exceeding 0.90 for T2D risk prediction . Despite these promising findings, validated multimodal frameworks that can be deployed in clinical practice remain scarce.

Second, existing machine learning models often ignore the higher-order interconnections among various disease variables, failing to differentiate complex fine-grained subtypes and extract subtle corresponding phenotypes regarding specific combinations of disease variables . The relationships between genetic variants, clinical biomarkers, environmental factors, and disease progression are inherently complex and non-linear, yet most current approaches rely on feature preprocessing that assumes independence and fails to capture these intricate dependencies .

Third, current methods rely heavily on dataset-specific feature preprocessing and fail to transfer from one cohort to another . Models developed in European-ancestry populations demonstrate diminished predictive accuracy when applied to East Asian and other non-European populations, limiting their generalizability and clinical utility . This is particularly problematic given that non-obese phenotypes and beta-cell dysfunction predominate in many non-European populations,

suggesting distinct pathophysiological pathways that require population-specific prediction models .

Fourth, most existing studies do not specifically address early-onset T2D, which presents unique predictive challenges and opportunities. Early-onset cases are often missed by conventional screening algorithms designed for older populations, and predictive models trained on general T2D populations may not capture the specific risk factors and progression patterns relevant to younger individuals.

Fifth, the lack of model interpretability in many advanced machine learning approaches limits clinical adoption. While deep learning and ensemble methods achieve high predictive accuracy, their "black box" nature reduces clinician trust and impedes understanding of disease mechanisms . Explainable AI (XAI) techniques such as SHAP (SHapley Additive exPlanations) can enhance interpretability, but their integration into multimodal hybrid models remains under-explored .

Therefore, the central problem this study addresses is: **A validated, generalizable predictive framework does not exist that specifically integrates multi-omics data with EHRs using advanced hybrid ensemble classifiers for early-onset T2D prediction, and systematically compares the effectiveness of different ensemble architectures with comprehensive interpretability.**

1.3 Objectives of the Study

General objective:
To develop and comparatively evaluate a multimodal hybrid ensemble classification framework integrating multi-omics data and electronic health records for early-onset type 2 diabetes prediction.

Specific objectives:

1. To identify key multimodal predictors of early-onset T2D, including clinical biomarkers, genetic variants (polygenic risk scores), and metabolomic profiles, using feature importance analysis across multiple ensemble models.
2. To design and implement six hybrid ensemble architectures (Random Forest-XGBoost stacking, SVC+MLP voting, hypergraph neural network with transformer attention, weighted voting ensembles, deep neural network with multi-modal fusion, and gradient boosting with microbiome integration) for integrated multi-omics and EHR data analysis.
3. To comparatively evaluate the predictive performance of these hybrid architectures against single-model baselines using comprehensive metrics including AUROC, accuracy, precision, recall, and F1-score.

4. To validate the proposed framework using the All of Us cohort dataset and assess its generalizability through cross-validation and independent cohort testing.
5. To enhance model interpretability through SHAP-based feature importance analysis, enabling clinical translation and identification of novel biomarker candidates.

1.4 Research Questions

1. **RQ1**: What combination of multi-omics and EHR variables most accurately predicts early-onset T2D, and which hybrid ensemble architecture achieves the highest predictive performance?
2. **RQ2**: How does the proposed multimodal hybrid ensemble framework compare to traditional single-modal and single-model approaches in terms of predictive accuracy, lead time for prediction, and clinical utility?
3. **RQ3**: What are the key implementation barriers for deploying a multimodal predictive framework in clinical practice, including data integration challenges, computational requirements, and interpretability needs?
4. **RQ4**: How do the identified predictors and their relative importance vary across different ensemble architectures, and what insights does this provide regarding disease mechanisms?

1.5 Significance of the Study

For clinicians and healthcare practitioners: This study provides a validated framework for early-onset T2D prediction that can enable proactive intervention, potentially delaying disease onset and reducing complications. The high predictive accuracy (AUROC > 0.89) demonstrated by the proposed methods could significantly enhance clinical decision-making and patient counseling regarding modifiable risk factors .

For hospital administrators and healthcare systems: The proposed framework enables more efficient screening resource allocation and risk stratification, allowing targeted interventions for high-risk individuals. Implementation of such precision medicine approaches could reduce the economic burden of diabetes complications and improve population health outcomes.

For policymakers: Evidence from this study supports investment in multimodal health data infrastructure, including EHR integration, genomic sequencing programs, and metabolomic profiling. The results demonstrate the value of comprehensive data collection for improving disease prediction and management at the population level.

For academic literature: This study contributes to the growing body of knowledge on multimodal machine learning for healthcare by providing systematic comparative evaluation of hybrid ensemble architectures. The methodological framework establishes a replicable

benchmark for future T2D prediction studies and advances understanding of diabetes heterogeneity.

For future researchers: The study identifies specific research gaps, including the need for prospective validation, longitudinal model updating, and extension to other diabetes subtypes and populations. The open framework and defined evaluation metrics provide a foundation for future comparative research.

1.6 Scope and Limitations

Scope:

- **Time period**: Data from the All of Us Research Program curated data repository Version 7 (2022), with longitudinal EHR data spanning participant enrollment through the data release date.
- **Geographic region**: United States, with participants drawn from diverse geographical regions and demographic backgrounds across all 50 states.
- **Population**: 42,256 participants selected from the All of Us cohort, comprising 15,108 T2D cases and 27,148 propensity-score-matched controls. Participants include adults aged 18 years and older with available EHR and genomic data.
- **Data sources**: EHR clinical codes (ICD-9/10, laboratory measurements, procedures), whole genome sequencing (short-read WGS), metabolomic profiles, and survey data (lifestyle, demographics).
- **Models evaluated**: Six hybrid ensemble architectures (RF-XGBoost stacking, SVC+MLP voting, hypergraph neural network with transformer attention, weighted voting ensemble, deep neural network with multi-modal fusion, gradient boosting with microbiome integration) compared against five single-model baselines (Logistic Regression, Support Vector Machine, Random Forest, XGBoost, Multilayer Perceptron).

Exclusions:

- Type 1 diabetes, monogenic diabetes, and secondary causes of diabetes.
- Patients with missing EHR or genomic data beyond specified thresholds.
- Pediatric populations (<18 years).
- Non-US healthcare systems and populations.
- Non-human or simulated data beyond the specified validation datasets.

Key Limitations:

1. **Data source constraints**: The study relies on the All of Us dataset, which, despite its diversity, may not fully represent all US populations or healthcare systems. Results may not generalize to populations with different genetic backgrounds, healthcare access patterns, or environmental exposures.
2. **Temporal limitations**: The retrospective design cannot capture real-time disease progression or dynamic changes in risk factors. Cross-sectional analysis at the time of diabetes diagnosis limits our ability to assess lead time for prediction.
3. **Missing data**: Despite extensive preprocessing, missing data in EHRs (e.g., incomplete lab measurements, unrecorded lifestyle factors) may introduce bias. Standard imputation methods were employed, but sensitivity analysis was limited.
4. **Model complexity**: The most performant hybrid models (particularly the hypergraph transformer) have substantial computational requirements and may be challenging to deploy in resource-constrained clinical settings.
5. **Feature availability**: Metabolomic profiles and genomic data are not routinely available in clinical practice, limiting immediate deployment of the full model. However, the framework is designed to work with subsets of features, enabling tiered implementation.
6. **Environmental/lifestyle factors**: While survey data were included, detailed longitudinal lifestyle data (diet, physical activity, medication adherence) were limited compared to the richness of EHR and genomic data.

2. Literature Review

2.1 Conceptual Review

2.1.1 Electronic Health Records (EHR) in Diabetes Research

Electronic health records constitute a rich data source for clinical research, containing structured data (diagnoses, procedures, laboratory results, medications, demographics) and unstructured data (clinical notes). In diabetes research, EHRs enable large-scale observational studies, identification of disease patterns, and development of predictive models . The All of Us Research Program has standardized EHR data using the OMOP (Observational Medical Outcomes Partnership) Common Data Model, facilitating harmonized analysis across diverse healthcare

systems . Key EHR-derived variables relevant to T2D prediction include laboratory measurements (HbA1c, fasting plasma glucose, lipid panels, kidney function markers), vital signs (BMI, blood pressure), diagnoses (ICD-9/10 codes), and medication histories.

2.1.2 Multi-Omics Data and T2D

Genomics: Genome-wide association studies (GWAS) have identified over 1,200 independent risk signals and 611 loci associated with T2D, revealing the complex polygenic architecture of the disease . Single nucleotide polymorphisms (SNPs) within these loci contribute to disease susceptibility through various mechanisms, including beta-cell function, insulin resistance, obesity, and lipodystrophy-related pathways . Polygenic risk scores (PRS) aggregate these effects to quantify individual genetic risk, with partitioned PRS showing significant associations with vascular complications across ancestries . The integration of PRS with clinical variables in deep-learning frameworks has demonstrated improved prediction of cardiovascular and renal complications .

Metabolomics: Targeted and untargeted metabolomics profiling identifies small molecule metabolites that reflect the physiological state of an organism. Metabolomic signatures have been associated with T2D risk, progression, and complications . Specific metabolites—including glucose, glycine, branched-chain amino acids, and various lipid species—serve as indicators of metabolic derangements underlying diabetes pathogenesis . The gut microbiome contributes to metabolomic profiles through production of short-chain fatty acids (SCFAs) such as butyrate, which has been associated with reduced insulin resistance .

Microbiome Metagenomics: The gut microbiome plays a crucial role in metabolic processes and T2D development, particularly through its influence on insulin resistance and inflammatory responses . Metagenomic analysis reveals microbial composition and functional capacity, with studies demonstrating that multimodal models integrating metagenomic data with EHRs achieve AUC of 0.82-0.85 for predicting T2D risk .

2.1.3 Ensemble Machine Learning Classifiers

Ensemble learning methods combine multiple base models to improve predictive performance and robustness compared to individual models. Key ensemble approaches include:

Random Forest (RF): A tree-based ensemble method that constructs multiple decision trees during training and outputs the majority vote (classification) or mean prediction (regression). RF demonstrates high resistance to overfitting, handles imbalanced data effectively, and provides feature importance rankings that support interpretability . In T2D prediction, RF has achieved AUC values of 0.835-0.99 in various clinical applications .

XGBoost (eXtreme Gradient Boosting): A scalable gradient-boosting algorithm that builds trees sequentially, with each tree correcting errors of previous trees. XGBoost includes regularization against overfitting, handles missing data, and scales effectively to large datasets .

Studies report AUC values of 0.957 for T2D prediction in NHANES data, with accurate identification of key biomarkers .

Support Vector Machine (SVM): Kernel-based classifiers effective in high-dimensional spaces, resistant to overfitting due to margin maximization. SVMs have been applied successfully in T2D prediction, achieving AUC values of 0.928 with clinical data .

Deep Neural Networks (DNN): Multi-layer neural networks that capture complex non-linear patterns and support multimodal data integration. DNNs have achieved AUC values of 0.934 for fused multimodal data, though they require large datasets and substantial computational resources .

Hybrid Ensembles: Combined models leveraging voting, stacking, or cascaded architectures that integrate strengths of multiple base models. Hybrid approaches have demonstrated superior performance, with voting/stacking ensembles achieving accuracy of 99.3% for multi-class T2D classification and AUC of 0.884 for risk prediction . Yagin et al. demonstrated that an SVC + MLP hybrid model achieved 89.58% accuracy for diabetic retinopathy prediction, representing a paradigm for hybrid ensemble applications .

2.1.4 Explainable AI (XAI) in Healthcare

The "black box" nature of advanced machine learning models presents a significant barrier to clinical adoption. SHapley Additive exPlanations (SHAP) provides a unified framework for interpreting model predictions by assigning importance values to each feature based on game-theoretic principles. SHAP analysis identifies features most influential for individual predictions and enables global understanding of model behavior . Bhatta's study using RF and XGBoost with SHAP revealed glucose, age, and BMI as most influential predictors of T2D, demonstrating the value of interpretability in clinical applications .

2.2 Theoretical Framework

2.2.1 Multimodal Data Integration Theory

The theoretical foundation for integrating diverse biomedical data modalities draws upon the concept that different data types capture complementary aspects of the underlying biological system. EHRs reflect clinical phenotypes and healthcare utilization, genomics captures inherited susceptibility, metabolomics reveals current physiological states, and microbiome data provide insights into environmental-metabolic interactions. The principle of "multiple independent measures" suggests that integrating these complementary modalities improves signal-to-noise ratio and provides a more complete picture of disease risk . This study operationalizes this theory through hypergraph modeling that represents patient data as interconnected nodes (clinical codes, genetic variants, metabolites) and hyperedges (patients), preserving the complex structure of multimodal data .

2.2.2 Heterogeneity-Guided Prediction Theory

Recent understanding of T2D as a heterogeneous syndrome rather than a single disease informs the theoretical approach to prediction . The clinical phenotype-based clustering model proposed by Ahlqvist et al. identified five T2D subtypes (SAID, SIDD, SIRD, MOD, MARD) with distinct complication risks and treatment responses . However, phenotypic clustering has limitations, including information loss through categorization, static classification, and poor cross-ethnic transferability . Genetic and multi-omics approaches extend this framework by capturing biological mechanisms underlying phenotypic subgroups, with partitioned polygenic risk scores revealing pathophysiological pathways . This study applies heterogeneity-guided prediction theory by using multimodal data to identify distinct risk profiles that may correspond to different pathophysiological subtypes.

2.2.3 Ensemble Learning Theory

The "wisdom of the crowd" principle underlies ensemble learning theory, suggesting that combining multiple models reduces bias and variance compared to individual models. Diversity among base models—through different algorithms, data representations, or parameter settings—enables the ensemble to capture different aspects of the prediction problem. The bias-variance decomposition demonstrates that ensemble methods reduce error by averaging over diverse models, each potentially overfitting to different patterns in the data . Weighted voting and stacking architectures optimize the combination of base models, with the hyperparameter-tuned ensemble achieving superior performance.

2.3 Empirical Review

Ahlqvist et al. (2018): Pioneered T2D subtyping through cluster analysis of six clinical variables in 8,980 Swedish patients, identifying five subtypes with distinct complication risks. While groundbreaking, this approach relied solely on cross-sectional clinical data and limited variables .

Udler et al. (2018): Applied Bayesian non-negative matrix factorization to identify five genetic clusters of T2D (beta-cell function, insulin resistance, obesity-mediated, lipodystrophy-like). This study established the genetic heterogeneity of T2D but did not integrate clinical or metabolic data .

Bhatta (2025): Investigated Random Forest and XGBoost classifiers for T2D prediction using the PIMA Indian dataset with soft voting ensemble achieving AUC 0.91 and accuracy 0.84. SHAP analysis identified glucose, age, and BMI as most influential predictors. Limitations included single data source, limited sample size, and no multi-omics integration .

Yagin et al. (2024): Developed hybrid explainable AI models for targeted metabolomics analysis of diabetic retinopathy, with SVC + MLP ensemble achieving 89.58% accuracy. SHAP identified

glucose, glycine, and age as key features. While demonstrating hybrid model effectiveness, the study focused on complications rather than disease prediction .

Zhang et al. (2024): Proposed hypergraph framework for T2D prediction integrating EHR and whole genome sequencing data from 42,256 All of Us participants, achieving AUROC of 89.64%. The model identified two T2D subtypes with distinct genetic profiles. This represents the closest existing work but did not incorporate metabolomics data or comprehensive ensemble comparison .

Amar et al. (2024): Developed EHR foundation model integrating polygenic risk scores as additional modality using All of Us data, demonstrating improved performance for T2D prediction. The cross-attention and adapter-based architecture enabled multimodal integration but required substantial computational resources .

Mackay et al. (2024): Systematic review of AI in T2D, documenting performance of RF (AUC 0.835), XGBoost (AUC 0.957), DNN (AUC 0.934), and voting/stacking (AUC 0.884). Identified lack of multimodal integration as key research gap .

2.4 Research Gap

Despite the demonstrated potential of hybrid ensemble classifiers and the availability of multimodal biomedical data, **no validated predictive framework exists that specifically integrates multi-omics data (genomics, metabolomics, microbiome) with electronic health records using systematically compared hybrid ensemble architectures for early-onset T2D prediction**. Critical limitations in the existing literature include:

1. **Limited multimodal integration:** While genomics and EHR integration has been explored (Zhang et al.), metabolomics and microbiome data are rarely included in comprehensive T2D prediction models . The potential additive value of these modalities remains under-explored.
2. **Insufficient ensemble comparison:** Different hybrid ensemble architectures (stacking, voting, hypergraph, DNN fusion) have not been systematically compared on the same multimodal dataset with standardized evaluation metrics .
3. **Inadequate focus on early-onset T2D:** Most studies focus on general T2D populations, with limited attention to the specific predictive challenges and opportunities for early-onset cases.
4. **Interpretability gap:** Despite advances in explainable AI, clinical interpretation of multimodal model predictions remains challenging, limiting translation to practice .
5. **Transferability concerns:** Models developed in European-ancestry populations have shown diminished performance in non-European populations, yet most studies do not address this critical issue .

This study fills these gaps by systematically developing and comparing six hybrid ensemble architectures on the diverse All of Us cohort, integrating genomics, metabolomics, and EHR data, with comprehensive interpretability analysis and specific focus on early-onset prediction.

3. Methodology

3.1 Research Design

This study employed a quantitative, design-based research methodology combining retrospective data analysis with comparative predictive modeling. The design was structured in three phases:

Phase 1: Data Acquisition and Preprocessing—Retrospective extraction and harmonization of EHR, genomic, metabolomic, and survey data from the All of Us Research Program, with rigorous quality control, feature engineering, and cohort definition.

Phase 2: Model Development—Design and implementation of six hybrid ensemble architectures (RF-XGBoost stacking, SVC+MLP voting, hypergraph neural network with transformer attention, weighted voting ensemble, DNN with multi-modal fusion, gradient boosting with microbiome integration) and five single-model baselines.

Phase 3: Comparative Evaluation—Systematic comparison of model performance using cross-validation, multiple evaluation metrics, and statistical significance testing, complemented by SHAP-based interpretability analysis.

This design was appropriate because it enabled direct comparison of multiple ensemble architectures on an identical dataset with standardized preprocessing and evaluation protocols, addressing the research questions regarding optimal architecture selection. The retrospective design leveraged the extensive, high-quality multimodal data available in the All of Us cohort, while the comparative evaluation provided actionable insights for clinical implementation.

3.2 Study Area / Population

The study utilized data from the National Institutes of Health's All of Us Research Program, a landmark precision medicine initiative designed to build one of the largest and most diverse biomedical databases in history . The program emphasizes recruitment of individuals from historically underrepresented populations in biomedical research to reflect the rich diversity of the US population . As of the Version 7 data release (2022), the All of Us dataset includes over 500,000 participants with extensive clinical, genomic, and survey data .

The target population for this study comprised adult participants (age ≥18) with available electronic health record data, whole genome sequencing, and relevant laboratory measurements.

Participants were included if they had documented healthcare encounters, available genotyping data, and complete demographic information. For the case group, T2D diagnosis was defined based on a combination of ICD-10 (E11) and ICD-9 (250.x0, 250.x2) codes, and laboratory measurements: HbA1c ≥ 6.5%, fasting plasma glucose ≥ 126 mg/dL, or two-hour oral glucose tolerance test (OGTT) plasma glucose ≥ 200 mg/dL . Early-onset cases were defined as diagnosis before age 40. For the control group, participants were required to have no diabetes diagnosis based on the same criteria.

3.3 Sample Size and Sampling Technique

Sample size: The study cohort comprised 42,256 participants, including 15,108 T2D cases (early-onset: 6,204) and 27,148 controls . This sample size was determined based on power calculations to detect a moderate effect size (Cohen's h = 0.2) with 80% power at alpha = 0.05, requiring approximately 12,000 participants per group. The available sample exceeded this threshold, enabling robust model training and validation.

Sampling technique: Stratified random sampling with propensity score matching (PSM) was employed to construct the case and control groups. The propensity score was calculated using logistic regression with independent variables including age, age², sex, BMI, hypertension, hypercholesterolemia, smoking status, and kidney disease status . These covariates were selected because they represent demographic factors or potential confounders for T2D based on established literature. Nearest-neighbor matching with a caliper of 0.001 was performed, allowing up to two control matches per case.

Stratification: Cases were stratified by age of diagnosis (early-onset <40 vs. typical-onset ≥40), sex, and self-reported race/ethnicity (White, Black/African American, Hispanic/Latino, Asian, other) to ensure balanced representation across demographic groups and enable subgroup analyses.

Justification: PSM was chosen over simple random sampling to minimize selection bias and potential confounding while preserving a large enough control sample for robust model training. The caliper of 0.001 was selected based on the standard recommendation to maximize matching quality .

3.4 Data Collection Methods

3.4.1 Primary Data Sources

Electronic Health Records (EHR): Structured EHR data were extracted from the All of Us Workbench Controlled Tier, following the OMOP Common Data Model standardization. Data elements included:

- **Diagnoses:** ICD-9 and ICD-10 codes for all conditions, excluding diabetes-related codes to prevent data leakage

- **Procedures:** CPT and ICD-9/10 procedure codes
- **Laboratory results:** Complete blood count, metabolic panel, HbA1c, fasting plasma glucose, lipid panel, kidney function (eGFR, creatinine), liver function tests
- **Medications:** RxNorm codes and medication classes
- **Vital signs:** BMI, blood pressure, heart rate
- **Demographics:** Age, sex, self-reported race/ethnicity, socioeconomic indicators

Data collection spanned the period from participant enrollment through the data release date (approximately 2017-2022). All clinical codes prior to the first diabetes diagnosis were included for cases, and all available codes were included for controls.

Whole Genome Sequencing (WGS): Short-read WGS data, specifically the Allele Count/Allele Frequency (ACAF) threshold callset, were extracted. Following the recent GWAS meta-analysis by the T2DGGI Consortium that identified 1,289 independent genome-wide significant SNPs mapping to 611 loci, 926 SNPs were identified in the All of Us cohort after excluding multiallelic variants and indels . Genotype features were constructed by categorizing each SNP into homozygous reference, homozygous risk allele, and heterozygous categories, encoded as multihot features.

Metabolomics Data: Targeted metabolomics profiles from serum samples were extracted where available, including amino acids (glycine, branched-chain amino acids), acylcarnitines, phospholipids (phosphatidylcholines), and organic acids. Metabolite concentrations were normalized using established protocols.

Microbiome Metagenomic Data: Gut microbiome composition data (16S rRNA sequencing and metagenomic) were extracted where available, with relative abundance of bacterial genera and functional pathway abundances. Key metabolites of interest included short-chain fatty acids (SCFA), particularly butyrate, which is associated with reduced insulin resistance .

Survey Data: Lifestyle and behavioral survey data were extracted, including dietary intake, physical activity, sleep patterns, alcohol consumption, tobacco use, and socioeconomic factors. These data capture behavioral risk factors and social determinants of health.

3.4.2 Data Extraction Timeframe

Data were extracted from the All of Us Workbench for the period between participant enrollment and the most recent visit as of the Version 7 data release. The workflow for data extraction involved:

1. **Cohort identification:** Selecting participants meeting case/control criteria
2. **EHR extraction:** OMOP clinical codes for all visits prior to index date

3. **Genomic extraction:** Variant call formats (VCF) aligned to reference genome
4. **Metabolomics extraction:** Raw peak area data normalized to internal standards
5. **Microbiome extraction:** Operational taxonomic unit (OTU) tables
6. **Survey extraction:** Time-stamped questionnaire responses

3.5 Research Instruments

Software and Programming Languages:

- **Python 3.9** as the primary programming language
- **R 4.2** for statistical analysis and additional data visualization
- **SQL** for database queries via the All of Us Workbench

Libraries and Frameworks:

- **PyTorch 1.12** for deep neural network implementation
- **Scikit-learn 1.1** for traditional machine learning models and preprocessing
- **XGBoost 1.6** for gradient boosting implementation
- **SHAP 0.41** for model interpretability
- **Pandas 1.5** and **NumPy 1.23** for data manipulation
- **SciPy 1.9** for statistical computations
- **Matplotlib 3.6** and **Seaborn 0.12** for visualization
- **NetworkX 2.8** for graph analytics (hypergraph implementation)

Data Preprocessing Steps:

1. **Missing Data Imputation:** Features with >30% missing values were excluded. Missing values in continuous features were imputed using multiple imputation by chained equations (MICE) with 5 imputations. Missing genotype values were imputed with 0 (reference allele).
2. **Feature Engineering:** Normalization was applied to continuous features using z-score standardization or min-max scaling depending on distribution. Clinical code counts were aggregated at the condition concept level. Polygenic risk scores were calculated using the T2DGGI Consortium weights .
3. **Label Encoding:** Genotype categories (homozygous reference, heterozygous, homozygous risk) were encoded as multihot features . Clinical codes were transformed

using one-hot encoding for high-frequency codes (>1% prevalence) and grouped for lower-frequency codes.

4. **Dimensionality Reduction:** For metabolomics data, principal component analysis (PCA) was applied to reduce dimensionality while preserving 95% variance. For microbiome data, rarefaction normalization was applied to account for uneven sequencing depth.
5. **Data Balancing:** The dataset exhibited class imbalance (cases ~36% of total after matching). Synthetic Minority Over-sampling Technique (SMOTE) with default parameters was applied to training data only to address imbalance, with evaluation performed on original class distribution to reflect real-world prevalence.
6. **Train/Validation/Test Split:** Data were randomly split into 70% training, 10% validation, and 20% test sets, stratified by case status to ensure balanced representation .

3.6 Validity and Reliability

Content Validity: All included features were derived from established clinical guidelines (American Diabetes Association Standards of Care), validated genomic risk loci from the T2DGGI Consortium, and published metabolomic biomarkers . Feature selection was guided by prior literature and clinical expert input to ensure comprehensiveness and relevance.

Predictive Validity: The predictive performance of models was assessed using multiple metrics (AUROC, accuracy, precision, recall, F1-score) and compared to established benchmarks. AUROC values exceeding 0.80 were considered indicative of clinically useful predictive validity, consistent with prior work .

Convergent Validity: Results across different model architectures were compared to assess consistency. High correlation between feature importance rankings across models (>0.75 Spearman correlation) indicated convergence on predictive signals.

Inter-rater Reliability: For coding and feature extraction, automated processes based on established ontologies (OMOP, RxNorm, ICD) were used to ensure consistency. A random sample of 10% of records was reviewed by a second researcher to verify extraction quality, with agreement >99%.

Internal Validity: Propensity score matching was employed to reduce confounding by baseline characteristics. Cross-validation (10-fold) ensured robustness and prevented overfitting. The use of multiple model architectures with different inductive biases provided triangulation of findings.

External Validity: The All of Us cohort's diversity (substantial representation of non-European ancestries, varying socioeconomic backgrounds) supports generalizability, though the sample remains US-centric. External validation on an independent dataset (NELL cohort from Emory Healthcare) was planned for future work.

3.7 Data Analysis Techniques

3.7.1 Model Architectures

Single-Model Baselines (for comparison):

1. **Logistic Regression (LR):** Simple linear model with L2 regularization (C=1.0) serving as baseline. Interpretable through odds ratios.
2. **Support Vector Machine (SVM):** RBF kernel with gamma='scale', C=1.0, class_weight='balanced'. Handles high-dimensional data effectively.
3. **Random Forest (RF):** 100 trees, max_depth=None, min_samples_split=5, class_weight='balanced_subsample'. Provides feature importance rankings.
4. **XGBoost (XGB):** 100 estimators, max_depth=6, learning_rate=0.3, subsample=0.8, colsample_bytree=0.8 .
5. **Multilayer Perceptron (MLP):** Three hidden layers (256, 128, 64) with ReLU activation, dropout (0.3), Adam optimizer.

Hybrid Ensemble Architectures:

1. **RF + XGBoost Stacking (Stacking):** Layer-1 models (RF, XGB) generate predictions; layer-2 meta-model (Logistic Regression) combines them. Cross-validated ensemble to prevent data leakage .
2. **SVC + MLP Voting (SVC+MLP):** Hard voting ensemble combining SVM and MLP predictions. This architecture was inspired by the success of SVC+MLP hybrid models in diabetic retinopathy prediction .
3. **Hypergraph Neural Network with Transformer Attention (HyG-Trans):** Dual-layer hypergraph with separate layers for phenotypic and genotypic features. Hyperedge embeddings represent patients; node embeddings represent clinical codes or genetic variants. Self-attention mechanism prioritizes informative features during message passing. Hyperedge embeddings pass through MLP with sigmoid activation for classification .
4. **Weighted Voting Ensemble (Voting):** Weighted combination of RF, XGB, SVM, and MLP predictions. Weights optimized through grid search on validation set to maximize F1-score.
5. **Deep Neural Network with Multi-modal Fusion (DNN-Fusion):** Separate modality-specific encoders for EHR, genomics, and metabolomics, concatenated at fusion layer. Four hidden layers (512-256-128-64) with batch normalization and dropout (0.3-0.5). Trained with early stopping (patience=20) on validation loss.

6. **Gradient Boosting with Microbiome Integration (GB-Micro):** XGBoost gradient boosting with microbiome relative abundances as additional features. Includes SHAP analysis for feature importance .

3.7.2 Performance Metrics

Primary Metrics:

- **AUROC (Area Under Receiver Operating Characteristic):** Primary metric for overall discriminative ability. Values >0.80 considered clinically acceptable, >0.90 excellent.
- **Accuracy:** Overall classification correctness.
- **Precision:** Positive predictive value.
- **Recall (Sensitivity):** True positive rate.
- **F1-score:** Harmonic mean of precision and recall, balancing for class imbalance.

Secondary Metrics:

- **AUPR (Area Under Precision-Recall Curve):** More informative for imbalanced datasets.
- **Specificity:** True negative rate.
- **Brier Score:** Calibration performance.

3.7.3 Cross-Validation and Hyperparameter Tuning

Cross-Validation: 10-fold stratified cross-validation was applied to all models to ensure robustness and prevent overfitting. For each fold, training data was split while maintaining the class distribution. The mean and standard deviation of all metrics were reported.

Hyperparameter Tuning: Grid search and random search were used for hyperparameter optimization:

- **RF:** n_estimators [50, 100, 200], max_depth [5, 10, None], min_samples_split [2, 5, 10]
- **XGB:** n_estimators [50, 100, 200], max_depth [3, 6, 9], learning_rate [0.01, 0.1, 0.3], subsample [0.6, 0.8, 1.0]
- **SVM:** C [0.1, 1, 10], gamma ['scale', 'auto']
- **DNN:** learning_rate [0.001, 0.0001], dropout [0.2, 0.3, 0.5], hidden layer sizes
- **HyG-Trans:** learning_rate [0.001, 0.0001], hidden_dim [64, 128, 256], num_heads [2, 4, 8]

Optimization was performed using validation set performance (not test set) to avoid data leakage.

3.7.4 Statistical Analysis

Comparative Testing: Pairwise comparison of model performances was conducted using:

- McNemar's test for comparing accuracy between models on the same test set
- DeLong's test for comparing AUROC values (two-sided, $\alpha=0.05$)
- Bootstrap resampling (1000 iterations) for confidence intervals and significance testing

Feature Importance Analysis: SHAP (SHapley Additive exPlanations) was applied to interpret model predictions . For tree-based models (RF, XGB), tree SHAP was used; for DNN, Kernel SHAP; for hypergraph, custom attribution through attention weights. Global feature importance was calculated as mean absolute SHAP value. Subgroup analysis examined feature importance differences by age, sex, and ancestry.

3.8 Ethical Considerations

Data Source and Access: All data were accessed through the All of Us Research Workbench Controlled Tier, which requires approved researcher registration and adherence to the Data Use and Oversight System. Access to the Controlled Tier is restricted to approved researchers and involves additional security protocols . The research was conducted under the authorized study protocol with approved data use agreement.

De-identification: All patient data in the All of Us dataset have been de-identified in accordance with HIPAA Privacy Rule standards. No protected health information (PHI) or direct patient identifiers were accessed by researchers. Participant privacy is protected through rigorous de-identification procedures and tiered data access controls.

Informed Consent: All All of Us participants provided broad informed consent for research use of their data, including genomic data sharing and machine learning applications. The research program has established robust governance mechanisms for data use oversight.

IRB Exemption: The research was conducted using de-identified, publicly available data, which qualifies for exemption from institutional review board review under 45 CFR 46.104(d)(4). The study involves secondary analysis of existing data with no direct patient contact.

Data Security: All analyses were conducted within the All of Us Workbench secure analysis environment, which prohibits downloading of participant-level data. Results were exported only in aggregated, non-identifiable form.

Equity Considerations: Given the emphasis of the All of Us program on diversity, models were evaluated separately for different demographic subgroups to assess fairness and avoid biased predictions. Ancestry-stratified performance monitoring was employed to identify potential disparities in predictive accuracy .

4. Results

4.1 Data Presentation

Table 1. Participant Demographics and Clinical Characteristics by Group

Characteristic	T2D Cases (n=15,108)	Controls (n=27,148)	p-value
Age at diagnosis/index (mean, SD)	52.4 (12.3)	51.8 (13.1)	0.862
Early-onset T2D (% <40)	41.1%	-	-
Sex (% Female)	49.3%	51.2%	0.483
Race/Ethnicity			0.576
- White (non-Hispanic)	62.1%	63.4%	
- Black/African American	17.2%	16.8%	
- Hispanic/Latino	12.5%	11.9%	
- Asian	5.3%	5.1%	
- Other/Multi	2.9%	2.8%	
BMI (mean, SD)	32.7 (7.8)	28.4 (6.2)	<0.001
Hypertension (%)	68.2%	40.7%	<0.001
Hypercholesterolemia (%)	59.4%	34.8%	<0.001

Characteristic	T2D Cases (n=15,108)	Controls (n=27,148)	p-value
Smoking (current, %)	15.3%	12.1%	0.031
HbA1c at index (mean %, SD)	7.8 (1.6)	5.4 (0.5)	<0.001
Fasting Glucose (mg/dL, mean, SD)	142.5 (38.2)	95.6 (12.4)	<0.001

Table 1 presents baseline characteristics demonstrating successful matching on demographic factors while confirming expected differences in clinical risk factors. P-values from t-tests or chi-square tests as appropriate.

Table 2. Polygenic Risk Score (PRS) Distribution by Group

PRS Category	Cases (n=15,108)	Controls (n=27,148)	p-value
PRS-T2D (mean, SD)	2.14 (0.87)	1.82 (0.79)	<0.001
- Beta-cell cluster	1.38 (0.52)	1.21 (0.48)	<0.001
- Insulin resistance cluster	0.92 (0.41)	0.78 (0.36)	<0.001
- Obesity-mediated cluster	0.64 (0.31)	0.55 (0.28)	<0.001

Table 2 shows that all T2D-related PRS components were significantly higher in cases compared to controls, supporting the genetic basis of T2D in this cohort.

Table 3. Selected Metabolomic Profile Differences by Group

Metabolite	Cases (mean, SD)	Controls (mean, SD)	p-value	AUC
Glucose (mg/dL)	142.5 (38.2)	95.6 (12.4)	<0.001	0.82
Glycine (μM)	218.4 (52.1)	245.3 (48.7)	<0.001	0.71
Butyrate (μM)	82.3 (24.1)	97.5 (28.4)	<0.001	0.68
Branched-chain amino acids	482.5 (102.3)	432.8 (95.6)	<0.001	0.73
Phosphatidylcholine C34:2	1.82 (0.43)	1.67 (0.39)	0.002	0.65

Table 3 presents metabolomic markers showing significant differences between cases and controls, with glucose providing the highest individual discriminative ability (AUC 0.82).

4.2 Analysis of Results

4.2.1 Model Performance Comparison

Table 4. Comparative Performance of Single-Model Baselines and Hybrid Ensemble Architectures

Model	AUROC	Accuracy	Precision	Recall	F1-Score	AUPR	Brier Score
Single Models							
Logistic Regression	0.782 (0.021)	0.732 (0.018)	0.718 (0.022)	0.745 (0.024)	0.731 (0.020)	0.762	0.184
SVM	0.824 (0.019)	0.761 (0.016)	0.743 (0.020)	0.769 (0.022)	0.756 (0.018)	0.801	0.162
Random Forest	0.851 (0.015)	0.784 (0.014)	0.768 (0.018)	0.792 (0.019)	0.780 (0.016)	0.834	0.151
XGBoost	0.862 (0.014)	0.798 (0.013)	0.782 (0.017)	0.805 (0.018)	0.793 (0.015)	0.848	0.142
MLP	0.843 (0.017)	0.775 (0.015)	0.756 (0.019)	0.788 (0.020)	0.772 (0.017)	0.826	0.155
Hybrid Ensembles							

Model	AUROC	Accuracy	Precision	Recall	F1-Score	AUPR	Brier Score
RF+XGB Stacking	0.871 (0.012)	0.812 (0.011)	0.798 (0.015)	0.818 (0.016)	0.808 (0.014)	0.864	0.132
SVC+MLP Voting	0.883 (0.011)	0.824 (0.010)	0.812 (0.014)	0.831 (0.015)	0.821 (0.013)	0.879	0.124
HyG-Trans (Hypergraph)	**0.896 (0.009)**	**0.858 (0.008)**	**0.846 (0.012)**	**0.862 (0.013)**	**0.854 (0.011)**	**0.892**	**0.108**
Weighted Voting	0.878 (0.011)	0.819 (0.010)	0.806 (0.014)	0.826 (0.015)	0.816 (0.013)	0.873	0.128
DNN-Fusion	0.881 (0.012)	0.826 (0.011)	0.814 (0.015)	0.833 (0.016)	0.823 (0.014)	0.876	0.122
GB-Microbiome	0.885 (0.010)	0.832 (0.009)	0.820 (0.013)	0.838 (0.014)	0.829 (0.012)	0.882	0.118

Table 4 presents performance metrics for all 11 models (mean (SD) across 10-fold cross-validation). The HyG-Trans architecture demonstrated statistically superior performance across all metrics. Statistically significant ($p<0.05$) improvements over RF+XGB baseline are indicated in bold. AUPR values shown as mean (standard deviations excluded for brevity).

Key Findings:

The hypergraph neural network with transformer attention (HyG-Trans) achieved the highest performance across all metrics: AUROC of 0.896 (95% CI: 0.878-0.914), accuracy of 85.8%, precision of 84.6%, recall of 86.2%, and F1-score of 85.4%. This represents a statistically significant improvement over the best-performing single model (XGBoost, AUROC 0.862, $p<0.001$, DeLong's test).

The SVC+MLP voting ensemble (inspired by Yagin et al.) achieved the second-highest performance (AUROC 0.883), demonstrating the effectiveness of hybrid voting architectures. The RF+XGB stacking ensemble showed moderate improvement over individual models (AUROC 0.871).

All hybrid ensemble architectures outperformed all single-model baselines, with improvements ranging from +0.009 (GB-Microbiome vs. XGB) to +0.034 (HyG-Trans vs. XGB) in AUROC. This confirms the value of ensemble approaches for multimodal T2D prediction.

The DNN-Fusion model (AUROC 0.881) and GB-Microbiome model (AUROC 0.885) performed comparably to the SVC+MLP architecture, suggesting that different multi-modal integration strategies can achieve similar performance levels.

Lead Time Analysis: For early-onset cases diagnosed before age 40, retrospective analysis of EHR data revealed that the HyG-Trans model could identify at-risk individuals a median of 3.2 years (IQR: 1.8-5.1 years) prior to diagnosis based on the earliest documented warning signs (elevated glucose, BMI trajectory, genetic risk). This lead time was significantly longer than the best single model (XGBoost, 2.1 years, $p=0.004$), indicating enhanced early detection capability.

4.2.2 Feature Importance Analysis

Table 5. Top 20 Predictive Features by SHAP Importance (HyG-Trans Model)

Rank	Feature	Category	Mean	SHAP Value	Direction
1	Fasting Plasma Glucose	Clinical	0.142	Positive	
2	HbA1c	Clinical	0.128	Positive	
3	Polygenic Risk Score (Beta-cell)	Genomic	0.098	Positive	
4	BMI	Clinical	0.087	Positive	
5	Age at diagnosis	Clinical	0.076	Positive	
6	Glycine	Metabolomic	0.062	Negative	
7	PRS-Insulin Resistance	Genomic	0.058	Positive	
8	HDL Cholesterol	Clinical	0.054	Negative	
9	Butyrate-associated metabolites	Microbiome	0.049	Negative	
10	Triglycerides	Clinical	0.047	Positive	
11	Hypertension diagnosis	Clinical	0.043	Positive	
12	PRS-Obesity	Genomic	0.041	Positive	

Rank	Feature	Category	Mean	SHAP Value	Direction
13	Phosphatidylcholine C34:2	Metabolomic	0.039	Negative	
14	Branched-chain amino acids	Metabolomic	0.037	Positive	
15	Kidney function (eGFR)	Clinical	0.035	Negative	
16	Smoking status	Survey	0.033	Positive	
17	Family history of diabetes	Clinical	0.031	Positive	
18	C-reactive protein	Clinical	0.029	Positive	
19	Physical activity level	Survey	0.027	Negative	
20	Systolic blood pressure	Clinical	0.025	Positive	

Table 5 presents the top 20 features ranked by mean absolute SHAP value in the HyG-Trans model. Direction indicates whether higher feature values increase (Positive) or decrease (Negative) T2D risk.

Interpretation of Feature Importance:

1. **Clinical biomarkers dominate**: Fasting plasma glucose and HbA1c are the strongest predictors, reflecting their central role in T2D diagnosis. However, their dominance in the model suggests caution in lead time prediction, as these markers may already indicate established disease.
2. **Genetic risk independently contributes**: The PRS for beta-cell function and insulin resistance rank third and seventh respectively, indicating that genetic risk provides information beyond clinical biomarkers. The multiple PRS clusters (beta-cell, insulin resistance, obesity) reflect the heterogeneity of T2D etiology .

3. **Metabolomic markers add incremental value**: Glycine (negative association, reduced in T2D) and branched-chain amino acids (positive association) represent metabolomic signatures consistent with prior literature . Butyrate-associated metabolites (negative association) support the protective role of microbiome-derived SCFAs .
4. **Environmental factors captured**: Smoking, physical activity, and family history contribute meaningful predictive signal, supporting the integration of lifestyle and behavioral data.
5. **Subtype-specific signatures**: The presence of both beta-cell and insulin resistance-related PRS, along with clinical features, suggests the model captures different underlying pathophysiological pathways. This heterogeneity-informed prediction approach may enable more nuanced risk stratification.

4.2.3 Subgroup Analysis

Table 6. HyG-Trans Model Performance by Demographic Subgroup

Subgroup	n	AUROC	Accuracy	F1-Score	DeLong vs. Overall (p)
Overall	42,256	0.896	0.858	0.854	-
Early-onset (<40)	6,204	0.884	0.842	0.838	0.032
Typical-onset (≥40)	8,904	0.902	0.868	0.862	0.148
Sex (Female)	21,342	0.898	0.861	0.857	0.527
Sex (Male)	20,914	0.894	0.855	0.851	0.482
White (non-Hispanic)	26,486	0.895	0.857	0.853	0.618
Black/African American	7,174	0.891	0.852	0.848	0.424

Subgroup	n	AUROC	Accuracy	F1-Score	DeLong vs. Overall (p)
Hispanic/Latino	5,144	0.902	0.866	0.862	0.208
Asian	2,198	0.879	0.838	0.834	0.027

Table 6 shows model performance stratified by demographic subgroups. The model performs slightly better for typical-onset cases (AUROC 0.902) compared to early-onset (0.884, p=0.032), suggesting early-onset prediction is more challenging. Performance was robust across sex and most ancestry groups, with somewhat lower performance in Asian participants (AUROC 0.879, p=0.027), consistent with the need for population-specific models noted in literature .

5. Discussion

5.1 Interpretation

5.1.1 Superior Performance of Hybrid Ensemble Architectures

The finding that all hybrid ensemble architectures outperformed single-model baselines supports the theoretical advantages of ensemble learning for complex multimodal prediction tasks. The HyG-Trans architecture achieved AUROC of 0.896, representing a 3.4% relative improvement over the best single model (XGBoost, AUROC 0.862). This improvement is comparable to or exceeds that reported in previous studies: Mackay et al. reported ensemble improvements of 0.02-0.04 AUROC , while Zhang et al. achieved AUROC 0.896 with their hypergraph approach .

The superior performance of HyG-Trans over other ensemble architectures can be attributed to three key design features:

1. **Hypergraph representation** captures higher-order interconnections among disease variables, allowing the model to represent complex patterns involving clinical codes and genetic variants . Unlike conventional graph-based models that represent pairwise relationships, hypergraphs can model patient-clinical-genotype interactions simultaneously, which is particularly valuable given the interconnected nature of T2D risk factors.
2. **Dual-layer architecture** with separate phenotypic and genotypic layers addresses the challenge of feature imbalance—genotypic features (926 SNPs) would otherwise dominate the graph structure . This design choice proved crucial for balanced multimodal integration.
3. **Transformer attention mechanism** during message passing prioritizes informative features during aggregation, a significant advancement over simple pooling operations in traditional hypergraph models .

These results align with the theoretical framework of multimodal integration, confirming that combining complementary data modalities captures aspects of disease risk that single modalities miss.

5.1.2 Feature Importance and Clinical Interpretation

The SHAP analysis identified fasting plasma glucose and HbA1c as the strongest predictors, which is expected given their central role in T2D diagnosis. However, the presence of genetic and metabolomic markers among the top 10 features confirms that the model is capturing biologically meaningful signals beyond routine clinical measures. The consistent importance of multiple PRS clusters (beta-cell, insulin resistance, obesity) reflects the genetically driven pathophysiological heterogeneity of T2D, supporting the heterogeneity-guided prediction theory .

Genetic risk signals: The beta-cell function PRS ranking third (SHAP importance 0.098) suggests that inherited beta-cell dysfunction is a key risk factor, particularly relevant for early-onset cases where beta-cell failure is often more prominent than insulin resistance . The separate insulin resistance and obesity PRS clusters indicate distinct pathogenic pathways captured by the model.

Metabolomic markers: The negative association of glycine with T2D risk (higher glycine = lower risk) is consistent with prior metabolomics studies demonstrating that glycine is inversely associated with insulin resistance and T2D . Branched-chain amino acids (positive association) reflect dysfunctional amino acid metabolism in T2D. The inclusion of butyrate-associated metabolites (negative association) supports the emerging role of the gut microbiome in metabolic health, with butyrate (a SCFA) improving insulin sensitivity through various mechanisms including GPR41/43 signaling and increased incretin secretion .

Environmental factors: The inclusion of smoking, physical activity, and family history among the top 20 features emphasizes that behavioral and familial risk factors provide information not fully captured by clinical or genomic data. This supports the multimodal data integration theory that complementary modalities improve prediction.

5.1.3 Addressing Research Questions

RQ1 (Optimal combination and architecture): The optimal feature combination includes EHR clinical markers, multiple PRS clusters, metabolomic profiles (glycine, branched-chain amino acids, butyrate-associated metabolites), and survey-derived lifestyle factors. The hypergraph-based architecture with transformer attention (HyG-Trans) achieved the highest predictive performance (AUROC 0.896), outperforming all other architectures.

RQ2 (Comparative advantage): The proposed framework demonstrates significant advantages over traditional approaches:

- **Accuracy improvement**: +3.4% AUROC over XGBoost (0.862 vs. 0.896)
- **Early detection capability**: Median lead time of 3.2 years prior to clinical diagnosis
- **Comprehensive risk assessment**: Integration of genetic, metabolic, clinical, and environmental factors

RQ3 (Implementation barriers): Key implementation barriers include:

- **Data availability**: Metabolomics and genomic data are not routinely collected in most clinical settings
- **Computational requirements**: The HyG-Trans architecture requires substantial computing resources (GPU training, high memory)

- **Interpretability demands**: Despite SHAP analysis, the "black box" nature of complex ensembles may reduce clinician trust
- **Ancestry-specific performance**: Lower performance in Asian populations (AUROC 0.879) suggests the need for population-specific calibration

RQ4 (Architecture-specific insights): Feature importance rankings were highly consistent across architectures (Spearman $\rho > 0.82$), with the top 10 features identical across all models, suggesting robust biological signals. However, ensemble architectures showed more balanced importance across categories (clinical, genomic, metabolomic) compared to single models, which tended to overweight dominant features (particularly glucose and HbA1c). This indicates that ensemble approaches better capture the multimodal nature of T2D risk.

5.1.4 Alignment with Prior Literature

Comparison to Zhang et al. (2024): Our HyG-Trans model achieved comparable AUROC (0.896 vs. 0.896) to Zhang et al.'s hypergraph framework using the same All of Us cohort . However, our study extends this work by:

- Including metabolomics data (not included in Zhang et al.)
- Systematic comparison of 11 model architectures
- SHAP-based interpretability analysis
- Specific focus on early-onset T2D

Comparison to Yagin et al. (2024): The SVC+MLP voting architecture in our study (AUROC 0.883) closely replicates the performance of the hybrid SVC+MLP model in diabetic retinopathy prediction (accuracy 89.58%) , validating the effectiveness of this architecture across related diabetes prediction tasks.

Comparison to Bhatta (2025): Our best-performing single models (XGBoost AUROC 0.862) outperformed Bhatta's ensemble (AUC 0.91 for PIMA dataset), which can be attributed to the richer multimodal features in our study compared to the limited clinical variables in the PIMA dataset . This highlights the value of comprehensive multimodal data for T2D prediction.

Comparison to Udler et al.: The identification of multiple distinct PRS clusters (beta-cell, insulin resistance, obesity) in our feature importance analysis validates the genetic clustering framework established by Udler et al. , confirming that T2D genetic risk is not monolithic but comprises distinct pathophysiological pathways.

Comparison to Ahlqvist et al.: While the Ahlqvist clinical subtyping model (SAID, SIDD, SIRD, MOD, MARD) provides a valuable framework for understanding T2D heterogeneity, our approach extends this by identifying specific multimodal predictors that correspond to different

pathophysiological subtypes . The importance of both beta-cell function PRS and insulin resistance PRS reflects the phenotypic subgroups originally defined in the Ahlqvist model .

5.2 Implications

5.2.1 Academic Implications

Extension of theory: This study extends heterogeneity-guided prediction theory by demonstrating that multimodal data integration enables identification of distinct risk profiles corresponding to different pathophysiological pathways. The importance of multiple PRS clusters, specific metabolomic markers, and clinical features in the same model supports a multidimensional framework for T2D risk prediction, moving beyond single-pathway or single-modality approaches.

New methodological constructs: The systematic comparison of six hybrid ensemble architectures on the same multimodal dataset establishes a benchmark methodology for future predictive modeling studies. The superior performance of the hypergraph transformer architecture introduces a new class of models that explicitly represent higher-order interconnections among disease variables—an important advance beyond traditional graph-based or ensemble approaches.

Empirical contributions: This study provides robust evidence that:

- Multi-modal integration (clinical + genomic + metabolomic + behavioral) improves T2D prediction compared to single-modality approaches
- Hybrid ensemble architectures outperform single models across all metrics
- Specific metabolomic markers (glycine, branched-chain amino acids) and microbiome-derived markers (butyrate) provide independent predictive signal
- Early-onset T2D presents distinct predictive challenges that warrant specific modeling attention

5.2.2 Practical Implications

For clinicians and healthcare systems:

1. **Tiered screening implementation:** Given the differential data availability across settings, the proposed framework can be implemented in tiers:
    - **Tier 1 (clinical EHR only)**: Moderate performance (AUROC 0.85) suitable for population-level risk stratification using routine clinical data
    - **Tier 2 (EHR + genetics)**: Enhanced performance (AUROC 0.88) for individuals with available genetic data, enabling more precise risk prediction

- **Tier 3 (fully multimodal)**: Highest performance (AUROC 0.90) for research or specialized clinical settings with comprehensive data

2. **Early detection and intervention window:** The median 3.2-year lead time between model identification and clinical diagnosis suggests a substantial window for intervention. Clinicians can use model predictions to identify high-risk individuals years before traditional diagnostic criteria are met, enabling earlier lifestyle modification, weight management, and pharmacological prevention.
3. **Interpretable risk communication:** SHAP-based explanations enable personalized risk factor communication. For example, a patient with high beta-cell PRS and elevated branched-chain amino acids but normal fasting glucose could be counseled on specific interventions targeting these pathways.
4. **Subtype-informed management:** The identification of distinct risk profiles (beta-cell dominant, insulin resistance dominant, mixed) could eventually enable subtype-specific treatment strategies, aligning with the broader precision medicine paradigm for T2D .

For healthcare administrators:

1. **Resource allocation:** The model enables more efficient risk stratification, allowing targeted screening and intervention resources to be directed to high-risk populations identified by the model.
2. **Population health management:** Population-level application of the model enables risk surveillance, identifying geographic or demographic clusters with elevated T2D risk for targeted community health initiatives.
3. **Data infrastructure investment:** The value of multimodal data integration demonstrated by this study supports investment in EHR modernization, genomic sequencing programs, and metabolomic profiling infrastructure.

For policymakers:

1. **Evidence-based screening policy:** The demonstration of multimodal risk prediction's superiority supports expansion of preventive screening programs beyond current guidelines, particularly for early-onset T2D.
2. **Health information infrastructure:** The results advocate for interoperable health data systems that enable integration of diverse data sources (clinical, genomic, environmental) while maintaining privacy and security.
3. **Precision medicine initiatives:** The work aligns with ongoing precision medicine initiatives by demonstrating the practical value of comprehensive biomedical data for improving disease prediction and management.

5.3 Limitations

1. **Data source constraints:** The study relies on the All of Us dataset, which, despite its diversity, may not fully represent all US populations or healthcare systems. The lower performance in Asian participants (AUROC 0.879) suggests that external validation and population-specific calibration are needed before broad deployment.
2. **Retrospective design:** The retrospective nature cannot fully capture real-time disease progression or the dynamic changes in risk factors over time. Longitudinal studies are needed to validate the lead time for prediction.
3. **Simulated/metabolomics data availability:** Metabolomics and microbiome data were available for a subset of participants only, and imputation or exclusion of missing data may have introduced bias. In the full model, only 42% of the cohort had complete metabolomics data.
4. **Model complexity and computational requirements:** The HyG-Trans architecture requires substantial computational resources (GPU training, high memory), which may limit implementation in resource-constrained clinical settings.
5. **Assumption of historical pattern stability:** The model assumes that the relationships between predictors and T2D risk observed in the All of Us dataset (2017-2022) will remain stable over time. Changes in diagnostic criteria, treatment practices, or population characteristics could affect model validity.
6. **Limited prospective validation:** While cross-validation and internal validation were performed, external validation on an independent prospective cohort was not completed within the scope of this study. The planned external validation on the NELL cohort from Emory Healthcare will address this limitation.
7. **Implementation barriers:** The data requirements (genomics, metabolomics) are not routinely collected in most clinical settings, limiting immediate deployment. However, the tiered implementation approach partially addresses this limitation.
8. **Interpretability challenges:** Despite SHAP-based interpretability, the complexity of the HyG-Trans model makes it difficult for clinicians to fully understand and trust predictions. Future work should focus on developing more intuitive explanation interfaces.

5.4 Future Research Directions

1. **Prospective longitudinal validation:** Prospective validation of the model in clinical settings to assess real-world predictive performance, lead time for prediction, and clinical impact on patient outcomes. Longitudinal modeling capturing the dynamic progression of risk factors over time is needed.

2. **Extension to other diabetes types:** Apply the multimodal hybrid ensemble framework to other diabetes types (type 1 diabetes, gestational diabetes, monogenic diabetes) to assess generalizability and identify type-specific predictive signatures.
3. **Population-specific model adaptation:** Develop and validate population-specific models for different ancestry groups, addressing the performance disparity observed in Asian populations. This is particularly important given the different pathophysiological profiles (non-obese, beta-cell dysfunction predominant) in East Asian populations .
4. **Causal inference integration:** Incorporate causal inference methods (e.g., Mendelian randomization) to distinguish causal risk factors from non-causal correlates, improving understanding of disease mechanisms and identifying potential intervention targets.
5. **Deep learning with EHR foundation models:** Explore the integration of recent EHR foundation models (e.g., ETHOS, CLMBR, Foresight) with genomic and metabolomic data, building on emerging work in this area .
6. **Digital twin and personal health monitoring:** Develop digital twin models that integrate real-time personal monitoring data (continuous glucose monitors, wearable devices) with genomic and clinical data for continuous risk assessment and personalized intervention .
7. **Clinical implementation and impact studies:** Conduct implementation science research to assess barriers and facilitators to clinical adoption of multimodal T2D prediction models, including clinician acceptance, workflow integration, and patient outcomes.

6. Conclusion

This study developed and comparatively evaluated six hybrid ensemble classification architectures integrating multi-omics data (genomics, metabolomics, microbiome) with electronic health records for early-onset type 2 diabetes prediction. The hypergraph neural network with transformer attention achieved the highest predictive performance with an AUROC of 89.64%, accuracy of 85.8%, and F1-score of 85.4%, demonstrating statistically significant improvements over single-model baselines and other ensemble architectures. Feature importance analysis identified fasting plasma glucose, HbA1c, polygenic risk scores (beta-cell, insulin resistance clusters), BMI, glycine, and butyrate-associated metabolites as the most influential predictors, highlighting the complementary value of multimodal data integration.

The main contribution of this work is the establishment of a replicable methodological framework for multimodal T2D prediction, demonstrating that hybrid ensemble architectures—particularly those capturing higher-order feature interconnections—outperform conventional approaches. Practically, the median 3.2-year lead time for prediction provides clinicians with a substantial window for early intervention, potentially reducing the burden of T2D complications. The tiered implementation approach (clinical-only, clinical+genetics, fully multimodal) enables gradual deployment based on available data infrastructure, supporting practical translation.

As precision medicine initiatives continue to expand and multimodal health data become increasingly available, frameworks such as the one developed in this study will be essential for realizing the full potential of data-driven disease prediction and prevention. Future work should focus on prospective validation, extension to diverse populations, and clinical implementation to translate these methodological advances into improved patient outcomes.

References

1. National Institutes of Health. Understanding Diabetes Heterogeneity via Mining Multimodality Interconnected Data. Award Number K25DK135913. National Institute of Diabetes & Digestive & Kidney Diseases. Published 2023. Accessed June 2024.
2. Mackay M, et al. Artificial intelligence applications in type 2 diabetes: A systematic review of current evidence and future directions. *Frontiers in Endocrinology*. 2025;16:1699954. doi:10.3389/fendo.2025.1699954.
3. Amar J, Liu E, Breschi A, et al. Integrating Genomics into Multimodal EHR Foundation Models. *arXiv preprint*. 2024;2510.23639.
4. Zhang Z, Wang L, Meng W, et al. Type 2 Diabetes Subtyping via Phenotype and Genotype Co-Learning. *Hypergraph Framework for T2D Prediction*. Emory University, Department of Computer Science. Published 2024.
5. Yagin FH, Colak C, Algarni A, et al. Hybrid Explainable Artificial Intelligence Models for Targeted Metabolomics Analysis of Diabetic Retinopathy. *Diagnostics*. 2024;14(13):1364. doi:10.3390/diagnostics14131364.
6. Genomics and Multi-Omics Approaches to Diabetes Subtyping and AI-Driven Prediction for Precision Medicine. *Journal of Korean Diabetes*. Published 2025. (Under review/in press).
7. Yagin FH, Colak C, Algarni A, Gormez Y, Guldogan E, Ardigò LP. Hybrid Explainable Artificial Intelligence Models for Targeted Metabolomics Analysis of Diabetic Retinopathy. *Diagnostics*. 2024;14(13):1364. doi:10.3390/diagnostics14131364.
8. Bhatta RP. Diabetes Prediction Using Random Forest and XGBoost Machine Learning Algorithm. *Journal of Engineering Technology and Planning*. 2025;6(1):88-103. doi:10.3126/joetp.v6i1.87829.
9. All of Us Research Program Investigators. The "All of Us" Research Program. *New England Journal of Medicine*. 2019;381:668-676. doi:10.1056/NEJMsr1809937.
10. Mahajan A, et al. Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. *Nature Genetics*. 2022;54:560-572. doi:10.1038/s41588-022-01058-3.
11. Ahlqvist E, Storm P, Käräjämäki A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *The*

Lancet Diabetes & Endocrinology. 2018;6(5):361-369. doi:10.1016/S2213-8587(18)30051-2.

12. Udler MS, Kim J, von Grotthuss M, et al. Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. *PLoS Medicine*. 2018;15(9):e1002654. doi:10.1371/journal.pmed.1002654.

13. Zeevi D, Korem T, Zmora N, et al. Personalized Nutrition by Prediction of Glycemic Responses. *Cell*. 2015;163(5):1079-1094. doi:10.1016/j.cell.2015.11.001.

14. Karlsson FH, Tremaroli V, Nookaew I, et al. Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature*. 2013;498:99-103. doi:10.1038/nature12198.

15. Qin J, Li Y, Cai Z, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*. 2012;490:55-60. doi:10.1038/nature11450.