Type 2 Diabetes Subtyping via Phenotype and Genotype Co-Learning

Ziyang ZHANG[a], Lily WANG[a], Weimin MENG[b], Chang LIU[c], Hui SHAO[c], Yan V. SUN[c], Jingchuan GUO[b], Jiang BIAN[d], Rui YIN[b] and Carl YANG[a, 1]
[a] *Department of Computer Science, Emory University, USA*
[b] *College of Medicine, University of Florida, USA*
[c] *Rollins School of Public Health, Emory University, USA*
[d] *Biostatistics and Health Data Science, School of Medicine, Indiana University, USA*

Abstract. Interpreting and subtyping type 2 diabetes (T2D) is challenging yet essential for achieving fine-grained pathophysiological insights and precise clinical stratification. Previous studies have primarily relied on a small number of pre-selected risk factors and biomarkers, neglecting the integration of multimodality data (e.g., phenotypic and genetic features) for more comprehensive analyses. In this study, we select a cohort of 42,256 participants from the National Institutes of Health's All of Us Research Program, where our hypergraph framework achieves an AUROC of 89.64% on predicting T2D when integrating phenotypic and genetic features. The proposed pipeline performs subtyping by clustering clinical concepts, genetic variants, and individuals in an end-to-end manner. Further analysis using genetic risk scores reveals distinct genetic profiles between T2D subtypes and highlights the potential applications of our solution in precision medicine.

Keywords. Type 2 diabetes, disease subtyping, WGS, EHR, multimodality

1. Introduction

Affecting over 500 million people worldwide, type 2 diabetes (T2D) poses a significant global health challenge largely due to its etiological heterogeneity influencing diabetes complications differently. While environmental and lifestyle factors are well-established risk factors of T2D, genetic information also plays a crucial role in understanding its complex biological mechanisms [1]. Deciphering this heterogeneity is crucial for uncovering the pathophysiological mechanisms from both phenotypic and genetic perspectives. In this study, we aim to interpret T2D by jointly analyzing phenotypic and genetic data. Accordingly, we propose a subtyping pipeline that simultaneously clusters individuals, phenotypes, and genotypes using electronic health record (EHR) and whole genome sequencing (WGS) data. The data is effectively integrated using a hypergraph backbone model, where the unsupervised subtyping process can be guided by a T2D prediction task in an end-to-end manner.

Previous efforts in data-driven T2D subtyping have often relied on single-modal approaches involving a panel of pre-selected risk factors and biomarkers, without the access to large-scale EHRs and genomics data in a same cohort, thereby restricting more comprehensive multimodal analyses. Additionally, traditional clustering techniques like K-means, which operate on pre-defined features, lack the adaptability to learn

[1] Corresponding Author: Carl Yang, j.carlyang@emory.edu

hierarchical or latent representations and are often challenging to integrate with deep learning frameworks. In contrast, our model performs effective and efficient phenotype-genotype co-learning guided by downstream prediction tasks. We use data from the National Institutes of Health (NIH)'s All of Us (AoU) Research Program, which has built one of the largest and most diverse biomedical databases including data such as clinical records (phenotypic), biosamples and bioassays (genomics), surveys, and physical measurements [2]. In this study, we select a cohort of 42,256 patients on the AoU Research Workbench[2]. Leveraging this extensive and ancestrally diverse dataset, our framework achieves an AUROC of 89.64% when predicting the risk of T2D. Our method results in two T2D subtypes characterized by distinct genetic profiles such as body fat and metabolic syndrome.

2. Method

2.1. Data Preprocessing

Our study cohort contains (1) a case group of 15,108 patients who were diagnosed with T2D and (2) a control group of 27,148 patients who shares similar demographic characteristics and clinical profiles with the case group but never diagnosed with diabetes. For the case group, we define the diagnostic criteria for T2D based on the guidelines provided in [3], including: (1) Medical code level: ICD-10 (E11) and ICD-9 (250.x0, 250.x2); and (2) Lab measurements level: HbA1c ≥ 6.5%, fasting plasma glucose (FPG) ≥ 126 mg/dL, and two-hour oral glucose tolerance test (OGTT) plasma glucose level ≥ 200 mg/dL. In this study, we use propensity score matching (PSM) to construct the control group while minimizing potential bias and the influence of variables that could confound the relationship between T2D and outcomes. The propensity score is calculated using logistic regression with independent variables including gender, age, age^2, BMI, hypertension, hypercholesterolemia, smoking status, and kidney disease status. These covariates are chosen because they are demographic factors or potential confounders for T2D [4]. We perform nearest-neighbor matching with a caliper of 0.001 on the propensity score, allowing up to two control matches per case.

In this study, all phenotypic features are derived from participants' EHRs, particularly from the standardized (OMOP CDM) clinical codes. For a patient in the case group, we identify the initial diagnosis of T2D and collect all unique clinical codes from any previous visits. These clinical codes are used to describe a patient's profile and will later be used as input phenotypic features. Note that any codes related to diabetes are excluded to prevent data leakage. For a patient in the control group, we simply utilize all patient's available visits and collect the unique clinical codes from those visits as input.

Following a recent genome-wide association study of T2D [1], which identified 1,289 independent genome-wide significant single nucleotide polymorphisms (SNPs)[3] that map to 611 loci, we identify 926 SNPs in the AoU cohort after excluding multiallelic variants and indels. To construct genotype features, we utilize short read whole genome sequencing (srWGS) data, specifically a smaller callset known as the Allele Count/Allele Frequency (ACAF) threshold callset. For each available SNP, we categorize genotypes into homozygous reference allele, homozygous risk allele, and heterozygous. These

[2] This study is conducted under the Controlled Tier and uses Version 7 of the Curated Data Repository (CDR) released in 2022. Access to the Controlled Tier is restricted to approved researchers.

[3] All SNPs have MAF over 5% and ORs below 1.05. Details of the 1,289 T2D-associated SNPs, including their genetic loci and summary statistics, are available in [1] and Supplementary Table 4.

genotype categories are then encoded into multihot features. Missing values in the genotype data are imputed with 0.

2.2. Hypergraph Modeling

We apply our previously developed hypergraph transformer model [5, 6] to capture the complex interactions among phenotypic and genotypic features. In our hypergraph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, nodes $\mathcal{V}$ represent clinical codes or genetic variants, and hyperedges $\mathcal{E}$ correspond to patients described by subsets of these nodes. Due to the typically higher density of genotypic features, directly modeling them within a single graph may result in imbalance in feature representation, where genotypic features dominate the graph structure. To mitigate this, we adopt a dual-layer hypergraph, allowing phenotypic and genotypic features to be modeled separately while enabling their interactions through a shared predictor. For each layer, node embeddings are initialized using DeepWalk [7], and updated iteratively via a two-step message-passing mechanism:

$$\boldsymbol{E}_e^{(l)} = f_{\mathcal{V}\to\mathcal{E}}\big(\boldsymbol{\mathcal{V}}_{e,\boldsymbol{X}^{(l-1)}}\big), \quad \boldsymbol{X}_v^{(l)} = f_{\mathcal{E}\to\mathcal{V}}\big(\boldsymbol{\mathcal{E}}_{v,\boldsymbol{E}^{(l)}}\big), \tag{1}$$

where $\boldsymbol{\mathcal{V}}_{e,\boldsymbol{X}}$ represents node features in hyperedge e, and $\boldsymbol{\mathcal{E}}_{v,\boldsymbol{E}}$ represents hyperedge features connected to node v. The function $f(\cdot)$ uses a standard self-attention mechanism to prioritize informative features during aggregation, where the input consists of feature representations from connected nodes or hyperedges. For T2D risk prediction, we aggregate hyperedge embeddings from phenotype and genotype layers and pass them through a multilayer perceptron (MLP) with a sigmoid activation to compute the predicted probability $\hat{y}$. The model is optimized using the binary cross-entropy loss: $\mathcal{L}_{CLS} = -y\log(\hat{y}) - (1-y)\log(1-\hat{y})$, where y is the ground truth.

2.3. Subtype Clustering

Self-supervised clustering is performed on learned embeddings to identify consistent subgroups of clinical codes, patients, and SNPs for the discovery of T2D subtypes. We employ a deep embedded approach [8] by iteratively learning cluster assignments Q_v and $\boldsymbol{Q}_{\mathcal{E}}$ for nodes and hyperedges, respectively. The soft assignment q_{ik} for a node (or hyperedge) i and cluster k is computed as follows, while the target distribution p_{ik} is designed to refine cluster purity:

$$q_{ik} = \frac{\left(1+\|\boldsymbol{x}_i-\boldsymbol{u}_k\|^2\right)^{-1}}{\sum_j\left(1+\|\boldsymbol{x}_i-\boldsymbol{u}_j\|^2\right)^{-1}}, \quad p_{ik} = \frac{q_{ik}^2/\sum_i q_{ik}}{\sum_j q_{ij}^2/\sum_i q_{ij}}, \tag{2}$$

where $\boldsymbol{x}_i$ is the embedding of i, and $\boldsymbol{u}_k$ is the centroid representation of cluster k. Cluster centroids are initialized using K-means. Then, we optimize Kullback-Leibler (KL) divergence between $\boldsymbol{P}$ and $\boldsymbol{Q}$:

$$\mathcal{L}_{CLU} = KL(\boldsymbol{P} \,||\, \boldsymbol{Q}) = \textstyle\sum_i \sum_k p_{ik}\, log(p_{ik}/q_{ik}). \tag{3}$$

Since clustering is performed simultaneously on phenotypes, genotypes, and patients, the pipeline incorporates three clustering losses as defined in Eq.3. The final optimization objective is the weighted sum of these three clustering losses and $\mathcal{L}_{CLS}$.

3. Results

3.1. T2D Risk Prediction

We evaluate the predictive capability of our proposed hypergraph (HyG) pipeline on a

T2D risk prediction task. To address label imbalance, we adopt metrics used in [6], including accuracy, AUROC, AUPR, and F1 score. We compare our model against traditional machine learning methods such as logistic regression (LR), support vector machine (SVM), random forest (RF), and XGBoost (XGB) [9], as well as a deep learning approach MLP. The dataset is split into training, validation, and test sets in a 7:1:2 ratio, and we fix the number of training epochs (if applicable) at 500.

Table 1 presents the prediction results using different data modalities. Our proposed framework consistently outperforms other baseline models across all metrics when only using EHR data. For models only using genotypic data, all methods exhibit suboptimal predictive results as expected [10], because T2D is not primarily driven by genetic factors. When combining EHR and genotypic data, HyG achieves the best overall performance, with an AUROC of 89.64% and an AUPR of 82.77%, exceeding the average by 2.56% and 3.19%, respectively.

Table 1. Performance of predicting T2D risk using different models on various data modalities. The results are the averages of metrics from 5 runs of the models. **Bold** numbers indicate the best results in each category.

Model	**EHR Only**				**Gene Only**				**EHR + Gene**			
	ACC	AUROC	AUPR	F1	ACC	AUROC	AUPR	F1	ACC	AUROC	AUPR	F1
LR	81.15	86.77	78.86	79.24	62.84	58.47	41.33	**53.38**	81.14	86.77	78.86	79.23
SVM	79.29	84.22	74.12	77.51	62.62	58.15	41.04	52.81	79.23	84.22	74.12	77.50
RF	81.77	87.33	80.65	79.49	64.58	55.89	39.44	44.60	81.85	87.30	80.76	79.53
XGB	80.76	87.10	80.63	78.25	63.67	57.26	40.43	49.92	80.76	87.71	80.63	78.25
MLP	81.53	87.65	80.75	79.87	**64.71**	53.45	37.26	42.31	80.88	87.46	80.31	79.29
HyG	**82.19**	**88.88**	**82.30**	**80.81**	64.41	**59.22**	**42.18**	49.32	**83.36**	**89.64**	**82.77**	**81.85**

3.2. Subtyping Analysis

We generate five T2D subtypes from our hypergraph pipeline. We compare the genetic risk scores (GRS) of these subtypes across eight functional clusters defined by [1]. The GRS is calculated by GRS $= \sum_{i=1}^{n} \beta_i \cdot g_i$, where β_i is the effect size of SNP i, and g_i is the number of risk alleles for the SNP, taking values of 0, 1, or 2. Subtype 2 stands out with the strongest association to body fat and obesity. Subtype 1 also shows elevated risks in these areas but to a lesser extent. Subtype 3 demonstrates weaker associations overall, particularly with metabolic and glycaemic traits. Subtype 4 is characterized by minimal variation, while subtype 5 shows mild positive associations, with slight elevations in lipid metabolism. Table 3 shows the representative clinical concepts for each subtype extracted from the node clustering process (illustrated in Fig.1).

Table 2. Comparison of standardized GRS across functional clusters in T2D subtypes.

Cluster	Subtype 1	Subtype 2	Subtype 3	Subtype 4	Subtype 5
Body fat	0.3671	0.7449	-0.0722	-0.0191	0.0264
Obesity	0.2827	0.3847	-0.0834	-0.0169	0.0830
Metabolic syndrome	-0.0465	-0.1417	-0.1284	0.0020	0.0439
Residual glycaemic	0.1730	0.2410	-0.1955	-0.0024	-0.0263
Beta cell -PI	0.0071	-0.0271	0.0683	-0.0030	0.0258
Beta cell +PI	0.2647	0.2853	0.1305	-0.0147	0.0166
Lipodystrophy	0.0560	0.1835	-0.1223	-0.0013	0.0005
Liver/lipid metabolism	0.0413	-0.0018	0.0247	-0.0060	0.0601

4. Discussion

The findings in subtyping emphasize the potential impact of integrating genetic factors into frameworks solely relying on phenotypic data. While the observed differences in GRS between subtypes are numerically small, the extracted clinical nodes from each subtype corroborate the biological functions observed from a genetic perspective. This multi-modal integration approach may provide deeper insights into the heterogeneity of disease profiles, particularly when more informative data components of T2D, such as proteomics and metabolomics are available. However, methodological limitations remain, as the iterative clustering approach based on representation learning may produce unstable results. Future work should provide cross-validation of the subtypes across diverse cohorts to establish stronger biological credibility.

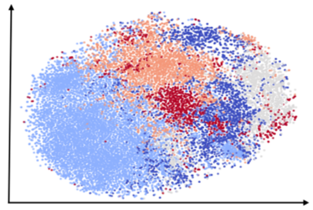

Figure 1. Node clustering illustration.

Table 3. Clinical concepts for each subtype.

Subtypes	Key clinical concepts
Subtype 1	Chronic kidney disease, Diabetic retinopathy
Subtype 2	Renal complications, Obesity, Gout-related symptoms
Subtype 3	Diabetic neuropathy, Pancreatic dysfunction
Subtype 4	Metabolic/inflammatory disorders
Subtype 5	Carbohydrate metabolism disorders, Kidney disease

5. Conclusions

We present a versatile framework that efficiently co-learns phenotypic and genotypic features. The framework achieves state-of-the-art predictive performance on the AoU dataset and guides the downstream subtyping process. It identifies novel T2D subtypes with biological differences across eight functionally distinct categories of SNPs and could therefore provide potential insights into the heterogeneity of T2D.

Acknowledgement

We gratefully acknowledge All of Us participants for their contributions, without whom this research would not have been possible. We also thank the National Institutes of Health's All of Us Research Program for making available the participant data examined in this study. This research was partially supported by the US National Science Foundation under Award Number 2319449 and Award Number 2312502, as well as the US National Institute of Diabetes and Digestive and Kidney Diseases of the US National Institutes of Health under Award Number K25DK135913.

References

[1] Ken Suzuki and et al. Genetic drivers of heterogeneity in type 2 diabetes pathophysiology. *Nature*, 627(8003):347–357, 2024.

[2] ALL of Us Research Program Investigators. The "all of us" research program. *New England Journal of Medicine*, 381(7):668–676, 2019.

[3] Rajeev Goyal and et al. Type 2 Diabetes. StatPearls. StatPearls Publishing, Treasure Island (FL), 2024.

[4] Brita Roy and et al. A propensity-matched study of the association of diabetes mellitus with incident heart failure and mortality among community-dwelling older adults. *The American journal of cardiology*, 108(12):1747–1753, 2011.

[5] Ran Xu and et al. Hypergraph transformers for ehr-based clinical predictions. *AMIA Summit on Translational Science Proceedings*, 2023:582, 2023.

[6] Ziyang Zhang and et al. Tacco: Task-guided co-clustering of clinical concepts and patient visits for disease subtyping based on ehr data. *In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6324–6334, 2024.

[7] Bryan Perozzi and et al. Deepwalk: Online learning of social representations. *In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710, 2014.

[8] Junyuan Xie and et al. Unsupervised deep embedding for clustering analysis. In International conference on machine learning, pages 478–487. PMLR, 2016.

[9] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[10] Jason L Vassy and James B Meigs. Is genetic testing useful to predict type 2 diabetes? *Best practice & research Clinical endocrinology & metabolism*, 26(2):189–201, 2012.