SHARK: MODELING SEMANTIC HIERARCHY OF MEDICAL CODE VIA RESIDUAL K-MEANS QUANTIZATION

Kaisong Zhang[1], *Hang Lv*[2], *Yanchao Tan*[2], *Zhigang Lin*[3], *Hengyu Zhang*[4], *Xiping Chen*[5], *Carl Yang*[6]

[1]Fuzhou University, Fuzhou, China
[2]Engineering Research Center of Big Data Intelligence, Ministry of Education, Fuzhou, China
[3]The First Affiliated Hospital of Fujian Medical University, Fuzhou, China
[4]Macquarie University, Sydney, Australia
[5]Hangzhou Bywin Technology Co., Ltd., Hangzhou, China
[6]Emory University, Atlanta, USA

ABSTRACT

Pretrained Language Models (PLMs) have advanced diagnosis prediction by leveraging the semantic understanding of medical concepts in Electronic Health Records (EHRs). However, existing PLM-based methods usually model disease relations by inheriting the ICD hierarchy, typically enforcing inclusive constraints along a strict single path. This rigid taxonomy produces semantic loss: clinically related conditions that diverge early in ICD (e.g., hypertension and hyperlipidemia) are assigned to completely different branches, despite sharing underlying pathophysiological mechanisms. To solve this limitation, we propose SHARK, which recursively extracts the flexible semantic hierarchy via Residual K-Means (RK-Means). Residual, acting as vector offsets, can capture the semantic factors of clinical concepts and each element in the residual part is a vector of K-Means clusters. Our semantic hierarchy allows diseases to have joint semantic factors across branches. We further obtain patient embeddings via a dual-RNN architecture for diagnosis prediction. Extensive experiments on two real-world EHR datasets show the competitive prediction performance of SHARK compared with various state-of-the-art models.

Index Terms— Diagnosis prediction, Residual K-Means

1. INTRODUCTION

Using pre-trained language models (PLMs) for diagnosis prediction has emerged as a crucial direction in Electronic Health Records (EHRs) [1, 2] research. A common approach is capturing the hierarchical relations inherent in the International Classification of Diseases (ICD) codes [3].

However, the ICD hierarchy is a rigid taxonomy, enforcing inclusive restrictions along a unique path. Once two codes diverge early, their fine-grained descendants cannot converge again. As a result, the ICD hierarchy produces semantic loss, where some clinically related conditions are far apart in the taxonomy [4]. For instance, as depicted in left side of Fig. 1, essential hypertension (I10) and hyperlipidemia (E78.5) frequently co-occur as cardiometabolic risk factors and share underlying pathophysiological mechanisms (e.g., endothelial dysfunction). But ICD classifies essential hypertension as a cardiovascular disease and hyperlipidemia as a metabolic disorder, assigning them to completely separate branches. This taxonomy therefore often fails to reflect more nuanced and similar relationships. This toy example reveals a broader issue: **ICD hierarchy, acting as a administrative taxonomy, fails to capture meaningful semantic hierarchy**. Therefore, this raises a natural question: *How can we model a more reasonable similarities across categories and coarse-to-fine semantic hierarchy that do not align with the rigid taxonomy*?

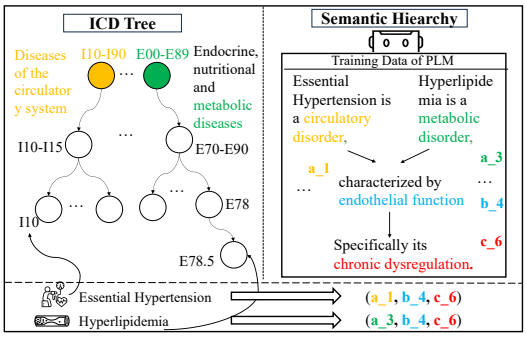

Fig. 1. An illustrated example where the ICD taxonomy tree separates related diseases (hypertension and hyperlipidemia), while a semantic hierarchy captures their convergence.

In this paper, we propose SHARK to capture the **S**emantic **H**ierarchy of medical codes vi**A** **R**esidual **K**-Means quantization. Our key insight comes from analogical reasoning in embedding spaces: differences between embedding vectors often correspond to interpretable semantic differences [5, 6]. A classic example is that the vector arithmetic $v(\text{king}) - v(\text{man}) \approx v(\text{queen}) - v(\text{woman})$ capturing the

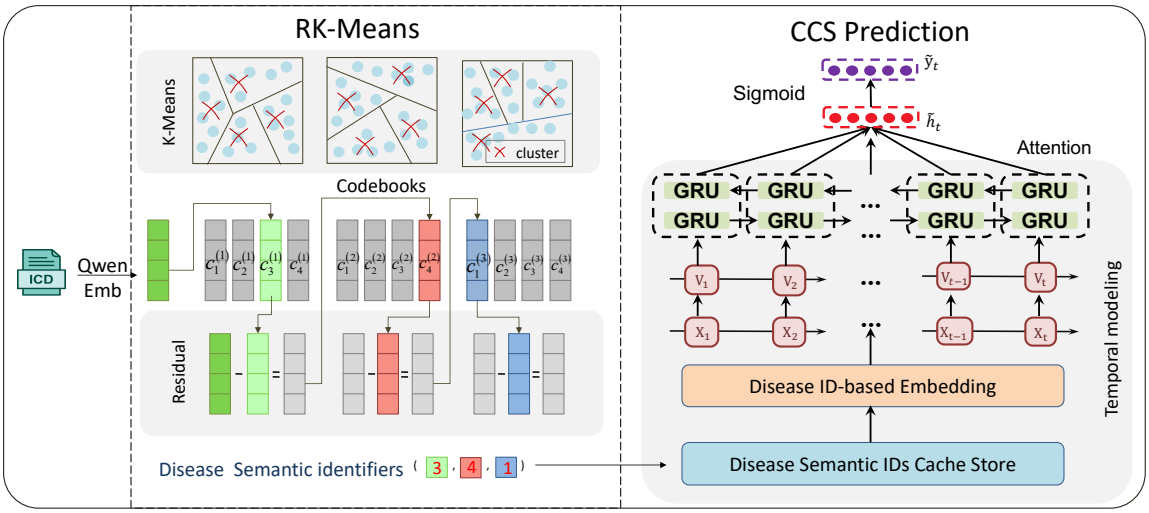

Fig. 2. The overall framework of SHARK, which consists of the RK-Means, temporal modeling, and predictor modules

king-to-man and queen-to-woman transformations (royalty). Our central assumption is that the disease embeddings from PLMs can capture similar transformation. For example, the difference between embeddings of two diagnoses might encode a specific clinical factor (eg., $v(\text{Diabetic nephropathy}) - v(\text{Diabetes}) \approx v(\text{Hypertensive nephropathy}) - v(\text{Hypertension})$, both sides express the nephropathy factor.).

Specifically, our work SHARK operates by recursively extracting semantic features from disease embeddings using Residual K-Means (RK-Means) [7]. In essence, we perform K-Means clustering on the set of all disease vectors from PLMs to obtain a set of coarse semantic prototype vectors for the first layer. Next, we cluster the residuals (the differences between each disease embedding and its first-layer approximation) to produce finer-grained codewords for a second layer, and so on. This recursive quantization continues for multiple layers, each capturing increasingly fine semantic details. Crucially, this representation is not a single branch in a tree, but rather a combination of codewords from each layer's codebook. As shown on the right side of Fig. 1, the two diseases can share some codewords even though differing in the coarse level. SHARK thereby allows arbitrary combinations of semantic features across granularities, in contrast to ICD's rigid inheritance. We then integrate RK-Means framework into a dual-attention temporal model and aggregate the disease representations to patient embeddings for diagnosis prediction. Extensive experimental results demonstrate that SHARK outperforms the state-of-the-art competitors.

2. METHOD

2.1. Problem Definition and SHARK Overview

Given a patient p_n's historical EHRs, represented as a sequence of visits $\mathcal{V}_n = (v_{n,1}, v_{n,2}, \ldots, v_{n,T_n})$, our goal is to predict the set of Clinical Classifications Software (CCS) codes for their next visit. Each visit $v_{n,t}$ consists of a set of diagnosis codes $\mathcal{D}_{n,t}$ from a universe $\mathcal{D}$. The prediction target is a multi-hot vector $\mathbf{y}_n \in \{0,1\}^C$, where C is the total number of possible CCS codes.

Our proposed SHARK framework operates in three main stages. As shown in Fig. 2, First, we introduce a novel semantic quantization module based on RK-Means to obtain semantic IDs, which leads to hierarchical representation of diseases and captures similarities between categories (Sect. 2.2). Second, using these powerful semantic IDs, we construct a vector for each patient visit and model the temporal dynamics of the visit sequence to generate a comprehensive patient representation (Sec. 2.3). Finally, this patient representation is used to predict the CCS for the subsequent visit (Sec. 2.4).

2.2. Semantic Hierarchy Quantization via RK-Means

We begin with a continuous embedding matrix $\mathbf{M} \in \mathbb{R}^{C \times d}$ for all diagnosis codes, where each row is a dense vector representation of a disease obtained from a PLM(Qwen3). Let $\mathbf{m}_c$ be the initial embedding for a diagnosis code c.

We employ Residual Quantization with K-Means [8] to iteratively extract semantic factors. This is performed in a coarse-to-fine manner over L layers.

For the first layer ($\ell = 1$), we apply K-Means clustering to the initial embeddings $\mathbf{M}^{(0)} = \mathbf{M}$. This yields a codebook of K centroids, $\mathbf{R}^{(1)} \in \mathbb{R}^{K \times d}$, which represent the coarsest semantic factors. For each disease c, we find the nearest centroid and assign its index $r_c^{(1)}$ as the first semantic ID:

$$r_c^{(1)} = \arg \min_{j \in \{1,\ldots,K\}} \left\| \mathbf{m}_c^{(0)} - \mathbf{R}_j^{(1)} \right\|_2 . \quad (1)$$

The crucial step is to compute the residual vector $\mathbf{m}_c^{(1)}$. This residual captures the specific semantic information of disease

c that is not explained by the coarse-grained factor $\mathbf{R}^{(1)}_{r_c^{(1)}}$:

$$\mathbf{m}_c^{(1)} = \mathbf{m}_c^{(0)} - \mathbf{R}^{(1)}_{r_c^{(1)}}. \quad (2)$$

For each subsequent layer $\ell = 2, \ldots, L$, we repeat this process on the residual embeddings from the previous layer, $\mathbf{M}^{(\ell-1)}$. This allows us to progressively extract finer-grained semantic factors from the remaining information. After L iterations, each diagnosis code c is represented by an L-tuple of semantic IDs: $\big(r_c^{(1)}, \ldots, r_c^{(L)}\big)$. This tuple represents a disease as a composition of L semantic factors. A single visit representation is then obtained by averaging the embeddings of all diagnosis codes within that visit.

2.3. Temporal Patient Modeling

To emphasize recent visits while preserving interpretability, we process the reversed sequence $\tilde{\mathbf{v}}_{n,i} = \mathbf{v}_{n,T_n-i+1}$, $i = 1, \ldots, T_n$, with two GRUs. The first GRU produces hidden states $\mathbf{g}_{n,i}$ which are used to generate visit-level scalars $\alpha_{n,i}$ as the first level attention weights, and the second GRU produces $\mathbf{h}_{n,i}$ used to generate feature-wise vectors $\beta_{n,i}$ as the second level attention weights.

$$\mathbf{g}_{n,i} = \mathrm{GRU}_\alpha(\tilde{\mathbf{v}}_{n,1:i}), \alpha_{n,i} = \mathrm{softmax}\big(\mathbf{w}_\alpha^\top \mathbf{g}_{n,i} + b_\alpha\big), \quad (3)$$

$$\mathbf{h}_{n,i} = \mathrm{GRU}_\beta(\tilde{\mathbf{v}}_{n,1:i}), \boldsymbol{\beta}_{n,i} = \tanh\big(\mathbf{W}_\beta \mathbf{h}_{n,i} + \mathbf{b}_\beta\big). \quad (4)$$

2.4. CCS Prediction and Training Objective

Using the generated attentions above, we aggregate visit representations into a patient context vector. Given the context vector $\mathbf{c}_n$, we predict the multi-label CCS vector as follows:

$$\hat{\mathbf{y}}_n = \mathrm{Sigmoid}(\mathbf{W}_c \mathbf{c}_n + \mathbf{b}_c) \in (0,1)^C. \quad (5)$$

We optimize the binary cross-entropy loss over the batch:

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{N}\sum_{n=1}^{N}\Big[\mathbf{y}_n^\top \log \hat{\mathbf{y}}_n + (\mathbf{1}-\mathbf{y}_n)^\top \log(\mathbf{1}-\hat{\mathbf{y}}_n)\Big]. \quad (6)$$

3. EXPERIMENTS

3.1. Experimental Settings

Datasets and Evaluation Metrics. We use two real-world EHR datasets the MIMIC-III and MIMIC-IV to evaluate SHARK. A summary of the dataset statistics is presented in Table 1. We select patients with multiple visits (of visits $\geq$ 2). We use the CCS codes in patients' next visit as labels and other visits as features. For evaluation metrics, we use visit-level Precision@k (P@k) and code-level Accuracy@k (Acc@k). These metrics are consistent with previous work [9–11].

Implementation Details. Both datasets are split into training/validation/test sets with the ratio of 7:1:2, and patients are the unit of segmentation. We use the existing Qwen3-Embedding-8B model and obtain each representation from the last states. We set the embedding dimension d of diseases and patients as 16. The layer number L of RK-Means is 4, and each layer has 128 clusters. This choice is a trade-off for the balance between capacity and stability. The model is optimized by using the AdamW optimizer with the learning rate 10^{-3}.

Table 1. Statistics of the used datasets.

Dataset	MIMIC-III	MIMIC-IV
# of patients	5,449	79,393
# of visits	14,141	408,990
Avg. # visits per patient	2.60	5.15
Max # visits per patient	29	170
# of unique diagnoses	3,874	37,917
# of CCS codes	285	842

Table 2. The impact of semantic IDs (%) on the MIMIC-III and MIMIC-IV datasets.

Metric	P@10	P@20	Acc@10	Acc@20
MIMIC-III				
SHARK SID	50.56	56.83	35.34	53.94
Random ID	48.39	53.32	34.53	51.43
MIMIC-IV				
SHARK SID	53.88	57.36	35.77	51.74
Random ID	47.03	54.12	33.85	49.91

3.2. Overall Performance Analysis

In this section, we compare our model with six state-of-the-art methods, including temporal-aware method Retain [14], Trans [10], and ICD hierarchy-aware methods KAME [9], CGL [15], BoxCare [16], Shy [17]. As shown in Table 3 SHARK consistently outperforms all other methods across all evaluation metrics on both datasets.

Compare with temporal-aware models. Compared with RETAIN and TRANS, SHARK combines temporal modeling with rich representations of medical codes via RK-Means mechanism. SHARK is particularly outstanding on the MIMIC-IV dataset with more medical entities (shown in Table 1), surpassing TRANS by 35.58% in P@10.

Compare with ICD hierarchy-aware models. Among these models, BoxCare is a strong hierarchy-aware method that leverages box embeddings to model structural relationships among diagnoses. In contrast, SHARK captures the semantic hierarchy, leveraging the capability of PLM embeddings. As a result, SHARK consistently outperforms BoxCare.

Table 3. Performance comparison (%) on the MIMIC-III [12] and MIMIC-IV [13] datasets.

Method	MIMIC-III				MIMIC-IV			
	P@10	P@20	Acc@10	Acc@20	P@10	P@20	Acc@10	Acc@20
CGL	47.84	54.78	34.12	52.88	49.37	55.06	<u>35.59</u>	<u>51.20</u>
RETAIN	46.83	53.85	34.26	51.53	47.91	54.14	34.90	50.28
KAME	47.11	53.04	33.18	50.41	51.94	55.33	34.91	49.22
BoxCare	<u>48.37</u>	<u>55.18</u>	<u>35.16</u>	<u>52.57</u>	<u>52.12</u>	<u>56.11</u>	35.25	51.02
TRANS	45.39	52.86	33.11	50.66	39.74	46.23	29.24	43.63
SHy	45.61	51.97	32.47	49.80	49.02	54.71	35.33	50.50
SHARK	**50.56**	**56.83**	**35.74**	**53.94**	**53.88**	**57.36**	**35.77**	**51.74**

3.3. Impact of Semantic IDs

We also compare the importance of Semantic IDs for our SHARK. To generate the Random ID baseline, we assign L random codewords to each disease. A Random ID of length L for an item is simply $(c_1, \ldots, c_L)$, where c_i is sampled uniformly at random from $\{1, 2, \ldots, K\}$. We set $L = 4$, and $K = 128$ for the Random ID baseline to make the cardinality similar to SHARK's Semantic IDs. The results are shown in Table 2. We see that Semantic IDs consistently outperform Random ID baseline, highlighting the importance of capture the semantic hierarchy.

Table 4. Interpretable semantic clusters.

Cluster ID	Semantic Summary	Purity
$[26, \_, \_, \_]$	Surgical procedures ("-tomy")	25/25
$[80, \_, \_, \_]$	Drug-related	43/43
$[\_, 2, \_, \_]$	Psychological disorders	17/18
$[\_, 126, \_, \_]$	Unspecified / NOS	105/106
$[\_, \_, 32, \_]$	Immune system diseases	33/41
$[\_, \_, 62, \_]$	Hepatitis-related	26/28
$[\_, \_, \_, 12]$	Tumors or masses	10/12
$[\_, \_, \_, 16]$	Abdominal conditions	13/13

3.4. Case Studies

Cluster Has Obvious Semantics. In this section, we found that each RK-Means cluster indeed corresponds to clear semantics. For example, the cluster $[\_, 126, \_, \_]$ groups 106 diagnosis codes, out of which 105 associated with "unspecified", "not elsewhere classified", or ambiguous categories—achieving a semantic purity of 99%. Here, purity is defined as the proportion of codes in a cluster that belong to its majority semantic category. More detail examples are shown in Table 4. These findings also suggest that residual vectors in RK-Means can extract coarse-grained clinical semantics, such as anatomical region, etiology, and modality (e.g., surgical, toxicological, pharmacological).

Capture Similarities Across Categories. In addition to the motivating example of essential hypertension (I10) and hyperlipidemia (E78.5) mentioned in the Introduction, we further observe similar cases where diseases are far apart in ICD ontology but have similar semantic hierarchy. For example, *Injury to duodenum, without open wound into cavity* (863.21) and *Duodenitis, with hemorrhage* (535.61) originate from different ICD categories, but they overlap on the identifiers $[22, 121, 64]$, revealing their semantic proximity. Although one describes a traumatic injury and the other an inflammatory disease, both involve pathological damage to the duodenum and may lead to hemorrhagic complications. More examples are shown in Table 5 These findings confirm that our proposed RK-Means based semantic quantization can successfully capture similarities across branches that are ignored by rigid ICD taxonomies.

Table 5. Case study examples of similarities across categories captured by RK-Means.

ICD Code	Disease Semantic	Shared IDs
079.6	RSV infection	[_, 46, 107, 0]
466.11	RSV bronchiolitis	
863.21	Duodenal injury	[_, 22, 121, 64]
535.61	Duodenitis w/ hemorrhage	
162.9	Lung neoplasm	[_, 46, 121, 64]
493.22	COPD exacerbation	
997.3	Respiratory complications	[_, 40, 101, 104]
E910.2	Drowning/submersion	
799.0	Asphyxia/hypoxemia	[_, 40, 79, 64]
994.7	Asphyxiation/strangulation	

4. CONCLUSION

This paper proposes a novel disease tokenizer that models the disease representation by capturing the semantic hierarchy. Extensive experiments on MIMIC-III and MIMIC-IV demonstrated that SHARK outperforms state-of-the-art baselines in EHR-based diagnosis prediction tasks. A set of case studies further confirm its effectiveness and interpretability.

5. ACKNOWLEDGEMENTS

This work was supported in part by the Fujian Provincial Artificial Intelligence Industry Development Technology Project under Grants (2025H0042), Fujian Provincial Natural Science Foundation of China under Grants (2025J01540), and National Natural Science Foundation of China under Grants (62302098). Carl Yang was not supported by any fund from China.

6. REFERENCES

[1] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi, "Behrt: Transformer for electronic health records," *Scientific Reports*, vol. 10, 2019.

[2] Yanchao Tan, Hang Lv, Yunfei Zhan, Guofang Ma, Bo Xiong, and Carl Yang, "Boxlm: Unifying structures and semantics of medical concepts for diagnosis prediction in healthcare," in *Forty-second International Conference on Machine Learning*.

[3] Emil Riis Hansen, Tomer Sagi, and Katja Hose, "Diagnosis prediction over patient data using hierarchical medical taxonomies," in *EDBT/ICDT Workshops*, 2023.

[4] Muhan Zhang, Christopher R. King, Michael S. Avidan, and Yixin Chen, "Hierarchical attention propagation for healthcare representation learning," *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020.

[5] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," in *International Conference on Learning Representations*, 2013.

[6] Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu, "Learning semantic hierarchies via word embeddings," in *Annual Meeting of the Association for Computational Linguistics*, 2014.

[7] Clark Mingxuan Ju, Liam Collins, Leonardo Neves, Bhuvesh Kumar, Louis Yufeng Wang, Tong Zhao, and Neil Shah, "Generative recommendation with semantic ids: A practitioner's handbook," *ArXiv*, vol. abs/2507.22224, 2025.

[8] Xinchen Luo, Jiangxia Cao, Tianyu Sun, Jinkai Yu, Rui Huang, Wei Yuan, Hezheng Lin, Yichen Zheng, Shiyao Wang, Qigen Hu, et al., "Qarm: Quantitative alignment multi-modal recommendation at kuaishou," *arXiv preprint arXiv:2411.11739*, 2024.

[9] Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao, "Kame: Knowledge-based attention model for diagnosis prediction in healthcare," in *Proceedings of the 27th ACM international conference on information and knowledge management*, 2018, pp. 743–752.

[10] Jiayuan Chen, Changchang Yin, Yuanlong Wang, and Ping Zhang, "Predictive modeling with temporal graphical representation on electronic health records," in *IJCAI: proceedings of the conference*, 2024, vol. 2024, p. 5763.

[11] Xianli Zhang, Buyue Qian, Shilei Cao, Yang Li, Hang Chen, Yefeng Zheng, and Ian Davidson, "Inprem: An interpretable and trustworthy predictive model for healthcare," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 450–460.

[12] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Mahdi Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific Data*, vol. 3, 2016.

[13] Alistair E. W. Johnson, David J. Stone, Leo Anthony Celi, and Tom J. Pollard, "The mimic code repository: enabling reproducibility in critical care research," *Journal of the American Medical Informatics Association : JAMIA*, vol. 25, pp. 32 – 39, 2017.

[14] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart, "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism," *Advances in neural information processing systems*, vol. 29, 2016.

[15] Chang Lu, Chandan K Reddy, Prithwish Chakraborty, Samantha Kleinberg, and Yue Ning, "Collaborative graph learning with auxiliary text for temporal event prediction in healthcare," *arXiv preprint arXiv:2105.07542*, 2021.

[16] Hang Lv, Zehai Chen, Yacong Yang, Guofang Ma, Tan Yanchao, and Carl Yang, "Boxcare: a box embedding model for disease representation and diagnosis prediction in healthcare data," in *Companion Proceedings of the ACM Web Conference 2024*, 2024, pp. 1130–1133.

[17] Leisheng Yu, Yanxiao Cai, Minxing Zhang, and Xia Hu, "Self-explaining hypergraph neural networks for diagnosis prediction," *arXiv preprint arXiv:2502.10689*, 2025.