Dual-Channel Semantic-Enhanced Combinatorial Medication Recommendation via Knowledge Distillation

Jiawei Wen[1], **Jiabao Guo**[2], **Zhaohai Bai**[3], **Kaishun Wu**[4], **Ma Lin**[5*], **Haodi Zhang**[1,4*]
[1]College of Computer and Software Engineering, Shenzhen University, Shenzhen, China
[2]College of Journalism and Communication, Shenzhen University, Shenzhen, China
[3]Institute of Genetics and Developmental Biology, Chinese Academy of Sciences
[4]Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
[5]School of Environment Nanjing University
wenjiawei2023@email.szu.edu.cn, 2023080024@email.szu.edu.cn, baizh1986@126.com,
malin1979@nju.edu.cn wuks@hkust-gz.edu.cn hdzhang@szu.edu.cn,

Abstract

As a vital task in healthcare, combinatorial medication recommendation aims to generate drug combinations tailored to patient health status. Precisely capturing the rich semantic information within clinical narratives is crucial for achieving this goal. However, existing approaches primarily rely on isolated identifiers (i.e. patient IDs, drug codes), failing to leverage the inherent semantic associations between patient conditions and medication descriptions. To fill this gap, we propose the Dual-Channel Semantics-Enhanced Network (DCSENet), a novel dual-channel framework that explicitly incorporates context-rich clinical narratives knowledge. DCSENet fine-tunes domain-adapted pre-trained language models (LMs) to capture semantic correlations between patient status and medication narratives. A transformer-based dual-channel module fuses the semantic information at the disease-level and the patient-level respectively. The disease-level channel focuses on the natural text semantic associations between diseases and drugs, while the patient-channel provides personalized features. To mitigate the prohibitive computational overhead of the LMs in clinical deployment, we introduce an attention-map-based knowledge distillation mechanism that efficiently transfers semantic knowledge from the LMs into an identifier-based (ID-based) target model. Extensive experiments on MIMIC-III and MIMIC-IV datasets demonstrate that DCSENet outperforms existing state-of-the-art methods in recommendation accuracy while maintaining a low computational cost.

1 Introduction

Combination medication recommendation systems aim to generate appropriate therapeutic prescriptions by comprehensively assessing a patient's health status. Such prescriptions currently rely heavily on physicians' manual experience. However, the expansion of medical knowledge and complexity in multimorbidity cases [Skou *et al.*, 2022] pose great challenges to clinical decision-making. Fortunately, the widespread adoption of electronic health records (EHRs) [Cowie *et al.*, 2017] has enabled large-scale digitization of clinical narratives, providing a crucial foundation for automated medication-recommendation systems. This vast clinical data repository has fueled robust research interest in AI-driven medication recommendation.

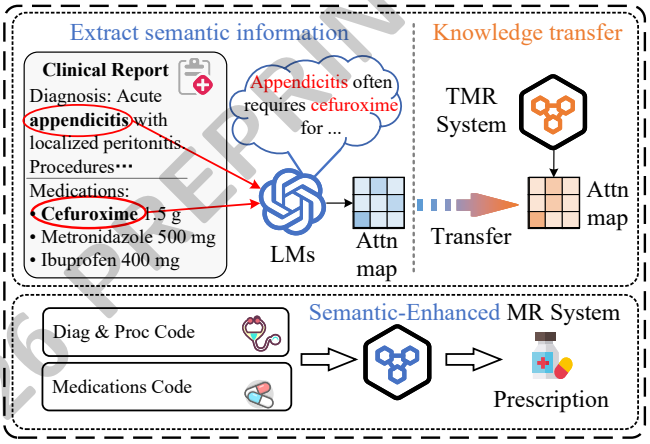

Figure 1: Illustration of transferring text semantic knowledge to the traditional medication recommendations system.

The predominant approach in current medication recommendation systems [Choi *et al.*, 2016; Wu *et al.*, 2022; Tan *et al.*, 2022] is ID-based. These methods primarily focus on architectural designs that attempt to better mimic the clinical decision-making process. However, they face three key limitations: (1) neglecting crucial external knowledge that guides real-world prescribing practices; (2) their reliance on identifiers limiting model generalizability; and (3) the inherent lack of interpretability, which hampers their utility in clinical decision support. Recent advances have explored various strategies to incorporate external knowledge, with three main directions: leveraging clinical en-

*Co-corresponding authors.

tity structures [Wang *et al.*, 2024; Li *et al.*, 2025], integrating drug chemical properties [Yang *et al.*, 2021; Yang *et al.*, 2023], and utilizing knowledge graphs [Zhang *et al.*, 2023b]. Although these approaches have achieved notable performance gains, they largely overlook the rich semantic information embedded in clinical narratives. More recently, some studies [Tan *et al.*, 2024; Liu *et al.*, 2024a; Zhang *et al.*, 2025b] have attempted to address this gap by employing natural language models to extract semantic information from clinical texts. Nonetheless, these efforts remain limited to surface-level semantic encoding, failing to comprehensively capture the complex relationships between medications and patient health status within the textual context.

In this paper, we propose a Dual-Channel Semantic-Enhanced Network (DCSENet[1]) that explores semantic associations between patient health status and medications. As illustrated in Figure 1, our framework employs LMs to extract meaningful semantic relationships from medical narratives. These relationships are modeled into attention maps that explicitly capture and highlight the semantic associations, serving as knowledge bridges. This facilitates the transfer of rich semantic representations to conventional recommendation systems. Consequently, the enhanced system leverages these semantic insights to generate prescriptions based on identifiers. Specifically, DCSENet fine-tunes domain-adapted LMs to effectively capture the semantic relationships between patient health status and medication texts. A dual-channel module further models hierarchical semantic information through disease- and visit-level pathways, fully revealing the semantic associations between patients' health condition texts and medication texts. Subsequently, attention map-based knowledge distillation transfers these semantic features to traditional ID-based recommendation models. This distillation approach transforms the knowledge of clinical texts to ID-based representations, while also reducing the computational overhead associated with LMs. We summarize our main contributions as follows:

- We propose to leverage semantic correlations in texts to enhance drug recommendation systems. Specifically, we fine-tune LMs to effectively extract semantic features from clinical texts and design a dual-channel module for comprehensive modeling of the semantic relationships between patient health conditions and medications.
- We introduce an attention-map-based knowledge distillation method to alleviate the semantic mismatch between clinical narratives and discrete identifiers. This approach effectively transforms the knowledge of textual narratives to ID-based representations.
- Extensive experiments on the MIMIC-III and MIMIC-IV datasets demonstrate that DCSENet outperforms state-of-the-art baselines in recommendation accuracy, while maintaining competitive computational efficiency.

2 Related Work

This section reviews work on medication recommendation and knowledge distillation for language models.

[1]https://github.com/ResearchGroupHdZhang/DCSENet

2.1 ID-Based Medication Recommendation

Existing medication recommendation approaches are predominantly ID-based, focusing on optimizing the recommendation process to improve performance. Many methods formulate this task as a multi-label classification problem. For instance, RETAIN [Choi *et al.*, 2016] employs bidirectional querying; GAMENet [Shang *et al.*, 2019] leverages memory augmentation; DPR-AG [Zheng *et al.*, 2021] uses interaction-aware graph induction; 4SDrug [Tan *et al.*, 2022] and SSPNet [Zhang *et al.*, 2025a] adopt a set-to-set paradigm; and DrugRec [Sun *et al.*, 2022] provides causal explanations. Other approaches treat medication recommendation as a sequential decision-making task. For example, LEAP [Zhang *et al.*, 2017] employs a recurrent decoder for sequential drug recommendations, while COGNet [Wu *et al.*, 2022] and VITA [Kim *et al.*, 2024] utilize Transformers [Vaswani *et al.*, 2017] to model sequential medication decisions. Despite their impressive performance, these methods largely neglect external knowledge integration, resulting in suboptimal recommendations with limited generalizability and poor explainability.

2.2 External-Knowledge-Enhanced Medication Recommendation

External knowledge is fundamentally important for achieving optimal medication recommendation. Existing approaches incorporate various forms of external information: SafeDrug [Yang *et al.*, 2021], KE-HMFNet [Zhang *et al.*, 2023a], and MoleRec [Yang *et al.*, 2023] use chemical-structure information; KG-MIML [Shang *et al.*, 2018], TOKEPER [Wang *et al.*, 2024], and MR-DTR leverage clinical entity structures; MedRec [Zhang *et al.*, 2023b] employs knowledge graphs to enhance patient and medication representations. Recently, approaches such as NLA-MMR [Tan *et al.*, 2024] leverage LMs to encode textual information. LEADER [Liu *et al.*, 2024a] and KEDRec-LM [Zhang *et al.*, 2025b] employ LMs to process textual data and utilize knowledge distillation from hidden states to transfer semantic information into smaller models. However, these approaches primarily exploit surface-level semantic information in clinical texts, limiting their ability to comprehensively leverage deep semantic relationships. To better leverage deep semantic relationships, we propose a dual-channel framework that explicitly mines and utilizes rich associations in clinical narratives.

2.3 Knowledge Distillation for LMs

Language models (LMs) exhibit powerful semantic encoding abilities that benefit various downstream tasks [Sun *et al.*, 2020; Xu *et al.*, 2024].Nevertheless, their heavy parameter scale and high computational cost hinder real-world deployment. To tackle this issue, numerous knowledge distillation techniques dedicated to LMs have been explored [Hinton *et al.*, 2015]. For example, DistilledBERT [Tang *et al.*, 2019] transfers BERT knowledge to task-specific models, while TinyBERT [Jiao *et al.*, 2019] uses attention-map distillation to derive a condensed BERT. Recent methods like TransKD [Liu *et al.*, 2024b] further improve distillation via a cross-interaction mechanism to fuse attention maps for cross-stage feature transfer. Motivated by these advances, we propose an

attention-based distillation approach to bridge the semantic gap between textual information and identifier embeddings.

3 Problem Formulation

This section formally defines electronic health records, clinical texts, and the medication-recommendation task.

3.1 Electronic Health Records (EHR)

For a given patient, the health information in the EHR is represented as a sequence of all historical visits, denoted as:

$$V = [v^{(1)}, v^{(2)}, ..., v^{(T)}], \tag{1}$$

where each visit $v^{(i)}$ encapsulates the clinical data recorded at the t-th encounter and T is the total number of visits. Let $\mathcal{D}_c$ denote diagnosis codes, $\mathcal{P}_c$ procedure codes, $\mathcal{M}_c$ medication codes, $\mathcal{D}_{txt}$ diagnostic textual descriptions, $\mathcal{P}_{txt}$ procedural textual descriptions, and $\mathcal{M}_{txt}$ medication textual descriptions. Each visit $v^{(i)}$ is a tuple comprising diagnosis codes, procedure codes, medication codes, and their associated textual descriptions.

$$v^{(i)} = (v_{dc}^{(i)}, v_{pc}^{(i)}, v_{mc}^{(i)}, v_{dtxt}^{(i)}, v_{ptxt}^{(i)}, v_{mtxt}^{(i)}) \tag{2}$$

where $v_{dc}^{(i)} \in \mathcal{D}_c$, $v_{pc}^{(i)} \in \mathcal{P}_c$, $v_{mc}^{(i)} \in \mathcal{M}_c$, $v_{dtxt}^{(i)} \in \mathcal{D}_{txt}$, $v_{ptxt}^{(i)} \in \mathcal{P}_{txt}$, $v_{mtxt}^{(i)} \in \mathcal{M}_{txt}$. We denote the entire EHR as $\mathcal{V}$ and the set of all patients as $\mathcal{P}$, and define $|\mathcal{V}| = |\mathcal{P}|$ to fully leverage the EHR information.

3.2 Medication Recommendation Problem

For patient $p \in \mathcal{P}$ with multiple visits V, provides the diagnoses $v_d^{(t)} = \{d_1, d_2, ..., d_k\}$, the procedure $v_p^{(t)} = \{p_1, p_2, ..., p_l\}$ and the previously prescribed medications $m_p^t = \{m_1, m_2, ..., m_r\}$. Our objective is to learn a prescription-prediction function $f(\cdot) : 2^{\mathcal{D}} \times 2^{\mathcal{P}} \mapsto 2^{\mathcal{M}}$, that generates a prescription for the given visit,

$$\hat{\mathbf{m}}^{(t)} = f(v_d^{(t)}, v_p^{(t)}, v_m^{(t)}). \tag{3}$$

where $\hat{\mathbf{m}}^{(t)} \subseteq \mathcal{M}$ is encoded into multi-hot vector.

4 Methodology

Our model operates in two distinct stages, as illustrated in Figure 2. In the first stage, we fine-tune a language model to extract semantic information from clinical narratives, utilizing dual channels at the disease and visit levels to fully leverage the semantic relationships between patient health status and medications. In the second stage, we enhance the output-layer hidden state distillation with attention-map-based distillation, transferring semantic knowledge from the LM-based framework to DCSENet.

4.1 LMs-Based Framework

The rich semantic information captured by LMs can effectively enhance downstream tasks [Qiu *et al.*, 2020; Han *et al.*, 2021]. Accordingly, we first fine-tune a language model to adapt it to the medication recommendation task and capture the semantic associations between patient health status and medications narratives. Then, we design a dual-channel module that separately models semantic information at both the disease and visit levels.

Clinical textual description

Drug–disease associations in clinical texts offer prior knowledge for medication recommendation.We extracted drug indication knowledge from DrugCentral [Avram *et al.*, 2023] and built direct links among drugs, diagnoses, and procedures. Using prompt templates, we further organized patients' historical visit records to supply personalized features for medication recommendation. The templates of prompt [Liu *et al.*, 2024a] for constructing the text is shown in Figure. 3. Given the EHR textual description V_{txt} of patient p, we partition the former into disease-level and visit-level segments to fully exploit the semantic information related to the patient's health status. At the disease level, we feed the patient's current diagnosis names v_{dtxt}^t and procedure names v_{ptxt}^t. At the visit level, we utilize all of the patient's visits. For i-th visit $v^{(i)}$, we concatenate the diagnostic text $v_{dtxt}^{(i)}$, procedural text $v_{ptxt}^{(i)}$, and medication text $v_{mtxt}^{(i)}$ into a single textual description $v_{txt}^{(i)}$, noting that the current visit v^t does not include any medication text. Then, we can obtain the textual descriptions at the visit level of healthcare $v_{txt} = [v_{txt}^{(1)}, v_{txt}^{(2)}, ..., v_{txt}^{(t)}]$.

LMs-based Clinical Narratives Encoder Module

For domain adaptation, we fine-tune the LMs by unfreezing only the final layer while keeping other parameters fixed. The adapted LMs then encode textual clinical data into dense representations. Given a text sequence *Text*, we obtain its representation $\boldsymbol{E}_{txt}$ by encoding it through the LMs.

$$\boldsymbol{E}_{txt} = LM(\textit{Text}) \tag{4}$$

Given the disease-level patient texts $v_{dtxt}^{(t)}$ and $v_{ptxt}^{(t)}$, the visit-level text V_{txt}, and the textual descriptions v_{mtxt} of all medications, we encode them via the LM to obtain the corresponding feature vectors $\boldsymbol{E}_{dtxt}^{(t)}$, $\boldsymbol{E}_{ptxt}^{(t)}$, $\boldsymbol{E}_{vtxt}$, and $\boldsymbol{E}_{mtxt}$.

Dual-Channel Semantic Fusion Module

The disease-level textual feature vector contains universal associations between drugs, diagnoses, and surgeries derived from the real world, while the visit-level textual feature vector incorporates patient-specific personalized information. To avoid conflicts between personalized conditions and universal knowledge, we first employ two separate cross-attention modules for modeling, respectively, and then utilize two MLPs (Multi-Layer Perceptrons) to project the information into an appropriate space, thereby resolving discrepancies and fusing the information. Our dual-channel module consists of a disease cross-attention and a visit cross-attention. In the disease channel, the textual medication features $\boldsymbol{E}_{mtxt}$ serve as the query, while the current diagnostic textual features $\boldsymbol{E}_{dtxt}$ and the procedural textual features $\boldsymbol{E}_{ptxt}$ are provided as two separate key–value pairs. In the visit channel, the same $\boldsymbol{E}_{mtxt}$ functions as the query, paired with the visit textual features $\boldsymbol{E}_{vtxt}$ provided as the key–value pair. as the key–value pair. Both channels use a Transformer Encoder (TransEec) [Vaswani *et al.*, 2017] to refine the medication representations $\boldsymbol{E}_{mtxt}$.

$$\boldsymbol{Z}_{dise} = \text{TransEec}\left(\boldsymbol{E}_{mtxt}; \left\{\boldsymbol{E}_{dtxt}^{(t)}, \boldsymbol{E}_{ptxt}^{(t)}\right\}\right) \tag{5}$$

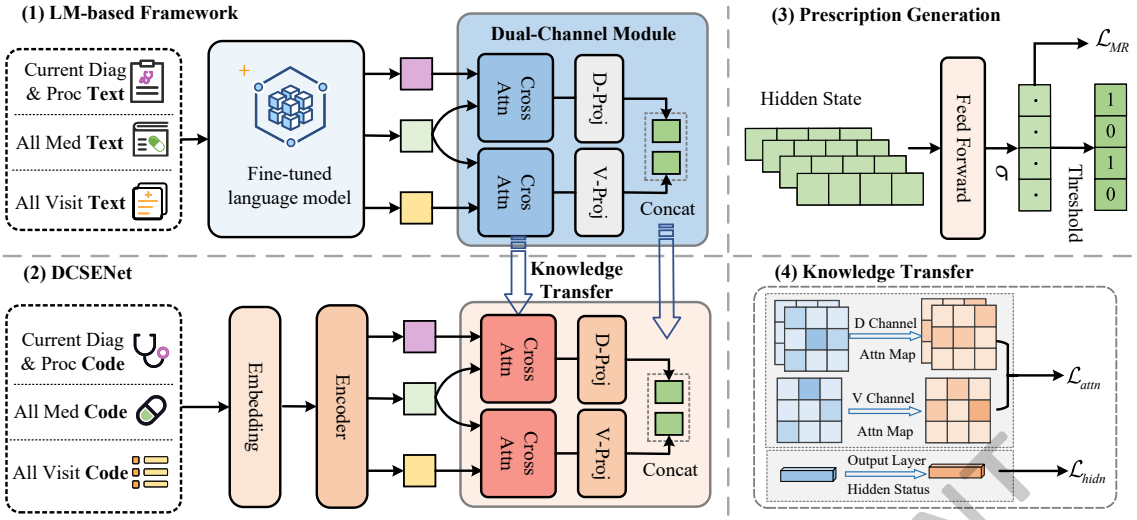

Figure 2: An overview of DCSENet: (1) A fine-tuned LM extracts deep semantic information from patient health and medication texts. A dual-channel module then models this information at both disease- and visit-level granularities. (2) DCSENet takes patient health and medication codes as input, encodes them, and performs inference using the same dual-channel module. Semantic knowledge is transferred to DCSENet through attention maps and output-layer hidden states.

Visit Text

In the i visit,
the patient had diagnosis:
<Diag_Name1>, <Diag_Name2>, <Diag_Name3>, ···;
the patient had procedures:
<Proc_Name1>, <Proc_Name2>, <Proc_Name3>, ···;
the patient had medications:
<Med_Name1>, <Med_Name2> , <Med_Name3> , ···.

Figure 3: Prompt for medication, diagnose, procedure and visit text.

$$\boldsymbol{Z}_{visit} = \text{TransEec}\left(\boldsymbol{E}_{mtxt}; \boldsymbol{E}_{vtxt}\right) \tag{6}$$

Subsequently, we project $\boldsymbol{Z}_{dise}$ and $\boldsymbol{Z}_{visit}$ into an appropriate space and concatenate them to obtain the hidden state of the output layer in the LM-based framework:

$$\boldsymbol{Z}^{'}_{dise} = \text{D-Proj}(\boldsymbol{Z}_{dise}),\ \boldsymbol{Z}^{'}_{visit} = \text{V-Proj}(\boldsymbol{Z}_{visit}) \tag{7}$$

$$\boldsymbol{H_T} = [\boldsymbol{Z}^{'}_{dise}; \boldsymbol{Z}^{'}_{visit}] \tag{8}$$

4.2 ID-Based Framework: DCSENet

While encoding patient health status and medication texts with LMs effectively captures rich semantic information, this approach also introduces significant redundancy in both data storage and computational complexity. Consequently, deploying such a medication recommendation system in real-world scenarios becomes more challenging. To mitigate this limitation, we design an identifier-based model , DCSENet, and transfer the semantic knowledge from LMs to DCSENet through knowledge distillation.

ID-based encoding

Unlike LM-based frameworks that utilize raw text inputs, DCSENet processes coded clinical information, including diagnoses, procedures, and medications. To represent these codes, we construct three embedding tables: $\mathbf{E}_{dc}$, $\mathbf{E}_{pc}$, and $\mathbf{E}_{mc}$, where each embedding vector corresponds to a diagnosis, procedure, or medication, respectively. For the i-th visit of a patient, we map the associated codes into their respective feature vectors, resulting in diagnosis matrix $\mathbf{D}_c^{(i)}$, procedure matrix $\mathbf{P}_c^{(i)}$, and medication matrix $\mathbf{M}_c^{(i)}$, respectively. Repeating this process across all visits, we obtain sequences $\left\{\mathbf{D}_c^{(1)}, \mathbf{D}_c^{(2)}, \ldots, \mathbf{D}_c^{(t)}\right\}$, $\left\{\mathbf{P}_c^{(1)}, \mathbf{P}_c^{(2)}, \ldots, \mathbf{P}_c^{(t)}\right\}$ and $\left\{\mathbf{M}_c^{(1)}, \mathbf{M}_c^{(2)}, \ldots, \mathbf{M}_c^{(t-1)}\right\}$, noting that the final visit does not include medication codes. To align with the visit-level textual input used in LM-based frameworks, we aggregate each feature matrix $\boldsymbol{D}_c^{(i)}$, $\boldsymbol{P}_c^{(i)}$, and $\boldsymbol{M}_c^{(i)}$ via summation,

$$\mathbf{h}_k^{(i)} = \sum_{c \in K_t} \mathbf{E}_k(c), \quad \text{for } k \in \{diag, proc, med\} \tag{9}$$

The resulting vectors $\mathbf{h}_{diag}^{(i)}$, $\mathbf{h}_{proc}^{(i)}$ and $h_{med}^{(i)}$ are concatenated and passed through a feed-forward network:

$$\mathbf{h}_{visit}^{(i)} = \text{FF}_1([\mathbf{h}_{diag}^{(i)}; \mathbf{h}_{proc}^{(i)}; \mathbf{h}_{med}^{(i)}]) \tag{10}$$

where the FF_1: $\mathbb{R}^{3h} \rightarrow \mathbb{R}^h$. By aggregating the per-visit representations, we form the visits feature matrix $\boldsymbol{V}_c$. Subsequently, the sequences $\boldsymbol{D}_c^{(i)}$, $\boldsymbol{P}_c^{(i)}$, $\boldsymbol{V}_c$ and $\boldsymbol{M}_c$ are encoded via a Transformer encoder to obtain the enhanced representations: $\boldsymbol{D}_c^{(i)'}$, $\boldsymbol{P}_c^{(i)'}$, $\boldsymbol{V}_c^{'}$ and $\boldsymbol{M}_c^{'}$.

Dual-Channel Fusion

For effective knowledge transfer, we employ a module architecture similar to the dual-channel textual semantic model module that was used in modeling the textual-information. Then, we can get $\boldsymbol{M}^{'}_{dise}$ and $\boldsymbol{M}^{'}_{visit}$. To fuse the feature information from the two channels, we employ two mapping functions to project the medication features from the two

channels into a shared feature space.

$$\boldsymbol{Z}_d = \text{D-Proj}(\boldsymbol{M}^{'}_{dise}),\ \boldsymbol{Z}_v = \text{V-Proj}(\boldsymbol{M}^{'}_{visit}) \tag{11}$$

where D-Proj and V-Proj denote two distinct feed-forward networks. Subsequently, we concatenate the $\boldsymbol{Z}_d$ and $\boldsymbol{Z}_v$:

$$\boldsymbol{H}_S = [\boldsymbol{Z}_d; \boldsymbol{Z}_v] \tag{12}$$

4.3 Prescription Generation

Given the drug hidden state representation refined from patient health information $\boldsymbol{H}$ we apply a linear projection to $\boldsymbol{H}$: $\mathbb{R}^{2h} \rightarrow \mathbb{R}$ to obtain a multi-hot vector $\hat{\boldsymbol{o}}$:

$$\hat{\boldsymbol{o}} = \sigma(FF_2(\boldsymbol{H})) \tag{13}$$

where FF_2 is a feed-forward network. Medications with predicted probabilities exceeding threshold δ in the multi-hot vector $\hat{\boldsymbol{m}}$ constitute the final prescription.

4.4 Loss Function

Following prior work, we employ a combined loss to optimize medication recommendation accuracy. For knowledge distillation, we utilize the Kullback-Leibler divergence (KL) [Van Erven and Harremos, 2014] to measure the divergence between attention maps, and cosine similarity to align the hidden states of the output layer.

Medication Recommendation Loss

Consistent with prior work [Yang *et al.*, 2023], we adopt multi-label loss $\mathcal{L}_{\text{bec}}$ and cross-entropy loss $\mathcal{L}_{\text{multi}}$ as the loss functions,

$$\mathcal{L}_{\text{bec}} = -\sum_{j=1}^{|M|} \mathbf{o}_j^{(i)} log(\hat{\mathbf{m}}_j^{(i)}) + (1 - \mathbf{o}_j^{(i)}) log(1 - \hat{\mathbf{m}}_j^{(i)}) \tag{14}$$

$$\mathcal{L}_{\text{multi}} = \sum_{p,q:\mathbf{o}_p^{(i)}=1,\mathbf{o}_q^i=0} \frac{max(0, 1 - (\hat{\mathbf{m}}_p^{(i)} - \hat{\mathbf{m}}_p^{(i)}))}{|M|} \tag{15}$$

where the superscript index (i) represents the i-th component of the vector. We combine these two losses as the overall loss for the medication recommendation task:

$$\mathcal{L}_{\text{MR}} = \alpha \mathcal{L}_{\text{bec}} + (1 - \alpha) \mathcal{L}_{\text{multi}} \tag{16}$$

where the α is a hyperparameter. Note that only the loss $\mathcal{L}_{\text{MR}}$ is first used to train the LM-based framework.

Knowledge Distillation Loss

Existing methods [Liu *et al.*, 2024a] perform distillation directly on the hidden states of the medication recommendation model.

$$\mathcal{L}_{\text{hidn}} = \text{Cosine}(\boldsymbol{H}_S \boldsymbol{W}_h,\ \boldsymbol{H}_T) \tag{17}$$

where $\boldsymbol{W}_h$ is an alignment projection matrix. However, the semantic representations derived from LM-based medication recommendation models reside in a feature space that is distinct from that of ID-based models. To address this issue, we introduce an attention-based distillation loss. Empirical studies [Clark *et al.*, 2019] have demonstrated that attention weights encode rich linguistic knowledge when processing textual data. The attention-based distillation loss function can be defined as:

$$\mathcal{L}_{\text{attn}} = \frac{1}{n} \sum_{i=1}^{n} \text{KL}(\boldsymbol{A}_i^S \parallel \boldsymbol{A}_i^T) \tag{18}$$

where $\text{KL}(\cdot \parallel \cdot)$ denotes the Kullback-Leibler divergence, and n is the number of attention heads. $\boldsymbol{A}_i^S$ and $\boldsymbol{A}_i^T$ represent the attention maps from the LM-based framework and DCSENet models, respectively, for the i-th attention head. We perform knowledge distillation on the attention maps within the cross-attention modules of the TransDec for multiple channels, resulting in losses $\mathcal{L}_{\text{attn}}^d$ (for medication-diagnose attention), $\mathcal{L}_{\text{attn}}^p$ (for medication-procedure attention), and $\mathcal{L}_{\text{attn}}^v$ (for medication-visit attention). These individual channel losses are combined through weighted summation to obtain the overall attention distillation loss:

$$\mathcal{L}_{\text{attn}} = \mathcal{L}_{\text{attn}}^d + \mathcal{L}_{\text{attn}}^p + \mathcal{L}_{\text{attn}}^v \tag{19}$$

Merged Loss

The final training objective for the DCSENet is a weighted sum of the $\mathcal{L}_{\text{MR}}$ and the $\mathcal{L}_{\text{KD}}$:

$$\mathcal{L} = (1 - \beta)\mathcal{L}_{\text{MR}} + \beta \mathcal{L}_{\text{KD}} \tag{20}$$

where β is a hyperparameter that balances the two loss.

5 Experiments

We conducted extensive experiments to demonstrate the effectiveness of DCSENet.

5.1 Datasets

We conducted experiments on the MIMIC-III [Johnson *et al.*, 2016] and MIMIC-IV [Johnson *et al.*, 2023]. Following the setting in [Yang *et al.*, 2023; Yang *et al.*, 2021], we split the dataset into training, validation and test as $4:1:1$. The statistical information of the preprocessed dataset is shown in Appendix A.

5.2 Settings

This section details the experimental setup, including hyperparameter configurations, baselines, and evaluation metrics.

Hyperparameter Configuration

Our proposed method is implemented using PyTorch 2.6.0 (with CUDA 12.4 support), which is built on the Python 3.10.16 environment. All experiments are conducted on an Intel Xeon Platinum 8260 server, equipped with 24 CPU cores, 188 GB of RAM, and a single NVIDIA L40 GPU with 46 GB of VRAM. In our model, the embedding dimension for LM-based framework is set to 768. For the compressed DCSENet, this dimension is configured to 64. The optimal hyper-parameters weight $\alpha = 0.95$, $\beta = 0.9$, and the threshold $\delta = 0.5$ are determined based on their performance on the validation set. We train our model with the Adam optimizer [Kingma and Ba, 2014] at an initial learning rate of 1×10^{-4}.

Model	MIMIC-III					MIMIC-IV				
	Jaccard↑	F1↑	PRAUC↑	DDI↓	Avg.#Med	Jaccard↑	F1↑	PRAUC↑	DDI↓	Avg.#Med
LR	0.4920 ±0.0028	0.6491 ±0.0026	0.7552 ±0.0031	0.0784 ±0.0008	16.4293 ±0.1487	0.4395 ±0.0010	0.5850 ±0.0010	0.7224 ±0.0009	0.0764 ±0.0005	8.2226 ±0.0278
ECC	0.4868 ±0.0031	0.6428 ±0.0028	0.7382 ±0.0026	0.0805 ±0.0007	16.0100 ±0.0972	0.4172 ±0.0013	0.5575 ±0.0014	0.7233 ±0.0015	0.0754 ±0.0004	7.4800 ±0.0346
LEAP	0.4441 ±0.0027	0.6068 ±0.0028	0.6475 ±0.0056	0.0730 ±0.0008	18.8987 ±0.0779	0.4085 ±0.0012	0.5584 ±0.0013	0.5306 ±0.0019	0.0673 ±0.0004	10.6090 ±0.0554
GAMENet	0.5160 ±0.0020	0.6713 ±0.0018	0.7642 ±0.0012	0.0796 ±0.0004	25.4674 ±0.1470	0.4480 ±0.0015	0.5995 ±0.0015	0.7099 ±0.0014	0.0838 ±0.0004	16.2131 ±0.0934
COGNet	0.5308 ±0.0018	0.6842 ±0.0017	0.7699 ±0.0014	0.0840 ±0.0005	28.2382 ±0.0580	0.4682 ±0.0006	0.6165 ±0.0006	0.6883 ±0.0005	0.0713 ±0.0003	18.0124 ±0.0384
VITA	0.5263 ±0.0016	0.6786 ±0.0027	0.7623 ±0.0028	0.0765 ±0.0008	29.2662 ±0.1565	0.4716 ±0.0013	0.6139 ±0.0013	0.6792 ±0.0012	0.0796 ±0.0005	18.4972 ±0.0613
SafeDrug	0.5120 ±0.0025	0.6685 ±0.0023	0.7640 ±0.0027	0.0619 ±0.0005	20.4573 ±0.1561	0.4457 ±0.0012	0.5972 ±0.0012	0.6917 ±0.0016	**0.0504** ±0.0003	12.3890 ±0.0364
MoleRec	0.5296 ±0.0031	0.6837 ±0.0025	0.7767 ±0.0021	0.0726 ±0.0009	21.2022 ±0.1639	0.4562 ±0.0019	0.6065 ±0.0019	0.6929 ±0.0022	0.0687 ±0.0004	12.9925 ±0.0488
MR-DTR	0.5414 ±0.0023	0.6942 ±0.0016	0.7922 ±0.0028	**0.0529** ±0.0006	19.6641 ±0.1348	0.4625 ±0.0015	0.6158 ±0.0019	0.7008 ±0.0026	0.0508 ±0.0007	12.5100 ±0.0691
NLA-MMR	0.5429 ±0.0027	0.6958 ±0.0024	0.7890 ±0.0016	0.0835 ±0.0008	22.5158 ±0.1684	0.4491 ±0.0044	0.6051 ±0.0042	0.6921 ±0.0045	0.0785 ±0.0006	13.6846 ±0.0761
DCSENet	**0.5632** ±0.0010	**0.7120** ±0.0008	**0.8048** ±0.0009	0.0796 ±0.0002	23.8144 ±0.0678	**0.4823** ±0.0006	**0.6319** ±0.0006	**0.7389** ±0.0006	0.0846 ±0.0003	14.8278 ±0.0333

Table 1: Performance Comparison on MIMIC-III and MIMIC-IV. The best result is highlighted in bold. The second-best result is underlined. ↑ indicates that higher values are preferable, while ↓ arrow signifies that lower values are more desirable. Avg.#Med denotes the average number of drugs recommended per visit.

model	Jaccard	F1	PRAUC
w/o Diag	0.5174	0.6724	0.7675
w/o Proc	0.5386	0.6912	0.7886
w/o Dise-channel	0.4938	0.6432	0.7422
w/o Visit-channel	0.5413	0.6939	0.7819
w/o Distillation	0.5277	0.6822	0.7747
w/o $\mathcal{L}_{\text{attn}}$	0.5444	0.6965	0.7920
w/o $\mathcal{L}_{\text{hidn}}$	0.5411	0.6933	0.7878
DCSENet	**0.5632**	**0.7120**	**0.8048**

Table 2: Comparison of DCSENet and its variants.

Baselines and Metrics

We select a diverse set of medication recommendation models as baselines to comprehensively evaluate both variants of DCSENet. These include:

- **Classical models**: Logistic Regression (LR) and Ensemble Classifier Chain (ECC) [Read *et al.*, 2011].
- **Traditional ID-based models**: GAMENet [Shang *et al.*, 2019], COGNet [Wu *et al.*, 2022] and VITA [Kim *et al.*, 2024].
- **Extra-information-enhanced model**: SafeDrug [Yang *et al.*, 2021], MoleRec [Yang *et al.*, 2023], NLA-MMR [Tan *et al.*, 2024] and MR-DTR [Li *et al.*, 2025].

We employ three widely-used metrics: Jaccard similarity, F1 score and PRAUC, to verify the accuracy and effectiveness of these models. Due to space constraints, details of the baselines and metrics are provided in the Appendix B.

5.3 Main Result

We compare the performance of DCSENet against the baseline models; the results are presented in Table 1. Overall, DCSENet outperforms all baseline models across three evaluation metrics, Jaccard, F1-score, and PRAUC. Traditional machine learning models such as LR and ECC fail to effectively leverage patients' historical visit information, resulting in lower performance. LEAP exhibits suboptimal performance due to its difficulty in modeling long sequences of patients' visit records. Although GAMENet incorporates a memory module, while COGNet and VITA apply Transformers that are more adept at modeling long sequences, these methods all overlook the critical role of external knowledge, resulting in suboptimal performance. SafeDrug and MoleRec incorporate the molecular structural information of medications, while MR-DTR introduces clinical entity structure information, leading to improved performance. NLA-MMR leverages language models to extract clinical textual information, which substantially enhances its performance; however, it lacks effective utilization of the semantic relationships between patient health status and medications, resulting in only moderate improvements. In contrast, the LM-based framework uncovers semantic associations between patient health status and medications at multiple levels and effectively transfers this knowledge to DCSENet through knowledge distillation. With enriched textual semantics, DCSENet achieves the best overall performance.

model	Jaccard	F1	PRAUC
NoLM	0.5277	0.6822	0.7747
BERT-base	0.5451	0.6969	0.7911
RoBERTa-base	0.5469	0.6985	0.7923
ClinicalBERT	0.5463	0.6981	0.7912
BioBERT	0.5555	0.7064	0.7984
BlueBERT	0.5493	0.7007	0.7922
PubMedBERT	**0.5632**	**0.7120**	**0.8048**

Table 3: Impact of different LMs on DCSENet performance.

5.4 Ablation Study

To validate the contribution of each component within our model, we conducted a comprehensive set of ablation experiments. The results are summarized in Table 2. Initially, we ablated the diagnostic and procedural information separately; the observed performance degradation confirms that both types of information are crucial for effective medication recommendation. Next, to evaluate the importance of our dual-channel architecture, we ablated the disease channel and the visit channel individually. The results reveal that the disease channel plays a more dominant role, while the visit channel further enhances overall performance. This aligns

with the intuition that medication recommendations primarily target the patient's current health status, whereas historical visit records serve as auxiliary information. To assess the effectiveness of knowledge transfer, we first removed the entire knowledge distillation module, which resulted in a significant decline in performance. Additionally, we ablated the individual loss components, $\mathcal{L}_{\text{attn}}$ and $\mathcal{L}_{\text{hidn}}$, separately. The results demonstrate that both losses are beneficial to the knowledge transfer process.

5.5 Impact of Different LMs on Performance

To assess how various textual encoding capabilities of different language models (LMs) affect the performance of DCSENet, we conducted experiments using six distinct LMs: BERT-base [Devlin *et al.*, 2019], RoBERTa-base [Warstadt *et al.*, 2020], ClinicalBERT [Huang *et al.*, 2019], BioBERT [Lee *et al.*, 2020], PubMedBERT [Gu *et al.*, 2021], and BlueBERT [Peng *et al.*, 2019]. As shown in Table 3, all variants outperform the baseline model without any LM. BERT-base and RoBERTa-base are general-purpose language models, whereas ClinicalBERT, despite its clinical text adaptation, exhibits relatively lower performance due to its insufficient coverage of extensive biomedical knowledge beyond standard clinical corpora. In contrast, BioBERT, BlueBERT, and PubMedBERT, all pre-trained on large-scale biomedical literature, outperform others in medication recommendation, highlighting the advantage of domain-specific pretraining. Of these domain-adapted models, PubMedBERT delivers optimal performance and is thus chosen as our foundational LM.

5.6 Effectiveness Analysis of Knowledge Transfer

To evaluate the effectiveness of knowledge transfer, we conducted the following experiments:

Attention Map Analysis

To illustrate the effectiveness of our attention-map-based knowledge transfer for medication recommendation, we visualize the attention maps of a representative patient during the distillation process, as shown in Figure 4. LM-based framework effectively captures the associations between diagnoses and medications. For example, diagnosis 568.81 (acute peritonitis) is characterized by a triad of symptoms: severe abdominal pain, signs of peritoneal irritation, and systemic inflammatory response. Correspondingly, the medications—A06A (Laxatives), N02A (Opioids), and N05B (Anxiolytics)—target the primary complications such as gut dysmotility, severe pain, and anxiety/stress. These medication selections represent a common combination in the comprehensive management of acute peritonitis. After knowledge transfer, DCSENet successfully acquires this semantic understanding, with its attention maps showing a high degree of similarity to LM-base framework. In contrast, the attention maps of the model without knowledge transfer display significant discrepancies, highlighting the effectiveness of the semantic transfer facilitated by our approach.

Parameter Count and Efficiency Analysis

As previously noted, employing LMs as encoders introduces significant computational overhead. The LM-based framework contains 510.22 MB of parameters and requires 453.29 seconds for inference. We compare several strong baseline models. As shown in Figure 5, MoleRec has the lowest computational overhead but exhibits subpar accuracy. COGNet and VITA, which rely on sequential decision processes, incur longer inference times. MR-DTR efficiently encodes clinical entity structures with relatively moderate computational cost while yielding suboptimal performance. NLA-MMR, which heavily leverages LM encoding, introduces substantial additional computational burden. In contrast, our DCSENet transfers LM-based semantic knowledge to an ID-based model, achieving superior accuracy while maintaining low computational costs, demonstrating an optimal balance between performance and efficiency.

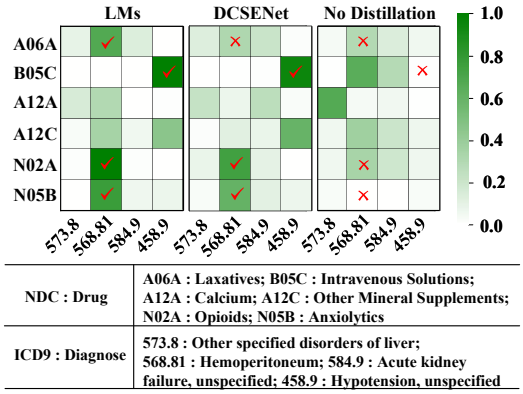

Figure 4: Visualization of attention maps. The horizontal axis displays ICD-9 diagnosis codes, the vertical axis shows NDC medication codes, with checkmarks indicating correct recommendations and crosses indicating incorrect ones.

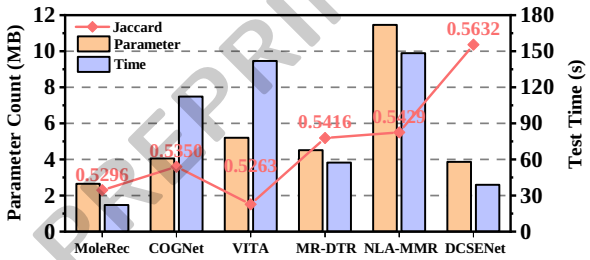

Figure 5: Comparison of model performance and cost.

6 Conclusion

In this paper, we propose DCSENet, a dual-channel semantics-enhanced framework for medication recommendation. DCSENet leverages language models to capture the semantic information embedded in patient health records and medication-related texts. The disease channel establishes drug-symptom associations, while the visit channel captures heterogeneous patient information; together, they enable accurate personalized medication recommendations. We further adopt a knowledge distillation strategy based on attention maps, which achieves superior knowledge transfer efficacy.Extensive experimental validations on the MIMIC-III and MIMIC-IV datasets consistently demonstrate the superiority of the proposed DCSENet over existing approaches.

Ethical Statement

This study employs the HIPAA-compliant, deidentified MIMIC dataset; its use does not constitute human subjects research under HIPAA rules. Ethical precautions were strictly followed to safeguard participant privacy and rights.

References

[Avram *et al.*, 2023] Sorin Avram, Thomas B Wilson, Ramona Curpan, Liliana Halip, Ana Borota, Alina Bora, Cristian G Bologa, Jayme Holmes, Jeffrey Knockel, Jeremy J Yang, et al. Drugcentral 2023 extends human clinical data and integrates veterinary drugs. *Nucleic acids research*, 51(D1):D1276–D1287, 2023.

[Choi *et al.*, 2016] 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*, 29, 2016.

[Clark *et al.*, 2019] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019.

[Cowie *et al.*, 2017] Martin R Cowie, Juuso I Blomster, Lesley H Curtis, Sylvie Duclaux, Ian Ford, Fleur Fritz, Samantha Goldman, Salim Janmohamed, Jörg Kreuzer, Mark Leenay, et al. Electronic health records to facilitate clinical research. *Clinical Research in Cardiology*, 106:1–9, 2017.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[Gu *et al.*, 2021] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

[Han *et al.*, 2021] Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, et al. Pre-trained models: Past, present and future. *Ai Open*, 2:225–250, 2021.

[Hinton *et al.*, 2015] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[Huang *et al.*, 2019] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

[Jiao *et al.*, 2019] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*, 2019.

[Johnson *et al.*, 2016] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[Johnson *et al.*, 2023] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

[Kim *et al.*, 2024] Taeri Kim, Jiho Heo, Hongil Kim, Kijung Shin, and Sang-Wook Kim. Vita:'carefully chosen and weighted less' is better in medication recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8600–8607, 2024.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Lee *et al.*, 2020] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[Li *et al.*, 2025] Yishuo Li, Qi Zhang, Wenpeng Lu, Xueping Peng, Weiyu Zhang, Jiasheng Si, Yongshun Gong, and Liang Hu. Time-aware medication recommendation via intervention of dynamic treatment regimes. In *Proceedings of the ACM on Web Conference 2025*, pages 5161–5172, 2025.

[Liu *et al.*, 2024a] Q Liu, X Wu, X Zhao, Y Zhu, Z Zhang, F Tian, and Y Zheng. Large language model distilling medication recommendation model. arxiv. *arXiv preprint arXiv:2402.02803*, 2024.

[Liu *et al.*, 2024b] Ruiping Liu, Kailun Yang, Alina Roitberg, Jiaming Zhang, Kunyu Peng, Huayao Liu, Yaonan Wang, and Rainer Stiefelhagen. Transkd: Transformer knowledge distillation for efficient semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[Peng *et al.*, 2019] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*, 2019.

[Qiu *et al.*, 2020] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. Pre-trained models for natural language processing: A survey. *Science China technological sciences*, 63(10):1872–1897, 2020.

[Read *et al.*, 2011] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Machine learning*, 85:333–359, 2011.

[Shang *et al.*, 2018] Junyuan Shang, Shenda Hong, Yuxi Zhou, Meng Wu, and Hongyan Li. Knowledge guided

multi-instance multi-label learning via neural networks in medicines prediction. In *Asian Conference on Machine Learning*, pages 831–846. PMLR, 2018.

[Shang *et al.*, 2019] Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. Gamenet: Graph augmented memory networks for recommending medication combination. In *proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1126–1133, 2019.

[Skou *et al.*, 2022] Søren T Skou, Frances S Mair, Martin Fortin, Bruce Guthrie, Bruno P Nunes, J Jaime Miranda, Cynthia M Boyd, Sanghamitra Pati, Sally Mtenga, and Susan M Smith. Multimorbidity. *Nature Reviews Disease Primers*, 8(1):48, 2022.

[Sun *et al.*, 2020] Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. Mobilebert: a compact task-agnostic bert for resource-limited devices. *arXiv preprint arXiv:2004.02984*, 2020.

[Sun *et al.*, 2022] Hongda Sun, Shufang Xie, Shuqi Li, Yuhan Chen, Ji-Rong Wen, and Rui Yan. Debiased, longitudinal and coordinated drug recommendation through multi-visit clinic records. *Advances in Neural Information Processing Systems*, 35:27837–27849, 2022.

[Tan *et al.*, 2022] Yanchao Tan, Chengjun Kong, Leisheng Yu, Pan Li, Chaochao Chen, Xiaolin Zheng, Vicki S Hertzberg, and Carl Yang. 4sdrug: Symptom-based set-to-set small and safe drug recommendation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3970–3980, 2022.

[Tan *et al.*, 2024] Jie Tan, Yu Rong, Kangfei Zhao, Tian Bian, Tingyang Xu, Junzhou Huang, Hong Cheng, and Helen Meng. Natural language-assisted multi-modal medication recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2200–2209, 2024.

[Tang *et al.*, 2019] Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. Distilling task-specific knowledge from bert into simple neural networks. *arXiv preprint arXiv:1903.12136*, 2019.

[Van Erven and Harremos, 2014] Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Wang *et al.*, 2024] Yanda Wang, Lin Yue, and Ying Li. Topological knowledge enhanced personalized ranking model for sequential medication recommendation. In *International Conference on Advanced Data Mining and Applications*, pages 109–124. Springer, 2024.

[Warstadt *et al.*, 2020] Alex Warstadt, Yian Zhang, Haau-Sing Li, Haokun Liu, and Samuel R Bowman. Learning which features matter: Roberta acquires a preference for linguistic generalizations (eventually). *arXiv preprint arXiv:2010.05358*, 2020.

[Wu *et al.*, 2022] Rui Wu, Zhaopeng Qiu, Jiacheng Jiang, Guilin Qi, and Xian Wu. Conditional generation net for medication recommendation. In *Proceedings of the ACM Web Conference 2022*, pages 935–945, 2022.

[Xu *et al.*, 2024] Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*, 2024.

[Yang *et al.*, 2021] Chaoqi Yang, Cao Xiao, Fenglong Ma, Lucas Glass, and Jimeng Sun. Safedrug: Dual molecular graph encoders for recommending effective and safe drug combinations. *arXiv preprint arXiv:2105.02711*, 2021.

[Yang *et al.*, 2023] Nianzu Yang, Kaipeng Zeng, Qitian Wu, and Junchi Yan. Molerec: Combinatorial drug recommendation with substructure-aware molecular representation learning. In *Proceedings of the ACM Web Conference 2023*, pages 4075–4085, 2023.

[Zhang *et al.*, 2017] Yutao Zhang, Robert Chen, Jie Tang, Walter F Stewart, and Jimeng Sun. Leap: learning to prescribe effective and safe treatment combinations for multimorbidity. In *proceedings of the 23rd ACM SIGKDD international conference on knowledge Discovery and data Mining*, pages 1315–1324, 2017.

[Zhang *et al.*, 2023a] Junjie Zhang, Xuan Zang, Hao Chen, and Buzhou Tang. E-hmfnet: A knowledge-enhanced hierarchical molecular representation fusion network for drug recommendation. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1690–1695. IEEE, 2023.

[Zhang *et al.*, 2023b] Yingying Zhang, Xian Wu, Quan Fang, Shengsheng Qian, and Changsheng Xu. Knowledge-enhanced attributed multi-task learning for medicine recommendation. *ACM Transactions on Information Systems*, 41(1):1–24, 2023.

[Zhang *et al.*, 2025a] Haodi Zhang, Jiawei Wen, Jiahong Li, Yuanfeng Song, Liang-Jie Zhang, and Lin Ma. Sspnet: Leveraging robust medication recommendation with history and knowledge. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 9465–9473, 2025.

[Zhang *et al.*, 2025b] Kai Zhang, Rui Zhu, Shutian Ma, Jingwei Xiong, Yejin Kim, Fabricio Murai, and Xiaozhong Liu. Kedrec-lm: A knowledge-distilled explainable drug recommendation large language model. *arXiv preprint arXiv:2502.20350*, 2025.

[Zheng *et al.*, 2021] Zhi Zheng, Chao Wang, Tong Xu, Dazhong Shen, Penggang Qin, Baoxing Huai, Tongzhu Liu, and Enhong Chen. Drug package recommendation via interaction-aware graph induction. In *Proceedings of the web conference 2021*, pages 1284–1295, 2021.