Recommendation for New Drugs with Limited Prescription Data

Zhenbang Wu[1], Huaxiu Yao[2], Zhe Su[3], David M Liebovitz[4], Lucas M Glass[5], James Zou[2], Chelsea Finn[2], Jimeng Sun[1]

[1] University of Illinois Urbana-Champaign, [2] Stanford University,
[3] Zhejiang University, [4] Northwestern University, [5] IQVIA

Abstract

Drug recommendation assists doctors in prescribing personalized medications to patients based on their health conditions. However, newly approved drugs do not have much historical prescription data and cannot leverage existing drug recommendation methods. To address this, we propose EDGE, which maintains a drug-dependent multi-phenotype few-shot learner to bridge the gap between existing and new drugs. Experiment results show that EDGE can adapt to the recommendation for a new drug with limited prescription data from a few patients.

1 Introduction

Drug recommendation assists doctors in recommending personalized medications to patients based on their health conditions. Existing drug recommendation methods typically formulate it as a supervised multi-label classification problem and train on massive prescription data (Zhang et al., 2017; Zitnik et al., 2018; Shang et al., 2019b; Yang et al., 2021; Rui Wu & Wu., 2022; Tan et al., 2022b). However, in reality, new drugs come to the market all the time (FDA, 2022). Most of these newly approved drugs do not have much historical data to support model training (Blass, 2021). As a result, existing drug recommendation methods are no longer applicable when new drugs appear.

In this paper, we formulate the recommendation of new drugs as a few-shot classification problem. Given a new drug with limited prescription data from a few support patients (e.g., from clinical trials (Duijnhoven et al., 2013)), the model should quickly adapt to the recommendation for this drug. However, directly applying existing meta-learning algorithms faces the following challenges. **Complex relations among diseases and drugs:** diseases and medicines can have inherent and higher order relations, which are not explicitly captured by general meta-learning algorithms. **Numerous false-negative patients:** there exist many false-negative patients who were eligible but did not yet use the new drug, which will substantially confuse the model learning.

To address this, we introduce EDGE, a drug-dependent multi-phenotype few-shot learner to quickly adapt to the recommendation for a new drug with limited support patients. Specifically, EDGE leverages the drug ontology to link new drugs to existing drugs with similar treatment effects and learns ontology-based drug representations. Such drug representations are used to customize the metric space of the phenotype-driven patient representations, which are composed of a set of phenotypes capturing complex patient health status. Lastly, EDGE eliminates the false-negative supervision signal using an external drug-disease knowledge base.

We evaluate EDGE on the public EHR data (MIMIC-IV) and private industrial claims data. Results show that EDGE achieves 7.3% improvement on the ROC-AUC score over the best baseline.

6th Workshop on Meta-Learning at NeurIPS 2022, New Orleans.

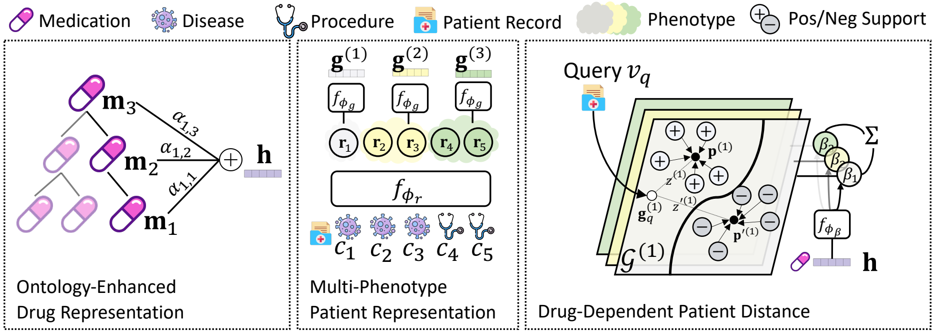

Figure 1: EDGE consists of the following modules: **(1) Ontology-enhanced drug encoder** that fuses ontology information into drug representation to link new drugs to existing drugs with similar treatment effects; **(2) Multi-phenotype patient encoder** that represents each patient with a set of phenotype-level representations to capture the complex patient's health status; **(3) Drug-dependent distance measures** that customizes the patient similarity with drug-dependent phenotype importance scores; **(4) knowledge-guide negative sampling** that eliminates the false-negative supervision signal.

2 Knowledge-driven New Drug Recommendation

2.1 Problem Formulation

Denote a patient with a list of diseases as $v = [c_1, \ldots, c_V]$. Assume the entire drug set $\mathcal{M}$ is partitioned into a set of existing drugs $\mathcal{M}^{old}$ and a set of new drugs $\mathcal{M}^{new}$, where $\mathcal{M}^{old} \cap \mathcal{M}^{new} = \emptyset$. Each existing drug $m_i \in \mathcal{M}^{old}$ has sufficient patients using the drug m_i (e.g., from EHR data). Each new drug $m_t \in \mathcal{M}^{new}$ is associated with a small support set $\mathcal{S}_t = \{v_j\}_{j=1}^{N_s}$ consisting of patients using the drug m_t, and an unlabeled query patient set $\mathcal{Q}_t = \{v_j\}_{j=1}^{N_q}$, where N_s and N_q are the number of patients in the support and query sets, respectively. The goal of new drug recommendation is to train a model $f_\phi(\cdot)$ parameterized by ϕ on existing drugs $\mathcal{M}^{old}$, such that it can adapt to new drug $m_t \in \mathcal{M}^{new}$ given the small support set $\mathcal{S}_t$, and make correct recommendation on the query set $\mathcal{Q}_t$.

2.2 Ontology-enhanced drug representation learning

Though many new drugs have not been used regularly in clinical practice, they still belong to the same drug category (from a drug ontology) as some existing drugs and share similar treatment effects, implicitly indicating similar patient populations. We here leverage the drug ontology to enrich the drug representation by attentively combing the drug itself and its corresponding ancestors.

Concretely, for the drug m_i, we obtain its basic embeddings $\mathbf{m}_i \in \mathbb{R}^e$ by feeding its description into Clinical-BERT (Alsentzer et al., 2019). Then, follow Choi et al. (2017), we use the basic embeddings of drug m_i and its ancestors to calculate the ontology-enriched drug representation as,

$$\mathbf{h} = \sum_{j \in \mathcal{A}_i} \alpha_{i,j} \mathbf{m}_j, \quad \mathbf{h} \in \mathbb{R}^e, \tag{1}$$

where $\mathcal{A}_i$ denotes the set of drug m_i and its ancestors, and the attention score $\alpha_{i,j}$ represents the importance of ancestor m_j for drug m_i, which is calculated as,

$$\alpha_{i,j} = \frac{\exp(f_{\phi_a}(\mathbf{m}_i \oplus \mathbf{m}_j))}{\sum_{k \in \mathcal{A}_i} \exp(f_{\phi_a}(\mathbf{m}_i \oplus \mathbf{m}_k))}, \quad \alpha_{i,j} \in [0, 1], \tag{2}$$

where $\oplus$ denotes the concatenation operator, and $f_{\phi_a}(\cdot) : \mathbb{R}^{2e} \mapsto \mathbb{R}$. In this way, we fuse the ontology information into the representation $\mathbf{h}$ for drug m_i, which is later used to customize the metric space of phenotype-driven patient representations, introduced next.

2.3 Multi-phenotype patient representation learning

Patient health status includes many factors, such as disease progression, comorbidities, ongoing treatments, individual drug response, and drug side effects. Encoding each patient into a single vector may not capture the complete information, especially for patients with complex health conditions. Therefore, we define a set of phenotypes and represent each patient with a set of phenotype vectors.

Specifically, for every support/query patient v with a list of diseases $[c_1, \ldots, c_V]$, EDGE first leverages domain knowledge to group diseases into different phenotypes. To obtain the representation for the l-th phenotype, we take the representations from all diseases that belong to that phenotype, project them to a lower dimension, and calculate their mean representations as,

$$\mathbf{g}^{(l)} = \frac{1}{|\mathcal{G}^{(l)}|} \sum_{j \in \mathcal{G}^{(l)}} f_{\phi_g}(\mathbf{r}_j), \;\; \mathbf{g}^{(l)} \in \mathbb{R}^g, \tag{3}$$

where $\mathbf{r}_j$ is the representation of disease c_j obtained via a bi-directional GRU; $\mathcal{G}^{(l)}$ represents the set of diseases whose phenotype is l; $f_{\phi_g}(\cdot) : \mathbb{R}^e \to \mathbb{R}^g$ and $g < e$. Further, based on the multi-phenotype patient representations, we calculate the phenotype-level prototypes from the support set $\mathcal{S}_i$ of drug m_i as,

$$\mathbf{p}^{(l)} = \frac{1}{|\mathcal{S}_i|} \sum_{j \in \mathcal{S}_i} \mathbf{g}_j^{(l)}, \;\; \mathbf{p}^{(l)} \in \mathbb{R}^g, \tag{4}$$

where $\mathbf{g}_j^{(l)}$ is the l-th phenotype representation for patient v_j from the support set $\mathcal{S}_i$ of drug m_i. In this way, we encode the support set $\mathcal{S}_i$ into a set of phenotype-level prototypes $\{\mathbf{p}^{(l)}\}_{l=1}^L$, which is further used to calculate the drug-dependent patient distance with the multi-phenotype representations $\{\mathbf{g}_q^{(l)}\}_{l=1}^L$ of the query patient v_q, described next.

2.4 Drug-dependent patient distance

We finally focus on defining a reasonable metric to measure the distance between query patient and support prototypes. Intuitively, different drugs may have different focuses when comparing patients. Therefore, we leverage the drug representation to learn a drug-dependent patient distance.

Specifically, given the phenotype-driven prototypes $\{\mathbf{p}^{(l)}\}_{l=1}^L$ from support patients $\mathcal{S}_i$, and the multi-phenotype representations $\{\mathbf{g}_q^{(l)}\}_{l=1}^L$ for the query patient v_q, we first calculate the per-phenotype distance between the query and support. For a specific phenotype l, the distance is defined as,

$$z^{(l)} = d(\mathbf{g}_q^{(l)}, \mathbf{p}^{(l)}), \;\; z^{(l)} \in \mathbb{R}, \tag{5}$$

where $d(\cdot)$ is the euclidean distance. Next, with the ontology-enhanced drug representation $\mathbf{h}$ for drug m_i, we calculate the drug-dependent importance weights over different phenotypes as,

$$\boldsymbol{\beta} = \sigma(f_{\phi_\beta}(\mathbf{h})), \;\; \boldsymbol{\beta} \in [0,1]^L, \tag{6}$$

where $\sigma(\cdot)$ denotes the sigmoid function, and $f_{\phi_\beta}(\cdot) : \mathbb{R}^e \mapsto \mathbb{R}^L$. According to the importance weight, the probability of recommending drug m_i to the query patient v_q is calculated as,

$$p_\phi(y_q = +|v_q) = \frac{\exp(-\boldsymbol{\beta}^\top \mathbf{z})}{\exp(-\boldsymbol{\beta}^\top \mathbf{z}) + \exp(-\boldsymbol{\beta}^\top \mathbf{z}')}, \tag{7}$$

where $\mathbf{z} \in \mathbb{R}^L$ combines all L per-phenotype distances, $\boldsymbol{\beta}^\top \mathbf{z}$ aggregates the distances depending on drug m_i, and $\mathbf{z}'$ is the distance vector between the query v_q and the negative (i.e., not using drug m_i) support patients obtained via negative sampling, introduced next.

2.5 Knowledge-Guided Negative Sampling

Negative sampling plays a vital role in model training. In real-world scenarios, many drugs can treat the same disease, but usually, only one of them is prescribed. For any given drug, there exist many false-negative patients that could but eventually did not use the drug for several reasons, such as availability, doctor preference, or insurance coverage. The number of false-negative supervision signals will confuse the model learning, especially in the few-shot learning setting.

To address this issue, EDGE introduces the following knowledge-guided negative sampling strategy. We leverage the MEDI (Wei et al., 2013) drug-disease relationship (i.e., pairs of drugs and their target diseases). When performing negative sampling for a given drug, we only sample from the negative patients whose diagnoses do not overlap with the listed target diseases of that drug. Note that we only use this strategy during training so that the information will not leak to testing.

Table 1: Results on new drug recommendation. ± denotes the 95% confidence interval. **Bold** indicates the best result and underline indicates second best. * indicates that EDGE achieves significant improvement over the best baseline method (i.e., the p-value is smaller than 0.05). Experiment results show that EDGE can adapt to new drugs and make correct recommendations.

Method		**MIMIC-IV**			**Claims**	
	ROC-AUC	**Precision@100**	**Recall@100**	**ROC-AUC**	**Precision@100**	**Recall@100**
RNN	0.4612 ± 0.0091	0.0279 ± 0.0062	0.0040 ± 0.0005	0.5012 ± 0.0062	0.0130 ± 0.0014	0.0047 ± 0.0006
GAMENet	0.6831 ± 0.0087	0.1047 ± 0.0121	0.0722 ± 0.0070	0.6361 ± 0.0077	0.0338 ± 0.0058	0.0203 ± 0.0044
SafeDrug	0.7488 ± 0.0102	0.1055 ± 0.0103	0.0655 ± 0.0064	0.4792 ± 0.0048	0.0129 ± 0.0016	0.0022 ± 0.0003
MAML	0.7197 ± 0.0105	0.0549 ± 0.0070	0.0356 ± 0.0045	0.6019 ± 0.0074	0.0223 ± 0.0044	0.0086 ± 0.0015
ProtoNet	0.7903 ± 0.0084	0.1426 ± 0.0116	0.1187 ± 0.0090	0.6740 ± 0.0070	0.0812 ± 0.0087	0.0436 ± 0.0053
FEAT	0.8020 ± 0.0081	0.1427 ± 0.0120	0.1080 ± 0.0077	0.6709 ± 0.0079	0.0718 ± 0.0101	0.0349 ± 0.0068
CTML	0.5960 ± 0.0113	0.0985 ± 0.0098	0.0829 ± 0.0072	0.5550 ± 0.0072	0.0102 ± 0.0014	0.0049 ± 0.0008
MeLU	0.7070 ± 0.0080	0.0600 ± 0.0089	0.0355 ± 0.0052	0.5932 ± 0.0068	0.0264 ± 0.0050	0.0060 ± 0.0009
TaNP	0.6093 ± 0.0093	0.0243 ± 0.0061	0.0092 ± 0.0019	0.5608 ± 0.0076	0.0134 ± 0.0021	0.0067 ± 0.0012
EDGE	**0.8608 ± 0.0069***	**0.2251 ± 0.0139***	**0.1907 ± 0.0126***	**0.7275 ± 0.0066***	**0.1254 ± 0.0102***	**0.0803 ± 0.0075***

3 Experiments

We evaluate EDGE on two datasets: **MIMIC-IV** (Johnson et al., 2016) are captured from the ICU or the emergency department at the Beth Israel Deaconess Medical Center between 2008 to 2019; **Claims** are sampled from payers and healthcare providers across the United States from 2015 to 2019. There are 49 new drugs in MIMIC-IV and 99 in Claims. We randomly generate 1000 episodes for the new drugs. Each episode contains a randomly sampled test drug, a support set consisting of 5 positive and 25 negative records for this drug, and a query set consisting of all the rest of testing records (over 15K for MIMIC-IV data and 29K for claims data). Each episode is a binary classification task: whether to prescribe the drug to query record or not. Detailed experiment settings and additional results can be found in the appendix.

3.1 Main Results

Table 1 evaluates EDGE on the performance of recommendations on new drugs. First, we observe that existing drug recommendation methods perform poorly for this task due to limited training examples. Among them, SafeDrug (Yang et al., 2021), achieves slightly better performance than RNN (Choi et al., 2016b) and GAMENet (Shang et al., 2019b), probably due to the usage of drug molecule information. General few-shot learning achieves inconsistent performance. Metric-based methods ProtoNet (Snell et al., 2017) and FEAT (Ye et al., 2021) performance better compared to optimization-based approaches MAML (Finn et al., 2017) and CTML (Peng & Pan, 2022). This may be due to the challenge in imbalanced classification setting, where the number of negatives is much larger than that of positives, and the optimization-based approach might get trapped in the local minimum. Among general recommendation methods, MeLU (Lee et al., 2019) and TaNP (Lin et al., 2021) do not perform very well for this task despite their strong performance reported in Tan et al. (2022a). We suspect that they adapt the model based on patient history for rare disease diagnoses. While the setting is different in our case, we need to adapt the model based on the drug (i.e., item) prescription history. Lastly, EDGE achieves the best performance in all metrics on both datasets. Specifically, compared to the best baseline, on MIMIC-IV, EDGE achieves 5.9% absolute improvements on ROC-AUC, 8.2% on Precision@100, and 7.2% on Recall@100; for claims data, EDGE achieves 5.3% on ROC-AUC, 4.4% on Precision@100, and 3.7% on Recall@100.

4 Conclusion

As new drugs get approved and come to the market, drug recommendation models trained on existing drugs can quickly become outdated. In this paper, we reformulate the task of drug recommendation for new drugs, which is largely ignored by prior works. We propose a meta-learning framework to solve the problem by modeling the complex relations among diseases and drugs, and eliminating the numerous false-negative patients with an external knowledge base. We evaluate EDGE on two medical datasets and show superior performance compared to all baselines.

References

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. Publicly available clinical bert embeddings, 2019. URL https://arxiv.org/abs/1904.03323.

Han Altae-Tran, Bharath Ramsundar, Aneesh S Pappu, and Vijay Pande. Low data drug discovery with one-shot learning. *ACS central science*, 3(4):283–293, 2017.

Benjamin E Blass. *Basic principles of drug discovery and development*. Academic Press, San Diego, CA, 2 edition, June 2021.

Maria Brbić, Marinka Zitnik, Sheng Wang, Angela O Pisco, Russ B Altman, Spyros Darmanis, and Jure Leskovec. Mars: discovering novel cell types across heterogeneous single-cell experiments. *Nature methods*, 17(12):1200–1206, 2020.

Kaidi Cao, Maria Brbic, and Jure Leskovec. Concept Learners for Few-Shot Learning, March 2021. URL http://arxiv.org/abs/2007.07375. arXiv:2007.07375 [cs, stat].

Edward Choi, Mohammad Taha Bahadori, Joshua A. Kulas, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 3512–3520, Red Hook, NY, USA, 2016a. Curran Associates Inc. ISBN 9781510838819.

Edward Choi, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, 24(2):361–370, 08 2016b. ISSN 1067-5027. doi: 10.1093/jamia/ocw112. URL https://doi.org/10.1093/jamia/ocw112.

Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. GRAM: Graph-based Attention Model for Healthcare Representation Learning, April 2017. URL http://arxiv.org/abs/1611.07012. arXiv:1611.07012 [cs, stat].

Edward Choi, Zhen Xu, Yujia Li, Michael W. Dusenberry, Gerardo Flores, Yuan Xue, and Andrew M. Dai. Learning the graphical structure of electronic health records with graph convolutional transformer, 2019. URL https://arxiv.org/abs/1906.04716.

Manqing Dong, Feng Yuan, Lina Yao, Xiwei Xu, and Liming Zhu. Mamo: Memory-augmented meta-optimization for cold-start recommendation, 2020. URL https://arxiv.org/abs/2007.03183.

Yuntao Du, Xinjun Zhu, Lu Chen, Ziquan Fang, and Yunjun Gao. Metakg: Meta-learning on knowledge graph for cold-start recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

Ruben G. Duijnhoven, Sabine M. J. M. Straus, June M. Raine, Anthonius de Boer, Arno W. Hoes, and Marie L. De Bruin. Number of patients studied prior to approval of new medicines: A database analysis. *PLOS Medicine*, 10(3):1–8, 03 2013. doi: 10.1371/journal.pmed.1001407. URL https://doi.org/10.1371/journal.pmed.1001407.

U.S. Food and Drug Administration FDA. Novel drug approvals for 2022, 2022. URL https://www.who.int/standards/classifications/classification-of-diseases.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1126–1135. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/finn17a.html.

Healthcare Cost & Utilization Project H. CUP. Clinical classifications software (ccs), 2010. URL https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp.

A Johnson, L Bulgarelli, T Pollard, S Horng, LA Celi, and R Mark. Mimic-iv (version 1.0), 2020.

Alistair E.W. 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):160035, December 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.35. URL `http://www.nature.com/articles/sdata201635`.

Hung Le, Truyen Tran, and Svetha Venkatesh. Dual memory neural computer for asynchronous two-view sequential learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*, KDD '18, pp. 1637–1645, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355520. doi: 10.1145/3219819.3219981. URL `https://doi.org/10.1145/3219819.3219981`.

Hoyeop Lee, Jinbae Im, Seongwon Jang, Hyunsouk Cho, and Sehee Chung. MeLU: Meta-Learned User Preference Estimator for Cold-Start Recommendation. *arXiv:1908.00413 [cs]*, July 2019. URL `http://arxiv.org/abs/1908.00413`. arXiv: 1908.00413.

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*, 10(1):7155, December 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-62922-y. URL `http://www.nature.com/articles/s41598-020-62922-y`.

Xixun Lin, Jia Wu, Chuan Zhou, Shirui Pan, Yanan Cao, and Bin Wang. Task-adaptive Neural Process for User Cold-Start Recommendation, February 2021. URL `http://arxiv.org/abs/2103.06137`. arXiv:2103.06137 [cs].

Yunan Luo, Jianzhu Ma, Xiaoming Zhao, Yufeng Su, Yang Liu, Trey Ideker, and Jian Peng. Mitigating data scarcity in protein binding prediction using meta-learning. In *RECOMB*, pp. 305–307. Springer, 2019.

Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms, 2018. URL `https://arxiv.org/abs/1803.02999`.

Boris N. Oreshkin, Pau Rodriguez, and Alexandre Lacoste. TADAM: Task dependent adaptive metric for improved few-shot learning, January 2019. URL `http://arxiv.org/abs/1805.10123`. arXiv:1805.10123 [cs, stat].

Danni Peng and Sinno Pan. Clustered task-aware meta-learning by learning from learning paths, 2022. URL `https://openreview.net/forum?id=hk3Cxc2laT-`.

Yeping Lina Qiu, Hong Zheng, Arnout Devos, Heather Selby, and Olivier Gevaert. A meta-learning approach for genomic survival analysis. *Nature communications*, 11(1):1–11, 2020.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86, May 2021. ISSN 2398-6352. doi: 10.1038/s41746-021-00455-y. URL `https://europepmc.org/articles/PMC8137882`.

Jiacheng Jiang Guilin Qi Rui Wu, Zhaopeng Qiu and Xian Wu. Conditional generation net for medication recommendation. In *WWW '22: The Web Conference 2022, Virtual Event, Lyon, France, April 25-29, 2022*, 2022.

Andrei A. Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=BJgklhAcK7`.

Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. Pre-training of Graph Augmented Transformers for Medication Recommendation, November 2019a. URL `http://arxiv.org/abs/1906.00346`. arXiv:1906.00346 [cs].

Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. GAMENet: Graph Augmented MEmory Networks for Recommending Medication Combination. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:1126–1133, July 2019b. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v33i01.33011126. URL `https://aaai.org/ojs/index.php/AAAI/article/view/3905`.

Yuxiang Shi, Yue Ding, Bo Chen, Yuyang Huang, Ruiming Tang, and Dong Wang. Task aligned meta-learning based augmented graph for cold-start recommendation. *arXiv preprint arXiv:2208.05716*, 2022.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical Networks for Few-shot Learning, June 2017. URL `http://arxiv.org/abs/1703.05175`. arXiv:1703.05175 [cs, stat].

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning, 2017. URL `https://arxiv.org/abs/1711.06025`.

Qiuling Suo, Jingyuan Chou, Weida Zhong, and Aidong Zhang. Tadanet: Task-adaptive network for graph-enriched meta-learning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*, KDD '20, pp. 1789–1799, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403230. URL `https://doi.org/10.1145/3394486.3403230`.

Yanchao Tan, Chaochao Chen, Carl Yang, Weiming Liu, Xiangyu Wei, Longfei Li, Jun Zhou, and Xiaolin Zheng. MetaCare++: Meta-Learning with Hierarchical Subtyping for Cold-Start Diagnosis Prediction in Healthcare Data. pp. 11, 2022a.

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. pp. 11, 2022b.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. URL `http://www.jmlr.org/papers/v9/vandermaaten08a.html`.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning, 2016. URL `https://arxiv.org/abs/1606.04080`.

Meng Wang, Mengyue Liu, Jun Liu, Sen Wang, Guodong Long, and Buyue Qian. Safe medicine recommendation via medical knowledge graph embedding. *CoRR*, abs/1710.05980, 2017. URL `http://arxiv.org/abs/1710.05980`.

Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv.*, 53(3), jun 2020. ISSN 0360-0300. doi: 10.1145/3386252. URL `https://doi.org/10.1145/3386252`.

Wei-Qi Wei, Robert M Cronin, Hua Xu, Thomas A Lasko, Lisa Bastarache, and Joshua C Denny. Development and evaluation of an ensemble resource linking medications to their indications. *Journal of the American Medical Informatics Association*, 20(5):954–961, 04 2013. ISSN 1067-5027. doi: 10.1136/amiajnl-2012-001431. URL `https://doi.org/10.1136/amiajnl-2012-001431`.

World Health Organization WHO. Anatomical therapeutic chemical (atc) classification system, 1976. URL `https://www.who.int/tools/atc-ddd-toolkit/atc-classification`.

World Health Organization WHO. International statistical classification of diseases and related health problems (icd), 1993. URL `https://www.who.int/standards/classifications/classification-of-diseases`.

Chaoqi Yang, Cao Xiao, Fenglong Ma, Lucas Glass, and Jimeng Sun. SafeDrug: Dual Molecular Graph Encoders for Safe Drug Recommendations. *arXiv:2105.02711 [cs]*, May 2021. URL `http://arxiv.org/abs/2105.02711`. arXiv: 2105.02711.

Huaxiu Yao, Long-Kai Huang, Linjun Zhang, Ying Wei, Li Tian, James Zou, Junzhou Huang, et al. Improving generalization in meta-learning via task augmentation. In *International Conference on Machine Learning*, pp. 11887–11897. PMLR, 2021a.

Huaxiu Yao, Yu Wang, Ying Wei, Peilin Zhao, Mehrdad Mahdavi, Defu Lian, and Chelsea Finn. Meta-learning with an adaptive task scheduler. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021b. URL `https://openreview.net/forum?id=MTs2adH_Qq`.

Huaxiu Yao, Ying Wei, Long-Kai Huang, Ding Xue, Junzhou Huang, and Zhenhui Jessie Li. Functionally regionalized knowledge transfer for low-resource drug discovery. *Advances in Neural Information Processing Systems*, 34:8256–8268, 2021c.

Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-Shot Learning via Embedding Adaptation with Set-to-Set Functions, June 2021. URL http://arxiv.org/abs/1812.03664. arXiv:1812.03664 [cs].

Xi Sheryl Zhang, Fengyi Tang, Hiroko H. Dodge, Jiayu Zhou, and Fei Wang. Metapred: Meta-learning for clinical risk prediction with limited patient electronic health records. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*, KDD '19, pp. 2487–2495, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362016. doi: 10.1145/3292500.3330779. URL https://doi.org/10.1145/3292500.3330779.

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*, KDD '17, pp. 1315–1324, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348874. doi: 10.1145/3097983.3098109. URL https://doi.org/10.1145/3097983.3098109.

Marinka Zitnik, Monica Agrawal, and Jure Leskovec. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13):i457–i466, 06 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty294. URL https://doi.org/10.1093/bioinformatics/bty294.

Table 2: Notations used in this paper.

Notation	Meaning
$\mathcal{M}, \mathcal{M}^{old}, \mathcal{M}^{new}$	set of all / old / new drugs
v, v_q	(query) patient record
c	disease / procedure
V	total number diseases and procedures in v
m, m_i, m_t	a single drug
$\mathcal{S}, \mathcal{S}_i, \mathcal{S}_t, \mathcal{Q}, \mathcal{Q}_i, \mathcal{Q}_t$	set of support & query set (for drug m_i / m_t)
$\mathcal{S}', \mathcal{S}'_i, \mathcal{Q}', \mathcal{Q}'_i$	set of negative support & query set (for drug m_i)
N_s, N_q	number of supports & queries
$*$	binary label
$\mathcal{L}$	loss
$\mathbf{h}$	ontology-enriched drug representation
$\alpha_{i,j}$	importance of ancestor m_j for drug m_i
$\mathbf{p}, \mathbf{p}^{(l)}, \mathbf{p}', \mathbf{p}'^{(l)}$	(l-th phenotype-driven) positive and negative prototype representation
$\mathcal{G}, \mathbf{g}^{(l)}$	l-th phenotype set and representation
$z^{(l)}, z'^{(l)}, \mathbf{z}, \mathbf{z}'$	distance to (l-th phenotype-driven) prototype of positive / negative supports
$\boldsymbol{\beta}$	drug-dependent importance weights over different phenotypes
$f_{\phi}(\cdot)$	drug recommendation model
$f_{\phi_a}(\cdot)$	drug-ancestor attention function
$f_{\phi_r}(\cdot)$	medical record embedding function
$f_{\phi_g}(\cdot)$	phenotype down-projection function
$f_{\phi_\beta}(\cdot)$	drug-dependent importance function
$\lvert \cdot \rvert$	cardinality
$\cap$	set intersection
$\emptyset$	empty set
$\sigma(\cdot)$	sigmoid function
$d(\cdot, \cdot)$	distance function
$\oplus$	concatentaion operator

A Related Work

Drug Recommendation. Drug recommendation aims at suggesting personalized medications to patients based on their specific health conditions. In practice, drug recommendation can act as a decision support system that helps doctors by supporting decision-making and reducing workload. Existing works typically formulate the problem as a multi-label classification task. LEAP (Zhang et al., 2017) proposes a sequence-to-sequence model to predict drugs given the patient's diagnoses. Later works try to model the longitudinal relations via knowledge graph (Wang et al., 2017), attention mechanism (Choi et al., 2016a), model pre-training (Shang et al., 2019a), and memory network (Le et al., 2018; Shang et al., 2019b). More recent works further incorporate more domain-related inductive bias. For example, SafeDrug (Yang et al., 2021) leverages the global and local molecule information, and 4SDrug (Tan et al., 2022b) utilizes the set-to-set (set of symptoms to set of medications) module. Despite the good performance on existing drugs, these models are no longer applicable when new drugs appear. That is why we study new drug recommendation problem in this paper.

Few-Shot Learning. Few-shot learning aims at quickly generalizing the model to new tasks with a few labeled samples (Wang et al., 2020). Existing works can be categorized into metric-learning based approaches that aims to establish similarity or dissimilarity between classes (Vinyals et al., 2016; Sung et al., 2017; Snell et al., 2017; Oreshkin et al., 2019; Cao et al., 2021; Ye et al., 2021), and optimization based approach seeks to learn a good initialization point that can adapt to new tasks within a few parameter updates (Finn et al., 2017; Nichol et al., 2018; Rusu et al., 2019; Yao et al., 2021b; Peng & Pan, 2022). For recommendation problem, few-shot learning is leveraged to solve the cold-start problem, where the goal is to quickly adapt the model to new users with limited history (Lee et al., 2019; Du et al., 2022; Dong et al., 2020; Lin et al., 2021; Shi et al., 2022). For healthcare problem, few-shot learning is applied to improve diagnosis of uncommon diseases (Zhang et al., 2019; Suo et al., 2020; Tan et al., 2022a), drug discovery (Altae-Tran et al., 2017; Yao et al., 2021a,c; Luo et al., 2019), novel cell type classification (Brbić et al., 2020), and genomic survival analysis (Qiu et al., 2020). However, when it comes to new drug recommendations, these works fail to capture the complex relationship between diseases and drugs, and will also be largely influenced by the noisy supervision signal.

B Detailed Experimental Settings

B.1 Dataset

MIMIC-IV (Johnson et al., 2020) covers more than 382K patients admitted to the ICU or the emergency department at the Beth Israel Deaconess Medical Center between 2008 to 2019. It provides the contemporary information which allows us to locate the year of each patient admission. The original data provides the timing information by a three year long date range (e.g., 2011-2013). We uniformly sample a year in the provided range as the admission year. We select our cohort by filtering out the following samples: (1) patients younger than 18-year-old; (2) admissions without disease or procedure; (3) admissions without medication. We filter out drugs with less than 20 admissions. We split the drugs by the time they were first prescribed. We use drugs first appeared in 2008 as the training (existing) drugs, drugs appeared in 2009 as the validation drugs, and drugs appeared after 2010 as the testing (new) drugs. The admissions are split accordingly: if an admission contains any test/validation/training drug, it will be in test/validation/training set. If an admission contains drugs from more than one split, priority is given to test, validation, and training. There are 49 new drugs in total.

Claims are captured from payers and healthcare providers across the United States. We randomly sample 30K patients and select all their claims from 2015 to 2019, resulting in 691K claims. We split the drugs into 70%/10%/20% training/validation/test sets. The admissions are split the same way as MIMIC-IV. There are 99 new drugs in total.

Table 3: Dataset statistics. We report the statistics for training / validation / testing separately.

Dataset	# Patients	# Admissions	# Drugs	# Diseases	# Procedures
MIMIC-IV	119K / 2K / 12K	216K / 3K / 15K	786 / 15 / 49	22K	12K
Claims	10K / 3K / 9K	37K / 6K / 29K	244 / 36 / 99	11K	-

B.2 Baselines

We compare EDGE to three categories of baselines. (1) The first category is **drug recommendation** methods, including RNN (Choi et al., 2016b), GAMENet (Shang et al., 2019b), SafeDrug (Yang et al., 2021). For a fair comparison, the drug recommendation is first trained on the training set of existing drugs, and then fine-tuned on the support set of new drugs. (2) The second category is **few-shot learning** approaches, including MAML (Finn et al., 2017), ProtoNet (Snell et al., 2017), FEAT (Ye et al., 2021), CTML (Peng & Pan, 2022). (3) Lastly, we include the **cold-start recommendation** methods, including MeLU (Lee et al., 2019) and TaNP (Lin et al., 2021). See below for details of all baselines.

- **RNN (Choi et al., 2016b)** formulates the task as a multi-label sequence prediction task. The patient record is encoder into vector representation and then make the prediction using a fully connected network.
- **GAMENet (Shang et al., 2019b)** adopts a memory network to model the drug-drug interaction information and to memorize the historical patient condition.
- **SafeDrug (Yang et al., 2021)** leverages dual molecular encoders to capture the global and local molecule patterns.
- **MAML (Finn et al., 2017)** tries to learn good initialization parameters such that it can adapt to new drugs with a few gradient updates.
- **ProtoNet (Snell et al., 2017)** is a metric learning based method which embeds the support and query samples into hidden space such that the samples belong to the same class are closer to each other.

Table 4: Additional results on new drug recommendation.

Method		MIMIC-IV			Claims	
	PR-AUC	Precision@500	Recall@500	PR-AUC	Precision@500	Recall@500
RNN	0.0288 ± 0.0057	0.0282 ± 0.0064	0.0243 ± 0.0017	0.0138 ± 0.0015	0.0138 ± 0.0014	0.0242 ± 0.0017
GAMENet	0.0727 ± 0.0076	0.0766 ± 0.0091	0.1913 ± 0.0121	0.0293 ± 0.0042	0.0286 ± 0.0043	0.0619 ± 0.0073
SafeDrug	0.0961 ± 0.0079	0.0887 ± 0.0079	0.2293 ± 0.0143	0.0135 ± 0.0014	0.0128 ± 0.0016	0.0134 ± 0.0009
MAML	0.0590 ± 0.0064	0.0543 ± 0.0059	0.1766 ± 0.0131	0.0234 ± 0.0036	0.0231 ± 0.0042	0.0415 ± 0.0051
ProtoNet	0.1164 ± 0.0098	0.0977 ± 0.0088	0.3371 ± 0.0197	0.0495 ± 0.0056	0.0572 ± 0.0061	0.1154 ± 0.0089
FEAT	0.0992 ± 0.0044	0.0873 ± 0.0061	0.3123 ± 0.0045	0.0473 ± 0.0023	0.0529 ± 0.0042	0.0962 ± 0.0025
CTML	0.0973 ± 0.0100	0.0871 ± 0.0100	0.2571 ± 0.0157	0.0137 ± 0.0012	0.0111 ± 0.0011	0.0249 ± 0.0022
MeLU	0.0577 ± 0.0074	0.0504 ± 0.0066	0.1245 ± 0.0102	0.0224 ± 0.0030	0.0231 ± 0.0038	0.0297 ± 0.0027
TaNP	0.0368 ± 0.0059	0.0280 ± 0.0042	0.0636 ± 0.0060	0.0170 ± 0.0015	0.0140 ± 0.0017	0.0275 ± 0.0034
EDGE	**0.1940 ± 0.0130***	**0.1459 ± 0.0107***	**0.4462 ± 0.0187***	**0.0781 ± 0.0068***	**0.0871 ± 0.0078***	**0.1975 ± 0.0121***

- **FEAT (Ye et al., 2021)** further proposes to use a set-to-set function to encode the support prototype instead of simple average.
- **CTML (Peng & Pan, 2022)** learns the task-specific representation from both support samples and the learning path.
- **MeLU (Lee et al., 2019)** builds upon the MAML (Finn et al., 2017) framework and can adapt to the new drugs with a few local updates.
- **TaNP (Lin et al., 2021)** incorporates a customization module which adapt the predictor parameters based on different tasks.

B.3 Evaluation Metrics

We randomly generate 1000 episodes for the test drugs. Each episode contains a randomly sampled test drug, a support set consisting of 5 positive and 25 negative records for this drug, and a query set consisting of all the rest of testing records (over 15K for MIMIC-IV data and 29K for claims data). Each episode is a binary classification task: whether to prescribe the drug to query record or not. We calculate the Area Under Receiver Operator Characteristic Curve (ROC-AUC), Area Under Precision-Recall Curve (PR-AUC), Precision@K, Recall@K. For each metric, we also report the 95% confidence interval calculated from the 1000 episodes. We also perform the independent two-sample t-test to evaluate if EDGE achieves significant improvement over baseline methods. Due to space limitations, we only show ROC-AUC, Precision@100, and Recall@100 in the main paper and leave other metrics in the appendix.

B.4 Implementation Details

We use ICD codes (WHO, 1993) to represent disease and procedure, and ATC-5 level codes (WHO, 1976) to represent medication. CCS category (H. CUP, 2010) is used to define the phenotype for diseases and procedures. ATC ontology WHO (1976) is used to build the drug ontology. We obtain the basic code embeddings by encoding code description with Clinical-BERT (Alsentzer et al., 2019). For all methods, we use the bi-directional GRU as the backbone encoder, with 768 as input dimension and 512 as the output dimension. We set the phenotype vector dimension g as 64. All models are trained for 100K episodes with 5 positive and 250 negative supports. We select the best model by monitoring the ROC-AUC score on the validation set. We use Adam as the optimizer with learning rate 1e-3. Linear warmup scheduler is used with the first 10% episodes as the warm-up episodes. Dropout is set to 0.5. We implement EDGE using PyTorch 1.11 and Python 3.8. The model is trained on a CentOS Linux 7 machine with 128 AMD EPYC 7513 32-Core Processors, 512 GB memory, and eight NVIDIA RTX A6000 GPUs.

C Additional Metrics for Main Results

See table 4 for additional metrics for the main results in table 1.

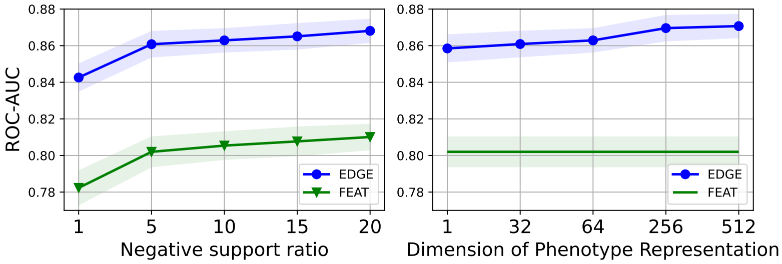

Figure 2: The effect of negative support ratio and dimension of phenotype representation.

D Quantitative Analysis of Performance Gains

This section presents quantitative analyses to understand the performance gains of EDGE.

D.1 Analysis on False Negatives

Table 5: Analysis on the influence of false negatives. Prev.: prevalence; Orig.: original; Adj: adjusted.

Dataset	Prev.	Orig. P@100	Adj. P@100
MIMIC-IV	31.5%	0.2251	0.4053
Claims	22.1%	0.1254	0.3317

Due to the one-to-many mapping between disease and drug, there exist many false-negative patients for a given drug. For each drug, we calculate the prevalence of false-negative records (i.e., percentage of records that could but did not use the drug) using the MEDI (Wei et al., 2013) drug-disease database. We report the average prevalence across all new drugs in table 5. The false-negative records range from 20 to 30% for both datasets. This level of noise can largely confuse the model if not appropriately handled. We then correct the false-negative labels (i.e., swap them from negative to positive) in the test query set, and re-calculate the adjusted Precision@100 for EDGE, which increases by 18.0% on MIMIC-IV data and by 20.6% on claims data.

D.2 Ablation Study of Module Importance

Table 6: Ablation study.

Method	MIMIC-IV			Claims		
	ROC-AUC	P@100	R@100	ROC-AUC	P@100	R@100
Remove ontology-enhanced drug representation	0.8239	0.1584	0.1411	0.6809	0.0985	0.0672
Remove multi-phenotype patient representation	0.7992	0.1555	0.1312	0.6910	0.0818	0.0551
Remove drug-dependent distance measure	0.8033	0.1687	0.1542	0.7078	0.1055	0.0758
Remove knowledge-guided negative sampling	0.8175	0.1899	0.1725	0.7149	0.1116	0.0789
EDGE	**0.8608**	**0.2251**	**0.1907**	**0.7275**	**0.1254**	**0.0803**

In the ablation study, we evaluate the contribution of each proposed module. According to the results in table 6, we observe that the multi-phenotype patient encoder contributes the most: switching it with a single vector patient representation will decrease the ROC-AUC by 16.9%. Removing drug-dependent patient distance also decreases the performance by a large margin, as the model cannot adapt the metric space to a specific drug. The ontology-enhanced drug representation and knowledge-guided negative sampling both slightly improve the model performance. Changing ontology-enhanced drug representation to the basic drug representation obtained from the drug name and description ignores the high-level category information shared between existing and new drugs. Changing knowledge-guided negative sampling to uniform random sampling might confuse the model with the prevalent false negatives.

D.3 Analysis on the Influence of Hyper-parameters

We evaluate the effect of two important hyper-parameters: the ratio of negative supports and the dimension of the phenotype vector on MIMIC-IV data. The results can be found in figure 2, where the results of both EDGE and FEAT (best baseline) are reported. The performance increases as the ratio of negative support records or the dimension of phenotype representation. Interestingly, we find the EDGE can outperform the best baseline method even if we set the dimension of phenotype to 1. That is to represent each patient with a single vector (the same as the prototypical network). This confirms the benefits of leveraging phenotype knowledge and drug ontology.

D.4 Analysis of Different Backbone Encoders

Table 7: Analysis on the influence of different backbone encoders.

Method	ROC-AUC	P@100	R@100
EDGE + MLP	0.7961 ± 0.0065	0.1258 ± 0.0100	0.0838 ± 0.0057
EDGE + Transformer	0.8417 ± 0.0073	0.1684 ± 0.0117	0.1482 ± 0.0109
EDGE + GRU	0.8418 ± 0.0073	0.2024 ± 0.0141	0.1779 ± 0.0119
EDGE + Bi-GRU	0.8608 ± 0.0069	0.2251 ± 0.0139	0.1907 ± 0.0126

We further incorporate EDGE with different backbone encoders, ranging the MLP to Transformer and single layer GRU. Results can be found in table 7. MLP simply feeds the disease through a feed-forward neutral network and then take the average. It gives the lowest result as it cannot capture the relationship among disease. Transformer leverages the attention mechanism and gets better results. However, it cannot outperform GRU and Bi-GRU. We suspect the main reason is that we skip the pre-training step for the Transformer. Theoretically, the performance of Transformer will be largely improved if we pre-train it via self-supervised learning and fine-tune it in our task, as shown in Li et al. (2020); Rasmy et al. (2021); Choi et al. (2019). We leave Transformer pre-training to future work.

D.5 Analysis of Different Distance Measures

Table 8: Analysis of the influence of different distance measures.

Method	ROC-AUC	P@100	R@100
EDGE + Cosine distance	0.8610 ± 0.0068	0.2118 ± 0.0135	0.1832 ± 0.0124
EDGE + Euclidean distance	0.8608 ± 0.0069	0.2251 ± 0.0139	0.1907 ± 0.0126

We test EDGE with two distance measures: cosine distance and euclidean distance. The results can be found in table 8. The two distance measures perform similarly on the MIMIC-IV dataset.

E Qualitative Analysis

This section presents qualitative analysis to further investigate the performance of different drugs and the learned importance scores and representations.

E.1 Analysis on Per-Category Performance

We first group new drugs by ATC-2 level and calculate the per-category performance of EDGE on MIMIC-IV data. We report the top-5 and bottom-5 categories in table 9. From the per-category result, the best-performing drugs are typically more targeting specific diseases. For example, *antineoplastic agents* are medications used to treat *cancer*, and *antivirals* are drugs targeting specific virals such as *influenza* and *rabies*. The support patients for such

Table 9: Performance by drug category.

Drug Category	ROC-AUC
Antineoplastic agents	0.9882
Antivirals	0.9754
Urologicals	0.9738
Anti-parkinson drugs	0.9611
Other alimentary tract products	0.9412
...	...
Antibacterials	0.7888
All other therapeutic products	0.7827
Antiinfective	0.7367
Calcium channel blockers	0.7285
Beta-adrenergic blockers	0.6074

drugs usually have specific diseases that act as a prominent indicator for the prescription of the drugs. On the other hand, the worst performing drugs are usually very general, such as *antibacterials* or *antiinfectives*, which applies to various circumstances. The limited number of support patients cannot capture the broad applications of these drugs, which leads to lower performance.

E.2 Analysis of the Importance of Phenotypes

We show three example drugs from the testing set with the top-3 phenotypes ranked by the learned importance scores. The result can be found in table 10. The top-ranked phenotypes match well with the usage of each drug. For example, for *dactinomycin* which treats a variety of *cancers*, the top-ranked phenotypes are all cancer-related. Interestingly, EDGE also learns to pay attention to some common comorbidities for a given disease. For example, for the drug *rasagiline* which is used for *Parkinson's disease*, EDGE also attends to *hypertension* and *anemia*, which are often seen in patients with *Parkinson's disease*.

Table 10: Top-3 phenotypes ranked by the learned importance scores.

Drug	**Usage**	**Top-3 Phenotypes**
Dactinomycin	Cancer	Other cancer Lung Cancer Neoplasms
Nebivolol	Hypertension	Essential hypertension Circulatory disease Surgical procedures
Rasagiline	Parkinson	Parkinson Hypertension Anemia

E.3 Analysis on Drug Representation

For qualitative analysis, we visualize the learned ontology-enhanced drug representation on MIMIC-IV data with t-SNE (van der Maaten & Hinton, 2008). As shown in figure 3, the representations for drugs with similar treatment effects cluster into smaller categories. Further, the representations for new drugs also align with existing drugs. This indicates that the model adapts to new drugs by linking the new drugs to existing drugs with similar treatment effects.

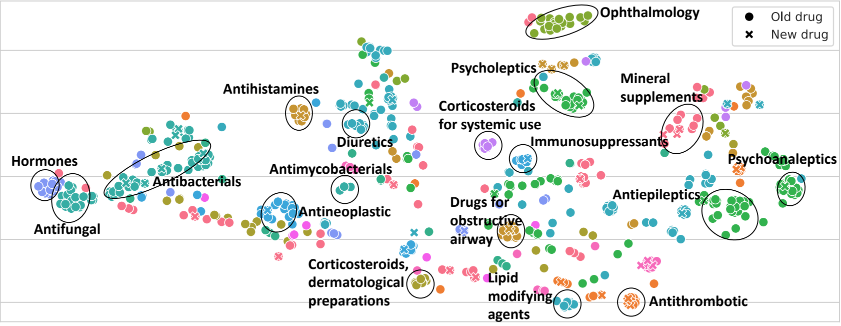

Figure 3: t-SNE plot of the learned ontology-enhanced drug representation. Node color represents ATC-2 level (coarse) categories while circle represents ATC-4 (fine-grained) level categories.