Hyperbolic Molecular Representation Learning for Drug Repositioning

Ke Yu
University of Pittsburgh
yu.ke@pitt.edu

Shyam Visweswaran
University of Pittsburgh
shv3@pitt.edu

Kayhan Batmanghelich
University of Pittsburgh
kayhan@pitt.edu

Abstract

Learning accurate drug representations is essential for task such as computational drug repositioning. A drug hierarchy is a valuable source that encodes knowledge of relations among drugs in a tree-like structure where drugs that act on the same organs, treat the same disease, or bind to the same biological target are grouped together. However, its utility in learning drug representations has not yet been explored, and currently described drug representations cannot place novel molecules in a drug hierarchy. Here, we develop a semi-supervised drug embedding that incorporates two sources of information: (1) underlying chemical grammar that is inferred from chemical structures of drugs and drug-like molecules (unsupervised), and (2) hierarchical relations that are encoded in an expert-crafted hierarchy of approved drugs (supervised). We use the Variational Auto-Encoder (VAE) framework to encode the chemical structures of molecules and use the drug-drug similarity information obtained from the hierarchy to induce the clustering of drugs in hyperbolic space. The hyperbolic space is amenable for encoding hierarchical relations. Our qualitative results support that the learned drug embedding can induce the hierarchical relations among drugs. We demonstrate that the learned drug embedding can be used for drug repositioning.

1 Introduction

The study of drug representation provides the foundation for computational drug repositioning. Drug repositioning, the process of finding new uses for existing drugs, is one strategy to shorten the time and reduce the cost of drug development [1]. Computational methods of drug repositioning typically aim to identify shared mechanism of actions among drugs that imply that the drugs may also share therapeutic applications [2]. However, such methods are limited when prior knowledge of drugs may be scarce or not available; for example, drugs that are in the experimental phase or have failed clinical trials. Therefore, it is appealing to map the chemical structure of a molecule to its pharmacological behavior.

A drug hierarchy encodes a broad spectrum of known drug relations. For example, a widely used drug hierarchy, Anatomical Therapeutic Chemical Classification System (ATC), groups drugs that are similar in terms of their mechanism of action and therapeutic, pharmacological and chemical characteristics. But its utility in learning drug representation has not yet been explored.

Here, we develop a drug embedding that integrates the chemical structures of drugs and drug-like molecules with a drug hierarchy such that the similarity between pairs of drugs is informed both

Machine Learning for Molecules Workshop at NeurIPS 2020. https://ml4molecules.github.io

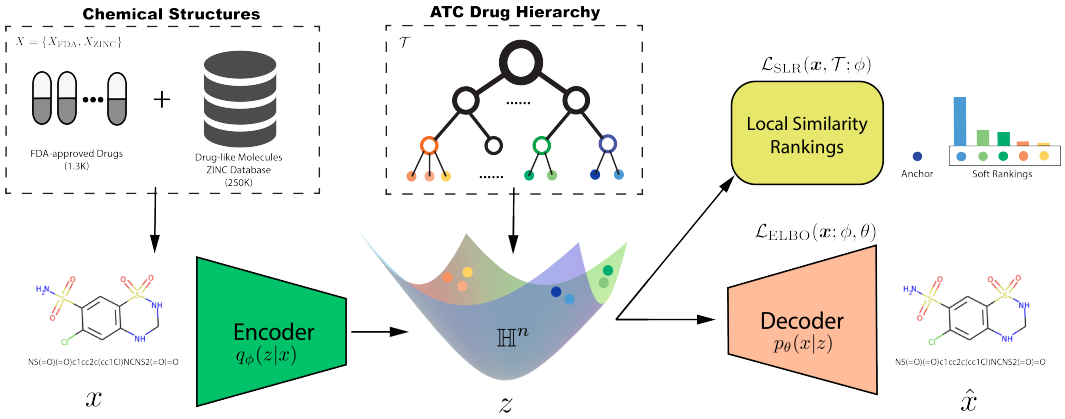

Figure 1: Schematic diagram of the proposed drug embedding method. Our semi-supervised learning approach integrates the chemical structures of a small number of FDA-approved drug molecules (X_{FDA}) and a larger number of drug-like molecules (X_{ZINC}) drawn from the ZINC database. We use VAE to encode molecules in hyperbolic space $\mathbb{H}^n$, and enforce the ATC drug hierarchy by preserving local similarity rankings of drugs. The symbols $\boldsymbol{x}$, $\boldsymbol{z}$, $\hat{\boldsymbol{x}}$ denote a molecule represented by its SMILES string, its embedding and its reconstruction; $q_\phi(\boldsymbol{z}|\boldsymbol{x}), p_\theta(\boldsymbol{x}|\boldsymbol{z})$ denote the encoder network and the decoder network respectively; $\mathcal{L}_{\text{ELBO}}(\boldsymbol{x}; \phi, \theta)$, $\mathcal{L}_{\text{SLR}}(\boldsymbol{x}, \mathcal{T}; \phi)$ denote the objective functions for the VAE and the local similarity rankings.

by the structure and groupings in the hierarchy (Figure 1). To learn the underlying grammar of chemical structures, we leverage a data set of drugs (about 1.3K) that are approved by the Food and Drug Administration (FDA) and a larger data set of drug-like molecules (about 250K) and use the simplified molecular-input line-entry system (SMILES) [3] structure representation. We obtain drug similarity relationships from the ATC drug hierarchy. We use the hyperbolic space for the embedding since it is amenable for learning continuous concept hierarchies [4, 5, 6, 7].

2 Method

Learning chemical grammar using VAE We use a Variational Auto-Encoder (VAE) to encode the chemical structure of molecules. More specifically, we model a molecule as a random variable generated by encoding a SMILES string through a encoder $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ into a code ($\boldsymbol{z}$), which is then decoded back to a reconstruction of the input by passing through a decoder $p_\theta(\boldsymbol{x}|\boldsymbol{z})$. In the classic VAE, the prior $p(\boldsymbol{z})$ is the standard normal distribution, the encoder $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ is modeled by a Gaussian distribution $\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_\phi, \Sigma_\phi)$. To extend VAE from a flat Euclidean space to a curved manifold, the Gaussian distribution needs to be extended to the hyperbolic space.

In this study, we adopt the so-called wrapped normal distribution proposed by Nagano et al., 2019 [8], which we denote by $\mathcal{N}_{\mathbb{H}}^{\text{W}}(\boldsymbol{z}|\boldsymbol{\mu}, \Sigma)$, where $\mathbb{H}^n$ is the Lorentz model of a n-dimensional hyperbolic space, $\boldsymbol{z} \in \mathbb{H}^n$, and $\boldsymbol{\mu}$ is the hyperbolic mean. The reparameterization trick in the hyperbolic VAE can be viewed as the composition of two operations $\exp_{\boldsymbol{\mu}}(\text{PT}_{\boldsymbol{\mu}_0 \to \boldsymbol{\mu}}(\boldsymbol{u}))$, where $\text{PT}_{\boldsymbol{\mu}_0 \to \boldsymbol{\mu}}(\boldsymbol{u})$ is the so-called *parallel transport*, $\exp_{\boldsymbol{\mu}}$ is the so-called *exponential map*, and $\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0}, \Sigma)$ is a random vector sampled from normal distribution. The inside operation $\text{PT}_{\boldsymbol{\mu}_0 \to \boldsymbol{\mu}}(\boldsymbol{u})$ shifts the tangent space, a linear approximation of the manifold around a point, from $\boldsymbol{\mu}_0$ to $\boldsymbol{\mu}$ analogous to the addition operation in the classic reparameterization trick. The $\exp_{\boldsymbol{\mu}}$ projects the shifted vector to the manifold. Note that, in the Lorentz model, both the *parallel transport* and the *exponential map* have analytical forms, and can be differentiated with respect to the hyperbolic mean $\boldsymbol{\mu}$ of the wrapped normal distribution $\mathcal{N}_{\mathbb{H}}^{\text{W}}(\boldsymbol{z}|\boldsymbol{\mu}, \Sigma)$. We compute the KL-divergence following the derivation in [8].

Integrating hierarchical knowledge The hyperbolic VAE learns an embedding for codes that are amenable to hierarchical representation. However, it only models $\boldsymbol{x}$ (the SMILES string of the drug), and it does not enforce our prior knowledge about drug hierarchy which defines similarity or

dissimilarity between drugs at various levels. Inspired by concept embedding in hyperbolic space [9], we incorporate the ATC hierarchy $\mathcal{T}$ in our model by using pairwise similarity between drugs. Let $t_{i,j}$ denote the path-length between two drugs, $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in $\mathcal{T}$, and let $\mathcal{D}(i,j) = \{k : t_{i,j} < t_{i,k}\} \cup \{j\}$ denote the set of drugs with path-lengths equal to or greater than $t_{i,j}$. We define the soft local ranking with respect to the anchor drug $\boldsymbol{x}_i$ as:

$$p(\boldsymbol{x}_i, \boldsymbol{x}_j; \phi) = \frac{\exp(-d_\ell(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j))}{\sum_{k \in \mathcal{D}(i,j)} \exp(-d_\ell(\boldsymbol{\mu}_i, \boldsymbol{\mu}_k))} \tag{1}$$

where $\boldsymbol{\mu}_i$ is the hyperbolic mean of $q_\phi(\boldsymbol{z}|\boldsymbol{x}_i) = \mathcal{N}_{\mathbb{H}}^{\mathrm{W}}(\boldsymbol{z}|\boldsymbol{\mu}_i, \Sigma_i)$ and $d_\ell(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)$ is the hyperbolic distance between $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$. The likelihood function of the soft local rankings is given by $\mathcal{L}_{\mathrm{SLR}}(\boldsymbol{x}_i, \mathcal{T}; \phi) = \sum_j \log p(\boldsymbol{x}_i, \boldsymbol{x}_j; \phi)$.

Note that the global hierarchy of $\mathcal{T}$ is decomposed into local rankings denoted by $\mathcal{D}(i,j) = \{k : t_{i,j} < t_{i,k}\} \cup \{j\}$. To train our model, we need to effectively sample $\mathcal{D}(i,j) \sim \mathcal{T}$, and the best sampling strategy supported by the empirical results of ablation study is as follows. For each anchor drug $\boldsymbol{x}_i$, we uniformly sample a positive example $\boldsymbol{x}_j$, such that the lowest common ancestor of $\boldsymbol{x}_i$, $\boldsymbol{x}_j$ has equal chance of being an internal node at any level, i.e., level 1, 2, 3, or 4, in the ATC tree. We then randomly sample k negative examples $\boldsymbol{x}_k$ from other leaf nodes that have greater path lengths than $t_{i,j}$.

Optimization We employ a semi-supervised learning approach that combines a small number of FDA-approved drugs X_{FDA} with a larger number of drug-like molecules X_{ZINC}. We upsample X_{FDA} to 20% in mini-batch to enhance the signal of the supervised learning task, which is maximizing the likelihood of the soft local rankings with respect to the ATC hierarchy $\mathcal{T}$. The unsupervised learning task is to maximize the variational evidence lower bound (ELBO) [10] of the marginal likelihood of the chemical structures of drugs and drug-like molecules $X = \{X_{\mathrm{ZINC}}, X_{\mathrm{FDA}}\}$. We then formulate the drug embedding problem as:

$$\underset{\phi,\theta}{\operatorname{argmax}} \Big(\mathcal{L}_{\mathrm{ELBO}}(\boldsymbol{x}; \phi, \theta) + c \cdot \mathcal{L}_{\mathrm{SLR}}(\boldsymbol{x}, \mathcal{T}; \phi) \Big) \tag{2}$$

where $c = 1$ when $\boldsymbol{x} \in X_{\mathrm{FDA}}$, $c = 0$ when $\boldsymbol{x} \in X_{\mathrm{ZINC}}$, and $|X_{\mathrm{ZINC}}| \gg |X_{\mathrm{FDA}}|$. The first term in the objective function captures the underlying chemical grammar of molecules, and the second term enforces the relative positions of the drugs in the latent space to correspond to their relative positions in the ATC hierarchy. Parameters are estimated using mini-batch gradient descent and gradients are straightforward to compute using the hyperbolic reparameterization trick.

3 Experimental Results

Datasets We obtained SMILES strings of 1,365 FDA-approved drugs and SMILES strings of 250,000 drug-like molecules extracted at random by [11] from the ZINC [12] database. We combine the 1,365 drug and the 250,000 drug-like molecules to create a single data set of chemical structures that we use in our experiments. The ATC hierarchy was created by the World Health Organization (WHO) [13] that leverages the location of action, therapeutic, pharmacological and chemical properties of drugs to group them hierarchically. We obtained the ATC hierarchy from the Unified Medical Language System (UMLS) Metathesaurus (version 2019AB) and mapped the FDA-approved drugs to the terminal nodes in the ATC tree that represents the active chemical substance.

Model visualization We visually explore the embedding in two dimensional hyperbolic space by mapping the embedding in the Lorentz model to the Poincaré disk via a diffeomorphism described in [14]. In Figure 2(a), we observe that most of the drugs are placed near the boundary of the Poincaré disk and form tight clusters that correspond to the drug groups at ATC level 1. The hyperbolic embedding exhibit a clear hierarchical structure where the clusters at the boundary can be viewed as distinct substrees with the root of the tree positioned at the origin. A small number of drugs (grey circles) are scattered around the origin and denote drugs that act on the sensory organs. This group of drugs mainly consist of anti-infectives, anti-inflammatory agents, and corticosteroids, most of which act on more than one system and have multiple therapeutic uses. We hypothesize that these sensory organ drugs are placed close to the center because minimizing the local ranking loss constrains them to be concurrently close to different drug groups in the latent space. Figure 2(b) and (c) demonstrate that embedding in hyperbolic space can effectively induce a multi-level tree and the embedding retains the hierarchical structure to the deepest levels.

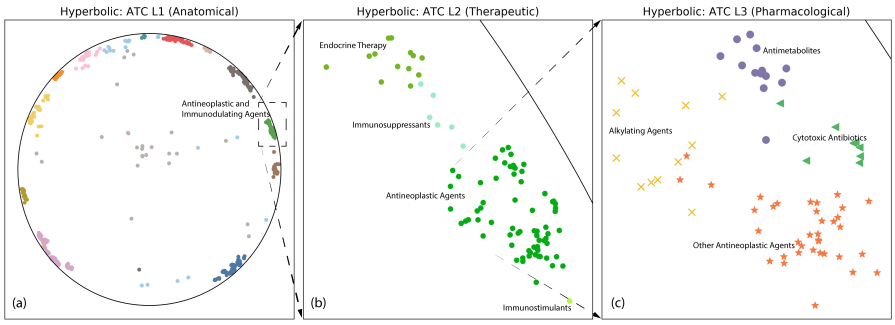

Figure 2: Visualization of hyperbolic drug embedding in two-dimensional Poincaré disk that shows drugs with colored symbols. In panel (a) drugs that belong to the same group at ATC level 1 are denoted by circles of the same color. Panel (b) shows drugs of one group from ATC level 1 namely, "Antineoplastic and Immunodulating Agents", and drugs that belong to the same group at ATC level 2 are denoted by circles with the same shade of green. Panel (c) shows drugs of one group from ATC level 2, namely, "Antineoplastic Agents", and drugs that belong to the same group at ATC level 3 are denoted by symbols of the same color.

Evaluating drug repositioning We evaluate the learned embedding for drug repositioning by deriving kNN models to discriminate between approved and unapproved drug-indication pairs in the repoDB [15] dataset, a benchmark data set that contains information on drug repositioning. We tag each drug-indication pair with the date when the drug was first approved by the FDA. We choose 2000 as the cutoff year to split the repoDB data set into training (earlier than year 2000) and test (year 2000 and later) sets. The ratio between the size of training and test data sets is about $85\% : 15\%$. For each drug $\boldsymbol{x}_i$ in the test set, we first encode it into the latent space using its SMILES string as the input, and then retrieve its k nearest neighbors $\{X_{kNN}\}$ from the training set in the latent space. We apply majority voting to the retrieved drug-indication pairs in $\{X_{kNN}\}$ to predict the status of each indication associated with $\boldsymbol{x}_i$. For indications of $\boldsymbol{x}_i$ that do not exist in $\{X_{kNN}\}$, we assume that it has an equal probability of being either being successfully approved or failed to be approved, and we assign an equal vote (0.5) to each class.

Because we are not aware of any other approach developed on the repoDB dataset with the same chronological split, we compare the performance of our drug embedding, denoted as Lorentz Drug Embedding (LDE), for drug repositioning using kNN to the following baselines: (1) kNN on RDKit-calculated [16] descriptors, (2) kNN on Morgan fingerprints (bit vector) [17], (3) kNN on count-based Morgan fingerprints, and (4) kNN on Lorentz drug embedding without ATC information. We use the Tanimoto coefficient [18] as the similarity metric for fingerprints-based representations. Performance is evaluated using area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC). Figure 3 shows that the LDE with ATC information outperforms other drug representations by a large margin. Averaging across different k values, the LDE with ATC information surpasses Morgan fingerprints, the second best representation, by 12% (AUROC) and 15.8% (AUPRC). Compared to LDE without ATC information, incorporating drug hierarchy in the embedding achieves a large gain of 33.6% (AUROC) and 48.8% (AUPRC). LDE's competitive performance on discovering repositioning opportunities are likely driven by the drug-drug similarity that is encoded in the ATC hierarchy.

4 Conclusion

We introduced a method for learning a high-quality drug embedding that integrates chemical structures of drug and drug-like molecules with local similarity of drugs implied by a drug hierarchy. We leveraged the properties of the Lorentz model of hyperbolic space and developed a novel hyperbolic VAE method that simultaneously encodes similarity from chemical structures and from hierarchical relationships. We showed qualitatively that our embedding recapitulates the hierarchical relationships in the ATC hierarchy. We showed empirically that the embedding can be used for drug repositioning.

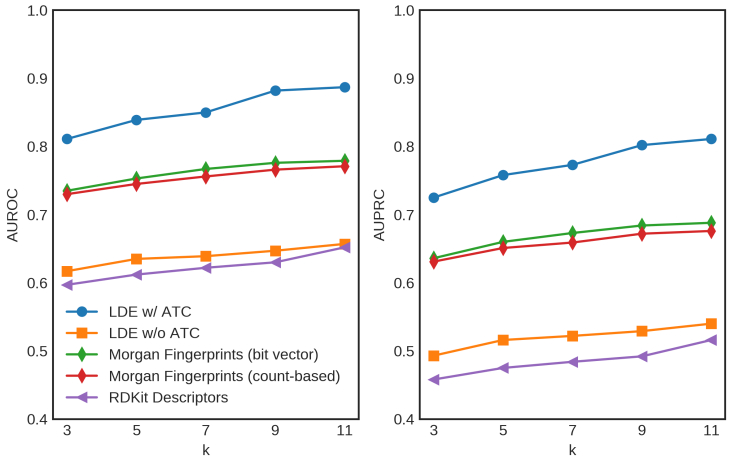

Figure 3: Comparison of representations for drug repositioning prediction using kNN ($k \in [3, 5, 7, 9, 11]$). The left panel shows AUROC scores and the right panel shows AUPRC scores.

References

[1] N. Nosengo, "New tricks for old drugs," *Nature*, vol. 534, no. 7607, pp. 314–316, 2016.

[2] S. Pushpakom, F. Iorio, P. A. Eyers, K. J. Escott, S. Hopper, A. Wells, A. Doig, T. Guilliams, J. Latimer, C. McNamee, A. Norris, P. Sanseau, D. Cavalla, and M. Pirmohamed, "Drug repurposing: progress, challenges and recommendations," *Nat. Rev. Drug Discovery*, vol. 18, no. 1, pp. 41–58, 2019.

[3] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," *J. Chem. Inf. Model.*, vol. 28, no. 1, pp. 31–36, 1988.

[4] M. Nickel and D. Kiela, "Poincaré embeddings for learning hierarchical representations," in *Advances in neural information processing systems*, 2017, pp. 6338–6347.

[5] E. Mathieu, C. Le Lan, C. J. Maddison, R. Tomioka, and Y. W. Teh, "Continuous hierarchical representations with poincaré variational auto-encoders," in *Advances in neural information processing systems*, 2019, pp. 12 544–12 555.

[6] C. De Sa, A. Gu, C. Ré, and F. Sala, "Representation tradeoffs for hyperbolic embeddings," *J. Mach. Learn. Res.*, vol. 80, p. 4460, 2018.

[7] N. Monath, M. Zaheer, D. Silva, A. McCallum, and A. Ahmed, "Gradient-based hierarchical clustering using continuous representations of trees in hyperbolic space," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 714–722.

[8] Y. Nagano, S. Yamaguchi, Y. Fujita, and M. Koyama, "A wrapped normal distribution on hyperbolic space for gradient-based learning," in *International Conference on Machine Learning*, 2019, pp. 4693–4702.

[9] M. Nickel and D. Kiela, "Learning Continuous Hierarchies in the Lorentz Model of Hyperbolic Geometry," *arXiv e-prints*, p. arXiv:1806.03417, Jun. 2018.

[10] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Mach. Learn.*, vol. 37, no. 2, pp. 183–233, 1999.

[11] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik, "Automatic chemical design using a data-driven continuous representation of molecules," *ACS Cent. Sci.*, vol. 4, no. 2, pp. 268–276, 2018.

[12] J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad, and R. G. Coleman, "Zinc: a free tool to discover chemistry for biology," *J. Chem. Inf. Model*, vol. 52, no. 7, pp. 1757–1768, 2012.

[13] W. H. Organization, "Who collaborating centre for drug statistics methodology, atc classification index with ddds," *World Health Organization Collaborating Centre for Drug Statistics Methodology*, 2014.

[14] M. Nickel and D. Kiela, "Learning continuous hierarchies in the lorentz model of hyperbolic geometry," *arXiv preprint arXiv:1806.03417*, 2018.

[15] A. S. Brown and C. J. Patel, "A standard database for drug repositioning," *Sci. Data*, vol. 4, no. 1, pp. 1–7, 2017.

[16] G. Landrum, "Rdkit: Open-source cheminformatics," 2006.

[17] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *J. Chem. Inf. Model.*, vol. 50, no. 5, pp. 742–754, 2010.

[18] D. Bajusz, A. Rácz, and K. Héberger, "Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations?" *J. Cheminf.*, vol. 7, no. 1, p. 20, 2015.